Multi-Channel Causal Variational Autoencoder

Safaa Al-Ali, Irene Balelli

Multi-Channel Causal Variational Autoencoder

Safaa Al-Ali[1,*], Irene Balelli[1]

[1]Centre Inria d'Université Côte d'Azur - Epione Team, 2004 Rte des Lucioles, 06902, Valbonne, France

Abstract

The multimodal nature of clinical assessment and decision-making, and the high rate of healthcare data generation, motivate the need to develop approaches specifically adapted to the analysis of these complex and potentially high-dimensional multimodal datasets. This poses both technical and conceptual problems: how can such heterogeneous data be analyzed jointly? How can modality-specific information be identified from shared information? Variational autoencoders (VAEs) offer a robust framework for learning latent representations of complex data distributions, while being flexible enough to adapt to different data types and structures, and have already been successfully applied for latent disentanglement of multimodal data. Identifying causal relationships between available modalities, beyond simple statistical associations, could provide valuable and actionable insights, but conventional causal discovery techniques suffer from the curse of dimensionality. To address these issues, we propose Multi-Channel Causal VAE (*MC^2 VAE*), a causal disentanglement approach for multichannel data, whose objective is to jointly learn modality-specific latent representations from a multichannel dataset, and identify a linear causal structure between the latent variables. Each modality is projected into its own latent space, where a causal discovery step is integrated to learn the hidden causal graph. Finally, the decoder takes into account the discovered graph to predict the data. We formally derive MC^{2}VAE and the optimization strategy for its parameters. Experiments on synthetically generated data-sets underline the ability of our model to identify ground-truth hidden causal relationships, opening up a viable avenue for actionable interventions on multichannel systems.

Keywords

Multi-channel, Causal discovery, Variational Autoencoder

1. Introduction

Multichannel data refers to data-sets where observations generated from multiple sources are gathered together: each of those sources is intended to capture a specific information of the phenomenon under study, and can contribute to its overall understanding. This type of data (sometimes called multimodal or multiviews) is becoming increasingly common in various fields, spanning *e.g.* from finance, to increase the predicting abilities of market trends analysis [1], to healthcare [2]. In the healthcare domain, clinical decision making, both for diagnosis, prognosis or therapeutics, is typically done considering as much information as possible on patients, which may come for instance from medical imaging, clinical scores, and medical reports. This simple statement, together with the increasing availability of healthcare data, motivates the development of methods specifically tailored to the joint analysis of multimodal data, which can plays a crucial role through personalized medicine [3].

safaa.al-ali@inria.fr (S. Al-Ali); irene.balelli@inria.fr (I. Balelli)
0000-0003-1864-8190 (S. Al-Ali); 0000-0002-4593-8217 (I. Balelli)

Unfortunately, multichannel data analysis is far from trivial, and comes with several challenges, due to their intrinsic heterogeneity, the possible high dimensionality of some channels, and the potential correlations between channels, each of one being a specific piece of a same puzzle. Finally, the effective integration of such multichannel data should be able to preserve each channel-specific information and highlight the cross-channels one.

Variational Autoencoders (VAEs) have gained significant attention for their ability to learn complex data distributions in an unsupervised manner [4]. VAEs are Bayesian generative models, consisting of an encoder, whose role is to project the input data into a lower-dimensional latent space, and the decoder which reconstructs the original data from its latent representation. Latent variables are sampled from their estimated posterior distributions. The VAE inference process is efficiently carried out using amortized inference [5], where the posterior moments are parameterized with neural networks. The flexibility of VAEs makes them particularly suitable for dealing with different types of data, hence they appear as good candidates to perform multichannel analysis. Causal learning is a very active and evolving area of research that aims to identify causal relationships between observations, going beyond simple statistical association, and ultimately improving interpretability, explainability and deployability. Classical methods to discover the underlined causal structure from a set of observations, are well suited to deal with relatively low-dimensional data. More recently, attempts to couple causal discovery and machine learning techniques to cope with higher dimensional and complex datasets, have shown promising results. Nevertheless, scaling up to multichannel data still appears as an extremely challenging task.

Our aim is to design a method, called Multi-Channel Causal Variational Autoencoder (*MC*^{2}*VAE*), for the joint analysis of multichannel data, able to identify meaningful causal relationships between each modality, enhancing our comprehension of the overall picture each channel is contributing to. To do so, MC^{2}VAE will rely on VAEs, which ensure the definition of a latent projection for each channel, and causal discovery techniques to identify the causal relationships between channels through their lower-dimensional representations: this knowledge will be valuable for helping the VAEs to reconstruct the data.

The rest of the paper is organised as follows. In Section 2, we summarize the state-of-the-art on multichannel data analysis, with a specific focus on VAEs, and causal disentanglement learning. In Section 3, we mathematically derive our method, MC^{2}VAE, and its optimisation strategy. Section 4 shows some results on synthetically generated data, showing the relevance of identifying relationships across channels to reconstruct the original data. Finally, Section 5 concludes the paper and propose some future research directions.

2. Related works

2.1. Multi-channel Representation Learning

The goal of disentangled representation learning is to disentangle the observed variables by projecting them into hidden lower-dimensional independent features, which correspond to distinct generative factors able to describe the data in a more compact manner [6, 7]. Classical models to solve this task includes Recurrent Neural Networks (RNNs) [8], Generative Adversarial Networks (GANs) [9], Deep Reinforcement Learning (DRL) [10], and VAEs [11].

In the multichannel context, VAEs have already been successfully deployed. For instance, Antelmi *et al.* [12] proposed a sparse VAE that jointly learns latent relationships across multiple channels, under the assumption that the latent space is shared across all channels. On the other side, in [13] the authors study the spatio-temporal dynamics of disease evolution through a multichannel VAE, where each channel is projected separately in its latent space, and a latent neural dynamical system describes the time evolution of the latent variables. A multichannel variational autoencoder (MVAE) based on conditional VAE (CVAE) [14] has also been proposed, and applied to analyze time signals, while a generalized multichannel conditional variational autoencoder [15] has been used for multichannel audio source separation under underdetermined conditions.

In [16], Bayesian Networks in conjunction with deep learning techniques (sparse autoencoders) are used to incorporate arbitrary multi-scale, multi-modal data without making specific distribution assumptions. A conditional generative modeling (CGM) approach for unsupervised disentangled learning based on variational autoencoder (VAE) was also proposed [17]. CGM employs a multimodal or categorical conditional prior distribution in the latent space to learn global uncertainty in the data. Finally, to enforce multimodal coherence, Wesego *et al.* [18] proposed to learn the correlation among the latent variables of unimodal VAEs using score-based models.

2.2. Causal Disentangled Representation Learning

Causal discovery [19, 20] is a branch of causal research that aims to unveil causal relationships between observed variables. Unlike correlation, which seeks to identify statistical associations between variables, causal links are defined as intrinsically asymmetric relationships, whose cause-and-effect directions are typically represented through causal graphs. Causal discovery has proven in being relevant in diverse research fields, including medicine [21], biology [22], physics [23], and economics [24]. It is a powerful tool, whose aim is try to understand the underlying mechanisms driving observed phenomena, when prior expert knowledge is not already fully established, enhancing decision-making processes. Indeed, causal discovery is a necessary preliminary step to perform causal inference, the basis for conterfactual reasoning. Nevertheless, classical approaches for causal discovery (*e.g.* [25, 26, 27, 28]) strongly suffer from the curse of dimensionality, and are not well adapted to deal with complex and high dimensional datasets.

Causal Disentangled Representation Learning (CDRL) [29] is an emerging field of research that seeks to solve this bottleneck by relying on the ability of machine and deep learning methods for feature extraction. Unlike conventional representation learning, which may merge multiple causal factors into a single representation, causal disentanglement strives to separate these factors enhancing the comprehension of the data generation process. Through the identification and separation of latent factors that contribute to the generation of the observed variables, CDRL enables a more comprehensive understanding of the underlying causal mechanisms governing complex data distributions. Moreover, causal disentanglement can facilitate the simulation of the effects of altering specific causal factors in a controlled manner, thus allowing for counterfactual reasoning.

In this context, CausalVAE [30] addresses the issue of learning causal disentanglement in

observational data. Considering external information about data named *concepts* or labels, the authors propose a VAE that includes a Causal Layer to learn a causal graph in the latent factors that mirrors the causal relationships between the considered *concepts*. The model shows good results on both synthetic and benchmark datasets. However, by leveraging the a priori known graph on the *concepts* of interest, they enforce the dimension of the latent space and the causal structure therein. An extension of CausalVAE has been later proposed by Komandouri *et al.* [31] to relax the linear assumption of causal relationships.

CausalGan [32], is designed to perform causal inference on images, but requires a prior causal graph. Another approach [33] incorporates dependent latent factors instead of assuming independent latent factors. For disentanglement, the authors utilize the principle of *independence mechanism* or *modularity* and design a layer with non-structured nodes that represents outputs of mutually independent causal mechanisms. These mechanisms contribute collectively to the final predictions, achieving disentanglement.

Finally, causalPIMA [34] has been recently proposed for causal representation learning, integrating multiple modalities and physics-based constraints. Each modality is encoded into its proper latent space and unimodal embeddings are joined through a product of experts. The latent space is then clustered by a Gaussian mixture prior. Notably, the algorithm permits also the integration of known physics constraints during the decoding stage.

Despite the significant achievements reached, we noticed that the already existing methods require to inject additional knowledge to drive the causal discovery in the latent space. In addition, with the exception of [34], the challenging multichannel scenario has still to be addressed.

To deal with both multichannel data and the need of recovering their causal relationships, accounting for the realistic unavailability of an established causal pattern across channels, we propose Multi-Channel Causal Variational Autoencoder, MC^{2}VAE. Our approach aims to uncover a causal graph between channels' latent variables in a fully unsupervised manner, *i.e.* without requiring any additional information beyond the data itself. In this way, we aim to obtain an interpretable and richer representation of such complex data.

3. Proposed method

The overall architecture of MC^{2}VAE is summarized in Figure 1. First, we jointly train the encoder of each channel to obtain the corresponding latent variable, which is supposed being one-dimensional to enforce the discovery of causal relationships *between channels*. Next, the latent variables undergo a *Latent Causal Layer* where a causal discovery step is performed to find the causal graph. Finally, the decoder considers the resulting transformed causal latent variables to reconstruct the original data.

In this section, we will detail the mathematical formulation of our model. We denote by $\mathbf{X} = (X_i)_{i=1,\dots,N}$ the dataset of observed variables, where N is the total number of subjects, and $X_i := (\mathbf{x}_i^m)_{m=1,\dots,M}$ is the dataset of the i^{th} subject, consisting of observations from a total of M channels, $\mathbf{x}_i^m$. Clearly, each $\mathbf{x}_i^m$ will lie in a specific d_m-dimensional space. The latent vector for subject i following the encoding operation will be denoted by $\mathbf{z}_i := (z_i^m)_{m=1,\dots,M} \in \mathbb{R}^M$, where each z_i^m corresponds to the latent representation of $\mathbf{x}_i^m$. After passing through the causal

layer, we will denote the transformed latent variables as $\mathbf{z}_i^c$: they will be feed to the decoder.

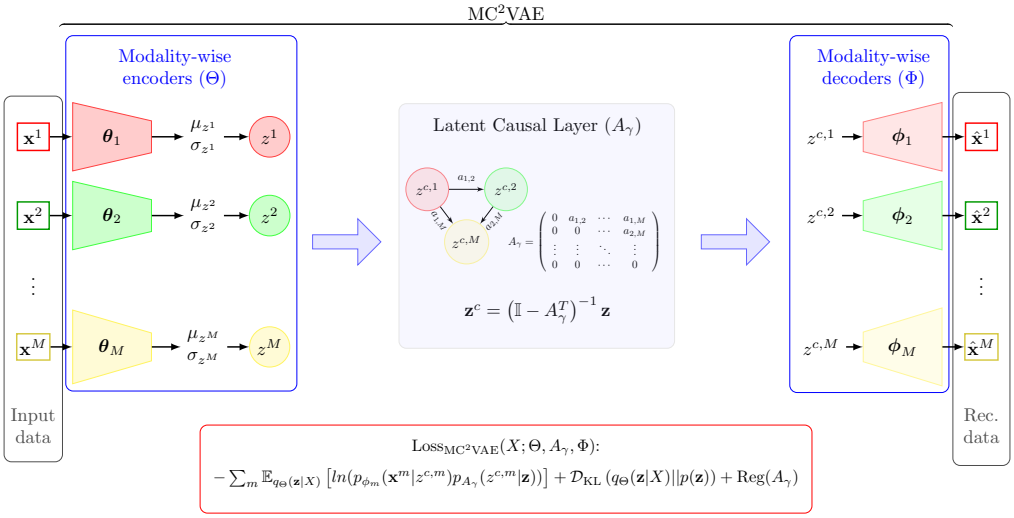

Figure 1: Workflow of MC^{2}VAE. Each channel m is encoded to a one-dimensional latent variable z^m through its channel-specific encoder. The latent variables $\mathbf{z} = (z^m)_{m=1,\dots,M}$ pass into the Latent Causal Layer, where the linear structural causal model (Eq. (3)) is learned. The decoder takes the transformed causal variables $\mathbf{z}^c = (z^{c,m})_{m=1,\dots,M}$ to reconstruct the observations for each channel, through a channel-specific decoding.

3.1. Latent structural causal model

MC^{2}VAE hypothesizes the existence of a structural causal model (SCM) relating the channel-specific latent variables. A SCM is typically defined by a triplet $\mathcal{M} = \{\mathbf{z}^c, \mathbf{f}_{\boldsymbol{\gamma}}, \mathbf{z}\}$ where $\mathbf{z}^c = \{z^{c,m}\}_{m=1,\dots,M}$ are the causally related variables, $\mathbf{f}_{\boldsymbol{\gamma}}$ are functionals, parameterized by $\boldsymbol{\gamma} = (\gamma_m)_{m=1,\dots,M}$, which describe in a deterministic manner the nature of the causal relationships between $z^{c,1}, \dots, z^{c,M}$, and $\mathbf{z} = (z^m)_{m=1,\dots,M}$ are the associated (independent) noise variables. We can write:

$$\mathbf{z}^c = \mathbf{f}_{\boldsymbol{\gamma}}(A^T\mathbf{z}^c) + \mathbf{z}, \tag{1}$$

where A is an adjacency matrix, strictly upper triangular up to some permutations, which describes the directed acyclic graph (DAG) structure of $z^{c,1}, \dots, z^{c,M}$. The superscript T denotes the matrix transpose.

In this work we restrict ourselves to the assumptions of linear relationships among the latent variables, and a normal prior for the noise terms. Therefore, the latent SCM writes:

$$\mathbf{z}^c = A_{\boldsymbol{\gamma}}^T\mathbf{z}^c + \mathbf{z} = (\mathbb{I} - A_{\boldsymbol{\gamma}}^T)^{-1}\mathbf{z}, \mathbf{z} \sim \mathcal{N}(0, \mathbb{I}), \tag{2}$$

where $\mathbb{I}$ denotes the identity matrix, and $A_{\boldsymbol{\gamma}}$ is the adjacency matrix A weighted by the parameter $\boldsymbol{\gamma}$: the ij^{th} term of $A_{\boldsymbol{\gamma}}$ provides the strength of the causal linear relationships of z_i (the parent variable) to z_j (the children).

Following the SCM in (2), we get the Markov factorization of the joint distribution of the latent variables:

$$p(\mathbf{z}_i^c|\mathbf{z}_i;\Theta_{\text{SCM}}) = \prod_{m=1}^{M} p(z_i^m|\text{Pa}_i^m, \mathbf{z}_i;\Theta_{\text{SCM}}), i \in \{1,\dots,N\}, \tag{3}$$

where Pa_i^m represent the set of parents of z_i in the latent causal graph and $\Theta_{\text{SCM}} := (A, \boldsymbol{\gamma})$ are the parameters of the structural causal model, to be learned.

3.2. Multichannel Causal Variational Auto-encoder

In this section, we state the probabilistic formulation of our proposed inference and generative model. The encoder and decoder functions are parameterized by $\Theta := (\boldsymbol{\theta}_m)_{m=1,\dots,M}$ and $\Phi := (\boldsymbol{\phi}_m)_{m=1,\dots,M}$ respectively. We assume a prior distribution $p(\mathbf{z}) = \prod_{m=1}^{M} \mathcal{N}(0,1)$ over the latent space. In order to optimise the parameters of MC^{2}VAE, we would like to maximize the marginal log likelihood:

$$\begin{aligned}\mathcal{L}(\mathbf{X};\Phi,\Theta,\Theta_{\text{SCM}}) &= \sum_{i=1}^{N} \log\left[p(X_i;\Phi,\Theta_{\text{SCM}})\right] \\ &= \sum_{i=1}^{N} \log\left[\int p(X_i|\mathbf{z}_i^c;\Phi)p(\mathbf{z}_i^c|\mathbf{z}_i;\Theta_{\text{SCM}})p(\mathbf{z}_i)d\mathbf{z}_i\right],\end{aligned}$$

where

$$p(X_i|\mathbf{z}_i^c;\Phi) := \prod_{m=1}^{M} p(\mathbf{x}_i^m|z_i^{c,m};\boldsymbol{\phi}_m). \tag{4}$$

We apply variational Bayes and introduce a tractable posterior $q(\mathbf{z}|X;\Theta)$ to approximate the true posterior $p(\mathbf{z}|X)$:

$$q(\mathbf{z}_i|X_i;\Theta) = \prod_{m=1}^{M} q(z_i^m|\mathbf{x}_i^m;\boldsymbol{\theta}_m) = \prod_{m=1}^{M} \mathcal{N}(\mu_m^{\text{E}}(\mathbf{x}_i^m,\theta_{m,1}), \sigma_m^{\text{E}^2}(\mathbf{x}_i^m,\theta_{m,2})), \tag{5}$$

where $\mu_m^{\text{E}}, \sigma_m^{\text{E}}$ are channel-specific functions depending on $\mathbf{x}_i^m$ and on the 2D-parameter $\boldsymbol{\theta}_m$. This leads us to the following lower bound:

$$\begin{aligned}\mathcal{L}(\mathbf{X};\Phi,\Theta,\Theta_{\text{SCM}}) \geq\ & \mathbb{E}_{\mathbf{z}\sim q(\mathbf{z}|X;\Theta)}\left\{\log\left[p(X|\mathbf{z}^c;\Phi)\right]\right\} + \mathbb{E}_{\mathbf{z}\sim q(\mathbf{z}|X;\Theta)}\left\{\log\left[p(\mathbf{z}_i^c|\mathbf{z}_i;\Theta_{\text{SCM}})\right]\right\} \\ & -\mathcal{D}_{\text{KL}}\left(q(\mathbf{z}|X;\Theta)||p(\mathbf{z})\right) := \mathcal{E},\end{aligned} \tag{6}$$

where $\mathcal{D}_{\text{KL}}$ denotes the Kullback-Leibler divergence.

The following result can be demonstrated:

Theorem 1. *Given the assumptions in Equations* (3),(5),(4)*, the lower bound* $\mathcal{E}$ *defined in Equation* (6) *is:*

$$\begin{aligned}\mathcal{E} = -\frac{1}{2}\sum_{i=1}^{N}\Bigg\{\sum_{m=1}^{M}\Bigg[d_m \log\left(2\pi\sigma_m^{\mathrm{D}^2}(z^{c,m},\phi_{m,2})\right) + \frac{||\mathbf{x}^m - \mu_m^{\mathrm{D}}(z^{c,m},\phi_{m,1})||^2}{\sigma_m^{\mathrm{D}^2}(z^{c,m},\phi_{m,2})} \\ - \log\left(\frac{\sigma_m^{\mathrm{E}^2}(\mathbf{x}_i^m,\theta_{m,2})}{2\pi}\right) + \sigma_m^{\mathrm{E}^2}(\mathbf{x}_i^m,\theta_{m,2}) + \mu_m^{\mathrm{E}^2}(\mathbf{x}_i^m,\theta_{m,1}) - 1\Bigg] + I_i\Bigg\},\end{aligned} \tag{7}$$

where

$$I_i \;=\; \det(B) + tr\left(B^{-1}\Sigma^{\mathrm{E}}(X_i,\boldsymbol{\theta})\right) + \left(\boldsymbol{\mu}^E(X_i,\boldsymbol{\theta})\right)^T B^{-1}\boldsymbol{\mu}^E(X_i,\boldsymbol{\theta}), \tag{8}$$

and

- $B := \left(\mathbb{I} - A_\gamma^T\right)^{-1}(\mathbb{I} - A_\gamma^T)$,
- $\boldsymbol{\mu}^E(X_i,\boldsymbol{\theta})$ *is the concatenation of all* $\mu_m^{\mathrm{E}^2}(\mathbf{x}_i^m,\theta_{m,1})$,
- $\Sigma^E(X_i,\boldsymbol{\theta}) = diag(\sigma_m^{\mathrm{E}^2}(\mathbf{x}_i^m,\theta_{m,2}))_{m=1,\dots,M}$.

Proof. For the sake of clarity, we denote $\mathbb{E}_{\mathbf{z}\sim q(\mathbf{z}|X;\Theta)}$ by simply $\mathbb{E}_q$, and we will drop the index i in what follows. The first term in Equation (6) gives (thanks to assumption (4)):

$$\begin{aligned}\mathbb{E}_q\left\{\log\left[p(X|\mathbf{z}^c;\Phi)\right]\right\} &= \sum_{m=1}^{M}\mathbb{E}_q\left\{\log\left[p(\mathbf{x}^m|\mathbf{z}^{c,m};\boldsymbol{\phi}_m)\right]\right\} \\ &= -\frac{1}{2}\sum_{m=1}^{M}\Bigg(d_m\log\left(2\pi\sigma_m^{\mathrm{D}^2}(z^{c,m},\phi_{m,2})\right) \\ &\quad + \frac{||\mathbf{x}^m - \mu_m^{\mathrm{D}}(z^{c,m},\phi_{m,1})||^2}{\sigma_m^{\mathrm{D}^2}(z^{c,m},\phi_{m,2})}\Bigg).\end{aligned} \tag{9}$$

The Kullback-Leibler divergence term can be factorised thanks to assumption (5), then analytically derived (it is the classical $\mathcal{D}_{\mathrm{KL}}$ between univariate Gaussians):

$$\begin{aligned}\mathcal{D}_{\mathrm{KL}}\left(q(\mathbf{z}|X;\Theta)||p(\mathbf{z})\right) &= \frac{1}{2}\sum_{m=1}^{M}\Bigg[\sigma_m^{\mathrm{E}^2}(\mathbf{x}_i^m,\theta_{m,2}) - \log(\sigma_m^{\mathrm{E}^2}(\mathbf{x}_i^m,\theta_{m,2}) \\ &\quad + \mu_m^{\mathrm{E}^2}(\mathbf{x}_i^m,\theta_{m,1}) - 1\Bigg].\end{aligned} \tag{10}$$

Let us focus on the second term of Equation (6).

Proposition 2. *Under the latent SCM assumption given by Equation* (2) *we have:*

$$\begin{aligned}\mathbb{E}_q\left\{\log\left[p(\mathbf{z}_i^c|\mathbf{z}_i;\Theta_{\mathrm{SCM}})\right]\right\} &= -\frac{1}{2}\Bigg(M\log(2\pi) + \det(B) + tr\left(B^{-1}\Sigma^{\mathrm{E}}(X_i,\boldsymbol{\theta})\right) \\ &\quad + \left(\boldsymbol{\mu}^E(X_i,\boldsymbol{\theta})\right)^T B^{-1}\boldsymbol{\mu}^E(X_i,\boldsymbol{\theta})\Bigg).\end{aligned}$$

Proof. Thanks to assumption (2):

$$\mathbb{E}_q\left\{\log\left[p(\mathbf{z}_i^c|\mathbf{z}_i;\Theta_{\text{SCM}})\right]\right\} \quad = \quad -\frac{1}{2}\left(M\log(2\pi)+\det\left(B\right)+\mathbb{E}_q\left(\mathbf{z}^T B^{-1}\mathbf{z}\right)\right).$$

The solution follows using assumption (5). □

The proof of theorem 1 follows by putting together Equations (9), (10) and (11). □

3.3. Acyclicity penalisation

We shall recall that our objective is to learn a directed acyclic graph relating the channel-specific latent variables. The acyclicity is encoded in the causal unweighted adjacency matrix A, which we expect to be strictly upper triangular. To enforce this requirement, we add a penalisation term to the loss function given in Equation (7), inspired by Zheng et al. [35], where they show the following result:

Theorem 3. *Let $A \in \mathbb{R}^{M\times M}$ be the unweighted adjacency matrix of a directed graph. For any $\alpha > 0$, the graph is acyclic if and only if*

$$\mathcal{H}_\alpha(A) := tr\left[(\mathbb{I}+\alpha A\circ A)^M\right]-M=0, \tag{11}$$

We use Theorem 3 with $\alpha = c/M$ for c big enough, and finally get our objective function to be optimized:

$$\min\left(-\mathcal{E}+\mathcal{H}_\alpha(A)\right). \tag{12}$$

3.4. MC^{2}VAE algorithm

We have implemented MC^{2}VAE using Python 3.10.9 and Pytorch 2.2.1. Algorithm 1 briefly outlines the steps performed at each training epoch. Our learning algorithm is efficiently carried out through minibatch stochastic gradient descent using the Adam optimizer [36]. The code used to run MC^{2}VAE (and to generate the synthetic datasets illustrated in Section 4) is made publicly available on https://gitlab.inria.fr/ibalelli/mc2vae.

4. Experimental results

We perform experiments on synthetically generated data, to challenge our method to 1) recover the ground truth causal graph and 2) correctly reconstruct all channels. In order to assess the relevance of accounting for the latent causal relationships we perform an ablation study and compare MC^{2}VAE with a simplified multichannel architecture, where the causal layer has been switched off. For each experiment and both models, encoders and decoders consist of a single linear layer. We performed 500 epochs with an initial learning rate of $1e^{-2}$, which allows to reach convergence for both models and all synthetic data-sets. The choice of privileging synthetic data-sets here is motivated by the need of a ground truth latent graph to compare with, which is not commonly known in most real worlds multichannel applications.

Algorithm 1 MC^{2}VAE

Require: Multimodal data $(\mathbf{x}_m)_{m=1,\dots,M}$; c (Eq. (11)); E (epochs); batch size; optimiser hyper-parameters

for $e = 1, \dots, E$ **do**
  for $m = 1, \dots, M$ **do**
    Sample $z^m \sim q(z^m|\mathbf{x}^m)$ (reparameterization trick)
  end for
  $\mathbf{z} \leftarrow concat(z^m)_{m=1,\dots,M}$
  Compute $\mathbf{z}^c = (\mathbb{I} - A_\gamma^T)^{-1}\mathbf{z}$ (Latent Causal Layer)
  Compute $\mathbb{E}_q\left\{\log\left[p(X|\mathbf{z}^c;\Phi)\right]\right\}$ (Eq. (9))
  Compute $\mathbb{E}_q\left\{\log\left[p(\mathbf{z}_i^c|\mathbf{z}_i;\Theta_{\text{SCM}})\right]\right\}$ (Eq. (11))
  Compute $\mathcal{D}_{\text{KL}}\left(q(\mathbf{z}|X;\Theta)||p(\mathbf{z})\right)$ (Eq. (10))
  Compute $\mathcal{E}$ (Eq. (7))
  Compute $\mathcal{H}_{c/M}(A)$ (Eq. (11))
  Return $Loss_{\text{MC}^2\text{VAE}}(X;\Phi,\Theta,\Theta_{\text{SCM}}) := -\mathcal{E} + \mathcal{H}_{c/M}(A)$
  Backpropagate
end for

We consider three different synthetically generated data (**D1**, **D2** and **D3**, see Table 1), which differs by the number of channels and channel dimensions. Data-sets **D1** and **D2** have been both generated from a 3-dimensional latent space which follows the causal graph given in Figure 2, top panel, where the arrows indicate a linear transformation of the child variable from its parent, as prescribed by Equation (2). Data-set **D3** consists of 5 channels of varying dimensions, generated from a 5-dimensional latent space which follows the causal graph in Figure 2, bottom panel. For all tests, a total of $N = 2000$ subjects was generated, using linear operators. Finally, centered Gaussian noises with random standard deviations was added to each channel.

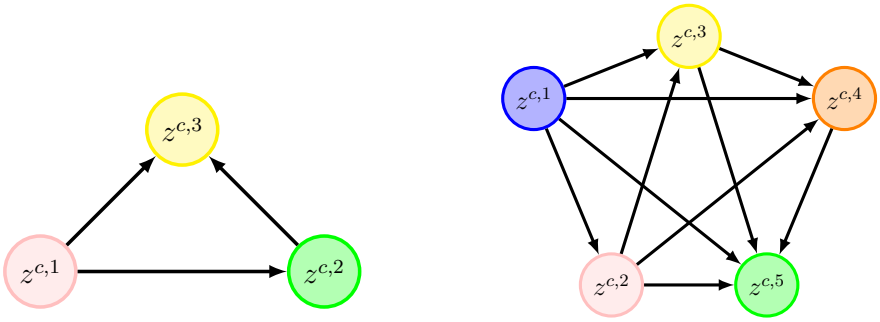

Figure 2: Ground truth latent causal graphs used to generate data-sets **D1** and **D2** (left graph) and **D3** (right graph).

We consider 500 epochs for all the tests, which allows to reach convergence of the algorithm.

Table 1
Number of channels and their respective dimension per generated data-set.

Data-sets	# of channels	# of features/channel
D1	3	$[10, 5, 7]$
D2	3	$[20, 15, 30]$
D3	5	$[20, 15, 30, 10, 5]$

In Figure 3, we show the results for data-sets **D1**, **D2** and **D3**, comparing the full MC^{2}VAE architecture (in red) with the one where the causal layer has been removed (in blue), meaning that each channel results in being fully independent from the others. One can see that considering this causal learning step within the VAE algorithm significantly improves the reconstruction of the data. When the causal learning step is removed, precious information coming for the interrelationships between channels is completely lost, highly affecting the reconstruction ability of the autoencoders, which in some case reach very scarce results. This occurs for almost all the channels regardless of the number of considered features, and is enhanced when increasing the number of channels and ground truth causal relationships across them.

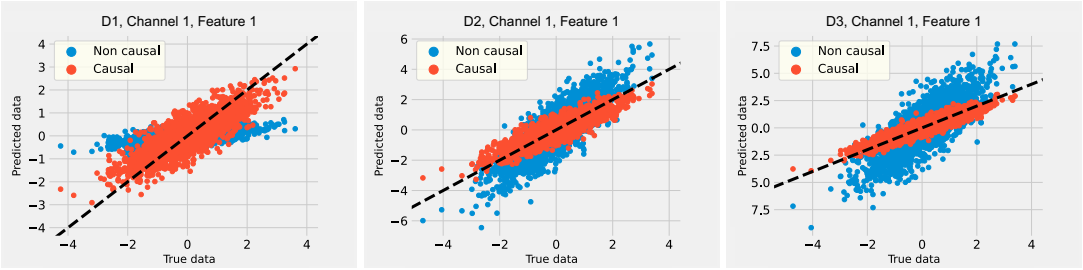

Figure 3: From left to right: results for data-set **D1**, **D2** and **D3**, respectively. For each data-set the first feature for the first channel is shown (similar results have been obtained for the remaining channels/features). The black dashed line corresponds to the diagonal.

In Table 2 we report some results concerning the comparison between the causal graphs discovered by our model for each data-set, and the ground truth ones from Figure 2. Our method can detect almost all the edges present in the ground truth graph in all considered data-sets. The number of missing arrows in the discovered graphs varies between 20% and 30% of total existing arrows. Interestingly, we observe an improvement in arrows detection when the number of channels is increased, and the causal graph is complexified. These very promising results highlight the ability of our model to correctly perform causal disentanglement form multichannel data, and discover the hidden causal relationships between the considered channels, enhancing both channel-specific and cross-channels information extraction.

To better show and quantify the reconstruction error, we computed the MSE between original and reconstructed data. In Figure 4, we show the MSE values for each channel (on the left) and each feature (on the right). Using causality, the average MSE per channel decreases by approximately 88% compared to the VAE trained without causality. Similar results are obtained

Table 2
Number of correctly identified edges and edge directions using MC^{2}VAE for all synthetic data-sets.

Metric	D1	D2	D3
Total edges number	3	3	10
# detected edges	3	2	10
# missing edges	1	1	2
# incorrect edges	1	0	2
# inverted arrows	0	0	3

for the features.

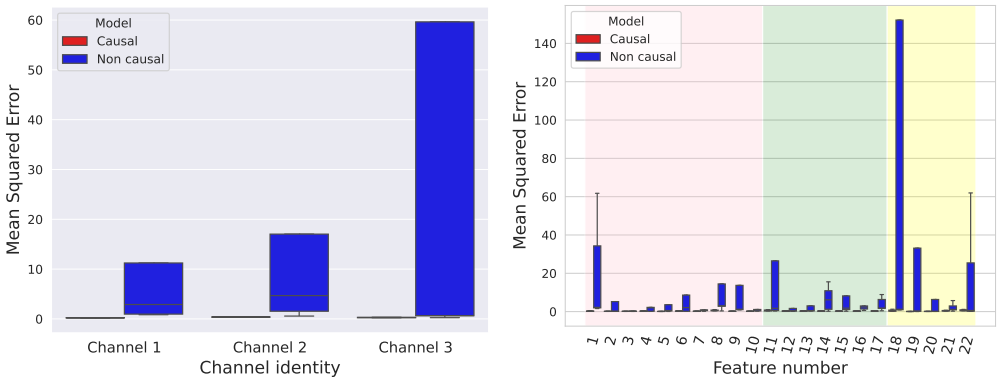

Figure 4: MSE values for each channel (on the left) and each feature (on the right) computed after training MC^{2}VAE five times over the same data-set **D1**. We highlight in pink the features of channel 1, in green features of channel 2 and in yellow features of channel 3.

5. Conclusions and future directions

In this paper, we present a novel model for causal learning from multichannel data. Our method leverage variational autoencoders to obtain a compact representation of each channel, but conversely to classical VAE structures, it integrates a causal discovery layer to unveil the underlying hidden causal relationships across channels. Our approach showed promising results when applied to synthetically generated data-sets. MC^{2}VAE displays an improved reconstruction ability thanks to the learned causal latent structure. In addition, MC^{2}VAE is able to detect almost all the causal relationships relating the channels. Application on real data-sets specifically from (but not restricted to) the healthcare domain is one of our short-term major interests. Beside new applications, this work opens to several exciting perspectives and extensions, including, among others, the possibility of modeling interventions on the latent graph, hence expand to the causal inference domain. This will open up a new perspective to get actionable information on multichannel systems.

References

[1] V. S. Pagolu, K. N. Reddy, G. Panda, B. Majhi, Sentiment analysis of twitter data for predicting stock market movements, in: 2016 international conference on signal processing, communication, power and embedded system (SCOPES), IEEE, 2016, pp. 1345–1350.

[2] J. N. Acosta, G. J. Falcone, P. Rajpurkar, E. J. Topol, Multimodal biomedical ai, Nature Medicine 28 (2022) 1773–1784.

[3] A. Kline, H. Wang, Y. Li, S. Dennis, M. Hutch, Z. Xu, F. Wang, F. Cheng, Y. Luo, Multimodal machine learning in precision health: A scoping review, npj Digital Medicine 5 (2022) 171.

[4] I. Higgins, L. Matthey, A. Pal, C. P. Burgess, X. Glorot, M. M. Botvinick, S. Mohamed, A. Lerchner, beta-vae: Learning basic visual concepts with a constrained variational framework., ICLR (Poster) 3 (2017).

[5] Y. Kim, S. Wiseman, A. Miller, D. Sontag, A. Rush, Semi-amortized variational autoencoders, in: International Conference on Machine Learning, PMLR, 2018, pp. 2678–2687.

[6] Y. Bengio, A. Courville, P. Vincent, Representation learning: A review and new perspectives, IEEE transactions on pattern analysis and machine intelligence 35 (2013) 1798–1828.

[7] X. Wang, H. Chen, S. Tang, Z. Wu, W. Zhu, Disentangled representation learning, arXiv preprint arXiv:2211.11695 (2022).

[8] L. R. Medsker, L. Jain, et al., Recurrent neural networks, Design and Applications 5 (2001) 2.

[9] M. Suzuki, K. Nakayama, Y. Matsuo, Joint multimodal learning with deep generative models, arXiv preprint arXiv:1611.01891 (2016).

[10] M. Sewak, Deep reinforcement learning, Springer, 2019.

[11] D. P. Kingma, M. Welling, Auto-encoding variational bayes, arXiv preprint arXiv:1312.6114 (2013).

[12] L. Antelmi, N. Ayache, P. Robert, M. Lorenzi, Sparse multi-channel variational autoencoder for the joint analysis of heterogeneous data, in: International Conference on Machine Learning, PMLR, 2019, pp. 302–311.

[13] C. Abi Nader, N. Ayache, G. B. Frisoni, P. Robert, M. Lorenzi, A. D. N. Initiative, Simulating the outcome of amyloid treatments in alzheimer's disease from imaging and clinical data, Brain communications 3 (2021) fcab091.

[14] H. Kameoka, L. Li, S. Inoue, S. Makino, Supervised determined source separation with multichannel variational autoencoder, Neural Computation 31 (2019) 1891–1914. doi:10.1162/neco_a_01217.

[15] S. Seki, H. Kameoka, L. Li, T. Toda, K. Takeda, Underdetermined source separation based on generalized multichannel variational autoencoder, IEEE Access 7 (2019) 168104–168115. doi:10.1109/ACCESS.2019.2954120.

[16] M. Sood, A. Sahay, R. Karki, M. A. Emon, H. Vrooman, M. Hofmann-Apitius, H. Fröhlich, Realistic simulation of virtual multi-scale, multi-modal patient trajectories using bayesian networks and sparse auto-encoders, Scientific reports 10 (2020) 10971.

[17] Improving generative modelling in vaes using multimodal prior, IEEE Transactions on Multimedia 23 (2021) 2153–2161. doi:10.1109/TMM.2020.3008053.

[18] D. Wesego, A. Rooshenas, Score-based multimodal autoencoders, arXiv preprint arXiv:2305.15708 (2023).

[19] J. Pearl, Causality, Cambridge university press, 2009.
[20] J. Pearl, D. Mackenzie, The book of why: the new science of cause and effect, Basic books, 2018.
[21] P. Sanchez, J. P. Voisey, T. Xia, H. I. Watson, A. Q. O'Neil, S. A. Tsaftaris, Causal machine learning for healthcare and precision medicine, Royal Society Open Science 9 (2022) 220638.
[22] S. Al-Ali, J. Llopis-Lorente, M. T. Mora, M. Sermesant, B. Trenor, I. Balelli, A causal discovery approach to streamline ionic currents selection to improve drug-induced tdp risk assessment, in: 2023 Computing in Cardiology (CinC), volume 50, 2023, pp. 1–4. doi:10.22489/CinC.2023.009.
[23] G. Camps-Valls, A. Gerhardus, U. Ninad, G. Varando, G. Martius, E. Balaguer-Ballester, R. Vinuesa, E. Diaz, L. Zanna, J. Runge, Discovering causal relations and equations from data, Physics Reports 1044 (2023) 1–68.
[24] B. Huang, K. Zhang, M. Gong, C. Glymour, Causal discovery and forecasting in nonstationary environments with state-space models, in: International conference on machine learning, Pmlr, 2019, pp. 2901–2910.
[25] P. Spirtes, C. Glymour, An algorithm for fast recovery of sparse causal graphs, Social science computer review 9 (1991) 62–72.
[26] P. Spirtes, C. N. Glymour, R. Scheines, Causation, prediction, and search, MIT press, 2000.
[27] S. Zhu, I. Ng, Z. Chen, Causal discovery with reinforcement learning, arXiv preprint arXiv:1906.04477 (2019).
[28] S. Shimizu, P. O. Hoyer, A. Hyvärinen, A. Kerminen, M. Jordan, A linear non-gaussian acyclic model for causal discovery., Journal of Machine Learning Research 7 (2006).
[29] A. G. Reddy, V. N. Balasubramanian, et al., On causally disentangled representations, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 36, 2022, pp. 8089–8097.
[30] M. Yang, F. Liu, Z. Chen, X. Shen, J. Hao, J. Wang, Causalvae: Disentangled representation learning via neural structural causal models, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 9593–9602.
[31] A. Komanduri, Y. Wu, W. Huang, F. Chen, X. Wu, Scm-vae: Learning identifiable causal representations via structural knowledge, in: 2022 IEEE International Conference on Big Data (Big Data), IEEE, 2022, pp. 1014–1023.
[32] M. Kocaoglu, C. Snyder, A. G. Dimakis, S. Vishwanath, Causalgan: Learning causal implicit generative models with adversarial training, arXiv preprint arXiv:1709.02023 (2017).
[33] M. Besserve, A. Mehrjou, R. Sun, B. Schölkopf, Counterfactuals uncover the modular structure of deep generative models, in: Eighth International Conference on Learning Representations (ICLR 2020), 2020.
[34] E. Walker, J. A. Actor, C. Martinez, N. Trask, Causal disentanglement of multimodal data, arXiv preprint arXiv:2310.18471 (2023).
[35] X. Zheng, B. Aragam, P. K. Ravikumar, E. P. Xing, Dags with no tears: Continuous optimization for structure learning, Advances in neural information processing systems 31 (2018).
[36] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980 (2014).