Model Selection for Detection of Directional Coupling from Time Series

Vejmelka, Martin
2008

Model Selection for Detection of Directional Coupling from Time Series

Post-Graduate Student:
ING. MARTIN VEJMELKA
Institute of Computer Science of the ASCR, v. v. i.
Pod Vodárenskou věží 2
182 07 Prague, Czech Republic
vejmelka@cs.cas.cz

Supervisor:
RNDR. MILAN PALUŠ, DRSC.
Institute of Computer Science of the ASCR, v. v. i.
Pod Vodárenskou věží 2
182 07 Prague, Czech Republic
mp@cs.cas.cz

Field of Study:
Biocybernetics and Artificial Intelligence

This work was supported by the EC FP6 NEST initiative project BRACCIA.

Abstract

This paper deals with the problem of selecting the conditioning model in the estimation of conditional mutual information in the context of detecting directional influence from raw time series. An approach similar to model selection in model fitting to time series is presented. A numerical study illuminating the problem and showing the effectivity of the proposed procedure is summarized at the end of the paper.

1. Introduction

The discipline of nonlinear dynamics has proven fruitful as many problems from meteorology [1, 2], geology [2], life sciences [3] and physics have been more satisfactorily understood in this framework. Time series analysis is a frequent tool used to process the activity records of dynamical oscillatory processes. Methods have been developed to detect various forms of synchronization and directional coupling from time series. The detection of directional influence is an important method of examining drive-response relationships in complex dynamical systems. Paluš [4, 5] has advocated the use of the conditional mutual information functional $I(X;Y_T|Y)$ between the two time series as a measure of "net information flow" between the process $\mathcal{X}$ and the process $\mathcal{Y}$ at some point of time in the future. Conditional mutual information has been applied in the context of phase dynamics to phase time series which simplify the analysis of signals [5, 6]. In this work the problem of discovering the directionality of coupling in amplitude time series is investigated and a method to solve one of the problems is presented.

The conditional mutual information can be decomposed into several terms which are interpretable in the context of time series analysis of nonlinear dynamical systems

$$I(X;Y_T|Y) = I(X;Y_T;Y) - I(X;Y) - I(Y;Y_T), \quad (1)$$

where X and Y are the time series of the processes $\mathcal{X}$ and $\mathcal{Y}$ respectively and Y_T is the time series of the process $\mathcal{Y}_T$, which is the process $\mathcal{Y}$ shifted by T samples into the future.

The term $I(X;Y_T;Y)$ of (1) represents the total common information in all the processes $\mathcal{X}$, $\mathcal{Y}$ and $\mathcal{Y}_T$.

The term $I(X,Y)$ represents the effect of common dynamics and common history. Common history can be brought about by the same noise or external influences on the two processes. If the two processes have narrowband spectra with close peaks, then their time series may have some common parameters (e.g. the period of oscillation), this increases the amount of mutual information in the first term and must be subtracted. If additionally the dynamics themselves, which are represented by the equations in case of models, share some common traits or the entire form then this may cause similar amplitude distributions. None of the above effects is brought about by the influence of directional coupling. It is therefore important to subtract these components from the term $I(X;Y_T;Y)$ to ensure that they are excluded from the estimation of "net information flow". We note here that the mutual information $I(X,Y)$ can be used to detect synchronization of the investigated processes.

The term $I(Y;Y_T)$ represents the action of the process upon itself and is connected to the predictability of the process. It is imperative that this term is estimated well and removed from the total common information. Underestimation of this term will result in false positive detections as strong action of the process $\mathcal{Y}$ upon itself

will be misinterpreted as directional influence from the process $\mathcal{X}$. Effective estimation of this term is crucial to the correct application of the framework for detecting directional influence and will be the goal of this work.

The variables X, Y represent the time series of the given processes and may in general be multidimensional. Multidimensional time series can be either directly measured by observing several aspects of the activity of a dynamical process or can be constructed from a single time series by means of an embedding technique. A frequently used embedding technique is that of time-delay embedding [7, 8], where equidistantly spaced samples of a given time series are used to construct a vector

$$\begin{aligned} \bar{x}(t) &= (x_1(t), x_2(t), ..., x_K(t)) \\ &= (x(t), x(t-\tau), ..., x(t-(K-1)\tau)), \end{aligned} \tag{2}$$

where $x(t)$ is the scalar time series of the activity of process $\mathcal{X}$ and τ is the *delay* between successive samples and K is the embedding dimension.

An important parameter is the number of samples the process $\mathcal{Y}$ is shifted into the future. In our previous work [9, 10] conditional mutual information is averaged for shifts T from 1 up to two periods of the faster process in the investigated system pair. For model systems or systems with simple structure improvements to this scheme are possible as there are clear patterns in the conditional mutual information with respect to the time shift. The estimation method used is equiquantal binning as it has shown the best properties in model tests and has been successfully applied to some experimental datasets [10, 6, 2].

1.1. The intersample delay

There are multiple established techniques for selecting the time delay τ to construct a vector representation of the state of a dynamical system from a univariate time series [8, 7, 11]. Kantz and Schreiber [12] have however argued that there is no optimal way of selecting the time delay in general. Rather the specific purpose with which the embedding is constructed allows one to discuss and gauge the optimality of an embedding method. The use of an intersample delay is a way to circumvent the problem of selecting samples that are highly correlated and thus as a set contain a lower amount of information about the structure of the system in state space [12]. The classical procedure requires the delay to be fixed first and then using another method the dimension is fixed by testing if adding more dimensions to the vector is reasonable [13]. This is simple because samples are considered sequentially. However we know of no apriori reason to restrict the selection procedure in this way.

It is important to produce a model which fits the dynamics of the time series as well as possible in a sense that will be described later. Selecting the delay greater than 1 in effect pre-filters the samples that can be included in the model. If the intersample delay is say $\tau = 2$ then only the time series samples $x(t-d), d \in \{2, 4, 6, 8, ...\}$ may be considered for inclusion in the conditioning model. Since the model search procedure is time intensive, it is advantageous to apriori restrict the set of possible delays for performance purposes. Additionally, a model utilizing samples close to each other will end up modeling the temporal structure of the time series instead of the geometrical structure of the attractor in the state space. However these are not rigorous arguments and counterexamples may be found where the optimal selected model contains samples close together.

The most frequently used method of selecting the intersample delay is to select the first minimum of the lagged mutual information $I(Y; Y_T)$ where T is the lag in samples between the original and shifted time series of the process [14]. This is the procedure that will be henceforth used to select an intersample delay. There have been many other suggestions in the literature (an overview can be found in [12]) but all of the suggested methods are based on heuristic arguments. Time lagged mutual information has been applied and found to work well in many practical settings although caution is advised as the first minimum may be spurious.

2. The model selection procedure

The purpose of this work is to select a proper vector representation of the process $\mathcal{Y}$ which enables a good estimation of the term $I(Y; Y_T)$ in (1) as explained in the Introduction. A good model is a model that maximizes the *expected* lagged mutual information $I(Y; Y_\tau)$, where τ is the intersample delay selected according to the method in the last paragraph. There are two reasons for this choice: a single lag is necessary because of the computational costs of computing the full, say 50, estimates and averaging them. Secondly, selecting too small a lag will result in temporal correlations guiding the selection and a lag too large will attenuate the deterministic structure between the lagged process and the original process. Because real dynamical processes are affected by external influences and usually are encumbered by noise, this means that the effects of the auto-structure of the process are attenuated for larger distances.

2.1. Model specification and criterion

Formally, each model M is completely specified by the indices of the samples used in constructing the state space vector $\bar{y}(t)$ as

$$\begin{aligned} M &= \{i_1, i_2, ..., i_K\} \text{ implies that} \\ \bar{y}(t) &= (y(t - i_1\tau), y(t - i_2\tau), ..., y(t - i_K\tau)), \end{aligned} \tag{3}$$

where K is the number of samples in the vector and depends on M. We will denote by Y_M the state space representation of the process Y using the vector specified by M. Then the best model M^* has the property

$$M^* = \text{argmax}_M \mathcal{E}[I(Y_M; Y_\tau)] \tag{4}$$

It is important to maximize the expectation of the mutual information over entire reconstructed space because the in-sample estimate would always increase if more samples were added to an existing state-space vector. This phenomenon is known as overfitting in the pattern recognition community. The problem can be converted to a problem of minimizing the conditional entropy

$$M^* = \text{argmin}_M \mathcal{E}[H(Y_\tau|Y_M)], \tag{5}$$

as $H(Y_\tau|Y_M) = I(Y_M; Y_\tau) + H(Y_\tau)$ and $H(Y_\tau)$ is a constant with respect to the optimization problem. In fact, due to the use of the equiquantal estimator the marginal entropy $H(Y_\tau) = H(Y) = \log B$. As usual, we assume the underlying processes to be ergodic for the duration of the analysis time window and this allows us to substitute expected values over time for expected values over the state space.

Any admissible model can be expressed as

$$M = (i_1, i_2, ..., i_K), \tag{6}$$

for $0 = i_1 < i_j < i_{j+1} \leq L, j \in \{2, .., K\}$ where L is some pre-selected maximum distance to the farthest considered sample and $K < K_{\max}$ is the number of elements in the model. It is important that $i_1 = 0$ is always included in the model because otherwise the random variables X and Y in term $I(X; Y)$ in (1) would not be taken at the same instant of time and would thus not represent the common history of the two processes. This would give the computed conditional mutual information different semantics and it would not reflect the "net information flow". This is not a significant restriction for dynamical systems because the action of noise, external influences and other factors causes the process to produce new information continuously and "forget" its initial conditions thus rendering samples further back in time less useful for constructing models. The threshold also limited by computational constraints and the number of models. The maximum size of the model $K_{\max}$ is also limited by computational constraints as the size of the model set grows combinatorially. A more important limit is the length of the time series itself which affects the maximum size K of the model M which can be reliably estimated. This however happens automatically during the estimation process as models with too many free parameters with respect to the length of the time series will be poorly estimated and the expected value of the conditional entropy will be high.

2.2. Conditional entropy and classification

It remains to show how the expectation of the criterion $[H(Y_\tau|Y_M)]$ can be computed for a given model M. First, given the number of bins B, the samples of the investigated time series are discretized using the equiquantal scheme into the B levels. The model specification M is then used to construct pairs

$$(\bar{y}_M(t), y(t + \tau)), \tag{7}$$

where the indices building the vector $\bar{y}^i_M(t)$ will be selected according to the model specification M. As the time when the training pair occurs in the time series will not be important, we will abbreviate the notation of the state vector $\bar{y}^i_M$ and the (predicted) future value to y^i_τ. When denoting the variable rather than a particular value, the index i will be omitted. The training pairs will be used to construct a classifier which will attempt to model the probability distribution function (PDF) of the state space of the underlying process. The classifier will be a simple multidimensional histogram which will aggregate all the training samples in its estimate of the PDF. The goal of the classifier is to predict the future state y^i_τ from the given vector $\bar{y}^i_M$. This process might seem crude but the key point is that in the estimation of the conditional mutual information functional (1), all the terms are estimated in exactly the same way. It follows that any problems that the classification process will have in estimating the PDF correctly are also expected in the estimation of CMI. It would thus not be useful to use a different classification scheme here because the model fitting procedure would yield a model which would not respect the advantages and disadvantages of this particular estimator and could potentially have a completely different number of free parameters.

It will now be shown that choosing a suitable loss function results in the error rate to be an estimate of the required criterion (conditional entropy)

$$L(y^i_\tau, \bar{y}^i_M) = -\log p(y^i_\tau|\bar{y}^i_M), \tag{8}$$

where the conditional entropy $p(y_\tau|\bar{y}_M)$ is unknown. We must substitute an estimate of the conditional

probability computed as

$$\hat{p}(y_\tau^i|\bar{y}_M^i) = \frac{N(y_\tau^i, \bar{y}_M^i)}{\sum_{y'_\tau} N(y'_\tau, \bar{y}_M^i)}, \tag{9}$$

where $N(\cdot,\cdot)$ is the number of occurrences of the pair in the training set. As the pair $(\bar{y}_M^i, y_\tau^i)$ is expected to be seen in a long sequence with probability $p(\bar{y}_M^i, y_\tau^i)$, the expected mean error over the state space will be

$$\begin{aligned} &- \sum_{(y_\tau, \bar{y}_M)} p(y_\tau, \bar{y}_M) \log \hat{p}(y_\tau|\bar{y}_M) \\ &- \sum_{(y_\tau, \bar{y}_M)} p(\bar{y}_M) p(y_\tau|\bar{y}_M) \log \hat{p}(y_\tau|\bar{y}_M) \\ &= \hat{H}(Y_\tau|Y_M) \end{aligned} \tag{10}$$

To further understand this result, let us relate it to the expected error assuming we would know the true distribution $p(y_\tau, \bar{y}_M)$:

$$\begin{aligned} &\hat{H}(Y_\tau|Y_M) - H(Y_\tau|Y_M) = \\ = &- \sum_{(y_\tau, \bar{y}_M)} p(\bar{y}_M) p(y_\tau|\bar{y}_M) \log \hat{p}(y_\tau|\bar{y}_M) + \\ + &\sum_{(y_\tau, \bar{y}_M)} p(\bar{y}_M) p(y_\tau|\bar{y}_M) \log p(y_\tau|\bar{y}_M) = \\ = &\; E_{Y_M} D(\hat{p}(y_\tau|\bar{y}_M) || p(y_\tau|\bar{y}_M)). \end{aligned} \tag{11}$$

The result shows that the expected error is equal to the value of the optimal expected error (conditional entropy) if the probability density function was known and the mean Kullback-Leibler divergence between the estimated and actual conditional probability density over all the states. It is clear that the conditional entropy is always overestimated. It is also clear that if the model contains a higher amount of free parameters (total histogram bins), the K-L divergence will increase as the estimate of the conditional probability density will be poorer and the bias will increase. This is behavior is favorable as it penalizes overfitting of the model.

Practically this procedure still has some unresolved problems. If a previously unseen pair $(y_\tau|\bar{y}_M)$ is encountered during the estimation of the criterion, the estimated conditional probability would be $\hat{p}(y_\tau|\bar{y}_M) = 0$ or undefined. The same would occur when a leave-one-out procedure is applied and the training pair exists only once in the training set. A regularization procedure is needed to deal with these pairs. Since the conditional probability estimate is computed from the accumulated histogram using (9). To resolve this a fixed term Δ is substituted for the unknown conditional probabilty in the loss function

$$L^*(y_\tau^i, \bar{y}_M^i) = \begin{cases} -\log \hat{p}(y_\tau^i, \bar{y}_M^i) & \text{if } N(y_\tau^i, \bar{y}_M^i) > 0 \\ -\log \Delta & \text{otherwise} \end{cases} \tag{12}$$

Obviously if no previously unseen states are encountered, the modified loss function gives identical results to the original loss function. When optimizing the model, we have elected to set $\Delta = \frac{1}{B}$, where B is the number of bins. This has the simple rationale that when the particular vector $\bar{y}_M^i(t)$ has not been seen in at all, then equal probability is assigned to all the possible future states y_τ.

Due to the form of the loss function, the same penalty is also assigned if the vector $\bar{y}_M^i(t)$ has been previously seen but not together with the given future state y_τ^i. In this case it is unclear whether $\frac{1}{B}$ is the best choice but no plausible argument has been found that would advocate selecting a different value for this situation.

A complete method for selecting a model for conditioning the CMI (1) from a given time series has now been constructed. The method connects a classification problem to the required criterion by using a suitably constructed loss function which is regularized for practical purposes.

We note here that there are many established methods for model selection in time series analysis (and elsewhere) such as the MDL principle [15], the Bayesian information criterion [16] or the Akaike information criterion [17]. These selection mechanisms however do not optimize the required criterion. These methods additionally assume a particular distribution family of the probability density function of the samples or the estimation of a likelihood function.

2.3. Including surrogates

It has been previously explained that the goal of the selection of the conditioning model was to be able to correctly determine directionality of coupling in as many cases as possible. To understand the influence of the surrogate time series on the usefulness of a particular conditioning model, it is necessary to recount the method of statistical testing of the estimates of conditional mutual information.

At the core of the directionality detection method is the estimation of conditional mutual information (1) for different lags T. These values are averaged over the selected lags to construct an index of directionality. This index reacts to an increase in coupling by increasing its value. However any directionality index also reacts to a change in other factors involving the underlying systems and the time series: noise levels, main frequencies, external influences on the systems and others. The *inverse problem* of determining directional influence is much more difficult: given a value of the index, can we infer that directional coupling exists ?

Surrogate testing is a method of verifying if there is sufficient evidence available to infer that directional coupling is present in a particular direction. The method is a simple one-sided hypothesis test with the null hypothesis of no directional coupling. The distribution of the index under the null hypothesis can be estimated by evaluating the index on as many surrogate time series as is deemed necessary and is computationally feasible. Usually 100 or 200 surrogates are used if the analysis is being performed offline. Surrogate time series are time series which preserve all of the properties of the original time series except the property being tested. Here, directional coupling is the tested property and surrogate time series are such time series that preserve the dynamical structure of the individual underlying processes but do not preserve the effect that coupling has on the time series. This is done by somewhat altering the temporal structure of the time series so that cause and effect of the coupling are separated and mixed in the time series. Common procedures which more or less accomplish this goal include Fourier transform surrogates [18], permutation surrogates [19], amplitude adjusted Fourier transform surrogates [20] or twin surrogates [21]. Each procedure is applicable in different situations and has it's advantages and disadvantages [10]. If a model of the underlying system is available, surrogate time series can be simply generated using the model by creating two pairs of time series of the coupled models and than taking the first time series from the first pair and the second time series from the second pair, these surrogates are called *equation-based surrogates*. These surrogates have the ideal properties and can be used as a standard against which other surrogate generation schemes are compared.

It is important to note that the hypothesis test is performed as if the surrogates had the ideal properties listed above. This is however only an approximation as the surrogate generation algorithm always destroys some of the dynamical structure it its random phase. This is a critical point for the model selection procedure.

Let this state of affairs now be related to the model selection procedure. Ideally when selecting a model, there would be enough data points in the source time series so that the set of data can be split into a *training* and *testing* set. The training set would be used to construct the models and the testing set would be used to obtain an unbiased estimate of the expectation of the criterion (4) for a given model. Assuming that models of the dynamical systems are available as much testing data as needed could be generated (this testing data would in fact be equivalent to the *equation-based surrogates*. This would seem to be fortuitous but in practice it is rarely the case that models of the underlying systems are available as the most interesting applications of the nonlinear dynamical framework are in areas where the physics of the analyzed systems is still poorly understood. If equation-based surrogates were available there would be no bias in the distribution under the null hypothesis stemming from the difference in the dynamics in the original and surrogate time series. In this case the conditioning model that would be optimal with respect to criterion (4) would also be optimal for use in the surrogate time series as they are for all practical purposes identical to the original time series. A leave-one-out procedure on the training set from the original time series would the suffice to select the best useable model.

In practice one of the above algorithms which does not need the underlying model is used to generate surrogates which are not identical in dynamical structure to the original time series. A possible exception to this rule are the twin surrogates which are difficult to apply in practice but do well in the preservation of the dynamical structure. Training and testing the model using a leave-one-out scheme would thus yield a model which is not the best possible for the evaluation of (1) as this model would not take into account the deformation of the dynamical structure due to the use of the surrogate generation algorithm. This is one of the most important practial caveats in the application of the above method for selecting conditioning models. It follows that creating the model on the original time series and testing the model (computing the criterion value) on the surrogates is what is required to obtain the best conditioning model. It has been found that the models selected using this procedure have less elements than those selected using a leave-one-out scheme. This is due to the fact that more complex models are more sensitive to the partial modification of the dynamical structure due to the surrogate generation algorithms. If the surrogates would have a dynamical structure identical to the original time series, then this procedure would be exactly the same as would be applied in a standard pattern recognition problem with a training and testing set.

2.4. The final procedure

The entire procedure for model selection can thus be summarized as:

- Input: time series with N points, no. of bins B, maximum model size $K_{\max}$, most distant sample L
- Compute intersample delay τ
- Generate r surrogate time series for testing

- For each possible model M:
- Build the histogram estimate $\hat{p}(y_\tau|\bar{y}_M)$ on the original time series
- Estimate the expected conditional entropy on the r surrogate time series and average the result: this is the criterion value
- Select the model M^* with the smallest criterion value

The more surrogates are used, the better will be the estimated conditional entropy. The generation of surrogates is usually fast for most surrogate generation algorithms but the estimation of the expected conditional entropy is expensive for long time series and must be repeated for each model of which there are $\binom{L-1}{K-1}$ as $i_1 = 0$ is always part of the model.

3. Numerical studies

In this section the effectivity of the presented procedure for selecting conditioning models will be shown on model systems the parameters and structure of which are known.

3.1. Rössler systems

In the first example, we will work with the famous Rössler system pair:

$$\begin{aligned} \dot{x}_{1,2} &= -\omega_{1,2}y_{1,2} - z_{1,2} + \epsilon_{1,2}(x_{2,1} - x_{1,2}) \\ \dot{y}_{1,2} &= \omega_{1,2}x_{1,2} + a_{1,2}y_{1,2} \\ \dot{z}_{1,2} &= b_{1,2} + z_{1,2}(x_{1,2} - c_{1,2}), \end{aligned} \tag{13}$$

where $a_{1,2} = 0.15$, $b_{1,2} = 0.2$, $c_{1,2} = 10$, $\omega_{1,2} = 1 \pm 0.015$ and $\epsilon_{1,2}$ is the coupling between the systems. The systems are integrated using a Runge-Kutta 4th order scheme with $dt = 0.05$ and the resulting time series is subsampled by a factor of 6 to yield 20 points per period of the system. Conditional mutual information (1) is computed for lags $T \in \{1, .., 50\}$ and averaged. The number of bins was set to 8 which is a value that works well for many systems [9, 10].

Fig. 1 shows the resulting curves of conditional mutual information against coupling strength for different selected models for the length of time series 32768 samples. The coupling strength $\epsilon_1 = 0$ while ϵ_2 was varied between $\{0, 0.2\}$. Such a long time series allows even CMI estimates with 3 elements in the conditioning model to be computed and thus negates any advantage a simpler model might have due to insufficient data. The intersample delay was set to $\tau = 5$. At the top, the model $M_0 = \{0\}$ was applied. It is clearly seen here, that a single condition is not sufficient as the CMI curve for the reverse direction is not constant but increases considerably towards $\epsilon_2 = 0.08$. In the middle the model $M* = \{0, 1\}$ was the result of the above optimization procedure. The bottom row is the model $M_L = \{0, 1, 7\}$ which was selected by using a leave-one-out estimation method without using the surrogate time series to test the model. The larger model M_L does not bring any improvement over model M^* recommended by the model selection procedure. The curve in the direction of coupling reacts to the coupling just as well as the more complicated model. In the reverse direction, the conditional mutual information is constant and close to 0 until the generalized synchronization threshold is reached. This is the desired behavior of the index.

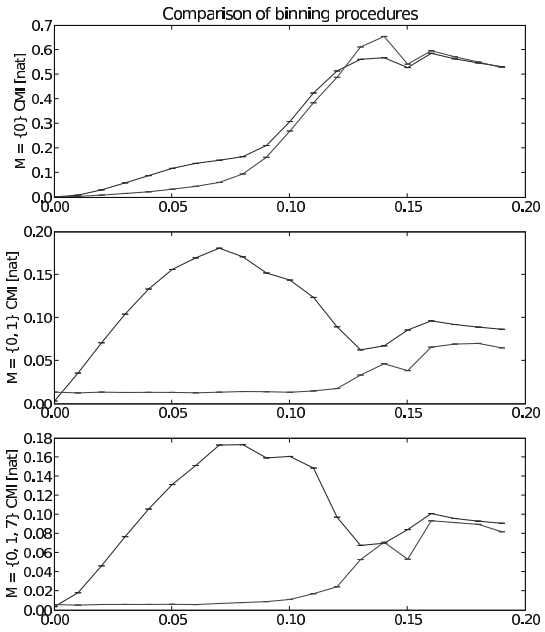

Figure 1: Conditional mutual information vs. strength of coupling for Rössler pair (13). Single condition model (top), optimal model per the selection procedure (center) and the model selected by the leave-one-out procedure on original time series data only (bottom).

Tests of detection of directionality in unidirectional coupling using all three models listed above have clearly shown that the model selected by the proposed procedure involving surrogate testing was the most effective. The proposed model has no false positives in the tested parameter range of window sizes from 256

points to 32768 points in powers of two and coupling strengths $\epsilon_1 = 0$ and $\epsilon_2 \in \{0, 0.1\}$. The model M_L had very low sensitivity and did not detect almost any directional coupling at all. The examination of the relevant histograms of the CMI indices has revealed that there is strong positive bias in the surrogates in the direction of coupling which renders all the detections negative. The model M_0 on the other hand has many false positive detections of coupling rendering the estimates unusable.

3.2. Van der Pol systems

The coupled Van der Pol equations are frequently used as example systems in nonlinear dynamics as they exhibit nonlinearity (and a stable limit cycle) but not deterministic chaos and complement other frequently used chaotic systems, such as the Rössler system or the Lorenz system. The nonlinearity of the Van der Pol system can be controlled by means of a parameter. The equations of the Van der Pol are given by

$$\ddot{x}_{1,2} - \mu(x_{1,2}^2 - 1)\dot{x}_{1,2} + \omega_{1,2}^2 x_{1,2} + \epsilon_{1,2}(x_{2,1} - x_{1,2}) + \eta_{1,2} = 0, \tag{14}$$

where $\mu = 0.2$ is the parameter affecting the nonlinearity of the model, $\omega_{1,2} = 1 \pm 0.1$ sets the main frequency of the model, $\eta_{1,2}$ are independent white zero-mean gaussian noise terms with standard deviation 0.1 and $\epsilon_{1,2}$ are the coupling strengths. The Van der Pol system pair was integrated with a Heun (reverse Euler) scheme with $dt = 0.01$ and subsampled by a factor of 20.

The intersample delay was comupted as $\tau = 5$ samples. With the parameters above the model selection procedure recommended the model $M^* = \{0, 6\}$, i.e. a two-dimensional model. The procedure was rerun without constraining the selected model to multiples of $\tau = 5$ and instead allowed to select any indices that are multiples of 2. Note that the prediction horizon $I(Y_\tau, Y_M)$ was the same in both runs. Using this less restrictive setting, the model selection procedure selected the model $M' = \{0, 12\}$ which is quite different to the previously chosen model. This shows that pre-selection can have adverse effects on the quality of the selected model.

Interestingly enough, the leave-one-out procedure selected a model $M = \{0, 12, 13\}$ with $\tau = 1$ (prediction horizon $I(Y_5, Y_M)$. The selected model is 3 dimensional, although the underlying dynamical model is only 2 dimensional. We note here that the model is not deterministic but stochastic and contains a noise input which is filtered by the dynamics of the system. Additionally, the model element 0 is forced to be a part of all models although it might not necessarily be useful in the prediction. Either of these considerations may explain why a 3 dimensional model was selected by the procedure.

The curves of conditional mutual information averaged for the lags $T \in \{17, 22\}$ (in case of the Van der Pols it is clear that coupling has most effect at these lags) is shown in Fig. 2.

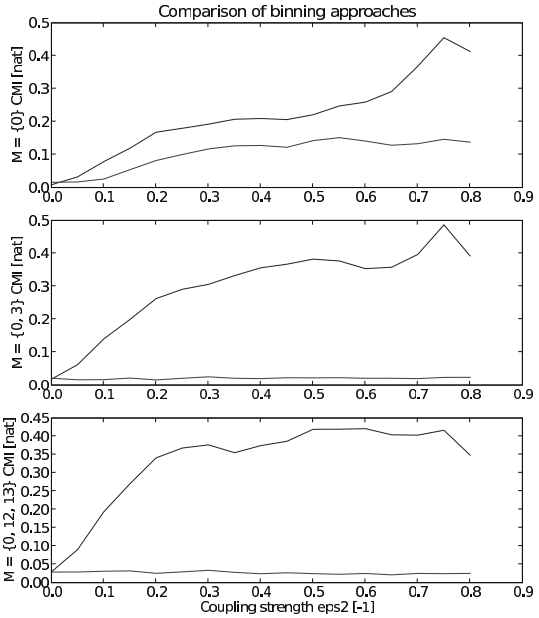

Figure 2: Conditional mutual information vs. strength of coupling for the Van der Pol pair (14). Single condition model (top), optimal model per the selection procedure (center) and the model selected by the leave-one-out procedure on original time series data only (bottom).

4. Conclusion

The selection of a conditioning model for processing amplitude series is a difficult problem and requires careful consideration. A method for selecting a conditioning model has been presented which attempts to select the optimal model with respect to the problem of detecting directional coupling.

The error of the prediction of a considered model was connected to the criterion (time lagged mutual information or conditional entropy) by selecting a suitable loss function. It has been shown that the error is positively biased with respect to the true expected value of the conditional entropy. The bias and variance that surrogates introduce into the directionality detection

method have been replicated in the model selection method by using generated surrogate time series to estimate the criterion instead of leave-one-out cross-validation or splitting the original time series into a training and testing set.

Some positive results have been shown on well-known and frequently used model systems. The recommended models have worked better than other reasonable choices. This has been verified by testing the conditioning models on the actual directionality detection problem for the considered systems. There are still however unresolved issues such as pre-filtering of the allowable samples to be included in the model and the methods is still very much a work in progress.

References

[1] D. Maraun and J. Kurths, "Epochs of phase coherence between El Niño/Southern Oscillation and indian monsoon," *Geophysical Research Letters*, vol. 32, no. 15, 2005.

[2] M. Paluš and D. Novotná, "Quasi-biennial oscillations extracted from the monthly NAO index and temperature records are phase-synchronized," *Nonlinear Processes in Geophysics*, vol. 13, pp. 287–296, 2006.

[3] C. Schäfer, M. Rosenblum, J. Kurths, and H.-H. Abel, "Heartbeat synchronization with ventilation," *Nature*, vol. 392, 1998.

[4] M. Paluš, V. Komárek, Z. Hrnčíř, and K. Štěrbová, "Synchronization as adjustment of information rates: Detection from bivariate time series," *Physical Review E*, vol. 63, 2001.

[5] M. Paluš and A. Stefanovska, "Direction of coupling from phases of interacting oscillators : An information-theoretic approach," *Physical Review E*, vol. 67, 2003.

[6] B. Musizza, A. Stefanovska, P. V. E. McClintock, M. Paluš, J. Petrovcic, S. Ribaric, and F. F. Bajrovič, "Interactions between cardiac, respiratory, and eeg-delta oscillations in rats during anaesthesia," *Journal of Physiology*, vol. 580, pp. 315–326, 2007.

[7] F. Takens, "Detecting strange attractors in turbulence," in *Dynamical systems and turbulence* (D. Rand and L. Young, eds.), vol. 898, (Berlin), pp. 366–381, Springer, 1981.

[8] T. Sauer, J. A. Yorke, and M. Casdagli, "Embedology," *Journal of Statistical Physics*, vol. 65, no. 3–4, pp. 579–616, 1991.

[9] M. Paluš and M. Vejmelka, "Directionality of coupling from bivariate time series: How to avoid false causalities and missed connections," *Physical Review E*, vol. 75, p. 056211, 2007.

[10] V. M. and M. Paluš, "Inferring the directionality of coupling with conditional mutual information," vol. 77, p. 026214, 2008.

[11] T. Sauer, "Reconstruction of dynamical systems from interspike intervals," *Physical Review Letters*, vol. 72, no. 24, pp. 3811–3814, 1994.

[12] H. Kantz and T. Schreiber, *Nonlinear Time Series Analysis*. Cambridge: Cambridge University Press, 1997.

[13] M. B. Kennel, B. R., and H. D. I. Abarbanel, "Determining embedding dimension for phase-space reconstruction using a geometrical construction," *Physical Review A*, vol. 45, p. 3403, 1992.

[14] A. M. Fraser and H. L. Swinney, "Independent coordinates for strange attractors from mutual information," *Physical Review A*, vol. 33, pp. 1134–1140, 1986.

[15] J. Rissanen, "Modeling by shortest data description," *Automatica*, vol. 14, pp. 465–471, 1978.

[16] G. Schwarz, "Estimating the dimension of a model," *Annals of Statistics*, vol. 6, pp. 461–464, 1978.

[17] H. Akaike, "A new look at the statistical model identification," *IEEE Transactions on Automatic Control*, vol. 19, pp. 716–723, 1974.

[18] J. Theiler, S. Eubank, A. Longtin, B. Galdrikian, and J. Farmer, "Testing for nonlinearity in time series: The method of surrogate data," *Physica D*, vol. 58, pp. 77–94, 1992.

[19] A. Stefanovska, H. Haken, P. V. E. McClintock, M. Hožič, F. Bajrovič, and S. Ribarič, "Reversible transitions between synchronization states of the cardiorespiratory system," *Physical Review Letters*, vol. 85, pp. 4831–4834, 2000.

[20] T. Schreiber and A. Schmitz, "Improved surrogate data for nonlinearity tests," *Physical Review Letters*, vol. 77, pp. 635–638, 1996.

[21] M. Thiel, M. Romano, J. Kurths, M. Rolfs, and R. Kliegl, "Twin surrogates to test for complex synchronisation," *Europhysics Letters*, vol. 75, pp. 535–541, 2006.