TOWARDS INFORMATION-THEORETIC PATTERN MINING IN TIME SERIES

Anonymous authors
Paper under double-blind review

ABSTRACT

Time series pattern discovery is one of the most fundamental tasks in data mining. Existing literature addressing this task often follows a generic paradigm in which a similarity metric is defined a priori and an extensive pattern-matching search is executed to find similar subsequences based on the metric. Algorithms developed under this paradigm therefore risk missing important patterns that do not meet the implicit biases within such pre-defined metrics. To mitigate this risk, we propose a new information-theoretic discovery paradigm that aims to find the most informative patterns on an embedding space that can learn to encode representative statistical variation trends in the time series. This paradigm is achieved by a probabilistic time-to-pattern mining algorithm, T2P, based on a biophysically-inspired adaptation of a variational auto-encoder (VAE). The adapted VAE incorporates a specific design for its latent space that learns to surface the most recurring and informative patterns without the need to run costly pattern-matching searches. Empirically, we demonstrate that our method is more scalable than existing works. Furthermore, T2P can find multiple diverse patterns that more effectively compress and represent the time series without relying on prior knowledge of the data.

1 INTRODUCTION

Finding informative and recurring patterns that summarize time series data effectively is a critical task in data-intensive applications (e.g., health monitoring, stock prediction) that require evidence-based decision-making and optimization. This is in fact one of the most fundamental tasks in a long-standing field of data mining with a vast and growing literature, reporting many successful findings that acquiring such temporal patterns often provides significant predictive and actionable data insights (Ali et al., 2019) for downstream data analytics. However, in spite of such findings, pattern discovery in time series still remains a notoriously difficult task due to:

C1. The explosive growth in the scale of data enabled by recent advances in IoT sensors which are capable of collecting data at high frequencies and creating very large time series that are voluminous, complex and unintuitive; and

C2. The lack of a robust notion or criteria that accurately and holistically incorporate both aspects of pattern informativeness and recurrence in the existing literature: whether a definition of pattern recurrence also implies pattern informativeness.

Both of these concerns severely hamper the scalability and robustness of most previous work in this field. To elaborate, existing literature (Dau & Keogh, 2017; Alaee et al., 2021) addressing this pattern mining task often follows a generic paradigm in which a similarity metric is defined a priori and an extensive pattern-matching search is executed to find similar subsequences based on the metric. Under such a paradigm, a pattern is associated with the occurrences of subsequences which are deemed sufficiently similar by a pre-defined metric.

However, without a carefully handcrafted metric, the occurrence frequencies of such patterns might not be a correct measurement of their encoded information. Algorithms developed under this view therefore risk missing important patterns that do not meet the implicit biases encoded by a pre-selected metric. For example, Euclidean distance (ED) (Dau & Keogh, 2017) biases towards minimal oscillation but depending on the specific nature of the data, patterns with minimal oscillation might not be most informative. Likewise, Dynamic Time Warping (DTW) distance (Alaee et al.,

2021) biases more towards warped patterns, which might also not be the most informative for some applications.

Furthermore, the definition of a pattern and its representativeness in terms of recurring subsequences deemed sufficiently similar by some distance function naturally leads to an extensive pattern-matching search to find all such subsequences. Despite substantial work to improve the scalability of the current methods by improving the computational complexity of the similarity function or representing the input in a compressed format, the performance on large-scale data remains a challenge: As the length of the time series increases, the number of distances needed to be calculated between each pair of subsequences inevitably increases quadratically. Alternatively, some methods instead propose to compress data in a pre-processing step (Keogh et al., 2005) to reduce the search space, but the compression is lossy and may cause important patterns to be undetected.

Motivated by the above shortcomings of previous works in light of challenges **C1** and **C2**, we develop and investigate a new approach, named *time-to-pattern* (T2P), to discover informative patterns in time series data that addresses both challenges via an information-theoretic framework. In particular, unlike previous works that use the frequencies of (artificial) similar occurrences as a heuristic measure of information, T2P rigorously learns an embedding or encoding of the time series in a latent space, which can be decoded back into the original space with minimized loss. The embedding loss is in fact structured in the spirit of the Minimum Description Length (MDL) principle (Wallace & Boulton, 1968), which formalizes Occam's razor and states that the best data model is the one that represents the data in the smallest number of bits.

The latent space is further devised to induce factorized encoding, which implicitly decomposes the time series into a diverse set of disengtangled patterns in the latent space spanning its information spectrum, which can be decoded back (with minimized loss) to the original space. We achieve this by adopting the seminal variational auto-encoder (VAE) framework (Honkela & Valpola, 2004; Rezaabad & Vishwanath, 2020; Higgins et al., 2017; Kirschbaum et al., 2019; Guo et al., 2021) which has been established as an effective probabilistic approach to unsupervised data embedding.

Furthermore, the design of the latent space for our adopted VAE is biophysically inspired by neuronal assemblies theory by (Hebb, 2005), which identifies subsets of neurons firing in a temporally coordinated way that gives rise to repeated patterns underlying neural representation and information processing. This corroborates a prior neuroscience work which has shown that neurons are able to learn a specific representation for each smell (e.g., a rose or a fertilizer) by capturing the neural spikes (Laurent, 1996). We implement this via imposing a Bernoulli prior on the latent space to creating stochastic nodes representing recurring spatio-temporally coordinated activation patterns.

To summarize, our technical contributions include:

1. We formulate *time-to-pattern* (T2P), a new information-theoretic approach to pattern discovery in time series data which defines a pattern as an optimized MDL encoding of the original time series data in an embedding space. This provides a coherent, unified view of both aspects of pattern recurrence and informativeness which is lacking in the existing literature.

2. We implement T2P via a learnable VAE with a Bernoulli prior implementing the principle of neuronal assemblies theory (Hebb, 2005) which help decomposes the latent encoding of time series into disentangled patterns spanning more effectively its information spectrum. This demonstrates an advantage of T2P in comparison to previous work: it does not depend on any specific distance function and is able to discover a set of diverse, informative patterns that summarize the time series better. Its incurred complexity (of learning a VAE) is also significantly better than that of a state-of-the-art pattern-matching approach (Imani et al., 2018).

3. We show experimentally that the running time of T2P remains almost constant as the size of the data increases. We also show that T2P outperforms previous work in discovering the most informative patterns in several real-life benchmark datasets.

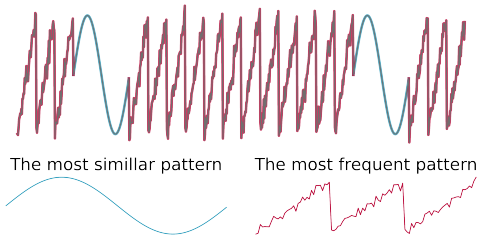


Figure 1: Plots of (top) synthetic data; and examples of (bottom left) the most similar pattern; and (bottom right) the most frequent pattern.

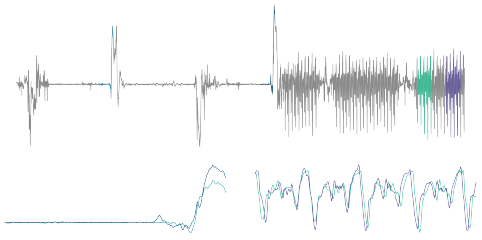


Figure 2: Plots of (top) real data from a hip-worn gyroscope Dau et al. (2018); (bottom left) top-1 Euclidean-distance subsequence pair; (bottom right) top-1 DTW-distance subsequence pair.

2 Related Work

2.1 Pattern Mining for Time Series

Prior work on discovering time series patterns can be broadly categorized into (1) similarity-based and (2) frequency-based methods. We review them below.

Similarity-based Methods (Zhu et al., 2016; Alaee et al., 2021). The main aim of these approaches is to find a pair of subsequences of fixed length that are most similar according to a pre-defined metric distance (e.g., Euclidean). The subsequences are usually referred to as motifs or patterns.

Formally, let $\mathbf{x} \triangleq (x_1, x_2, \ldots, x_n)$ denote a time series of length n and let $\mathbf{x}_t^\tau \triangleq (x_t, \ldots, x_{t+\tau})$ denote a sub-sequence from timestep t to $t+\tau$. The most similar pattern of length τ is defined as a pair of sub-sequences $p_\tau = (\mathbf{x}_a^\tau, \mathbf{x}_b^\tau)$ such that $\ell(\mathbf{x}_a^\tau, \mathbf{x}_b^\tau) \leq \ell(\mathbf{x}_u^\tau, \mathbf{x}_v^\tau)\ \forall(u, v)$ where $u, v \in [n-\tau]$ and ℓ denotes a distance measure (e.g., Euclidean) which is set depending on particular applications.

Frequency-based Methods (Imani et al., 2018). Unlike similarity-based methods, the aim of frequency-based methods is to find the largest set of sub-sequences (of fixed length) such that any two sub-sequences are sufficiently close according to a distance measure.

Formally, a frequent pattern is defined as $p_{\tau,\epsilon} = (\mathbf{x}_{a_1}^\tau, \ldots, \mathbf{x}_{a_k}^\tau)$ such that (1) $\ell(\mathbf{x}_a^\tau, \mathbf{x}_{a'}^\tau) \leq \epsilon$ for all pairs of $a, a' \in \{a_1, a_2, \ldots, a_k\}$ and (2) if $a' \notin \{a_1, a_2, \ldots, a_k\}$, then $\ell(\mathbf{x}_{a'}, \mathbf{x}_{a_i}) > \epsilon$ for $i \in [k]$. That is, the pattern $p_{\tau,\epsilon}$ is a maximal set of all sub-sequences with length τ whose pairwise distance is within a pre-defined similarity threshold ϵ. Its frequency is therefore denoted by the size of $|p_{\tau,\epsilon}|$. The most frequent pattern is the pattern with the largest size.

Discussion. The difference between these two pattern definitions is illustrated above in Figure 1, which shows examples of most-similar and most-frequent patterns in a synthetic time series. Both approaches are restricted in terms of the informativeness of the discovered patterns (i.e., how much information of the original time series that is encoded within the discovered patterns) and the incurred computation cost to find them, as further discussed below. In addition, relying on a specific distance function produces biased results. As illustrated in Figure 2, the most-similar pair is different when using z-normalized ED versus DTW.

Pattern Informativeness. Characterizing a time series in terms of its patterns induces a compressed representation, which can be used for a wide range of downstream applications, including data interpretation and predictive analytics. For such a representation to be effective, it must be able to encode the most salient information of the time series, which can be formalized either in terms of mutual information or perhaps more interestingly, the reconstructability of the time series based on the patterns. This aspect, however, is seemingly unaccounted for by most existing approaches, which is evident in their dependence on a pre-defined distance metric ℓ as well as the similarity threshold ϵ as mentioned above: without learning the embedding space of the time series, the preset distance and threshold might not reflect well the subsequences' intrinsic similarities. This could potentially result in the accidental exclusion of important patterns which do not meet the implicit biases encoded by a preset metric. For example, Euclidean distance (Dau & Keogh, 2017) biases

towards minimal oscillation but depending on the specific nature of the data, patterns with minimal oscillation (e.g., flat line) might not be most informative.

Recently, Noering et al. (2021) proposed a new approach that accounts for both the frequency and informative aspects focusing on the reconstructability of a time series based on its patterns. This is achieved via a convolutional autoencoder (CAE), which converts the time series to a gray-scale image by discretizing the data into a given number of bins and then converting each bin to one-hot encoding (Noering et al., 2021). The induced image is then embedded into a latent space via a (learnable) convolutional neural net (i.e., the encoder). Its latent embedding vector is optimized so that it can be used to reconstruct the original input via a (learnable) decoder net. Each of its latent components can then be used to generate a separate, representative pattern using the decoder. CAE however suffers a significant overhead of converting the time series to discrete images. Further, CAE is not able to discover diverse patterns and tends to discover patterns with low complexity as a result of converting the time series to discrete images, as shown by the authors' experiments (Noering et al., 2021).

Scalability of Pattern Mining. Furthermore, most of the current time series pattern discovery methods are based on exhaustive search, which is the natural approach given the definition of a pattern and its representativeness in terms of recurring subsequences deemed sufficiently similar by some distance function. For example, using the Euclidean distance as the similarity function, the cost of finding the most similar pattern with length τ from a time series of length n is $O(\tau n^2)$. The cost of finding the most frequent pattern is even worse, scaling exponentially in the size of the time series $O(\tau n^2)+O(Kn)$ for top-K patterns. As such, it can be seen that the complexity of exhaustive search often scales very poorly in the size of the time series.

To sidestep this, two approaches have been proposed to reduce computational complexity. The first approach (Ding et al., 2008) focuses on compressing the time series length, thus decreasing the search space. However, the compression methods are lossy, which might omit important information that results in the exclusion of some key patterns. To avoid this, the second approach (Zhu et al., 2016; Alaee et al., 2021) focuses instead on decreasing the computational complexity of similarity functions. For example, STOMP (Zhu et al., 2016) proposed an approximation that decreases the cost of computing the Euclidean distance (ED) between subsequences from $O(\tau)$ to $O(1)$, and likewise, SWAMP (Alaee et al., 2021) reduces the cost of computing the Dynamic Time Warping (DTW) distance from $O(\tau^2)$ to $O(\tau)$. Nonetheless, the complexities of such approaches are still dependent on a factor of $O(n^2)$, which is the main source of the computation bottleneck.

2.2 Variational Autoencoders

Introduced by Kingma & Welling (2014); Rezende et al. (2014), VAEs are probabilistic embedding models. A VAE characterizes a distribution $p_\theta(\mathbf{x}, \mathbf{z}) = p(\mathbf{z})p_\theta(\mathbf{x}|\mathbf{z})$ over both the observed data $\mathbf{x}$ and its latent embedding $\mathbf{z}$. Its parameter is characterized by θ, which is often the weight tensors of a deep neural net (DNN). In this view, $p(z)$ is often characterized by a parameter-free prior distribution, e.g. a Gaussian $\mathbb{N}(\mathbf{0}, \mathbf{I})$, over latent variable $\mathbf{z}$. Its likelihood function (or the decoder) $p_\theta(\mathbf{x}|\mathbf{z})$ is learned to generate authentic data $\mathbf{x}$ given latent variables $\mathbf{z}$. Its posterior $p_\theta(\mathbf{z}|\mathbf{x})$ is conversely designated as the encoder and needs to be computed from the above generative scheme. Such posterior inference in most cases is however intractable, so VAE bypasses this by approximating it with a surrogate $q_\phi(\mathbf{z}|\mathbf{x})$ such that $\mathrm{D}_{\mathrm{KL}}(q_\phi(\mathbf{z}|\mathbf{x})\|p_\theta(\mathbf{x}|\mathbf{z}))$ is minimized. This is interestingly equivalent to maximizing the following lower-bound $\mathbf{L}(q_\phi)$ of the model evidence

$$\log p(\mathbf{x}) \;\geq\; \mathbf{L}(q_\phi) \;\triangleq\; \mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x})}\Big[\log p_\theta(\mathbf{x}|\mathbf{z})\Big] - \mathrm{D}_{\mathrm{KL}}\Big(q_\phi(\mathbf{z}|\mathbf{x}) \;||\; p(\mathbf{z})\Big) \tag{1}$$

This makes sense since $\log p(\mathbf{x}) = \mathbf{L}(q_\phi) + \mathrm{D}_{\mathrm{KL}}(q_\phi(\mathbf{z}|\mathbf{x})\|p_\theta(\mathbf{x}|\mathbf{z}))$. Intuitively, the first term in Eq. 1 reduces the reconstruction error and the second term minimizes the Kullback–Leibler divergence between the posterior distribution $q_\phi(z|x)$ and the prior distribution $p_a(z)$. However, the gradient w.r.t. ϕ cannot be easily computed. To overcome this problem and generate samples from $q_\phi(\mathbf{z})$, a re-parameterization trick has been proposed. The random variable $\mathbf{z} \sim q_\phi(\mathbf{z}|\mathbf{x})$ can be re-parameterized using a differentiable transformation $\mathbf{z} = g_\phi(\boldsymbol{\epsilon}, \mathbf{x})$, where $\boldsymbol{\epsilon}$ is an auxiliary variable with independent, parameter-free marginal $p(\boldsymbol{\epsilon})$ (that can be sampled from) and $g_\phi(.)$ is a (learnable) vector-valued function parameterized by ϕ characterizing the transformation from $\boldsymbol{\epsilon}$ to $\mathbf{z}$. Appendix A provides more information on VAEs.

3 INFORMATION-THEORETIC PATTERN MINING

This section formalizes a new notion of the information-theoretic pattern (Section 3.1) and proposes a scalable solution (Section 3.2) to discover it via augmenting the aforementioned VAE model for data embedding with a concept of neuronal assembly by (Hebb, 2005), which (1) optimizes a probabilistic embedding of the time series with minimal information loss; and (2) extracts a diverse set of informative patterns from such latent embedding.

3.1 INFORMATION-THEORETIC PATTERN

Let $\mathbf{x} \triangleq (x_1, x_2, \ldots, x_n)$ denote a time series of length n. Let $p_\tau = \{\mathbf{x}_1^\dagger, \mathbf{x}_2^\dagger, \ldots, \mathbf{x}_m^\dagger\}$ denote a set of m sequences of length τ. We consider p_τ a set of informative patterns if it can be used as a sufficient statistic to infer any predictable (but unobservable) statistics $c(\mathbf{x})$ of $\mathbf{x}$. That is, p_τ satisfies $p(c(\mathbf{x}) \mid p_\tau, \mathbf{x}) = p(c(\mathbf{x}) \mid p_\tau)$ or $c(\mathbf{x}) \perp \mathbf{x} \mid p_\tau$ for any predictable statistics $c(\mathbf{x})$.

Motivated by the above VAE model (Section 2.2), a natural approach to finding such p_τ is perhaps to associate p_τ with the probabilistic latent encoding $\mathbf{z}$ of $\mathbf{x}$, which can provably be used to reconstruct $\mathbf{x}$ with high-fidelity as a result of maximizing Eq. 1. However, such embedding $\mathbf{z}$ does not live in the same space as $\mathbf{x}$ and it is therefore unclear how one would associate components of $\mathbf{z}$ with representative sub-sequences of $\mathbf{x}$. This raises the question of whether we can devise a latent space for VAE that induces factorized encoding, which helps decompose the time series $\mathbf{x}$ into a diverse set of disengtangled latent patterns spanning its information spectrum, which can be decoded back (with minimized loss) into legitimate time series $\{\mathbf{x}_1^\dagger, \mathbf{x}_2^\dagger, \ldots, \mathbf{x}_m^\dagger\}$ living in the original space.

We will address this question via an adaptation of the VAE framework inspired by the concept of neuronal assembly of (Hebb, 2005), which is discussed next.

3.2 DISCOVERING INFORMATION-THEORETIC PATTERNS

This section presents *time-to-pattern* (T2P), which is a scalable information-theoretic framework to discover the pattern defined above. As mentioned above, it is based on a VAE whose latent space implements the principle of neuronal assembly, which is concisely described below.

Neuronal Assemblies. The concept of *neuronal assemblies*, which is also known as *cortical motif* or *neuronal ensemble*, was originally introduced by Hebb (2005). Neuronal assemblies are subsets of neurons firing in a temporally coordinated way that gives rise to repeated motifs underlying neural representation and information processing. There is evidence that all cognitive domains can benefit from learning based on the Hebbian principle (Pulvermüller et al., 2021). For example, the fire-together-wire-together principle leads to long-term potentiation (LTP) of connections between co-activated neurons (Panahi et al., 2021). LTP is a process that synaptic connections between neurons become stronger with frequent activation. Different methods have been proposed for capturing the temporal structure of embedded assemblies in neural networks. PCA (Nicolelis et al., 1995) and ICA (Comon, 1994) are the simplest methods for detecting cell motifs. More sophisticated methods have been proposed recently, such as the utilization of sparse convolutional coding for reconstructing the neural spike matrix as a convolution of motifs and their activations time points (Peter et al., 2017). This concept has in fact been recently adopted in a recent work of (Kirschbaum et al., 2019), which implements a VAE with discrete latent space which enables effective extraction of repetitive, summarizing patterns in imaging videos of calcium activity in the brain.

VAE with Neuronal Assembly Prior. Inspired by such prospect of combining principles of neuronal assemblies with probabilistic data embedding model, we further develop T2P, a VAE framework for discovering informative patterns in time series data with neuronal assembly prior. We hypothesize that if repetitive patterns exist in time series data, the neural network should learn the same representation for them based on neuronal assembly theory.

Previous research (Alemi et al., 2016; Higgins et al., 2017; Achille et al., 2018; Achille & Soatto, 2018; Kirschbaum et al., 2019) that explores optimal representations also draws from information theory. They utilize VAEs with a discrete latent space to learn patterns that describe image data. VEAs with a discrete latent space can capture the neural spikes that result in discovering neuronal assemblies, as shown by Kirschbaum et al. (2019). Further, VAEs with a discrete space can learn disentangled representations (Dupont, 2018).

Hence, we utilized a Bernoulli-VAE to implement this principle. However, Bernoulli is a discrete distribution that poses a significant challenge because it cannot be re-parameterized using the conventional reparameterization trick developed by (Kingma & Welling, 2014). Several methods have been proposed to treat Bernoulli discrete nodes continuously, such as Binary Concrete (Maddison et al., 2016), Beta-Bernoulli (Singh et al., 2017), and Continuous Bernoulli (Loaiza-Ganem & Cunningham, 2019) distributions. In T2P, the Binary Concrete distribution is chosen as the Bernoulli relaxation since it produced superior results in our initial testing. This is detailed below.

3.3 Derivation of T2P Framework

For each subsequent $\mathbf{x}^{\dagger}$ a latent random variable $z \in \{0, 1\}$ is drawn from a prior distribution $p_a(z)$. The variable z indicates whether pattern p is activated in the subsequence. The subsequence $\mathbf{x}^{\dagger}$ is then generated from the conditional distribution $p_\theta(x|z)$ with parameters θ. Consider an unnormalized parameterization (α_1, α_2) and let $\alpha = \alpha^1/\alpha^2$ where $\alpha_1, \alpha_2 \in (0, \infty)$. Here, the reparameterization trick for the Binary Concrete distribution proceeds as follows:

$$z = \sigma(Y) = \frac{1}{1 + \exp(-Y)} \tag{2}$$

with

$$Y = \frac{\log \alpha + \log U - \log(1 - U)}{\lambda} \tag{3}$$

where U is drawn from uniform random numbers $U \sim Uniform(0, 1)$. Correspondingly, Eq. (1) can be written as

$$MSE(x, \hat{x}) - KL(g_{\alpha,\lambda_1}(Y|x)||f_{a,\lambda_2}(Y)) \tag{4}$$

where $MSE(x, \hat{x})$ represents the mean square error between input x and decoder output $\hat{x}$, $g_{\alpha,\lambda_1}(Y|x)$ is the reparameterized Binary Concrete relaxation of the variational posterior $q_\phi(z|x)$ with temperature λ_1, and $f_{a,\lambda_2}(Y)$ is the is the density of a Logistic random variable sampled via Eq. 3 with temperature λ_2 and the location of the prior distribution, a. The density of the Binary Concrete distribution is computed as follows:

$$\frac{\lambda \alpha x^{-\lambda-1}(1-x)^{-\lambda-1}}{(\alpha x^{-\lambda} + (1-x)^{-\lambda})^2}. \tag{5}$$

Combining the equations, we can specify that the objective of T2P is to minimize

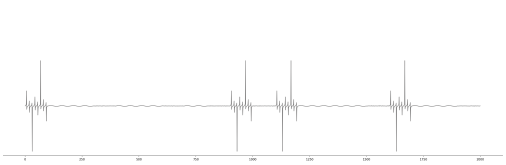
$$\beta_{MSE}.MSE(x, \hat{x}) + \beta_{KL}.\mathbb{E}_{Y \sim g_{\alpha,\lambda_1}(y|x)}\left[\log \frac{f_{a.\lambda_2(Y)}}{g_{\alpha,\lambda_1}(Y|x)}\right] \tag{6}$$

where β_{MSE} and β_{KL} are hyperparameters that control the effects of MSE and KL, respectively. Learning a particular pattern suited to represent a set of similar subsequences is difficult. In complex time series data, representative patterns exhibit slight variations with each occurrence. To capture such interesting and approximate occurrences, we need to choose $\beta_{MSE} < 1$ and $\beta_{KL} > 1$. To illustrate the effect of the β_{MSE} temperature, consider a synthetic sequence containing normal noise, a noisy sine wave pattern, and another pattern, as shown in Figure 3d. When $\beta_{MSE} = 1$ as in the conventional VAE model, T2P is not able to learn the embedded pattern, as shown in Figure 3c. In contrast, T2P can learn all the existing patterns by reducing the value to $\beta_{MSE} = 0.8$, as shown in Figure 3d. Similarly, as illustrated by (Higgins et al., 2017), choosing $\beta_{KL} > 1$ results in a more disentangled latent representation z.

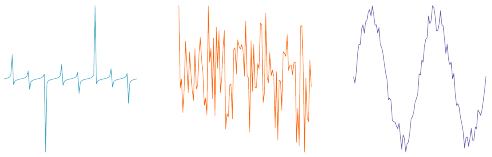
3.4 Component Parameterization for T2P

The T2P encoder network consists of convolution layers with 1D filters that extract features from the time series subsequences. Next, the learned features feed to a convolution layer with 3D filters to learn parameters α utilized by Eq. 3 for reparametrization. The decoder is a single deconvolution layer that enforces the reconstructed data $\hat{x}$ to be an additive mixture of the decoder filters. Thus, the decoder filters contain the detected patterns after minimizing the objective function in Eq. 6.

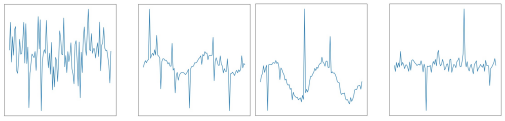
For a given time series T with window length L, we propose to train the model multiple subsequence windows. Given N windows, training T2P consists of positioning i randomly in the time series and



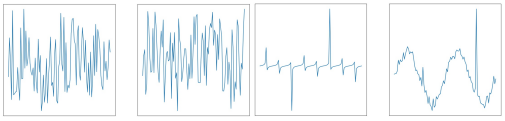
(a) The synthetic data.



(b) Three existing patterns in the data.



(c) Learned pattern by T2P with $\beta_{MSE} = 1$.



(d) Learned pattern by T2P with $\beta_{MSE} = 0.8$.

Figure 3: Effect of β_{MSE} on learning correct patterns from the data.

training the model on windows starting at those locations. We note that the choice of N does not impact training quality – T2P uses this parameter to report which patterns were activated.

Training T2P is straightforward. Parameter a influences the sparsity of the latent space z. A smaller value of a will penalize the activations harder and result in cleaner and more meaningful patterns. Based on our experiments, we recommend $a = 0.1$ or $a = 0.05$, which worked well in all cases. B_{KL} is another parameter that improves the diversity of patterns learned by T2P. In our experiments, we observe $B_{KL} = 2$ works in all cases, and changing to a higher value does not help with learning more diverse patterns. For the temperature parameters, the default values $\lambda_1 = 0.6$ and $\lambda_2 = 0.5$ worked well in most cases, and changing them is usually unnecessary. Figure 4 displays the plate diagram of the proposed generative and recognition models.

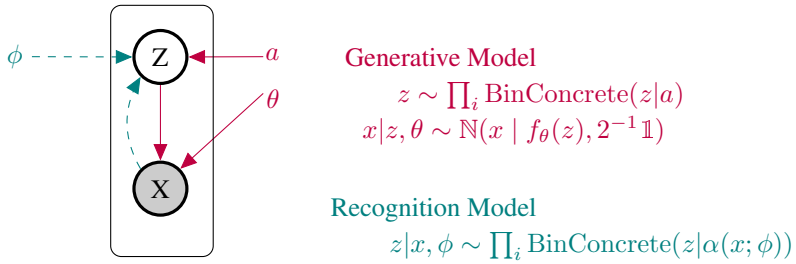


Figure 4: Plate diagram and proposed generative and recognition model. Solid lines denote the generative model $p_\theta(x|z)$, dashed lines denote the variational approximation $q_\phi(z|x)$.

4 Pattern Discovery Metrics

It is important to be able to quantify the performance of time series pattern discovery algorithms. However, designing specific metrics is not straightforward. Previous articles report anecdotal performance, such as preferring a method that discovers patterns with more oscillation over a method that discovers flat patterns. We begin by defining the properties that we expect discovered patterns should exhibit. We then describe our proposed solution for quantifying the presence of such properties.

Based on the definitions in Section 3, we are able to measure the performance of time series pattern discovery algorithms by utilizing Minimum Description Length (MDL). Introduced by Rissanen, MDL states that the best theory to describe a set is to minimize the description length of the entire dataset more than other theories Rissanen (1978). Hence, a pattern discovery method is evaluated based on how well it can compress the data. According to the MDL principle, a method that minimizes $DL(P) + DL(X|P)$ is preferred, where P is the discovered patterns, X is the input time series, $DL(P)$ is the number of units (e.g., bits) required to encode the discovered pattern, and $DL(X|P)$ is the number of units required to encode the time series X with respect to P.

5 EXPERIMENTAL RESULTS

We validate the performance of T2P using a series of experiments. To our knowledge, T2P is the first probabilistic model that learns an aggregate summary of patterns from raw time series data. Hence, there is no directly-comparable baseline to consider. To provide a basis for comparison, we include results using a current state-of-the-art time series pattern discovery algorithm, matrix profile snippets (MP-snippets) Imani et al. (2018). We first test our method on synthetic datasets containing predefined, embedded patterns.

We hypothesize that T2P offers advantages over prior discovery methods. First, because T2P does not utilize a fixed similarity measure, the algorithm holds the potential to find more complex patterns. Second, because T2P is governed by the information-theoretic preference for compressive patterns, T2P's discoveries do not focus exclusively on frequent patterns but instead balance pattern size and frequency. Third, unlike earlier approaches, T2P does not rank the patterns but instead finds a set of disentangled patterns, eliminating the need for priors. Finally, T2P replaces the traditional search-based approach with a VAE learning model, offering improved scalability to larger time series.

5.1 DATASETS

To provide data for analysis of pattern discovery methods, we start with real data containing known patterns and combine these time series to generate larger series with multiple embedded patterns. Additionally, we performed experiments on the MixedBag Imani et al. (2018) and UCR Dau et al. (2018) time series collections.

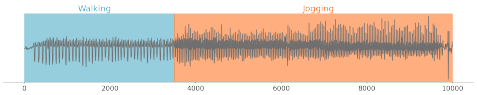
5.2 HUMAN ACTIVITY ANALYSIS

Analyzing human behavior is one of the most popular applications of pattern discovery Alaee et al. (2021). For this experiment, we consider data collected for one person living in a semi-structured environment Imani et al. (2018). As Figure 5a shows, there are two embedded patterns corresponding to the two scripted activities. MP-snippets yields the patterns shown in Figure 5b. To successfully find both patterns, the user must have prior information that there are two patterns within the data. We also note that the top MP-snippets pattern corresponds to the walking activity (highlighted blue). Although the goal of MP-snippets is to find frequent patterns, this top pattern is not the most frequent for this dataset. MP-snippets overlooks the more-frequent jogging pattern because of the algorithm's use of a fixed Euclidean distance measure. In contrast to MP-snippets, T2P does not require this information to discover all embedded patterns, as shown in Figure 5c. Here, the user receives a summary of both patterns. Furthermore, T2P will highlight the variations of the pattern that exist in the data, unlike MP-snippets.

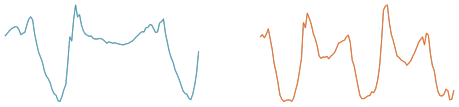
We posited that the information-theoretic principle of minimizing description length offers a guide to discovering the most informative patterns within time series data. In these experiments, we compute the description length that results from compressing the data using the discovered pattern and compare these results for alternative methods. In this case, each occurrence of the pattern is replaced by a single "placeholder" bit, and the description length of the pattern definition is included in the total. Another advantage is T2P compared to the traditional method is the ability to show different variations of the same pattern, while MP-Snippets show only a subsequence from the data as the best representation for each pattern. Both T2P and MP-snippets have the same description length when the patterns are used to compress the data. While the original description length is 10,000, the compressed data using either MP-Snippets or T2P have a description length of 300 (compression ratio = 0.03).

5.3 UMD SYNTHETIC DATA

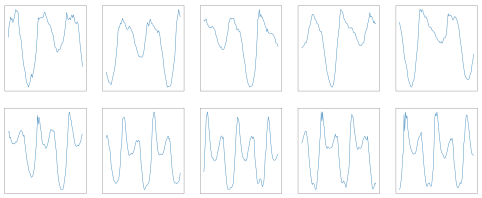
The UMD synthetic dataset contains three embedded time series patterns: one pattern is characterized by an up-bell shape arising at the initial or final period (Up); one does not have any bell (Middle); and one exhibits a down-bell shape arising at the initial or final period (Down), as shown in Figure 12. When MP-snippets is used to discover the top-3 patterns, as shown in Figure 6b, one of the patterns is presented twice. This is a result of employing the fixed, Euclidean Distance measure.



(a) Y-axis acceleration from a hip-mounted accelerometer.

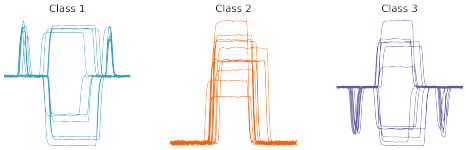


(b) The top two patterns discovered by MP-snippets.

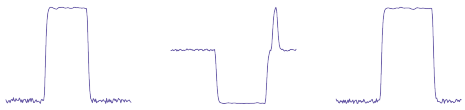


(c) Discovered patterns by T2P with a 2×5 decoding filter size.

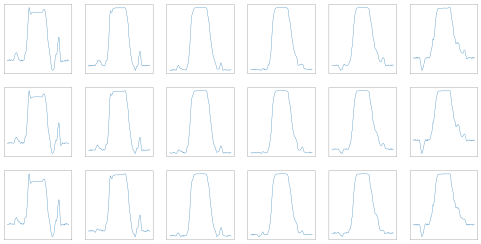
Figure 5: A sample of the data was collected from a participant performing a walking activity followed by a jogging activity.



(a) UMD synthetic data.



(b) The top three patterns discovered by MP-snippets.



(c) Discovered patterns by T2P with a 3×6 decoding filter size.

Figure 6: UMD synthetic data.

In contrast, T2P discovered a set of diverse patterns, as shown in Figure 6c. Additionally, the T2P pattern resulted in a smaller description length than MP-snippets. The description length (DL) of the original dataset is 5,400. The compressed DL of MP-snippets-compressed data is 3,615 (compression ratio of 0.67), while the DL of the T2P-compressed data is 2,274 (compression ratio of 0.42). This analysis provides evidence that T2P offers greater compression of the original data and thus, following information theory, offers more informative and comprehensive patterns. See Appendix C for additional analyses and comparison of discovery results.

5.4 Scalability

Finally, we perform a scalability comparison between MP-snippets and T2P. Figure 7 plots the run time of MP-snippets and T2P as a function of each dataset size. As illustrated in the graph, the run time of MP-snippets increases superlinearly. In contrast, the run time of T2P remains close to constant despite the increasing size of the time series.

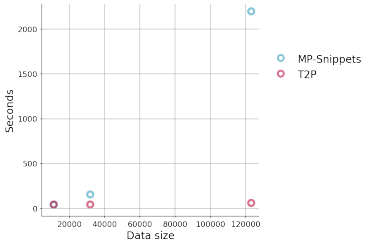


Figure 7: The run time comparison of T2P and MP-Snippets in seconds.

6 Conclusion

Motivated by information theory, T2P aims to discover informative patterns in time series data that compress the data. T2P differs from existing methods because the algorithm does not rely on a fixed similarity function or an exhaustive search radius. T2P learns a set of prototypes from training data, where each pattern is represented by a set of points in the feature space. The extracted patterns provide an intuitive way to summarize the data.

In our experiments, we observe that the ability of T2P to identify patterns is equivalent or better than a state-of-the-art method that requires an exhaustive search. Moreover, T2P identifies the pattern variations that occur in the original time series. We anticipate that T2P will also be effective for pattern discovery in multivariate time series data by analyzing each dimension separately.

ACKNOWLEDGMENTS

Acknowledgements will be added in the final version of the paper.

REFERENCES

Alessandro Achille and Stefano Soatto. Information dropout: Learning optimal representations through noisy computation. *IEEE transactions on pattern analysis and machine intelligence*, 40 (12):2897–2905, 2018.

Alessandro Achille, Tom Eccles, Loic Matthey, Chris Burgess, Nicholas Watters, Alexander Lerchner, and Irina Higgins. Life-long disentangled representation learning with cross-domain latent homologies. *Advances in Neural Information Processing Systems*, 31, 2018.

Sara Alaee, Ryan Mercer, Kaveh Kamgar, and Eamonn Keogh. Time series motifs discovery under dtw allows more robust discovery of conserved structure. *Data Mining and Knowledge Discovery*, 35(3):863–910, 2021.

Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. Deep variational information bottleneck. *arXiv preprint arXiv:1612.00410*, 2016.

Mohammed Ali, Mark W Jones, Xianghua Xie, and Mark Williams. Timecluster: dimension reduction applied to temporal data for visual analytics. *The Visual Computer*, 35(6):1013–1026, 2019.

Pierre Comon. Independent component analysis, a new concept? *Signal processing*, 36(3):287–314, 1994.

Hoang Anh Dau and Eamonn Keogh. Matrix profile v: A generic technique to incorporate domain knowledge into motif discovery. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 125–134, 2017.

Hoang Anh Dau, Eamonn Keogh, Kaveh Kamgar, Chin-Chia Michael Yeh, Yan Zhu, Shaghayegh Gharghabi, Chotirat Ann Ratanamahatana, Yanping, Bing Hu, Nurjahan Begum, Anthony Bagnall, Abdullah Mueen, Gustavo Batista, and Hexagon-ML. The ucr time series classification archive, October 2018. https://www.cs.ucr.edu/~eamonn/time_series_data_2018/.

Hui Ding, Goce Trajcevski, Peter Scheuermann, Xiaoyue Wang, and Eamonn Keogh. Querying and mining of time series data: experimental comparison of representations and distance measures. *Proceedings of the VLDB Endowment*, 1(2):1542–1552, 2008.

Emilien Dupont. Learning disentangled joint continuous and discrete representations. *Advances in Neural Information Processing Systems*, 31, 2018.

Xiaojie Guo, Yuanqi Du, and Liang Zhao. Property controllable variational autoencoder via invertible mutual dependence. In *9th International Conference on Learning Representations, ICLR*, 2021.

Donald Olding Hebb. *The organization of behavior: A neuropsychological theory*. Psychology Press, 2005.

Irina Higgins, Loïc Matthey, Arka Pal, Christopher P. Burgess, Xavier Glorot, Matthew M. Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *5th International Conference on Learning Representations, ICLR*, 2017.

Antti Honkela and Harri Valpola. Variational learning and bits-back coding: an information-theoretic view to bayesian learning. *IEEE transactions on Neural Networks*, 15(4):800–810, 2004.

Shima Imani, Frank Madrid, Wei Ding, Scott Crouter, and Eamonn Keogh. Matrix profile xiii: Time series snippets: a new primitive for time series data mining. In *2018 IEEE international conference on big knowledge (ICBK)*, pp. 382–389. IEEE, 2018.

Eamonn Keogh, Jessica Lin, and Ada Fu. Hot sax: Efficiently finding the most unusual time series subsequence. In *Fifth IEEE International Conference on Data Mining (ICDM'05)*, pp. 8–pp. Ieee, 2005.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations, ICLR*, 2014.

Elke Kirschbaum, Manuel Haußmann, Steffen Wolf, Hannah Sonntag, Justus Schneider, Shehabeldin Elzoheiry, Oliver Kann, Daniel Durstewitz, and Fred A Hamprecht. Lemonade: Learned motif and neuronal assembly detection in calcium imaging videos. In *7th International Conference on Learning Representations, ICLR*, 2019.

Gilles Laurent. Dynamical representation of odors by oscillating and evolving neural assemblies. *Trends in neurosciences*, 19(11):489–496, 1996.

Gabriel Loaiza-Ganem and John P Cunningham. The continuous bernoulli: fixing a pervasive error in variational autoencoders. *Advances in Neural Information Processing Systems*, 32, 2019.

Chris J Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. *arXiv preprint arXiv:1611.00712*, 2016.

Miguel AL Nicolelis, Luiz A Baccala, Rick CS Lin, and John K Chapin. Sensorimotor encoding by synchronous neural ensemble activity at multiple levels of the somatosensory system. *Science*, 268(5215):1353–1358, 1995.

Fabian Kai-Dietrich Noering, Yannik Schroeder, Konstantin Jonas, and Frank Klawonn. Pattern discovery in time series using autoencoder in comparison to nonlearning approaches. *Integrated Computer-Aided Engineering*, 28(Preprint):1–20, 2021.

Mahta Ramezanian Panahi, Germán Abrevaya, Jean-Christophe Gagnon-Audet, Vikram Voleti, Irina Rish, and Guillaume Dumas. Generative models of brain dynamics–a review. *arXiv preprint arXiv:2112.12147*, 2021.

Sven Peter, Elke Kirschbaum, Martin Both, Lee Campbell, Brandon Harvey, Conor Heins, Daniel Durstewitz, Ferran Diego, and Fred A Hamprecht. Sparse convolutional coding for neuronal assembly detection. *Advances in Neural Information Processing Systems*, 30, 2017.

Friedemann Pulvermüller, Rosario Tomasello, Malte R Henningsen-Schomers, and Thomas Wennekers. Biological constraints on neural network models of cognitive function. *Nature Reviews Neuroscience*, 22(8):488–502, 2021.

Ali Lotfi Rezaabad and Sriram Vishwanath. Learning representations by maximizing mutual information in variational autoencoders. In *2020 IEEE International Symposium on Information Theory (ISIT)*, pp. 2729–2734. IEEE, 2020.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *International conference on machine learning*, pp. 1278–1286. PMLR, 2014.

Jorma Rissanen. Modeling by shortest data description. *Automatica*, 14(5):465–471, 1978.

Rachit Singh, Jeffrey Ling, and Finale Doshi-Velez. Structured variational autoencoders for the beta-bernoulli process. In *NIPS 2017 Workshop on Advances in Approximate Bayesian Inference*, 2017.

Chris S Wallace and David M Boulton. An information measure for classification. *The Computer Journal*, 11(2):185–194, 1968.

Yan Zhu, Zachary Zimmerman, Nader Shakibay Senobari, Chin-Chia Michael Yeh, Gareth Funning, Abdullah Mueen, Philip Brisk, and Eamonn Keogh. Matrix profile ii: Exploiting a novel algorithm and gpus to break the one hundred million barrier for time series motifs and joins. In *2016 IEEE 16th international conference on data mining (ICDM)*, pp. 739–748. IEEE, 2016.

A VARIATIONAL AUTOENCODER

The variational autoencoder (VAE) Kingma & Welling (2014); Rezende et al. (2014) is a deep generative model consisting of a true posterior (generator) $p_\theta(x|z)$, a prior $p_\theta(z)$, and an approximate posterior (inference) $q_\phi(z|x)$. Let us consider a dataset $x = \{x^{(i)}\}_{i=1}^N$ consisting of N i.i.d samples of some continues or discrete variable x. We would like to compute the true posterior $p_\theta(z|x) = \frac{p_\theta(x,z)}{p_\theta(x)}$ but computing the $p_\theta(x) = \int p_\theta(x|z)p(z)dz$ is intractable in many cases. To overcome this obstacle, we can approximate $p_\theta(z|x)$ by selecting a tractable distribution $q(z)$ such as Gaussian. In order to approximate $p_\theta(z|x)$, we minimize the KL-divergence between the approximate and true posterior $D_{\mathrm{KL}}(q_\phi(z|x)||p_\theta(z|x))$. Thus the marginal likelihoods can be written as:

$$\log p_\theta(x^{(i)}) = D_{\mathrm{KL}}(q_\phi(z|x^{(i)})||p_\theta(z|x^{(i)})) + \mathcal{L}(\theta, \phi; x^{(i)}) \tag{7}$$

Because the KL divergence is non-negative, the second RHS term $\mathcal{L}(\theta, \phi; x^{(i)})$ is called the (variational) lower bound on the marginal likelihood of data point i. Instead of minimizing the KL divergence between the approximate and true posterior, we can maximize the variational lower bound.

$$\log p_\theta(x) \geq \mathcal{L}(\theta, \phi; x) = \mathbb{E}_{q_\phi(z|x)}\big[-\log q_\phi(z|x) + \log p_\theta(x, z)\big] \tag{8}$$

which can also be written as:

$$\mathcal{L}(\theta, \phi; x) = -D_{\mathrm{KL}}(q_\phi(z|x)||p_\theta(z)) + \mathbb{E}_{q_\phi(z|x)}\big[\log p_\theta(x|z)\big] \tag{9}$$

We need to compute the gradient of the lower bound $\mathcal{L}(\theta, \phi; x)$ w.r.t the generative parameter θ and the inference parameter ϕ, in order to optimize.

$$\nabla_{\phi,\theta}\mathcal{L}(\theta, \phi; x) = \nabla_{\phi,\theta}\mathbb{E}_{q_\phi(z|x)}\big[\log p_\theta(x|z)\big] - \nabla_{\phi,\theta}D_{\mathrm{KL}}(q_\phi(z|x)||p_\theta(z)) \tag{10}$$

For the first part of the lower bound, the gradient w.r.t θ can be easily computed using Monte Carlo sampling

$$\nabla_\theta\mathbb{E}_{q_\phi(z|x)}\big[\log p_\theta(x|z)\big] = \mathbb{E}_{q_\phi(z|x)}\big[\nabla_\theta \log p_\theta(x|z)\big] \approx \frac{1}{S}\sum_{s=1}^{S}\nabla_\theta \log p_\theta(x|z^s) \tag{11}$$

But the gradient w.r.t ϕ can not be sampled that easily

$$\nabla_\phi\mathbb{E}_{q_\phi(z|x)}\big[\log p_\theta(x|z)\big] = \nabla_\phi \int q_\phi(z|x)\log p_\theta(x|z)dz = \int \log p_\theta(x|z)\nabla_\phi q_\phi(z|x)dz. \tag{12}$$

To overcome this problem, we use reparameterization trick. The function $g_\phi(.)$ is chosen such that it maps a data point $x^{(i)}$ and a random noise vector $\epsilon^{(l)}$ to a sample from the approximate posterior for that datapoint: $z^{(i,l)} = g_\phi(\epsilon^{(l)}, x^{(i)})$ where $z^{(i,l)} \sim q_\phi(z|x^{(i)})$. A vector of latent variables z in a high-dimensional space Z sample according to some probability density function (PDF) $P(z)$ defined over Z. Thus, the reparameterized lower bound $\tilde{\mathcal{L}}(p, q; x) \approx \mathcal{L}(p, q; x)$ can be written as

$$\tilde{\mathcal{L}}(p, q; x) = \frac{1}{S}\sum_{s=1}^{S}\log p_\theta(x|z^s) - D_{\mathrm{KL}}(q_\phi(z|x)||p_\theta(z)) \tag{13}$$

The first term of RHS is a negative reconstruction error and KL-divergence act as a regularizer in auto-encoder parlance.

B T2P NETWORK ARCHITECTURE AND IMPLEMENTATION DETAILS

The details of T2P networks are shown in Algorithm 1, Figure 8, and Table 1

Input: raw time series x, architecture f_θ, α_ϕ, hyperparameter $\lambda_1, \lambda_2, a, \beta_{KL}, \beta_{MSE}$.
`/* trained` f_θ, α_ϕ `*/`
$\theta, \phi \leftarrow$ Initialize network parameters;
`/* Sample subset of frames from time series` T `*/`
while *convergence of* θ, ϕ **do**
  $x \leftarrow$ Randomly chosen sequence of consecutive frames from X `/* Encoding step */`
  Encode x to get α Compute z following Eq. 2 `/* Decoding step */`
  $x \leftarrow$ decode via $f_\theta(z)$ `/* Update Parameters */`
  Compute the gradient of loss
end

Algorithm 1: T2P algorithm

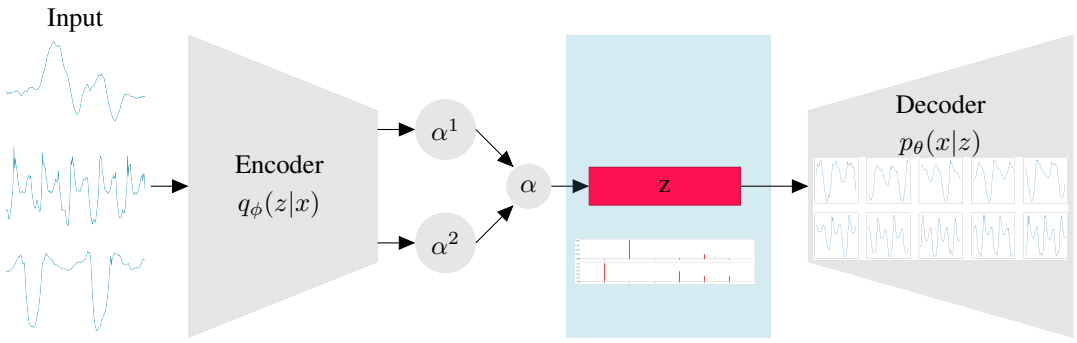


Figure 8: Schematic T2P framework. In this example, three frames of time series feed to T2P. The encoder network contains several layers of 1D convolution followed by a 3D convolution layer where $\alpha = \alpha^1/\alpha^2$ is the encoded space. The latent space z is sampled utilizing the reparametrization trick. The discrete z space illustrates what filters have been utilized to reconstruct the input. The decoder is a 3D convolution layer that employs to recreate the input. The learned filters for the reconstruction of the input are shown inside the decoder.

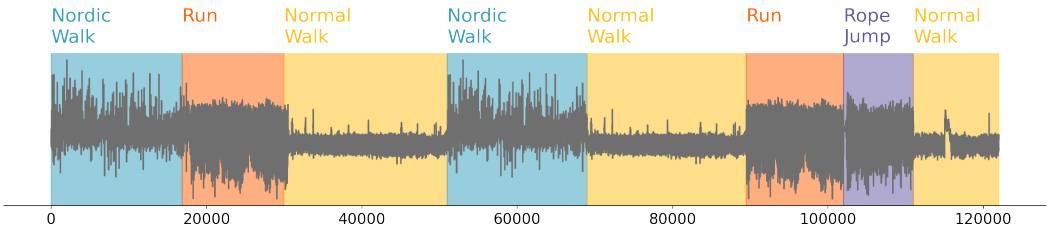
Table 1: T2P network architecture details.

Operation	Kernel	Feature map	Padding	Stride	Nonlinearity
Input: F time series sequence of length L					
1D Convolution	3	24	0	1	ELU
1D Convolution	3	48	0	1	ELU
1D Max-Pooling	2	-	0	2	-
1D Convolution	3	72	0	1	ELU
1D Convolution	3	96	0	1	ELU
1D Max-Pooling	2	-	0	2	-
1D Convolution	3	120	0	1	ELU
1D Convolution	3	48	0	1	ELU
Output: F time series sequence of length $\tilde{L} = (((L-3+1-3+1)/2-3+1-3+1)/2-3+1-1+1)$					
Input: F time series sequence of length $\tilde{L}$					
3D Convolution	$latentdim_2 \times \tilde{L}$	$2 \times latentdim_1$	$(latentdim_2 - 1) \times 0 \times 0$	1	SoftPlus
Output: $2 \times latentdim_1, (F + latentdim_2 - 1) \times 1 \times 1$					
Input: $2 \times latentdim_1, (F + latentdim_2 - 1) \times 1 \times 1$					
Reparametrization	-	-	-	-	-
Output: M activations, $(F + latentdim_2 - 1) \times 1 \times 1$					
Input: M activations, $(F + latentdim_2 - 1) \times 1 \times 1$					
3D Transposed Convolution	$latentdim_2 \times L$	M	$(latentdim_2 - 1) \times 0 \times 0$	1	-
Output: F time series sequence of length L					

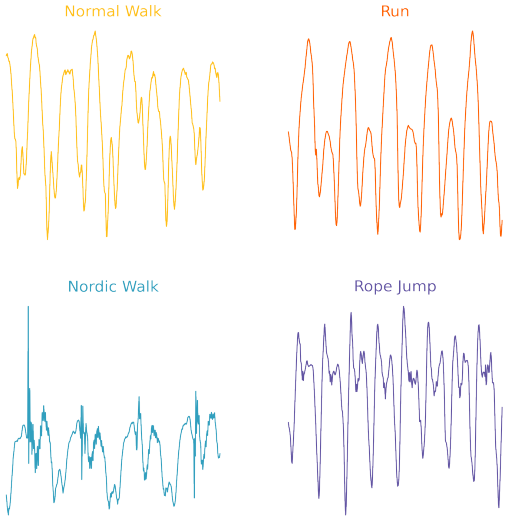
C Experiments and Results

C.1 Human Activity

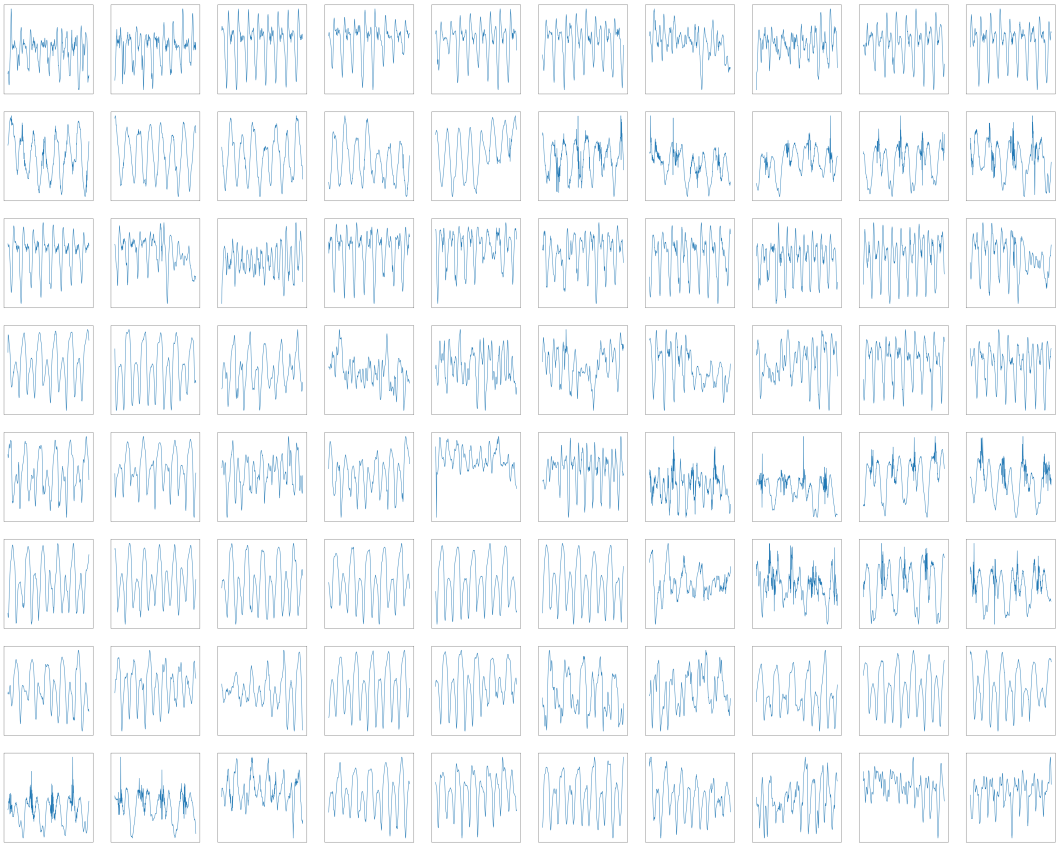
We repeat the experiment with similar data from the PAMAP dataset, a widely used benchmark. We created a larger time series by concatenating several sequences collected from the same person, as shown in Figure 10a. The top-4 patterns discovered by MP-snippets ranked by normal walk, run, nordic walk, and rope jump, as shown in Figure 10b. To learn all existing patterns from this data, we need to have a greater number of decoding filters, as shown in Figure 5c. Because there are several patterns similar to each other (normal walk, run, and rope jump), T2P is able to recreate most of the data using filters by learning only these patterns. However, T2P is able to learn nordic patterns by increasing the number of filters, as shown in Figure 10c. We include more experiments to illustrate the effect of the number of decoding filters in the appendix.



(a) Y-axis acceleration from a hip-mounted accelerometer.



(b) The top four patterns discovered by MP-snippets.

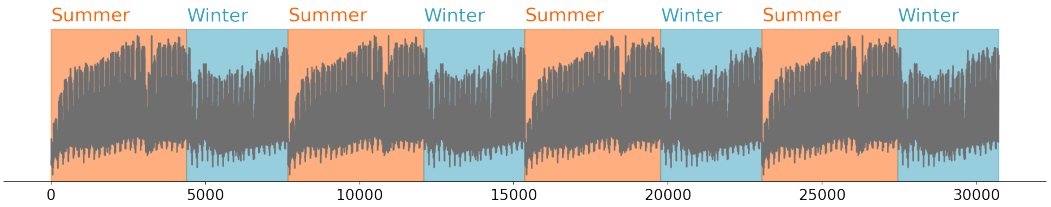


(c) Discovered patterns by T2P.

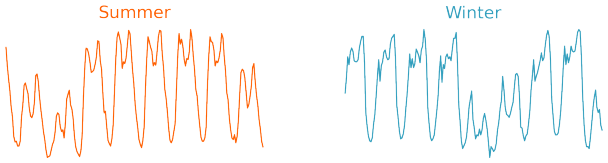
Figure 9: Samples of the data were collected from a participant doing nordic walk followed by run, normal walk, nordic walk, normal walk, run, rope jumping, and normal walk.

C.2 Electrical Power Demand Analysis

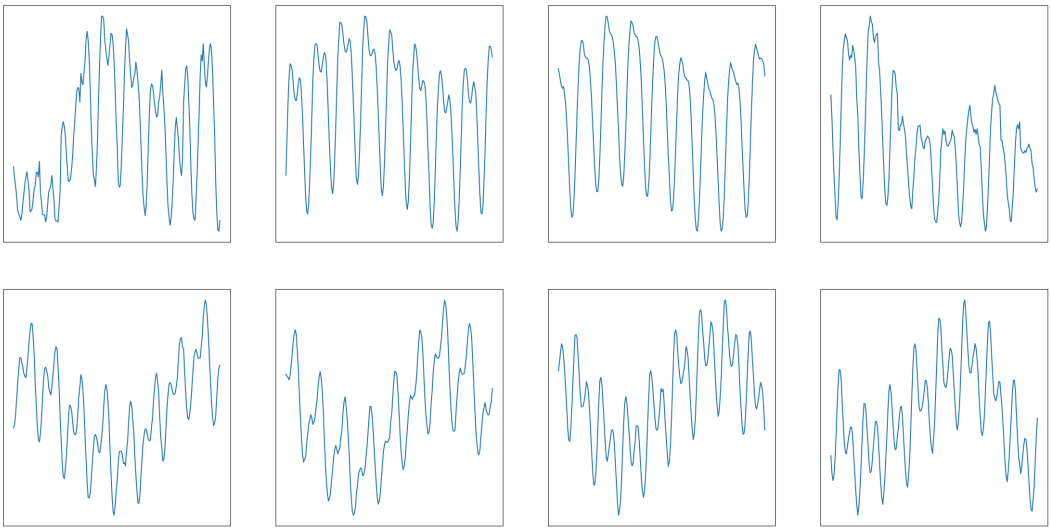
To demonstrate that T2P is domain agnostic, we evaluate the versatility of T2P on the Italian power demand dataset, as shown in Figure. The data represent a small Italian city's electrical power demand during summer and winter. The electrical demand is higher in the summer because more people use their HVAC system than in the winter. We create this data set by repeating the data four ti,e to create a larger dataset. MP-snippets rank summer and then winter patterns as illustrated in Figure. T2P also is able to find similar patterns, as shown in Figure. Further, the patterns' learned by T2P show the curve exists in each pattern as repeated in the data more frequently.



(a) The power demand dataset corresponds to summer/winter.



(b) The top two patterns discovered by MP-snippets.

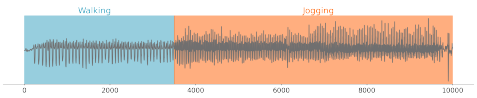


(c) Discovered patterns by T2P.

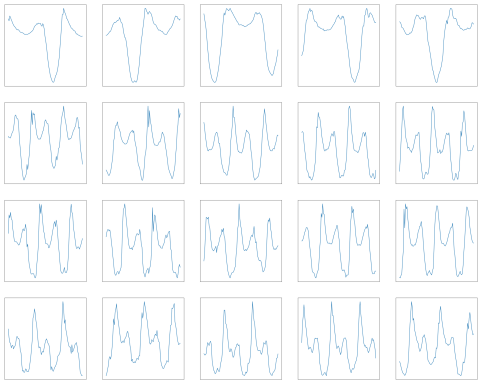
Figure 10: The Italian power dataset.

C.3 Extra Human Activity Analysis and UMD Synthetic Data

In our experiments, we tried larger latent space to observe the effect. Increasing the number of latent space result in learning repetitive patterns.

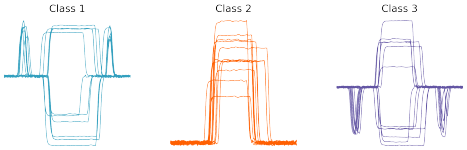


(a) Y-axis acceleration from a hip-mounted accelerometer.

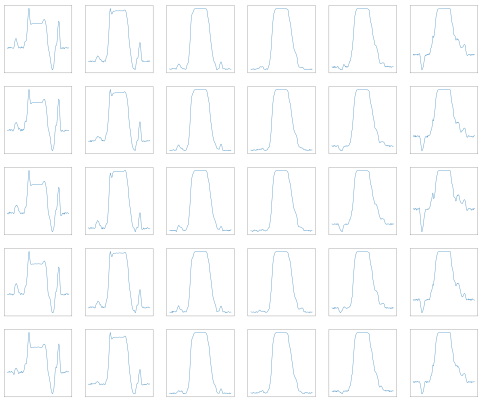


(b) Discovered patterns by T2P with 4 by 5 decoding filter size.

Figure 11: A small sample of the data was collected from a participant doing walking followed by jogging.



(a) UMD synthetic data.



(b) Discovered patterns by T2P with 5 by 6 decoding filter size.

Figure 12: UMD synthetic data.