
ON THE AUTOMATIC PARAMETER SELECTION FOR PERMUTATION ENTROPY

A PREPRINT

Audun D. Myers

Department of Mechanical Engineering
Michigan State University
East Lansing, MI
myersau3@msu.edu

Firas A. Khasawneh

Department of Mechanical Engineering
Michigan State University
East Lansing, MI
khasawn3@egr.msu.edu

May 17, 2019

ABSTRACT

Permutation Entropy (PE) has been shown to be a useful tool for time series analysis due to its low computational cost and noise robustness. This has drawn for its successful application in many fields. Some of these include damage detection, disease forecasting, detection of dynamical changes, and financial volatility analysis. However, to successfully use PE, an accurate selection of two parameters is needed: the permutation dimension n and embedding delay τ . These parameters are often suggested by experts based on a heuristic or by a trial and error approach. Unfortunately, both of these methods can be time-consuming and lead to inaccurate results. To help combat this issue, in this paper we investigate multiple schemes for automatically selecting these parameters with only the corresponding time series as the input. Specifically, we develop a frequency-domain approach based on the least median of squares and the Fourier spectrum, as well as extend two existing methods: Permutation Auto-Mutual Information Function (PAMI) and Multi-scale Permutation Entropy (MPE) for determining τ . We then compare our methods as well as current methods in the literature for obtaining both τ and n against expert-suggested values in published works. We show that the success of any method in automatically generating the correct PE parameters depends on the category of the studied system. Specifically, for the delay parameter τ , we show that our frequency approach provides accurate suggestions for periodic systems, nonlinear difference equations, and ECG/EEG data, while the mutual information function computed using adaptive partitions provides the most accurate results for chaotic differential equations. For the permutation dimension n , both False Nearest Neighbors and MPE provide accurate values for n for most of the systems with a value of $n = 5$ being suitable in most cases.

1 Introduction

Permutation Entropy (PE) has its origins in information entropy, which is a tool to quantify the uncertainty in an information-based system. Information entropy was first introduced by Shannon [38] in 1948 as Shannon Entropy. Specifically, Shannon entropy measures the uncertainty in future data given the probability distribution of the data types in the original, finite dataset. Shannon entropy is calculated as $H_s(n) = -\sum p(x_i) \log p(x_i)$, where x_i represents a data type, and $p(x_i)$ is the probability of that data type. In recent years information entropy has been heavily applied to the time series of dynamic systems. To better accommodate these applications, several new variations of information entropy have been proposed. Some of these include approximate entropy [32], sample entropy [35], and permutation entropy (PE) [2]. These methods measure the predictability of a sequence through the entropy of the relative data types. However, PE considers the ordinal position of the data (permutations), which have been shown to be very effective for analyzing the dynamic state of a time series [8, 26, 29]. PE is quantified in a similar fashion to Shannon entropy with only a change in the data type to permutations (see Fig. 2), which we symbolically represent as π_i . PE has two parameters: the permutation dimension n and embedding delay τ , which are used when selecting the permutation size and spacing, respectively. In practice, the results of PE have been shown to be highly dependent on these parameters [23, 36, 39]. Unfortunately, there is no single, accurate approach for selecting them. This introduces the motivation for this paper: investigate automatic methods for selecting both PE parameters. There are currently three main methods for selecting PE parameters: (1) use parameters suggested by experts for a specific application, (2) use trial and error to find suitable parameters, or (3) use methods developed for phase space reconstruction.

Many domain scientist who apply permutation entropy make general suggestions for n and τ [14, 44]. However, these suggested parameters can result in inaccurate results. As an example of this shortcoming, Popov et al. [33] applied PE to EEG data and showed that the sampling frequency influences the selection of τ . As for the dimension n , there are only general suggestions on how to choose its value. For example, it has been shown that the appropriate permutation dimension lies in the range $3 < n < 8$ for the vast majority of applications [36]. Additionally, Bandt and Pompe [2] suggest that $N \gg n$, where N is the length of the time series. However, these general outlines for the selection of n (and τ) do not allow for an application specific suggestion for n .

If we assume that suitable PE parameters correspond to optimal phase space reconstruction parameters, then a common approach for selecting τ and n is to implement one of the existing methods for estimating the optimal Takens' embedding [41] parameters. Based on this practice, some of the common methods for determining τ include the mutual information function approach [15], the first folding time of the autocorrelation function [5, 16], and phase space methods [7]. Additionally, some common phase space reconstruction methods for determining n include box-counting [4], correlation exponent method [16], and false nearest neighbors [18]. However, although the parameters in PE have similar names to their delay reconstruction counterpart, there are innate differences between ordinal patterns and phase space reconstruction which can lead to inaccurate n or τ values.

1.1 Our contribution

The problem we address in this paper is the following: Given a sufficiently sampled/oversampled, noisy time series $X = \{x_t\}_{\mathcal{R}^+}$, how can we reliably and systematically define appropriate dimension n and time delay τ values for computing the corresponding PE?

Our first contribution towards answering this question is detailed in Section 2 and it is related to the automatic selection of the time delay τ . Specifically, assuming that X is contaminated by Gaussian measurement noise, in Section 2.1 we combine the Least Median of Squares (LMS) approach for outliers detection with Fourier transforms theorem to derive a formula for the maximum significant frequency in the Fourier spectrum. This formula allows obtaining a cutoff value where the only input, besides the signal, is a desired percentile from the Probability Density Function (PDF) of the Fourier spectrum. Once this value is obtained, Nyquist's sampling theorem is used to compute an appropriate τ value.

The second contribution is through an approach that we develop in Section 2.2, which uses Multi-scale Permutation Entropy (MPE) for finding τ . We show how MPE can be used to find the main period of oscillation for a time series derived from a periodic system. Building upon this, we show how the method can be extended to find τ for a chaotic time series by using the first maxima in the MPE. Specifically, we suggest using the delay τ at the first maxima for PE calculations as it satisfies the Nyquist's sampling theorem.

Our third contribution to the automatic selection of τ is through the analysis of Permutation Auto-Mutual Information [24] (PAMI). PAMI is an existing method for measuring the mutual information of permutations. However, in this paper we tailor this method to specifically select τ for PE. We found that PAMI reaches a first minima at a specific window size that is approximately independent of n . We then use this window size to find a suitable τ based on the common selection

of $4 \leq n \leq 6$. We also conclude that finding the window size using $n = 2$ is the least computationally expensive method when using PAMI.

Our final contribution towards answering the posited question is our evaluation of the ability of existing tools for computing an embedding dimension to provide an appropriate value for the PE parameter n . We compare n values computed from False Nearest Neighbors or (FNN—Section 3.1), Singular Spectrum Analysis (SSA—Section 3.2), and Multi-scale Permutation Entropy (MPE—Section 2.2). While we use existing methods for performing the FNN and the SSA analyses, for the MPE-based approach, we use a criteria established in prior works [36], which requires finding τ first, for selecting n . However, for the latter, we further automate the selection process by using the procedure that we developed in Section 2.2 for finding τ first. The resulting τ and n values from all the methods are then listed in Tables 2 and 3 and compared to values suggested in the literature for a variety of examples.

This paper is organized as follows. In Section 1.2 we provide a simple PE calculation example to illustrate the main concepts of PE and to introduce the influence of the two parameters n and τ . We then go into detail on some existing methods for selecting both τ and n . Specifically, in Section 2 we provide a detailed explanation for selecting τ using existing, automatic methods such as autocorrelation in Section 2.3 and Mutual Information (MI) in Section 2.4. Additionally, we modified and develop/tailor methods to automatically select τ . These methods include a frequency approach in Section 2.1, Multi-scale Permutation Entropy (MPE) in Section 2.2, and PAMI in Section 2.5. In Section 3 we expand on the process for selecting n using False Nearest Neighbors (FNN) in Section 3.1 and Singular Spectrum Analysis in Section 3.2. In Section 3.3, we explain our algorithm for automatically selecting n using MPE. After introducing each method, in Section 4 we contrast all of these methods and make conclusions on their viability by comparing the resulting parameters to those suggested by PE experts.

1.2 Permutation Entropy Example

Bandt and Pompe [2] defined PE according to

$$H(n) = - \sum p(\pi_i) \log p(\pi_i), \tag{1}$$

where $p(\pi_i)$ is the probability of a permutation π_i and $H(n)$ is the permutation entropy for dimension n , which has units of bits when the logarithm is of base 2. The permutation entropy parameters τ and n are used when selecting the motif size, with τ as the time difference between two consecutive points in a uniformly sub-sampled time series and n as the permutation length or motif dimension. To form a permutation, begin with with an element of the x_i of the series X . Using this element, the dimension n , and delay τ , define the vector $v_i = [x_i, x_{i+\tau}, x_{i+2\tau}, \dots, x_{i+(n-1)\tau}]$. The corresponding permutation π_i of this vector is determined using its ordinal pattern. For example, consider the third degree $n = 3$ permutation shown in Fig. 1. The permutation type, which categorizes the permutation, is found by first ordering the n values of the permutation smallest to largest, and then accounting for the order received. For the given permutation in Fig. 1, the resulting permutation is categorized as the sequence $\pi_i = (1, 0, 2)$, which is one of $n!$ possible permutations for a dimension n , see Fig. 2.

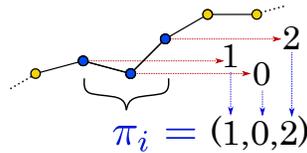


Figure 1: Sample permutation formation for $n = 3$ and $\tau = 1$.

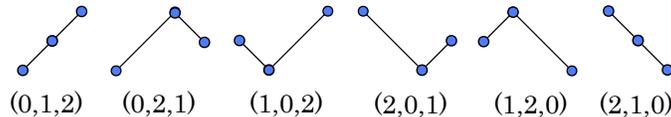


Figure 2: All possible permutation Configurations for $n = 3$.

To demonstrate how PE is calculated, consider another sequence $X = [4, 7, 9, 10, 6, 11, 3, 2]$ with PE parameters $n = 3$ and $\tau = 1$. The left side of Fig 3 shows how the sequence can be broken down into the following permutations: two $(0, 1, 2)$, one $(1, 0, 2)$, two $(1, 2, 0)$, and one $(2, 1, 0)$ for a total of 6 permutations. This makes each permutation type have a probability out of 6. The permutation distribution can be visually understood by illustrating the probabilities of each permutation as separate bins. To accomplish this, the right side of Fig. 3 shows the abundance of each permutation.

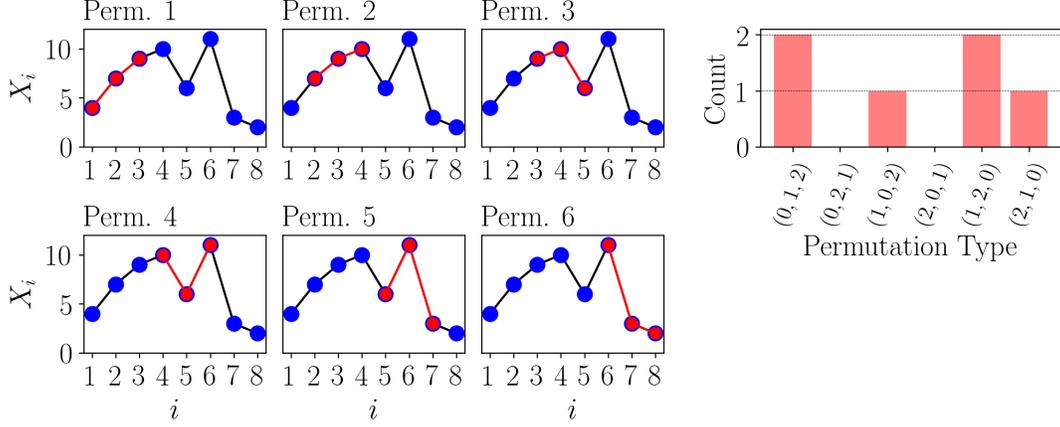


Figure 3: Permutations 1 through 6 shown for example sequence X (left) with $n = 3$ and $\tau = 1$ and the relative abundance of each permutation (right). Permutation 1 corresponds to a $(0, 1, 2)$, permutation 2 is of type $(0, 1, 2)$, permutation 3 is of type $(1, 2, 0)$, permutation 4 is of type $(1, 0, 2)$, permutation 5 is of type $(1, 2, 0)$, and permutation 6 is of type $(2, 1, 0)$.

Applying Eq. (1) to the probabilities of each permutation for our example sequence X yields

$$H(3) = -\frac{2}{5} \log \frac{2}{5} - \frac{2}{5} \log \frac{2}{5} - \frac{1}{5} \log \frac{1}{5} - \frac{1}{5} \log \frac{1}{5} = 1.918 \text{ bits.}$$

We can normalize PE using the possible PE value, which occurs when all $n!$ possible permutations are equiprobable according to $p(\pi_1) = p(\pi_2) = \dots = p(\pi_{n!}) = \frac{1}{n!}$. The resulting normalized PE is

$$h_n = -\frac{1}{\log_2 n!} \sum p(\pi_i) \log_2 p(\pi_i). \quad (2)$$

Continuing with our example series X , Eq. (2) gives a normalized PE $h_3 \approx 0.742$.

2 Embedding Delay Parameter Selection

The delay embedding parameter τ is used to uniformly subsample the original time series. To elaborate, consider the time series $X = \{x_i \mid i \in \mathbb{N}\}$. By applying the delay $\tau \in \mathbb{N}$, a new subsampled series is defined as $X(\tau) = [x_0, x_\tau, x_{2\tau}, \dots]$. In order to obtain a stable and automatic method for estimating an optimal value for τ we investigate: a novel frequency-based analysis that we describe in Section 2.1, Multi-scale Permutation Entropy (MPE) (Section 2.2), autocorrelation (Section 2.3), and Mutual Information function (MI) (Section 2.4). We recognize, but do not investigate, some other methods for finding τ such as diffusion maps [3] and phase space expansion [7].

2.1 Frequency Approach: Fourier Spectrum Noise Cutoff using 1-D Least Median of Squares

In this section, we develop a method for finding the noise floor in the Fourier spectrum using Least Median of Squares (LMS) [25]. We then use the noise floor to find the maximum significant frequency of a signal contaminated with additive Gaussian white noise (GWN). Our method is based on finding the maximum significant frequency in the Fourier spectrum and the Nyquist sampling frequency criteria. To motivate the development of this approach, we begin by working with the frequency criteria developed by Melosik and Marszalek [27], which agrees with Nyquist sampling theorem [20], for choosing a suitable sampling frequency f_s :

$$2f_{\max} < f_s < 4f_{\max}, \quad (3)$$

where f_{\max} is the maximum significant frequency in the signal. Melosik and Marszalek [27] showed that a sampling frequency within this range is appropriate for subsampling an oversampled signal, thus mitigating the effect of temporal correlations of neighboring points in densely sampled signals. They also showed that the time delay obtained using this criteria can also be used for delay reconstruction. However, the automatic identification of f_{\max} from an oversampled signal is not trivial. Melosik and Marszalek [27] selected a maximum significant frequency by inspecting the normalized Fourier spectrum and using a threshold cutoff of approximately 0.01 for a noise-free chaotic Lorenz system. This made *visually* finding the maximum frequency significantly easier but did not provide guidance on how to algorithmically

find f_{\max} . Further, Attempting to algorithmically adopt the approach in [27] resulted in large errors especially in the presence of a low signal to noise ratio. Therefore, this motivated the search for an automatic, data-driven approach for identifying the noise floor, which could then be used to find the maximum significant frequency. This led us to develop a method that is based on 1-D least median of squares applied to the Fourier spectrum, which we describe below. The assumptions inherent to our method are

1. The time series is not undersampled. The purpose of the methods is to determine a suitable delay parameter for subsampling the signal, which would be meaningless if the signal is undersampled.
2. The time series has less than 50% outliers. By outliers, it is meant that the Fourier transform of the signal needs to have less than 50% of the points as significant peaks. This requirement stems from the limitations of the least median of squares regression.
3. The noise in the signal is approximately Gaussian white noise; otherwise, the ensuing statistical analysis becomes inapplicable. Violating this assumption can yield false peak detections, which would lead to an incorrect delay parameter.

Using the noise floor determined from 1-D least median of squares, we find suitable cutoffs for obtaining f_{\max} of the signal and compute a suitable embedding delay according to

$$\tau = \frac{f_s}{\alpha f_{\max}}, \quad (4)$$

where we set $\alpha = 2$, thus agreeing with the range in Eq. (3) and the Nyquist sampling criterion.

Figure 4 summarizes the frequency approach for τ with the use of our 1-D LMS method for finding a noise floor in the Fourier spectrum. This process begins with computing the signal’s Fourier spectrum. We then fit an LMS regression line to noise in the Fourier spectrum to provide statistical information about the Probability Distribution Function (PDF) of the noise. Next, we use the PDF to determine the Cumulative Distribution Function (CDF), which can be used to determine a meaningful noise cutoff in the Fourier spectrum. However, if noise is present, it must be approximately GWN for this method to hold statistical significance. We then use this cutoff for separating the highest significant frequency in the Fourier spectrum f_{\max} , which is used to find a suitable embedding delay τ based on the frequency criteria in Eq. (4). In the following paragraphs we review our use of the LMS and the derivation of the PDF of the Fourier spectrum of GWN. We then show how to combine the LMS method with the resulting PDF expression to find a suitable noise floor cutoff and the corresponding maximum significant frequency.



Figure 4: Overview of our frequency domain approach for finding the maximum significant frequency f_{\max} using LMS for a signal contaminated with GWN.

Least Median of Squares: LMS is a robust regression technique used when up to 50% of the data is corrupted by outliers. The algorithm for LMS regression was initially published in 1984 by Massart et al. [25]. In comparison to the widely used least sum of squares (LS) algorithm, the LMS replaces the sum for the median, which makes LMS resilient to outliers. The difference between LS and LMS can be described by

$$\begin{aligned} LS &: \min \sum r_i^2, \\ LMS &: \min \left(\text{median}_i(r_i^2) \right), \end{aligned} \quad (5)$$

where r is the residual. Figure 5 shows an example application of the linear LMS regression.

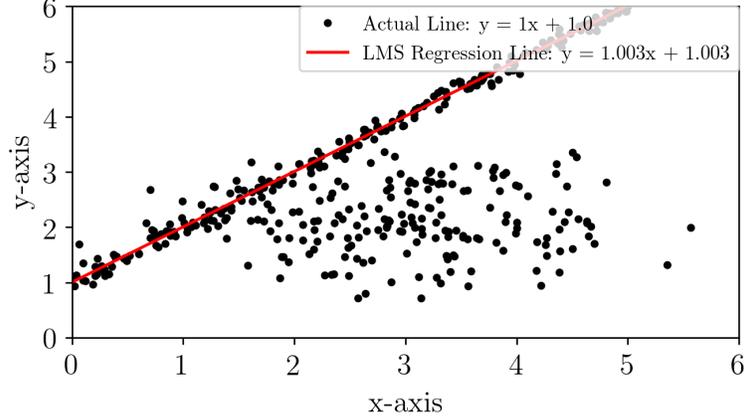


Figure 5: LMS linear regression with 45% outliers. Results match those found in [25].

Specifically, this figure shows 110 data points drawn from the line $y = x + 1$ with added GWN of zero mean and 0.1 standard deviation. The data is corrupted with 90 outliers centered around $(3, 2)$ with a normal distribution of 1.0 along x and 0.6 along y . Figure 5 shows that the linear regression results closely match the actual trend line with the fitted line being $y = 1.003x + 1.003$ in comparison to the actual $y = x + 1$. These results agree well with the ones found in [25].

PDF and CDF of the magnitude of the Fast Fourier Transform of GWN: This section reviews the probability distribution function (PDF) and cumulative density function (CDF) for the Fourier Transform (FT) of GWN. Additionally, this section derives the location of the theoretical maximum of the PDF. The distribution of the FT of GWN is described by [34]

$$P_{|X|}(|X|) = \frac{2|X|}{E_w \sigma_x^2} e^{-\frac{|X|^2}{E_w \sigma_x^2}}, \quad (6)$$

where $|X|$ is the magnitude of the FT of GWN, $P_{|X|}$ is the probability density function of $|X|$, σ_x is the standard deviation of the GWN, and E_w is the window energy or number of discrete transforms taken during the FT. By setting the first derivative of $P_{|X|}$ with respect to $|X|$ equal to zero, the theoretical maximum of the PDF is

$$|X|_{\max} = \sqrt{\frac{E_w \sigma_x^2}{2}}. \quad (7)$$

We calculate the CDF corresponding to the PDF described in Eq. (7) by combining the PDF in Eq. 6 with the CDF for a Rayleigh distribution in [30]. This results in

$$CP_{|X|}(|X|) = 1 - e^{-\frac{|X|^2}{E_w \sigma_x^2}}, \quad (8)$$

where $CP_{|X|}$ is the cumulative probability of $|X|$.

Finding the Noise Floor: Our approach for finding the noise floor combines LMS with Eqs. (6) and (7). Specifically, we utilize LMS to obtain a 1-D fit of the Fast Fourier Transform (FFT) of the signal, which results in an approximate value of $|X|_{\max}$, which is $|X|$ at the maximum of $P_{|X|}$. Using $|X|_{\max}$ from the LMS fit, we then find the standard deviation of the distribution σ_x from Eq. 7, which is used to find a cutoff based on a set cumulative probability in Eq. (8).

We begin by showing the accuracy of the LMS fit for finding $|X|_{\max}$. Our example uses GWN with a mean of zero and standard deviation of 0.035 with 1000 data points. Taking the FFT of the GWN (see Fig. 6A) results in the distribution shown in Fig. 6B. The distribution shows a 1-D LMS fit of 8.215 compared to the theoretical maximum of the PDF from Eq. 7 of 7.826, which is approximately 4.67% greater. This shows that the 1-D LMS fit accurately locates $|X|_{\max}$. Additionally, the theoretical shape of the PDF in Fig. 6B is shown to be very similar to the actual distribution.

Next, our approach utilizes Eq. (8) and σ_x derived from Eq. 7 for finding the cutoff value $|X|_{\text{cutoff}}$. The $|X|_{\text{cutoff}}$ for a desired cumulative probability CP is found by solving Eq. (8) for $|X|$ as

$$|X|_{\text{cutoff}} = \sqrt{-E_w \sigma_x^2 \ln(1 - CP)}. \quad (9)$$

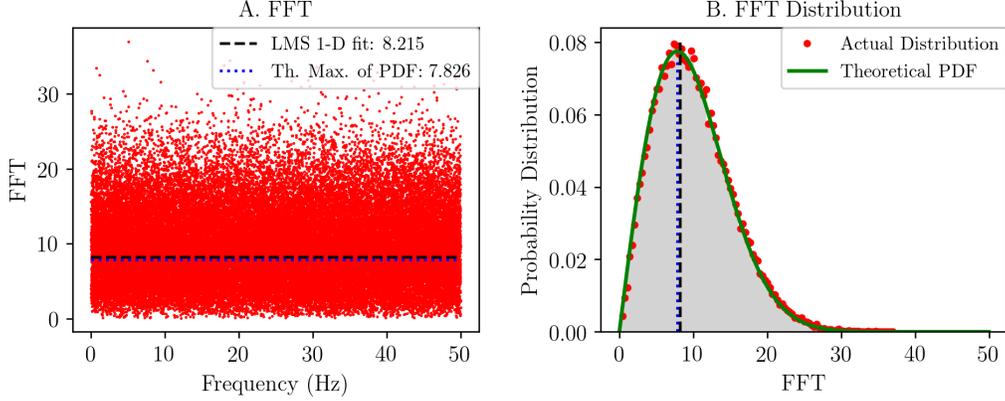


Figure 6: (A) FFT of GWN with 0.035 standard deviation and zero mean with the location of the theoretical maximum of the PDF and one-dimensional LMS regression value. (B) Distribution of GWN in the Fourier Spectrum with overlapped theoretical PDF and location of the theoretical maximum of the PDF and one-dimensional LMS regression value.

In order to make $|X|_{\text{cutoff}}$ robust to normalization and scaling of the FFT, we define the ratio C between the suggested cutoff from Eq. (9) and the maximum of the PDF from Eq. (7) as

$$C = \frac{|X|_{\text{cutoff}}}{|X|_{\text{max}}} = \sqrt{-2 \ln(1 - CP)}. \quad (10)$$

Example Cutoff: An example of how Eqs. (7) and (9) are used is shown in Fig. 7, where the maximum of the PDF and the cutoff for $CP = 99\%$ are marked in Fig. 7a and 7b, respectively. For this example, we find the ratio C to be approximately 3.03 for a 99% probability. In addition, we suggested a cutoff ratio $C = 6$ to be used for signals with less than 10^4 data points. This yields an expected probability of approximately $10^{-8}\%$ for a point in the FFT of the GWN attaining a magnitude greater than $|X|_{\text{cutoff}}$. Alternatively, Eq. (10) can be used to calculate a different value of C based on the desired probability and length of the signal.

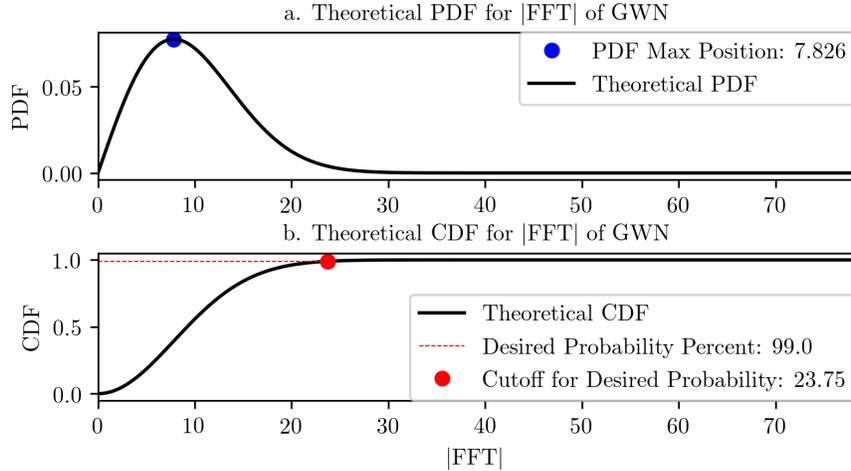


Figure 7: (a) Theoretical PDF for GWN. (b) CDF for GWN with an example cutoff at the 99% CP .

2.2 Multi-scale Permutation Entropy for Embedding Delay

In this section, we develop a method based on Multi-scale Permutation Entropy (MPE) to find the periodicity of a signal, which is then used to find a suitable delay parameter. MPE is a method of applying permutation entropy over a range of delays, which was first introduced by Costa et al. [10] for analyzing physiological time series. Applying MPE

results in a plot of permutation entropy versus delay, which reveals several useful trends. More specifically, Zunino et al. [45] showed how the first maxima in the MPE plot is developed when τ matches the characteristic time delay τ_r . Furthermore, the periodicity can be captured by the first dip in the MPE plot as shown in Fig. 8 at the location d_2 .

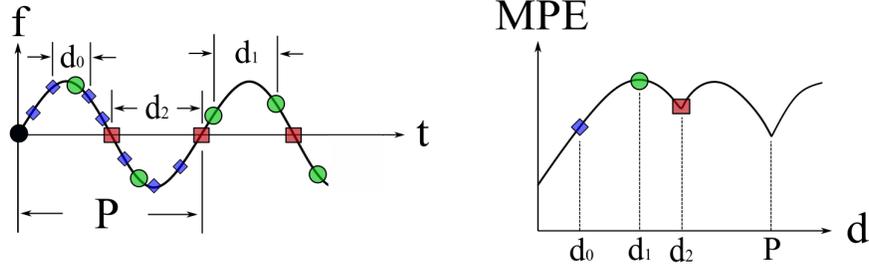


Figure 8: Resulting MPE plot for $2P$ periodic time series with example embedding vector sizes d_0 , d_1 , and d_2 .

Figure 8 shows embedding delays d_0 , d_1 , and d_2 calculated as $d = \frac{\tau}{f_s}$ as well as their corresponding locations on a normalized MPE plot. This MPE plot shows that the normalized MPE reaches its first maximum when the delay is roughly d_1 , which corresponds to an even distribution of motifs. A second observation, as mentioned previously, is that at d_2 (or the first dip in the MPE plot) there is a resonance or aliasing effect, which can be used to determine the period of the time series. This is based on the embedding delay size at d_2 causing the embedding vector size $V = d(n - 1)$ to be approximately half the of the periodicity P , which can be expressed as

$$d_2 = \frac{1}{2}P = \frac{1}{f_s}\tau_r = \frac{1}{2f}, \quad (11)$$

where τ_r is the delay that causes aliasing, P is the main period of oscillation, f is the main frequency of the time series corresponding to P , and f_s is the sampling frequency. The reason for the dip in the permutation entropy (PE) when the condition from Eq. (11) is met is caused from an aliasing effect, which induces more regularity in the motif distribution.

Calculating Embedding Delay from MPE We use the criteria of Melosik and Marszalek [27] to determine a suitable delay from the location of the first dip at d_2 . Their criteria states that the sampling frequency must fall within the range shown in Eq. (3). This range led to Eq. (4), which is used to calculate τ . However, for MPE, we substitute f_s and f_{\max} in Eq. (3) with $f_s = 2f\tau_r$ from Eq. (11) and $f_{\max} = f$. These substitutions allow Eq. (4) to reduce to

$$\tau = \frac{2}{\alpha}\tau_r, \quad (12)$$

where $\alpha \in [2, 4]$. These simplifications show that τ is only dependent on the delay which causes resonance τ_r when applying MPE.

In the following subsections we will further explain the method, trends, and limitations for using MPE to calculate τ . Specifically, Section 2.2.1 explains in detail the trends in the MPE plot we use, Section 2.2.2 shows the MPE plot for an example chaotic time series from the Lorenz system, Section 2.2.3 investigates the robustness of the method to noise, and Section A details our algorithm (Algorithm 1) for finding τ using MPE.

2.2.1 MPE Regions

Riedl et al. [36] showed that the MPE plot can be separated into three distinct regions as described below and shown in Fig. 9.

Region A Region A shows a gradual increase in the permutation entropy until reaching a maxima at the transition between regions A and B. Oversampling or a low value of τ causes the motif distribution corresponding to the permutation entropy to be heavily weighted on just increasing or decreasing motifs (motifs (0,1,2) and (2,1,0) for $n = 3$ from Fig. 2). This effect was coined as the ‘‘Redundancy Effect’’ by De Micco et al. [13], which means sufficiently low values of τ result in redundant motifs. However, as τ increases, the motif distribution becomes more equiprobable. Additionally, when the motif probability reaches a maximum equiprobability, the permutation entropy is at a maxima, which is the point of transitions from region A to B.

Region B Region B shows a slight dip to the first minima. This reduction in permutation entropy is caused by the aliasing or resonance from the value of d approaching half the main period length. At the transition from B to C, the resonance is reached, which provides information on the main frequency and period of the time series.

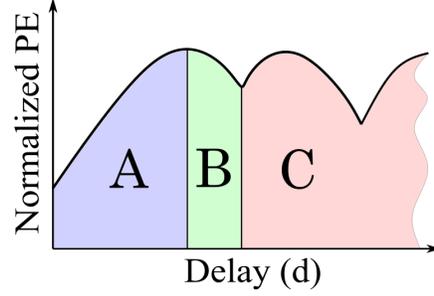


Figure 9: The three regions of the MPE plot for a periodic signal: (A) redundant, (B) resonant, and (C) irrelevant.

Region C Region C has possible additional minima and maxima from additional alignment of the embedding vector d with multiples of the main period. This region was referred to as the “Irrelevant Region” by De Micco et al. [13] due to effectively large values of τ forcing the delayed sampling frequency to fall below the Nyquist sampling rate as described by the lower bound in Eq. (3).

2.2.2 Example with Chaotic Time Series

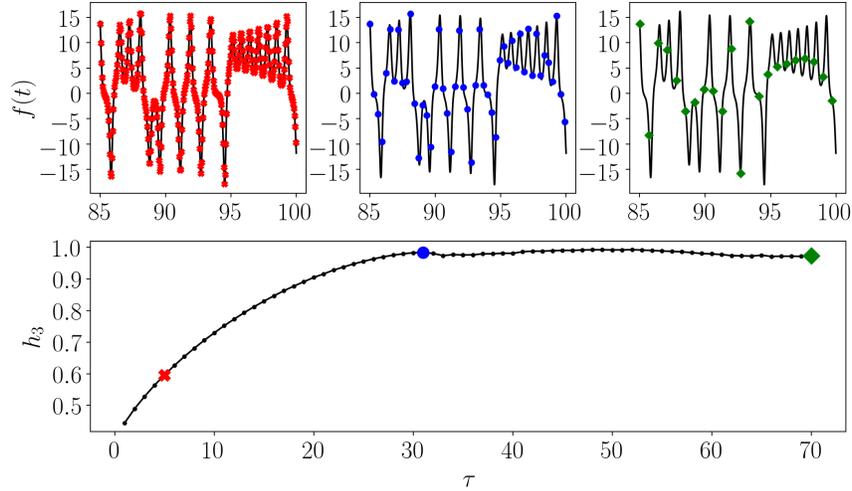


Figure 10: MPE plot for the x coordinate of the Lorenz system. Additionally, points in the MPE plot with their corresponding subsampled time series are shown for the redundant, resonant, and irrelevant regions as described in Section 2.2.1.

In Sections 2.2 and 2.2.1, we used a periodic time series to show and explain the regions developed in an MPE plot as well as an MPE-based method for determining a suitable embedding delay τ . In this section we further show the applicability of this approach to chaotic signals using the x -coordinate of the Lorenz System as an example. We simulate the Lorenz equations

$$\frac{dx}{dt} = \sigma(y - x), \quad \frac{dy}{dt} = x(\rho - z) - y, \quad \frac{dz}{dt} = xy - \beta z, \quad (13)$$

with a sampling rate of 100 Hz and using the parameters $\rho = 28.0$, $\sigma = 10.0$, and $\beta = 8.0/3.0$. This system was solved for 100 seconds and only the last 15 seconds from the time series are used. Figure 10 shows the result of applying MPE to the simulated Lorenz system.

Figure 10 shows similarities to Fig. 9 with a clear maxima at the boundary between regions A and B, albeit with no obvious minima. Therefore, a new distinct feature needs to be used to determine τ_r . We suggest using the first maxima to find τ because this delay is likely to fall within the region described by Eq. (12).

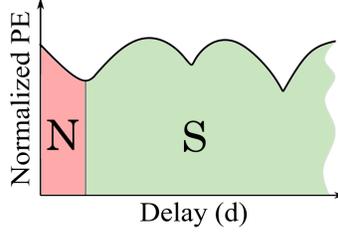


Figure 11: Region N is affected by noise in the MPE plot, and region S is unaffected.

2.2.3 Effects of Noise

We found that the main advantage of using MPE for determining the embedding delay is its robustness to noise. Noise on an MPE plot has minimal effects on regions B and C from Fig. 9, while only significantly affecting region A as shown in Fig 11. Furthermore, depending on the signal to noise ratio, there will only be an effect at the beginning of region A. Figure 11 shows the first region N where noise is affecting the permutation entropy. The effect of noise causes the MPE plot to start at a maxima and decrease to a local minima. When the time delay becomes large enough, the permutations are no longer influenced by the noise causing this minima. We found that the location of the minima is based on the condition

$$m_{\text{avg}}\tau_N \approx A_{\text{noise}}f_s, \quad (14)$$

where m_{avg} is the average of the absolute value of the slope and A_{noise} is approximately the maximum amplitude of the noise, τ_N is the value of τ great enough to surpass the noise amplitude. We derived this condition from the need for, on average, $|f(t) - f(t + \tau)| > A_{\text{noise}}$. This shows that MPE is robust to noise as long as the noise amplitude does not exceed the amplitude of the signal.

2.3 Autocorrelation

Autocorrelation uses the correlation coefficient between the time series and its τ -lagged version to find a suitable delay for phase space reconstruction. This method was first introduced by Box et al. [5] in 1970. Typically, the autocorrelation function is computed as a function of τ and, as a rule of thumb, a suitable delay τ is found when the correlation between $x(t)$ and $x(t + \tau)$ reaches the first folding time, i.e., when $\rho \leq 1/e$ [17]. We will overview the two prominent correlation techniques that can be used when implementing an autocorrelation-based approach for finding τ . These methods include Pearson Correlation in section 2.3.1 and Spearman's Correlation in section 2.3.2. After these two methods are introduced, we provide an example demonstrating the use of autocorrelation in section 2.3.3.

2.3.1 Pearson Correlation

The Pearson correlation coefficient $\rho_{xy} \in [-1, 1]$ measures the linear correlation of two time series x and y . Using these two data sets the correlation coefficient is calculated as

$$\rho_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}, \quad (15)$$

The possible values of ρ_{xy} represent the relationship between the two data sets, where $\rho_{xy} = 1$ represents a perfect positive linear correlation, $\rho_{xy} = 0$ represents no linear correlation, while $\rho_{xy} = -1$ represents a perfect negative linear correlation. However, Pearson correlation is limited because it only detects linear correlations. This limitation is somewhat alleviated by using Spearman's Correlation which operates on the ordinal ranking of the two time series instead of their numeric values.

2.3.2 Spearman's Correlation

Spearman's correlation is also calculated using Eq. (15) with the substitution of x and y for their ordinal ranking. This substitution allows for detecting nonlinear correlation trends to be represented as long as the correlation is monotonic. To demonstrate the difference, Fig. 12 shows two sequences x and y calculated from $y = x^4$ with $x \in [0, 10]$. Using this example, the Pearson correlation is calculated as $\rho \approx 0.86$, while Spearman's ranked correlation yields $\rho = 1.0$. This result demonstrates how Spearman's correlation coefficient accurately detects the non-linear, monotonic correlation between x and y whereas Pearson correlation may miss it.

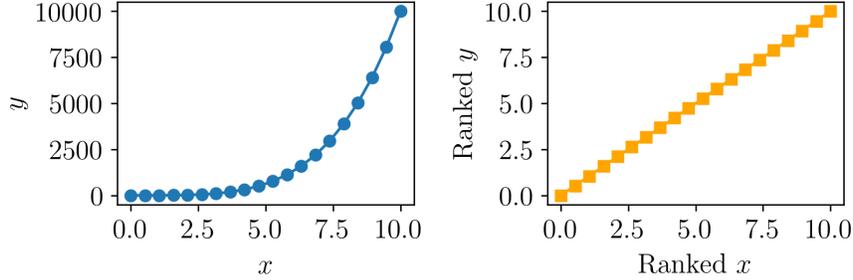


Figure 12: A comparison between (left) unranked values and (right) ranked values for calculating correlation coefficients. Using the ranked x and y , Spearman’s correlation coefficient can be used to accurately reveal existing nonlinear monotonic correlations.

2.3.3 Autocorrelation Example

We can use the concept of correlation to select a delay τ by calculating the correlation coefficient using Eq. (15) between a time series and its τ -lagged version. As an example, take the time series $x(t) = \sin(2\pi t)$, with $t \in [0, 5]$ having a sampling frequency of 100 Hz. This results in a suggested delay $\tau = 20$ at the first folding time using both Spearman’s and Pearson correlation. In section 4 we will implement Spearman’s version of autocorrelation to account for the possibility of non-linear correlations.

2.4 Mutual Information

Mutual information (MI) can be used to select the embedding delay τ based on a minimum in the joint probability between two sequences. The mutual information between two discrete sequences was first realized by Shannon et al. [37] as

$$I(X; Y) = \sum_{x \in X} \sum_{y \in Y} p(x, y) \log \frac{p(x, y)}{p(x)p(y)}, \quad (16)$$

where X and Y are the two sequences, $p(x)$ and $p(y)$ are the probability of the element x and y separately, and $p(x, y)$ is the joint probability of x and y . Fraser and Swinney [15] showed that for a chaotic time series the MI between the original sequence $x(t)$ and delayed version $x(t + \tau)$ will decrease as τ increases until reaching a first minimum. At this minima, the delay τ allows for the individual data points to share a minimum amount of information, which indicates sufficiently separated data points. While this delay value was specifically developed for phase space reconstruction, it is also commonly used for the selection of the PE parameter τ . We would like to point out that, in general, there is no guarantee that local minima exist in the mutual information, which is a serious limitation for computing τ using this method. All MI methods can be applied to either ranked or unranked data, and we will investigate the resulting delay values using both of these scenarios. As an overview, we plan to investigate four methods for estimating τ for PE using MI. These methods include MI with equal-sized partitions in Section 2.4.1, adaptive partitions in Section 2.4.2, and two permutation-based MI estimation methods in Section 2.4.3. After presenting each method, In Section 2.4.4 the optimal will be chosen by comparing the τ values from each MI method to those suggested by PE experts for various systems.

2.4.1 MI using Equal-sized Partitions

For the calculation of MI, the joint and independent probabilities of the original $x(t)$ and time lagged $x(t + \tau)$ time series are needed. However, since x is a discrete time series, we approximate these probabilities using bins, which segment the range of the series into discrete groups. The simplest method for approximating the probabilities using this discretization method is to use equal sized bins. However, the size of these bins is dependent on the number of bins k . We investigated various methods for estimating an appropriate number of bins using the length of the time series N . These methods include the common square-root choice $k = \lceil \sqrt{N} \rceil$, Sturge’s formula [40] $k = \lceil \log_2(N) \rceil + 1$, and Rice Rule [21] $k = \lceil 2N^{1/3} \rceil$. After comparing each method using a variety of examples, we found that the use of Sturge’s formula provided the best results for selecting τ for PE using MI.

2.4.2 MI using Adaptive Partitions

Darbellay and Vajda [12] introduced a multistep, adaptive partitioning scheme to select appropriate binning sizes in the observation space formed by the plane $x(t)$ and $x(t + \tau)$. Their method is often considered state-of-the-art for estimating

the mutual information function [19]. In this approach, the bins are recursively created where in the first function call, the space of the signal and its τ -lagged version is divided into an equal number of 2D bins. Then a A chi-squared test is used to test the null hypothesis that the data within the newly created bins are independent. Any segment that fails the test is further divided until the resulting sub-segments contain independent data (or a certain number of divisions is satisfied). Using this partitioning method, the MI is calculated using Eq. (16).

2.4.3 Kraskov MI

Kraskov et al. [19] developed a method for approximating the MI using entropy estimates using partition sizes based on k -nearest neighbors. Specifically, the method begins by first calculating the MI using entropy [11] as

$$I(X; Y) = H(X) + H(Y) - H(X, Y), \quad (17)$$

where H is the Shannon entropy. Next, an approximation of $H(X)$ with digamma functions is done, but the probability density of X and Y still needs to be estimated. To do this, adaptive partitions using the k -nearest neighbor are formed. Specifically Kraskov et al. develop two different partitioning methods with similar results. The first method uses the maximum Chebyshev distance to the $k = 1$ nearest neighbor j to form square bins as shown in Fig. 13-a, and the second method in Fig. 13-b uses rectangular partitions using the horizontal and vertical distances to the $k = 1$ nearest neighbor j . To continue with the example shown in Fig. 13, the density probability is estimated using the strips formed from

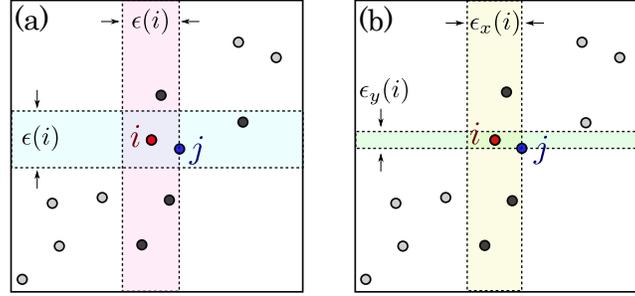


Figure 13: Example showing two different partition methods for Mutual Information estimation using an using $k = 1$ nearest neighbors adaptive partitioning.

these bins. To highlight the difference, Fig. 13-a shows a horizontal strip of width $\epsilon(i)$ encapsulating $n_x(i) = 2$ points (strip does not include the point i), while in Fig. 13-b only $n_x(i) = 1$ point is enclosed. Using these probability density approximations and the digamma function ψ , MI between X and Y can be estimated. Using the partitioning method shown in Fig. 13-a the MI is estimated as

$$I^{(1)}(X; Y) = \psi(k) - (\psi(n_x + 1) + \psi(n_y + 1) + \psi(N)). \quad (18)$$

Using the partitioning method shown in Fig. 13-b the MI is estimated as

$$I^{(2)}(X; Y) = \psi(k) - 1/k - [\psi(n_x) + \psi(n_y)] + \psi(N). \quad (19)$$

For the results shown in Section 4, we use the $k = 3$ nearest neighbor to generate the partitions.

2.4.4 MI Methods Comparison

To determine the optimal MI approximation method for selecting τ for PE, Table 1 shows a comparison between the τ values computed from each of the MI methods in Sections 2.4.1–2.4.3 and the corresponding values suggested by experts. The table shows that the adaptive partitioning method of Section 2.4.2 results in an accurate selection of τ for the majority of systems. Therefore, we will use the adaptive partitioning estimation method when making comparisons to other methods in Section 4.

2.5 Permutation Auto-mutual Information

As shown in Section 2.4, Mutual information (MI) is a useful method for selecting τ for phase space reconstruction. However, it does not account for permutations when selecting τ , which can lead to inaccuracies in computing the PE. To circumvent this issue, we develop a new method for selecting τ using Permutation Auto-Mutual Information (PAMI). PAMI was first introduced by Liang et al. [24] as a tool for detecting dynamic changes in brain activity. However, in this

Table 1: A comparison between the calculated and suggested values for the delay parameter τ for multiple MI approximation methods. The shaded cells highlight the methods that yielded the closest match to the suggested delay. The equal-sized partition method is described in Section 2.4.1, Kraskov et al. methods 1 and 2 in Section 2.4.3, and the adaptive partitioning approach in Section 2.4.2.

System	Mutual Information				Suggested Delay tau	Ref.
	Equal-sized Partitions	Kraskov et al. Method 1	Kraskov et al. Method 2	Adaptive Partitions		
White Noise	1	3	3	1	1	[36]
Lorenz	13	9	9	9	10	[36]
Rosler	14	13	11	9	9	[42]
Bi-directional Rosler	16	14	14	15	15	[36]
Mackey-Glass	7	8	7	7	1 to 700	[36]
Sine Wave	4	17	13	1	15	[42]
Logistic Map	5	8	11	5	1 to 5	[36]
Henon Map	12	15	13	8	1 to 5	[36]
ECG	22	16	9	8	1 to 4	[36]
EEG	6	5	5	5	1 to 3	[36]

paper we are tailoring its application for the first time to the selection of the permutation entropy parameter τ . PAMI is defined as

$$I_p(\tau, n) = H_{x(t,n)} + H_{x(t+\tau,n)} - H_{x(t,n),x(t+\tau,n)}, \quad (20)$$

where H is the permutation entropy described in Eq. (1). We suggest an optimal delay τ for a given dimension n when PAMI is at a minimum. This delay corresponds to minimum shared information between the time series and its time lagged version. By applying this method for the simple sinusoidal function described in Section A.2.3 we can form Fig. 14 with $n \in [2, 5]$ and $\tau \in [1, 50]$. As shown, the window size is approximately independent of the dimension n ,

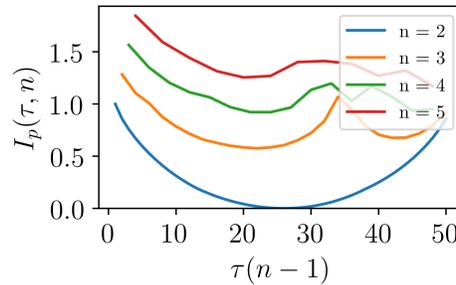


Figure 14: PAMI results for the sinusoidal function described in Section A.2.3 with $n \in [2, 5]$ and $\tau \in [1, 50]$. The figure shows an optimal window size $\tau(n-1) \approx 25$.

with an optimal window $\tau(n-1) \approx 25$ for the example. Through our analysis of the minimum PAMI as a function of the window size, we have developed a new method for selecting the optimal embedding window. However, we need the embedding dimension to suggest an optimal delay. To do this, we implement the common choice for n ranging from $4 \leq n \leq 6$ for PE. To reduce the computational demand, we suggest using permutation dimensions $n = 2$ to find an optimal window size. In addition to the reduced computational demand of using $n = 2$, we found that $I_p(n = 2) \approx 0$ at the first minima. This also helps making this first minima even more simple.

3 Motif Dimension

The second parameter for permutation entropy that needs to be automatically identified is the embedding dimension n . The methods for determining n fall into one of two categories: (1) independently determining n and τ , and (2) simultaneously determining n and τ based on the width of the embedding window. For the first category, we investigate using the method of False Nearest Neighbors (FNN) [18] in Section 3.1, and Singular Spectrum Analysis (SSA)[6] in Section 3.2. For the second category, we contribute to the selection of n by developing an automatic method using MPE from Section 3.3. This method combines the results for finding τ through MPE in Section 2.2 with the work

of Riedl et al. [36]. We recognize that our work does not include other commonly used methods for independently calculating n such as box-counting [9], largest Lyapunov exponent [43], and Kolmogorov–Sinai entropy [31]. However, our comparisons include many of the most common ones.

3.1 False Nearest Neighbors

False Nearest Neighbors (FNN) is one of the most commonly used methods for determining the minimum embedding dimension n for state space reconstruction. It was originally developed by Kennel et al. [18] as a geometric approach for determining the minimum needed embedding dimension. In this method the time series is repeatedly embedded into an a sequence of m -dimensional Euclidean spaces for a range of increasing values of m . The idea is that when the minimum embedding dimension m is reached or $m \geq n$, the distance between neighboring points does not significantly change as we keep increasing m . In other words, the Euclidean distance $d_m(i, j)$ between the point $\mathbf{P}_i \in \mathbb{R}^m$ and its nearest neighbor $\mathbf{P}_j \in \mathbb{R}^m$ minimally changes when the embedding dimension increases to $m + 1$. If the dimension m is not sufficiently high, then the points are false neighbors if their pairwise distance significantly increases when incrementing m . This ratio of change in the distance between nearest neighbors embedded in \mathbb{R}^m and \mathbb{R}^{m+1} is quantified using the ratio of false nearest neighbors

$$R_i = \sqrt{\frac{d_{m+1}^2(i, j) - d_m^2(i, j)}{d_m^2(i, j)}}. \quad (21)$$

R_i is compared to the tolerance threshold R_{tol} to distinguish false neighbors when $R_i > R_{\text{tol}}$. In this paper, we select $R_{\text{tol}} = 15$ as used by Kennel et al. [18]. By applying this threshold over all points, we can find the number of false neighbors as a percent FNN P_{FNN} . If there is no noise in the system, P_{FNN} should reach zero when a sufficient dimension is reached. However, with additive noise present, P_{FNN} may never reach zero. Thus, it is commonly suggested to use a percent FNN cutoff for finding a sufficient dimension n . We use the typically chosen cutoff $P_{\text{FNN}} < 10\%$, which is suitable for most applications when moderate noise is present.

3.2 Singular Spectrum Analysis

The singular spectrum analysis method was first introduced by Broomhead and King [6] as a tool to find the trends and most prominent periods of a time series. Leles et al. [22] summarized the SSA procedure as (1) immersion, (2) Singular Value Decomposition (SVD), (3) grouping, and (4) diagonal averaging. Specifically, immersion embeds the time series into a dimension L to form a Hankel matrix, SVD factors all the matrices, grouping combines the matrices that are similar in structure, and diagonal averaging reconstructs the time-series using the combined matrices. The needed embedding dimension is determined from the SVD by calculating the ratio D

$$D = \frac{g_L}{g_r} \quad (22)$$

of the sum of the L th diagonal entries g_L to the sum of the total diagonal entries g_r . When D exceeds 0.9, we consider the dimension to be high enough and set $n = L$, which can then be used as the embedding dimension for permutation entropy.

3.3 Multi-scale Permutation Entropy

Riedl et al. [36] showed how MPE can be used to determine an embedding dimension n . This method requires the embedding delay τ to be set to the length of the main period of the signal as shown in Section 2.2. The theory behind the method is based on normalizing the MPE according to

$$h'_n = \frac{-1}{n-1} H(n), \quad (23)$$

where h'_n is the PE normalized using the embedding dimension, and H_n is the PE calculated from Eq. (1). Riedl et al. [36] determine the embedding dimension by incrementing n to find the largest corresponding normalized PE h'_n with an embedding delay τ heuristically determined from the main period length. They concluded that the h'_n with the highest entropy accurately accounts for the needed complexity of the time series, and therefore suggests a suitable embedding dimension. Riedl et al. [36] show how this method provides an accurate embedding dimension for the Van-der-Pol-oscillator, Lorenz system, and the logistic map. However, the method is not automatic due to the reliance on a heuristically chosen τ .

To make the process automatic, we introduce algorithm 1 based on Section 2.2 to automatically select the correct τ , which we then use in conjunction with Eq. (23) to find n corresponding to the maximum h'_n . Additionally, we suggest scaling n from 3 to 8 as we have not yet found a system requiring $n > 8$ using this method.

4 Results and Discussion

To make conclusions about the described methods for determining τ and n , we made comparisons to values suggested by experts. The majority of the suggested parameters are taken from the work of Riedl et al. [36]. Tables 2 and 3 show the calculated and suggested values for τ and n , respectively.

Embedding Delay Table 2 highlights the automatically computed τ which match relatively well with the expert-identified value for a variety of systems. These systems fall within several categories, which include the following: noise, chaotic differential equations, periodic systems, nonlinear difference equations, and medical data. The methods presented in Table 2 include PAMI from Section 2.5, MI calculated using adaptive partitioning from Section 2.4.2, Spearman’s Autocorrelation from Section 2.3, MPE from Section 2.2, and the frequency approach from Section 2.1. For the noise category we only investigated GWN, and all the methods accurately suggest an embedding delay. For the second category of chaotic differential equations, Mutual Information approximated using adaptive partitions accurately provided suitable delay values. For the third category, periodic systems, we only investigated a simple sinusoidal function. This resulted in both MPE and the Frequency approach providing accurate suggestions. For difference equations we found that PAMI, autocorrelation, MPE, and the frequency approach provide accurate suggestions for the delay. Finally, when testing each method on medical data with intrinsic noise, we found that the noise-robust frequency approach yielded the optimal parameter selection for τ . As a generalization of the results found, we suggest the use of MI with adaptive partitioning when selecting τ for chaotic differential equations. However, for periodic systems, nonlinear difference equations, or ECG/EEG data we suggest the use of the frequency approach that we developed in this paper.

Embedding Dimension Table 3 shows parameters falling within the region specified by experts as highlighted in green. It can be seen that both MPE and FNN commonly had parameters within the range specified for all categories. However, SSA failed to provide a consistently suitable embedding dimension n . This leads to the conclusion that either MPE or FNN are sufficient methods for determining the embedding dimension for the majority of the considered applications. Although, these results also show that the dimension $n = 6$ works well for almost all applications.

Table 2: A comparison between the calculated and suggested values for the delay parameter τ . The shaded cells highlight the methods that yielded the closest match to the suggested delay. The following conditions or abbreviations were used in the table: the range under PAMI results is from using the range ($4 < n < 6$), AP under MI is an abbreviation for adaptive partitioning, and AC is an abbreviation for autocorrelation.

Category	System	Method					Suggested Delay τ	Ref.
		PAMI ($4 < n < 6$)	MI using AP	Spearman’s AC	MPE	Freq. App.		
Noise	White Noise	1	1	1	1	1	1	[36]
Chaotic Differential Equation	Lorenz	5 to 9	9	15	17	6	10	[36]
	Rosler	6 to 10	9	12	19	7	9	[42]
	Bi-directional Rosler	6 to 10	15	12	20	7	15	[36]
	Mackey-Glass	2 to 4	7	5	8	3	1 to 700	[36]
Periodic System	Sine Wave	5 to 8	1	10	16	21	15	[42]
Nonlinear Difference Eq.	Logistic Map	1	5	1	1	1	1 to 5	[36]
	Henon Map	1	8	1	1	1	1 to 5	[36]
Medical Data	ECG	1 to 2	8	21	13	2	1 to 4	[36]
	EEG	2 to 4	5	4	4	1	1 to 3	[36]

5 Conclusion

In this paper we demonstrated various methods for automatically determining the PE parameters τ and n when supplied with a sufficiently sampled/oversampled time series. The goal is to find, in an automatic way, the most accurate method in comparison to expert suggested parameters. The methods we investigated for calculating τ include autocorrelation, mutual information, permutation-auto-mutual information, frequency analysis, and multi-scale permutation entropy. Additionally, the methods we investigated for determining the embedding dimension n include false nearest neighbors, singular spectrum analysis, and multiscale permutation entropy.

Table 3: A comparison between the calculated and suggested values for the embedding dimension n . The shaded cells highlight the methods that yielded the closest match to the suggested dimension.

Category	System	Method			Suggested Dimension n	Ref.
		MPE	FNN	SSA		
Noise	White Noise	5	4	23	3 to 6	[36]
Chaotic Differential Equations	Lorenz	5	3	4	5 to 7	[36]
	Rössler	4	4	4	6	[42]
	Bi-directional Rössler system	4	4	4	6 to 7	[36]
	Mackey-Glass	4	4	6	4 to 8	[36]
Periodic System	Sine Wave	3	4	2	4	[42]
Difference Equations	Logistic Map	5	4	3	2 to 16	[36]
	Henon Map	5	4	2	3 to 10	[36]
Medical Data	ECG	5	7	8	3 to 7	[36]
	EEG	6	5	11	3 to 7	[36]

Our first contribution was developing a new frequency approach analysis and extending two existing methods, PAMI, and MPE, to automatically determine τ . For the frequency approach, we developed an automatic algorithm for finding the maximum significant frequency using a cutoff greater than the noise floor. The noise floor was found using one dimensional least median of squares applied to the Fourier spectrum in conjunction with the theoretical probability distribution function for the Fourier transform of Gaussian white noise. For PAMI, we showed how the minimum in the PAMI can be used to find an optimal embedding window, which is approximately independent of the dimension n . We then suggested using the range $4 \leq n \leq 6$ to find τ from the first minima in PAMI for $n = 2$. For MPE, we showed how it can be used to find the main period of oscillation from a periodic time series, which we then use to find τ . Additionally, we expanded upon this method by showing how the main period of oscillation can also be found for non-periodic time series, which we implemented into an automatic algorithm.

Our second contribution was implementing the automatic selection of τ using MPE to also find n using MPE. We also collected and compared some of the most popular methods for obtaining n including false nearest neighbors, and singular spectrum analysis. We applied these methods to various categories including difference equations, chaotic differential equations, periodic systems, EEG/ECG data, and Gaussian noise. We then compared the generated parameters to values suggested by experts to determine which methods consistently found accurate values for τ and n . We found that SSA did not provide suitable values for n . However, both FNN and MPE provided accurate values for n for most of the systems. We also concluded that, for the majority applications, a permutation dimension $n = 5$ is suitable. For determining τ , we showed that our frequency approach provided accurate suggestions for τ for periodic systems, nonlinear difference equations, and medical data, while the mutual information function computed using adaptive partitions provided the most accurate results for chaotic differential equations.

Acknowledgment

FAK acknowledges the support of the National Science Foundation under grants CMMI-1759823 and DMS-1759824.

References

- [1] Ralph G Andrzejak, Klaus Lehnertz, Florian Mormann, Christoph Rieke, Peter David, and Christian E Elger. Indications of nonlinear deterministic and finite-dimensional structures in time series of brain electrical activity: Dependence on recording region and brain state. *Physical Review E*, 64(6):061907, 2001.
- [2] Christoph Bandt and Bernd Pompe. Permutation entropy: a natural complexity measure for time series. *Physical review letters*, 88(17):174102, 2002.
- [3] T. Berry, J. R. Cressman, Z. Gregurić-Ferenček, and T. Sauer. Time-scale separation from diffusion-mapped delay coordinates. *SIAM Journal on Applied Dynamical Systems*, 12(2):618–649, jan 2013.
- [4] A Block, W Von Bloh, and HJ Schellnhuber. Efficient box-counting determination of generalized fractal dimensions. *Physical Review A*, 42(4):1869, 1990.
- [5] George EP Box, Gwilym M Jenkins, Gregory C Reinsel, and Greta M Ljung. *Time series analysis: forecasting and control*. John Wiley & Sons, 2015.

- [6] David S Broomhead and Gregory P King. Extracting qualitative dynamics from experimental data. *Physica D: Nonlinear Phenomena*, 20(2-3):217–236, 1986.
- [7] Th Buzug and G Pfister. Optimal delay time and embedding dimension for delay-time coordinates by analysis of the global static and local dynamical behavior of strange attractors. *Physical review A*, 45(10):7073, 1992.
- [8] Yinhe Cao, Wen-wen Tung, JB Gao, Vladimir A Protopopescu, and Lee M Hively. Detecting dynamical changes in time series using the permutation entropy. *Physical review E*, 70(4):046217, 2004.
- [9] Septima Poinsette Clark. Estimating the fractal dimension of chaotic time series. *Lincoln Laboratory Journal*, 3(1), 1990.
- [10] Madalena Costa, Ary L Goldberger, and C-K Peng. Multiscale entropy analysis of complex physiologic time series. *Physical review letters*, 89(6):068102, 2002.
- [11] Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2012.
- [12] G.A. Darbellay and I. Vajda. Estimation of the information by an adaptive partitioning of the observation space. *IEEE Transactions on Information Theory*, 45(4):1315–1321, may 1999.
- [13] Luciana De Micco, Juana Graciela Fernández, Hilda A Larrondo, Angelo Plastino, and Osvaldo A Rosso. Sampling period, statistical complexity, and chaotic attractors. *Physica A: Statistical Mechanics and its Applications*, 391(8):2564–2575, 2012.
- [14] Birgit Frank, Bernd Pompe, Uwe Schneider, and Dirk Hoyer. Permutation entropy improves fetal behavioural state classification based on heart rate analysis from biomagnetic recordings in near term fetuses. *Medical and Biological Engineering and Computing*, 44(3):179, 2006.
- [15] Andrew M Fraser and Harry L Swinney. Independent coordinates for strange attractors from mutual information. *Physical review A*, 33(2):1134, 1986.
- [16] Peter Grassberger and Itamar Procaccia. Measuring the strangeness of strange attractors. *Physica D: Nonlinear Phenomena*, 9(1-2):189–208, 1983.
- [17] Holger Kantz and Thomas Schreiber. *Nonlinear time series analysis*, volume 7. Cambridge university press, 2004.
- [18] Matthew B Kennel, Reggie Brown, and Henry DI Abarbanel. Determining embedding dimension for phase-space reconstruction using a geometrical construction. *Physical review A*, 45(6):3403, 1992.
- [19] Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. *Physical Review E*, 69(6), jun 2004.
- [20] HJ Landau. Sampling, data transmission, and the nyquist rate. *Proceedings of the IEEE*, 55(10):1701–1706, 1967.
- [21] David Lane, Joan Lu, Camille Peres, Emily Zitek, et al. Online statistics: An interactive multimedia course of study. Retrieved January, 29:2009, 2008.
- [22] Michel CR Leles, João Pedro H Sansão, Leonardo A Mozelli, and Homero N Guimarães. Improving reconstruction of time-series based in singular spectrum analysis: A segmentation approach. *Digital Signal Processing*, 77:63–76, 2018.
- [23] Duan Li, Zhenhu Liang, Yinghua Wang, Satoshi Hagihira, Jamie W. Sleight, and Xiaoli Li. Parameter selection in permutation entropy for an electroencephalographic measure of isoflurane anesthetic drug effect. *Journal of Clinical Monitoring and Computing*, 27(2):113–123, dec 2012.
- [24] Zhenhu Liang, Yinghua Wang, Gaoxiang Ouyang, Logan J Voss, Jamie W Sleight, and Xiaoli Li. Permutation auto-mutual information of electroencephalogram in anesthesia. *Journal of Neural Engineering*, 10(2):026004, feb 2013.
- [25] Desire L Massart, Leonard Kaufman, Peter J Rousseeuw, and Annick Leroy. Least median of squares: a robust method for outlier and model error detection in regression and calibration. *Analytica Chimica Acta*, 187:171–179, 1986.
- [26] Michael McCullough, Michael Small, Thomas Stemler, and Herbert Ho-Ching Iu. Time lagged ordinal partition networks for capturing dynamics of continuous dynamical systems. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 25(5):053101, 2015.
- [27] Michał Melosik and W Marszałek. On the 0/1 test for chaos in continuous systems. *Bulletin of the Polish Academy of Sciences Technical Sciences*, 64(3):521–528, 2016.
- [28] George B Moody and Roger G Mark. The impact of the mit-bih arrhythmia database. *IEEE Engineering in Medicine and Biology Magazine*, 20(3):45–50, 2001.
- [29] Audun Myers, Elizabeth Munch, and Firas A Khasawneh. Persistent homology of complex networks for dynamic state detection. *arXiv preprint arXiv:1904.07403*, 2019.
- [30] Athanasios Papoulis and S Unnikrishna Pillai. *Probability, random variables, and stochastic processes*. Tata McGraw-Hill Education, 2002.
- [31] Yakov Borisovich Pesin. Characteristic lyapunov exponents and smooth ergodic theory. *Uspekhi Matematicheskikh Nauk*, 32(4):55–112, 1977.
- [32] Steven M Pincus. Approximate entropy as a measure of system complexity. *Proceedings of the National Academy of Sciences*, 88(6):2297–2301, 1991.
- [33] Anton Popov, Oleksii Avilov, and Oleksii Kanaykin. Permutation entropy of eeg signals for different sampling rate and time lag combinations. In *Signal Processing Symposium (SPS), 2013*, pages 1–4. IEEE, 2013.
- [34] Mark A Richards. The discrete-time fourier transform and discrete fourier transform of windowed stationary white noise. *Georgia Institute of Technology, Tech. Rep*, 2013.

- [35] Joshua S Richman and J Randall Moorman. Physiological time-series analysis using approximate entropy and sample entropy. *American Journal of Physiology-Heart and Circulatory Physiology*, 278(6):H2039–H2049, 2000.
- [36] Müller Riedl, A Müller, and N Wessel. Practical considerations of permutation entropy. *The European Physical Journal Special Topics*, 222(2):249–262, 2013.
- [37] Claude E Shannon, Warren Weaver, and Arthur W Burks. *The mathematical theory of communication*. 1951.
- [38] Claude Elwood Shannon. A mathematical theory of communication. *ACM SIGMOBILE mobile computing and communications review*, 5(1):3–55, 2001.
- [39] MATTHÄUS STANIEK and KLAUS LEHNERTZ. PARAMETER SELECTION FOR PERMUTATION ENTROPY MEASUREMENTS. *International Journal of Bifurcation and Chaos*, 17(10):3729–3733, oct 2007.
- [40] Herbert A Sturges. The choice of a class interval. *Journal of the american statistical association*, 21(153):65–66, 1926.
- [41] Floris Takens. Detecting strange attractors in turbulence. In *Dynamical systems and turbulence, Warwick 1980*, pages 366–381. Springer, 1981.
- [42] Mei Tao, Kristina Poskuviene, Nizar Alkayem, Maosen Cao, and Minvydas Ragulskis. Permutation entropy based on non-uniform embedding. *Entropy*, 20(8):612, 2018.
- [43] Alan Wolf, Jack B Swift, Harry L Swinney, and John A Vastano. Determining lyapunov exponents from a time series. *Physica D: Nonlinear Phenomena*, 16(3):285–317, 1985.
- [44] Hong Zhang and Xuncheng Liu. Analysis of parameter selection for permutation entropy in logistic chaotic series. In *Intelligent Transportation, Big Data & Smart City (ICITBS), 2018 International Conference on*, pages 398–402. IEEE, 2018.
- [45] Luciano Zunino, Miguel C Soriano, Ingo Fischer, Osvaldo A Rosso, and Claudio R Mirasso. Permutation-information-theory approach to unveil delay dynamics from time-series analysis. *Physical Review E*, 82(4):046212, 2010.

A Appendix

A.1 Dynamic System Models

A.1.1 Lorenz System

The Lorenz system used is defined as

$$\frac{dx}{dt} = \sigma(y - x), \quad \frac{dy}{dt} = x(\rho - z) - y, \quad \frac{dz}{dt} = xy - \beta z. \quad (24)$$

The Lorenz system had a sampling rate of 100 Hz with parameters $\sigma = 10.0$, $\beta = 8.0/3.0$, and $\rho = 95$. This system was solved for 100 seconds and the last 24 seconds were used.

A.2 Rössler System

The Rössler system used was defined as

$$\frac{dx}{dt} = -y - z, \quad \frac{dy}{dt} = x + ay, \quad \frac{dz}{dt} = b + z(x - c), \quad (25)$$

with parameters of $a = 0.1$, $b = 0.1$, $c = 14$, which was solved over 400 seconds with a sampling rate of 10 Hz. Only the last 1500 data points of the x-solution were used in the analysis.

A.2.1 Bi-Directional Coupled Rössler System

The Bi-directional Rössler system is defined as

$$\begin{aligned} \frac{dx_1}{dt} &= -w_1 y_1 - z_1 + k(x_2 - x_1), \\ \frac{dy_1}{dt} &= w_1 x_1 + 0.165 y_1, \\ \frac{dz_1}{dt} &= 0.2 + z_1(x_1 - 10), \\ \frac{dx_2}{dt} &= -w_2 y_2 - z_2 + k(x_1 - x_2), \\ \frac{dy_2}{dt} &= w_2 x_2 + 0.165 y_2, \\ \frac{dz_2}{dt} &= 0.2 + z_2(x_2 - 10), \end{aligned} \quad (26)$$

with $w_1 = 0.99$, $w_2 = 0.95$, and $k = 0.05$. This was solved for 4000 seconds with a sampling rate of 10 Hz. Only the last 400 seconds of the x-solution were used in the analysis.

A.2.2 Mackey-Glass Delayed Differential Equation

The Mackey-Glass Delayed Differential Equation is defined as

$$\dot{x}(t) = -\gamma x(t) + \beta \frac{x(t - \tau)}{1 + x(t - \tau)^n} \quad (27)$$

with $\tau = 2$, $\beta = 2$, $\gamma = 1$, and $n = 9.65$. This was solved for 400 seconds with a sampling rate of 100 Hz. The solution was then downsampled to 5 Hz and only the last 1500 terms of the x-solution were used in the analysis.

A.2.3 Periodic Sinusoidal Function

The sinusoidal function is defined as

$$x(t) = \sin(2\pi t) \quad (28)$$

This was solved for 10 seconds with a sampling rate of 50 Hz.

A.2.4 EEG Data

The EEG signal was taken from andrzejak et al. [1]. Specifically, the first 2000 data points from the EEG data of a healthy patient from set A, file Z-093 was used.

A.2.5 ECG Data

The Electrocardiogram (ECG) data was taken from SciPy's misc.electrocardiogram data set. This ECG data was originally provided by the MIT-BIH Arrhythmia Database [28]. We used data points 3000 to 4500 during normal sinus rhythm.

A.2.6 Logistic Map

The logistic map was generated as

$$x_{n+1} = rx_n(1 - x_n), \quad (29)$$

with $x_0 = 0.5$ and $r = 3.95$. Equation 29 was solved for the first 500 data points.

A.2.7 Hénon Map

The Hénon map was solved as

$$\begin{aligned} x_{n+1} &= 1 - ax_n^2 + y_n, \\ y_{n+1} &= bx_n, \end{aligned} \quad (30)$$

where $b = 0.3$, $x_0 = 0.1$, $y_0 = 0.3$, and $a = 1.4$. This system was solved for the first 500 data points of the x-solution.

A.3 Algorithm

A.3.1 MPE delay Algorithm

In Section 2.2.2 we showed that choosing τ using MPE should be based on the position of the first peak after the noise in the MPE plot for an embedding dimension of $n = 3$. At this maximum, the normalized PE hits a maximum of approximately 1. From this methodology, we developed Algorithm 1 to determine the delay τ using the location of the first peak, while ignoring the noise region in Fig. 11.

Result: τ

```

begin
  set  $n = 3$ ;
  start with a delay of  $\tau = 1$ ;
  start with initial normalized PE as  $h_0 = 0$ ;
  while first peak not found do
    calculate normalized PE as  $h(\tau)$ ;
    if  $h(\tau) < 0.9$  then
      | Set outside of noise section flag as true ( $f_n = \text{True}$ )
    end
    if  $h(\tau) > 0.9$  and  $f_n = \text{True}$  then
      | if  $h(\tau) < h_0$  then
      |   |  $\tau = \tau - 1$ ;
      |   | First peak found;
      | end
      | Set  $h_0 = h(\tau)$ 
    end
     $\tau = \tau + 1$ 
  end
  return  $\tau$ ;
end

```

Algorithm 1: Algorithm using MPE for τ .