Available online at www.sciencedirect.com

ScienceDirect

IFAC Papers Online CONFERENCE PAPER ARCHIVE

IFAC PapersOnLine 55-20 (2022) 43–48

EVOLVE·INFOMAX: A New Criterion for Slow Feature Analysis of Nonlinear Dynamic System from an Information-Theoretical Perspective

Xinrui Gao*, 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).*

Abstract: Slow feature analysis (SFA) has attracted much attention as a method for dynamic modelling. However, SFA has an inherent limitation in that it assumes that the dynamic behaviour is linear. In this paper, a new criterion for SFA in general dynamic systems is defined based on the motivation of maximising the information retained during system evolution, which is called EVOLVE·INFOMAX. The theoretical properties of this new criterion are rigorously justified, the optimisation function under EVOLVE·INFOMAX is proposed, and a tailored algorithm based on neural networks is designed. The case study on a simulated data set and the Tennessee Eastman process benchmark shows that the proposed method has better performance to extract slow features of nonlinear dynamical systems.

Keywords: **slow feature analysis, variational inference, mutual information, dynamic modelling, neural networks**

1. INTRODUCTION

In the big-data era, one of the crucial tasks is to represent the high dimensional and large volume data in a compact and parsimonious way, while retaining meaningful information. The extracted features should be optimal in the sense of a predefined criterion that measures the information of interest, while disentangled, and thus, containing different aspects of information. This makes the resulting model parsimonious. For data with time dependence, the dynamic structure is useful information to understand the data and the underlying system. Slow feature analysis (SFA) is a dynamic modelling method that seeks to find the slowly varying latent features (Wiskott & Sejnowski, 2002). It has gained much attention and been used in many fields, *e.g.*, understanding the complex mechanism of the visual cortex cell reaction to visual signals (Berkes & Wiskott, 2005), online time-series analysis (Kompella, et al., 2012), blind source separation (Sprekeler, et al., 2014), process monitoring and fault diagnosis (Gao & Shardt, 2021), as well as soft sensing and quality forecasting (Shang, et al., 2015). However, the original definition of SFA has the inherent limitation, in that it is based on the assumption of linear dynamic behaviour.

In the paper, a new criterion for SFA of general dynamic systems, which is called EVOLVE·INFOMAX, is defined based on the motivation of maximizing the evolutionary information during system evolution process. The theoretical properties of EVOLVE·INFOMAX are justified and the equivalence to the original SFA under linear Gaussian systems is analysed. The objective function of the generalised SFA is derived based on EVOLVE·INFOMAX and a tailored algorithm based on artificial neural networks (NN) is designed to solve the problem. Finally, the results on a simulated data set and the Tennessee Eastman process (TEP) benchmark are presented.

2. ORIGINAL SLOW FEATURE ANALYSIS

SFA is an unsupervised model to extract dynamic latent features $\boldsymbol{s}(t) = [s_1(t) \cdots s_k(t)]$ from the observation signal $\boldsymbol{x}(t) = [x_1(t) \cdots x_k(t)]$. The extracted variables are called slow features (SFs) due to their slowly time-varying property. This can be achieved by minimising the expectation of the squared first-order difference of the SF sequence, that is (Wiskott & Sejnowski, 2002)

$$\begin{aligned} &\min_{g_j(\cdot)} \left\langle \dot{s}_j^2(t) \right\rangle \\ \text{s.t.} \quad &1).\ \left\langle s_j(t) \right\rangle = 0, \\ &2).\ \left\langle s_j{}^2(t) \right\rangle = 1, \\ &3).\ \forall i \neq j,\ \left\langle s_i(t) s_j(t) \right\rangle = 0 \end{aligned} \tag{1}$$

where $g_j(\cdot)$ is the input-output function that needs to be found, $\dot{s}(t) = d(s(t))/dt \approx s(t) - s(t-1)$ is the first-order derivative/difference, and $\langle\cdot\rangle$ is the time averaging operator. Constraint 1) centres the SF sequences to simplify the problem. Constraints 2) and 3) whiten the SF variables to be isotropic and with unit variances. The slowness $\Delta(s)$ in Equation (1) can be translated into (Blaschke, et al., 2006)

$$\Delta(s) = \left\langle \dot{s}^2(t) \right\rangle = \left\langle \left(s(t) - s(t-1) \right)^2 \right\rangle = 2 - 2\lambda \tag{2}$$

where λ is the first-order autocorrelation of sequence $\{s(t)\}$, which is a measure of the linear autocorrelation. Equation (2) clearly shows that the definition of the slowness in the original SFA is only suitable for modelling linear dynamic behaviour. It has inherent limitations for general dynamic systems with both linear and nonlinear time dependencies. Furthermore, SFA treats the SF sequence as a deterministic

10.1016/j.ifacol.2022.09.069

time series rather than a stochastic process. The effect of system noise is not taken into consideration. This paper proposes the EVOLVE·INFOMAX criterion from the information-theoretic perspective for slow feature extraction of general dynamic systems, which can grasp the full structure of the temporal relationships and automatically incorporate the influence of noise on the system dynamics. Unlike the original definition that is a moment-based approach, EVOLVE·INFOMAX directly uses the underlying distributions.

3. SLOWNESS REDEFINITION: INVARIABILITY OF A SYSTEM DURING THE EVOLUTION PROCESS

This section defines EVOLVE·INFOMAX, justifies its properties, and shows its equivalence to the original SFA.

3.1 EVOLVE·INFOMAX: A New Criterion for SFA for General Dynamic Systems

Intuitively, slow means some amount of invariability related to time. For a dynamic system that evolves in the time domain, slowness can be regarded as a measurement of the invariability of system states over time. To be more specific, this invariability can be interpreted as the amount of information the system states retained during system evolution, or in other words, the information transmitted from the previous state to the current one. Generally speaking, the more information from the historical states is contained in the current states, the more significant the invariability of the system, and thus, the slower the system changes. Therefore, the slowness of a dynamic system state can be conceptually defined as *the invariability of the system state in the time domain, or the amount of state information preserved during process of system evolution*.

Corresponding to the Pearson's correlation for linear dependency, mutual information (MI) measures the complete structure of the dependency including linear and nonlinear components between two random variables. Hence, for a general, time-discrete dynamic system, slowness can be defined as maximising the amount of information of a state transmitted between successive time instants. Given two random variables X and Y, MI is defined as the Kullback-Leibler (KL) divergence between the joint and the product of the marginals, that is,

$$I(X;Y) = \mathbb{E}_{p(x,y)} \log \frac{p(y \mid x)}{p(y)} = H(Y) - H(Y \mid X) \quad (3)$$

where $p(\cdot)$, $p(\cdot,\cdot)$, $p(\cdot|\cdot)$, $H(\cdot)$, and $H(\cdot|\cdot)$ are the corresponding marginal, joint, conditional distributions, entropy, and conditional entropy. It has been proved that $0 \le I(X; Y) \le \log |\mathcal{X}| < +\infty$ (Cover, 1999).

Define a general dynamic system

$$\begin{cases} \boldsymbol{s}(t) = \boldsymbol{f}\left(\boldsymbol{s}(t-1)\right) + \boldsymbol{e}(t), \boldsymbol{e}(t) \sim p(\boldsymbol{e}) \\ \boldsymbol{x}(t) = \boldsymbol{g}\left(\boldsymbol{s}(t)\right) + \boldsymbol{\varepsilon}(t), \boldsymbol{\varepsilon}(t) \sim p(\boldsymbol{\varepsilon}) \end{cases} \quad (4)$$

where $\boldsymbol{s}$ is the system states that are mutually independent, $\boldsymbol{x}$ is the observation measurements, $\boldsymbol{e}$ and $\boldsymbol{\varepsilon}$ are the system and observation white noise with general distribution forms and independent elements, and the system and observation functions $\boldsymbol{f}$ and $\boldsymbol{g}$ are both vector-valued functions. We assume $\boldsymbol{f}$ to be a deterministic function such that, given the previous states, the uncertainty of the system comes only from the additive system noise. As well, $\boldsymbol{f}$ is assumed invertible and continuous. Then, the information retention of a state between the successive time instant during evolution can be measured by $I(s(t); s(t-1))$, where s is one of the states in $\boldsymbol{s}$, and the subscript is omitted for simplicity.

For the convenience of analysis and comparison with the original SFA, the generalized slowness can also be defined as $\Delta(s) = e^{-2I(s(t);\, s(t-1))}$. It follows that $\Delta(s) \in [e^{-2H(s(t))}, 1]$. When $\Delta(s)$ equals one, this means $s(t)$ is not the slow feature, while as $\Delta(s)$ approaches $e^{-2H(s(t))}$, $s(t)$ varies slowly. This is similar to the original definition.

3.2 Properties of the EVOLVE·INFOMAX Criterion

After giving the formal definition of the slowness for the generalized SFA, several important results of the EVOLVE·INFOMAX criterion are presented in the following lemma and theorems, from which the optimisation objective is derived. The only assumption for the properties to hold is that the dynamic system (4) be stationary.

Lemma 1: *For the general dynamic system given in (4), the uncertainty of the system state* H(s(t)) *at time instant* t *consists of two parts: the uncertainty reduction* I(s(t); s(t–1)) *through system evolution given the previous state* s(t – 1)*, and the uncertainty of the system noise* H(e)*.*

Proof: Since $p(s(t)|s(t-1)) = p(e)$ and Equation (3), we have

$$\begin{aligned} I\left(s(t); s(t-1)\right) &= H\left(s(t)\right) - H\left(s(t) \mid s(t-1)\right) \\ &= H\left(s(t)\right) - H(e) \end{aligned} \quad (5)$$

Thus, Lemma 1 is proved. *Q.E.D.*

Lemma 1 gives a coarse skeletal relationship between the system evolution and the system noise. Theorem 1 will further describe the noise suppression property and the extreme conditions of EVOLVE·INFOMAX. It bridges the system characteristics and the model behaviour.

Theorem 1: *(Noise suppression property and the extreme conditions) Under the assumption of stationarity, the generalised SFA based on EVOLVE·INFOMAX suppresses noise. The slowness follows that* $\Delta(s) \in [e^{-2H(s(t))}, 1]$*. Under the extreme condition of* $\Delta(s) = 1$*, the extracted SF is a white noise process. In the other extreme of* $\Delta(s) = e^{-2H(s(t))}$*, the system function* f *is a general identical map whose Jacobian equals one, and the state process will become* $\{s(t) = s(0)\}$ *or* $\{s(t) = (-1)^t s(0)\}$.

Proof: If the system state $s(t)$ is stationary, the entropy of the state at each time instant is a constant $H(s(t)) = \mathcal{C}$ (see

Chapter 4 of (Cover, 1999)). From Lemma 1, the slower the state varies (namely, the smaller the slowness $\Delta(s)$), the larger the uncertainty reduction $I(s(t), s(t-1))$ is, and thus, the smaller the system noise uncertainty $H(e)$.

Since $I(s(t); s(t-1)) \in [0, H(s(t))]$, the slowness follows $\Delta(s) \in [e^{-2H(s(t))}, 1]$. In the extreme of maximum slowness, that is $\Delta(s) = 1$ (the fastest feature), we have $I(s(t); s(t-1)) = 0$, which means $s(t)$ is independent of $s(t-1)$ and $H(s(t)) = H(e)$. The corresponding system equation of the system (4) becomes $s(t) = e(t)$, namely, a white noise process. For the other extreme $\Delta(s) = e^{-2H(s(t))}$, it implies that $I(s(t), s(t-1)) = H(s(t))$, and thus, $H(e) = 0$. Then, the system equation in (4) will become a deterministic transition process $\{s(t) = f(s(t-1))\}$. For an invertible, continuous function f, the density of $s(t)$ can be obtained as

$$p\left(s\left(t\right)\right) = p\left(f\left(s\left(t-1\right)\right)\right) = p\left(s\left(t-1\right)\right)\frac{1}{\left|J\right|_{s(t-1)=f^{-1}(s(t))}} \tag{6}$$

where $J = \dfrac{ds(t)}{ds(t-1)}$ is the Jacobian of the system function f.

Since $\{s(t)\}$ is stationary, we have $H(s(t)) = H(s(t-1)) = \mathcal{C}$. Expanding the entropy expression using Equation (6) gives

$$H\left(s\left(t\right)\right) = H\left(s\left(t-1\right)\right) + \log\left|J\right| = H\left(s\left(t-1\right)\right) \tag{7}$$

Therefore, we have $|J| = 1$, which further limits f to a general identical map that yields the state process $\{s(t) = s(t-1)\}$ or $\{s(t) = -s(t-1)\}$. Therefore, given the initial state $s(0)$, the state at any time instant t will be determined as $s(t) = s(0)$ or $s(t) = (-1)^t s(0)$. *Q.E.D.*

Lemma 1 and Theorem 1 can be schematically presented in Figure 1. The line with a gradient colour represents Equation (5). The generalized SFA tends to extract the features lying on the green part of the line.

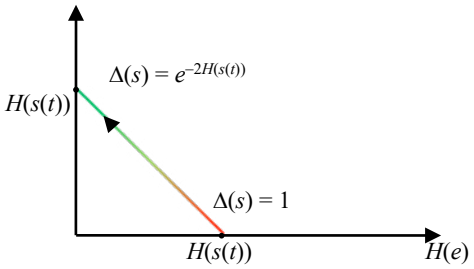

Figure 1: Schematic interpretation of Lemma 1 and Theorem 1.

Theorem 2 states another important property of the system transition function f under EVOLVE·INFOMAX.

Theorem 2: *If the entropy* H(s(t)) *of a stationary state* s(t) *of system (4) exists, the corresponding system function* f *is Lipschitz continuous, that is,* $|J| \le K < e^{H(s(t))}$.

Proof: The independence of $s(t-1)$ and $e(t)$ in (4) implies that

$$p\left(s\left(t\right)\right) = p\left(f\left(s\left(t-1\right)\right) + e\left(t\right)\right) = \frac{1}{\left|J\right|}p\left(s\left(t-1\right)\right) * p\left(e\right) \tag{8}$$

where $*$ is convolution. Then, $H(s(t))$ can be transformed to

$$\begin{aligned} H\left(s\left(t\right)\right) &= \log\left|J\right| - \mathbb{E}_{p(s(t-1),e)}\left(\log\left(p\left(s\left(t-1\right)\right) * p\left(e\right)\right)\right) \\ &= I\left(s\left(t\right); s\left(t-1\right)\right) + H\left(e\right) \end{aligned} \tag{9}$$

with the last equals sign satisfied due to Lemma 1. Hence, we obtain Equation (12). It can be concluded that $\log|J| \le I(s(t); s(t-1))$ as the KL divergence is nonnegative. Therefore, due to $I(s(t); s(t-1)) \le H(s(t))$, the Jacobian of f is bounded as

$$\left|J\right| \le e^{I\left(s(t);s(t-1)\right)} \le e^{H\left(s(t)\right)} \tag{10}$$

with the first inequality being equality if

$$p\left(e\right) = p\left(s\left(t-1\right)\right) * p\left(e\right) \tag{11}$$

$$\begin{aligned} &\log\left|J\right| = I\left(s\left(t\right); s\left(t-1\right)\right) + H\left(e\right) + \mathbb{E}_{p(s(t-1),e)}\left(\log\left(p\left(s\left(t-1\right)\right) * p\left(e\right)\right)\right) \\ &= I\left(s\left(t\right); s\left(t-1\right)\right) - \mathbb{E}_{p(s(t-1),e)}\left(\log\frac{p\left(e\right)}{p\left(s\left(t-1\right)\right) * p\left(e\right)}\right) = I\left(s\left(t\right); s\left(t-1\right)\right) - \mathbb{E}_{p(s(t-1))}\left(D_{KL}\left(p\left(e\right)\middle\|p\left(s\left(t-1\right)\right) * p\left(e\right)\right)\right) \end{aligned} \tag{12}$$

At such a point, D_{KL} reaches its minimum $D_{KL} = 0$. Let the characteristic functions of $p(s(t-1))$ and $p(e)$ be $\varphi_{s(t-1)}(\gamma) = \mathbb{E}_{p(s(t-1))}(\exp(i\gamma s(t-1)))$ and $\varphi_e(\gamma) = \mathbb{E}_{p(e)}(\exp(i\gamma e))$, respectively. Performing the Fourier transform (with a reversal in the sign of the exponent) on both sides of Equation , gives the identical critical condition of the Jacobian upper bound in Equation as

$$\varphi_e\left(\gamma\right) = \varphi_{s(t-1)}\left(\gamma\right)\varphi_e\left(\gamma\right) \tag{13}$$

However, this critical condition does not necessarily meet the critical point of the second inequality in Equation (10). Hence, we presume the Jacobian of the system function f is loosely upper bounded by $|J| < e^{H(s(t))}$. If $H(s(t))$ exists, namely, $H(s(t)) < +\infty$, there exists a finite constant $K > 0$ that satisfies

$$\left|J\right| \le K < e^{H\left(s(t)\right)} \tag{14}$$

which implies that f is Lipschitz continuous. *Q.E.D.*

Note: Since the two terms on the right side of the last equality in Equation (12) are both related to $\Delta(s)$ (especially the complicated relationship between $\Delta(s)$ and the D_{KL}-term), the direct relationship between $\Delta(s)$ and $|J|$ is difficult to

obtain in the general dynamic system (4). This is unlike the original SFA for a linear dynamic system, where the direct relationship is given by Equation (2). Also, the critical conditions or (13) does not necessarily guarantee the maximum of log |J|, since $I(s(t); s(t-1))$ might also be minimised when the D_{KL}-term reaches its minimum.

Lemma 1, as well as Theorems 1 and 2, introduce several related quantities. They are presented in Figure 2 and summarized as

$$(1)\ H\left(s\left(t\right)\right) = I\left(s\left(t\right); s\left(t-1\right)\right) + H\left(e\right)$$

$$(2)\ I\left(s\left(t\right); s\left(t-1\right)\right) = \log\left|J\right| + \mathbb{E}_{p\left(s\left(t-1\right)\right)}\left(D_{KL}\text{-term}\right)$$

$$(3)\ \log\left|J\right| \le I\left(s\left(t\right); s\left(t-1\right)\right)$$

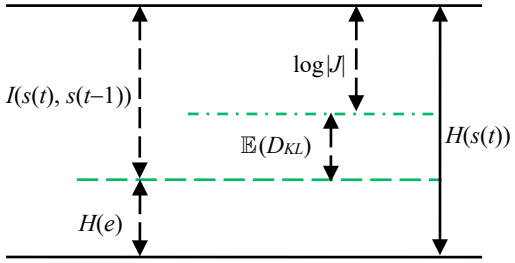

Figure 2: Relationships between the variables in EVOLVE·INFOMAX

The black lines show $H(s(t))$, which do not change. The green dashed and dash-dot lines indicate respectively relationships (1) and (2), whose values can change during model optimisation. Please note that the green dashed line will not cross the two black lines since MI and $H(e)$ are nonnegative, while the green dash-dot line can cross the upper black line but not the green dashed line due to relationship (3).

3.3 Equivalence Analysis between EVOLVE·INFOMAX and the Original Slowness

In the case of a linear Gaussian process $\{s(t) = \lambda s(t-1) + e\}$, the original SFA constrains the variance of $s(t)$ to one, that is,

$$\mathbb{E}\left(s\left(t\right)^2\right) = \mathbb{E}\left(\left(\lambda s\left(t-1\right) + e\right)^2\right) = \lambda^2 + \sigma^2 = 1 \qquad (15)$$

where σ is the standard deviation of e. From Equation (2), the objective of SFA is

$$\min \Delta\left(s\right) = 2 - 2\lambda \qquad (16)$$

Based on Theorem 1, the entropy of $s(t)$ should be constrained to a constant $\mathcal{C}$ in the generalized SFA. In the linear Gaussian case, this can be expressed as

$$H\left(s\left(t\right)\right) = I\left(s\left(t\right), s\left(t-1\right)\right) + H\left(e\right) = \frac{1}{2}\log\frac{2\pi e\sigma^2}{1-\lambda^2} = \mathcal{C} \qquad (17)$$

After normalization, Equation (17) is the same as (15). As well, in this situation, the objective of the generalised SFA simplifies to

$$\min \Delta\left(s\right) = e^{-2I\left(s(t); s(t-1)\right)} = 1 - \lambda^2 \qquad (18)$$

Hence, for linear Gaussian systems, from Equations (18) and (16), the generalized SFA is equivalent to the original version if $\lambda > 0$. Nevertheless, the former also pays attention to the oscillation signals when $\lambda < 0$. Similar characteristics applies to LTSFA (Gao & Shardt, 2021), another extension of SFA.

4. GENERALISED SFA

In the previous section, the EVOLVE·INFOMAX criterion of SFA for a general dynamic system is defined mathematically, followed by a rigorous justification of its properties. In this section, the objective function of the generalised SFA under EVOLVE·INFOMAX is proposed. The optimization algorithm is designed based on NN to address the intractable distributions and integrations.

4.1 Optimization Objective of the Generalised SFA

Based on EVOLVE·INFOMAX, an implementation instance of the generalized SFA can be formulated as

$$\begin{aligned} &\max_{\boldsymbol{g}\in\mathcal{G}} I\left(\boldsymbol{s}\left(t\right); \boldsymbol{s}\left(t-1\right)\right) \\ \text{s.t.}\quad &1).\, H\left(s_j\right) = \mathcal{C}, j = 1, \cdots, K \\ &2).\, D\left(\boldsymbol{s}\right) = D_{KL}\left(p\left(\boldsymbol{s}\middle|\boldsymbol{x}\right)\middle\|\prod_j^K p\left(s_j\middle|\boldsymbol{x}\right)\right) \le \delta \end{aligned} \qquad (19)$$

where $\boldsymbol{s}(t) = \boldsymbol{g}(\boldsymbol{x}(t))$ is the SF vector, $\boldsymbol{g} = [g_1, \cdots, g_K] \in \mathcal{G}$ is a vector-valued function that needs to be found, $D(\boldsymbol{s})$ is the complete dependency between the elements of $\boldsymbol{s}$ given the data $\boldsymbol{x}$, and $\mathcal{C}$ and δ are both constants. The function g_j can be a stochastic map $p_j(s_j|\boldsymbol{x})$ between $\boldsymbol{x}$ and s_j. The first constraint comes from Theorem 1 and aligns each SF based on the level of uncertainty. In some circumstances, the uncertainty can be equivalently measured by entropy and variance, *e.g.*, a unimodal Gaussian model. Thus, this constraint is equivalent to the second constraint of the original SFA in Equation (1) for the Gaussian assumption. The second constraint is used to force SFs to be mutually independent. The complete dependency $D(\boldsymbol{s})$ is defined based on the property of conditional independence of the elements in $\boldsymbol{s}$ given $\boldsymbol{x}$. Thus, the conditional distribution $p(\boldsymbol{s}|\boldsymbol{x})$ can be factorised as

$$p\left(\boldsymbol{s}\middle|\boldsymbol{x}\right) = \prod_j^K p\left(s_j\middle|\boldsymbol{x}\right) \qquad (20)$$

if s_j is conditionally independent of each other. Then, the distance between the conditional distribution with mutually independent elements and those with arbitrary elements can be defined as the *complete dependency* $D(\boldsymbol{s})$ of the elements of $\boldsymbol{s}$ given data $\boldsymbol{x}$, which gives the second constraints of the generalised SFA in Equation (19).

It can be seen that $D(\boldsymbol{s})$ is actually a high-dimensional generalisation of the conditional MI $I(s_1; s_2|\boldsymbol{x})$ when $K = 2$. $D(\boldsymbol{s})$ is constrained to be no larger than a small constant δ to enhance the independence. If δ is set to zero, then constraint 2) will exactly result in the conditional independence condition (20), which is equivalent to the constraint 3) of the original SFA in the case of a Gaussian distribution. Since the generalised SFA in this paper is based on underlying distributions rather than moments, the first constraint of the original SFA (1), which centres the first moment and will not change the uncertainty, is no longer needed.

4.2 Algorithm Based on Neural Networks

The optimization objective of the generalised SFA (19) is very complex, which contains distributions and integrations that are in general intractable. In the present paper, NNs are used to tackle the intractable terms in the problem. First, the input-output function is parameterised by a NN considering its approximation capacity, denoted as $\boldsymbol{g}_\theta(\cdot)$. To simplify the problem, the first constraint is relaxed to unit variance, which can be achieved by constructing a normalisation layer after $\boldsymbol{g}_\theta(\cdot)$ to normalize the output. Hence, the problem (19) can be re-arranged to

$$\max_{\boldsymbol{g}_\theta} I\left(\boldsymbol{s}(t);\boldsymbol{s}(t-1)\right)-\beta D(\boldsymbol{s}) \tag{21}$$

where β is the trade-off weight. For the first term in Equation (21), the lower bound of MI is calculated based on contrastive predictive coding (NCE) by another NN $f_\alpha(\cdot)$ (Van den Oord, et al., 2018) (Poole, et al., 2019), that is,

$$\begin{aligned} &I\left(\boldsymbol{s}(t);\boldsymbol{s}(t-1)\right)\geq I_{f_\alpha}\left(\boldsymbol{s}(t);\boldsymbol{s}(t-1)\right)\\ &\equiv \mathbb{E}\left(\frac{1}{N-1}\sum_{t=2}^{N}\log\frac{\exp\left(f_\alpha\left(\boldsymbol{s}(t),\boldsymbol{s}(t-1)\right)\right)}{\frac{1}{N-1}\sum_{j=1}^{N}\exp\left(f_\alpha\left(\boldsymbol{s}(t),\boldsymbol{s}(j)\right)\right)}\right) \end{aligned} \tag{22}$$

where the expectation is over N independent samples from the joint distribution $p(s(t), s(t-1))$. Thus, MI can be estimated by iteratively maximising the lower bound.

For the complete dependency $D(\boldsymbol{s})$ term in (21), an independence testing trick called permutation is used to obtain the samples from the product of marginals $\prod_j^K p\left(s_j|\boldsymbol{x}\right)$ (Arcones & Gine, 1992). Specifically, randomly permuting each latent dimension across the set of $\boldsymbol{s}$ to obtain a new set of $\hat{\boldsymbol{s}}$. The product of the marginals can be closely approximated by $\hat{\boldsymbol{s}}$ if the set is sufficiently large. For the detailed permutation algorithm, one can refer to Algorithm 1 in (Kim & Mnih, 2018). After obtaining the samples of the two distributions within the KL divergence term, $p\left(\boldsymbol{s}|\boldsymbol{x}\right)$ and $\prod_j^K p\left(s_j|\boldsymbol{x}\right)$, the density-ratio trick (Nguyen, et al., 2010) is used to estimate $D(\boldsymbol{s})$, which actually approximates the density ratio of the two distributions by a discriminator d_ψ. The discriminator d_ψ, a binary classifier (in this paper, a NN), seeks to distinguish the categories that the samples come from, whose output $d_\psi(\boldsymbol{s})$ is the probability of $\boldsymbol{s}$ coming from the two categories. Hence, given two sets of $\boldsymbol{s}$ and $\hat{\boldsymbol{s}}$, d_ψ can be obtained by minimising the cross-entropy of the outputs, that is,

$$\mathcal{L}(\psi)=-\left(\mathbb{E}_{p_1(s)}\log d_\psi\left(\boldsymbol{s}\right)+\mathbb{E}_{1-p_1(\hat{s})}\log\left(1-d_\psi\left(\hat{\boldsymbol{s}}\right)\right)\right) \tag{23}$$

Having trained the discriminator d_ψ, the complete dependence $D(\boldsymbol{s})$ of an arbitrary $\boldsymbol{s}$ can be estimated as

$$\begin{aligned} D(\boldsymbol{s}) &= D_{KL}\left(p\left(\boldsymbol{s}|\boldsymbol{x}\right)\middle\|\prod_j^K p\left(s_j|\boldsymbol{x}\right)\right)\\ &= \mathbb{E}_{p(\boldsymbol{s}|\boldsymbol{x})}\log\frac{p\left(\boldsymbol{s}|\boldsymbol{x}\right)}{\prod_j^K p\left(s_j|\boldsymbol{x}\right)}=\mathbb{E}_{p(\boldsymbol{s}|\boldsymbol{x})}\log\frac{d_\psi\left(\boldsymbol{s}\right)}{1-d_\psi\left(\boldsymbol{s}\right)} \end{aligned} \tag{24}$$

where we assume $d_\psi(\boldsymbol{s})$ is the probability of $\boldsymbol{s}$ coming from $p(\boldsymbol{s}|\ \boldsymbol{x})$, and thus, $1 - d_\psi(\boldsymbol{s}|\ \boldsymbol{x})$ is the probability from the product of marginals. Thus, we can now access all the intractable terms contained in Equation (21). Combining Equations (21), (22), and (24), the optimization problem can be translated into

$$\begin{aligned} &\min_\theta \mathcal{L}(\theta)\\ &=-\left(I_{f_\alpha}\left(\boldsymbol{s}_\theta(t);\boldsymbol{s}_\theta(t-1)\right)-\beta\mathbb{E}_{p(\boldsymbol{s}|\boldsymbol{x})}\log\frac{d_\psi\left(\boldsymbol{s}_\theta\right)}{1-d_\psi\left(\boldsymbol{s}_\theta\right)}\right) \end{aligned} \tag{25}$$

The detailed algorithm is presented in Algorithm 1.

Algorithm 1 The algorithm for generalised SFA

Input: Observation $\mathbf{X}=\left\{\boldsymbol{x}^{(i)}\right\}_{i=1}^{N}$, where $\boldsymbol{x}=\begin{bmatrix}x_1 & \cdots & x_m\end{bmatrix}^\top$,

Batch size M, latent dimension K, weight β;

Initialise the parameters θ, α, and ψ of the SF extractor $\boldsymbol{g}_\theta(\cdot)$, MI estimator $f_\alpha(\cdot)$, and discriminator d_ψ.

Repeat:

1. Randomly sample $\left\{\boldsymbol{x}^{(i)}\right\}_{i=1}^{M}$, $\left\{\left(\boldsymbol{x}^{(i)}(t),\boldsymbol{x}^{(i)}(t-1)\right)\right\}_{i=1}^{M}$ from $\mathbf{X}$.

   Obtain $\mathcal{B}=\left\{\boldsymbol{s}^{(i)}\right\}_{i=1}^{M}$, $\mathcal{B}_{pair}=\left\{\boldsymbol{s}^{(i)}(t),\boldsymbol{s}^{(i)}(t-1)\right\}_{i=1}^{M}$ by $\boldsymbol{s}=\boldsymbol{g}_\theta(\boldsymbol{x})$.

2. Iteratively update d_ψ using Equation (23):

   permute $\mathcal{B}$ to get $\hat{\mathcal{B}}=\left\{\hat{\boldsymbol{s}}^{(i)}\right\}_{i=1}^{M}=\left\{\boldsymbol{s}^{\pi(i)}\right\}_{i=1}^{M}$;

   $\min_\psi \mathcal{L}(\psi)=-\frac{1}{2M}\left(\sum_{s\in\mathcal{B}}\log d_\psi\left(\boldsymbol{s}^{(i)}\right)+\sum_{\hat{s}\in\hat{\mathcal{B}}}\log\left(1-d_\psi\left(\hat{\boldsymbol{s}}^{(i)}\right)\right)\right)$.

3. Iteratively update $f_\alpha(\cdot)$ by $\max_\alpha I_{f_\alpha}\left(\boldsymbol{s}^{(i)}(t);\boldsymbol{s}^{(i)}(t-1)\right)_{\mathcal{B}_{pair}}$.

4. Iteratively update $\boldsymbol{g}_\theta(\cdot)$ based on Equation (25):

   $\min_\theta \mathcal{L}(\theta)=-\left(I_{f_\alpha}\left(\boldsymbol{s}_\theta^{(i)}(t);\boldsymbol{s}_\theta^{(i)}(t-1)\right)_{\mathcal{B}_{pair}}-\beta\frac{1}{M}\sum_{\mathcal{B}}\log\frac{d_\psi\left(\boldsymbol{s}_\theta^{(i)}\right)}{1-d_\psi\left(\boldsymbol{s}_\theta^{(i)}\right)}\right)$.

Until convergence.

5. CASE STUDY

In this section, two cases are conducted to verify the validity of the proposed method. Three ordinary, multilayer perception (MLP) NNs are used as the SF extractor $\boldsymbol{g}_\theta(\cdot)$, MI estimator $f_\alpha(\cdot)$, and discriminator d_ψ shown in Algorithm 1. No other advanced deep NN structure is used considering the potential instability caused by the interactions between the

three nets during the training process. The trade-off weight β is 0.1. The first case is the exponential transformation of the trigonometric polynomials

$$\boldsymbol{x}(t) = \exp\left(\varepsilon_t + \sum_{k}^{K} \boldsymbol{\alpha}_k \cos(kt)\right) \tag{26}$$

where $\boldsymbol{x}$, ε, $\boldsymbol{\alpha} \in \mathrm{R}^m$, m = 100, coefficients $\boldsymbol{\alpha} \sim N(0, \mathbf{I})$, and the noise $\varepsilon \sim N(0, 0.1)$, the degree of the polynomials K = 300. A total of 1000 data samples are generated. For ease of comparison, the original slowness defined in Equation (1) is used to evaluate the performance of the generalised SFA and the original SFA. Figure 3 shows the generalised SFA has slower and smoother SFs than SFA. The second case uses the open-sourced TEP data, and the results in Figure 4 also show a better ability to extract slower and smoother SFs.

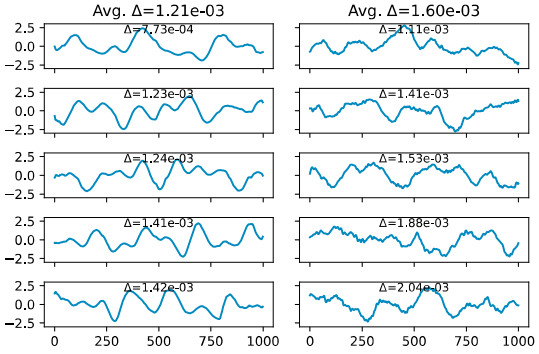

Figure 3: SFs by (left) the generalised SFA and (right) SFA

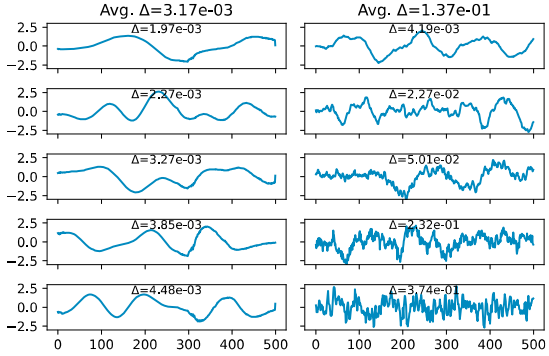

Figure 4: SFs by (left) the generalised SFA and (right) SFA

The two cases show that the NNs in the model are harnessed as we expected, that is, they do pay attention to the slowly time-varying features. The noise is filtered out. Nevertheless, the generalised SFA seems to lack the ability of decomposing the observational signals into regularly varying components with distinct frequencies, namely, different frequency components still mixed in the extracted SFs. This might be due to the insufficient disentanglement capacity provided by constraining the complete dependency. Further, the instability resulting from the interaction of the three NNs also matters.

6. CONCLUSIONS

A new criterion for SFA for general dynamic systems is defined in this paper, which is called EVOLVE·INFOMAX. The theoretical properties are rigorously justified. An implementation objective function is derived and the optimization algorithm based on the training of NNs is designed. The case study shows the feasibility of the NN-based implementation of the generalized SFA. In the future, the technique of stabilising the NN's joint optimisation, as well as more advanced and powerful NN structures will be considered, e.g, recurrent neural networks might work better as an input-output function. In addition, stable and accurate estimation methods of MI should be investigated, especially for the high-dimensional data with relatively large MI values. To improve disentanglement ability, the theoretical properties of complete dependency should be investigated, and thus, propose potential improvement. Furthermore, different implementations of objective functions based on EVOLVE·INFOMAX can be developed.

REFERENCES

Arcones, M. A. & Gine, E., 1992. On the bootstrap of U and V statistics. *The Annals of Statistics,* pp. 655-674.

Berkes, P. & Wiskott, L., 2005. Slow feature analysis yields a rich repertoire of complex cell properties. *Journal of vision,* 5(6), pp. 9-9.

Blaschke, T., Berkes, P. & Wiskott, L., 2006. What is the relation between slow feature analysis and independent component analysis?. *Neural computation,* 18(10), pp. 2495-2508.

Cover, T. M., 1999. *Elements of information theory.* s.l.:John Wiley & Sons.

Gao, X. & Shardt, Y. A., 2021. Dynamic system modelling and process monitoring based on long-term dependency slow feature analysis. *Journal of Process Control,* Volume 105, pp. 27-47.

Kim, H. & Mnih, A., 2018. *Disentangling by factorising.* Stockholmsmässan, PMLR.

Kompella, V. R., Luciw, M. & Schmidhuber, J., 2012. Incremental slow feature analysis: Adaptive low-complexity slow feature updating from high-dimensional input streams. *Neural Computation,* 24(11), pp. 2994-3024.

Nguyen, X., Nguyen, M. J. & Jordan, M. I., 2010. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory,* 56(11), pp. 5847-5861.

Poole, B. et al., 2019, May. *On variational bounds of mutual information.* Long Beach, PMLR.

Shang, C., Huang, B., Yang, F. & Huang, D., 2015. Probabilistic slow feature analysis-based representation learning from massive process data for soft sensor modelling. *AIChE Journal,* 61(12), pp. 4126-4139.

Sprekeler, H., Zito, T. & Wiskott, L., 2014. An extension of slow feature analysis for nonlinear blind source separation. *The Journal of Machine Learning Research,* 15(1), pp. 921-947.

Van den Oord, A., Li, Y. & Vinyals, O., 2018. Representation learning with contrastive predictive coding. *arXiv e-prints,* pp. pp.arXiv-1807.

Wiskott, L. & Sejnowski, T. J., 2002. Slow feature analysis: Unsupervised learning of invariances. *Neural computation,* 14(4), pp. 715-770.