4.1 Initial Considerations

As introduced in Chapter 2, reconstructing the phase space from a time series is a key step in dynamical systems analysis. For this purpose, the embedding theorem proposed by Takens (1981), also known as method of delays, is the most employed approach in the literature (Packard et al., 1980). This chapter further discusses methods for performing the phase-space reconstruction that are associated with Takens’ theorem.

Before describing actual methods to compute embedding parameters \((m, \tau)\), let us overview the general considerations that underlie the computation of such embeddings. We outline three points of interest: the use of a single vs multiple time delays \(\tau\); the effects of overestimating the embedding dimension \(m\); and the impact for the estimation if \(m\) and \(\tau\) are correlated or not.

**Single vs multiple time delay:** Methods for computing phase-space optimal embeddings of time series rely on the fact that a time series \(T_i\), whose observations are the components of a single dimension \(i\) of \(S^d\), is directly or indirectly influenced by other non-measured variables \(j \in [1,d], \forall j \neq i\). While studying natural phenomena, however, such influences among variables may happen at different timestamps and intervals, leading to nonlinear recurrences and chaos (Kantz and Schreiber, 2004). Thus, one could use multiple time delays \(\tau_1, \tau_2, \cdots\) to reconstruct the attractor of \(S\) that resembles the features of \(S^d\), as first investigated by Breedon and Packard (1992) and more recently by Manabe and Chakraborty (2007). However, this Ph.D. (as most studies) is mainly concerned in estimating a single \(\tau\), which may be interpreted as the most common or maximal time delay. Thus, implications of using more than one time delay in phase-space reconstruction is out of the scope of this thesis.

**Overestimated embedding dimension:** In order to reconstruct the phase space \(S \subset S^d\), an adequate pair of embedding parameters \(m\) and \(\tau\) is required to unfold the system dynamics. In that sense, given that most real-world phenomena present low-dimensional phase spaces (as shown in Chapter 3), underestimated embeddings are difficult to occur in practice, at least in terms of \(m\). Conversely, estimations leading to an overestimated embedding dimension will
not impact the quality of the analysis (as the phase space will be already unfolded), but will increase the computational effort needed for modeling, which is undesirable for real-time applications on data streams. Thus, a minimal embedding pair \((m, \tau)\) is highly desired. Complementary, an overestimated time delay may generate misleading conclusions, especially when dealing with maps as it is the case of the Hénon and Logistic time series. For instance, a second-order polynomial regression (Björck, 1996) applied over the phase space of Figure 1.3(a) gives

\[
x(t + 1) = -3.8x(t)^2 + 3.8x(t) - 1.49236 \times 10^{-7},
\]

which is enough to get an approximation of the actual generating rule built with \(r = 3.8\) (Equation 3.2). With the generating rule available, simpler and more reliable analyses can be performed than when using the raw time series itself. Nonetheless, the attractor structure is lost in Figure 1.3(b), which resembles a phase space from a stochastic process (Alligood et al., 1996). In such cases, no simple regression model is capable of providing the generating rule.

**Relation of the two parameters:** Regarding embedding parameters, despite Takens proved that \(m\) and \(\tau\) are independent for infinite-length free-noise time series, there is no consent about the true relation among such parameters for finite-noisy observations (Martinerie et al., 1992; Kim et al., 1999; Ma and Han, 2006; Cai et al., 2008, etc.). Some researchers, for instance, focused on finding one of those parameters at a time, assuming they are uncorrelated. Conversely, others believe that these parameters are bounded by the time delay window \(t_w = (m - 1)\tau\) (a number representing the total time spanned by the components of each embedded point), thus making \(m\) and \(t_w\) independent instead.

Next, we discuss methods that estimate these parameters assuming that they are uncorrelated (Section 4.2) or dependent (Section 4.3). Lastly, Section 4.4 concludes the chapter with our final considerations.

### 4.2 Assuming Independence of Embedding Parameters

Methods to estimate the optimal time delay \(\tau\) (Section 4.2.1) can be mainly grouped in two categories: the ones that use series correlation (Fraser and Swinney, 1986; Albano et al., 1987, 1991; Ma and Han, 2006, etc.), and the ones based on the phase-space expansion (Kember and Fowler, 1993; Rosenstein et al., 1994; Chen et al., 2016, etc.). Despite their differences, such methods do not attempt to estimate \(m\) in the process, usually performing computations on the time series itself (which, roughly speaking, is the same as using
an embedding dimension \( m = 1 \) or on some predefined \( m \). For instance, when the generating rule \( R(\cdot) \) is given, some authors decide to use the minimal embedding dimension that satisfies Takens’ theorem \((m = 2d + 1)\).

When estimating the embedding dimension \( m \) (Section 4.2.2), on the other hand, authors typically use \( \tau = 1 \) or the time delay found by one of the previously-mentioned methods (detailed in Section 4.2.1). Also, those algorithms usually take decisions based on phase-states distances. Therefore, the phase state is always assumed to be endowed in the Euclidean space, unless stated differently.

4.2.1 Estimating The Time Delay

Several methods exist for estimating the time delay \( \tau \): Autocorrelation Function (Section 4.2.1.1), Auto-Mutual Information (Section 4.2.1.2), high-order correlations (Section 4.2.1.3), Singular Value Fraction (Section 4.2.1.4), Average Displacement (Section 4.2.1.5), Multiple Autocorrelation Function (Section 4.2.1.6), and Dimension Derivation (Section 4.2.1.7). These are described next.

4.2.1.1 Autocorrelation Function

Let \( S = \{x(t), x(t+1), \ldots \} \) and \( Q = \{x(t+\tau), x(t+1+\tau), \ldots \} \) denote, respectively, the time series \( T_i \) and the same series \( \tau \)-units shifted along time, namely \( T_{i,\tau} \). The Autocorrelation Function (ACF) quantifies the amount of linear independence in the form

\[
R_{xx}(\tau) = \frac{E[(S - \mu)(Q - \mu)]}{\sigma^2},
\]

where \( E[\cdot] \) is the expected value, \( \mu \) is the mean, and \( \sigma^2 \) is the variance of \( T_i \). In the scope of phase-space reconstruction, ACF was used to estimate the value of \( \tau \) that better unfolded the attractor, since authors believed that the more independent states were from each other (up to a certain limit), the greater it is the probability of describing a deterministic rule and form a representative structure. Then, several approaches were proposed on \( R_{xx}(\cdot) \) to measure this limit. For instance, Albano et al. (1987); Abarbanel et al. (1993) chose the first zero-crossing \( R_{xx}(\cdot) \) to estimate \( \tau \), while King et al. (1987) suggested the first inflection point of \( R_{xx}(\cdot) \) instead. Later, Albano et al. (1988) achieved more robust results by choosing the time delay that causes \( R_{xx}(\cdot) \) to first drop to a certain fraction of its initial value. However, despite that autocorrelation methods provide a good initial estimation for the time delay in some scenarios, such approaches showed to be inconsistent.
for general systems, probably due to the low correlation among linear dependencies on the time series and the nonlinear structures of the underlying dynamical system (Fraser and Swinney, 1986; Martinerie et al., 1992).

4.2.1.2 Auto-Mutual Information

Fraser and Swinney (1986), already knowing the problems involving the linear dependencies of ACF (even earlier than the publication of corresponding articles), proposed a method to measure the general dependencies among signals based on the Auto-Mutual Information (AMI)

$$I(\tau) = \int_S \int_Q p_{SQ}(s,q) \log_2 \left( \frac{p_{SQ}(s,q)}{p_S(s)p_Q(q)} \right) dxdy, \quad (4.3)$$

where $p_S$ and $p_Q$ are marginal probability densities from the continuous variables $S$ e $Q$ (representing $T_i$ and $T_{i+\tau}$), respectively, and $p_{SQ}$ is the joint probability density function. Given a measurement $s \in S$, the mutual information is the number of bits of $q \in Q$, on average, that can be predicted. Thus, $p_S$ and $p_Q$ can be estimated by respectively partitioning $S$ and $Q$ into bins and counting occurrences on them. Similarly, $p_{SQ}$ is usually computed by recursively partitioning the plane $SQ$ until each cell is uniformly distributed according to some statistical criteria. As observed later by Rosenstein et al. (1994), the authors used the auto-mutual information in attempt to measure the shift from redundancy to irrelevance on time series. As the irrelevance error is more difficult to compute, Fraser and Swinney (1986) associated the first local minimum of AMI to the optimal $\tau$ (less redundant time delay).

Liebert and Schuster (1989) later showed that such minima coincide with those found using the correlation integral (Theiler, 1987; Wong et al., 2005). However, Martinerie et al. (1992) empirically observed that neither ACF (using the previously described measurements) nor AMI were consistent to find the optimal value of $t_w$ (and, as consequence, a bound for $m$ and $\tau$), as illustrated in Figure 4.1.

4.2.1.3 High-Order Correlation

Albano et al. (1991) noticed that time series deriving from multivariate systems should carry multivariate cumulants and moments. By investigating the relations of high-order correlations on the columns of the trajectory matrix (formed by states of the phase space, as described in Equation 2.12), they observed that a number of correlation functions, although not consistently the same ones, have extrema occurring at the same time, later referred to as $t_c$.
4.2 Assuming Independence of Embedding Parameters

![Figure 4.1](image)

As there are no previous reason for such coincidence, the authors empirically defined $t_c$ as the time delay window $t_w = (m - 1)\tau$, which defines important attractor features. However, no patterns were found to deterministically correlate $t_c$ with $t_w$ (and $\tau$).

4.2.1.4 Singular Value Fraction

Kember and Fowler (1993) proposed the Singular Value Fraction (SVF) method, which estimates the time delay when the attractor is mostly expanded in all dimensions. As the expansion of an attractor can be described by the spreading rate of its phase states, i.e., in function of the singular values $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m \geq 0$ of the trajectory matrix (Equation 2.12), they attempt to find to the (ideal) case when all eigenvalues are equal. Thus, given the function

$$F_{sv}(k) = \frac{\sum_{j=1}^{k} \lambda_j^2}{\sum_{j=1}^{m} \lambda_j^2}, \quad (4.4)$$

SVF is defined as

$$f_{sv}(k) = \frac{mF_{sv}(k) - k}{m - k}, \quad (4.5)$$
so that $0 \leq f_{sv}(k) \leq 1$ and $f_{sv}(m) = 1$. By this definition, Equation 4.5 only approaches zero when $F_{sv} = k/m$, meaning $\lambda_2 = c, \forall i \in [1, m]$, where $c$ is a constant. For general attractors, however, hardly ever $f_{sv} = 0$, and, consequently, $\tau$ was defined as the first timestamp SVF reaches a local minimum. Moreover, the authors empirically observed that $f_{sv}(1)$, *i.e.*, the average of eigenvalues in function of the most spread dimension, led to better estimations of the time delay. From our point of view, such approach was an early attempt to overcome the well-known “curse of dimensionality” (Chen, 2009).

One interesting point about this method is that the plot of $f_{sv}(1)$ versus $\tau$ gives a simple (but important) idea about the time-series recurrences and the interval of its periodicities, as illustrated in Figure 4.2(a). As it can be seen, this figure shows that the Rössler system has cyclical behavior each 29 units of time, providing empirical evidence that the first peak of SVF could be used to estimate the time delay window. Despite consistent results for different dimensions (including noisy data), the same strategy failed for the Lorenz system, as shown in Figure 4.2(b). As a counter-argument to this limitation, the authors were able to enhance results for the Lorenz series by powering the observations in the form $T^\alpha = \{x(t)^2, x(t+1)^2, \cdots\}$, where $\alpha$ is the manifold genus (number of voids or holes). However, this approach may be difficult to be reproduced when there is no *a priori* knowledge about the system.

![Figure 4.2](image)

Figure 4.2: (a) The SVF method found $\tau = 10$ for the Rössler system when embedded using $m = 3$, which is a good estimate for the time delay. Also, a constant time-delay-window length was found for different dimensions, empirically reinforcing the belief that $m$ and $\tau$ are independent to each other. (b) On the other hand, SVF does not provide any useful information for the Lorenz system: there is neither a minimum nor a peak. From outer to inner curves, $m = 2, \cdots, 10$ in both plots.

Similarly, Chen et al. (2016) proposed Singular Entropy (SE), a modified version of SVF. They based their algorithm on the same
4.2 Assuming Independence of Embedding Parameters

principle of equality as SVF, but used the ratio of the entropy of eigenvalues instead

\[ E(k) = \sum_{j=1}^{k} \Delta E_j, \quad (4.6) \]

where

\[ \Delta E_j = -P_j \log P_j, \quad P_j = \frac{\lambda_j}{\sum_{m=1}^{\infty} \lambda}. \quad (4.7) \]

Given Equation 4.6 was based on Shannon’s entropy (Hammer et al., 2000) of the singular values, the major difference between SE and SVF is the sign of their results, and, therefore, the concavity of their plots. Thus, similar results given by SVF, such as \( \tau \) and \( t_w \), were also obtained while analyzing inverse features of SE, i.e., the minimum of SVF becomes the maximum of SE and vice-versa, as illustrated in Figure 4.3.

![Figure 4.3: SE leads to similar conclusions as the SVF method for the Rössler (a) and Lorenz (b) systems. From outer to inner curves, \( m = 10, \cdots, 2 \).](image)

4.2.1.5 Average Displacement

An interesting discussion about irrelevance and redundancy, as well as the errors associated with both structural conditions, were conducted by Rosenstein et al. (1994). Despite the difficulties in measuring the irrelevance error (there are no guarantees the reconstruction will become less space-filling as the lag increases beyond the optimal time delay), the authors assumed the irrelevance error to be lower than the redundancy error, therefore focusing on minimizing the latter while searching for a proper time delay. In this scenario, they inversely measure the relation between the redun-
dance error and the attractor expansion in form of the Average Displacement (AD) of phase states, as function of the time delay

\[ S_m(\tau) = \frac{1}{N_i} \sum_{t=0}^{N_i-1} \sqrt{\sum_{j=1}^{m-1} (x(t + j\tau) - x(t))^2}. \]  
\( (4.8) \)

As illustrated in Figure 4.4, it was observed that AD increases until it reaches a plateau on the plot \( S_m(\tau) \) versus \( \tau \), which indicates the attractor is sufficiently expanded. Thus, the authors empirically observed that a good estimation for \( \tau \) was given when the slope of such relation decreased by less than 40\% of its initial value. Additionally, the authors also noticed that this criterion corroborated with the assumption that \( \tau \) and \( m \) are correlated within the time delay window: increasing the values of \( m \) contributes to the early formations of the plateau (and early estimations for \( \tau \)). Indeed, the pairs \((m, \tau)\) found for the Lorenz system (shown in the figure) were \{\((14, 4), (8, 5), (5, 8), (3, 14)\)\}. Despite those contributions, some authors argued that the dependency on the slope of the AD method may consider non-ignorable errors (Ma and Han, 2006).

![Figure 4.4: Results of the average displacement for the Lorenz system. Diagonal lines indicate the first time delay in which the slope of the curve decreases by less than 40\% of its initial value. From outer to inner curves: \( m = 14, 8, 5, 3 \).](image)

### 4.2.1.6 Multiple Autocorrelation Function

Rosenstein et al. (1994) also investigated the relation between AD and ACF. They squared Equation 4.8 as follows

\[ S^2_m(\tau) = \frac{1}{N_i} \sum_{t=0}^{N_i-1} \sum_{j=1}^{m-1} (x(t + j\tau) - x(t))^2, \]
\( (4.9) \)
so that $S_m^2(\tau)$ could be interpreted as a scaled version of $S_m(\tau)$ as $S_m^2(\tau) = f(\tau)S_m(\tau)$. Moreover, as ACF (Equation 4.2) from a finite set can be approximated by

$$R_{xx}(\tau) \approx \frac{1}{n_i - \tau} \sum_{t=0}^{n_i-\tau} x(t)x(t+\tau),$$

it can be shown that, under some assumptions

$$S_m^2(\tau) \approx c - 2R_{xx}^m(\tau),$$

where $c$ is a constant and $R_{xx}^m(\tau) = \sum_{j=1}^{m-1} R_{xx}(j\tau)$. After this definition, Rosenstein et al. (1994) noticed that $R_{xx}(\cdot)$ has the same shape as $S_m^2$ when $m = 2$, indicating the easier-to-compute ACF should be used in the place of $S_m^2$ for two-dimensional systems. However, as most systems are at least three-dimensional and $S_m^2$ tends to become more sensitive to variations as $m$ is increased, the authors concluded that the Multiple Autocorrelation Function (MACF), as later referred to $S_m^2$, should lead to the poor estimation of attractors for high-dimensional systems.

### 4.2.1.7 Dimension Derivation

More recently, Tamma and Lachman Khubchandani (2016) proposed a method to find both embedding parameters $m$ and $\tau$ from a time series. However, as the method to find the embedding dimension $m$ was very similar to FNN (Kennel et al., 1992), the main contribution of their article was the estimation of $\tau$, which is based on the attractor expansion and the pointwise correlation dimension, defined as

$$D_p(t, \tau) = \lim_{\epsilon \to 0} \frac{\log(P(t, \tau, \epsilon))}{\log(\epsilon)},$$

where $P(t, \tau, \epsilon)$ is the probability of two phase states, reconstructed using a predefined $m$ and the chosen $\tau$, to be closer than the open ball centered in $\phi_i(t)$ with radius $\epsilon$, similar to (Equation 2.21). Since $D_p(t, \tau)$ should be invariant to any chosen state $\phi_i(t)$ (Ott, 2002), a well-reconstructed phase space should present zero Dimension Deviation (DD) among their pointwise correlations, i.e.

$$f(\tau) = \frac{1}{N_i} \sum_{t=0}^{N_i-1} (D_p(t, \tau) - \overline{D_p}(t, \tau))^2 = 0,$$

where $\overline{D_p}(t, \tau) = \sum_{t=0}^{N_i-1} D_p(t, \tau)/N_i$.

Due to noise and floating point fluctuations, however, Equation 4.13 would hardly be equal to zero. Hence, the authors defined
the first minimum of $f(\tau)$ (over a predefined range of time delays) as the optimal time delay. However, results showed that neither this method nor the choice of the first minimum are robust to estimate $\tau$ for general attractors, as illustrated in Figure 4.5. There, the fourth minimum would lead to the closest optimal estimation for the Rössler system. Lastly, the authors claimed that a representative set of states could be used instead of all phase states in order to improve performance.

To check this out, we tested the DD algorithm using (roughly) 1000 phase states and 200 (representative) centroids. Following our implementation, we found a smoother curve for the plot $f(\tau)$ versus $\tau$. This actually hindered the estimation for the time delay, as depicted in Figure 4.6.

![Figure 4.5: Results of the DD method for the Rössler system (a) and Logistic map (b) over the range of $\tau \in [1, 20]$. Filled circles represent the number of minima until the closest estimation for the optimal time delay.](image)

### 4.2.2 Estimating The Embedding Dimension

Methods to estimate the embedding dimension $m$ comprise False Nearest Neighbors (Section 4.2.2.1), Gamma test (Section 4.2.2.2), and others based on the fractal dimension (Section 4.2.2.3), as follows.

#### 4.2.2.1 False Nearest Neighbors

Kennel et al. (1992) proposed False Nearest Neighbors (FNN) to estimate the optimal embedding dimension $m$. By using the time delay found after AMI (Section 4.2.1.2), the method reconstructs the attractor using different dimensions $m \in [m_{\text{min}}, m_{\text{max}}]$ and computes, for each of them, the index set of the $k$-nearest neighbors $N_m(\phi_i(t), k)$ for each phase state $\phi_i(t) \in S^m$ (the number of phase states is kept constant for different embeddings). The opti-
4.2 Assuming Independence of Embedding Parameters

Figure 4.6: The DD method applied to the Lorenz system using \( m = 3 \).

(a) 200 centroids led to an acceptable time delay \( \tau = 10 \), but only when the fourth minimum of \( f(\tau) \) is used. (b) The first minimum yielded \( \tau = 17 \) when all phase states are considered. As in Figure 4.5, filled circles represent the number of minima until the closest estimation for the optimal time delay.

The embedding dimension \( m \) is defined as the one in which a fraction of the nearest neighbors in the attractor remains constant as the dimension increases, i.e., \( N_m(\phi(t), k) = N_{m-1}(\phi(t), k) \). This assumption could be relaxed, however, by analyzing how many neighbors (usually 30%) remained close to each other according to a threshold \( r_{\text{tol}} \), as \( m \) is increased.

In other words, if the \( k \)th nearest neighbor of the phase state \( \phi_i(t) \) is \( \phi_i(t') \), then the Euclidean distance between them is given by

\[
R^2_m(t, \tau, k) = \sum_{j=1}^{m} (x(t+j\tau) - x(t'+j\tau))^2.
\] (4.14)

After adding one more dimension, we get

\[
R^2_{m+1}(t, \tau, k) = R^2_m(t, \tau, k) + (x(t+(m+1)\tau) - x(t'+(m+1)\tau))^2.
\] (4.15)

If the distance between such states become greater than \( r_{\text{tol}} \) after increasing one dimension, i.e.

\[
\left( \frac{R^2_{m+1}(t, \tau, k) - R^2_m(t, \tau, k)}{R^2_{m+1}(t, \tau, k)} \right)^{1/2} = \frac{x(t+(m+1)\tau) - x(t'+(m+1)\tau)}{R^2_{m+1}(t, \tau, k)} > r_{\text{tol}},
\] (4.16)

then the states are considered false neighbors of each other, as illustrated in Figure 4.7.
Kennel et al. (1992) stated that as long as there is some variation in the attractor, the phase space is not completely unfolded and, thus, the number of false neighbors is still too large. According to their experiments, $r_{tol} \geq 10$ and $k = 1$ were enough to correctly identify false neighbors\(^1\). A problem raised by the authors is the lack of effectiveness of FNN in the presence of noise. A simple example was performed using a white noise following the Normal distribution $N(0, 1^2)$, which is known to have a high-dimensional (but finite) attractor. In this scenario, FNN tends to find greater values of $m$ as the number of states increases, i.e., $N = \infty \rightarrow m = \infty$. This occurs mainly due to the curse of dimensionality (Chen, 2009), in which all states tend to become dissimilar to each other as the dimension increases, never dropping the rate of false neighbors. As observed, this is a contrast to commonly found low-dimensional attractors, where more states in the phase space should contribute to a better formation of its structure (and to the confirmation of its dimension).

Although this method does not achieve consistent results for general time series (which usually contain noisy observations), due to seminal contributions and simplicity, FNN still remains as one of the most used methods to estimate the embedding dimension.

---

\(^1\) It is not uncommon, however, to find algorithms that use $r_{tol} \geq 20$. 

---

Figure 4.7: Example of false nearest neighbors. Representation of 100 states of the Rössler system. For the sake of simplicity, three states are depicted by letters $A, B, C$. In the example, $A$ and $C$ remain neighbors from each other after expanding the space in one dimension. State $B$ is labeled as a false neighbor.
4.2 Assuming Independence of Embedding Parameters

4.2.2.2 Gamma Test

Otani and Jones (1997) proposed to use the Gamma test (or \( \Gamma \) test) (Stefánsson et al., 1997) to find the optimal embedding dimension \( m \). Let

\[
y = f(X) + r \Rightarrow [y_0, y_1, \cdots, y_N] = f([x_0, x_1, \cdots, x_N]) + r, \tag{4.17}
\]

be a function applied over \( m \)-dimensional phase states \( \phi_i(t) \), \( \forall t \in [0, N_i - 1] \), such that every pair \( (x_t, y_t) \) is defined as

\[
([x(t), \cdots, x(t + (m - 2))\tau], x(t + (m - 1)\tau)). \tag{4.18}
\]

The adjust parameter \( r \) can be seen as the amount of noise in the system or the lack of determinism of \( f \). Therefore, if one assumes a controlled level of noise, the variance of \( r \), namely \( \sigma_r^2 \), estimates the quality of the reconstructed phase space. Given the statistic

\[
\gamma = \frac{1}{2N_i} \sum_{t=0}^{N_i - 1} (y_{N_m(y_t,1)} - y_t)^2, \tag{4.19}
\]

where \( N_m(y_t,1) \) is the index of the nearest neighbors of \( y_t \), one can show that \( \gamma \to \sigma_r^2 \) as \( N \to \infty \). Let \( N^x = N_m(x_t, k) \) and \( N^y = N_m(y_t, k) \) be the index vectors of the \( k \)-nearest neighbors of \( x_t \) and \( y_t \), respectively. If \( \| \cdot \|_2 \) is the Euclidean norm, then

\[
\Delta(k) = \frac{1}{k} \sum_{h=1}^{k} \frac{1}{N_i - 1} \sum_{t=0}^{N_i - 1} \| x_{N^x_h} - x_t \|_2^2, \tag{4.20}
\]

computes the mean square distance of the \((h \leq k)\)-nearest neighbors and

\[
\Gamma(k) = \frac{1}{k} \sum_{h=1}^{k} \frac{1}{2N_i} \sum_{t=0}^{N_i - 1} \| y_{N^y_h} - y_t \|_2^2, \tag{4.21}
\]

is an estimation of \( \gamma \) for the \((h \leq k)\)-nearest neighbors. Based on both quantities, the \( \Gamma \)-test algorithm estimates \( \gamma \to \sigma_r^2 \) by means of the linear correlation of \( \Delta(k) \) and \( \Gamma(k) \). In a perfect deterministic scenario, where \( \sigma_r^2 = 0 \), \( x_t \) should lead to exactly one \( y_t \), and \( N^x = N^y \). Therefore, the ideal plot of \( \Delta(k) \) versus \( \Gamma(k) \) should form a straight diagonal line for different values of \( k \). Due to noise and fluctuations, this plot becomes irregular in practice. After a linear regression, the line \( y = mx + \Gamma \) that best fits the plot is an estimation of Equation 4.17, so that \( \Gamma \approx \sigma_r^2 \).

Otani and Jones (1997) noted that the phase space should be described by non-deterministic states for small values of \( m \), yielding
greater values for $\sigma^2_r$. For a sufficient $m$, the attractor is unfolded and should present the most deterministic structure (lower $\sigma^2_r$). For greater values of $m$, neighbors suffer interferences from different orbits, again increasing $\sigma^2_r$. Therefore, the optimal embedding was defined as the dimension when $\Gamma$ reaches its first minimum.

As the authors also observed, the less $\Gamma$ approaches to zero, the more non-deterministic are the state changes, and there is no guarantee in reconstructing the attractor accurately. This may happen if the signal-to-noise ratio is low or when the choice of the time delay is poor. With this in mind, rather than using an elsewhere estimated time delay, one can propose an extended algorithm to find $(m, \tau)$ using an acceptable trade-off between the smaller values of $\Gamma$ and $m$, assuming $\Gamma$ initially decreases as $m$ is increased.

4.2.2.3 Methods Based On The Fractal Dimension

As already explained at various points, the reconstruction of attractors using the method of delays relies on the mathematical framework proposed by Takens (1981). His theorem states that a sufficient reconstruction is achieved by using at least $m \geq 2d + 1$ dimensions for a $d$-dimensional system. For common phenomena, however, $m$ is (usually) considerably smaller, i.e., $m \leq 2d + 1$. In addition, as stated by Otani and Jones (1997), the effective dimensionality of some dissipative system is that of its attractor, which may be lower than the number of variables specialists use to describe it. Therefore, if the dimension of the attractor is $d' \leq d$, then $m$ is at least $d'$ (Farmer and Sidorovich, 1987). Based on both facts, the relation

$$d' \leq m \leq 2d' + 1,$$

(4.22)
can be used to estimate or validate the embedding dimension $m$. However, in order to apply Equation 4.22, one needs to rely on two assumptions: i) embedding the phase space using $m \leq 2d + 1$ is enough to unfold the dynamics of the underlying system; and ii) the attractor dimension $d'$ is well estimated by fractal-based methods (see Section 2.6.1). Thus, methods based on this approach may be inconsistent for, and not applicable to, general systems.

4.3 Assuming Dependence To Estimate The Embedding Parameters

A separate class of algorithms does not use previous estimations or any other information about the underlying dynamical system. Based on the assumption $\tau$ and $m$ are correlated by the time delay window $t_w = (m - 1)\tau$, some authors believe that important
4.3 Assuming Dependence to Estimate the Embedding Parameters

System features (e.g., correlation dimension) can be revealed using different combinations of $(\tau, m)$. Therefore, as some authors strive on defining such window, others focus on directly finding the simpler, but sufficient, set of parameters $m$ and $\tau$ that provides a good reconstruction. Therefore, most algorithms rely on Monte Carlo simulations (Landau and Binder, 2005; Rubinstein and Kroese, 2007) to find both parameters. This makes such methods quite computationally expensive.

Methods that estimate $m$ and $\tau$ assuming some dependence between them include: Wavering Product (Section 4.3.1), Fill Factor (Section 4.3.2), C-C method (Section 4.3.3), Entropy Ratio (Section 4.3.4), Non-Biased MACF, Gamma test (Section 4.3.5), and neural networks (Section 4.3.6), as follows.

4.3.1 Wavering Product

The Wavering Product (WP) was one of the first methods designed to find the embedding dimension $m$ and the time delay $\tau$ at the same time. Similarly to FNN, this method uses an expansion-based approach relying on the fact that a well-reconstructed embedding should have consistent topological structure, such as neighborhood relationship. Differently from FNN, Liebert et al. (1991) used the distances between the old $S^m$ and new $S^{m+1}$ embeddings in their method, as explained next.

In order to keep consistency with the previous presented nomenclature, let

$$R^2_m(t, \tau, k, m + 1)$$

be the Euclidean distance from the phase state $\phi_i(t) \in S^m$ to its $k$th nearest neighbor, but measured in the embedding $\Phi_i \in S^{m+1}$, having the inverse applied to $R^2_{m+1}(t, \tau, k, m)$. In this context, the authors defined the ratio

$$Q_1(t, \tau, k, m) = \frac{R^2_{m+1}(t, \tau, k, m)}{R^2_{m+1}(t, \tau, k, m + 1)},$$

so that $Q_1(t, \tau, k, m) = 1$ when the distance of $\phi_i(t)$ to its $k$th nearest neighbor remains equal in both embeddings, and $Q_1(t, \tau, k, m) > 1$ otherwise. Nonetheless, even when $\phi_i(t)$ has the same set of $p$-nearest neighbors for greater dimensions, Equation 4.24 becomes sensitive to the order of how neighbors are rearranged. Such an order may oscillate from one dimension to another as well as due to the presence of noise (hence the name wavering). In order to mitigate this, the authors applied

$$P_i(m, \tau) = \left( \prod_{k=1}^{p} Q_1(t, \tau, k, m) \right)^{1/p}$$

(4.25)
over all \( p \)-nearest neighbors. However, as \( P_i \) does not tend to 1 even for sufficiently-known embeddings, the authors introduced a second ratio

\[
Q_2(t, \tau, k, m) = \frac{R_{m}^2(t, \tau, k, m)}{R_{m}^2(t, \tau, k, m + 1)},
\]

(4.26)
analogous to \( Q_1 \), but calculated using the old embedding. Thus, the wavering product, defined as

\[
W_i(m, \tau) = \left( \prod_{k=1}^{p} Q_1(t, \tau, k, m)Q_2(t, \tau, k, m) \right)^{1/2p},
\]

(4.27)
should be approximately equal to 1 for sufficient reconstructions, as states in such attractor should have the same set of neighbors in both old and new embeddings.

Liebert et al. (1991) defined \( W(m, \tau) = \log \langle W_i(m, \tau) \rangle_s \) as the WP average deviation over a sufficient number of \( s \) reference points, such as 10% of the whole dataset. Finally, they applied \( W(m, \tau)/\tau \) versus \( \tau \) over a range of embedding dimensions, and according to their experiments, the first minimum of each curve provided a good estimation for \( \tau \). In addition, as \( m \) was increased, curves tended to assume a constant value, such that the first dimension before the constant curves should be used as an estimation for \( m \) (and the minimum of this curve should define \( \tau \)). Although curves converge to constant values for greater time delays (allowing the definition of upper bounds for \( \tau \)), no further patterns were found to correlate curves to the embedding dimension \( m \).

### 4.3.2 Fill Factor

Buzug and Pfister (1992a) proposed two methods to find the embedding parameters for a time series: Fill Factor (FF) and Local Integral Deformation (LID). While both rely on the expansion of states, the former is based on the volume of the attractor, whereas the latter on the trajectories evolution\(^2\).

Given a candidate pair \((m, \tau)\), the FF method selects \( m+1 \) phase states \( \phi_i(t_{r_1}), \phi_i(t_{r_2}), \ldots, \phi_i(t_{r_{m+1}}) \), where the subscript \( r_j \) is a randomly selected index \( j \in [1, N_i] \). Then, defining a pivot state, namely \( \phi_i(t_{r_1}) \), the \( m \)-dimensional vector

\[
d_j(t_{r_1}) = \phi_i(t_{r_j}) - \phi_i(t_{r_1}), \ j \in [1, m + 1],
\]

(4.28)

\(^2\) LID is just cited in here, as FF is more robust and well-known. In addition, LID is more complicated to be implemented and computationally expensive.
4.3 Assuming Dependence to Estimate the Embedding Parameters

can be seen as one edge of the hyperparallelepiped described by the
$m \times m$ matrix $M_{m,r_1}(\tau) = (d_1(t_{r_1}), d_2(t_{r_1}), \ldots, d_m(t_{r_1}))$, whose
volume is given by the determinant

$$V_{m,r_1}(\tau) = |\det[M_{m,r_1}(\tau)]|.$$  \hfill (4.29)

To compute an average volume over $s$ parallelepipeds from the
reconstructed attractor (the authors used $s = 2\%$ of states), the
quantity

$$F_m(\tau) = \frac{1}{s} \sum_{k=1}^{s} V_{m,r_k}(\tau)$$

is used, so that the logarithmic fraction of volume

$$f_m(\tau) = \log_{10} F_m(\tau)$$ \hfill (4.31)

is defined as the fill factor. By analyzing the plot $f_m(\tau)$ versus $\tau$,
results empirically led the authors to chose the first non-constant
line as $m$, and the maximum of FF in that dimension as the best
time delay. Despite that Buzug and Pfister (1992a) achieved good
estimations for the Duffing system (Korsch et al., 2008) using FF,
they noticed this method is not robust when applied to attractors
with more than one unstable focus, as the Lorenz system (see result
in Figure 4.8). Based on this last figure, it is clear that neither
patterns nor effective maximum were found to proper estimate the
embedding dimension and the time delay. As those attractors are
not uncommon, FF becomes, at least as proposed, infeasible to be
applied for general systems. Additional drawbacks of FF and LID
are found in (Rosenstein et al., 1994).

![Figure 4.8: Fill Factor applied on the Lorenz system. No patterns were generated by our implementation testing this method, which raises concerns about its reproducibility in all cases. Due to the difference of magnitudes while assessing FF results for different dimensions $m$, values were normalized in range $[m, m + 1]$.](image-url)
4.3.3 C–C Method

Kim et al. (1999) proposed the C–C method to find the optimal time delay $\tau$ and the time delay window $t_w$. Nonetheless, both parameters can also be used to estimate the embedding dimension as well, since $t_w = (m-1)\tau$. The authors were inspired by the BDS statistic (Brock et al., 1992), which can be used to test the null hypothesis that a time series is independently and identically distributed (i.i.d.). Briefly speaking, Brock et al. (1992) noticed that if a $n_i$-length time series $T_i$ was generated from a completely stochastic process, then the reconstructed phase space $\Phi_i$ respects the following power rule

$$C(m, \tau, r, n_i) = C^m(1, \tau, r, n_i), \quad r > 0,$$

where $C(m, \tau, r, n_i)$ is the correlation dimension (some extra arguments were put in evidence, when compared to Equation 2.23) of $\Phi_i$ embedded with $(m, \tau)$, measured with decreasing radius $\epsilon$. Therefore, the C–C method quantifies

$$S(m, \tau, \epsilon, n_i) = C(m, \tau, \epsilon, n_i) - C^m(1, \tau, \epsilon, n_i),$$

which yields zero for infinite i.i.d. time series. However, for finite-length observations collected from natural phenomena, $S(m, \tau, r, n_i) \neq 0$ due to nonlinear correlations. Therefore, the C–C method is concerned to measure dependencies on a time series. In order to eliminate spurious temporal correlations, the approach subdivided $T_i$ in $\tau$ disjoint sub-series: $\{[x(t), x(t+\tau), \cdots], [x(t+1), x(t+1+\tau), \cdots]\}$, and computes

$$S(m, \tau, r, n_i) = \frac{1}{\tau} \sum_{s=1}^{\tau} [C_s(m, \tau, r, \frac{n_i}{\tau}) - C^m_s(1, \tau, r, \frac{n_i}{\tau})],$$

where $C_s$ is the correlation integral for the $s$th sub-series. Then, if $T_i$ is i.i.d., any fixed values of $m$ and $\tau$ respect the relation

$$S(m, \tau, \epsilon) = \frac{1}{\tau} \sum_{s=1}^{\tau} [C_s(m, \tau, \epsilon) - C^m_s(1, \tau, \epsilon)] = 0,$$

for all $\epsilon$.

On the other hand, for general time series, Kim et al. (1999) defined that the best representation of independence among a finite set of observations is either given by the first zero-crossing of $S(m, \tau, r)$ or when $S(m, \tau, r)$ shows the least variation of $\epsilon$, in the form

$$\Delta S(m, \tau) = \max\{S(m, \tau, \epsilon)\} - \min\{S(m, \tau, \epsilon)\}. $$
In addition, the zero-crossing of \( S(m, \tau, \epsilon) \) should be nearly the same for all \( m \) and \( \epsilon \), as well as the minimum of \( \Delta S(m, \tau) \) for all values of \( m \). With this in mind and after defining representative ranges for \( m \in [m_{\text{min}}, m_{\text{max}}] \) and \( \epsilon \in [\epsilon_{\text{min}}, \epsilon_{\text{max}}] \), the averages

\[
\overline{S}(\tau) = \frac{\sum_{m=m_{\text{min}}}^{m_{\text{max}}} \sum_{\epsilon=\epsilon_{\text{min}}}^{\epsilon_{\text{max}}} S(m, \tau, \epsilon)}{(m_{\text{max}} - m_{\text{min}})(\epsilon_{\text{max}} - \epsilon_{\text{min}})},
\]

and

\[
\Delta \overline{S}(\tau) = \frac{\sum_{m=m_{\text{min}}}^{m_{\text{max}}} \Delta S(m, \tau)}{m_{\text{max}} - m_{\text{min}}},
\]

were used to estimate \( \tau \) and \( t_w \), respectively. As \( t_w \) should describe the interval for optimal independence on series observations, and \( \tau \) should be the first locally optimal time, the authors defined \( \tau \) as the first zero-crossing of \( \overline{S}(\tau) \) or the first local minimum of \( \Delta \overline{S}(\tau) \), and \( t_w \) as the time instant when both measures are close to zero, \( i.e., \) the minimum of the quantity

\[
S_{\text{cor}} = \Delta \overline{S}(\tau) + |\overline{S}(\tau)|.
\]

Despite the contribution on proposing a different point of view to find \( \tau \) and \( t_w \), the results presented by Kim et al. (1999) overestimated \( \tau \) and \( m \). For instance, they reported that the C–C method estimated \( m = 8 \) and \( \tau = 18 \) for the Lorenz system, which is known under the presented circumstances to have a representative attractor with \( m = 3 \) and \( d = 8 \). Later, Cai et al. (2008) listed more drawbacks of the C–C method and proposed improvements which led to a new strategy called C–C–1. However, there is, to our knowledge, no evidence that any of the above two methods can handle general purpose systems.

### 4.3.4 Entropy Ratio

Gautama et al. (2003) proposed Entropy Ratio (ER), a method based on minimizing the ratio between the entropy of the phase spaces from the original series and its surrogates\(^4\). The authors realized that a deterministic attractor should have a well-formed structure and, therefore, low entropy. In this scenario, they use the Kozachenko-Leonenko entropy to measure the amount of disorder in the phase space

\[
H(\Phi_i, m, \tau) = \sum_{t=0}^{N_i-1} \ln(N_i R_m^2(t, \tau, 1)) + \ln 2 + 0.5572, \quad (4.40)
\]

\(^3\) In that article, \( \epsilon \) was defined in function of the standard deviation of the time series.

\(^4\) Due to its relation to entropies, the concept of the ER method was the closest to our hypothesis.
where $R^2_{m}(t, \tau, 1)$ is the distance of the $t$th phase space to its nearest neighbors (Equation 4.14).

In attempt to overcome the curse of dimensionality, the method chooses the pair $(m, \tau)$ that minimizes the entropy ratio after superimposing the Minimum Description Length (Rissanen, 1978)

$$R_{\text{ent}}(m, \tau) = I(m, \tau) \left( 1 + \frac{m \ln N_i}{N_i} \right),$$

(4.41)

where

$$I(m, \tau) = \frac{H(\Phi_i, m, \tau)}{\langle H(\Phi_{ij}, m, \tau) \rangle_s},$$

(4.42)

and $\langle \cdot \rangle_s$ is the mean of all the $s$ surrogates of $T_i$, referred to as $T_{ij}, j \in [1, s]$.

Initially creating surrogates based on random permutation of $T_i$, the authors observed that the ER usually led to a time delay equals to one. To improve this result, they applied the iterative Amplitude Adjusted Fourier Transform (Schreiber and Schmitz, 1996) instead, as this method attempts to preserve both the distribution and the spectrum of the original series, as illustrated in Figure 4.9.

However, according to our own implementation of the ER method, the entropy ratio is still prone to estimate $\tau = 1$ even when using the iAAFT.

Figure 4.9: ER creates surrogates using the iAAFT method, which iteratively creates a new series (b) in attempt to preserve the same distribution and spectrum from the original one (a). The example shows the sinusoidal function (Section 3.2.1).

4.3.5 Non-Biased MACF And Gamma Test

Although the $\Gamma$ test (Section 4.2.2.2) was adapted by Otani and Jones (1997) to estimate the optimal embedding dimension $m$, Ma and Han (2006) noticed that it also could be used to validate $\tau$. In this sense, based on the fact that AD (Section 4.2.1.5) can be
4.3 Assuming Dependence to Estimate the Embedding Parameters

applied to reveal correlations on higher-order systems, overcoming restrictions on ACF-based methods, the authors initially used it to have a “guessimation” on the time delay. Nonetheless, the authors later noticed that AD may struggle with inconsistencies while computing the slope of the plateau. To overcome this issue, they proposed the Non-Biased Multiple Autocorrelation Function (NB-MACF)

\[ C^m_{xx}(\tau) = R^m_{xx}(\tau) - (m - 1)\mu_i^2, \]  

(4.43)

where \( \mu_i \) is the mean value of the time series \( T_i \). The algorithm then was divided in three steps based on the interval \( m \in [m_{\text{min}}, m_{\text{max}}] \). Firstly, for each dimension, \( \tau \) is estimated using the NB-MACF. Secondly, the \( \Gamma \) test is used to estimate the corresponding embedding dimension \( m \) for the previously found time delay. Finally, the optimal embedding pair is defined as the first one to reach the minimum value of Equation 4.43. By using this approach, Ma and Han (2006) achieved good results for the Hénon map and the Lorenz system. However, this method still needs further validation on other systems.

4.3.6 Neural Networks

Taking advantage of Multilayer Perceptron (MLP) (Delashmit and Manry, 2005; Haykin, 2009), Karumasinghe and Liong (2006) proposed to train an MLP network in order to find the best model to predict chaotic time series. Due to chaos, they tried to predict an observation \( \rho = 1, 3, 5 \) step(s) in the future, so that \( \hat{x}(t) \) denotes the predicted value for \( x(t) \), given a chosen leap-time \( \rho \). In this context, they implemented \( s \) feed-forward, back-propagation sigmoidal MLPs for each leap time and selected the best pair resulting in the smallest prediction error, as illustrated in Figure 4.10. Denoting by \( \mu_i \) the mean value of the time series \( T_i \) with \( n_i \) observations, they used the Normalized Root-Mean-Square Error (NRMSQ)

\[ \sqrt{\frac{\sum_{t=0}^{n_i-1} (x(t) - \hat{x}(t))^2}{\sum_{t=0}^{n_i-1} (x(t) - \mu_i)^2}} \]  

(4.44)

as loss function in order to train the network. Additionally, they reported the Mean-Absolute Error (MAE)

\[ \frac{\sum_{t=0}^{n_i-1} |x(t) - \hat{x}(t)|}{n_i} \]  

(4.45)

\[ \text{Although ACF measures linear dependencies between } x(t) \text{ and } x(t+\tau), \text{ nothing is inferred on } x(t) \text{ and } x(t+2\tau). \]
in their experiments to rely on both global (NRMSE) and local (MAE) measurements. They search space considered varies the range of embedding dimension $m \in [1, 10]$ with combination of fixed values of $\tau = \{1, 3, 6, 9\}$.

As reported by its authors, this method found $(m = 7, \tau = 6)$ for free-noised observations from the Lorenz system, which yielded to better predictions, (according to the NRMSE and MAE errors) when compared to other methods (Farmer and Sidorowich, 1987; Abarbanel et al., 1993). While sufficient, the embedding dimension is known to be overestimated. In addition, three other (usual) drawbacks of neural networks must be mentioned: i) long computational time required to train MLPs (they varied $\tau$ over a small set of values in attempt to overcome this issue and used $s = 5$); ii) costly and/or delicate hyperparameter tunning, including the number of epochs, error threshold, and adaptive learning rate. In order to mitigate such issues, after some trial-and-error procedure, the MLPs were defined with the following architecture: $m$ units at the input layer, 100 units at the hidden layer, and 1 unit at the output layer; and (iii) no guarantees that forecasting results are due to the embedding parameters or to other reasons. For instance, the output could vary depending on the random initialization of the layers weights.

Later, Bhardwaj et al. (2010) proposed a similar approach for training a neural network, but using Hidden Markov Models (Meyn and Tweedie, 2009) instead. More recently, Manabe and Chakraborty (2007) estimated $m$ and $\tau$ directly from the neural-network architecture after training interactions. Due to its importance and relatively good results (compared to other estimation methods), such article has inspired us when tackling RQ5 (more details in Chapter 9).
4.4 Final Considerations

This chapter provided a detailed and, to our knowledge, comprehensive overview of existing methods designed to estimate the optimal embedding dimension $m$ and time delay $\tau$ to reconstruct phase spaces based on Takens (1981), either assuming they are independent (Section 4.2) or not (Section 4.3) on each other. For completeness, we should mention that a number of additional methods, not mentioned here, exist in the literature. We did not cover those as they are variants of methods already discussed here; or methods which present less validation, or had otherwise lower impact, than the methods discussed here. For the interested reader, we direct further reading to Rosenstein et al. (1994), who discuss additional approaches to estimate $\tau$; and to Otani and Jones (1997), who mention additional strategies to compute $m$. A summary of algorithms to find both parameters (either simultaneously or not) is also found in (Buzug and Pfister, 1992b; Ma and Han, 2006).

From this overview, we can draw several conclusions, as follows.

**Validation difficulty:** It is difficult to define the set of attributes (e.g., homogeneity and distribution of phase states) that generally describes the optimal phase space, especially since this phase space can manifest itself in notorious different ways (as shown in Chapter 3). The computation of metrics based on entropy, correlation dimension or attractor expansion struggle with several inconsistencies related to the partition of the space, computation of probabilities, curse of dimensionality, and the presence of noise. These aspects reinforce the difficulty in defining a set of properties to look at when validating phase spaces. Those issues are the main factors for the lack of robustness in state-of-the-art methods that, despite contributions and good estimations for some datasets, are not adequate to be use in general scenarios. Nonetheless, FNN and AMI are still among the most common methods to estimate $m$ and $\tau$, respectively. These difficulties are one of the main reason why we chose a small set of systems, which are well-understood in the literature, and for which consensus exists on optimal embeddings (described in Chapter 3) to compare our work (described in the next chapters) against.

**Complexity:** This chapter has overviewed different methods to estimate $m$ and $\tau$. All methods are based on extensive sets of heuristics and involve numerous parameters. Some are also computationally expensive. Hence, the quest is still open for designing simple to understand (and use), robust, and fast estimation methods. These requirements are among the main drivers that underlie our work presented in the next chapters.