Bayesian Delay Embeddings for Dynamical Systems

Neil Dhir†∗, Adam R. Kosiorek† and Ingmar Posner
Oxford Robotics Institute
Department of Engineering Science
University of Oxford

Abstract

Selecting a suitable embedding space is a key issue in modelling nonlinear dynamics. In classical phase-space reconstruction, which relies on time-delay vectors, the embedding space is highly dependent on two discrete parameters (for the univariate case), the values of which greatly affect model performance. They also determine the complexity of the dynamics topology. In consequence, the parameters and dynamics are intimately linked. Thus we propose a modelling framework that jointly models the embedding dynamics and parameters, in a Bayesian fashion by framing the learning problem in terms of variational inference over model parameters. We compare our methods to other models on noisy synthetic observations.

1 Introduction

Takens’ theorem provides the theoretical basis for an experimental approach for autoregressive modelling of nonlinear dynamical systems [10]. We will use this strategy as outlined below.

Motivation

Suppose we are faced with some physical system, such as an aircraft engine or the ocean surface. The state of this system is deterministically changing with time and changes according to (generally) smoothly varying Newtonian dynamics. Engines for example do not have abrupt stops or process discontinuities, but are fluently powered up and placidly powered down. Because we assume that the system is deterministic, it is true that the state \( x_t \) is uniquely determined by the state \( x_{t-1} \), where time is discretely indexed by \( t \) [12]. Equivalently there is a smooth diffeomorphism \( \xi : M \to M \) which maps the state \( x_{t-1} \) to \( x_t \). Takens’ theorem (see theorem 1) tells us that if we have time-series measurements \( y_1:T \equiv \{ y_1, \ldots, y_T \} \) of this system, and compose \( n = T - (\tau - 1) \) time-delayed vectors \( x_t = [y_t, y_{t-\tau}, \ldots, y_{t-(d_E-1)\tau}]^\top \) s.t. \( \{ x_t \mid t = 1, \ldots, n \} \) from those observations, these vectors lie on a subset of \( \mathbb{R}^n \) which is an embedding of \( M \) iff \( d_E \geq 2m + 1 \). We now use a nonparametric proxy (the delay vector) for the true system state, which lives in a space of size \( m \).

Hence, if we have sufficient measurements, and the state-space is well covered, it is possible to reconstruct the true dynamics of a system using only a univariate stream of information. We seek principled ways of finding the embedding dimension \( d_E \), the time delay \( \tau \) and the system dynamics \( \xi \). Under Takens’ theorem, it is assumed that these properties and quantities exist [20][25]. Patel & Eng [20] explain, however, that this is an ill-posed inverse problem, since the quality of the observations determine if any solutions are stable, unique or whether they exist at all.

∗Corresponding author: neild@robots.ox.ac.uk
†Equal contribution.
Problem Consider the general, discrete-time, non-linear or linear, Gaussian or non-Gaussian state-space model:

\[
\begin{align*}
    x_1 &\mid \omega \sim \mu_\omega(x_1) & \text{Initial state distribution} \\
    x_t \mid x_{t-1}, \phi &\sim f_\phi(x_t \mid x_{t-1}) & \text{System process} \\
    y_t \mid x_t, \lambda &\sim g_\lambda(y_t \mid x_t) & \text{Observation process}
\end{align*}
\]

where the states and the observed measurements are denoted by \(x_t \in \mathbb{R}^d\) and \(y_t \in \mathbb{R}^D\) respectively, \(\forall t \in T \equiv \{1, \ldots, T\}\) and \(d < D\). The dynamics and the observations are modelled by probability density functions \(f_\phi(\cdot)\) and \(g_\lambda(\cdot)\) respectively, parametrised by \(\phi \in \Phi \subseteq \mathbb{R}^{n_\phi}\) and \(\lambda \in \Lambda \subseteq \mathbb{R}^{n_\lambda}\) respectively. The initial state distribution has parameters \(\omega \in \Omega \subseteq \mathbb{R}^{n_\omega}\).

In this paper we assume that \(f(\cdot)\) and \(g(\cdot)\) are unknown and have to be inferred or nonparametrically approximated. Our problem formulation injects Takens’ embedding theorem and uses it instead of the system model, while and employing the imposed mappings. Despite this assumption, we are still faced with finding the functional form of \(f(\cdot)\) and the parameters of the delay-coordinate vectors.

Solution We propose a solution by way of prediction. In our model, with inspiration from Basharat & Shah [3], under the state-space modelling formalism information is passed as follows:

\[
\begin{align*}
\{y_1, y_2, \ldots, y_t\} \quad &\quad \rightarrow \quad \{y_2, y_3, \ldots, y_{t+1}\} \\
\end{align*}
\]

\[
\begin{align*}
    x_t &= \left[ y_t, \ldots, y_t-(d_E-1)\tau \right]^T & \text{Encoded by VAE} \\
    x_{t+1} &= f(x_t) & \text{\tau, \(d_E\) and \(f(\cdot)\) inferred by VAE} \\
    x_{t+1} &= \left[ y_{t+1}, \ldots, y_{t+1}-(d_E-1)\tau \right]^T & \text{Decoded by VAE}
\end{align*}
\]

We *jointly* learn the functional form of \(f(\cdot)\) and the embedding parameters \(d_E\) and \(\tau\), by employing a novel form of the variational autoencoder [15] which includes the delay coordinate method [25].

![Phase-space reconstruction using the method of delays.](image)

We jointly learn the functional form of \(f(\cdot)\) and the embedding parameters \(d_E\) and \(\tau\), by employing a novel form of the variational autoencoder [15] which includes the delay coordinate method [25].

![Reconstructed phase-space.](image)

Figure 1: Phase-space reconstruction using the method of delays. The \(x\) coordinate from the original Lorenz-63 system [17] (which is three dimensional with coordinates \((x, y, z)\)) is time-delayed, thereby creating phase-space coordinates, with parameters \(\tau = 17\) and \(d_E = 3\).

2 Method

The first step to forecast a chaotic time-series is to employ the history of the observations to reconstruct the state-space [26], which requires finding appropriate values of the embedding dimension and time-delay parameters. The second step builds the forecasting model itself. Our method combines these steps, typically disjoint and performed separately, into a single model.

2.1 Variational autoencoder

We seek to learn an encoder and a decoder for mapping observations \(y = y_{1:T}\) to and from values \(x\) that live in a continuous space. By introducing a parametric inference model of the latent variables
Kingma & Welling [15] showed that $x$ can be interpreted as the latent variables in a probabilistic generative model. They introduce a probabilistic decoder via the likelihood function $p_\theta(y \mid x)$, as well as a probabilistic encoder with the posterior distribution $p(x \mid y) \propto p_\theta(x)p_\theta(y \mid x)$. Where $p_\theta(x)$ is the prior distribution over the latent variables. Kingma & Welling [15] demonstrated that the variational Bayesian approach admits efficient inference by maximising the evidence lower bound (ELBO)

$$\mathcal{L}(y; \theta, \phi) = \mathbb{E}_{q(x \mid y)} \left[ \log p_\theta(y, x) - \log q_\phi(x \mid y) \right] \leq \log p_\theta(y)$$

by simultaneously learning the parameters of the encoder and the approximate posterior distribution $q_\phi(x \mid y)$. If the decoder and the encoder can be computed point-wise, and are both differentiable with respect to their parameters, a monte-carlo approximation of the ELBO can be maximised with gradient descent.

### 2.2 Combining Takens’ embedding theorem with a VAE

The state of a deterministic dynamical system, and consequently its future evolution, is fully specified by a point in phase-space \(21\). Correspondingly, a point in phase-space will always yield the same image in reconstruction space. Theorem \[1\] provides us with a rigorous framework for analysing the information content of non-linear dynamics of the reconstruction space by enriching a measurement $y_t$ with time-shifted copies of itself $y_{t-\tau}$ i.e. the delay coordinates \[6\]. The price (viz. the condition), however, for this result is an exhaustive search over the embedding parameters (we discuss this process in the next section). Currently, to our knowledge, there exists no direct method for learning these parameters from observation, in a principled manner such that the measured space can be accurately and faithfully encoded into and decoded (reconstructed) from the phase-space \[23\], \$9.2].

However, if we treat the embedding task as an autoencoding problem, a vast domain of results becomes available. Further, we are not aware of any methods which approach this inference problem jointly; that is, interleaving inference over embedding parameters and dynamics topology (dynamics function).

In previous work, dynamics have been found using kernel regression \[3\], feed-forward neural networks, recurrent radial basis functions as well as fuzzy models \[22, 2\]. All of these current methods are reliant on empirical selection of $\tau$ and $d_E$ as theorem \[1\] is “silent on the choice of” \[1\], \$3.1] these parameters. There is a large literature on the ‘optimal’ choice of the time delay parameter(s) \[1\]. It turns out that what constitutes the optimal choice largely depends on the application \[1\]. Still, practitioners turn to one metric above all others: average mutual information (AMI). This measure tells us how much information can be learned from a measurement taken at one time compared to a measurement taken at a different time \[1\], \$3.3]. What such linear choice has to do with the nonlinear process relating $y_t$ and $y_{t+\tau}$ is not clear, and hence this choice is one which “we cannot recommend at all” \[1\], \$3.3]. That being said, practically speaking, for some systems it might work, but it is not clear why. Kantz & Schreiber \[14\] take the view that good embeddings are best “found by trial and error”, conditioned on the problem at hand. Further, theorem \[1\] gives us a sufficient condition for integer dimension $d_E > 2m$, which tells us, purely from geometrical considerations, the magnitude of $d_E$ necessary to render a good projection i.e. without phase-space trajectories crossing (i.e. to ‘unfold’ the phase-space). Currently, the false nearest neighbour (FNN) method is the most popular method for finding the $d_E$ given a chaotic system \[1\]. It is a very empirical operation, but the main idea is simple: examine how the number of nearest neighbours of a point along a phase-space orbit changes when varying $d_E$. The FNN, like the AMI, are currently the most common attempts at solving what is generally considered an open problem \[2, 5, 8, 4\].

We structure our model as a VAE with both discrete and continuous latent variables. We learn an approximate posterior distribution over the number of dimensions $d_E$ and the sampling delay $\tau$. Given this approximation, we construct a stochastic embedding function used for encoding observations $y$ into the phase-space $x$. Finally, we employ a multi-layer perceptron (MLP) to predict a time-shifted sequence from its latent encoding. We arrive at an ELBO, whose derivation can be found in appendix \[A\]

$$\mathcal{L}(y; \theta, \phi, \tau, d_E) = \mathbb{E}_{x, \tau, d_E \sim q_\phi(x, \tau, d_E \mid y)} \left[ \log p_\theta(y \mid x, \tau, d_E) \right] - D_{KL} \left( q_\phi(x, \tau, d_E \mid y) \parallel p_\theta(x, \tau, d_E) \right).$$

\[5\] We only consider univariate input sequences, but the multivariate case follows trivially, in which case a different $\tau$ is prescribed for each dimension of the input sequence. For further details see the work by Cao \[7\].
The first term can be seen as a probabilistic analogue of the reconstruction error. The second term pulls the approximate posterior distributions to their priors, thereby acting as a regulariser. Optimising the ELBO requires specifying priors for the latent variables. Since $\tau$ and $d_E$ are both discrete and can be thought of as population statistics, we employ the Poisson distribution. If we do not have prior information about a dynamical system, it is natural to expect that we need many and densely spaced observations to characterise its behaviour. Therefore, we use Poisson distributions with high values for its parameters for $d_E$ and low for $\tau$. The time-delay $\tau$ defines the interval with which we sample and $d_E$ defines the number of intervals we take samples from. We have one additional degree of freedom, namely, which point in the interval given by $\tau$ should we take. We model this as a categorical distribution, and to this end we use Gumbel-Softmax [13], which can be easily used with gradient-based optimisation methods. Sampling of $\tau$ and $d_E$ introduces discontinuities and we employ the NVIL estimator [18] to work around them.

Since we are learning to predict shifted sequences, we are implicitly learning the dynamics $f(\cdot)$ of the phase-space. We encode $y_{1:T}$ to $x_{1:T}$ and reconstruct back to $\hat{x}_{2:T+1}$. Consequently we arrive at a sample with future predictions included, and also obtain the embedding parameters in the process. Provided that these are known, we can again encode the sequence to obtain $x_{T+1}$. We have therefore found a functional relationship from $x_t$ to $x_{t+1}$. The full algorithm can be found in appendix B.

3 Experiments

We consider $N$ corrupted time-series sampled from the Lorenz-63 system [17] – the canonical chaotic system. For results on the easier problem using clean time-series see e.g. [24, 19]. The goal is to accurately predict the dynamics of $x$ in a one-step prediction regime with the addendum that our model has to learn the parameters of the embedding mapping. In other words the VAE has to learn the parametric form of mapping into and out of the phase-space. The parameters for the time-delay vectors are typically assumed known for toy problems and estimated through the AMI and FNN methods – the drawbacks of which were discussed in the previous section. We compare our latent variable model to two other specifications on two data corrupted prediction tasks.

Table 1: Summary of one-step prediction results and comparisons.

<table>
<thead>
<tr>
<th>Latent variable model</th>
<th>$\tau$</th>
<th>$d_E$</th>
<th>RMSE</th>
<th>SNR</th>
<th>$N$</th>
<th>Nonlinear dynamics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kalman-Takens filter† [9]</td>
<td>20</td>
<td>4</td>
<td>$\sim 1.900^*$</td>
<td>$\sim 5$</td>
<td>6000</td>
<td>X</td>
</tr>
<tr>
<td>EKF-RMLP† [10, §4]</td>
<td>4</td>
<td>3</td>
<td>0.378</td>
<td>25</td>
<td>5000</td>
<td>✓</td>
</tr>
<tr>
<td>Our method</td>
<td>3.9*</td>
<td>7.8*</td>
<td>1.078</td>
<td>25</td>
<td>5000</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>3.5*</td>
<td>7.5*</td>
<td>3.270</td>
<td>10</td>
<td>5000</td>
<td>✓</td>
</tr>
</tbody>
</table>

* Expected value: $E[\cdot]$

† The authors do not specify how the embedding parameters were selected.

‡ The authors used a different parametrisation for the Lorenz system, where AMI and FNN were calculated before noise was added.

4 Discussion and conclusion

We do not propose yet another state-of-the-art prediction model, as there are already many impressive systems. Instead, we seek to update the methodology for robustly learning the embedding parameters and the system dynamics from the observations, whilst taking nonlinearities into account. The current, most common, analysis pipeline starts with 1) AMI then 2) FNN and finally 3) phase-space reconstruction and dynamics learning. We combine these steps into one coherent Bayesian model, which introduces dependencies between all the constituent parts.

By leveraging deep-learning with dynamical systems theory, we jointly learn system dynamics and the mapping from phase-space to observed space – by maximising our contributed VAE’s ELBO (derived in appendix A for our model). Whilst our prediction performance (see table 1) is not as competitive as the EKF-RMLP, we can instead point to Bayesian embeddings i.e. we have posterior estimates of our embedding parameters, fully conditioned on the observations. Whilst one-step prediction is an important task, having strong information regarding the topology of the entire phase-space is arguably more valuable, since we are then in a position to deterministically reason about the present, past and future of our system, provided the space has been adequately covered through our measurements [12]. This insight is difficult to glean with the current methods for finding the embedding (see appendix D).
References


A ELBO derivations

Derivation of the expected lower bound (ELBO) of the variational autoencoder infused with Takens’ embedding theorem. This derivation usually takes places under the auspices of a different observation indexing convention, than the one we have used in the paper proper. Hence, for this derivation we shall adhere to the convention $D = \{x^{(n)}\}_{n=1}^N$.

A.1 Standard VAE

![Graphical model of the standard variational autoencoder.](image)

Figure 2: Graphical model of the standard variational autoencoder.

We seek to maximise the marginal likelihood of our observations under our model. Because all observations are taken to be I.I.D. we can express this as

$$\log p_\theta (y^{(1)}, \ldots, y^{(N)}) = \log \sum_{n=1}^{N} p_\theta (y^{(n)})$$  \hspace{1cm} (6)

from whence maximisation is given by

$$\theta^* = \arg \max_{\theta} \sum_{n=1}^{N} p_\theta (y^{(n)})$$

$$= \arg \max_{\theta} \sum_{n=1}^{N} \log p_\theta (y^{(n)}).$$

Consequently our problem boils down to one of finding an expression for $\log p_\theta (y^{(n)})$ under our model. For notational brevity and to reduce clutter, we will remove the observation index $n$ in the derivation.

$$\log p_\theta (y) = E_{x \sim q_\phi(x|y)} [\log p_\theta (y)]$$

Since $p_\theta (y) \perp x$

$$= E_x \left[ \log \frac{p_\theta(y|x)p_\theta(x)}{p_\theta(x|y)} \right]$$

$$= E_x \left[ \log \frac{p_\theta(y|x)p_\theta(x)}{p_\theta(x)q_\phi(x|y)} \right]$$

$$= E_x \left[ \log \frac{q_\phi(x|y)}{p_\theta(x)} \right] + E_x \left[ \log \frac{p_\theta(x|y)}{p_\theta(x)} \right]$$

$$= E_x \left[ \log \frac{p_\theta(y|x)}{p_\theta(x)} \right] - D_{KL} (q_\phi(x|y) \parallel p_\theta(x)) + D_{KL} (q_\phi(x|y) \parallel p_\theta(x)).$$  \hspace{1cm} (7)

Consequently the expected lower bound (ELBO) of the standard VAE is given by $\log p_\theta (y^{(n)}) \geq L(y^{(n)}, \theta, \phi)$.

Then, in the standard model, training amounts to maximising

$$\theta^*, \phi^* = \arg \max_{\theta, \phi} \sum_{n=1}^{N} L(y^{(n)}, \theta, \phi).$$  \hspace{1cm} (8)
We establish some basic probabilistic relations to aid our derivation:

\[ p(y, x, \tau, d_E) = p(x, \tau, d_E \mid y)p(y) \iff p(y) = \frac{p(y, x, \tau, d_E)}{p(x, \tau, d_E \mid y)} \]

That is the basic machination required for a standard VAE. In our model the ELBO takes a different form but follows the same approach as that in eq. (7).

### A.2 ELBO derivation for VAE combined with the method of delays

First let the phase-space coordinates be defined as so

\[ x = \Phi(y, \tau, d_E). \] (9)

We establish some basic probabilistic relations to aid our derivation:

\[ E = \log p(y, x, \tau, d_E \mid y) \]

As in appendix A.1, we seek to maximise the marginal likelihood of our observations, w.r.t. to the model parameters. We have omitted subscripts and superscripts where their nature is obvious, for brevity and to reduce clutter. We proceed as before by letting

\[
\log p_\theta(y) = \log p_\theta(y) \iint q_\phi(\tau, d_E, x \mid y) \, dx \, d\tau \, dd_E \quad \text{Recall that } p(x \mid \tau, d_E, y) = \delta(x - \Phi(y, \tau, d_E))
\]

\[
= \log p_\theta(y) \iint q_\phi(\tau, d_E \mid y) \, d\tau \, dd_E
\]

\[
= \int \int \int q_\phi(\tau, d_E \mid y) \log p_\theta(y) \frac{q_\phi(\tau, d_E \mid y)}{q_\phi(\tau, d_E \mid y)} \, d\tau \, dd_E \, dy \quad \text{Use eq. (9)}
\]

\[
= \int q_\phi(\tau, d_E \mid y) \left[ \log \frac{p(y, x, \tau, d_E)}{p(x, \tau, d_E \mid y) q_\phi(\tau, d_E \mid y)} \right]_{x = \Phi(y, \tau, d_E)} \, dy
\]

\[
= \int q_\phi(\tau, d_E \mid y) \left[ \log \frac{p(y, x, \tau, d_E)}{q_\phi(\tau, d_E \mid y)} - \log q_\phi(\tau, d_E \mid y) \right]_{x = \Phi(y, \tau, d_E)} \, dy
\]

\[
= \int q_\phi(\tau, d_E \mid y) \left[ \log \frac{p(y, x, \tau, d_E)}{q_\phi(\tau, d_E \mid y)} \right]_{x = \Phi(y, \tau, d_E)} \, dy - \int q_\phi(\tau, d_E \mid y) \left[ \log \frac{q_\phi(\tau, d_E \mid y)}{p(x, \tau, d_E \mid y)} \right]_{x = \Phi(y, \tau, d_E)} \, dy
\]

\[
= \mathbb{E}_{\tau, d_E \sim q_\phi(\tau, d_E \mid y)} \left[ \log \frac{p_\theta(y, x, \tau, d_E)}{q_\phi(\tau, d_E \mid y)} \right]_{x = \Phi(y, \tau, d_E)} \]

\[
\geq 0
\]

Having arrived at an expression for the ELBO, we can now factorise it further, to arrive at an expression which we can reduce via stochastic gradient descent methods. We omit ELBO subscripts for brevity:

\[
\mathbb{E} \left[ \log \frac{p_\theta(y, x, \tau, d_E)}{q_\phi(\tau, d_E \mid y)} \right] = \iint \int q_\phi(\tau, d_E \mid y) \log \frac{p_\theta(y, x, \tau, d_E)}{q_\phi(\tau, d_E \mid y)} \, dx \, d\tau \, dd_E
\]

\[
= \iint \int q_\phi(\tau, d_E \mid y) \log p_\theta(y, x, \tau, d_E \mid x) \mid_{x = \Phi} \, d\tau \, dd_E
\]

\[
= \iint \int q_\phi(\tau, d_E \mid y) \log p_\theta(y, \tau, d_E \mid x) \mid_{x = \Phi} \, d\tau \, dd_E
\]

\[
= \iint \int q_\phi(\tau, d_E \mid y) \log p_\theta(y, \tau, d_E, x) \mid_{x = \Phi} \, d\tau \, dd_E
\]

\[
+ \iint \int q_\phi(\tau, d_E \mid y) \log p_\theta(\tau, d_E) \, d\tau \, dd_E
\]

\[
= \int \int q_\phi(\tau, d_E \mid y) \log q_\phi(\tau, d_E \mid y) \, d\tau \, dd_E
\]

\[
= \mathbb{E}_{\tau, d_E \sim q_\phi(\tau, d_E \mid y)} \left[ \log p_\theta(y \mid x, \tau, d_E) \mid_{x = \Phi} \right]
\]

\[
- \text{KL} (q_\phi(\tau, d_E \mid y) \mid \mid p_\theta(\tau, d_E)) \mid_{x = \Phi}
\]

(11)

(12)
We arrive at the final expression, which is given by

$$L(y; \theta, \phi, \tau, d_E) = E_{x, \tau, d_E \sim q_{\phi}(x, \tau, d_E \mid y)} \left[ \log p_{\theta}(y \mid x, \tau, d_E) \right] - D_{\text{KL}}(q_{\phi}(x, \tau, d_E \mid y) \mid \mid p_{\theta}(x, \tau, d_E)) .$$  

(13)
Algorithm 1: Bayesian Delay Embeddings

Input: Slices $y_{1:t} \sim y_{1:T}$ of length $t$ of observations.

for $i = 1, 2, \ldots$ do
  // Encoder
  $\tau \sim \text{Poisson}(\theta_1)$ and $d_E \sim \text{Poisson}(\theta_2)$
  $\Phi \leftarrow \Phi(f, \tau, d_E)$ \hspace{1cm} \triangleright Forward transform as a sample from Gumbel-Softmax.
  $x_{1:t-(\tau-1)} \leftarrow \Phi^T y_{1:t}$ \hspace{1cm} \triangleright Encoded phase-space orbit as a matrix product.
  // Decoder
  $\mu_{2:t+1} \leftarrow \text{MLP}(x_{1:t})$ \hspace{1cm} \triangleright Multilayered perceptron
  $\tilde{y}_{2:t+1} \sim \mathcal{N}(\mu_{2:t+1}, \hat{\sigma}^2)$
  $\mathcal{L}(x_{1:t}; \theta, \phi) = \log \mathcal{N}(x_{2:t}; \mu_{2:t}, \hat{\sigma}^2) - D_{KL}$
  $\Psi = \Psi + \alpha \nabla_\Psi \mathcal{L}(y_{1:t}; \theta, \phi)$ \hspace{1cm} \triangleright Update model parameters to maximise ELBO.

Output: $\tau, d_E, \tilde{y}_{2:t+1}, \mathcal{L}(y_{1:t}; \theta, \phi)$
C Raw measurements

(a) Full coordinate measurements from the Lorenz-63 [17] system.

(b) Noisy observations of the x-coordinate of the system, with SNR: 25dB.

(c) Noisy observations of the x-coordinate of the system, with SNR: 10dB.

Figure 3: The full Lorenz system is depicted in fig. 3a followed by single-coordinate observations with noise in the following plots in fig. 3b and fig. 3c respectively. The vertical red bars mark the train-test cut-off points. The observations to the left of the bar are used for training.
D Average mutual information and false nearest neighbours

![Graph](image)

(a) Average mutual information.  
(b) False nearest neighbours.

Figure 4: Average mutual information (fig. 4a) and false nearest neighbours (fig. 4b) calculated for both corrupted synthetic datasets. When the SNR was set to 25dB, then \((\tau, d_E) = (20, 5)\) and for 10dB \((\tau, d_E) = (24, -)\). For the latter case, the signal was found to be too corrupted for a dimension to be found within the given search range \((d_E \in [0, 30] \text{ was unsuccessfully explored as well}) – for this method.
E  Taken’s embedding theorem

Taken’s theorem (theorem 1), though the most celebrated, is not the only method for embedding scalar observations in a multivariate dynamical spaces [1]. But, it is the only systematic schema to do so (see also the works by [21, 16]). It tells us that information about the phase-space of a dynamical system can be preserved in a time-series output. In and of itself, it forms a bridge between the theory of non-linear dynamical systems and the analysis of experimental time-series [12]. We state the theorem for completeness and refer the reader to [25] for a formal proof:

**Theorem 1** (Takens’ embedding theorem [25, Theorem 1]). Let \( M \) be a compact manifold of dimension \( m \). For pairs \((\xi, y)\), where \( \xi : M \rightarrow M \) is a smooth diffeomorphism and \( y : M \rightarrow \mathbb{R} \) a smooth function, it is a generic property that the \((2m + 1)\)–delay observation map \( \Phi_{(\xi, y)} : M \rightarrow \mathbb{R}^{2m+1} \) given by

\[
\Phi_{(\xi, y)}(x) = (y(x), y \circ \phi(x), \ldots, y \circ \xi^{2m}(x))
\]

is an embedding; by ‘smooth’ we mean at least \( C^2 \).

Where the set of pairs \((\xi, y)\) ∈ \( C^2(M, M) \times C^2(M, \mathbb{R}) \). Then, applied to our domain, theorem 1 provides us with a delay coordinate-map by stacking \( d_E \) previous entries\(^4\) of a uniformly sampled time-series, such that

\[
\Phi_{(\xi, \psi)}(x) : x \mapsto [\psi(x), \psi \circ \xi(x), \ldots, \psi \circ \xi^{d-1}(x)]^T
\]

explicitly

\[
\Phi \circ (x(n)) = [x(n), x(n - \tau), \ldots, x(n - (d - 1)\tau)]^T = [\psi \circ x(n), \psi \circ \xi(x(n)), \ldots, \psi \circ \xi^{d-1}(x(n))]^T
\]

where we have dropped the subscripts on \( \Phi \) for brevity. Alas, theorem 1 gives us a lower bound for the cardinality of \( d_E \). Specifically if \( d_E > 2m + 1 \) (where \( m \) can also be seen as the “true” size of the attractor space), then the attractor, as seen in the space with lagged coordinates, will be smoothly related to the attractor as viewed in the original physical coordinates [1, §I.V]. Sauer et al. [21] showed that this was but a sufficient but not necessary condition. The attractor could, it was shown, be reconstructed with a dimension as low as \( m + 1 \) as explained by Patel & Eng [20].

Practically, if we set \( d_E \) large enough, physical properties of the attractor we wish to extract from the measurements, will be the same when computed on the representation in lagged coordinates and when computed in the physical coordinates [1 §I.V]. Thus, if we can find this deterministic mapping from time-series data, we can also predict the future of the system. For a thorough discussion on the consequences of theorem 1 see Huke [12 §5]. This means that for a large class of observation functions \( \psi \), \( \Phi \) will preserve the topology of \( M \), consequently information about \( M \) can be retained in the time-series output.

\(^4\)An invertible function that maps one differentiable manifold to another such that both the function and its inverse are smooth.

\(^5\)In theory we could also use future entries but for the sake of causality, we only consider past entries.