12th IFAC International Symposium on
Advanced Control of Chemical Processes
July 14-17, 2024. Toronto, Canada

Discovering Latent Causal Variables Using a Trade-Off Between Compression and Causality[♦]

Xinrui Gao*, Yiman Huang*, Yuri A.W. Shardt*

* *Department of Automation Engineering, Technical University of Ilmenau, Ilmenau, Thuringia, Germany, 98684*

(e-mail: {xinrui.gao, yuri.shardt}@tu-ilmenau.de, echoihuang@gmail.com)

Abstract: Causality is a fundamental relationship in the physical world, around which almost all activities of human life revolve. Causal inference refers to the process of determining whether an event or action caused a specific outcome, which involves the evaluation of cause-and-effect relationships in data. This paper presents a new approach to discover latent causal representations of crucial variables in easy-to-obtain data. The proposed method takes a form of trade-off between compression of input data and the causality between the learnt latent variables and critical variables, thereby removing the irrelevant information contained in input data and obtaining the decoupled, strongest causal factors. By introducing variational bounds and specific configurations, the optimisation objective is relaxed to a tractable problem. The approach compacts causal discovery and inference into one model, which is flexible to downstream tasks and parsimonious in the parameters. A case study on an exhaust-emission dataset shows that the proposed method improves the predictive performance over the baseline model, which is a variational information bottleneck model with the same hyperparameters.

Keywords: causal inference, latent causal variables, soft sensors, information bottleneck, variational inference.

1. INTRODUCTION

Throughout history, the vast majority of human activities have revolved around the fundamental question of causal analysis. Many scientific research problems can be put down to a causal inference problem (Imbens & Rubin, 2015), for instance, on the macrolevel, how an industrial policy affects the national economy, how minimum wages and immigration affect the labour market, and what impact a public policy has on the crime rate. Prominent examples of using causal analysis to solve such problems are Prof. Joshua D. Angrist and Prof. Guido W. Imbens, who won the Nobel Prize for Economics in 2021 for "their methodological contributions to the analysis of causal relationships" (The Royal Swedish Academy of Sciences, 2021). At the individual microlevel, people always wonder how the patient's situation would have been if different treatments had been taken. In the area of process monitoring and fault diagnosis, a very critical question is what is the root cause of a failure (Duan, Chen, Shah, & Yang, 2014), and, for soft sensing, which factors cause fluctuations in the key indicator (Yu, Xiong, Cao, & Fan, 2022).

The most commonly used concept for inferring associations from data is correlation, which is the core of statistical analysis. It is an associational concept that can be defined completely by a joint distribution of the observed variables. However, correlation quantifies only superficial associations wrapped in observations during data generation process that are determined by causal structures of the underlying system. Contrasting with statistical analysis, which in principle estimates the likelihood of events given certain environments, causal analysis seeks to infer the likelihood of events and its dynamics under changing environments (Pearl, 2008). From this point of view, the standard statistical analysis can be regarded as a projection of causal analysis onto one single environment. Whereas, only by knowing the complete information of the likelihood and its dynamics under the whole changing process of environments that the system interact with, can the real causal structures of the system be identified from observations.

In many cases, identifying the causal factors of some critical, hard-to-measure variables is of great interest, since this will result in an accurate and exact model to infer the critical variables from easy-to-measure variables. In this paper, a new method of discovering latent causal factors of critical variables from input variables is proposed. It is formulated as a trade-off between maximising the transfer entropy between the latent causal representations and the critical variables and minimising the mutual information (MI) between the input and the latent causal variables. Hence, the resulting model tends to forget irrelevant information in the massive input data and maximise the causal relationship between the latent representations and the critical variables. The method is verified using an exhaust-emission dataset.

[♦] Xinrui Gao acknowledges funding from the Swiss National Science Foundation under NCCR Automation, Grant # 180545.

2. INFORMATION BOTTLENECK

The proposed method is motivated by the information-bottleneck (IB) theory. Hence, IB and one of variants, the variational information bottleneck (VIB), are reviewed.

2.1 Information Bottleneck

The problem of extracting concise representations of one set of variables to predict target variables is widespread in the machine-learning area. IB solves it by a trade-off between the complexity of the representations and the prediction accuracy of the target variable, which can be formulated mathematically as a variational problem in terms of the stochastic mapping $p(z|x)$, that is (Tishby, Pereira, & Bialek, 1999),

$$\min_{p(z|x)} \mathcal{L} = I(X;Z) - \beta I(Z;Y) \tag{1}$$

where X is the model input variables that are going to be compressed, Z is the resulting representation, Y is the critical variable, β is the trade-off coefficient for the two factors, and $I(\cdot;\cdot)$ is the MI between the random variables defined as

$$I(X;Y) = \mathbb{E}_{p_{X,Y}(x,y)} \log \frac{p_{X,Y}(x,y)}{p_X(x)\,p_Y(y)} \tag{2}$$

To exclude the degenerate solution, β is no larger than 1.

Equation (1) shows that the information of X passed to Y through representation Z is compressed as much as possible and is therefore interpreted as the information "bottleneck". Coefficient β controls the trade-off between compression and accuracy, and can be seen as the Lagrange multiplier of the equivalent problem

$$\begin{aligned} &\min_{p(z|x)} \mathcal{L} = I(X;Z) \\ &s.t. \qquad I(Z;Y) = I_c \end{aligned} \tag{3}$$

where I_c is the least amount of information that is shared between Z and Y.

2.2 Variational Information Bottleneck

Equation (3) is difficult to solve because MI is hard to calculate in general, which limits the application of IB theory to special cases, *e.g.*, discrete variables (Tishby, Pereira, & Bialek, 1999), and jointly variables (Chechik, Globerson, Tishby, & Weiss, 2005). To overcome this, VIB builds a variational bound on Objective (1), making the problem tractable (Alemi, Fischer, Dillon, & Murphy, 2017, April).

Based on Theorem 2.7.4 in (Cover & Thomas, 2006), MI is a concave function of the marginal distribution for a fixed conditional and a convex function of the conditional distribution for a fixed marginal. This gives the bounds of $I(Z;Y)$ and $I(X;Z)$, that is,

$$\begin{aligned} I(Z;Y) &= \mathbb{E}_{p(z,y)} \log \frac{p(\boldsymbol{y}|\boldsymbol{z})}{p(\boldsymbol{y})} \geq \mathbb{E}_{p(z,y)} \log \frac{q(\boldsymbol{y}|\boldsymbol{z})}{p(\boldsymbol{y})} \\ &= \mathbb{E}_{p(z,y)} \log q(\boldsymbol{y}|\boldsymbol{z}) + H(Y) \end{aligned} \tag{4}$$

$$I(X;Z) = \mathbb{E}_{p(x,z)} \log \frac{p(\boldsymbol{z}|\boldsymbol{x})}{p(\boldsymbol{z})} \leq \mathbb{E}_{p(x,z)} \log \frac{p(\boldsymbol{z}|\boldsymbol{x})}{r(\boldsymbol{z})} \tag{5}$$

where $q(\boldsymbol{y}|\, \boldsymbol{z})$ is a variational approximation of the real prediction $p(\boldsymbol{y}|\, \boldsymbol{z})$, and $r(\boldsymbol{z})$ is a known distribution used to approximate (or constrain) the real marginal $p(\boldsymbol{z})$. Since the output entropy $H(Y)$ cannot be influenced by modelling or data processing, we can omit this term. Substituting Equations (4) and (5) into Equation (1), gives the upper bound on the objective function of IB, that is,

$$\begin{aligned} &\mathcal{L} = I(X;Z) - \beta I(Z;Y) \\ &\leq \int p(\boldsymbol{x},\boldsymbol{z}) \log \frac{p(\boldsymbol{z}|\boldsymbol{x})}{r(\boldsymbol{z})} \mathrm{d}\boldsymbol{z}\mathrm{d}\boldsymbol{x} - \beta \int p(\boldsymbol{z},\boldsymbol{y}) \log q(\boldsymbol{y}|\boldsymbol{z}) \mathrm{d}\boldsymbol{z}\mathrm{d}\boldsymbol{y} = \mathcal{L}^u \end{aligned} \tag{6}$$

Then, minimising $\mathcal{L}$ can be relaxed to minimising its upper bound $\mathcal{L}^u$, and is further transformed into

$$\begin{aligned} \min_{p(z|x)} \mathcal{L}^u &= D_{KL}\left(p(\boldsymbol{z}|\boldsymbol{x}) \| r(\boldsymbol{z})\right) \\ &\quad - \beta \int p(\boldsymbol{x},\boldsymbol{y})\, p(\boldsymbol{z}|\boldsymbol{x}) \log q(\boldsymbol{y}|\boldsymbol{z}) \mathrm{d}\boldsymbol{z}\mathrm{d}\boldsymbol{y}\mathrm{d}\boldsymbol{x} \\ &= D_{KL}\left(p(\boldsymbol{z}|\boldsymbol{x}) \| r(\boldsymbol{z})\right) \\ &\quad - \frac{1}{N} \beta \sum_{t=1}^{N} \int p(\boldsymbol{z}|\boldsymbol{x}_t) \log q(\boldsymbol{y}_t|\boldsymbol{z}) \mathrm{d}\boldsymbol{z} \end{aligned} \tag{7}$$

where D_{KL} is the Kullback-Leibler (KL) divergence. The first equality of Equation (7) holds since $p(\boldsymbol{z},\boldsymbol{y}) = \int p(\boldsymbol{z}, \boldsymbol{y}, \boldsymbol{x})\mathrm{d}\boldsymbol{x} = \int p(\boldsymbol{x})p(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{z}|\boldsymbol{x})\mathrm{d}\boldsymbol{x}$ from the Markovian assumption $Y \leftrightarrow X \leftrightarrow Z$ (Alemi, Fischer, Dillon, & Murphy, 2017, April). The second equality is derived based on the empirical distribution approximation $p(\boldsymbol{x}, \boldsymbol{y}) = 1/N \sum_{t=1}^{N} \delta_{\boldsymbol{x}_t}(\boldsymbol{x}) \delta_{\boldsymbol{y}_t}(\boldsymbol{y})$ where δ is the Dirac distribution (Murphy, 2022). To solve Problem (7), three quantities need to be defined: the encoder output distribution $p(\boldsymbol{z}|\, \boldsymbol{x})$, the prior marginal $r(\boldsymbol{z})$ of the latent representation $\boldsymbol{z}$, and the prediction distribution $q(\boldsymbol{y}|\, \boldsymbol{z})$.

3. PROPOSED METHOD

In this section, the details of the proposed method are explained. The objective function is formulated based on a physically meaningful motivation. The original objective is then relaxed to a tractable problem by introducing variational bounds on related quantities and using appropriate configurations.

3.1 Transfer Entropy for Measuring Causality

From the definition in Equation (2), MI quantifies the overlap of the information content between two systems. However, it is incapable of understanding the asymmetric statistical coherence between systems that evolve over time. To answer this question, transfer entropy is proposed to quantify the directional information transfer from one system to another by taking the dynamics of information transport into account (Schreiber, 2000). Mathematically, the transfer entropy from

X to *Y* exclusively measures the response of *Y* to *X*, while excluding that due to confounders and its own history, by conditioning on appropriate transition probabilities, that is (Kaiser & Schreiber, 2002),

$$\begin{aligned} T_{X\to Y}^{k,l} &= \mathbb{E}_{p\left(\boldsymbol{x}_{t-k:t-1},\boldsymbol{y}_{t-l:t-1},y_t\right)}\left(\log\frac{p\left(y_t\middle|\boldsymbol{x}_{t-k:t-1},\boldsymbol{y}_{t-l:t-1}\right)}{p\left(y_t\middle|\boldsymbol{y}_{t-l:t-1}\right)}\right) \\ &= \int\left(\begin{array}{l} \mathrm{d}\boldsymbol{x}_{t-k:t-1}\mathrm{d}\boldsymbol{y}_{t-l:t-1}\mathrm{d}y_t\,p\left(\boldsymbol{x}_{t-k:t-1},\boldsymbol{y}_{t-l:t-1},y_t\right) \\ \log\dfrac{p\left(y_t\middle|\boldsymbol{x}_{t-k:t-1},\boldsymbol{y}_{t-l:t-1}\right)}{p\left(y_t\middle|\boldsymbol{y}_{t-l:t-1}\right)} \end{array}\right) \\ &= I\left(X_{t-k:t-1};Y_t|Y_{t-l:t-1}\right) \end{aligned} \tag{8}$$

where *k* and *l* are time delays, Y_t is the random response variable at time *t*, $X_{t-k:t-1} = [X_{t-k}, \cdots, X_{t-1}]^\top$ is the random driving vector, $Y_{t-l:t-1}$ is also a random vector with similar form, and the letters in lowercase are realisations of the corresponding random variables. It is shown in the definition that the transfer entropy is a special case of the MI between the history of the driving element and the current response element, conditioning on the history of the response element. In short, transfer entropy is a conditional MI.

Based on Wiener's idea, "causality" is defined as follows: one variable (element) could be called "causal" to another if the prediction of the latter is improved by incorporating the information about the former (Wiener, 1956). This conceptual idea of causality is mathematically formulated by Granger in the context of a linear autoregression of stochastic processes (Granger, 1969). Transfer entropy is the nonlinear analogy of Granger causality (Barnett, Barrett, & Seth, 2009). This arises by transforming the definition of the transfer entropy in Equation (8) using the Shannon entropy, that is,

$$\begin{aligned} T_{X\to Y}^{k,l} &= I\left(X_{t-k:t-1};Y_t|Y_{t-l:t-1}\right) \\ &= H\left(Y_t|Y_{t-l:t-1}\right)-H\left(Y_t|Y_{t-l:t-1},X_{t-k:t-1}\right) \end{aligned} \tag{9}$$

In Equation (9), the conditional entropy $H(Y_t| Y_{t-l:t-1})$ indicates the uncertainty in Y_t given its history $Y_{t-l:t-1}$. Likewise, $H(Y_t| Y_{t-l:t-1}, X_{t-k:t-1})$ shows the uncertainty in Y_t given $Y_{t-l:t-1}$ and $X_{t-k:t-1}$. Therefore, the difference between the two terms implies the uncertainty reduction, *i.e.*, the improvement of the prediction of *Y* when introducing the information about *X*.

3.2 Objective Function and System Structure

Intuitively, the motivation of the proposed method is to extract the causal factors of the critical variables and remove the irrelevant information contained in the massive input data, thereby giving a more stable and accurate forecast of the critical variables. This idea can be interpreted as a trade-off between maximising the cause-and-effect relationship between the easy-to-obtain variables and the critical variables, and minimising the complexity of the latent causal variables. Mathematically, the problem can be formulated as seeking a mapping $p(z| x)$ to maximise the transfer entropy between the critical and the input variables, while minimising the MI between the input and the latent variables, that is,

$$\min_{p(z|x)} \mathcal{L} = I\left(X;Z\right)-\beta T_{Z\to Y}^{k,l}\left(Z;Y\right) \tag{10}$$

where β is the trade-off coefficient.

To simplify Problem (10), the underlying system structure is assumed to be similar to a hidden Markov model (HMM) (Eddy, 1998), as shown in Figure 1. The relationship between *X* and *Z*, shown by arrows in Figure 1, is symmetric as it is modelled by MI, while the information flow from *Z* to *Y* is directed by transfer entropy $T_{Z\to Y}$. As well, the flow of time is also unidirectional from the past to the future. Several substructures of conditional independence are contained in the graphical model, *e.g.*, $Y_t \perp X_t \,|\, Z_t$, $Y_t \perp Z_{t-1} \,|\, Z_t$, and $Z_{t+1} \perp Z_{t-1} \,|\, Z_t$, which reduce largely the complexity of the decomposition of the joint distribution.

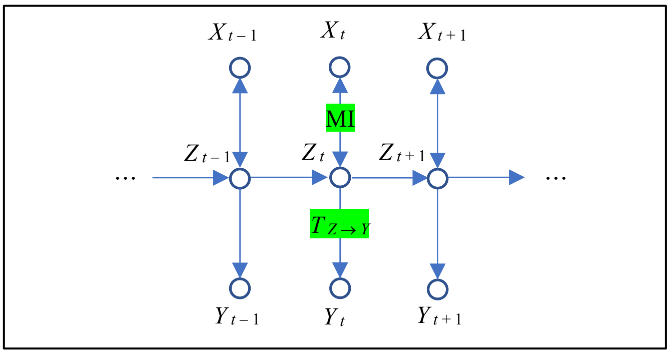

Figure 1. Graphical structure of the underlying systems

In this system structure, all information that drives the response element *Y* comes from *Z*, which means *l* should not be larger than *k*. These conditions simplify the transfer entropy to

$$\begin{aligned} T_{Z\to Y}^{k,l}\left(Z;Y\right) &= \mathbb{E}_{p\left(\boldsymbol{y}_{t-l:t},\boldsymbol{z}_{t-k:t-1}\right)}\left(\log\frac{p\left(y_t\middle|\boldsymbol{z}_{t-k:t-1},\boldsymbol{y}_{t-l:t-1}\right)}{p\left(y_t\middle|\boldsymbol{y}_{t-l:t-1}\right)}\right) \\ &= \mathbb{E}_{p\left(y_{t-1},y_t,z_{t-1}\right)}\left(\log\frac{p\left(y_t\middle|z_{t-1},y_{t-1}\right)}{p\left(y_t\middle|y_{t-1}\right)}\right) = I\left(Z_{t-1};Y_t|Y_{t-1}\right) \end{aligned} \tag{11}$$

Equation (11) shows that the HMM-like structure indicates that $l = k = 1$. Substituting $p(y_t| z_{t-1}, y_{t-1}) = p(y_t| z_{t-1})$ into Equation (11) gives

$$\begin{aligned} T_{Z\to Y}^{k,l}\left(Z;Y\right) &= \mathbb{E}_{p\left(y_{t-1},y_t,z_{t-1}\right)}\left(\log\frac{p\left(y_t\middle|z_{t-1}\right)p\left(y_t\right)}{p\left(y_t\right)p\left(y_t\middle|y_{t-1}\right)}\right) \\ &= I\left(Y_t;Z_{t-1}\right)-I\left(Y_t;Y_{t-1}\right) \end{aligned} \tag{12}$$

Since $I(Y_t; Y_{t-1})$ cannot be changed by data processing, it is omitted, thereby simplifying the objective function (10) to

$$\min_{p(z|x)} \mathcal{L} = I\left(X;Z\right)-\beta I\left(Y_t;Z_{t-1}\right) \tag{13}$$

3.3 Relaxation of the Objective Function

The simplified objective in Equation (13) cannot be calculated directly since the distributions are unknown in general. To solve the problem, a tractable upper bound $\mathcal{L}^u$ on the objective function is obtained, and the problem is relaxed to minimise the upper bound. The relaxed problem is a sufficient condition for the original, that is,

$$\min_{p(z|x)} \mathcal{L} = I(X;Z) - \beta I(Y_t;Z_{t-1}) \Leftarrow \min_{p(z|x)} \mathcal{L}^u \tag{14}$$

Using

$$\begin{aligned} p(y_t|z_{t-1}) &= \int p(y_t|z_t)\, p(z_t|z_{t-1})\,\mathrm{d}z_t \\ &= \mathbb{E}_{p(z_t|z_{t-1})} p(y_t|z_t) \end{aligned} \tag{15}$$

the second MI in Equation (13) can be rewritten as

$$\begin{aligned} I(Y_t;Z_{t-1}) &= \mathbb{E}_{p(y_t,\, z_{t-1})} \log \frac{p(y_t|z_{t-1})}{p(y_t)} \\ &= \mathbb{E}_{p(y_t,\, z_{t-1})} \log \frac{\mathbb{E}_{p(z_t|z_{t-1})} p(y_t|z_t)}{p(y_t)} \\ &\geq \mathbb{E}_{p(y_t,\, z_{t-1})} \mathbb{E}_{p(z_t|z_{t-1})} \log \frac{p(y_t|z_t)}{p(y_t)} = \mathbb{E}_{p(z_t|z_{t-1})} I(Y_t;Z_t) \end{aligned} \tag{16}$$

Since $p(y_t|z_t)$ is unknown, analogous to Equation (4), a variational approximation $q(y_t|z_t)$ is introduced. This gives a looser lower bound on $I(Y_t; Z_{t-1})$, that is,

$$\begin{aligned} I(Y_t;Z_{t-1}) &\geq \mathbb{E}_{p(z_t|z_{t-1})} I(Y_t;Z_t) \\ &\geq \mathbb{E}_{p(z_t|z_{t-1})} \mathbb{E}_{p(y_t,\, z_t)} \log \frac{q(y_t|z_t)}{p(y_t)} \\ &= \mathbb{E}_{p(z_t|z_{t-1})} \mathbb{E}_{p(y_t,\, z_t)} \log q(y_t|z_t) + H(Y_t) \end{aligned} \tag{17}$$

The entropy $H(Y_t)$ can be ignored since it is a constant. Recalling the bound on $I(X; Z)$ derived by Equation (5), the upper bound appearing in Equation (14) is

$$\begin{aligned} \mathcal{L} &= I(X;Z) - \beta I(Y_t;Z_{t-1}) \\ &\leq D_{KL}\big(p(z|x) \| r(z)\big) - \beta \mathbb{E}_{p(z_t|z_{t-1})} \mathbb{E}_{p(y_t,\, z_t)} \log q(y_t|z_t) = \mathcal{L}^u \end{aligned} \tag{18}$$

This upper bound is tractable after configuring the related distributions, which is presented in the next section.

3.4 Configuration of the Proposed Method

The proposed method is implemented using the structure of auto-encoders, which consists of an encoder $p_\phi(Z| X)$ and a decoder $q_\theta(Y| Z)$, where ϕ and θ are the parameters of the corresponding neural networks f_ϕ and f_θ. Since

$$\begin{aligned} p(y_t, z_t) &= \int p(y_t, z_t, x_t)\,\mathrm{d}x_t \\ &= \int p(y_t, x_t)\, p(z_t|x_t)\,\mathrm{d}x_t \end{aligned} \tag{19}$$

the second term of $\mathcal{L}^u$ can be rewritten as

$$\begin{aligned} &\mathbb{E}_{p(z_t|z_{t-1})} \mathbb{E}_{p(y_t,\, z_t)} \log q(y_t|z_t) \\ &= \mathbb{E}_{p(z_t|z_{t-1})} \int p(y_t, x_t)\, p(z_t|x_t) \log q(y_t|z_t)\,\mathrm{d}x_t \mathrm{d}y_t \mathrm{d}z_t \\ &= \mathbb{E}_{p(z_t|z_{t-1})} \frac{1}{N} \sum\nolimits_{i=1}^{N} \int p\left(z_t \middle| x_t^{(i)}\right) \log p\left(y_t^{(i)} \middle| z_t\right) \mathrm{d}z_t \end{aligned} \tag{20}$$

The last equation holds because of the empirical approximation (Murphy, 2022)

$$p(y_t, x_t) = \frac{1}{N} \sum\nolimits_{i=1}^{N} \delta\left(y_t^{(i)}\right) \delta\left(x_t^{(i)}\right) \tag{21}$$

Unlike the independently distributed Z in VIB, $\{Z_t\}$ is a stochastic process in this method, which means the dynamics need to be considered. Hence, it is a sum of two sources: the input X_t from the encoder $p_\phi(Z_t| X_t)$, and its history from the dynamic evolution $p(Z_t| Z_{t-1})$. The first part can be expressed using the reparameterisation trick (Kingma & Welling, 2014), that is,

$$Z_{t,\phi} = f_\phi(X_t, \boldsymbol{e}) = \boldsymbol{u}_\phi(X_t) + \boldsymbol{\sigma}_\phi(X_t) \circ \boldsymbol{e} \tag{22}$$

where $\boldsymbol{u}_\phi$ and $\boldsymbol{\sigma}_\phi$ are the actual deterministic outputs of f_ϕ, and refer, respectively to the centre and scale of $p_\phi(Z_t| X_t)$, the symbol $\circ$ refers to the elementwise product, and $\boldsymbol{e}$ is white noise. The reparameterisation trick equates a Monte Carlo estimation of Equation (20) by sampling Z_t over $p_\phi(Z_t| X_t)$ and the estimation by sampling $\boldsymbol{e}$ over $p(\boldsymbol{e})$, thereby overcoming the difficulty of gradient backpropagation of directly sampling over $p_\phi(Z_t| X_t)$. Similar to Equation (22), the dynamic evolution for the second part is written as

$$Z_{t,t-1} = f(Z_{t-1}, \boldsymbol{\varepsilon}) = \boldsymbol{u}(Z_{t-1}) + \boldsymbol{\sigma}(Z_{t-1}) \circ \boldsymbol{\varepsilon} \tag{23}$$

where $f(\cdot)$ is the dynamic transition function, $\boldsymbol{u}$ and $\boldsymbol{\sigma}$ are the centre and scale of $p(Z_t| Z_{t-1})$, and $\boldsymbol{\varepsilon}$ is white noise.

Given a standard Gaussian distribution as the prior marginal $r(Z_t) = \mathcal{N}(0, \mathbf{I})$, the latent process $\{Z_t\}$ in Figure 1 is

$$\left\{ Z_t \left| \begin{array}{l} Z_t = f_\phi(X_t, \boldsymbol{e}) + f(Z_{t-1}, \boldsymbol{\varepsilon}) \\ \boldsymbol{\varepsilon} \sim \mathcal{N}\left(0, \Sigma_\varepsilon = \mathrm{diag}\left\{\sigma_{\varepsilon_1}^2, \cdots, \sigma_{\varepsilon_m}^2\right\}\right), \\ \boldsymbol{e} \sim \mathcal{N}\left(0, \Sigma_e = \mathrm{diag}\left\{\sigma^2, \cdots, \sigma^2\right\}\right), \end{array} \right. \right\} \tag{24}$$

Finally, combining Equations (18), (20), and (24) gives

$$\begin{aligned} \mathcal{L}^u &= D_{KL}\big(p(z|x) \| r(z)\big) \\ &- \frac{1}{N} \beta \sum\nolimits_{i=1}^{N} \int p(z_t|z_{t-1})\, p\left(z_t \middle| x_t^{(i)}\right) \log p\left(y_t^{(i)} \middle| z_t\right) \mathrm{d}z_t \mathrm{d}z_t \\ &= D_{KL}\big(p(z|x) \| r(z)\big) \\ &- \frac{1}{N} \beta \sum\nolimits_{i=1}^{N} \mathbb{E}_{p(\varepsilon)} \mathbb{E}_{p(e)} \log p\left(y_t^{(i)} \middle| z_t = f_\phi\left(x_t^{(i)}, \boldsymbol{e}\right) + f(z_{t-1}, \boldsymbol{\varepsilon})\right) \end{aligned} \tag{25}$$

In Equation (25), the first term forces the extracted latent variable to approach the predefined prior, which can be regarded as a regularisation. The second term is actually the negative prediction error of the target variable Y. While implementing the model, the unknown dynamic transition function $f(\cdot)$ can be cancelled out by integrating it into the decoder net f_θ. This causes no loss in terms of the prediction of target variable Y. Specifically, the prediction of Y can be the integration of the two parts shown in Equations (22) and (23) by the decoder, that is,

$$\begin{aligned}\hat{y}_t &= \mathbb{E}_{p(Y|Z)} g\left(z_t\right) = \mathbb{E}_{p(e),p(\varepsilon)} g\left(z_t = f_\phi\left(x_t^{(i)}, \boldsymbol{e}\right) + f\left(z_{t-1}, \boldsymbol{\varepsilon}\right)\right) \\ &= \mathbb{E}_{p(e)} f_\theta\left(\begin{bmatrix} z_{t,\phi} & z_{t-1,\phi} \end{bmatrix}\right)\end{aligned} \tag{26}$$

where $g(\cdot)$ is the underlying mapping function, and $[\cdot\ \ \cdot]$ is an augmentation of the corresponding variables. If the decoder is Gaussian, the second term of Equation (25) is equivalent to mean squared error, which can be calculated based on Equation (26). In addition, the noise is $\boldsymbol{e} \sim \mathcal{N}(0, \mathbf{I})$, and the distribution of the encoder is specified as $p_\phi\ (Z|\ X) = \mathcal{N}(\boldsymbol{u}_\phi(X), \boldsymbol{\Sigma}_\phi(X))$, which can be reparameterised in the form of Equation (22). Thus, Equation (25) is finally a tractable objective for the proposed method.

4. CASE STUDY

In this section, the proposed method is used to model the exhaust emission of a gas turbine, that is, predict the concentration of pollutants in the exhaust emission.

4.1 Dataset at a Glance

The dataset is collected from a gas turbine in a power plant, which is composed of hourly average measurements of eleven variables. The first nine variables are input variables and the remaining two are critical target variables. Here, we only focus on the first one, *i.e.*, the carbon-dioxide (CO_2) concentration in the exhaust. The complete dataset contains 36,733 samples over five years, of which the first 60% are used for model training, the next 20% for adjusting the hyperparameters, and the remainder as the test set. The basic information about the dataset is shown in Table 1. The dataset is downloaded from http://www.e-adys.com/datasets/pp_gas_emission.zip. More information can be found in Kaya, Tüfekci, and Uzun (2019).

Table 1. Information on the exhaust dataset

Variables	Min	Mean	Max
AT: ambient temperature (°C)	–6.23	17.71	37.10
AP: ambient pressure (mbar)	985.85	1013.07	1036.56
AH: ambient humidity (%)	24.08	77.87	100.20
AFDP: air filter difference pressure (mbar)	2.09	3.93	7.61
GTEP: gas turbine exhaust pressure (mbar)	17.70	25.56	40.72
TIT: turbine inlet temperature (°C)	1000.85	1081.43	1100.89
TAT: turbine after temperature (°C)	511.04	546.16	550.61
DCP: compressor discharge pressure (mbar)	9.85	12.06	15.16
TEY: turbine energy yield (MWH)	100.02	133.51	179.50
CO: carbon monoxide (mg/m^3)	0.00	2.37	44.10
NO_X: nitrogen oxides (mg/m^3)	25.90	65.29	119.91

4.2 Model Implementation and Analysis of the Results

The proposed method is implemented by an auto-encoder, of which the encoder is a gated-recurrent-unit (GRU) followed by a fully connected (FC) net, and the predictor[1] is a simple FC net. For comparison, a VIB model is also built, which is used as the baseline. Figure 2 shows the pipelines of information flow of the VIB and the proposed method. All hyperparameters for both models are described in Table 2. The two models have the same hyperparameters and architectures except for the input dimension of the predictor, which is double that of the baseline. Hence, the only difference in the two models is the configuration for the latent variable/process $\{Z_t\}$, which is reflected in the input dimension of f_θ, and results from the fundamental difference in the approach to modelling.

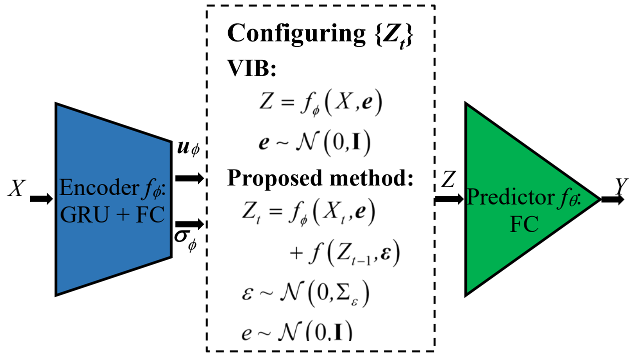
Configuring $\{Z_t\}$
VIB:
$Z = f_\phi(X, \boldsymbol{e})$
$\boldsymbol{e} \sim \mathcal{N}(0, \mathbf{I})$
Proposed method:
$Z_t = f_\phi(X_t, \boldsymbol{e}) + f(Z_{t-1}, \boldsymbol{\varepsilon})$
$\varepsilon \sim \mathcal{N}(0, \Sigma_\varepsilon)$
$e \sim \mathcal{N}(0, \mathbf{I})$
X
Encoder f_ϕ: GRU + FC
$\boldsymbol{u}_\phi$
$\boldsymbol{\sigma}_\phi$
Z
Predictor f_θ: FC
Y

Figure 2. Model pipelines of VIB and the proposed method

Model performance is evaluated using the coefficient of determination (R^2) and the mean squared error (MSE). The results under different trade-off coefficients β are shown in Table 3. It can be seen that the optimal β for the proposed model is 1×10^5, while for VIB, it is either 1×10^5 or 1×10^3 depending on the metric. Furthermore, the proposed method has a better performance based on the two metrics than VIB under almost all the β. Because the two models have the same architectures and hyperparameters, this indicates that incorporating causality improves predictive performance.

[1] The name *predictor* is used instead of *decoder*, since here, the decoder is used for predicting the critical variable.

Table 2. Hyperparameters of VIB and the proposed model

Hyperparameters	Value
Iteration epoch	500
Learning rate	1×10^{-4}
Batch size	150
Sequence length	20
Number of GRU layers	2
Dimension of GRU output	128
Number of FC layers in f_ϕ	2
Dimension of the output of f_ϕ	128
Number of FC layers in f_θ	2
Dimension of the output of f_θ	1

Table 3. Model performance of the two models for different β

β	Metrics	VIB	Proposed model
1×10^2	R^2	0.470	0.587
	MSE	0.121	0.108
1×10^3	R^2	0.592	0.602
	MSE	**0.100**	0.109
1×10^4	R^2	0.600	0.618
	MSE	0.106	0.103
1×10^5	R^2	**0.608**	**0.637**
	MSE	0.106	**0.092**

5. CONCLUSIONS

This paper proposes a new method of discovering latent causal representations of crucial variables using easy-to-obtain data. The optimisation objective is relaxed to a tractable problem by introducing variational bounds and appropriate configurations. The resulting model tends to forget irrelevant information in the massive input data and maximise the causal relationship between the latent representations and the critical variables. A case study on an exhaust dataset shows that the proposed method improves the predictive performance. For the future, the performance of causal discovery needs to be verified. As well, different system structures should be explored to obtain more general relaxations of the objective function.

REFERENCES

Alemi, A. A., Fischer, I., Dillon, J. V., & Murphy, K. (2017, April). Deep Variational Information Bottleneck. *The 5th International Conference on Learning Representations.* Toulon, France.

Barnett, L., Barrett, A. B., & Seth, S. K. (2009). Granger causality and transfer entropy are equivalent for Gaussian variables. *Physical review letters, 103*(23), 238701.

Chechik, G., Globerson, A., Tishby, N., & Weiss, Y. (2005). Information bottleneck for Gaussian variables. *The Journal of Machine Learning Research, 6*(1), 165-188.

Cover, T. M., & Thomas, J. A. (2006). *Elements of information theory* (2nd Edition ed.). Wiley-Interscience.

Duan, P., Chen, T., Shah, S. L., & Yang, F. (2014). Methods for root cause diagnosis of plant-wide oscillations. *AIChE Journal, 60*(6), 2019-2034.

Eddy, S. R. (1998). Profile hidden Markov models. *Bioinformatics, 14*(9), 755-763.

Granger, C. W. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica, 37*(3), 424-438.

Imbens, G. W., & Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences.* Cambridge University Press.

Kaiser, A., & Schreiber, T. (2002). Information transfer in continuous processes. *Physica D: Nonlinear Phenomena, 166*(1), 43-62.

Kaya, H., Tüfekci, P., & Uzun, E. (2019). Predicting CO and NOx emissions from gas turbines: novel data and a benchmark PEMS. *Turkish Journal of Electrical Engineering and Computer Sciences, 27*(6), 4783-4796.

Kingma, D. P., & Welling, M. (2014). Auto-Encoding Variational Bayes. *The 2nd International Conference on Learning Representations, ICLR 2014.* Banff, AB, Canada.

Murphy, K. P. (2022). *Probabilistic machine learning: an introduction.* MIT Press.

Pearl, J. (2008). Causal Inference. *In Proceedings of Workshop on Causality: Objectives and Assessment at NIPS 2008* (pp. 39-58). PMLR.

Schreiber, S. (2000). Measuring information transfer. *Physical review letters, 85*(2), 461-464. doi:10.1103/PhysRevLett.85.461

Shannon , C. E. (1959). Coding Theorems for a Discrete Source with a Fidelity Criteria. *IRE National Convention Record, 4*(1), 142-163.

The Royal Swedish Academy of Sciences. (2021, October 11). *The Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel 2021.* Retrieved from The Nobel Prize: https://www.nobelprize.org/prizes/economic-sciences/2021/press-release/

Tishby, N., Pereira, F. C., & Bialek, W. (1999). The information bottleneck method. *In Proceedings of the 37th Annual Allerton Conference on Communications, Control, and Computing*, (pp. 368-377).

Wiener, N. (1956). The Theory of Prediction. In E. F. Beckenbach (Ed.), *Modern Mathematics for the Engineer.* New York: McGraw-Hill.

Yu, F., Xiong, Q., Cao, L., & Fan, Y. (2022). Stable soft sensor modeling based on causality analysis. *Control Engineering Practice, 122*, 105109.