Encoding Causal Macrovariables

Benedikt Höltgen*
External collaborator to OATML Group
Department of Computer Science
University of Oxford
Oxford, United Kingdom
benedikt.hoeltgen@mailbox.org

Abstract

In many scientific disciplines, coarse-grained causal models are used to explain and predict the dynamics of more fine-grained systems. Naturally, such models require appropriate macrovariables. Automated procedures to detect suitable variables would be useful to leverage increasingly available high-dimensional observational datasets. This work introduces a novel algorithmic approach that is inspired by a new characterisation of causal macrovariables as information bottlenecks between microstates. Its general form can be adapted to address individual needs of different scientific goals. After a further transformation step, the causal relationships between learned variables can be investigated through additive noise models. Experiments on both simulated data and on a real climate dataset are reported. In a synthetic dataset, the algorithm robustly detects the ground-truth variables and correctly infers the causal relationships between them. In a real climate dataset, the algorithm robustly detects two variables that correspond to the two known variations of the El Niño phenomenon.

1 Introduction

With graphical and structural causal models becoming increasingly popular, both scientists and philosophers are starting to look more closely at the main ingredient to these models: causal variables. Finding such causal variables has long been a neglected area of research that has only recently started to motivate a growing literature in both machine learning ([5, 7, 6]) and philosophy of science [34, 9]. Especially in the machine learning community, there are now calls for further research into causal representation learning [27, 28] and variable construction in particular [10]. Beyond artificial intelligence research, this is particularly relevant for scientific disciplines – such as climate science, neuroscience, or economics – in which higher-level models need to be constructed based on high-dimensional observational data. The two central challenges are the identification of suitable macrovariables and the inference of causal relationships from purely observational data.

The approach proposed in this work is based on a novel characterisation of causal macrovariables as information bottlenecks. Building on the information bottleneck framework, we show how neuron activations in artificial neural networks can be interpreted and optimised as coarse-grained causal variables over high-dimensional data. To this end, we introduce a novel neural network structure loosely based on Variational Autoencoders, which we call the Causal Autoencoder. It can be applied to settings where two high-dimensional datasets are available. This framework allows to establish a connection between the often separately studied problems of causal inference (where both cause and effect variables are investigated) and learning disentangled representation (where only one dataset is given). With the novel approach, the causal relationships between detected macrovariables can be

*Work done while a master student at the MCMP/LMU Munich.

Workshop on Causal Inference & Machine Learning: Why now? (WHY-21@NeurIPS2021).

investigated through additive noise models, after applying an additional transformation step. The methodology is tested on both simulated and natural data. For the simulated dataset, the ground-truth generative model is recovered, including the direction of causality. For the natural climate dataset, sensible macrovariables are detected that are in line with corresponding domain knowledge.

2 Background: Causal macrovariable detection

2.1 Causal macrovariables

On a strict notion of micro- and macrovariables, they stand in a deterministic functional relationship: there needs to be some deterministic function between a (often high-dimensional) microvariable space and a macrovariable space. This function assigns a macrovariable state to each microvariable state, the former 'supervene' on the latter. It should be noted that micro- and macrovariable are relative notions: While temperature is a macrovariable in relation to the kinetic energy of molecules, it is a microvariable in the context of large-scale climate models with hundreds of temperature measurements. In general, different scientific goals often require different scientific ontologies for the same system [8, 25].

A good way to think about causal structures in the world is as relationships between patterns that supervene on microvariable states [2, 25]. There need, however, not be a unique causal structure within a given system: different ways of carving up the system into patterns can often yield a variety of causal structures within the system [25]. Furthermore, as Spirtes [29] and Eberhardt [9] show, even one specific causal structure "can be equivalently described by two different sets of variables that stand in a non-trivial translation-relation to each other" [9]. In general, there is no a uniquely 'correct' choice of appropriate variables for the representation of causal systems; this is sometimes explicitly acknowledged in the machine learning literature [33].

However, it does not mean that any model is as good as the other. In fact, coming up with sensible scientific ontologies is one of the main tasks that scientists are concerned with. Deciding between them can be guided by different criteria whose importance varies with context and goal. James Woodward [34] has recently proposed a tentative list of criteria for causal variable selection, although conceding that eventually it might turn out "that there is nothing systematic to say about this issue" (p. 1048); among his criteria, the following four are particularly interesting for the present work (p. 1054f):

a) "there is a clear answer to the question of what would happen if they were to be manipulated or intervened on"

b) they "lead to causal representations that are relatively sparse"

c) they "exhibit strong correlations between cause and effect"

d) the relationships between them "continue to hold under changes in background conditions"

2.2 Problem setup

For this work we are considering a very general setup where two high-dimensional and dependent variables X and Y are given. X and Y might have common causes C and there might be a one-directional causal path either from X to Y or vice versa.[2] As the setup is not restricted to deterministic cases, we also allow that X and Y can be influenced by individual and independent noise N^X and N^Y, respectively. This leads to the setup depicted in fig. 1 and one of the following pairs of structural assignments:

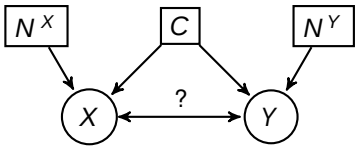

Figure 1: Schematic causal diagram of the setup with variables X and Y, noise, and common causes. It is assumed that only X and Y are observed and that the graph is acyclic.

$$X := (N^X; C) \quad Y = (N^Y; C; X) \quad (X \not\perp\!\!\!\perp Y)$$
$$X := (N^X; C; Y) \quad Y = (N^Y; C) \quad (X \not\perp\!\!\!\perp Y)$$
$$X := (N^X; C) \quad Y = (N^Y; C) \quad (X \leftrightarrow Y)$$

[2] Throughout this work, uppercase letters denote multidimensional variables while lowercase letters denote one-dimensional ones.

2.3 Related work

A task that is closely related to one addressed here is learning disentangled representations. The goal of this area of research is to find representations that correspond to the (perhaps causal) factors of variation [4, 19]. Various approaches to this have been based on Variational Autoencoders (VAEs), thus encoding samples into a noisy bottleneck layer and decoding it to 'predict' the input again. Although the concept of mutual information (MI) is a cornerstone of many of these approaches, the relevance of MI for unsupervised representation learning algorithms is still unclear [32]. A central difference in the present approach is the aim of encoding high-dimensional datasets into bottlenecks that stand in causal relation to each other.

In other work, machine learning researchers have started to investigate causal relationships between neuron activations and image classification outputs [28]. However, this type of causal relationship concerns the algorithm's classification mechanism rather than dependencies in the data. The authors investigate claims like "the presence of cars cause the presence of wheels" rather than actual causal mechanisms in the world. While they also use neuron activations, their activations do not provide information bottlenecks between two datasets.

The closest work to the present one, focusing on the same setup, is the 'causal feature learning' approach developed by Krzysztof Chalupka and colleagues [5, 7]. Their aim is to find categorical variables representing different causal macrostates in each of the two datasets X and Y. The approach assumes that the causal direction is known beforehand. The key idea is that X-microstates belong to the same causal X-macrostate iff, when the result of an intervention, they induce the same probability distribution over Y-microstates and analogous for Y-microstates. In [6], they report results from applying their algorithm scheme to climate data, interpreting them as an "unsupervised discovery of El Niño". The resulting causal macrovariables are categorical, which are strictly less informative than continuous ones. Another limitation, which they concede in the context of the mentioned climate data, is that without (perhaps infeasible) real climate experiments or "large-scale climate experiments with detailed climate models" [6], no causal claims can be justified. While their work provides a potentially very fruitful avenue, we will in the following suggest a novel approach aiming to overcome these limitations by detecting continuous macrovariables.

3 Method

3.1 Causal macrovariables as information bottlenecks

The starting point for our approach to automated causal macrovariable detection is the insight that all dependencies are due to causation. This was famously formulated by Reichenbach [26] in his principle of the common cause: If events – or, rather, random variables – A and B are correlated – or, rather, dependent –, then either A caused B, B caused A, or A and B are both effects of a shared common cause. This also implies that causal macrovariables provide an information bottleneck (sometimes also called sufficient statistics) between the microvariables. To illustrate this, assume that the simple causal diagram of fig. 2, including the causal macrovariables x_1, y_1, x_2, and y_2, exhaustively represents the causal structure of the system depicted in fig. 1. Now, by exhausting all causal connections between X and Y, the principle of the common cause implies that they also exhaust all mutual information shared by X and Y. In formal terms, $I(\overline{X};Y) = I(X;Y)$ and $I(\overline{Y};X) = I(Y;X)$ for $\overline{X} := x_1; x_2$ and $\overline{Y} := y_1; y_2$, where $I(\,;\,)$ denotes the mutual information. As the macrovariables are functions of the microstates, they also cannot contain more information than the respective micro description. This leads to the equalities $I(\overline{X};Y) = I(X;Y)$ and $I(\overline{Y};X) = I(Y;X)$.[3]

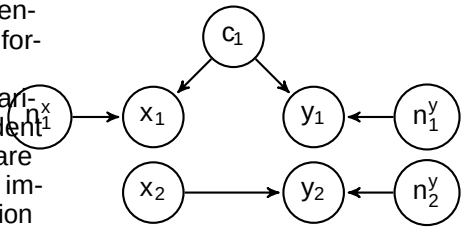

Figure 2: Example of a simple causal diagram with the macrovariables x_1, y_1, x_2, and y_2, where the former two have a common cause and the third causes the fourth.

[3] Note that the example of fig. 2 has a particularly simple structure. If there was a pair of variables which both stand in a direct causal relationship and have common causes, this would not affect the information theoretic

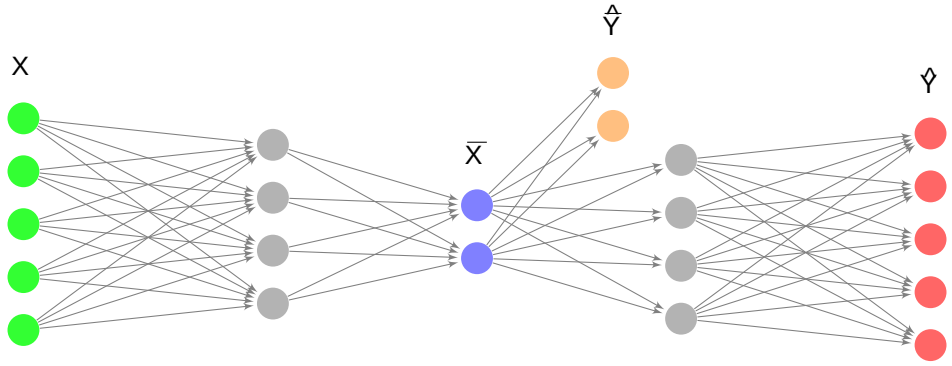

Figure 3: Structure of net_X, which constitutes one half of the CAE: for input X, net_X learns a lower-dimensional embedding $\bar{X}$ (providing the causal macrovariables), from which it predicts Y as well as the current bottleneck layer $\bar{Y}$ of net_Y. The latter is the the second half of the CAE, predicting X and $\bar{X}$ from Y, and has the same structure as net_X.

Recall that we want to find representations f_i and g_i yielding macrovariables $x_i = f_i(X)$ and $y_i = g_i(Y)$ which ideally capture all and only the mutual information between X and Y. We noted earlier that abstracting to higher-level descriptions generally comes with a loss of information. Precisely for this trade-off between compressing a signal and "preserving the relevant information about another variable", Tishby et al. [31] developed the information bottleneck (IB) framework. In this framework, the "optimal assignment" of the bottleneck $\bar{X}$ (and, analogously, $\bar{Y}$) can be found by minimising the functional

$$L[p(\bar{x}|x)] = I(\bar{X};X) \quad \beta I(\bar{X};Y); \tag{1}$$

where the Lagrange multiplier β governs the trade-off. Later, it has been observed that neural networks can be fruitfully analysed within this framework: considering the mutual information between the layers and the input and output variables, the training task can be seen as "an information theoretic trade-off between compression and prediction" [30]. In related work, certain stochastic neural nets – of which VAEs are a special case – have been shown to minimise the IB functional [1]. One can, thus, train a VAE-like stochastic neural net to learn a compression of the input and then use the encoder function without noise to construct macrovariables that supervene on the input.

3.2 Causal Autoencoders

The approach presented here consists in training two stochastic neural nets simultaneously. In the following, we will describe only the net that takes X as input (net_X), for simplicity of presentation; the second net (net_Y) has the same structure and together they form the Causal Autoencoder (CAE). The most significant difference between VAEs and net_X is that the latter does not aim to decode the bottleneck layer $\bar{X}$ back into X; in line with the bottleneck's identification with causal macrovariables and the IB analysis discussed above, net_X is instead trained to predict Y (cf. fig. 3). In this regard, net_X is more akin to supervised learning algorithms than autoencoders. In another deviation from VAEs, net_X has a second output layer, which is trained to predict the current bottleneck layer of net_Y (cf. fig. 3). This allows, first, to ensure that the macrovariables $\bar{X}$ and $\bar{Y}$ indeed stand in some functional relation and, second, to enforce constraints on this functional relation: in the example of fig. 3, the single fully connected layer between $\bar{X}$ and $\hat{\bar{Y}}$ makes the CAE learn macrovariables which stand in a linear relation to each other. I will return to the topic of functional constraints in section 3.3.

As net_X should be trained to predict the bottleneck layer of net_Y and vice versa, both nets need to be trained simultaneously. Taking the loss function of conventional VAEs and adding a third term for the second output layer, the loss function of net_X takes the form

$$loss_{net_X} = d_1(Y;\hat{Y}) + D_{KL}(N(0;1)||q(\bar{X}|X)) + d_2(\bar{Y};\hat{\bar{Y}}) \tag{2}$$

description given above, as the variables would still contain all the shared information. The causal inference techniques discussed below can, however, only handle cases with a simple structure.

where d_1 and d_2 are appropriate metrics – like MSE – and $q(\bar{X}|X)$ denotes the (noisy) distribution of the bottleneck neurons. While the loss function proposed in the original VAE paper [17] did not contain any parameters, later work by Higgins et al. [14] introduced a parameter β, as in eq. (2). Note that this parameter governs a similar trade-off as the parameter β in the IB framework. Higgins et al. [14] consider it a limitation of their approach that it is not possible to estimate the optimal value of β directly; in the present context, it allows to accommodate the objectives of different scientific goals. In general, as discussed above, there is no unique causal structure in a given system and different representations might be better for different goals. Hence, the possibility to find different models – which are, for example, more or less detailed – is actually desirable. Intuitions on the meaning of the terms are given in appendix A.1 and examples of how choices between models can be informed are given in the experiments section below.

A naive approach for training net_X and net_Y would be to train each of them for one minibatch in alternation. In appendix A.2, we explain why it is better to instead combine the loss functions for net_X and net_Y into a sum with six terms and actually treat both parts as constituents of the same neural network, the CAE.

3.3 Architectural constraints on function classes

The task of net_X is essentially to learn three functions f, a, and γ such that $\bar{X} = f(X)$, $\hat{Y} = a(\bar{X})$ and $\hat{Y} = \gamma(X)$. This section concerns different constraints that can be imposed on these functions through choices about the neural network architecture. With the architecture depicted in fig. 3, the only constraint (beyond complexity constraints depending on the size of the layers) is that a must be linear, given that there is no hidden layer between the bottleneck layer $\bar{X}$ and the variable output layer $\hat{Y}$. This implies that the function a is defined by a matrix A, with $\hat{Y} = A\bar{X}$. While this constraint can be dropped by adding hidden layers or an activation function, other constraints can be put in place by altering the architecture in other respects. In our experiments, we use the combination of two constraints:

First, $\hat{y}_i = a_i x_i + b_i$ implies that each Y-macrovariable y_i (or bottleneck neuron of net_Y) is predicted based on the value of only one corresponding X-macrovariable x_i, and the prediction must be linear. This constraint leads to variables that fulfill two of Woodward's desiderata (section 2.1), namely sparse representations and strong correlations, to a very high degree. Another advantage of this constraint is that it facilitates the investigation of causal relationships, e.g. through additive noise models (section 3.4).

Second, $\hat{Y} = \sum_i \gamma_i(x_i)$ implies that the full Y-microstate is predicted as a linear combination of transformations of individual X-variables. This can help to make the CAE learn variables that lend themselves to more straightforward causal interpretations. Such a constraint in necessary to prevent the CAE from learning variables that are arbitrary recombinations of a given set of variables: For a 'good' model with $\bar{X} := x_1, x_2$ and $\bar{Y} := y_1, y_2$ s.t. $y_1 = a_1 x_1$ and $y_2 = a_2 x_2$, the variables $\bar{X}' := x_1, x_2'$ and $\bar{Y}' := y_1, y_2'$ with $x_2' = a_1 x_1 x_2$ and $y_2' = y_1 y_2$ satisfy the same pair of equations. The constraint proposed here is a way to select the more interpretable models.

3.4 Investigating the causal direction with ANMs

One advantage of the novel CAE approach is that continuous macrovariables allow to draw on an ample causal inference literature for the investigation of relationships between the detected variables [24]. In this section, we sketch the idea behind additive noise models (ANMs) before describing how they can be adapted to overcome difficulties faced in the present context.

3.4.1 Additive Noise Models

In the general framework of structural causal models, a relationship of causing y_i can be represented by some assignment $y_i := \hat{\gamma}_i(x_i, n)$, with the noise n being independent from x_i. Now the idea behind ANMs is that by imposing plausible restrictions on $\hat{\gamma}_i$, it can be possible to infer causal relationships from observational data. As suggested by their name, ANMs assume that the independent noise is additive, i.e. we can reformulate the assignment as $y_i := \gamma_i(x_i) + n$, where x_i and n are again independent. Given two dependent variables x_i and y_i, we can try to find such an ANM either for the causal direction $x_i \to y_i$ or the reverse $x_i \leftarrow y_i$. Mooij et al. [22] compare the performance

of some approaches to this causal inference task on a benchmark dataset and come to the conclusion that "the original ANM method ANM-pHSIC proposed by Hoyer et al. (2009) turned out to be one of the best methods overall" (p. 45). On the mentioned approach, a prediction $\hat{y}_i$ for y_i is made based on the predictor x_i, in order to compute the residuals $y_{i;\text{res}} = y_i - \hat{y}_i$ that serves as a proxy for the noise n. Then it is tested whether $y_{i;\text{res}}$ and x_i are independent; for this, Hoyer and colleagues suggest to use the Hilbert-Schmidt Independence Criterion (HSIC) [21]. The same steps are applied with x_i and y_i reversed; if the independence hypothesis (and thus the ANM) is accepted in one direction but rejected in the other, we infer that this is the causal direction.

3.4.2 CAEs and ANMs

In the causal inference literature, it is usually assumed that the causal variables are given and come with a natural scale. However, the correct numerical representation of causal variables is often not clear in complex applications like neuro- or climate science. Here, the causal patterns we want to investigate might not have a privileged numerical representation and it might not be obvious whether some macrovariable x_i is better than e.g. a transformed version $\log(x_i)$. This immediately leads to a problem for directly investigating the macrovariables detected by a CAE through ANMs.

Even with the architectural constraints described in section 3.3, the CAE is still agnostic with respect to monotonic transformations of the detected variables. Therefore, the CAE should not be seen as learning macrovariable pairs but equivalence classes of macrovariable pairs, where pairs are equivalent if their variables can be transformed into each other by monotonic transformations. This is not surprising from an information-theoretic perspective, as mutual information is invariant under invertible (and, thus, monotononic) transformations. Now the issue for applying ANMs is that the independence of residuals is affected by such transformations. Therefore, the ANM approach can yield very different results for two pairs of variables even if both pairs are equivalent from the CAE's perspective. This means that we cannot naively plug the detected macrovariables directly into ANM algorithms for causal inference.

In order to check whether there are transformations of detected variables x_i and y_i that are compatible with an ANM in one of the two directions, we can check the transformations which minimise the dependence between residual and predictor. The first task is, thus, to find two variable pairs satisfying $x_i^0, y_i^0 = \arg\min_{x_i, y_i} \text{HSIC}(x_i, y_{i;\text{res}})$ and $x_i^{00}, y_i^{00} = \arg\min_{x_i, y_i} \text{HSIC}(y_i, x_{i;\text{res}})$, where all variables are monotonic transformations of x_i and y_i, respectively. One way to do this is to again use VAEs, one for each variable (fig. 4). To find the variables that minimise the dependence between x_i and $y_{i;\text{res}}$, the VAEs are trained on the loss

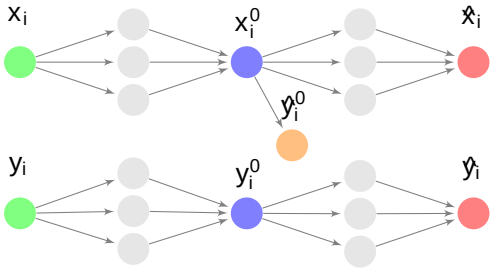

Figure 4: A pair of VAEs is used to find the monotonically transformed variables x_i^0 and y_i^0 (from x_i and y_i) that minimise the dependence (given by the HSIC score) between x_i^0 and $y_{i;\text{res}}^0 = y_i^0 - \hat{y}_i^0$.

$$\text{loss}_{X \to Y} = \text{loss}_{\text{VAE};x} + \text{loss}_{\text{VAE};y} + \frac{\text{MSE}(y_i, \hat{y}_i)}{\text{Var}(y_i)} + \text{HSIC}(x_i, y_{i;\text{res}}): \tag{3}$$

After finding the optimal transformations, the HSIC score of x_i^0 and $y_{i;\text{res}}^0$ is computed[4]. The same procedure is applied with x_i and y_i reversed, to find a monotonically transformed variable pair with minimal dependence between y_i^{00} and $x_{i;\text{res}}^{00}$. The two scores are then compared to see whether there is a strong case for a causal influence in either direction. A threshold can be computed that gives a criterion deciding whether to accept or reject the independence hypothesis [21]. However, as whether the threshold is exceeded depends on both the sample size and the details of the loss function, it seems advisable to also take into account the disparity between $\text{HSIC}(x_2^0, y_{2;\text{res}}^0)$ and $\text{HSIC}(y_2^{00}, x_{2;\text{res}}^{00})$.

[4] The idea of minimising the HSIC score directly before testing it has already been explored in [21], although for regression and not in the setting of an autoencoder.

4 Experiments

4.1 Simulated data

We first report experiments on simulated data with a known ground truth model.[5] Here, X and Y are random variables over $\mathbb{R}^{64}$, so they can be thought of as quadratic grey-scale images of 8×8 pixels. The underlying ground truth model used for generating the data is that of fig. 2: There are four macrovariables x_1, y_1, x_2, y_2 where x_1 and y_1 have a common cause, x_2 and y_2 is caused by x_2. The four macrovariables correspond to averages in the left/right half of X and the top/bottom half of Y, respectively. The data is generated according to the following three structural assignments:

$$x_1 := c_1 + n_1^X \qquad y_1 := c_1^3 + n_1^Y \qquad y_2 := \tanh(x_2) + n_2^Y$$

with $c_1, x_2 \sim U([-1, 1])$ and $n_1^X, n_1^Y, n_2^Y \sim U([-0.2, 0.2])$ all uniformly distributed and mutually independent. Pairs of low-level states (x, y) are generated by starting with states that satisfy the high-level descriptions exactly and then adding pixel-wise uniform noise $\sim U([-0.2, 0.2])$.

4.1.1 Variable detection

We use the constrained CAE structure discussed above with bottleneck dimension 4. The results of different hyperparameter settings are reported in table 1. It shows that both net_X and net_Y effectively use two bottleneck neurons for all hyperparameter settings (i.e. the other neurons are dominated by the injected noise and cannot carry information), corresponding to the number of macrovariables in the ground truth model (see previous paragraph), attesting the robustness of the approach when there is a clear ground truth. However, the CAE's ability to learn variables that can predict each other (e.g. y_1 from x_1 and vice versa) varies greatly. Although the network structure generally yields pairs of variables x_i and y_i that predict each other, this is not the case for low choices of β, this sometimes leads to negative explained variance (EV) scores in predicting the variables. The CAE with the best performance according to the EV measures is the one with the highest and lowest β, i.e. the one where the relative weight of the second term of the loss function is the lowest. This is not surprising, as it means that the training is less noisy (no noise was injected during the evaluation on the validation set) and, thus, more accurate – yet still generalisable – predictions can be learned.

Table 1: Explained variance (EV) in predicting Y from X and $\bar{Y}$ from $\bar{X}$ through $\text{net}_X = \text{net}_Y$ on the test set and the number of detected variables $|\bar{X}|, |\bar{Y}|$ (i.e. number of informative/non-random bottleneck neurons) after 1500 epochs, for different values of α and β in eq. (2). The number of detected macrovariables is always two, corresponding to the number of ground-truth variables. For some settings with low β, the learned variables do not allow a prediction of their counterparts, resulting in negative EV scores.

	$\beta = 1$	$\beta = 0.1$	$\beta = 0.01$
$\alpha = 1$	$\lvert\bar{X}\rvert = 2, \lvert\bar{Y}\rvert = 2$ $EV(Y \mid X) = .62 \mid .57$; $EV(\bar{Y} \mid \bar{X}) = .62 \mid .63$;	$\lvert\bar{X}\rvert = 2, \lvert\bar{Y}\rvert = 2$ $EV(Y \mid X) = .80 \mid .77$; $EV(\bar{Y} \mid \bar{X}) = .89 \mid .88$;	$\lvert\bar{X}\rvert = 2, \lvert\bar{Y}\rvert = 2$ $EV(Y \mid X) = .81 \mid .78$; $EV(\bar{Y} \mid \bar{X}) = .89 \mid .90$;
$\alpha = 0.1$	$\lvert\bar{X}\rvert = 2, \lvert\bar{Y}\rvert = 2$ $EV(Y \mid X) = .57 \mid .51$; $EV(\bar{Y} \mid \bar{X}) = < 0 \mid < 0$;	$\lvert\bar{X}\rvert = 2, \lvert\bar{Y}\rvert = 2$ $EV(Y \mid X) = .80 \mid .77$; $EV(\bar{Y} \mid \bar{X}) = .86 \mid .86$;	$\lvert\bar{X}\rvert = 2, \lvert\bar{Y}\rvert = 2$ $EV(Y \mid X) = .81 \mid .78$; $EV(\bar{Y} \mid \bar{X}) = .89 \mid .87$;
$\alpha = 0.01$	$\lvert\bar{X}\rvert = 2, \lvert\bar{Y}\rvert = 2$ $EV(Y \mid X) = .59 \mid .53$; $EV(\bar{Y} \mid \bar{X}) = < 0 \mid < 0$;	$\lvert\bar{X}\rvert = 2, \lvert\bar{Y}\rvert = 2$ $EV(Y \mid X) = .80 \mid .77$; $EV(\bar{Y} \mid \bar{X}) = < 0 \mid < 0$;	$\lvert\bar{X}\rvert = 2, \lvert\bar{Y}\rvert = 2$ $EV(Y \mid X) = .81 \mid .78$; $EV(\bar{Y} \mid \bar{X}) = .61 \mid .26$;

4.1.2 Causal direction

For analysing the causal direction, the variables detected by the CAE with hyperparameters $\alpha = 0.01$ and $\beta = 1$ were chosen since this setting showed the best performance w.r.t. the explained variance metrics (cf. table 1). Recall that to satisfy an ANM, the residual in predicting one variable should be independent from the respective predictor variable. After applying the respective transformations,

[5] The code for all experiments is available at https://github.com/benedikthoeltgen/causal-macro .

the HSIC scores for the second pair differ by a factor of ten, i.e. $\mathrm{HSIC}(x_2^0, y_{2;\mathrm{res}}^0) = 0.32$ and $\mathrm{HSIC}(y_2^{00}, x_{2;\mathrm{res}}^{00}) = 3.15$. In particular, the former drops below the threshold while the latter remains above it. This implies that the ANM $y_2^0 = f_2(x_2^0) + n_2^Y$ is accepted while the reverse model $x_2^{00} = f_2(y_2^{00}) + n_2^X$ is rejected. By the criterion suggested in [16], we correctly infer that x_2 causes y_2.

For the first variable pair (x_1, y_1), neither variable causes the other in the generative model. After transforming the detected variables, we get $\mathrm{HSIC}(x_2^0, y_{2;\mathrm{res}}^0) = 3.88$ and $\mathrm{HSIC}(y_2^{00}, x_{2;\mathrm{res}}^{00}) = 13.92$. All scores are well above the threshold, so we infer that no variable causes the other, in line with the ground truth. It should also be noted that one value is about four times as high as the other, both before and after the transformation. This is less salient than the disparity for the causal variables x_2 and y_2; still, it shows that the present approach might run into problems for more complex datasets.

4.2 Natural data: El Niño

Here, we report results from running the algorithm on the climate dataset investigated in [6], retrieved from the first author's website. It comprises 13140 weekly averaged measurements of zonal winds (ZW) and sea surface temperatures (SST) in the equatorial pacific, each on a 9×55 grid spanning from 140°E to 80°W and 10°N to 10°S, from the years 1979-2014. A well-known climate phenomenon repeatedly appearing in this region is El Niño. According to the National Oceanic and Atmospheric Administration (NOAA), it is defined as a "three-month average of sea surface temperature departures from normal for a critical region of the equatorial Pacific" [23]. A good reason to study El Niño in the context of causal macrovariable detection is that it provides strong causal links between various measurable quantities (such as ZW and SST) which allow both a high- and a low-level description. Chalupka et al. [6] apply their causal feature learning algorithm to this dataset and interpret their results as an unsupervised discovery of the El Niño phenomenon.

The CAE again uses the constrained structure, now with 16 bottleneck neurons. On this dataset, different choices of hyperparameters not only lead to differences in the accuracy of predictions but also in the number of detected variables (table 2). Low choices of α and, to a smaller degree, low choices of β – led to higher numbers of variables, i.e. non-noise bottleneck neurons. This higher model complexity comes with higher predictive power, as discussed in appendix A.1. Low choices of α again lead to negative EV scores in predicting the variables. One can also see that $|\bar{Y}|$ is often greater than $|\bar{X}|$. This is probably because X is harder to predict from Y than vice versa (see the EV scores).

Table 2: Explained variance (EV) in predicting Y from X and $\bar{Y}$ from $\bar{X}$ through $\mathrm{net}_X = \mathrm{net}_Y$ on the test set and the number of detected variables $|\bar{X}|, |\bar{Y}|$ for different values of α and β in eq. (2). Here, the number of detected variables strongly depends on the choice of hyperparameters, especially α. For some settings with low α, no prediction of the variables is learned, resulting in negative EV scores.

	$\beta = 1$	$\beta = 0.1$	$\beta = 0.01$
$\alpha = 1$	$\lvert\bar{X}\rvert = 3, \lvert\bar{Y}\rvert = 2$ $EV(Y/X) = .70 / < 0$; $EV(\bar{Y}/\bar{X}) = .82 / .43$;	$\lvert\bar{X}\rvert = 4, \lvert\bar{Y}\rvert = 5$ $EV(Y/X) = .83 / .52$; $EV(\bar{Y}/\bar{X}) = .82 / .85$;	$\lvert\bar{X}\rvert = 7, \lvert\bar{Y}\rvert = 7$ $EV(Y/X) = .86 / .64$; $EV(\bar{Y}/\bar{X}) = .88 / .88$;
$\alpha = 0.1$	$\lvert\bar{X}\rvert = 3, \lvert\bar{Y}\rvert = 3$ $EV(Y/X) = .71 / .15$; $EV(\bar{Y}/\bar{X}) = < 0 / < 0$;	$\lvert\bar{X}\rvert = 6, \lvert\bar{Y}\rvert = 9$ $EV(Y/X) = .85 / .67$; $EV(\bar{Y}/\bar{X}) = < 0 / .41$;	$\lvert\bar{X}\rvert = 13, \lvert\bar{Y}\rvert = 13$ $EV(Y/X) = .90 / .67$; $EV(\bar{Y}/\bar{X}) = .46 / .46$;
$\alpha = 0.01$	$\lvert\bar{X}\rvert = 3, \lvert\bar{Y}\rvert = 4$ $EV(Y/X) = .71 / .21$; $EV(\bar{Y}/\bar{X}) = < 0 / < 0$;	$\lvert\bar{X}\rvert = 7, \lvert\bar{Y}\rvert = 10$ $EV(Y/X) = .86 / .69$; $EV(\bar{Y}/\bar{X}) = < 0 / < 0$;	$\lvert\bar{X}\rvert = 16, \lvert\bar{Y}\rvert = 16$ $EV(Y/X) = .90 / .78$; $EV(\bar{Y}/\bar{X}) = < 0 / < 0$;

The detected variables allow a more nuanced analysis than the cluster analysis in [6], as they not only represent a warm area in general, but also allow to distinguish between two different variations of El Niño: For any tested configuration of α and β, there is one neuron tracking a very warm tongue from the east and one tracking a warm area in the western centre (fig. 5). Remarkably, these patterns correspond to the two known variations of El Niño, the warm pool (WP) and cold tongue (CT) El Niño [18]. As shown in appendix C, "the warm events can be well separated into the WP El Niño and CT El

Figure 5: Deviation from the mean temperature for the inputs with the highest values for two of the five learned macrovariables over the temperature dataset ($N = 6000$; $\alpha = 0.1$; $\beta = 1$). For all settings of α and β, the CAE detects two variables that allow to distinguish between the eastern 'warm pool' (WP, left) and the western/central 'cold tongue' (CT, right) El Niño variations (see appendix C).

Niño based on their SST anomaly patterns". Note that while the macrovariables assign a value to the strength of some pattern for each week, the labels CT, WP, or mixed are assigned for each year. As El Niño is defined as an anomaly over several months, assigning macrovariables based on potential causal relationships on a weekly scale (as done both here and in [6]) may appear problematic. For these reasons, we did not expect to find causal relationships between macrovariables through the ANM approach in the climate dataset. And indeed, the ANM analysis yielded no evidence for a direct causal relationship (see appendix D).

5 Discussion

As causal macrovariables form information bottlenecks, they can be captured by bottleneck layers of a novel neural network called Causal Autoencoder (CAE). Different hyperparameter settings in the loss function allow to weight the model's predictive power, simplicity, and accuracy against each other. In experiments, the CAE recovers the ground-truth variables from data generated by a simple four variable model. On natural data, it detects sensible variables that align with known variations of the El Niño phenomenon. It is also possible to apply additive noise models to the detected variables after a transformation step and thus investigate the causal relationship. For the simulated data, this has been shown to correctly identify the causal relationships. It is, however, not yet clear whether the application of ANMs will prove to be useful in scientific practice. In an experiment performed on natural climate data, the ANM approach did not lead to additional insights. It is not clear whether this is due to particularities of the used data. Another worry relates to the general difficulty of inferring causal relationships from observational data; Woodward's first criterion for causal variables cited in section 2.1 demands that interventions on these variables have distinctive effects. While the other three cited criteria are reasonably well met, this one is hard to assess.

When two variables both stand in a direct causal relationship and share a common cause, the ANM procedure can be expected to fail. In such cases, other approaches like the Neural Causation Coefficient [20] might give better results. On this issue, further theoretical and empirical investigations are required. Another avenue for future research is the investigation of variations of the CAE, in particular of other architectural constraints. Another exciting possibility would be to build CAEs in the form of a recurrent (RNN) or convolutional neural net (CNN). CAE-RNNs might be better suited for the application to time-series data, such as the climate data used in this work. Extending the CAE to CNNs could be particularly useful for domains where the location of a pattern is not important, i.e. where the same causal phenomenon can appear in different parts of a dataset. Lastly, it would be very interesting to investigate potential theoretical connections between the mutual information in observational distributions, as used for discovery here, and information-theoretic measures of causality. A variety of such measures can be found in the literature, often employing the mutual information between interventional distributions, e.g. [3, 13, 15]. This might provide a more thorough theoretical foundation of the approach proposed in this work.

Acknowledgements

I would like to thank Stephan Hartmann, Moritz Grosse-Wentrup, Frederick Eberhardt, Gunnar König, Timo Freiesleben, and Lood van Niekerk for helpful discussions and thorough feedback at different stages of this project.

References

[1] A. Achille and S. Soatto. Information dropout: learning optimal representations through noisy computation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(12): 2897–2905, 2018.

[2] H. K. Andersen. Patterns, information, and causation. *The Journal of Philosophy*, 114(11): 592–622, 2017.

[3] N. Ay and D. Polani. Information flows in causal networks. *Advances in complex systems*, 11 (01):17–41, 2008.

[4] Y. Bengio. Deep learning of representations: Looking forward. *International Conference on Statistical Language and Speech Processing*, 2013.

[5] K. Chalupka, P. Perona, and F. Eberhardt. Visual causal feature learning. *31st Conference on Uncertainty in Artificial Intelligence*, pages 181–190, 2015.

[6] K. Chalupka, T. Bischoff, P. Perona, and F. Eberhardt. Unsupervised discovery of El Niño using causal feature learning on microlevel climate data. *32nd Conference on Uncertainty in Artificial Intelligence*, pages 72–81, 2016.

[7] K. Chalupka, P. Perona, and F. Eberhardt. Multi-level cause-effect systems. In *19th International Conference on Artificial Intelligence and Statistics*, volume 41, pages 361–369, 2016.

[8] D. Danks. Goal-dependence in (scientific) ontology. *Synthese*, 192(11):3601–3616, 2015.

[9] F. Eberhardt. Green and grue causal variables. *Synthese*, 193(4):1029–1046, 2016.

[10] F. Eberhardt. Introduction to the foundations of causal discovery. *International Journal of Data Science and Analytics*, 3(2):81–91, 2017.

[11] A. Gretton, O. Bousquet, A. Smola, and B. Scölkopf. Measuring statistical dependence with Hilbert-Schmidt norms. In J. Jain, H. U. Simon, and E. Tomita, editors, *ALT 2005*, volume LNAI 3734, pages 63–77, 2005.

[12] A. Gretton, K. Fukumizu, C. H. Teo, L. Song, B. Schölkopf, and A. J. Smola. A kernel statistical test of independence. In *Advances in Neural Information Processing Systems 20*, pages 1–8, 2008.

[13] P. E. Griffiths, A. Pocheville, B. Calcott, K. Stotz, H. Kim, R. Knight, P. E. Grif, A. Pocheville, B. Calcott, K. Stotz, H. Kim, and R. Knight. Measuring Causal Specificity. *Philosophy of Science*, 82(4):529–555, 2015.

[14] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner. -VAE: Learning basic visual concepts with a constrained variational framework. *5th International Conference on Learning Representations*, pages 1–13, 2017.

[15] E. P. Hoel. When the map is better than the territory. *Entropy*, 19(5):188, 2017.

[16] P. O. Hoyer, D. Janzing, J. Mooij, J. Peters, and B. Schölkopf. Nonlinear causal discovery with additive noise models. In *Advances in Neural Information Processing Systems 21*, pages 689–696, 2009.

[17] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *2nd International Conference on Learning Representations*, pages 1–14, 2014.

[18] J. S. Kug, F. F. Jin, and S. I. An. Two types of El Niño events: Cold tongue El Niño and warm pool El Niño. *Journal of Climate*, 22(6):1499–1515, 2009.

[19] F. Locatello, S. Bauer, M. Lucie, G. Rätsch, S. Gelly, B. Schölkopf, and O. Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. *36th International Conference on Machine Learning*, pages 7247–7283, 2019.

[20] D. Lopez-Paz, R. Nishihara, S. Chintala, B. Schölkopf, and L. Bottou. Discovering causal signals in images. *30th IEEE Conference on Computer Vision and Pattern Recognition*, 2017-Janua: 58–66, 2017.

[21] J. M. Mooij, D. Janzing, J. Peters, and B. Schölkopf. Regression by dependence minimization and its application to causal inference in additive noise models. *26th International Conference On Machine Learning*, pages 745–752, 2009.

[22] J. M. Mooij, J. Peters, D. Janzing, J. Zscheischler, and B. Schölkopf. Distinguishing cause from effect using observational data: Methods and benchmarks. *Journal of Machine Learning Research*, 17:1–102, 2016.

[23] NOAA. North American countries reach consensus on El Niño definition. *NOAA News Announcement*, 2005. URL `https://www.nws.noaa.gov/ost/climate/STIP/ElNinoDef.htm`.

[24] J. Peters, D. Janzing, and B. Schölkopf. *Elements of causal inference: foundations and learning algorithms*. MIT press, 2017.

[25] A. Potochnik. *Idealization and the Aims of Science*. University of Chicago Press, 2017.

[26] H. Reichenbach. *The direction of time*. Univ of California Press, 1956.

[27] B. Schölkopf. Causality for machine learning. *arXiv preprint*, pages 1–20, 2019.

[28] B. Schölkopf, F. Locatello, S. Bauer, N. R. Ke, N. Kalchbrenner, A. Goyal, and Y. Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

[29] P. Spirtes. Variable definition and causal inference. *Proceedings of the international Congress for Logic, Methodology and Philosophy of Science*, 2007.

[30] N. Tishby and N. Zaslavsky. Deep learning and the information bottleneck principle. *2015 IEEE Information Theory Workshop*, 2015.

[31] N. Tishby, F. C. Pereira, and W. Bialek. The information bottleneck method. In *Proceedings ofthe 37-th Annual Allerton Conference on Communication, Control and Computing*, pages 368–377, 1999.

[32] M. Tschannen, J. Djolonga, P. K. Rubenstein, S. Gelly, and M. Lucic. On mutual information maximization for representation learning. *8th International Conference on Learning Representations*, pages 1–16, 2020.

[33] S. Weichwald. *Pragmatism and variable transformations in causal modelling*. PhD thesis, ETH Zurich, 2019.

[34] J. Woodward. The problem of variable choice. *Synthese*, 193(4):1047–1072, 2016.

A The CAE loss function

A.1 Loss terms and trade-offs

To demonstrate the benefits of the freedom to choose the parameters, we can give a high-level description of the role that each of the terms in the loss function (equation 2)

$$\text{loss}_{\text{net}_X} = d_1(Y; \hat{Y}) + \quad D_{KL}(N(0;1)||q(\overline{X}|X)) + \quad d_2(\overline{Y}; \hat{\overline{Y}})$$

plays for the resulting causal macrovariable model.

The *first term* simply measures the distance (e.g. via MSE) between the prediction $\hat{Y}$ and the known Y (for net_X). A low loss in this first term signifies that the bottleneck neurons, and hence the macrovariables, contain much of the information necessary for predicting the output. In other words, the resulting model has a high predictive power as it captures large parts of the system in question.

The *second term* measures the Kullback-Leibler Divergence of the noisy bottleneck distribution from the standard normal distribution. D_{KL} is zero for a single neuron if its activation is always drawn completely at random from that distribution and thus carries no information about X. It increases when the activation is drawn from other normal distributions, in particular when the mean varies among samples and the noise is small, which allows the neuron to carry information. The more neurons carry information, and the more information they carry, the higher is the loss. I denote by $|\overline{X}|$ and $|\overline{Y}|$ the number of neurons that carry information, which can be lower than the number of neurons in the bottleneck. A low loss in this second term signifies that the model is fairly simple (comprising few variables) and fairly robust, as it has been trained on noisy variables.

The *third term* is introduced in this work for the novel bottleneck neuron output layer. Similar to the first term, it measures (for net_X) the distance between the prediction $\hat{\overline{Y}}$ and $\overline{Y}$, the latter being calculated from the current net_Y. A low loss in this third term signifies that the X-variables allow a good prediction of the Y-variables, implying that the model is fairly accurate and may have closely causally connected variables.

A.2 Combining the loss functions

An issue that comes up especially – though not exclusively – for cases of asymmetric information (i.e. when X allows a better prediction of Y than the other way round or vice versa)[6], is the coordination between net_X and net_Y. Such cases motivate the use of a combined loss function $\text{loss}_{\text{net}_X} + \text{loss}_{\text{net}_Y}$ instead of training the two nets in alternation. This is best illustrated through an example.

Consider a system where y_1 and x_1 are given by the average of Y and X, respectively, and where $x_1 \quad U([-1;1])$ and $y_1 = x_1^2$. Here, for each Y-sample with $y_1 = q$ for some value $q \in [0;1]$, the corresponding X-sample has a x_1-value of either $\sqrt{q}$ or $-\sqrt{q}$ with equal probability. So net_Y has no chance of learning a useful prediction of x_1 or X; in fact, the third term of its loss function is minimal if it outputs the prediction $\hat{x}_1 = 0$ for every sample (given a loss function like MSE). As detecting y_1 does not help net_Y to decrease its loss, it would come up with a completely noisy bottleneck layer. This keeps the CAE from finding a model with useful variables like x_1 and y_1, as it relies on net_Y to detect y_1. It is possible to overcome this problem by using a combined loss function for gradient descent, by treating net_X and net_Y as *one* neural network. This way, both parts of the CAE can adapt also in order to reduce each other's loss. In the mentioned example, net_X could learn $x_1' = x_1^2$ – which is sufficient to predict both y_1 and Y –, such that net_Y would be able to reduce its third loss term by learning y_1 and thereby predicting $x_1' = y_1$. net_X, on the other hand, can now predict y_1, and thus $\overline{Y}$ in general, more accurately, and reduce its third loss term. Whether such an optimal solution will be found eventually still depends on the learning process and on local minima in particular. But as the huge successes of neural networks have shown, this issue often turns out to be less grave than expected. As this simple example shows, a combined loss function can help the CAE to find better models for a given specification of hyperparameters.

[6] Note that the formulation 'asymmetric information' might be misleading as mutual information is symmetric, that is, $\mathcal{I}(X;Y) = \mathcal{I}(Y;X)$. What is asymmetric, then, is how much information each of the variables contain *in addition* to the information shared between the two.

B Transformations of detected variables for simulated data

Here, we give more details on the transformation of the variables and the application of ANMs in the simulated data experiments. As fig. 6 (a) indicates, $y_{2;res}$ and x_2 (left side) appear to be less dependent than $x_{2;res}$ and y_2 (right side), in line with the true causal direction $x_2 \not\to y_2$. This impression is partly confirmed by calculating the HSIC scores (as suggested by Hoyer et al. [16]): The test statistics are $\mathrm{HSIC}(x_2; y_{2;res}) = 5.18$ and $\mathrm{HSIC}(y_2; x_{2;res}) = 13.76$ respectively, with a threshold of 0.65. This difference is, arguably, not sufficient for accepting the causal model $x_2 \not\to y_2$ (especially as both HSIC scores are far above the threshold). This suggests that it is necessary to first apply the VAE-based transformation step to minimise dependence as described in section 3.4.2.

Naturally, this yields variables with lower HSIC test statistics (see also fig. 6 (b), (c)). They now differ by a factor of ten, as $\mathrm{HSIC}(x'_2; y'_{2;res}) = 0.32$ and $\mathrm{HSIC}(y''_2; x''_{2;res}) = 3.15$. In particular, the former drops below the threshold while the latter remains above it. This means that the ANM $y'_2 = {}_2(x'_2) + n_2^Y$ is accepted while the reverse model $x''_2 = {}_2(y''_2) + n_2^X$ is rejected. By the criterion suggested in [16], we infer that x_2 causes y_2; this is in line with the generative model, i.e. with the known ground truth.

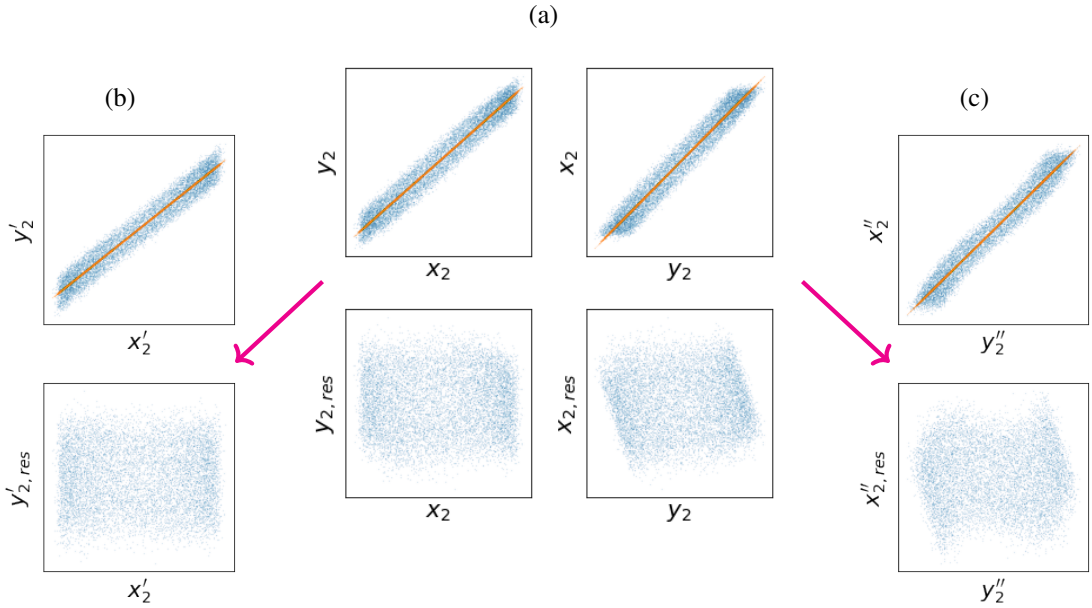

Figure 6: Values and residuals of the variables initially detected by the CAE **(a)**, as well as of the transformed variables, after minimising either $\mathrm{HSIC}(x_2; y_{2;res})$ **(b)** or $\mathrm{HSIC}(y_2; x_{2;res})$ **(c)**. Top row: Scatter plots of variable values (blue) and their predictions (orange). Bottom row: Scatter plots of the residuals (actual value minus prediction) against the variable on which the prediction is based. While transformations with independent x'_2 and $y'_{2;res}$ can be found **(b)**, the same does not hold true for y''_2 and $x''_{2;res}$ **(c)**. This indicates that x_2 causes y_2, in line with the known ground-truth model.

C El Niño

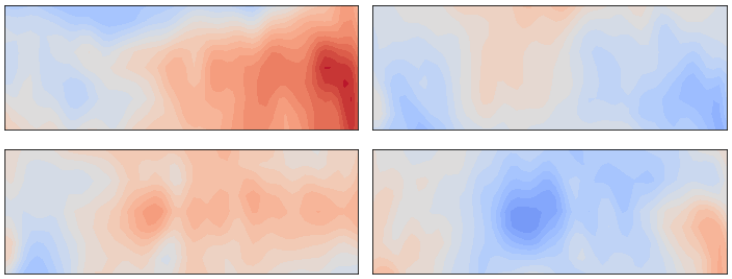

Figure 7: Deviation from mean temperature of inputs with highest (left) and lowest (right) values for two of the detected macrovariables (cf. fig. 5). The top row variable tracks the temperature in the east (with high values corresponding to WP El Niño) while the bottom row variable tracks the temperature in the center-west (with high values corresponding to CT El Niño), compare fig. 8 below.

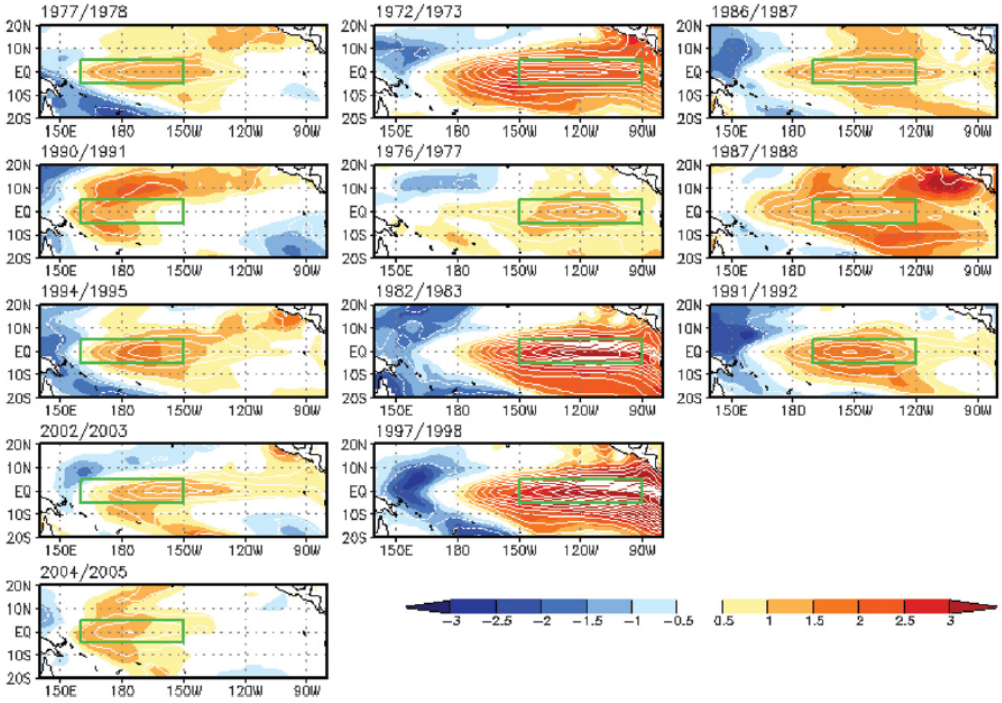

FIG. 1. SST anomalies of El Niño events during 1970–2005. The anomalies are averaged from September to the following February. Shading indicates normalized anomalies; contour interval is 0.3 K. The El Niño events are classified into (left) WP El Niño, (middle) CT El Niño, and (right) mixed El Niño. The green boxes indicate (left) Niño-4, (middle) Niño-3, and (right) Niño-3.4 regions.

Figure 8: Variations of El Niño, taken from Kug et al. [18, 1501]. Note that the depicted segment of the Pacific Ocean is the same as in the dataset investigated in the experiments in east-west expanse while it is greater in its north-south expanse: the samples in the investigated dataset only extend from 10°N to 10°S, both of which are marked here in the sub-figures by horizontal dotted lines. Also note that this figure shows averages over several months rather than a single week.