Journal of Machine Learning Research 23 (2022) 1-27 Submitted 7/21; Revised 10/21; Published 2/22

Innovations Autoencoder and its Application in One-class Anomalous Sequence Detection

Xinyi Wang XW555@CORNELL.EDU
Department of Electrical and Computer Engineering
Cornell University
Ithaca, NY 14850, USA

Lang Tong LT35@CORNELL.EDU
Department of Electrical and Computer Engineering
Cornell University
Ithaca, NY 14850, USA

Editor: Sayan Mukherjee

Abstract

An innovations sequence of a time series is a sequence of independent and identically distributed random variables with which the original time series has a causal representation. The innovation at a time is statistically independent of the history of the time series. As such, it represents the new information contained at present but not in the past. Because of its simple probability structure, the innovations sequence is the most efficient signature of the original. Unlike the principle or independent component representations, an innovations sequence preserves not only the complete statistical properties but also the temporal order of the original time series. An long-standing open problem is to find a computationally tractable way to extract an innovations sequence of non-Gaussian processes. This paper presents a deep learning approach, referred to as Innovations Autoencoder (IAE), that extracts innovations sequences using a causal convolutional neural network. An application of IAE to the one-class anomalous sequence detection problem with unknown anomaly and anomaly-free models is also presented.

Keywords: Innovations sequence. Generative adversary networks. Autoencoder. Out-of-distribution detection. Non-parametric anomaly detection.

1. Introduction

At the heart of modern machine learning is the notion of "signature". A data signature should capture essential characteristics of the data and has low dimensionality for easy processing. This work focuses on extracting signatures from a random process (time series) for real-time machine learning applications such as anomaly detection, target tracking, control, and system monitoring.

Assume that, at time t, we have observations $\mathscr{X}_t = \{x_t, x_{t-1}, \cdots\}$ of a stationary random process (x_t), and there are infinitely many data samples to arrive, one at a time, in the future. We are interested in real-time statistical learning problems that make inference or control decisions based on $\mathscr{X}_t$ under the general framework that the underlying probability model of (x_t) is partially or entirely unknown. The real-time (or online) nature of such machine learning problems makes it essential to generate, *causally*, a signature sequence (ν_t) from past

samples such that it captures all statistical information of (x_t) up to time t. Specifically, we are interested in low dimensional but *sufficient* statistics with a simple probability structure.

A classic example of an efficient signature is the *innovations process* introduced by Wiener (1958). Wiener and Kallianpur pioneered the idea of encoding and decoding of a stationary ergodic process (x_t) via an independent and identically distributed (*i.i.d.*) sequence (ν_t) referred to as an *innovations sequence* by Masani (1966). In particular, the *innovations representation* of (x_t) is defined by an encoder G that produces an *i.i.d.* sequence (ν_t) by a *causal transform*:

$$\nu_t = G(x_t, x_{t-1}, \cdots), \tag{1}$$

and a decoder H that recovers (x_t) from (ν_t) also via a causal transform[1]:

$$x_t = H(\nu_t, \nu_{t-1}, \cdots). \tag{2}$$

Without loss of generality, the marginal distribution of ν_t can be assumed to take uniform distribution on $[-1, 1]$.

The existence of such a representation notwithstanding, the interpretation of (ν_t) as an *innovations* sequence of (x_t) is immediate: ν_t, being independent of $\mathscr{X}_{t-1} = (x_{t-1}, x_{t-2}, \cdots)$, represents the new information in x_t but not in $\mathscr{X}_{t-1}$. Equally important is the decoding process, which makes the innovations sequence a sufficient statistics for all decisions made based on $\mathscr{X}_t$. From a modern machine perspective, the encoder-decoder construct of the innovations representation is simply a *causal autoencoder* with innovations sequence as the latent process.

The classical theory on innovations has a long and illustrious history, starting from the work of Kolmogorov (1992), Wiener (1958), and Kalman (1960) on prediction, filtering, and control, and it soon became popular in the engineering community; see Kailath (1970, 1974). The innovations approach is particularly powerful for statistical inference and decision making when the innovations process in (1) can be easily computed in real-time. One prominent case is the stationary Gaussian process for which the innovations process (ν_t) is simply the error sequence of the linear minimum mean-squared error (MMSE) predictor. Another case is the continuous-time (possibly non-Gaussian) process (x_t) under the additive white Gaussian noise model, for which the innovation process is the error sequence of a *nonlinear* MMSE predictor (Kailath, 1971). In general, however, there is no computationally tractable way to extract an innovations sequence when an innovations representation exists. This paper aims to bridge this gap using modern machine learning techniques and demonstrate the potential of innovations representation in an open challenge of one-class anomalous sequence detection.

1.1 Summary of results

We develop a deep learning approach, referred to as Innovations Auto-Encoder (IAE), that provides a practical way to extract innovations of discrete-time stationary random processes with unknown probability structures, assuming the existence of an innovations representation and the availability of historical training samples. To this end, we propose a causal convolutional neural network, *a.k.a* time-delayed neural network (Waibel et al., 1989), and a Wasserstein generative adversary network (GAN) learning algorithm for extracting

1. The stationary random process is defined for $-\infty < t < \infty$.

innovations sequences (Goodfellow et al., 2014; Arjovsky et al., 2017). Because the implementation of an IAE involve only a finite number of current and past samples, a convergence property is needed that ensures a sufficiently high-dimensional implementation will lead to a close approximation of the actual innovations representation. Under ideal training and implementation conditions, we establish a finite-block convergence property in Theorem 1, which ensures that a sufficiently high-dimensional implementation of an IAE produces a close approximation of the ideal innovations representation.

Next, we apply the idea of IAE to a "one-class" anomalous sequence detection problem [2] where neither the anomaly-free nor the anomaly model is known, but anomaly-free training samples are given. By *anomaly sequence detection*, we mean to distinguish the underlying probability distributions of the anomaly-free model and that of the anomaly. Although there are many practical machine learning techniques for outlier and out-of-distribution (OoD) detection, to our best knowledge, the result presented here is the first one-class anomalous sequence detection approach for time series models with unknown underlying probability and dynamic models.

The problem of detecting anomalous sequence brings considerable computational and learning-theoretic challenges. Although one expects that taking a large block of consecutive samples can reasonably approximate statistical properties of a random process, applying existing detection schemes to such high-dimensional vectors with unknown (sequential) dependencies among its components is nontrivial. The main contribution of this work is leveraging of the innovations representation to transform the anomaly-free time series to a sequence of (approximately) uniform *i.i.d.* innovations, thus making the anomalous sequence detection problem the classic problem of testing uniformity, for which we apply versions of coincidence test (David, 1950; Viktorova and Chistyakov, 1964; Paninski, 2008; Goldreich, 2017). We then demonstrate, using field-collected and synthetic datasets, the effectiveness of the proposed approach on detecting system anomalies in a microgrid (Pignati *et al.*, 2015).

1.2 Notations

Notations used are standard. All variables and functions are real. We use $\mathbb{R}^m$ and $\mathbb{N}$ for the m-dimension of real vector space and the set of integers, respectively. Vectors are in boldface, and $\mathbf{x} = (x_1, \cdots, x_m) \in \mathbb{R}^m$ is a *column vector*. A time series is denoted as (x_t) with $t \in \mathbb{N}$. Denote by $\mathbf{x}_t^{(m)} := (x_t, \cdots, x_{t-m+1})$ the column vector of current and $(m-1)$ past samples of (x_t).

Suppose that F is an m-variate scaler function. Let $\mathbf{F}^{(n)} : \mathbb{R}^{m+n-1} \to \mathbb{R}^n$ be the n-fold time-shifted mapping of F defined by $\mathbf{F}^{(n)}(\mathbf{x}_t^{(n+m-1)}) := (F(\mathbf{x}_t^{(m)}), \cdots F(\mathbf{x}_{t-n+1}^{(m)}))$. We drop the superscripts when the dimensionality is immaterial or obvious from the context.

2. Background and Related Work

2.1 Innovations representation and estimation

We refer the readers to Kailath (1970) for an exposition of historical developments of the innovations approach. In general, an innovations representation of a stationary and

2. The terminology of one-class detection comes from the notion of "one-class classification" introduced by Khan and Madden (2014).

ergodic process may not exist. The existence of a causal encoder that maps (x_t) to a uniform *i.i.d* sequence holds quite generally for a large number of popular nonlinear time series models (Wiener, 1958; Rosenblatt, 1959, 2009; Wu, 2005, 2011). The existence of a causal decoder that recovers the original sequence from an innovations sequence requires additional assumptions (Rosenblatt, 1959, 2009; Wu, 2011). Whereas general conditions for the existence of an innovations representation are elusive, a relaxation on the requirement for the decoder to produce a random sequence $(\hat{x}_t)$ with the same conditional (on past observations) distributions as that of (x_t) makes the innovations representation applicable to a significantly larger class of practical applications as shown in Wu (2005, 2011). In this paper, we shall side-step the question of the existence of an innovations representation and focus on learning the innovations representation (H, G) in (1-2) when such a representation does exist.

Although there are no known ways to extract (or estimate) innovations when the underlying probability model is unknown, several existing techniques can be tailored for this purpose. One way is to estimate (ν_t) by the error sequence of a linear or nonlinear MMSE predictor, which can be implemented and trained using a causal convolutional neural network (CNN). Such an approach can be motivated by viewing (x_t) as a sampled process from a *continuous-time process* $\tilde{x}(t) = \tilde{z}(t) + w(t)$ in some interval, where $w(t)$ is a white Gaussian noise and $z(t)$ a possibly non-Gaussian but strictly stationary process. Under mild conditions (Kailath, 1971), the continuous-time innovations process $\tilde{\nu}(t)$ of $\tilde{x}(t)$ turns out to be the MMSE prediction error process. Unfortunately, there is no guarantee that the discrete-time version of the nonlinear MMSE predictor will be an innovation sequence for (x_t).

If we ignore the requirement that the innovation process (ν_t) needs to be a *causally invertible transform* of (x_t), one can view the innovations sequence (ν_t) as independent components of (x_t), for which there is an extensive literature since the seminal work of Jutten and Herault (1991) and Comon (1994) on independent component analysis (ICA). Originally proposed for linear models, ICA is a generalization of the principal component analysis (PCA) by enforcing statistical independence on the latent variables. A line of approaches akin to modern machine learning is to pass x_t through a nonlinear transform to obtain an estimate of an *i.i.d.* sequence (ν_t), where the nonlinear transform can be updated based on some objective function that enforces independence conditions. Examples of objective functions include information-theoretic and higher-order moment based measures (Comon, 1994; Karhunen et al., 1997; Naik and Kumar, 2011).

The ICA approach most related to this paper is ANICA proposed by Brakel and Bengio (2017), where a deep learning approach to nonlinear ICA via an autoencoder is trained by the Wasserstein GAN (Arjovsky et al., 2017) technique. The main difference between IAE and ANICA lies in how causality and statistical independence are enforced in the learning process. Different from IAE, ANICA does not enforce causality in training and its implementation. It achieves statistical independence among extracted components through repeated re-samplings.

Also relevant is NICE (Dinh et al., 2015) where a class of bijective mappings with unity Jacobian is proposed to transform blocks of (x_t) to Gaussian *i.i.d.* components. The property of unity Jacobian makes NICE a particularly attractive architecture capable of evaluating relevant likelihood functions for real-time decisions. However, the special form of bijective

mappings destroys the causality of the data sequence, and the requirement of *i.i.d.* training samples in the maximum-likelihood-based learning is difficult to satisfy for time series models.

It is natural to cast the problem of extracting independent components as one of designing an autoencoder where the latent variables are constrained to be statistically independent. Several variational auto-encoder (VAE) techniques (Kingma and Welling, 2014; Kingma et al., 2017; Tucker et al., 2018; Maaløe et al., 2019; Vahdat and Kautz, 2021) have been proposed to produce generative models for the observation process using independent latent variables. Without requiring that the latent variables are directly encoded from and capable of reproducing the original process, these VAE-based techniques do not guarantee that the latent independent components are part of an innovations sequence.

If we relax the condition of $\{\nu_t\}$ being *i.i.d*, the structure of IAE is similar to a class of Koopman operator based auto-encoders (Takeishi et al., 2017; Morton et al., 2018; Omri et al., 2020). These papers proposed auto-encoder based algorithms to learn a set of finite observable functions that span a Koopman invariant subspace. The novelty of IAE lies in the enforcement of independence and a parametric distribution where the latent variable $\{\nu_t\}$ are generated from.

Unlike the non-parametric approach to obtaining innovations representations considered in this work, there is the literature on parametric techniques to extract innovations by assuming that (x_t) is a causal transform of an innovations sequence (ν_t). By identifying the transform parameters, one can construct a causal inverse of the transform (if it exists). From this perspective, extracting innovations can be solved in two steps: estimating first the parameter estimation of a time series model, and (ii) constructing a causal inverse of the time series model. Under relatively general conditions, parameters of multivariate moving-average and auto-regressive moving average models can be learned by moment methods involving high-order statistics (Cardoso, 1989; Swami and Mendel, 1992; Tong, 1996).

2.2 One-class anomalous sequence detection

The one-class anomalous sequence detection problem[3] is a special instance of semi-supervised anomaly detection that classifies a data sequence as anomalous or anomaly-free when training samples are given only for the anomaly-free model. To our best knowledge, there is no machine learning techniques specifically designed for time series, although there is an extensive literature on the related problem of detecting outlier or OoD samples.

A well-known technique for the one-class anomaly detection is the the one-class support vector machine (OCSVM) (Schölkopf et al., 1999) and its many variants for different applications (Khan and Madden, 2014). OCSVM finds the decision region for the anomaly-free samples by fencing in training samples with a ceratin margin that takes into account unobserved anomaly-free samples. A related idea is to separate the anomaly and anomaly-free model in the latent variable space of an autoencoder. One such technique is f-AnoGAN proposed by Schlegl et al. (2019) where the OoD detection is made by fencing out all samples that result in large decoding error by an autoencoder trained with anomaly-free samples. Other similar techniques include (Bergmann et al., 2019; Gong et al., 2019). These discriminative methods rely, *implicitly*, on an assumption that the support of the anomaly distribution

3. In this paper, we do not consider the broader class of one-class anomaly detection problems where unlabeled training data or scarcely labeled anomalous training data are also used.

is disjoint from that of the anomaly-free, *i.e.,* the anomaly distribution concentrates in the region that the anomaly-free training samples do not or less likely to appear. Therefore, they perform poorly when the domains of anomaly and anomaly-free models overlap completely.

Another set of techniques attach some kind of confidence scores on samples in the feature space, leveraging the neural network's ability to learn the posterior distribution of the anomaly-free model (Hendrycks and Gimpel, 2018; Lakshminarayanan et al., 2017; Liang et al., 2018; Lee et al., 2018b). These techniques construct confidence scores from the learned anomaly-free model without attempting to learn or infer the anomaly model. As shown in (Lan and Dinh, 2021), even with the perfect density estimate, OoD detection may still perform poorly.

The third type of techniques simulate OoD samples in someway, often around the anomaly-free models, as proposed in (Lee et al., 2018a; Hendrycks et al., 2019; Ren et al., 2019). With simulated OoD samples, it is possible, in principle, to capture fully the difference between the anomaly and anomaly-free distributions and derive a likelihood ratio test as proposed by Ren et al. (2019). In practice, however, there could be uncountably many OoDs, and simulating OoD samples are highly nontrivial. A heuristic solution is to perturb training samples from the anomaly-free model and use them to create a proxy of OoD samples.

Existing OoD detection techniques, under the most favorable conditions when training samples are unlimited, learning algorithm most powerful, and the complexity of neural network unbounded, are fundamentally limited in two aspects. First, in general, there does not exist a uniformly most power test (even asymptotically) for all possible anomaly models. This means that, for every detection rule, there are anomaly cases for which the power of detection is suboptimal. Second, they do not provide Chernoff consistency (Shao, 2017) defined as the type I (false positive) and type II (false negative) errors approach to zero as the number of observations increases. For detecting anomalies in times series, Chernoff consistency is essential.

The source of such apparently fundamental limits arises, perhaps, from the lack of a clear characterization of the OoD model; the standard notion of OoD being something other than the anomaly-free distribution is simply not precise enough to provide a theoretical guarantee. Indeed, Lan and Dinh (2021) argue that even perfect density models for the anomaly-free data cannot guarantee reliable anomaly detection, and there are ample examples that demonstrate OoD samples can easily fool standard OoD detectors (Goodfellow et al., 2014). To achieve Chernoff consistency or the asymptotic uniformly most power performance, there needs to be a positive "distance" between the distributions of the anomaly-free model and those of the anomaly. It is the constraint on anomaly models being ϵ-distance away in our formulation makes it possible, under ideal training, implementation, and sampling conditions, to achieve Chernoff consistency. See Sec. 4 for one such approach, building on an earlier result of universal anomaly detection (Mestav and Tong, 2020).

3. Innovations Autoencoder

3.1 Parameterization and Dimensionality of Innovations Autoencoder

A parameterized innovations autoencoder (IAE), denoted by $\mathcal{A}_{(\theta,\eta)} = (G_\theta, H_\eta)$, is defined by an *innovations encoder* G_θ and an *innovations decoder* H_η shown in Figure 1(left), both implemented by causal CNNs (Waibel et al., 1989) with parameters θ and η respectively,

as shown in Figure 1(right). Once trained, the innovations sequence (v_t) is produced by the encoder G_θ, and decoded time series $(\hat{x}_t)$ by H_η. We note here that the causal encoder and decoder can also be implemented using causal recurrent neural networks with suitably defined internal states.

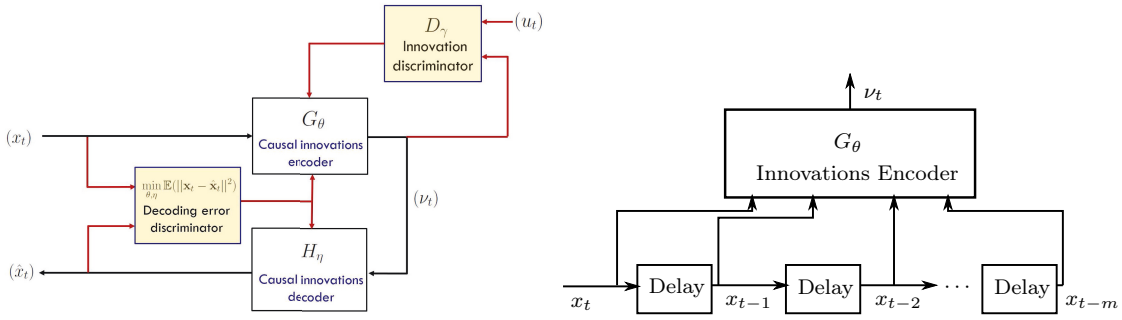

Figure 1: Structure of IAE (left) and a causal CNN implementation (right).

In a practical implementation, at time t, only the current and a finite number of past samples are used to generate the output. We define the *dimension* of an IAE by the input dimension of its encoder. For an m-dimension IAE $\mathcal{A}_{(\theta_m,\eta_m)}$, or $\mathcal{A}_m$ in abbreviation, the input of the encoder[4] G_{θ_m} is $\mathbf{x}_t^{(m)} := (x_t, \cdots, x_{t-m+1})$ and the output is a scaler output ν_t causally produced by G_{θ_m}. The decoder H_{η_m} also takes a finite dimensional input vector $\boldsymbol{\nu}_t^{(n_d(m))} = (\nu_t, \cdots, \nu_{t-n_d(m)+1})$ and causally produces a scaler output $\hat{x}_t$ as an estimate of x_t. Herein, we assume that $n_d(m) = k_\nu m$ for some design parameter[5] k_ν.

The structure of IAE is similar to some of the existing VAEs in modeling (causally) sequential data (Bayer and Osendorfer, 2014; Chung et al., 2015; Goyal et al., 2017). The main difference between IAE and these VAEs is the way IAE is trained and the objective of training. Unlike IAE, these VAEs aims at obtaining a generative model that does not enforce matching encoder input realizations with those of the decoder output; their objective is to produce the underlying stochastic representations in the forms of probability distributions.

3.2 Training of IAE: Dimensionality and the Training Objective

The shaded boxes in Figure 1(left) represent algorithmic functionalities used in the training process, and the red lines represent input variables from the data flow and output variables used in adapting neural network coefficients. Two discriminators are used for acquiring the encoder-decoder neural networks. The *innovations discriminator* is trained via a Wasserstein GAN that evaluates the Wasserstein distance between the estimated innovations (ν_t) and the standard (uniform *i.i.d.*) sequence. The *decoding error discriminator* evaluates the Euclidian distance between input (x_t) and the decoder output $(\hat{x}_t)$. The two discriminators generate stochastic gradients in updating the encoder and decoder neural networks.

In learning an m–dimensional IAE $\mathcal{A}_m$, the two discriminators can only take finite-dimensional samples as their inputs. In practice, the two discriminators may have different

4. Note that the index m of θ_m is associated with the dimension of the autoencoder. The dimension of θ_m can be arbitrarily large.
5. The choice of k_ν doesn't affect the convergence proof.

dimensions. For presentation convenience, we assume both discriminators have the same sample dimension ($k_\nu = 1$). We define the *dimension of training* as the dimension of the training vectors used by the discriminators to derive updates of the encoder and decoder coefficients.

For an n-dimensional training of an m-dimensional IAE, the innovations discriminator $\mathcal{D}_{\gamma_m}$ compares a set of n-dimensional encoder output samples $\{\boldsymbol{\nu}_{m,t}^{(n)} := \mathbf{G}_{\theta_m}^{(n)}(\mathbf{x}_t^{(n+m-1)})\}$ with a set of uniformly distributed *i.i.d.* samples $\{\mathbf{u}_t^{(n)}\} \in [-1,1]^n$ and produces the empirical gradients of the Wasserstein distance $W(\boldsymbol{\nu}_{m,t}^{(n)}, \mathbf{u}_t^{(n)})$. The n-dimensional decoding error discriminator takes decoder outputs $\hat{\mathbf{x}}_{m,t}^{(n)} := \mathbf{H}_{\eta_m}^{(n)}(\boldsymbol{\nu}_t^{(n+k_\nu m-1)})$ and computes the decoding error $||\hat{\mathbf{x}}_{m,t}^{(n)} - \mathbf{x}_t^{(n)}||_2$. The two discriminators compute stochastic gradients and update encoder, decoder, and discriminator parameters $(\theta_m, \eta_m, \gamma_m)$ jointly.

The learning objective of IAE is minimizing a weighted sum of the Wasserstein distance between the probability distributions of $\boldsymbol{\nu}_{m,t}^{(n)}$ and $\mathbf{u}_t^{(n)}$ and the decoding error of the autoencoder. By the Kantorovich-Rubinstein Duality, the training algorithm can be derived from the min-max optimization:

$$\min_{\theta,\eta}\max_{\gamma}\left(L_m^{(n)}(\theta,\eta,\gamma) := \mathbb{E}[D_\gamma(\boldsymbol{\nu}_{m,t}^{(n)}, \mathbf{u}_t^{(n)})] + \lambda\mathbb{E}[||\hat{\mathbf{x}}_{m,t}^{(n)} - \mathbf{x}_t^{(n)}||_2]\right), \tag{3}$$

where the first term measures how close the innovation estimated $\boldsymbol{\nu}_{m,t}^{(n)} = \mathbf{G}_{\theta_m}^{(n)}(\mathbf{x}_t^{(n+m-1)})$ is to a standard reference vector with a uniformly distributed *i.i.d* random vector $\mathbf{u}_t^{(n)}$. The second term measures how well $\boldsymbol{\nu}_{m,t}^{(n)}$ serves as an innovations sequence in reproducing $\mathbf{x}_t^{(n)}$. A pseudo code that implements the IAE learning is shown in the Appendix.

3.3 Convergence Analysis

We will not deal with the convergence of the learning algorithms, which is more or less standard; we shall assume that the learning algorithm converges to its global optimum. Here we address a "structural" convergence issue of some theoretical significance.

A practical implementation of IAE can only be of a finite dimension m. So is the dimension n of the training process. Such a finite dimensional training can only enforce properties of a finite set of variables of the random process. Let $\mathcal{A}_m^{(n)} = (G_{\theta_m^{(n)}}, H_{\theta_m^{(n)}})$ be the encoder and decoder output sequences of the m-dimensional autoencoder optimally trained with an n-dimensional training process according to (3). Let $\mathcal{A} = (G, H)$ be the ideal IAE with encoder G and decoder H. Let (ν_t) and (x_t) be the output sequences of G and H, respectively. We are interested in how $\mathcal{A}_m^{(n)}$ converges to $\mathcal{A}$ in some fashion.

Ideally, we would like to have $\nu_{m,t} \xrightarrow{d} \nu_t$ and $x_{m,t} \xrightarrow{L_2} x_t$, which, unfortunately, is not achievable with finite dimensional training. Our goal, therefore, is to achieve a *finite-block convergence* defined as follows:

Definition 1 (Finite training-block convergence) *An m-dimensional IAE $\mathcal{A}_m^{(n)}$ trained with n-dimensional training samples converges in training block size n to $\mathcal{A} = (G, H)$ if, for all t,*

$$\boldsymbol{\nu}_{m,t}^{(n)} \xrightarrow{d} \boldsymbol{\nu}_t^{(n)}, \quad \mathbf{x}_{m,t}^{(n)} \xrightarrow{L_2} \mathbf{x}_t^{(n)}, \quad \text{as } m \to \infty. \tag{4}$$

Note that, even though the dimension n of the learning process can be arbitrarily large, the finite-block convergence does not guarantee the innovations vector of block size greater than n consists of uniform *i.i.d.* entries, nor does it ensure that a block of the decoder output of size greater than n can approximate the block of encoder inputs probabilistically unless the process (x_t) has a short memory. In practice, unless there is an adaptive procedure to train IAE with increasingly higher dimensions, one may have to be content with a weaker measure of convergence as defined above.

Note also that, by the definition of innovations sequence, it suffices to require that the encoded vector $\boldsymbol{\nu}_{m,t}$ converges in distribution to a vector of uniform *i.i.d.* random variables. A stronger mode of convergence of the decoded sequence is necessary, however. Herein, we restrict ourselves to the L_2 (mean-square) convergence.

We make the following assumptions on $\mathcal{A}_m^{(n)}$ and $\mathcal{A}$:

A1 **Existence:** The random process (x_t) has an innovations representation defined in (1-2), and there exists a causal encoder-decoder pair (G, H) satisfying (1-2) with H being uniform continuous.

A2 **Feasibility:** There exists a sequence of finite-dimensional IAE encoding-decoding functions $(G_{\tilde{\theta}_m}, H_{\tilde{\eta}_m})$ that converges uniformly to (G, H) as $m \to \infty$.

A3 **Training:** The training sample sizes are infinite. The training algorithm for all finite dimensional IAE using finite dimensional training samples converges almost surely to the global optimal.

With these assumptions, we have the following structural convergence. See Appendix A for a proof.

Theorem 1 *Let $\mathcal{A}_m^{(n)} = (G_{\theta_m^{(n)}}, H_{\theta_m^{(n)}})$ be the m-dimensional autoencoder optimally trained with training sample dimension n according to (3). Under (A1-A3), $\mathcal{A}_m^{(n)}$ converges (in finite block size n) to $\mathcal{A}$.*

We now consider the special case of an autoregressive process of finite order K to gain insights into assumptions A1-A3 and Theorem 1. It is sufficient to demonstrate the case for the AR(1) process defined by

$$x_t = \alpha x_{t-1} + \nu_t, \quad \alpha \in (0, 1),$$

where $\nu_t \sim \mathcal{U}(-1, 1)$ is a uniformly distributed on $[-1, 1]$ *i.i.d.* sequence. A natural IAE $\mathcal{A} = (G_\theta, H_\eta)$ is given by

$$G_\theta : \ \nu_t = G_\theta(\mathbf{x}_t^{(\infty)}) = \boldsymbol{\theta}^{\mathrm{T}} \mathbf{x}_t^{(\infty)}, \quad \boldsymbol{\theta} = (1, -\alpha, 0, 0, \cdots), \tag{5}$$

$$H_\eta : \ x_t = H_\eta(\boldsymbol{\nu}_t^{(\infty)}) = \boldsymbol{\eta}^{\mathrm{T}} \boldsymbol{\nu}_t^{(\infty)}, \quad \boldsymbol{\eta} = (1, \alpha, \alpha^2, \cdots). \tag{6}$$

It is readily verified that both H and G are uniform continuous. It is also obvious that the assumption A1 is satisfied.

Now consider the m-dimensional IAE $\tilde{A}_m = (G_{\tilde{\theta}_m}, H_{\tilde{\eta}_m})$ defined by

$$G_{\tilde{\theta}_m}: \ \tilde{\nu}_{m,t} = G_{\tilde{\theta}_m}(\mathbf{x}_t^{(m)}) = \tilde{\boldsymbol{\theta}}_m^{\mathrm{T}} \mathbf{x}_t^{(m)}, \quad \tilde{\boldsymbol{\theta}}_m = (1, -\alpha, 0, 0, \cdots). \tag{7}$$

$$H_{\tilde{\eta}_m}: \ \tilde{x}_{m,t} = H_\eta(\tilde{\boldsymbol{\nu}}_{m,t}^{(\kappa_\nu m)}) = \tilde{\boldsymbol{\eta}}_m^{\mathrm{T}} \boldsymbol{\nu}_t^{(\kappa_\nu m)}, \quad \tilde{\boldsymbol{\eta}}_m = (1, \alpha, \alpha^2, \cdots, \alpha^{\kappa_\nu m - 1}). \tag{8}$$

It is immediate that $G_{\tilde{\theta}_m} \to G_\theta$ and $H_{\tilde{\eta}_m} \to H_\eta$ uniformly as $m \to \infty$. Therefore the assumption A2 is met.

From (3), we have, for all N, $m > 2$ and γ,

$$L_m^{(n)}(\tilde{\theta}_m, \tilde{\eta}_m, \gamma) := \mathbb{E}[D_\gamma(\tilde{\boldsymbol{\nu}}_{m,t}^{(n)}, \mathbf{u}_t^{(n)})] + \lambda \mathbb{E}[||\tilde{\mathbf{x}}_{m,t}^{(n)} - \mathbf{x}_t^{(n)}||_2] = \lambda \mathbb{E}[||\tilde{\mathbf{x}}_{m,t}^{(n)} - \mathbf{x}_t^{(n)}||_2].$$

Since $H_{\tilde{\eta}_m}$ is the best l_2 approximation of H, $\mathbb{E}(||\tilde{\mathbf{x}}_{m,t}^{(n)} - \mathbf{x}_t^{(n)}||_2) = \min_{\theta,\eta} \mathbb{E}(||\hat{\mathbf{x}}_{m,t}^{(n)} - \mathbf{x}_t^{(n)}||_2)$. Therefore, $\tilde{A}_m = (G_{\tilde{\theta}_m}, H_{\tilde{\eta}_m})$ is a global optimum of (3). Therefore, Theorem 1 is verified. Further, with $\tilde{A}_m$, we have strong convergence of $\tilde{\nu}_{m,t} = \nu_t$ for all $m \geq 2$ and $(\tilde{x}_{m,t}) \xrightarrow{L_2} (x_t)$ as $m \to \infty$.

4. Anomalous Sequence Detection via Innovations Autoencoder

We develop an IAE-based approach to nonparametric anomalous sequence detection (or simply anomaly detection for brevity) and demonstrate in Sec. 5.2 the proposed approach in a smart power grid application using data collected in a microgrid.

4.1 A Nonparametric Anomaly Model

We consider the problem of real-time detection of anomalies in a time series (x_t) modeled as a random process with unknown temporal dependencies and probability distributions. In particular, we assume that the anomaly-free time series is stationary whereas the anomalous time series can be arbitrary, even non-stationary.

Let $\mathbf{x}_t$ be a vector consisting of a finite block of current and past sensor measurements. Under the null hypothesis $\mathcal{H}_0$ that models the anomaly-free measurements and the alternative $\mathcal{H}_1$ for the anomalies, we consider the following hypothesis testing problem

$$\mathcal{H}_0 : \mathbf{x}_t \sim f_0 \quad vs. \quad \mathcal{H}_1 : \mathbf{x}_t \sim f_1 \in \mathscr{F}_1 = \{f : ||f_1 - f_0|| \geq \epsilon\} \tag{9}$$

with unknown probability distribution f_0 under the anomaly-free hypothesis and a collection $\mathscr{F}_1$ of unknown anomaly distributions under $\mathcal{H}_1$. Parameter $\epsilon > 0$ represents the degree of separation between the anomaly and anomaly-free models where $\|\cdot\|$ is the l_1 distance although other measures such as the Shannon-Jensen distance and the Kullback-Leibler divergence are equally applicable. We assume that an anomaly-free dataset $\mathscr{X}_0$ is available for offline or online training.

Note that $\mathcal{H}_0$ is a simple hypothesis with a single distribution f_0 whereas $\mathcal{H}_1$ is a composite hypothesis that captures all possible anomalies in $\mathscr{F}_1$. Prescribing a positive distance $\epsilon > 0$ between the anomaly-free and the collections of anomaly models is crucial to establish Chernoff consistency that drives the false positive and false negative rates to zero as the dimension of $\mathbf{x}_t$ increases.

4.2 Anomaly detection via IAE and Uniformity Test

A defining feature of IAE is that, under $\mathcal{H}_0$, the innovations encoder transforms the measurement time series with an unknown probability model to the standard uniform *i.i.d.* sequence. Through an IAE trained with anomaly-free data, the anomaly detection problem is transformed to testing whether the transformed sequence is *i.i.d.* uniform, for which Chernoff-consistent detectors can be constructed. Implicitly assumed in this approach is that an anomaly process ϵ-distance away from that of anomaly-free will not be mapped to a uniform *i.i.d.* process which is reasonable in practice.

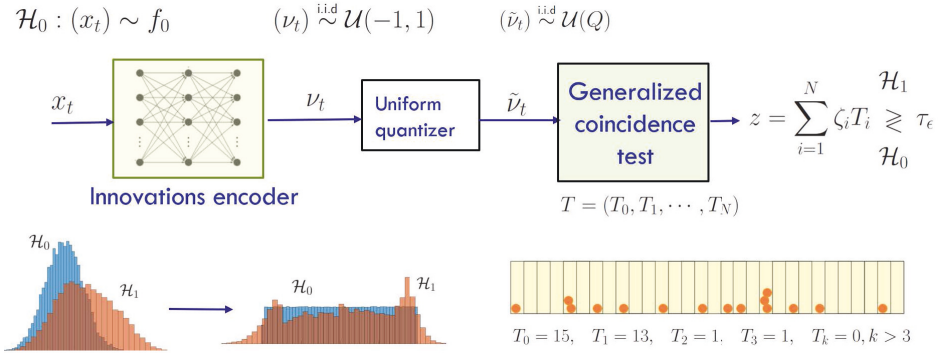

Figure 2: IAE Uniformity Test. Top: Implementation schematic and test statistics under $\mathcal{H}_0$. Bottom left: Histogram of (x_t) and (ν_t) under $\mathcal{H}_0$ and $\mathcal{H}_1$. Bottom right: an example of coincidence statistics (T_i) with 30 quantization levels and 15 samples.

Fig. 2 shows a schematic of the proposed IAE anomaly detection: the sensor measurements (x_t) is passed through an innovations encoder that, under the anomaly-free hypothesis $\mathcal{H}_0$, generates a uniform $\mathcal{U}(-1,1)$ *i.i.d.* innovations sequence (ν_t). The innovation sequence is then passed through a uniform quantizer that puts ν_t in one of the Q equal-length quantization bins to produce a sequence of discrete random variables $\tilde{\nu}_t \in \{1, \cdots, Q\}$:

$$\tilde{\nu}_t = \begin{cases} 1, & \nu_t \le -1 + 2/Q, \\ i, & -1 + 2(i-1)/Q < \nu \le -1 + 2i/Q, \quad i = 2, \cdots, Q-1, \\ Q, & \nu \ge 1 - 2/Q. \end{cases}$$

Under $\mathcal{H}_0$, we thus have a uniform Q-ary *i.i.d.* sequence $\tilde{\nu}_t$, transforming the original hypothesis testing problem to the following derived hypotheses from (9):

$$\mathcal{H}_0' : \tilde{\nu}_t \sim P_0 = (\frac{1}{Q}, \cdots, \frac{1}{Q}) \ \textit{vs.}\ \mathcal{H}_1' : \tilde{\nu}_t \sim P_1 \in \mathscr{P}_1 = \{(p_1, \cdots, p_Q), ||P_1 - P_0||_1 \ge \epsilon'\}. \quad (10)$$

where P_0 and P_1 are Q-ary probability mass functions. Testing $\mathcal{H}_0'$ against $\mathcal{H}_1'$ is a classic problem (David, 1950; Viktorova and Chistyakov, 1964; Paninski, 2008; Goldreich, 2017).

Consider (10) with N samples $\tilde{\boldsymbol{\nu}}_t^{(N)} = (\tilde{\nu}_t, \cdots, \tilde{\nu}_{t-N+1})$, a sufficient statistic equivalent to the histogram is the *coincidence statistic* $T = (T_0, \cdots, T_N)$ where T_i is the number of quantization bins containing exactly i samples. See Fig. 2 (bottom right) for an example for $Q = 30$ and $N = 15$. The coincidence statistics such as T_0 and T_1 characterize the uniformity property particularly well when samples are "sparse" relative to the quantization level. For

instance, when Q is considerably larger than N, the T_1 value of a uniformly distributed $\tilde{\nu}_t$ tends to be large, and T_0 tends to be small, relatively. It is shown in Paninski (2008) that using T_1 alone achieves Chernoff consistency, and the sample complexity is optimal at the order of $\sqrt{Q}$.

A general form of a coincidence test is a linear test given by

$$\sum_{i=1}^{N} \zeta_i T_i \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \tau_{\epsilon'}, \tag{11}$$

where the threshold parameter, a function of ϵ' (and N), controls the level of false positive rate. Paninski gives τ_ϵ for the sparse sample case when only T_1 is used, whereas Viktorova and Chistyakov (1964) showed the coefficients for the asymptotically most powerful linear detector.

5. Performance Evaluation

We present two sets of evaluations based on a combination of field collected datasets from actual systems and synthetic datasets designed to test specific properties. Sec. 5.1 focuses on the IAE encoder-decoder performance using the runs up and down test (Gibbons and Chakraborti, 2003) for the independency of the encoder output sequence and the reconstruction error of the decoded sequence. Sc. 5.2 focuses on the IAE-based anomaly detection in a smart power grid.

5.1 Training and testing datasets

We used two field-collected datasets of continuous point-on-wave (CPOW) measurements from two actual power systems. The BESS dataset contained direct bus voltage measurements sampled at 50 kHz at a medium-voltage (20kV) substation collected from the EPFL campus smart grid as described by Sossan et al. (2016). As shown in Fig. 3 (left), several circuits were connected at a bus via a medium voltage switchgear. Also connected to the same bus was a battery energy storage system (BESS) used to emulate physically anomaly power injections. The BESS dataset captured anomaly-free measurements and anomaly power injections that varied from 0 to 500 (kW). Fig. 3 (right) shows segments of anomaly and anomaly-free measurements of a single phase CPOW voltage waveforms. The CPOW samples exhibited narrow-band (sinusoidal-like) characteristics with strong temporal correlations. Because of the frequency and voltage regulation mechanisms in a power system, the voltage magnitudes and frequencies were tightly controlled such that both anomaly-free and anomaly voltage CPOW data were very similar although a zoomed-in plots exhibited differences in high-order harmonics, as shown in the zoom-in plot of Fig. 3 (right). The detection of anomaly in such voltage CPOW measurements was quite challenging.

The second field-collected dataset (UTK) contained direct samples of voltage waveform at 6 kHz collected at the University of Tennessee. The dataset contains 180,000 voltage samples collected by a sensor in the real power system by University of Tennessee, Knoxville. Although high-order harmonics were observed in the voltage samples, only anomaly-free CPOW measurements were available. Similar to the BESS dataset, the UTK dataset contained strongly correlated narrow-band samples.

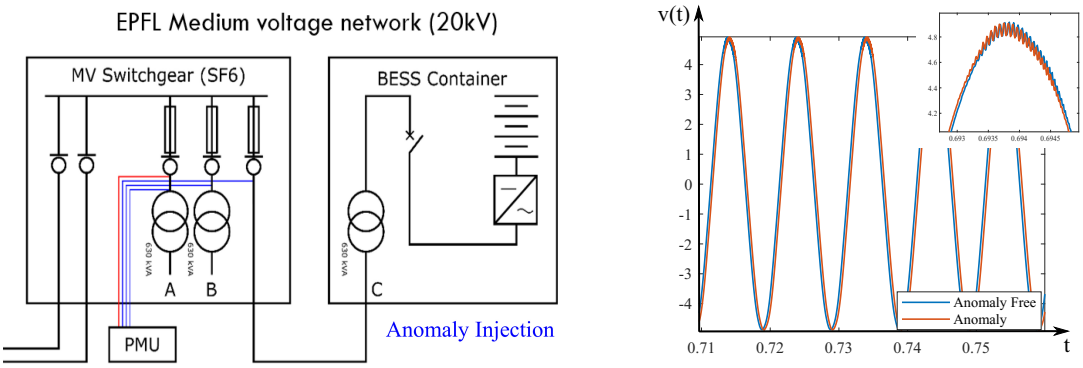

Figure 3: Battery Energy System at EPFL. See Sossan et al. (2016) for detailed description.

Besides the two field datasets (BESS and UTK), we also designed several synthetic datasets to evaluate specific properties of IAE and IAE-based anomaly detections. These datasets are described in Sec. 5.2 and Sec. 5.3.

5.2 IAE performance

We evaluated the performance of IAE and several of benchmarks for extracting innovations sequences. In particular, we examined whether the estimated innovation sequences passed the test of being statistically independent and identically distributed. We also evaluated the mean-squared error (MSE) of the reconstructed signal.

5.2.1 Benchmarks

Since there were very few techniques specifically designed for extracting innovation sequences, we compared IAE with four benchmarks adapted from existing techniques aimed at extracting independent components. Among these benchmarks, three were deep learning solutions (NLLS, ANICA, f-AnoGAN) and two of those (ANICA, f-AnoGAN) were autoencoder based.

- *LLS:* LLS was a linear least-squares prediction error filter that generated the one-step prediction error sequence. For stationary Gaussian time series, a perfectly trained LLS predictor would produce a true innovations sequence.
- *NLLS:* NLLS was a nonlinear least-squares prediction error filter that generated the one-step production error time series. If the measurement time series was obtained from a sampled (possibly) non-Gaussian process with additive Gaussian noise, the NLLS prediction error sequence would be a good approximation of an innovations process.
- *ANICA:* ANICA was an adaption of the nonlinear ICA autoencoder proposed in (Brakel and Bengio, 2017). Aimed to extract independent components from a block of measurements, the original design did not enforce causality and was not intended to generate an innovations sequence.
- *f-AnoGAN:* f-AnoGAN proposed by Schlegl et al. (2019) was an autoencoder technique involving convolutional neural networks. The goal was to extract low-dimensional

latent variables as features from which the decoder could recover the original. Since the autoencoder was trained with anomaly-free samples, the intuition was that anomaly data would have high reconstruction errors. Such a construction could be viewed as a nonlinear principle component analysis (PCA).

IAE was implemented by adapting the Wasserstein GAN with a few modifications [6]. For all cases in this paper, similar neural network structures were used: the encoder and decoder both contained three hidden layers with 100, 50, 25 neurons respectively with hyperbolic tangent activation. The discriminator contained three hidden layers with 100, 50, and 25 neurons, of which the first two used hyperbolic tangent activation and the last one the linear activation. The tunning parameter used for each case is presented in the Appendix.

Dataset	Model
Moving Average (MA):	$x_t = \frac{1}{10}\sum_{i=1}^{10} \nu_{t-i}$
Linear Autoregressive (LAR)	$x_t = 0.5x_{t-1} + \nu_t$
Nonlinear Autoregressive (NLAR)	$x_t = 0.5x_{t-1} + 0.4\mathbb{1}(x_{t-2} < 0.7) + \nu_t$

Table 1: Test Synthetic Datasets. $\nu_t \overset{\text{i.i.d}}{\sim} \mathcal{U}[0,1]$. $\mathbb{1}(\cdot)$ is the indicator function.

5.2.2 Test datasets

Besides the two field datasets (BESS and UTK) described in Sec. 5.1, we included three synthetic datasets shown in Table. 1 to produce different levels of temporal dependencies and probability distributions in test data. In particular, the linear autoregressive (LAR) dataset was chosen such that a properly trained LLS approach would produce an innovations sequence. For the moving average (MA) and nonlinear autoregressive (NLAR) datasets, sufficiently complex neural network implementation of NLLS and ANICA could produce approximations of the innovations sequences.

For all the synthetic cases, we used the memory size m = 20 dimensional IAE and training sample dimension n = 60 in the neural network training, and 100,000 samples were used for training for all cases. The neural network memory size for real data cases were chosen to be m = 100, n = 250, due to stronger temporal dependency.

5.2.3 Performance and discussion

In evaluating the performance of the benchmarks, we adopted the runs up and down test that used the numbers of consecutively ascending or descending samples as test statistics of statistical independence. Its null hypothesis assumes that the sequence of samples consists of *i.i.d* samples, and the alternateive hypothesis assumes the opposite. The test was shown (empirically) to have the best performance in (Gibbons and Chakraborti, 2003). We also evaluated the empirical mean-squared error of the reconstruction.

Table. 2 shows the p-values of the runs up and down test. NLLS prediction method was not implemented for LAR case because the linear least-square was sufficient for demonstration

6. `https://keras.io/examples/generative/wgan_gp/`

Method	MA	LAR	NLAR	UTK	BESS
IAE (p-value)	0.9492	0.8837	0.7498	0.9990	0.8757
LLS (p-value)	0.3222	0.9697	0.0186	<0.0001	<0.0001
NLLS (p-value)	0.2674	N/A	0.5116	<0.0001	<0.0001
ANICA (p-value)	<0.0001	0.1502	<0.00019	<0.0001	<0.0001
F-anoGAN (p-value)	<0.001	0.0106	<0.0001	<0.0001	<0.0001
IAE(MSE)	6.3849	8.5366	9.398425	14.5641	21.0144
Anica(MSE)	137.2839	274.3765	283.31250	315.6521	319.9284
F-anoGAN(MSE)	6.7421	12.4379	11.6458	11.8630	11.8821

Table 2: p-value of the runs test and the mean-squared error (MSE) of the reconstruction.

purposes. As autoencoder based methods, F-anoGAN and IAE achieved the comparable reconstruction error, with F-anoGAN performing better. ANICA failed to obtain a competitive reconstruction error.

As for the independence test for the BESS dataset, IAE achieved the highest p-values for all the scenarios except the synthetic LAR dataset designed specifically for the LLS algorithm. For the synthetic datasets, LLS and NLLS produced sequences that the runs tests could not easily reject the independence hypothesis. For the field datasets, LLS and NLLS failed the run tests. Not specifically designed for extracting innovations, ANICA failed the run tests for statistical independence.

5.3 Detection of anomalies in power systems

We evaluated the performances of several benchmarks in detecting system anomalies in field-collected dataset BESS, the UTK dataset with synthetically generated anomalies, and two synthetic time series datasets. We compared benchmark techniques using their receiver characteristic curves (ROC) that plotted true positive rates (TPR) over a range of the false positive rate (FPR). The area under ROC (AUROC) was also calculated for all techniques.

5.3.1 Test datasets

In addition to the BESS dataset that included both anomaly and anomaly-free measurements, we also considered three additional datasets shown in Table 3, two synthetic datasets (`SYN1`, `SYN2`) and one semi-synthetic dataset with anomaly waveforms added to the field-collected anomaly-free samples (Wang et al., 2021). `SYN1` and `SNY2` had the identical anomaly-free models of AR(1) Gaussian. Under the anomaly hypothesis, `SYN1` was AR(2) Gaussian, whereas `SYN2` was an AR(1) with uniform innovations. Because only the anomaly-free training samples were assumed and the anomaly waveforms and probability distributions were arbitrary, the same anomaly detector trained based on the anomaly-free samples were tested under `SYN1` and `SYN2`.

5.3.2 Benchmarks

Few benchmark techniques were suitable for the anomaly sequence detection problem considered here. Most relevant prior techniques that could be applied directly were the one-class

Test Case	Anomaly Free Samples	Anomaly Samples	Block Size (N)
Syn1	$x_t = 0.5x_{t-1} + \nu_t$	$x_t = 0.3x_{t-1} + 0.3x_{t-2} + \nu_t$	1000
Syn2	$x_t = 0.5x_{t-1} + \nu_t$	$x_t = 0.5x_{t-1} + \nu'_t$	1000
UTK	Real Data	GMM Noise	200
BESS	Real Data	Real Data	500

Table 3: Data Detection Test Cases. $\nu_t \overset{i.i.d}{\sim} \mathcal{N}(0,1)$, $\nu'_t \overset{i.i.d}{\sim} \mathcal{U}[-1.5, 1.5]$

support vector machine (OCSVM) proposed in (Schölkopf et al., 1999) and f-AnoGAN in (Schlegl et al., 2019). OCSVM, a semisupervised classification technique, was implemented with radial basis functions as its kernel and was trained with anomaly-free samples. Although not designed as an anomaly detection solution, ANICA (Brakel and Bengio, 2017) was adapted to be a preprocessing algorithm (similar to IAE) before applying a uniformity test described in Sec. 4.

We have also included the Quenouille test (Priestley, 1981) designed to test the goodness of fit of an AR(k) model (with `SYN1` dataset.) Because of the asymptotic equivalence of the Quenouille test and the maximum likelihood test of Whittle (1951), we used Quenouille test as a way to calibrate how well IAE and other nonparametric tests would perform under AR(k) time series models with dataset `SYN1` for which the Quenouille test is asymptotically optimal.

5.3.3 Performance on the BESS dataset

The BESS dataset was used to test the proposed testing technique's ability to detect system anomalies. As shown in Fig. 4, the anomaly and anomaly-free voltage signals were very similar due to the voltage regulation of the bus voltage in the power system. The detection based solely on the raw voltage signal can be very challenging. Fig. 4 (right) shows the ROC curves obtained using 500-sample blocks. Since the anomaly and anomaly-free samples are hard to distinguish, all the other methods apart from IAE didn't seem to work well in this case.

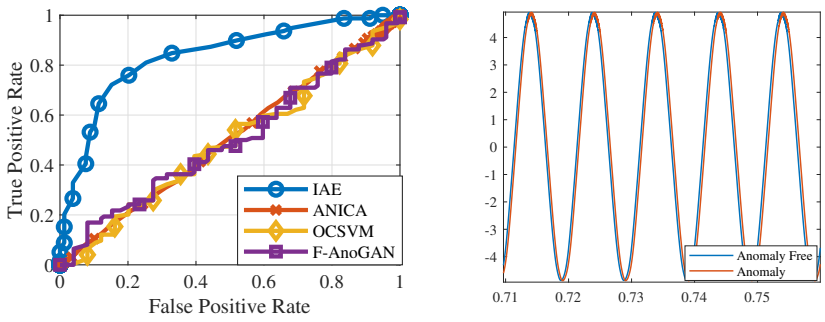

Figure 4: Detection performance for the BESS dataset. Left: ROC curves. (AUROC: IAE:0.8354, ANICA:0.5027 OCSVM:0.4903 F-AnoGAN:0.4993) Right: Anomaly and anomaly-free traces.

5.3.4 Performance on the UTK dataset

We evaluated benchmark performance on the UTK dataset with synthetic anomaly test samples. To construct the anomaly samples, we added a comparably small Gaussian Mixture noise on the anomaly-free measurements. The signal to noise ratio of the Gaussian Mixture noise to the anomaly free signal is roughly 40dB, and the time-domain trajectories of anomaly and anomaly-free signals are shown in Fig. 5 (right), which demonstrate the level of similarity between anomaly and anomaly-free samples. Seen from Fig. 5 (left), IAE was the only detection method that was able to make reliable decisions, with ANICA performing slightly better than the rest.

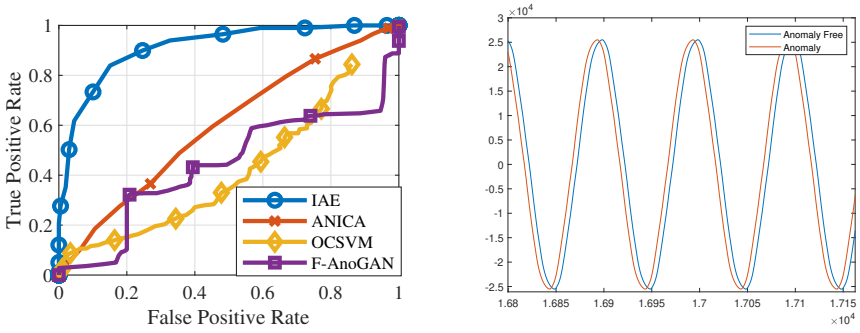

Figure 5: Detection performance for the UTK dataset. Left: ROC curves. (AUROC: IAE:0.9135, ANICA:0.5967, OCSMV:0.2978, F-AnoGAN:0.4393) Right: Anomaly and anomaly-free traces.

5.3.5 Performance on synthetic datasets `SYN 1` and `SYN 2`

We also conducted anomaly detection based on synthetic data generated by auto-regressive models (`SYN1` and `SYN2`). For both `SYN1` and `SYN2`, the anomaly free datasets were the same, and the two anomaly datasets were designed to highlight the the performance of the detectors facing different anomalies.

In `SYN1`, we designed the anomaly to have the same marginal distribution as the anomaly-free data ($x_t \sim \mathcal{N}(0, 4/3)$), as shown in Fig. 6 (right), intentionally making detection based on a single sample ineffective. As shown by Fig. 6 (left), IAE performed similarly well as the asymptotically optimal detector (Quenouille). ANICA performed better (with AUROC above 0.5) than the other two machine learning-based detection methods. Because the marginal distributions of the measurements are the same for both hypotheses, the other two machine learning-based techniques were not competitive under this setting.

`SYN2` adopted two auto-regressive models with the same parameters in temporal dependencies. The marginal distributions of the measurements, however, were slightly different under the anomaly and anomaly-free models. See. Fig 7 (right), which made it very challenging for OCSVM and F-AnoGAN. ANICA was also not effective in extracting independent components, causing failures in the uniformity test. Only IAE was able to capture the difference between the two datasets through the extraction of innovations and made reasonably reliable decisions, as shown in Fig 7 (left).

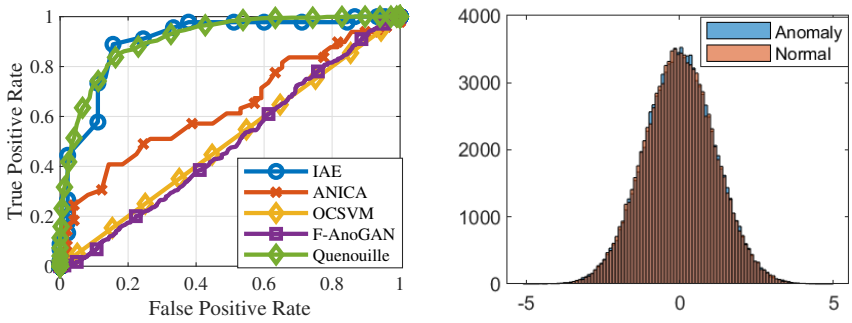

Figure 6: Anomaly Detection for `SYN1`. Left: ROC curves. (AUROC: IAE:0.9021, ANICA:0.6337, OCSVM:0.4455, F-AnoGAN:0.4881, Quenouille:0.9112) Right: Anomaly and anomaly-free histograms.

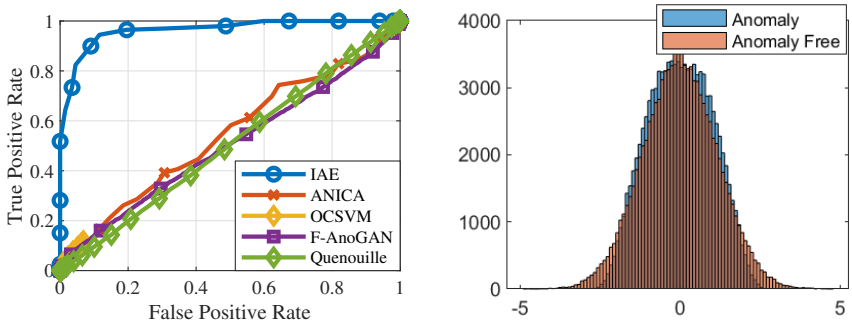

Figure 7: Anomaly Detection for `SYN1`. Left: ROC curves. (AUROC: IAE:0.9635, ANICA:0.6337, OCSVM:0.5407, F-AnoGAN:0.5020, Quenouille:0.5026) Right: Anomaly and anomaly-free histograms.

6. Conclusion

IAE is a machine learning technique that extracts innovations sequence from real-time observations. When properly trained, IAE can serve as a font-end processing unit that transforms processes of unknown temporal dependency and probability structures to a standard uniform *i.i.d* sequence. (Extension to other marginal distributions is trivial.) IAE is, in someway, an attempt to realize Wiener's original vision of encoding stationary random processes with the simplest possible form, although the existence of such an autoencoder is not guaranteed (Rosenblatt, 1959). From an engineering perspective, however, the success of Wiener and Kalman filtering in practice is a powerful testament that many applications in practice can be approximated by innovations representations. It is under such an assumption that IAE serves to remove the modeling assumptions in Wiener and Kalman filtering and pursues a data-driven machine learning solution. As an example, the IAE-based anomaly detection is shown to be quite effective for the one-class anaomalous time series detection problem, for which there are few solutions.

Acknowledgments

This work was supported in part by the National Science Foundation under Grants 1932501 and 1816397. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

M. Arjovsky, S. Chintala, and L.Bottou. Wasserstein GAN, January 2017. arXiv:1701.07875.

J. Bayer and Christian Osendorfer. Learning Stochastic Recurrent Networks. March 2014. arXiv:1411.7610.

P. Bergmann, S. Löwe, M. Fauser, D. Sattlegger, and C. Steger. Improving Unsupervised Defect Segmentation by Applying Structural Similarity to Autoencoders. *Proceedings of the 14th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, 2019. doi: 10.5220/0007364503720380. URL `http://dx.doi.org/10.5220/0007364503720380`.

P. Brakel and Y. Bengio. Learning Independent Features with Adversarial Nets for Non-linear ICA, October 2017. arXiv:1710.05050.

J.-F. Cardoso. Source Separation Using Higher Order Moments. In *Intermmanational Conference on Acoustics, Speech, and Signal Processing,*, pages 2109–2112 vol.4, 1989. doi: 10.1109/ICASSP.1989.266878.

J. Chung, K. Kastner, L. Dinh, K. Goel, A. Courville, and Y. Bengio. A Recurrent Latent Variable Model for Sequential Data. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'15, page 2980–2988, Cambridge, MA, USA, 2015. MIT Press.

P. Comon. Independent Component Analysis, A New Concept? *Signal Processing*, 36(3): 287–314, 1994. ISSN 0165-1684. doi: https://doi.org/10.1016/0165-1684(94)90029-9. URL `https://www.sciencedirect.com/science/article/pii/0165168494900299`. Higher Order Statistics.

F. N. David. Two Combinatorial Test of Whether a Sample has Come from a Given Population. *Biometrika*, 37(1/2):97–110, 1950. ISSN 00063444. URL `http://www.jstor.org/stable/2332152`.

L. Dinh, D. Krueger, and Y. Bengio. NICE: Non-linear Independent Components Estimation, April 2015. URL `https://arxiv.org/abs/1410.8516`. arXiv:arXiv:1410.8516.

J. D. Gibbons and S. Chakraborti. *Nonparametric Statistical Inference. (4th ed., rev. and expanded.)*. M. Dekker, New York, 2003.

O. Goldreich. *Introduction to Property Testing*. Cambridge University Press, 2017. doi: 10.1017/9781108135252.

D. Gong, L. Liu, V. Le, B. Saha, M. R. Mansour, S. Venkatesh, and A. Hengel. Memorizing Normality to Detect Anomaly: Memory-augmented Deep Autoencoder for Unsupervised Anomaly Detection, 2019.

I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative Adversarial Networks, June 2014. arXiv:2012.05163.

A. Goyal, A. Sordoni, M. Côté, N. R. Ke, and Y. Bengio. Z-Forcing: Training Stochastic Recurrent Networks. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/900c563bfd2c48c16701acca83ad858a-Paper.pdf.

D. Hendrycks and K. Gimpel. A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks, 2018. URL https://arxiv.org/abs/1610.02136. arXiv:1610.02136.

D. Hendrycks, M. Mazeika, and T. Dietterich. Deep Anomaly Detection with Outlier Exposure. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=HyxCxhRcY7.

C. Jutten and J. Herault. Blind Separation of Sources, Part I: An Sdaptive Slgorithm Based on Neuromimetic Architecture. *Signal Processing*, 24(1):1–10, 1991. ISSN 0165-1684. doi: https://doi.org/10.1016/0165-1684(91)90079-X. URL https://www.sciencedirect.com/science/article/pii/016516849190079X.

T. Kailath. The Innovations Approach to Detection and Estimation Theory. *Proceedings of the IEEE*, 58(5):680–695, 1970. doi: 10.1109/PROC.1970.7723.

T. Kailath. Some Extensions of the Innovations Theorem*. *Bell System Technical Journal*, 50(4):1487–1494, 1971. doi: https://doi.org/10.1002/j.1538-7305.1971.tb02563.x.

T. Kailath. A View of Three Decades of Linear Filtering Theory. *IEEE Transactions on Information Theory*, 20(2):146–181, 1974. doi: 10.1109/TIT.1974.1055174.

R. E. Kalman. A New Approach to Linear Filtering and Prediction Problems. *Trans. ASME J. of Basic Engineering*, 82(1):35–45, 1960.

J. Karhunen, E. Oja, L. Wang, R. Vigario, and J. Joutsensalo. A Class of Neural Networks for Independent Component Analysis. *IEEE Transactions on Neural Networks*, 8(3):486–504, 1997. doi: 10.1109/72.572090.

S. S. Khan and M. G. Madden. One-class Classification: Taxonomy of Study and Review of Techniques. *The Knowledge Engineering Review*, 29(3):345–374, 2014. doi: 10.1017/S026988891300043X.

D. P. Kingma and M. Welling. Auto-Encoding Variational Bayes, May 2014. URL https://arxiv.org/abs/1312.6114. arXiv:1312.6114.

D. P. Kingma, T. Salimans, R. Jozefowicz, X. Chen, I. Sutskever, and M. Welling. Improving Variational Inference with Inverse Autoregressive Flow, January 2017. URL https://arxiv.org/abs/1606.04934. arXiv:1606.04934.

A. N. Kolmogorov. Stationary sequences in Hubert space. In A. N. Shiryayev, editor, *Selected Works of A. N. Kolmogorov: Volume II Probability Theory and Mathematical Statistics*, pages 228–271. Springer Netherlands, Dordrecht, 1992. ISBN 978-94-011-2260-3. doi: 10.1007/978-94-011-2260-3_27.

B. Lakshminarayanan, A. Pritzel, and C. Blundell. Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/9ef2ed4b7fd2c810847ffa5fa85bce38-Paper.pdf.

C. L. Lan and L. Dinh. Perfect Density Models Cannot Guarantee Anomaly Detection, June 2021. URL https://arxiv.org/abs/2012.03808. arXiv:2012.03808.

K. Lee, H. Lee, K. Lee, and J. Shin. Training Confidence-calibrated Classifiers for Detecting Out-of-Distribution Samples. In *International Conference on Learning Representations*, 2018a. URL https://openreview.net/forum?id=ryiAv2xAZ.

K. Lee, K. Lee, H. Lee, and J. Shin. A Simple Unified Framework for Detecting Out-of-Distribution Samples and Adversarial Attacks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018b. URL https://proceedings.neurips.cc/paper/2018/file/abdeb6f575ac5c6676b747bca8d09cc2-Paper.pdf.

S. Liang, Y. Li, and R. Srikant. Enhancing The Reliability of Out-of-distribution Image Detection in Neural Networks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=H1VGkIxRZ.

L. Maaløe, M. Fraccaro, V. Liévin, and O. Winther. BIVA: A Very Deep Hierarchy of Latent Variables for Generative Modeling, November 2019. URL https://arxiv.org/abs/1902.02102. arXiv:1902.02102.

P. Masani. Wiener's Contributions to Generalized Harmonic Analysis, Prediction Theory and Filter Theory. *Bulletin of the American Mathematical Society*, 72(1.P2):73 – 125, 1966. doi: bams/1183527586. URL https://doi.org/.

K. R. Mestav and L. Tong. Universal Data Anomaly Detection via Inverse Generative Adversary Network. *IEEE Signal Processing Letters*, 27:511–515, 2020. doi: 10.1109/LSP.2020.2978462.

J. Morton, A. Jameson, M. J. Kochenderfer, and F. Witherden. Deep Dynamical Modeling and Control of Unsteady Fluid Flows. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/file/2b0aa0d9e30ea3a55fc271ced8364536-Paper.pdf.

G. Naik and D. Kumar. An Overview of Independent Component Analysis and Its Applications. *Informatica*, 35:63–81, 01 2011.

A. Omri, N. B. Erichson, V. Lin, and M. W. Mahoney. Forecasting Sequential Data Using Consistent Koopman Autoencoders. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pages 475–485, 2020.

L. Paninski. A Coincidence-Based Test for Uniformity Given Very Sparsely Sampled Discrete Data. *IEEE Transactions on Information Theory*, 54(10):4750–4755, Oct 2008. doi: 10.1109/TIT.2008.928987.

M. Pignati *et al.* Real-time State Estimation of the EPFL-Campus Medium-Voltage Grid by using PMUs. In *2015 IEEE Power Energy Society Innovative Smart Grid Technologies Conference (ISGT)*, pages 1–5, Feb 2015. doi: 10.1109/ISGT.2015.7131877.

M. B. Priestley. *Spectral Analysis and Time Series*. Academic Press, San Diego, CA, 1981.

J. Ren, Peter J. Liu, E. Fertig, J. Snoek, R. Poplin, M. A. DePristo, J. V. Dillon, and B. Lakshminarayanan. Likelihood Ratios for Out-of-Distribution Detection, December 2019. URL `https://arxiv.org/abs/1906.02845`. arXiv:1906.02845.

M. Rosenblatt. Stationary Processes as Shifts of Functions of Independent Random Variables. *Journal of Mathematics and Mechanics*, 8(5):665–681, 1959.

M. Rosenblatt. A Comment on a Conjecture of N. Wiener. *Statistics & Probability Letters*, 79(3):347–348, 2009. ISSN 0167-7152. doi: https://doi.org/10.1016/j.spl.2008.09.001. URL `https://www.sciencedirect.com/science/article/pii/S0167715208004173`.

T. Schlegl, P. Seeböck, S. M. Waldstein, G. Langs, and U. Schmidt-Erfurth. f-AnoGAN: Fast Unsupervised Anomaly Detection with Generative Adversarial Networks. *Medical Image Analysis*, 54:30 – 44, 2019. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2019.01.010.

B. Schölkopf, R. Williamson, A. Smola, J. Shawe-Taylor, and J. Platt. Support Vector Method for Novelty Detection. *Proceedings of the 12th International Conference on Neural Information Processing Systems*, pages 582–588, 1999.

J. Shao. *Mathematical Statistics*. Springer-Verlag, 2017.

F. Sossan, E. Namor, R. Cherkaoui, and M. Paolone. Achieving the Dispatchability of Distribution Feeders Through Prosumers Data Driven Forecasting and Model Predictive Control of Electrochemical Storage. *IEEE Transactions on Sustainable Energy*, 7(4):1762–1777, Oct 2016. ISSN 1949-3037. doi: 10.1109/tste.2016.2600103. URL `http://dx.doi.org/10.1109/TSTE.2016.2600103`.

A. Swami and J.M. Mendel. Identifiability of the AR Parameters of an ARMA Process using Cumulants. *IEEE Transactions on Automatic Control*, 37(2):268–273, 1992. doi: 10.1109/9.121633.

N. Takeishi, Y. Kawahara, and T. Yairi. Learning Koopman Invariant Subspaces for Dynamic Mode Decomposition. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL `https://proceedings.neurips.cc/paper/2017/file/3a835d3215755c435ef4fe9965a3f2a0-Paper.pdf`.

L. Tong. Identification of Multichannel MA Parameters using Higher-order Statistics. *Signal Processing*, 53(2):195–209, 1996. ISSN 0165-1684. doi: https://doi.org/10.1016/0165-1684(96)00086-2. Higher Order Statistics.

G. Tucker, D. Lawson, S. Gu, and C. J. Maddison. Doubly Reparameterized Gradient Estimators for Monte Carlo Objectives, November 2018. URL `https://arxiv.org/abs/1810.04152`. arXiv:1810.04152.

A. Vahdat and J. Kautz. NVAE: A Deep Hierarchical Variational Autoencoder, January 2021. URL `https://arxiv.org/abs/2007.03898`. arXiv:2007.03898.

I. I. Viktorova and V. P. Chistyakov. On the Calculation of the Power of the Test of Empty Boxes. *Theory of Probability & Its Applications*, 9(4):648–653, 1964. doi: 10.1137/1109089.

A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. J. Lang. Phoneme Recognition Using Time-delay Neural Networks. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 37(3):328–339, 1989. doi: 10.1109/29.21701.

X. Wang, Y. Liu, and L. Tong. Adaptive Subband Compression for Streaming of Continuous Point-on-Wave and PMU Data. *IEEE Transactions on Power Systems*, pages 1–1, 2021. doi: 10.1109/TPWRS.2021.3072882.

P. Whittle. *Hypothesis Testing in Time Series Analysis*. PhD thesis, Uppsala, 1951.

N. Wiener. *Nonlinear Problems in Random Theory*. Technology Press of Massachusetts Institute of Technology, Cambridge, MA, 1958.

W. Wu. Nonlinear System Theory: Another Look at Dependence. *Proceedings of National Academy of Sciences*, 102(40):14150–14154, 2005. URL `https://doi.org/10.1073/pnas.0506715102`.

W. Wu. Asymptotic Theory for Stationary Processes. *Statistics and Its Interface*, 0:1–20, 01 2011. doi: 10.4310/SII.2011.v4.n2.a15.

Appendix A. Proof of Theorem 1

Proof: By A2, there exists a sequence of IAEs $\tilde{\mathcal{A}}_m = (G_{\tilde{\theta}_m}, H_{\tilde{\eta}_m})$ of dimension m that converges to $\mathcal{A}$. Let $(\tilde{\nu}_{m,t})$ be the output sequences of $G_{\tilde{\theta}_m}$ and $(\tilde{x}_{m,t})$ the output sequence of the decoder H_{η_m}. Similarly defined are vector outputs of the encoder and decoder $\tilde{\mathbf{x}}_{m,t}^{(n)}$ and $\tilde{\boldsymbol{\nu}}_{m,t}^{(n)}$.

We prove Theorem 1 in two steps:

1. **Finite-block convergence of $\tilde{\mathcal{A}}_m$:** By assumption A2, the uniform convergence of $G_{\tilde{\theta}_m} \to G$ implies that, for every $\epsilon_1 > 0$, there exists $M_1 \in \mathbb{N}^+$ such that, for all realizations $\mathbf{x}_t^{(\infty)}$, $m > M_1$ and $t \in \mathbb{N}$,

$$|\tilde{\nu}_{m,t} - \nu_t| < \epsilon_1 \quad \Rightarrow \quad \mathbb{E}\left(\|\tilde{\boldsymbol{\nu}}_{m,t}^{(n)} - \boldsymbol{\nu}_t^{(n)}\|^2\right) \le n\epsilon_1.$$

Therefore, for all finite n, we have the following uniform convergence as $m \to \infty$:

$$\tilde{\boldsymbol{\nu}}_{m,t}^{(n)} \xrightarrow{L_2} \boldsymbol{\nu}_t^{(n)} \quad \Rightarrow \quad \tilde{\boldsymbol{\nu}}_{m,t}^{(n)} \xrightarrow{d} \boldsymbol{\nu}_t^{(n)}, \quad \Rightarrow \quad W(\tilde{\boldsymbol{\nu}}_{m,t}^{(n)}, \boldsymbol{\nu}_t^{(n)}) \to 0. \tag{12}$$

Next we consider the decoder convergence. Fix $\epsilon_2 > 0$.

$$\begin{aligned}|x_t - \tilde{x}_{m,t}| &= |H \circ G(\mathbf{x}_t^{(\infty)}) - H_{\tilde{\eta}_m} \circ G_{\tilde{\theta}_m}(\mathbf{x}_t^{(m)})| \\ &\le |H \circ G(\mathbf{x}_t^{(\infty)}) - H \circ G_{\tilde{\theta}_m}(\mathbf{x}_t^{(m)})| \ + |H \circ G_{\tilde{\theta}_m}(\mathbf{x}_t^{(m)}) - H_{\tilde{\eta}_m} \circ G_{\tilde{\theta}_m}(\mathbf{x}_t^{(m)})|.\end{aligned}$$

Because H is uniform continuous, there exists an $M_2(\epsilon_2)$ such that for all $m > M_2(\epsilon_2)$ and $\mathbf{x}_t^{(\infty)}$,

$$|H \circ G(\mathbf{x}_t^{(\infty)}) - H \circ G_{\tilde{\theta}_m}(\mathbf{x}_t^{(m)})| \ < \epsilon_2/2.$$

Because H_{η_m} converges to H uniformly, there exists an $M_2'(\epsilon_2)$ such that

$$|H \circ G_{\tilde{\theta}_m}(\mathbf{x}_t^{(m)}) - H_{\tilde{\eta}_m} \circ G_{\tilde{\theta}_m}(\mathbf{x}_t^{(m)})| < \epsilon_2/2,$$

for all $m > M_2'(\epsilon_2)$ and $\mathbf{x}_t^{(\infty)}$. Therefore, for all $m > \max\{M_2(\epsilon_2), M_2'(\epsilon_2)\}$,

$$|\tilde{x}_{m,t} - x_t| < \epsilon_2 \quad \Rightarrow \quad \mathbb{E}\left(\|\tilde{\mathbf{x}}_{m,t}^{(n)} - \mathbf{x}_t^{(n)}\|^2\right) \le n\epsilon_2 \quad \Rightarrow \quad \tilde{\mathbf{x}}_{m,t}^{(n)} \xrightarrow[m\to\infty]{L_2} \mathbf{x}_t^{(n)}.$$

The risk $\tilde{L}_m^{(n)}$ converges uniformly:

$$\tilde{L}_m^{(n)} := W(\tilde{\boldsymbol{\nu}}_{m,t}^{(n)}, \boldsymbol{\nu}_t^{(n)}) + \mathbb{E}(\|\tilde{\mathbf{x}}_{m,t}^{(n)} - \mathbf{x}_t^{(n)}\|_2^2) \xrightarrow{m\to\infty} 0. \tag{13}$$

2. **Finite-block convergence of $\mathcal{A}_m^{(n)}$:** Fix the dimension of training at n. From the finite-block convergence of $\tilde{\mathcal{A}}_m$, $\forall\epsilon$, there exists M_ϵ such that, for all $m > M_\epsilon$,

$$\tilde{L}_m^{(n)} = W(\tilde{\boldsymbol{\nu}}_{m,t}^{(n)} - \boldsymbol{\nu}_t^{(n)}) + \mathbb{E}\left(\|\tilde{\mathbf{x}}_{m,t}^{(n)} - \mathbf{x}_t^{(n)}\|_2^2\right) \le \epsilon. \tag{14}$$

By the Kantorovich-Rubinstein duality theorem,

$$W(\tilde{\boldsymbol{\nu}}_{m,t}^{(n)}, \boldsymbol{\nu}_t^{(n)}) = \max_{\gamma} \mathbb{E}(D_\gamma(\tilde{\boldsymbol{\nu}}_{m,t}^{(n)}, \boldsymbol{\nu}_t^{(n)})).$$

From (3), let (without loss of generality assuming $\lambda = 1$)

$$L_m^{(n)} := \min_{\theta,\eta} \max_{\gamma} \left(\mathbb{E}(D_\gamma(\tilde{\mathbf{x}}_{m,t}^{(n)}, \boldsymbol{\nu}_t^{(n)}) + \mathbb{E}(\|\mathbf{x}_{m,t}^{(n)} - \mathbf{x}_t^{(n)}\|_2^2) \right).$$

We therefore have, for all $m \geq M_\epsilon$,

$$L_m^{(n)} \leq \tilde{L}_m^{(n)} \leq \epsilon \quad \Rightarrow \quad \begin{cases} W(\boldsymbol{\nu}_{m,t}^{(n)}, \boldsymbol{\nu}_t^{(n)}) \leq \epsilon, \\ \mathbb{E}(\|\mathbf{x}_{m,t}^{(n)} - \mathbf{x}_t^{(n)}\|_2^2) \leq \epsilon, \end{cases}$$

which completes the proof. □□□

Appendix B. Pseudocode

Algorithm 1 Training the Innovations Autoencoder

Input: data (x_t), encoder H_η, generator G_θ, discriminator D_γ. λ is the gradient penalty coefficient, μ the weight for auto-encoder, and α, β_1, β_2 hyper-parameters for Adam optimizer.

while Not converged **do**
  for $t = 1, \cdots, n_c$ **do**
    for $i = 1, \cdots, B$ **do**
      Sample $\mathbf{x}_i$ from the input matrix (x_t)
      Sample $\mathbf{u}= [u_1, \cdots, u_n]^T \overset{i.i.d}{\sim} \mathcal{U}[-1, 1]$
      Sample $\epsilon \sim \mathcal{U}[0, 1]$
      $\hat{\boldsymbol{\nu}} \leftarrow \mathbf{G}_\theta(\mathbf{x}_i)$
      $\bar{\boldsymbol{\nu}} \leftarrow \epsilon\mathbf{u} + (1-\epsilon)\hat{\boldsymbol{\nu}}$
      $L^{(i)} \leftarrow \mathbf{D}_\gamma(\hat{\boldsymbol{\nu}}) - \mathbf{D}_\gamma(\mathbf{u}) + \lambda(\|\nabla_\gamma \mathbf{D}_\gamma(\bar{\boldsymbol{\nu}})\|_2 - 1)^2$
    end for
    $\gamma \leftarrow Adam(\nabla_\gamma \frac{1}{B} \sum_{i=1}^B L^{(i)}, \alpha, \beta_1, \beta_2)$
  end for
  Sample a batch of $\{\mathbf{x}_i\}_{i=1}^B$ from the input matrix (x_t)
  $\theta \leftarrow Adam\left(\nabla_\theta \frac{1}{B} \sum_{i=1}^B \Big[-\mathbf{D}_\gamma(\mathbf{G}_\theta(\mathbf{x}_i)) + \mu\|\mathbf{H}_\eta(\mathbf{G}_\theta(\mathbf{x}_i)) - \mathbf{x}_i\|_2 \Big], \alpha, \beta_1, \beta_2 \right)$
  $\eta \leftarrow Adam\left(\nabla_\eta \frac{1}{B} \sum_{i=1}^B \left[\mu\|\mathbf{H}_\eta(\mathbf{G}_\theta(\mathbf{x}_i)) - \mathbf{x}_i\|_2 \right], \alpha, \beta_1, \beta_2 \right)$
end while

Appendix C. Neural Network Parameter

All the neural networks (encoder, decoder and discriminator) in the paper had three hidden layers, with the 100, 50, 25 neurons respectively. The input dimension for the generator was chosen such that $n = 3m$. In the paper, $m = 20$ was used for synthetic case, and $m = 100$ for real data cases. The encoder and decoder both used hyperbolic tangent activation. The first two layers of the discriminator adopted hyperbolic tangent activation, and the last one linear activation.

The tuning parameter was chosen to be the same for all synthetic cases, with $\mu = 0.1$, $\lambda = 5$, $\alpha = 0.0002$, $\beta_1 = 0.9$, $\beta_2 = 0.999$. For the two real data cases, the hyper-parameters were set to be $\mu = 0.01$, $\lambda = 3$, $\alpha = 0.001$, $\beta_1 = 0.9$, $\beta_2 = 0.999$.