Journal of Statistical Physics, Vol. 124, No. 5, September 2006 (© 2006)
DOI: 10.1007/s10955-006-9131-x

Whitening as a Tool for Estimating Mutual Information in Spatiotemporal Data Sets

Andreas Galka,[1,2] Tohru Ozaki,[1] Jorge Bosch Bayard[3] and Okito Yamashita[4]

Received September 1, 2004; accepted May 3, 2005
Published Online: August 1, 2006

We address the issue of inferring the connectivity structure of spatially extended dynamical systems by estimation of mutual information between pairs of sites. The well-known problems resulting from correlations within and between the time series are addressed by explicit temporal and spatial modelling steps which aim at approximately removing all spatial and temporal correlations, i.e. at whitening the data, such that it is replaced by spatiotemporal innovations; this approach provides a link to the maximum-likelihood method and, for appropriately chosen models, removes the problem of estimating probability distributions of unknown, possibly complicated shape. A parsimonious multivariate autoregressive model based on nearest-neighbour interactions is employed. Mutual information can be reinterpreted in the framework of dynamical model comparison (i.e. likelihood ratio testing), since it is shown to be equivalent to the difference of the log-likelihoods of coupled and uncoupled models for a pair of sites, and a parametric estimator of mutual information can be derived. We also discuss, within the framework of model comparison, the relationship between the coefficient of linear correlation and mutual information. The practical application of this methodology is demonstrated for simulated multivariate time series generated by a stochastic coupled-map lattice. The parsimonious modelling approach is compared to general multivariate autoregressive modelling and to Independent Component Analysis (ICA).

KEY WORDS: time series analysis, autoregressive modelling, innovations, mutual information, maximum likelihood, whitening, spatiotemporal modelling, independent component analysis

[1] Institute of Statistical Mathematics (ISM), Minami-Azabu 4-6-7, Tokyo 106-8569, Japan; e-mail: galka@physik.uni-kiel.de
[2] Institute of Experimental and Applied Physics, University of Kiel, 24098 Kiel, Germany
[3] Cuban Neuroscience Center, Ave 25 No. 5202 esquina 158 Cubanacán, POB 6880, 6990, Ciudad Habana, Cuba
[4] ATR Computational Neuroscience Laboratories, Hikaridai 2-2-2, Kyoto 619-0288, Japan

1. INTRODUCTION

Recently there is growing consensus that for the investigation of complex extended systems new approaches to the analysis of dynamical multivariate data sets are required.(1,2) Such data sets will typically arise in the guise of multivariate time series, such that the temporal dimension of the data reflects the dynamical nature of the underlying processes, while other aspects pertinent to these processes are accommodated in further dimensions. If the data, as a typical and very important case, in addition to the dimension of time also depends on physical space, it is commonly called *spatiotemporal* data.

The temporal dimension is characterised by the non-equivalence between the two possible directions of the time arrow, which gives rise to the principle of causality, according to which events in the past may influence events in the future, whereas the opposite situation never occurs. The spatial dimension is (in most cases) characterised by the absence of such asymmetries, and by the concept of locality, i.e. the existence of a neighbourhood for each point in space, such that, for sufficiently short time intervals, direct interactions involving this point will typically be confined to points within this neighbourhood, but not extend to spatially remote points.

In contemporary scientific research a vast number of systems are investigated which fall into the class of spatially extended dynamical systems. They cover a wide range of disciplines, including hydrodynamics,(3) optics,(4) meteorology, geophysics,(5) ecology,(6) biology, medicine(7,8) and engineering.(9) Presently, it has become standard in these disciplines to routinely record spatiotemporal data sets in large quantities, either through active experimentation or through field observations. Despite the diversity of these disciplines and of the corresponding data sets, it seems likely that for many cases a unified approach to analysing and modelling the basic properties of the underlying dynamics can be formulated.

As a central notion relevant for this field we mention *connectivity*, i.e. the presence of direct (and possibly directed) dynamical interactions between spatially distinct locations within the system. Information about connectivity in a system may reach far beyond the level of pure data description since it addresses an important aspect of the functional composition of the system. Correlations within multivariate time series can be described by measures such as linear and nonlinear correlation functions or mutual information;(10) especially mutual information has attracted considerable attention recently since it promises a very general quantification of statistical dependence. Its practical estimation from real data, however, is known to be a difficult task because of the need to estimate probability densities.(11) This is true in particular for the case of short time series; in many situations it may be difficult or impossible to record long time series from a system, e.g. due to technical limitations or limited stationarity of the system.

In this paper we discuss the estimation of mutual information from temporally and spatially correlated time series, and aim at clarifying the relations between predictive modelling, maximum-likelihood estimation, model comparison, linear correlation and mutual information. The close relationship between mutual information and likelihood ratio testing was recently also observed by Brillinger.(12,13) As our starting point, we interpret modelling as a process of spatial and temporal *whitening* of the raw data, i.e. transforming it into spatially and temporally uncorrelated *innovations*. We will argue that this step removes the need for estimating unknown and possibly complicated probability distributions, since they are replaced by Gaussian distributions. While it is well known that fitting a time series by a predictive model corresponds to temporal whitening, i.e. the identification of the noise process driving the (multivariate) dynamics, we will show that spatial whitening can be interpreted as fitting the covariance matrix of this noise process.

The main focus of this paper lies on the presentation of the methodology and its theoretical background, and the discussion is kept at a very general level, assuming just the presence of a spatially extended dynamical system, from which spatiotemporal data has been recorded. Although our interest in this subject was initially triggered by data sets recorded in brain research, a full discussion of this particular field of application would be beyond the scope of this paper. Instead we will illustrate the practical application of the methodology by showing results of analysing simulated data sets generated by a system of coupled stochastic oscillators; this system can be regarded as an example of a coupled map lattice.(14)

The structure of this paper is as follows. In Sec. 2 we will briefly review the definitions of the coefficient of linear correlation and of mutual information. Both quantify dependencies between pairs of data sets, i.e. they may be estimated from bivariate time series; while it is also possible to estimate the mutual information of a set of more than two data sets, we will not use this case in this paper.

In Sec. 3 we will introduce a different viewpoint on the definition and estimation of mutual information from bivariate time series; for this purpose we will discuss the concept of whitening by predictive autoregressive (AR) modelling and refer to an important theorem from the theory of Markov processes which provides the justification of this approach. The derivation of a central result on the relation between the coefficient of linear correlation and mutual information will be given in Appendix A. We will also briefly review linear modelling of bivariate time series by likelihood maximisation.

In Sec. 4 we will generalise the discussion by considering modelling of multivariate time series representing spatially extended systems; in such situation the issue of modelling the instantaneous correlations between the various spatial locations has to be addressed, which will lead us to the concept of spatial whitening. Also in this case linear correlation and mutual information will be regarded as properties of bivariate time series within the multivariate data set, i.e. of pairs of spatial locations. In this section we will also briefly discuss the wider field of AR modelling of multivariate time series and one of its well-established manifestations, the Principal Oscillation Pattern (POP) method;(41) in contrast to this method, in this paper we prefer to employ *parsimonious* multivariate autoregressive (MAR) models, i.e. models with sparse transition matrix.

In Sec. 5 we will discuss how from the innovations resulting from multivariate time series modelling, deeper layers of correlation can be extracted; this will lead us to a fully parametric estimator of mutual information, providing an alternative to the common nonparametric estimators. The possibility to define this estimator is a direct consequence of the fact that, for sufficiently good models, the distribution of the innovations is known to be Gaussian. The detailed derivation of the estimator will be deferred to Appendix B. The parametric estimator itself is valid for general situations, not only for the case of spatiotemporal data.

In Sec. 6 the modelling approach, as proposed in this paper, will be briefly compared with a widely applied class of algorithms for the analysis of multivariate data, known as Independent Component Analysis (ICA).

In Sec. 7 a simulation study will be presented, and a concluding discussion will be given in Sec. 8.

2. LINEAR CORRELATION AND MUTUAL INFORMATION

In this section we will review the definition of the two commonly used statistics for measuring mutual dependencies between two time series, linear correlation and mutual information, and we will briefly discuss in which way the latter differs from the former.

2.1. Linear Correlation

For a given pair of time series x_t and y_t, $t = 1, \ldots, N_t$, the linear correlation structure can be quantified by the symmetric 2×2 sample covariance matrix

$$\mathsf{S}_{xy} = \frac{1}{N_t} \sum_{t=1}^{N_t} \left((x_t - \langle x_t \rangle), (y_t - \langle y_t \rangle) \right)^\dagger \left((x_t - \langle x_t \rangle), (y_t - \langle y_t \rangle) \right), \qquad (1)$$

where $\langle x_t \rangle = \frac{1}{N_t} \sum_t x_t$. We choose to define the elements of this matrix as

$$\mathsf{S}_{xy} = \begin{pmatrix} \sigma_x^2 & \sigma_x \sigma_y r(x, y) \\ \sigma_x \sigma_y r(x, y) & \sigma_y^2 \end{pmatrix}, \qquad (2)$$

where σ_x^2 and σ_y^2 denote the (sample) variances of x_t and y_t, respectively, while the off-diagonal element quantifies the linear cross-covariance between the two variables; here the coefficient of linear correlation $r(x, y)$ has been defined as[13]

$$r(x, y) = \frac{\sum_t (x_t - \langle x_t \rangle)(y_t - \langle y_t \rangle)}{N_t \sigma_x \sigma_y} = \frac{\sum_t (x_t - \langle x_t \rangle)(y_t - \langle y_t \rangle)}{\sqrt{\sum_t (x_t - \langle x_t \rangle)^2 \sum_t (y_t - \langle y_t \rangle)^2}}. \tag{3}$$

Note that by this definition $r(x, y)$ is normalised to $-1 \leq r(x, y) \leq 1$.

2.2. Mutual Information

In contrast to the coefficient of linear correlation, mutual information, as introduced in 1948 by Shannon,[15] does not directly refer to the case of a pair of time series. Rather it is based on the probability distributions of two random variables x and y which can assume values out of a set of states; here we limit our attention to the case of the number of possible states being finite, say S. Let the index $i, i = 1, \ldots S$, label these states, denote the corresponding values by x_i and y_i and assume that joint and marginal probability distributions $p(x_i, y_j)$, $p(x_i)$ and $p(y_i)$ for the occurrence of these states exist. Then the mutual information between x and y is defined by

$$I(x, y) = \sum_{i=1}^{S} \sum_{j=1}^{S} p(x_i, y_j) \log \frac{p(x_i, y_j)}{p(x_i) p(y_j)}. \tag{4}$$

Note that i and j do not label time, but states. Here we mention that it is possible also to define linear correlation by a summation over states instead over time,[10] but in this paper we will not need this variant. We rewrite Eq. (4) as

$$I(x, y) = \langle \log p(x_i, y_j) - \log(p(x_i) p(y_j)) \rangle_{p(x_i, y_j)}, \tag{5}$$

where $\langle . \rangle_{p(x_i, y_j)}$ denotes the average over (i, j) with respect to $p(x_i, y_j)$.

We remark that both mutual information and Shannon entropy are special cases of a general measure of discrepancy between two probability distributions p_i and q_i, introduced by Boltzmann already in 1877[16] and sometimes known as Boltzmann entropy,[17] which is defined by

$$B(p; q) = -\sum_i p_i \log \frac{p_i}{q_i}, \tag{6}$$

where i may be an index vector. The negative of $B(p; q)$ is known as Kullback–Leibler information or relative entropy.[18]

When estimating $I(x, y)$ from paired data $(x^{(t)}, y^{(t)})$, $t = 1, \ldots N_t$, where t denotes simply an index for labelling the data, the probability distributions $p(x_i, y_j)$, $p(x_i)$ and $p(y_j)$ have to be estimated numerically; while the most common method for this purpose is based on histograms,[13] recently various alternative approaches have been explored, in particular methods based on kernels,[19] correlation integrals(20) and k-nearest neighbour distances.(21) Obviously, the same situation is given for the case of the estimation of Shannon entropy.(11)

Note that the concepts of "time" and "dynamics" are not involved in the definition of $I(x, y)$. The index t just serves the purpose of determining which values in the data sets for x and y form pairs; this information is needed in the estimation of $p(x_i, y_j)$. Apart from this constraint the sampled data $(x^{(t)}, y^{(t)})$ are assumed to be independently drawn from the corresponding true probability distributions, and $I(x, y)$ will be invariant with respect to any shuffling of the order of the data, as long as the pairs are preserved.

If the data comes in the guise of time series (such that t in fact denotes time), the assumption of independence will typically be invalid, since most time series show temporal correlations. It has been observed that such correlations pose a problem for the estimation of mutual information; while recently a few authors have begun to address this problem and to develop remedies,(22,23) still in most applications the presence of temporal correlations is ignored. But by ignoring these correlations the dynamics of the underlying process is ignored. In this paper we suggest to regard these correlations as a source of valuable information, rather than a nuisance.

3. INNOVATION APPROACH TO MUTUAL INFORMATION

In this section we will review the main theoretical foundations of this paper, namely the concept of whitening and its implications for the likelihood of the data; furthermore we will briefly discuss univariate and bivariate linear autoregressive models.

3.1. The Likelihood of Innovation Time Series

If we decide to take the temporal correlations in time series serious, we have to regard x_t and y_t as different random variables for each value of t; then Eq. (5) should be replaced by

$$I(x, y) = \log p\left((x_1, y_1), \ldots, (x_{N_t}, y_{N_t})\right) - \log\left(p(x_1, \ldots, x_{N_t})p(y_1, \ldots, y_{N_t})\right), \tag{7}$$

where in the joint distribution of all elements of both time series we have ordered the elements as pairs. If several time series aligned by an external trigger are available, also the average with respect to the joint distribution (see Eq. (5)) could be maintained, but this step is not essential and can be omitted. Reinterpreting Eq. (7) from the viewpoint of time series analysis, it can be seen that mutual information can be regarded as a difference between two terms representing log-likelihoods, the first referring to the bivariate time series (x_t, y_t), the second being the sum of the log-likelihoods of the two univariate time series x_t and y_t. Let log-likelihood be denoted by $\mathcal{L}$, then Eq. (7) corresponds to

$$I(x,y) = \mathcal{L}(x,y) - (\mathcal{L}(x) + \mathcal{L}(y))\,. \tag{8}$$

For later use we note that Eq. (7) can alternatively be written by using conditional probabilities as

$$\begin{aligned} I(x,y) = {} & \log(p(x_1,\ldots,x_{N_t}|y_1,\ldots,y_{N_t})\,p(y_1,\ldots,y_{N_t})) \\ & - \log(p(x_1,\ldots,x_{N_t})\,p(y_1,\ldots,y_{N_t})), \end{aligned} \tag{9}$$

corresponding to

$$I(x,y) = \mathcal{L}(x|y) - \mathcal{L}(x). \tag{10}$$

In order to estimate mutual information from Eq. (7) high-dimensional joint distributions need to be evaluated which, due to correlations between the elements of the two time series, generally will have very complicated structure; these correlations will occur both between x and y and within the values of each of these two variables at different points of time t. In order to simplify the structure of these distributions we propose to describe these correlations by the corresponding optimal predictors of x_t and y_t, based on the set of previous values of the two time series, $\{(x_{t_x}, y_{t_y})|\, t_x, t_y < t\}$, i.e. we perform three different modelling steps, thereby estimating the conditional means for x_t, y_t and (x_t, y_t); conditional mean will be denoted by $\mathcal{E}(.)$. By limiting the information available for prediction to the past we ensure that we stay in the domain of causal modelling. The residuals of these predictions are given by

$$\epsilon_t(x|x) = x_t - \mathcal{E}(x_t|x_{t-1}, x_{t-2}, \ldots), \tag{11}$$

$$\epsilon_t(y|y) = y_t - \mathcal{E}(y_t|y_{t-1}, y_{t-2}, \ldots), \tag{12}$$

$$\begin{aligned} (\epsilon_t(x|x,y), \epsilon_t(y|x,y))^\dagger &= (x_t, y_t)^\dagger \\ & -\mathcal{E}((x_t, y_t)^\dagger|(x_{t-1}, y_{t-1})^\dagger,\ (x_{t-2}, y_{t-2})^\dagger, \ldots). \end{aligned} \tag{13}$$

Note that in general the residuals for x and y in Eq. (13) will be different from those in Eqs. (11) and (12), since in Eq. (13) additional information is employed for the predictions; this fact is expressed by the notation $\epsilon_t(x|x)$, $\epsilon_t(x|x,y)$, explicitly stating the conditioning on one or both of the variables x and y.

Following a suggestion of Wiener, residuals are also called *innovations*. The theoretical framework for the transformation of time series data into independent innovations is provided by the theory of stochastic differential equations(24,25) and by filtering theory,(26,27) alternatively also known as Markov process theory. The most general results were given by Lévy(28) (see Theorem 41 in Ref. 29). He has shown that, under mild conditions, any continuous-time Markov process can be modelled such that the corresponding innovations can be represented as the sum of two white noise processes, which have Gaussian and Poisson distributions, respectively; in the case of continuous dynamics only the Gaussian noise process will be present. The case of additional observation noise has been treated by Frost and Kailath.[30] Consequently, we expect that, under the assumption of continuous dynamics, for optimal predictors the time series of resulting innovations will be uncorrelated (in fact, independent) and Gaussian, even if, due to nonlinearities in the dynamics, the x_t and y_t are non-Gaussian.

Eqs. (11), (12) and (13) represent mappings from the original data to the corresponding innovations; let the Jacobians of these mappings be denoted by $\frac{\partial \boldsymbol{\varepsilon}(\boldsymbol{x})}{\partial \boldsymbol{x}}$, $\frac{\partial \boldsymbol{\varepsilon}(\boldsymbol{y})}{\partial \boldsymbol{y}}$ and $\frac{\partial \boldsymbol{\varepsilon}(\boldsymbol{x},\boldsymbol{y})}{\partial (\boldsymbol{x},\boldsymbol{y})}$, respectively, then the likelihoods of the original data and of the innovations are related according to[31]

$$p(x_1, \ldots, x_{N_t}) = \left| \frac{\partial \boldsymbol{\varepsilon}(\boldsymbol{x})}{\partial \boldsymbol{x}} \right| \; p(\epsilon_1(x|x), \ldots, \epsilon_{N_t}(x|x)), \tag{14}$$

$$p(y_1, \ldots, y_{N_t}) = \left| \frac{\partial \boldsymbol{\varepsilon}(\boldsymbol{y})}{\partial \boldsymbol{y}} \right| \; p(\epsilon_1(y|y), \ldots, \epsilon_{N_t}(y|y)), \tag{15}$$

$$\begin{aligned} p((x_1, y_1), \ldots, (x_{N_t}, y_{N_t})) = \left| \frac{\partial \boldsymbol{\varepsilon}(\boldsymbol{x}, \boldsymbol{y})}{\partial (\boldsymbol{x}, \boldsymbol{y})} \right| & p((\epsilon_1(x|x, y), \epsilon_1(y|x, y)), \ldots, \\ & (\epsilon_{N_t}(x|x, y), \epsilon_{N_t}(y|x, y))), \end{aligned} \tag{16}$$

where the notation $|.|$ denotes the absolute value of the determinant of a matrix.

Now it can be seen from Eq. (11) that $\frac{\partial \epsilon_t}{\partial x_t} = 1$ and $\frac{\partial \epsilon_t}{\partial x_{t'}} = 0$ for $t' > t$ (according to the principle of causality), therefore the determinant in Eq. (14) is unity, and in fact $p(x_1, \ldots, x_{N_t})$ is equal to $p(\epsilon_1(x|x), \ldots, \epsilon_{N_t}(x|x))$ for the specific series of innovations $\epsilon_1(x|x), \ldots, \epsilon_{N_t}(x|x)$ corresponding to the data $x_1, \ldots, x_{N_t}$, although the shapes of these two distributions itself may differ very much. The same argument applies to Eqs. (15) and (16).

Fortunately, this result removes the need to estimate the unknown and possibly complicated probability distributions in Eqs. (4) and (7), since, if sufficiently good models can be found, it follows from Markov process theory, as discussed above, that the distributions of the innovations are products of Gaussians. Then the corresponding log-likelihoods are given for x by

$$\begin{aligned} \mathcal{L}(x) &= \log p(x_1, \ldots, x_{N_t}) = \log p(\epsilon_1(x|x), \ldots, \epsilon_{N_t}(x|x)) \\ &= -\frac{1}{2} \left(N_t \log \sigma^2_{\epsilon(x|x)} + \sum_{t=1}^{N_t} \frac{\epsilon_t^2(x|x)}{\sigma^2_{\epsilon(x|x)}} + N_t \log(2\pi) \right), \end{aligned} \tag{17}$$

for y by a corresponding expression and for (x, y) by

$$\begin{aligned} \mathcal{L}(x, y) &= \log p((x_1, y_1), \ldots, (x_{N_t}, y_{N_t})) \\ &= \log p((\epsilon_1(x|x, y), \epsilon_1(y|x, y)), \ldots, (\epsilon_{N_t}(x|x, y), \epsilon_{N_t}(y|x, y))) \end{aligned}$$

$$= -\frac{1}{2}\Bigg(N_t \log |\mathsf{S}_{\epsilon(x,y|x,y)}| + \sum_{t=1}^{N_t} (\epsilon_t(x|x, y), \epsilon_t(y|x, y))$$
$$\times \mathsf{S}^{-1}_{\epsilon(x,y|x,y)} \ (\epsilon_t(x|x, y), \epsilon_t(y|x, y))^{\dagger} + 2N_t \log(2\pi)\Bigg). \qquad (18)$$

Here $\sigma^2_{\epsilon(x|x)}$ denotes the variance of the innovation process for the case of a predictive model for x only; and $\mathsf{S}_{\epsilon(x,y|x,y)}$ denotes the covariance matrix of the bivariate innovation process for the case of a predictive model for (x, y). Note that, if x and y are replaced by the corresponding innovations, the sample covariance matrix defined in Eqs. (1) and (2) corresponds to $\mathsf{S}_{\epsilon(x,y|x,y)}$, therefore the same structure is chosen:

$$\mathsf{S}_{\epsilon(x,y|x,y)} = \begin{pmatrix} \sigma^2_{\epsilon(x|x,y)} & \sigma_{\epsilon(x|x,y)}\,\sigma_{\epsilon(y|x,y)}\, r(\epsilon(x), \epsilon(y)) \\ \sigma_{\epsilon(x|x,y)}\,\sigma_{\epsilon(y|x,y)}\, r(\epsilon(x), \epsilon(y)) & \sigma^2_{\epsilon(y|x,y)} \end{pmatrix}. \qquad (19)$$

Here again we have defined a normalised linear correlation coefficient $r\big(\epsilon(x), \epsilon(y)\big)$, analogous to Eq. (3), in order to describe instantaneous correlations between the innovations $\epsilon(x|x, y)$ and $\epsilon(y|x, y)$.

If we replace in Eqs. (17), (18) and (19) the parameters $\sigma_{\epsilon(x|x)}$, $\sigma_{\epsilon(y|y)}$, $\sigma_{\epsilon(x|x,y)}$, $\sigma_{\epsilon(y|x,y)}$ and $r\big(\epsilon(x), \epsilon(y)\big)$ by their appropriate maximum-likelihood estimators and insert the results into Eq. (8), we obtain

$$I(x, y) = -\frac{1}{2}N_t\big(\log(1 - r^2(\epsilon(x), \epsilon(y)))$$
$$+\big(\log \sigma^2_{\epsilon(x|x,y)} - \log \sigma^2_{\epsilon(x|x)}\big) + \big(\log \sigma^2_{\epsilon(y|x,y)} - \log \sigma^2_{\epsilon(y|y)}\big)\big); \qquad (20)$$

here for notational convenience we have abstained from denoting the estimators of the parameters in a different way than the parameters themselves. A detailed derivation of Eq. (20) can be found in Appendix A.

Strictly speaking, we have been ignoring so far that for the first data points $x_1, x_2, \ldots, x_p$ (where p is the model order) no optimal predictions can be performed due to lack of a sufficient number of previous points, so the corresponding contribution to the likelihood has to be calculated by different methods;(32) but for small p and sufficiently large N_t this missing contribution can usually be neglected.

Note that Eq. (7) can be regarded as the likelihood ratio test (LRT) statistic of the null hypothesis of independence of the time series x_t and y_t.(13) By Eq. (20) possible deviations from independence are decomposed into three components, the first describing instantaneous correlations between the innovations of x and y (quantified by $r(\epsilon(x), \epsilon(y))$), while the second and the third describe dependence of x on the past of y and vice versa. If knowing the past of y does not improve predictions of x, and vice versa, the mutual information can still be non-zero, as given by the first term on the rhs of Eq. (20); however, since any estimates of the mutual information obtained from actual finite samples will follow a χ^2-distribution (as it is usually the case for any LRT statistics in the null case), it is to be expected that even in this case there will be a small positive bias resulting from the second and third terms on the rhs of Eq. (20). In contrast to this, in the non-null case the estimate of mutual information can be expected to follow a Gaussian distribution).[33]

3.2. Calculation of Innovations

The time series of innovations $\epsilon_t(x|x)$, $\epsilon_t(y|y)$, $\epsilon_t(x|x, y)$ and $\epsilon_t(y|x, y)$ are obtained by fitting causal dynamical models to the data (x_t, y_t); linear or nonlinear model classes may be employed, depending on the properties of the data. Nonlinear modelling will be necessary, if the distribution of the data displays deviations from Gaussianity; this nonlinearity can be incorporated either directly in the model for the dynamics, or in a static observation function through which the underlying dynamics was observed, or in a combination of these two cases. The main classes of deviation from Gaussianity are given by heavy-tailed distributions, thin-tailed distributions and asymmetric distributions. In its full generality the task of identifying an optimal model may be very demanding, if not infeasible, but it can be expected that in many cases sub-optimal approximations will be sufficient. The class of linear models represents a convenient first-order approximation; experience has shown that in a substantial number of cases even this model class provides sufficient whitening.[34] In this paper we will focus exclusively on the case of linear modelling.

A specific model is characterised by its structure (or, equivalently, by belonging to a specific model class) and by a set of data-dependent parameters, collected in a parameter vector $\boldsymbol{\vartheta}$. The univariate linear AR model for the dynamics of a state variable ξ is given by

$$\xi_t = \mu + \sum_{t'=1}^{p} a_{t'} \xi_{t-t'} + e_t, \tag{21}$$

where p denotes the positive integer model order, μ denotes a constant intercept term (allowing for non-zero mean of ξ_t), and e_t represents the dynamical noise term which drives the dynamics; the variance of e_t is denoted by σ_e^2. The parameter vector for this model is given by $\boldsymbol{\vartheta} = (\mu, a_1, \ldots, a_p, \sigma_e^2)$. On the basis of this model the innovations can be estimated from given data x_t by

$$\epsilon_t(x|x) = x_t - \left(\mu + \sum_{t'=1}^{p} a_{t'} x_{t-t'} \right); \tag{22}$$

the variance of the innovations $\epsilon_t(x|x)$ provides a sample estimate for σ_e^2. For the case of a linear bivariate AR model for the state variables $(\xi, \eta)^\dagger$ the corresponding model is given by

$$\begin{pmatrix} \xi_t \\ \eta_t \end{pmatrix} = \begin{pmatrix} \mu_\xi \\ \mu_\eta \end{pmatrix} + \sum_{t'=1}^{p} \mathsf{A}_{t'} \begin{pmatrix} \xi_{t-t'} \\ \eta_{t-t'} \end{pmatrix} + \begin{pmatrix} e_t(\xi) \\ e_t(\eta) \end{pmatrix}, \tag{23}$$

where the dynamical noise term is represented by $(e_t(\xi), e_t(\eta))^\dagger$; its (2×2) covariance matrix is denoted by $\mathsf{S}_{e(\xi,\eta)}$. The parameter vector $\boldsymbol{\vartheta}$ for this model consists of the intercept terms (μ_ξ, μ_η), all elements of the transition matrices $\mathsf{A}_{t'}$ and the three independent elements from $\mathsf{S}_{e(\xi,\eta)}$. The estimator of the innovations $(\epsilon_t(x|x, y), \epsilon_t(y|x, y))^\dagger$ for data x_t and y_t (representing ξ and η) follows in analogy to Eq. (22). From now on we will express models always in terms of observed data and estimated innovations (instead of the not directly accessible "true" driving noises e_t), as in the example of Eq. (22).

The preferred method for choosing an appropriate model class and for fitting the parameters is maximisation of the likelihoods, as given by Eqs. (17) and (18). However, it should be mentioned that when comparing the performance of different model classes, having different numbers of data-dependent parameters (i.e. different dimension of $\boldsymbol{\vartheta}$), the model class with the larger number of data-dependent parameters will typically achieve the better likelihood, and therefore overfitted models will result. This phenomenon can be interpreted by stating that the likelihood represents a biased estimator of Boltzmann entropy, Eq. (6) (where p_i represents the true distribution of the quantity to be predicted, and q_i represents the predictive distribution).(35) Based on estimating this bias, corrections to the likelihood have been proposed, such as the Akaike Information Criterion (AIC) or the Bayesian Information Criterion (BIC).(36) Cross-validation represents an alternative approach to avoiding overfitting; it has been proved that minimisation of AIC and cross-validation are asymptotically equivalent.(37) In this paper we will employ minimisation of AIC for the purpose of model comparison.

Remembering that in general there is the possibility of choosing between various different model classes for whitening, such as classes based on different nonlinear functions or different state space representations, we note that consequently the innovations are not uniquely defined, a phenomenon which at first sight may appear problematic. We would like to remark that the situation is not different for the case of the nonparametric estimators of probability distributions mentioned earlier, such as histogram or kernel estimators: Also for these methods the results will depend on various design choices, such as kernel functions, bandwidths, bin widths, etc.

4. MODELLING SPATIOTEMPORAL DATA

So far we have discussed the case of only two random variables x and y; now we shall turn to the general case of spatiotemporal dynamics. Data which depend both on time and on space can be modelled by MAR models; we will now describe how such models can be formulated and simplified for high-dimensional situations (i.e. data covering a large number of grid points), and we will discuss an approximative approach to modelling the covariance matrix of the driving noise of MAR models.

4.1. Multivariate AR Modelling: General Case

We assume that a discretisation of physical space into a rectangular grid is used, and that the grid points are labelled by an index v, such that the measurement consists of scalar time series $x_t^{(v)}$ for each grid point. Let $\mathbf{x}_t$ and $\boldsymbol{\epsilon}_t(\mathbf{x})$ denote the column vectors formed of the data $x_t^{(v)}$ and the innovations $\epsilon_t(x^{(v)})$ for all grid points, respectively, and let N_v denote the total number of grid points.

In general form the dynamics of such systems may be described by nonlinear MAR models; the linear case is given by the generalisation of Eq. (23), such that the corresponding innovations are estimated by (see Eq. (22))

$$\boldsymbol{\epsilon}_t(\mathbf{x}) = \mathbf{x}_t - \left(\boldsymbol{\mu} + \sum_{t'=1}^{p} \mathsf{A}_{t'} \mathbf{x}_{t-t'} \right), \tag{24}$$

where the N_v-dimensional intercept vector $\boldsymbol{\mu}$, the set of the $N_v \times N_v$ transition matrices $\mathsf{A}_{t'}$, $t' = 1, \ldots, p$, and the $N_v(N_v - 1)/2$ independent elements of the covariance matrix $\mathsf{S}_{\boldsymbol{\epsilon}(\mathbf{x})} = \mathcal{E}\left(\boldsymbol{\epsilon}_t(\mathbf{x})\boldsymbol{\epsilon}_t(\mathbf{x})^\dagger\right)$ form the parameter vector $\boldsymbol{\vartheta}$. For each pair of grid points (u, v) the corresponding elements of the transition matrices $(\mathsf{A}_{t'})_{uv}$ and $(\mathsf{A}_{t'})_{vu}$ describe the dynamical (i.e., delayed) interactions between these two grid points, while the instantaneous correlations are described by $(\mathsf{S}_{\boldsymbol{\epsilon}(\mathbf{x})})_{uv} \equiv (\mathsf{S}_{\boldsymbol{\epsilon}(\mathbf{x})})_{vu}$. The total number of parameters in this model (i.e. the dimension of $\boldsymbol{\vartheta}$), to be estimated from the data, is given by

$$N_{\text{par}} = N_v + pN_v^2 + \frac{1}{2}N_v(N_v - 1). \tag{25}$$

Efficient methods for estimating the parameters of Eq. (24) have been proposed by Levinson,[38] Whittle[39] and Neumaier and Schneider.[40] The special case of a model of first order, $p = 1$, forms the basis for the method known as *Principal Oscillation Pattern* (POP) analysis[41] which has found widespread application in geophysics and related fields. This method is based on the analysis of the eigenvalues of the first-order transition matrix A_1; depending on whether these eigenvalues are real numbers or complex conjugated pairs, they correspond to (stochastically driven) relaxator or oscillator modes of the underlying spatiotemporal dynamics. This method has been generalised to MAR models of higher order, $p > 1$, by Neumaier and Schneider.[(40)]

However, it is a well-known disadvantage of MAR models that the number of parameters N_{par} from Eq. (25) may easily become very large, thereby leading to overparametrised, non-parsimonious models. Depending on the resolution of the spatial discretisation (e.g., the number of spatially distinct measurement sites), N_v may be a large number; if furthermore the length of the available time series N_t is only short (a common situation in many fields of application), it may easily occur that N_{par} assumes a value comparable to (or even excessing) the total number of data values, thereby rendering reliable estimation of parameters infeasible. According to a generally accepted guideline obtained from practical experience, the number of parameters to be estimated should not exceed a threshold of approximately 10% of the number of available data values. For this reason we will now discuss a parsimonious variant of MAR modelling.

4.2. Multivariate AR Modelling: Parsimonious Approach

For most spatially extended physical systems the assumption is justified that at sufficiently short time scales interactions will take place only over short distances, therefore we propose to restrict the dynamics to local neighbourhoods: As a first-order approximation we assume that each grid point will interact directly only with its immediate spatial neighbours on the grid. Most elements of A_1 become zero by this assumption, i.e. A_1 assumes a *sparse* structure. As a further simplification, motivated by regarding the spatially and temporally discrete model as an approximation of a (stochastic) partial differential equation, it is reasonable to limit the interaction between different grid points to the first lag, i.e. to set all off-diagonal elements of $\mathsf{A}_{t'}$, $t' > 1$, to zero. The diagonal elements, describing the dependence of each grid point on its own past, may in general be non-zero up to a lag (a model order) $t' = p > 1$. This situation corresponds to the approximation of a pth-order time derivative in a partial differential equation.

For a single grid point v the temporally whitened innovations result as

$$\epsilon_t\big(x^{(v)}\big) = x_t^{(v)} - \left(\mu^{(v)} + \sum_{t'=1}^{p} a_{t'}^{(v)} x_{t-t'}^{(v)} + \sum_{u \in \mathcal{N}(v)} b_1^{(v,u)} x_{t-1}^{(u)}\right), \tag{26}$$

where $a_{t'}^{(v)}$, $t' = 1, \ldots, p$, and $b_1^{(v,u)}$ denote the parameters for self-interaction and neighbour interaction, respectively, $\mathcal{N}(v)$ denotes the set of labels of the neighbours of grid point v, and $\mu^{(v)}$ is the constant intercept for time series $x_t^{(v)}$. In analogy with Eq. (22), we should have denoted the innovations in Eq. (26) by $\epsilon_t(x^{(v)}|x^{(v)})$ or possibly $\epsilon_t(x^{(v)}|x^{(v)}, x^{\mathcal{N}(v)})$ instead of $\epsilon_t(x^{(v)})$, but in order to avoid excessively complicated notation we will from now on refrain from explicitly stating the conditioning on the past of the same grid point, and possibly its neighbours.

As a generalisation we mention the possibility that in Eq. (26) also the model order p could be chosen differently for each grid point, $p = p(v)$, but in this paper we will not explore this option further. Rather will we employ a common value for p for all grid points, which could be obtained by minimisation of an information criterion such as AIC or BIC; alternatively it may be argued that p should be set at most to 2, since higher orders of a time derivative would be unusual in partial differential equations describing excitable media.

This model can be interpreted as a decomposition of the high-dimensional dynamical system described by Eq. (24) into a set of coupled low-dimensional dynamical systems, each of which is focussed on one grid point; the influence of the neighbouring grid points is treated as an additional external disturbance, represented by the neighbourhood term. It should be stressed that when analysing spatiotemporal data with a view at the connectivity structure of the underlying system, we are not interested in the correlations between pairs of neighbouring grid points, since we regard their presence as natural; but rather will we be interested in correlations between pairs of grid points separated by larger distances, since they may represent some faster mode of interaction which is not described by the basic MAR model.

Given the data $x_t^{(v)}$, the parameters $a_1^{(v)}, \dots, a_p^{(v)}$, $b_1^{(v,u)}$ and $\mu^{(v)}$ can be estimated separately for each grid point v by the linear least-squares method (which under the assumptions of Gaussian innovations and linear dynamics is equivalent to full likelihood maximisation), and a series of innovations $\epsilon_t(x^{(v)})$ will result as an estimate of the noise process driving the local dynamics of grid point v. This simple and efficient pointwise model fitting approach replaces the usual methods for fitting of full MAR models, as mentioned above, but it does so at the cost of neglecting the need to estimate also the off-diagonal elements of the covariance matrix of the driving noise, $\mathsf{S}_{\boldsymbol{\epsilon}(\mathbf{x})}$. We will deal with this point in the next section.

4.3. Spatial Whitening

In the previous section simplifications have been introduced by which most of the elements of the transition matrices $\mathsf{A}_{t'}$ could be set to zero. Now a similar step has to be accomplished for the covariance matrix $\mathsf{S}_{\boldsymbol{\epsilon}(\mathbf{x})}$. In general, it has to be expected that the off-diagonal elements of $\mathsf{S}_{\boldsymbol{\epsilon}(\mathbf{x})}$ are non-zero, i.e. that instantaneous mutual correlations between the innovations $\epsilon_t(x^{(v)})$ at different grid points are present; this will be true especially for neighbouring grid points. But if this is the case, the decomposition approach described in the previous section is invalid, since it assumes uncorrelated noises and provides estimates only for the diagonal elements $\sigma^2_{\epsilon(x^{(v)})}$.

Note that here we are dealing with instantaneous (as compared to the sampling time of the data), purely spatial correlations, which cannot be described by dynamical models like Eq. (26); therefore a separate step of spatial whitening is needed. We describe this step by an instantaneous linear transform applied to the original data

$$\tilde{\mathbf{x}}_t = \mathsf{L}\mathbf{x}_t. \tag{27}$$

Various approaches for the choice of the matrix L may be explored, and again it will depend on the particular data which choice provides best spatial whitening; as a first approximation we propose to employ a *Laplacian* matrix

$$\mathsf{L} = \left(\mathsf{I}_{N_v} - \frac{1}{k}\mathsf{N}\right), \tag{28}$$

where I_{N_v} denotes the $N_v \times N_v$ unity matrix, and N denotes a $N_v \times N_v$ matrix having $\mathsf{N}_{vu} = 1$ if u belongs to $\mathcal{N}(v)$, and 0 otherwise. This transform corresponds to a discrete second-order spatial derivative; derivatives are natural tools for whitening. The parameter k in Eq. (28) should be chosen as $k = 6$, if for each grid point six spatial neighbours are considered; but the parameter may also be chosen in a data-adaptive way, e.g. by maximum-likelihood, if a corresponding correction term is added to the likelihood (see Eq. (31)). A practical advantage of choosing a value for k different from the number of neighbours is that otherwise L may be very close to singular.

We remark that the class of autoregressive integrated moving-average (ARIMA) models for time series modelling[32] results from a similar approach of taking a discrete derivative prior to any further analysis, but in the case of ARIMA this derivative is applied in the time domain, while here we are applying it to the spatial domain.

If consequently we replace $\mathbf{x}$ by $\tilde{\mathbf{x}}$, Eq. (24) becomes

$$\boldsymbol{\epsilon}_t(\mathbf{x}) = \mathsf{L}^{-1}\boldsymbol{\epsilon}_t(\tilde{\mathbf{x}}) = \mathbf{x}_t - \left(\mathsf{L}^{-1}\boldsymbol{\mu} + \sum_{t'=1}^{p}\mathsf{L}^{-1}\mathsf{A}_{t'}\mathsf{L}\mathbf{x}_{t-t'}\right), \tag{29}$$

where $\mathsf{S}_{\boldsymbol{\epsilon}(\tilde{\mathbf{x}})} = \mathcal{E}\left(\boldsymbol{\epsilon}_t(\tilde{\mathbf{x}})\boldsymbol{\epsilon}_t(\tilde{\mathbf{x}})^\dagger\right)$ is expected to be diagonal or at least closer to diagonal than $\mathsf{S}_{\boldsymbol{\epsilon}(\mathbf{x})}$. Therefore this approach to spatial whitening is equivalent to modelling the *non-diagonal* covariance matrix of the innovations corresponding to the original data by

$$\mathsf{S}_{\boldsymbol{\epsilon}(\mathbf{x})} = \mathsf{L}^{-1}\mathsf{S}_{\boldsymbol{\epsilon}(\tilde{\mathbf{x}})}(\mathsf{L}^{-1})^\dagger; \tag{30}$$

recently a similar model has also been applied successfully for estimating unobserved brain states through spatiotemporal Kalman filtering.[42,43] Future research may succeed in identifying superior approaches to model instantaneous spatial correlations in spatiotemporal data. Clearly, depending on the data, this transformation will not remove all correlations from the $\epsilon_t(x^{(v)})$, but it removes those correlations which are merely an artifact of spatial neighbourhood and therefore may be regarded as "trivial." Remaining correlations can be expected to contain more relevant information about the underlying system, i.e. its connectivity structure, as will be demonstrated in the next section.

Note that after the application of the Laplacian transform to the data, a correction needs to be applied to the log-likelihood, which is given by

$$\mathcal{L}(\mathbf{x}_1, \ldots, \mathbf{x}_{N_t}) = \mathcal{L}(\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_{N_t}) + (N_t - p) \log |\mathsf{L}|. \tag{31}$$

The number of parameters of the model given by Eq. (26) follows as

$$N_{\text{par}} = N_v \left(\left\langle \text{card} \left(\mathcal{N}(v) \right) \right\rangle + p + 1 \right) + 1, \tag{32}$$

where $\left\langle \text{card} \left(\mathcal{N}(v) \right) \right\rangle$ denotes the average number of neighbours of a grid point, and the final +1 counts the parameter k of the Laplacian.

In practice, once the data has been spatially whitened by multiplication with the Laplacian L, the model fitting can be performed as already described above, without paying any further attention to the spatial whitening step. This remains true for the case that within the multivariate data set the time series of a given pair of (typically non-neighbouring) grid points v and w are modelled, in addition to the model terms already present in Eq. (24) for each of the two grid points, by some kind of direct interaction terms, e.g. as decribed by Eq. (23). In this case the resulting series of conditional innovations can be denoted by $\epsilon_t(\tilde{x}^{(v)}|\tilde{x}^{(w)})$ (and vice versa). From now on we will omit the tilde and use the following shorthand notations: $\epsilon_t(v) := \epsilon_t(x^{(v)})$ and $\epsilon_t(v|w) := \epsilon_t(x^{(v)}|x^{(w)})$; in the same way, the corresponding variances will be abbreviated as $\sigma^2_{\epsilon(v)} := \sigma^2_{\epsilon(x^{(v)})}$ and $\sigma^2_{\epsilon(v|w)} := \sigma^2_{\epsilon(x^{(v)}|x^{(w)})}$.

5. ANALYSING CONNECTIVITY IN INNOVATION TIME SERIES

With this section we will complete the theoretical part of this paper by discussing how further information can be extracted from the innovations resulting from spatiotemporal modelling. For pairs of grid points within spatiotemporal data sets a modification of the modelling presented so far will be introduced; based on this modified model we will derive a parametric estimator of mutual information.

5.1. Spatial Correlations in the Innovations

If a spatiotemporal time series would be transformed to completely independent innovations (with respect to time and space), this would mean that all available dynamical information had successfully been extracted from this time series and condensed into the model, as represented by both the dynamical model and the spatial whitening transformation (i.e. the covariance matrix of the innovations). Clearly for real-world data, this can never be achieved completely. As an example, there may exist very fast connections between grid points which are not neighbours, but separated by some larger distance; these would give rise to additional instantaneous spatial correlations that are not removed by multiplication with the Laplacian matrix L. We propose to regard the possibility of such incidents not as a weakness of the parsimonious modelling approach discussed so far, but rather as a strength, since it enables us to explore deeper layers of the spatiotemporal dynamics. By removing a first layer of spatial and temporal correlations we obtain a transformed representation of the original data set, within which we may be able to detect much more subtle correlations. This line of argumentation can be regarded as an illustration of the mathematical argument of Sec. 3.1 which established the equality of $p\left(\epsilon_1(x), \ldots, \epsilon_{N_t}(x)\right)$ and $p(x_1, \ldots, x_{N_t})$.

Given the innovation time series $\boldsymbol{\epsilon}_t(\mathbf{x})$, we may look for remaining instantaneous spatial correlations between pairs of grid points (v, w) by computing the coefficient of linear correlation $r\left(\epsilon(v), \epsilon(w)\right)$ according to Eq. (3), or by estimating the mutual information $I\left(\epsilon(v), \epsilon(w)\right)$ according to Eq. (4), thereby making use of the fact that temporal correlations have been removed by the temporal whitening step.

Alternatively, as mentioned already at the end of Sec. 4.3, the modelling of the data (i.e. the generation of the innovations) may be performed by including additional direct interaction terms, either by performing a full bivariate model fitting step for a given pair of grid points, following Eq. (23), or at least by allowing for a non-zero value of the covariance term $\mathcal{E}\left(\epsilon_t(v)\epsilon_t(w)\right)$ within $\mathsf{S}_{\boldsymbol{\epsilon}(\mathbf{x})}$. If our aim is to capture instantaneous spatial correlations, the latter approach is sufficient; in this case those elements within $\mathsf{S}_{\boldsymbol{\epsilon}(\mathbf{x})}$ which correspond to grid points v and w can be defined as in Eq. (19).

Again the model fitting should be done by maximisation of likelihood, but in this case the linear least-squares method cannot be applied, and estimating the additional covariance term in $\mathsf{S}_{\boldsymbol{\epsilon}(\mathbf{x})}$ requires numerical optimisation. Since performing numerical optimisation steps for all pairs of grid points would consume considerable time, we will in the next section present an approximative method which renders this problem accessible to the linear least-squares method.

5.2. Approximative Model for Spatially Correlated Innovations

Consider for a moment again the case of modelling a pair of (typically non-neighbouring) grid points v and w within a full spatiotemporal model, *without* assuming a non-zero covariance term of the corresponding innovations $\mathcal{E}\left(\epsilon_t(v)\epsilon_t(w)\right)$; then each grid point is modelled according to Eq. (26), and we have a bivariate sub-model with innovations

$$\begin{aligned}\epsilon_t(v) &= x_t^{(v)} - \left(\mu^{(v)} + \sum_{t'=1}^{p} a_{t'}^{(v)} x_{t-t'}^{(v)} + \sum_{u\in\mathcal{N}(v)} b_1^{(v,u)} x_{t-1}^{(u)}\right)\\ \epsilon_t(w) &= x_t^{(w)} - \left(\mu^{(w)} + \sum_{t'=1}^{p} a_{t'}^{(w)} x_{t-t'}^{(w)} + \sum_{u\in\mathcal{N}(w)} b_1^{(w,u)} x_{t-1}^{(u)}\right).\end{aligned} \tag{33}$$

Motivated by the work of Geweke,[44] we suggest to represent instantaneous correlations between $x_t^{(v)}$ and $x_t^{(w)}$, i.e. non-zero $\mathcal{E}\,(\epsilon_t(v)\epsilon_t(w))$, by introducing an additional *instantaneous* coupling term into the bivariate model, which yields a new bivariate sub-model with innovations

$$\begin{aligned}\epsilon_t(v) &= x_t^{(v)} - \left(\mu^{(v)} + \sum_{t'=1}^{p} a_{t'}^{(v)} x_{t-t'}^{(v)} + \sum_{u\in\mathcal{N}(v)} b_1^{(v,u)} x_{t-1}^{(u)}\right)\\ \epsilon_t(w|v) &= x_t^{(w)} - \left(\mu^{(w)} + \sum_{t'=1}^{p} a_{t'}^{(w)} x_{t-t'}^{(w)} + \sum_{u\in\mathcal{N}(w)} b_1^{(w,u)} x_{t-1}^{(u)} + c_{vw}\, x_t^{(v)}\right).\end{aligned} \tag{34}$$

Note that, unlike with the standard definition of autoregressive models, here we use the state value of one grid point at time t in order to model the value of another grid point *at the same time* t.

The additional coupling parameter c_{vw} can be estimated conveniently by the same linear least-squares method which is also employed for estimating the other parameters in Eq. (34), thus avoiding computationally more demanding numerical optimisation. In this model the covariance matrix of the innovations is guaranteed to be diagonal; its diagonal elements shall be denoted by $\sigma^2_{\epsilon(v)}$ and $\sigma^2_{\epsilon(w|v)}$. However, due to the inclusion of the instantaneous interaction term this covariance matrix does not directly refer to the original state variables $x_t^{(v)}$ and $x_t^{(w)}$, but to transformed variables; see Appendix B for a discussion of the corresponding non-diagonal covariance matrix of the original $x_t^{(v)}$ and $x_t^{(w)}$.

It has to be emphasised that the model underlying Eq. (34) is not equivalent to the original bivariate modelling, corresponding to Eq. (23), but rather it represents a useful approximation; in Appendix B we will show that the coupling parameter c_{vw} approximately corresponds to the coefficient of linear correlation $r\,(\epsilon(v), \epsilon(w))$. Also the result of the least-squares model fit does not represent a maximum-likelihood fit, although we expect that in most cases it will have very similar properties.

In Eq. (34) the two grid points v and w are no longer treated in a symmetrical way, since $x_t^{(v)}$ is modelled only by its own past (and the past of its neighbours), whereas $x_t^{(w)}$ is modelled by the past of both $x_t^{(v)}$ and $x_t^{(w)}$. Experience has shown that if the roles of v and w are interchanged, in most cases the results remain essentially unchanged; however, in certain situations this remark does no longer apply, e.g. when the variances of the original data $x_t^{(v)}$ and $x_t^{(w)}$ differ very much. In such cases we have adopted the approach to choose out of the two possibilities the model which achieves the better likelihood. In Appendix B an explicit expression for the likelihood of model Eq. (34) is derived.

As already mentioned, it would also be possible to augment the model Eq. (34) by time-lagged interaction terms, corresponding to the full bivariate modelling of Eq. (23); by such terms further deviations of the innovations $\boldsymbol{\epsilon}_t(\mathbf{x})$ from the ideal condition of being completely white and mutually independent could be modelled. In particular, this generalisation would offer the possibility of quantifying *directed* correlations between grid points, which can be interpreted as measuring causal connectivity (while by definition instantaneous correlations are non-directed).

Through direct comparison of the likelihoods obtained by models (33) and (34) it can be decided whether for given data the inclusion of the interaction term improves the model; this step can be regarded as a Likelihood Ratio Test (LRT).[45] Based on Eqs. (9) and (10), the LRT test statistic can also be used for deriving an estimator of mutual information; following this path we obtain

$$I\left(x^{(v)}, x^{(w)}\right) = (N_t - p)\frac{2c_{vw}\,\sigma^2_{\epsilon(v),\epsilon(w|v)} - c^2_{vw}\,\sigma^2_{\epsilon(v)}}{2\sigma^2_{\epsilon(w|v)}}, \tag{35}$$

where we have defined $\sigma^2_{\epsilon(v),\epsilon(w|v)} = \mathcal{E}\left(\epsilon_t(v)\epsilon_t(w|v)\right)$; see Appendix B for details on the derivation of this result. This estimator differs from most other currently known estimators by the complete absence of histograms or similar non-parametric elements.

However, it has to be mentioned that this estimator has the disadvantage of sometimes producing negative estimates of mutual information; some discussion on this point can be found in Appendix B. For small values of mutual information also nonparametric estimators are known to occasionally produce negative results which are due to statistical fluctuations.[21]

6. COMPARISON WITH INDEPENDENT COMPONENT ANALYSIS

The concept of whitening is also playing an important role in other approaches for the analysis of multivariate time series, such as Independent Component Analysis (ICA):[46] As a part of this method commonly a linear preprocessing step is applied to the multivariate data, also called “whitening,” such that the channels become mutually uncorrelated; they are also rescaled in order to have unit variance, such that the sample covariance matrix of the transformed data becomes a unity matrix. While the rescaling step has to be regarded as problematic, since in most cases it will deteriorate the signal-noise ratio of the data, it can be said that otherwise this step roughly corresponds to the multiplication with the Laplacian and the inclusion of instantaneous terms in Eq. (34), i.e. to the identification of the instantaneous correlation structure of the driving noise.

However, the concept of *temporal* whitening does not have a counterpart in the standard ICA methodology, since the description of data, as provided by ICA, ignores the constraint of causal modelling, i.e. the direction of time. The numerous ICA algorithms which have been proposed, differ in whether they take the temporal ordering of the data into account; while the majority of algorithms ignores the temporal ordering completely, some approaches aim at simultaneously diagonalising instantaneous and lagged covariance matrices,(47) but also these matrices do not depend on the direction of time.

Most ICA algorithms are based on the assumption of the existence of a set of source components $\mathbf{s}_t = (s_t^{(1)}, \ldots, s_t^{(N_s)})$, which are assumed to be mutually independent and to have non-Gaussian distributions, such that the data is modelled as

$$\mathbf{x}_t = \mathsf{C}\mathbf{s}_t, \tag{36}$$

where C denotes the $N_v \times N_s$ *mixing* matrix; frequently $N_v = N_s$ is chosen, but the case $N_v > N_s$ is also possible, as in Factor Analysis.(51) Both C and $\mathbf{s}_t$ are unknown and need to be estimated; in many ICA algorithms the assumption of non-Gaussianity of all components $s_t^{(v)}$ (except for at most one) is essential.

Obviously, the approach proposed in this paper differs distinctively from ICA, since it rests on the theorem from Markov process theory mentioned above, according to which a wide class of dynamical systems can be (temporally) whitened into *Gaussian* innovations, while deviations from Gaussianity in the data are assumed to be a result of nonlinear elements in the dynamics or in the observation process (see Sec. 3.1). Innovations at different grid points are not assumed to be independent or uncorrelated, but rather an explicit model for their non-diagonal covariance matrix is formulated. In the light of the theorem on Gaussian innovations, we would regard the whitening as unsuccessful if the resulting innovations were found to have a clearly non-Gaussian distribution (but, as mentioned in Sec. 3.1, a Poisson noise component may be accepted).

7. APPLICATION TO SIMULATED SPATIOTEMPORAL DYNAMICS

We will now illustrate the ideas presented in this paper by application to data generated by a simulated dynamical system of moderate complexity. The design of our simulation is motivated by earlier work of Schreiber,(48) but while he employed a strongly nonlinear deterministic system, we choose a stochastic system which is based on applying a sigmoid nonlinearity to a linear autoregressive structure;

the case of employing a stochastic version of Schreiber's system will be briefly discussed in Sec. 8.

This section discusses the definition of the system, the method to produce simulated data from the system and the results of analysing this data by the whitening approach, using the parsimonious MAR model, as presented above; for comparison we will also discuss the cases of analysing the data with a full MAR model (having a non-sparse transition matrix) and with standard ICA algorithms.

7.1. Chain of Coupled Stochastical Oscillators

The system consists of a one-dimensional chain of N_v nodes (representing "grid points") with periodic boundary conditions. At each node v, $v = 1, \dots, N_v$, a local dynamical process evolves, defined by

$$y_t^{(v)} = \tanh\left(a_1^{(v)} y_{t-1}^{(v)} + a_2^{(v)} y_{t-2}^{(v)} + b_1^{(v,v-1)} y_{t-1}^{(v-1)} + b_1^{(v,v+1)} y_{t-1}^{(v+1)}\right) + \eta_t^{(v)}. \tag{37}$$

By $\boldsymbol{H}_t = (\eta_t^{(1)}, \dots, \eta_t^{(N_v)})$ the vector of driving noise terms shall be denoted; its covariance is chosen as nondiagonal, following the form given by Eq. (30); the diagonal matrix corresponding to $\mathsf{S}_{\epsilon(\tilde{\mathbf{x}})}$ is created from a set of node-dependent standard deviations $\sigma_{\text{dyn}}^{(v)}$ which are generated by drawing N_v numbers from a Gaussian distribution $\mathcal{N}(\chi_\eta, \sigma_\eta)$ with $\chi_\eta \gg \sigma_\eta$ and additionally smoothing this set along the chain of nodes (observing periodic boundary conditions).

The sets of left and right neighbour interaction parameters $b^{(v,v-1)}$, $b^{(v,v+1)}$, $v = 1 \dots, N_v$ (again observing periodic boundary conditions), are created in the same way as the set of standard deviations $\sigma_{\text{dyn}}^{(v)}$. By choosing slightly different means χ_{b-}, χ_{b+} for the Gaussian distributions of left and right neighbour interaction parameters, asymmetries can be introduced into the system, and more "interesting" dynamics results. The local dynamical parameters $a_1^{(v)}$, $a_2^{(v)}$ are chosen according to the rules for designing stable second-order autoregressive processes, i.e. such that the roots of the corresponding characteristic equation

$$\left(\lambda^{(v)}\right)^2 - a_1^{(v)}\lambda^{(v)} - a_2^{(v)} = 0, \tag{38}$$

which shall be written as

$$\lambda_\pm^{(v)} = r^{(v)}\, e^{\pm i\varphi^{(v)}}, \tag{39}$$

lie within the unit circle in the complex plane (but the resulting dynamics will become more "interesting," if a few moduli are allowed to be slightly larger than one). Moduli $r^{(v)}$ and phases $\varphi^{(v)}$ can be chosen independently from suitably chosen distributions; we have decided to choose $r^{(v)}$ from a Gaussian distribution

$\mathcal{N}(\chi_r, \sigma_r)$ and $\varphi^{(v)}$ from the asymmetric distribution

$$p(\varphi) = -\log\left(p_{\mathcal{U}}(\varphi)(1 - e^{-\pi}) + e^{-\pi}\right), \tag{40}$$

where $p_{\mathcal{U}}(\varphi)$ is the uniform distribution on the interval [0, 1]. Parameters then follow by

$$a_1^{(v)} = 2\, r^{(v)} \cos\varphi^{(v)}, \qquad a_2^{(v)} = -\left(r^{(v)}\right)^2. \tag{41}$$

Again smoothing along the chain is applied, in order to enforce smoothly varying properties of local dynamics.

The hyperbolic tangent in Eq. (37) has been introduced following the example of the sigmoid nonlinearity in Neural Network models; its main purpose is to prevent divergence of the dynamics which may arise due to some roots $\lambda_{\pm}^{(v)}$ lying outside the unit circle or due to feedback between neighbours. Furthermore this pronounced nonlinearity contributes to creating "interesting" dynamics.

Instantaneous long-distance correlations can be introduced into this system by choosing two (non-neighbouring) nodes v and w and driving them by a common noise process (but preserving the local variances of the two nodes)

$$\tilde{\eta}_t^{(m)} = \frac{\sigma_{\text{dyn}}^{(m)}}{2}\left(\frac{\eta_t^{(v)}}{\sigma_{\text{dyn}}^{(v)}} + \frac{\eta_t^{(w)}}{\sigma_{\text{dyn}}^{(w)}}\right), \quad m = v, w; \tag{42}$$

the resulting instantaneous long-distance correlations may be termed "intrinsic," in contrast to those which occur either due to the dynamical coupling of nodes within local neighbourhoods, or as a result of pure coincidence, as it is not uncommon in short time series.

This system implements a coupled-map lattice consisting of stochastically driven oscillators;(14) after allowing for a transient to die out, data is sampled from the system by recording noisy observations of the states of each node:

$$x_t^{(v)} = y_t^{(v)} + n_t^{(v)}, \tag{43}$$

where the observational noise $n_t^{(v)}$ is sampled from $\mathcal{N}(0, \sigma_{\text{obs}}^{(v)})$; the node-dependent standard deviations $\sigma_{\text{obs}}^{(v)}$ are drawn from $\mathcal{N}(\chi_n, \sigma_n)$ (where $\chi_n \gg \sigma_n$) and afterwards smoothed along the chain. Observation noise is uncorrelated between nodes.

7.2. Results of Simulation Study

We implement the dynamical system described in the previous section with $N_v = 64$ nodes. Parameters of the Gaussian distributions are chosen as $(\chi_\eta, \sigma_\eta) = (0.25, 0.05)$, $(\chi_{b-}, \sigma_{b-}) = (0.55, 0.25)$, $(\chi_{b+}, \sigma_{b+}) = (0.45, 0.2)$, $(\chi_r, \sigma_r) = (0.5, 0.2)$ and $(\chi_n, \sigma_n) = (0.1, 0.01)$. Grid points 16 and 41 are correlated according to Eq. (42). The system is simulated, starting from random

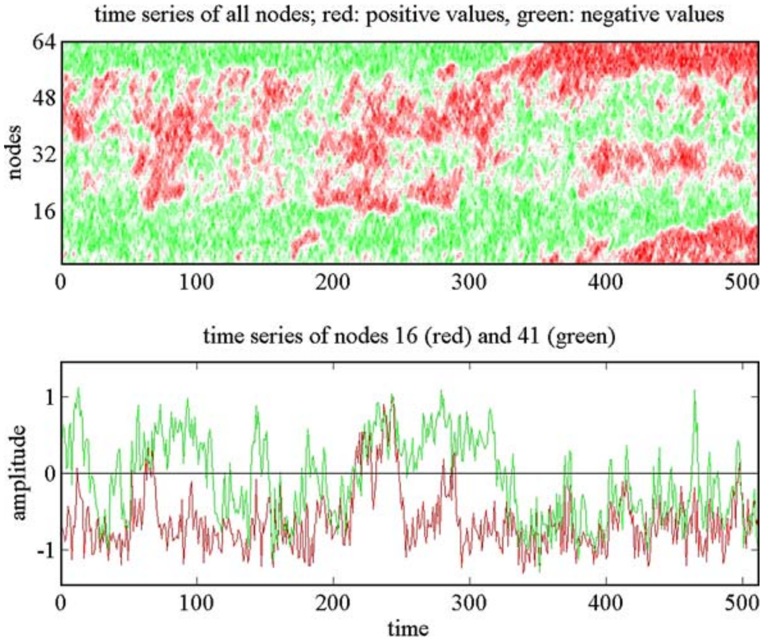

Fig. 1. Simulated dynamics for a stochastic coupled-map chain consisting of 64 nodes; upper panel: amplitude (colour-coded) vs. node number (vertical axis) and time (horizontal axis); periodic boundary conditions apply to the vertical axis. Lower panel: time series of amplitudes at nodes 16 (red) and 41 (green) vs. time.

initial conditions, and after waiting for 1000 time points in order to allow for transient behaviour, a multivariate data set of $N_t = 512$ points length is recorded from all 64 nodes, including observational noise. The data is displayed in Fig. 1; the upper panel shows the time series of all nodes along the chain, while the time series of the intrinsically correlated nodes 16 and 41 are shown explicitly in the lower panel. In the figure it can be seen that for the selected values of parameters groups of nodes switch randomly between predominantly positive or negative values, thereby forming coherent spatiotemporal patterns. Furthermore it can be seen that, at least with respect to visual inspection, the correlation between nodes 16 and 41 is not very pronounced.

The same point is demonstrated in Fig. 2 where for all pairs of nodes (v, w) the linear correlation measure $r(v, w) := r(x^{(v)}, x^{(w)})$ according to Eq. (3) (upper panel) and the mutual information $I(v, w) := I(x^{(v)}, x^{(w)})$ according to Eq. (4) (lower panel) are shown. The estimate of $I(v, w)$ is obtained by using a MATLAB implementation of a histogram-based estimator, provided by Moddemeijer.(22) Note that in this figure both quantities are calculated directly for the data $x_t^{(v)}$. The

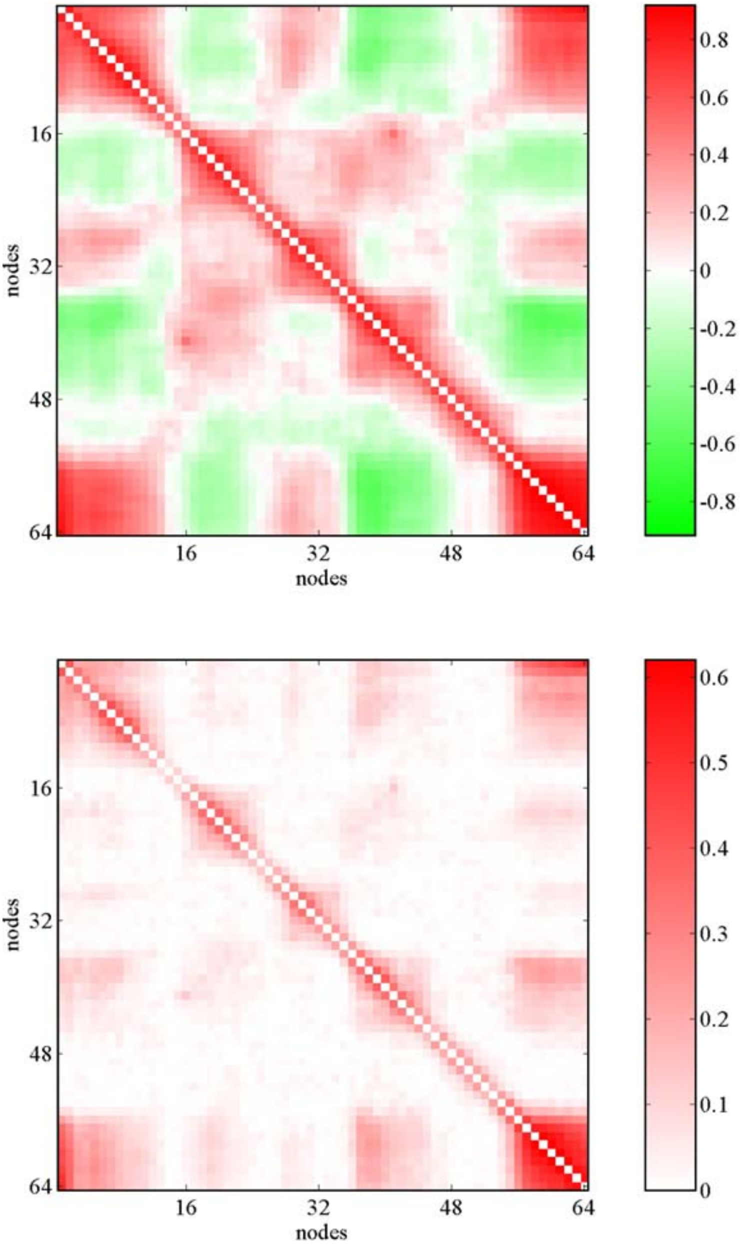

Fig. 2. Linear correlation matrix (upper panel) and mutual information matrix (lower panel) for all pairs of nodes, estimated from the multivariate time series shown in Fig. 1. Mutual information was estimated by a histogram estimator. In both panels values on the diagonal have been omitted.

matrix of linear correlations shows broad areas of correlated or anti-correlated nodes, and the pair (16, 41), producing a value of $r(16, 41) = 0.5006$ (while r is bounded by definition to the interval $[-1,1]$), cannot be easily distinguished from these areas, although its correlation is of completely different origin than all other correlations which are visible in the figure. The matrix of mutual information reproduces some of the features of the matrix of linear correlations, but without distinguishing between correlation and anti-correlation; in this case the intrinsically correlated pair produces a rather low value of $I(16, 41) = 0.1286$ (here it should be mentioned that the typical scaling of this estimate depends on certain design choices of the mutual information estimator which we shall not discuss here in detail).

Clearly the particular estimates obtained for linear correlation and mutual information are to be regarded as not very reliable, since the underlying time series is rather short, $N_t = 512$. Repeating the analyses 1000 times for different realisations generated by the same system (each of 512 points length) yields, in terms of means $m(.)$ and standard deviations $s(.)$, for linear correlation $m\,(r(16, 41)) = 0.2536$ and $s\,(r(16, 41)) = 0.1741$, and for mutual information $m\,(I(16, 41)) = 0.0841$ and $s\,(I(16, 41)) = 0.0344$, i.e. fairly broad distributions. Note that in real life frequently the option of repeating analyses for many equivalent time series recorded under exactly preserved conditions may not be given.

We would like to briefly discuss the distribution of the simulated data. In Fig. 3 skewness and kurtosis are shown for the time series recorded for each node (open circles). It can be seen that some nodes display pronounced skewness, i.e. asymmetric distributions; and while many nodes display negative values of kurtosis, some reach sizable positive values. These results show a substantial deviation from Gaussianity, which is a result of the sigmoid nonlinearity in the

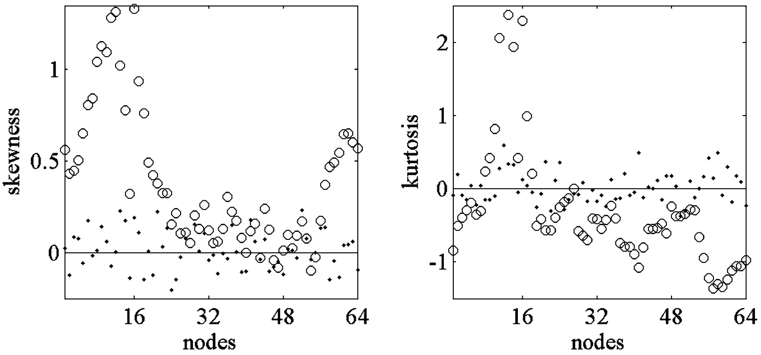

Fig. 3. Skewness (left panel) and kurtosis (right panel) vs. node number (horizontal axis) calculated from the multivariate time series shown in Fig. 1 (open circles) and from the corresponding innovations (black dots).

dynamics. Negative values of kurtosis correspond to thin-tail behaviour, as it should be expected for a sigmoid function, but it is obvious that not all nodes follow this pattern, since the distribution of some nodes is dominated by asymmetries.

Next we perform spatial whitening by applying the transformation according to Eqs. (27) and (28), and temporal whitening by fitting to each node the model Eq. (26); note that this is a completely linear model, in contradiction to the true model used for generating the data, Eq. (37). We are employing an incorrect model class, nevertheless very good whitening can be achieved. The parameter k in Eq. (28) is chosen as 9, although one would rather expect 2 in a one-dimensional system; but we found that too strong spatial whitening produces spurious temporal correlations (or, more precisely, anticorrelations), especially for pairs of neighbouring nodes, therefore we recommend to avoid too high values for $1/k$.

The success of the whitening transformation can be see in Fig. 4, where again linear correlation (upper panel) and mutual information (lower panel) are shown, but now for the innovations $\epsilon_t(v)$ of the spatial and temporal whitening steps instead for the original data $x_t^{(v)}$. In both panels the pair (16, 41) stands out clearly against a background of very small values, assuming a correlation value of $r(16, 41) = 0.6959$, while the second largest value in the matrix is 0.1560, and a mutual information value of $I(16, 41) = 0.2485$, while the second largest value is 0.0310. The values for repeating the analysis for 1000 different realisations are for linear correlation $m\,(r(16, 41)) = 0.6735$ and $s\,(r(16, 41)) = 0.0262$, and for mutual information $m\,(I(16, 41)) = 0.2404$ and $s\,(I(16, 41)) = 0.0290$; these distributions are much narrower than obtained above in the case without whitening, thereby demonstrating the degree of precision that can be achieved on the basis of time series of only 512 points length. Note that if the innovations could be estimated without any error, we would expect $r(16, 41) = 1.0$, but due to observational noise, model imperfection and finite-sample effects this value cannot be attained. The theoretical optimal value for $I(16, 41)$ is given by the entropy of the common innovations.

If we investigate the distribution of the innovations, we obtain the results shown by dots in Fig. 3. It can be seen that for most nodes both skewness and kurtosis are much closer to zero now; especially the pronounced asymmetries have been removed. The values of these two quantities scatter around zero for the set of all nodes, with mean values very close to zero, 0.0105 for skewness and 0.0175 for kurtosis. These results confirm that the whitening step has produced approximately Gaussian innovations, despite fitting a linear model to nonlinear data.

Furthermore we apply to each pair of nodes the model comparison given by Eqs. (33) and (34) and compute the corresponding parametric estimate of mutual information, given by Eq. (35); the result is shown in Fig. 5. From the figure it can be seen that also by this method the intrinsic correlation of the pair (16, 41) is clearly

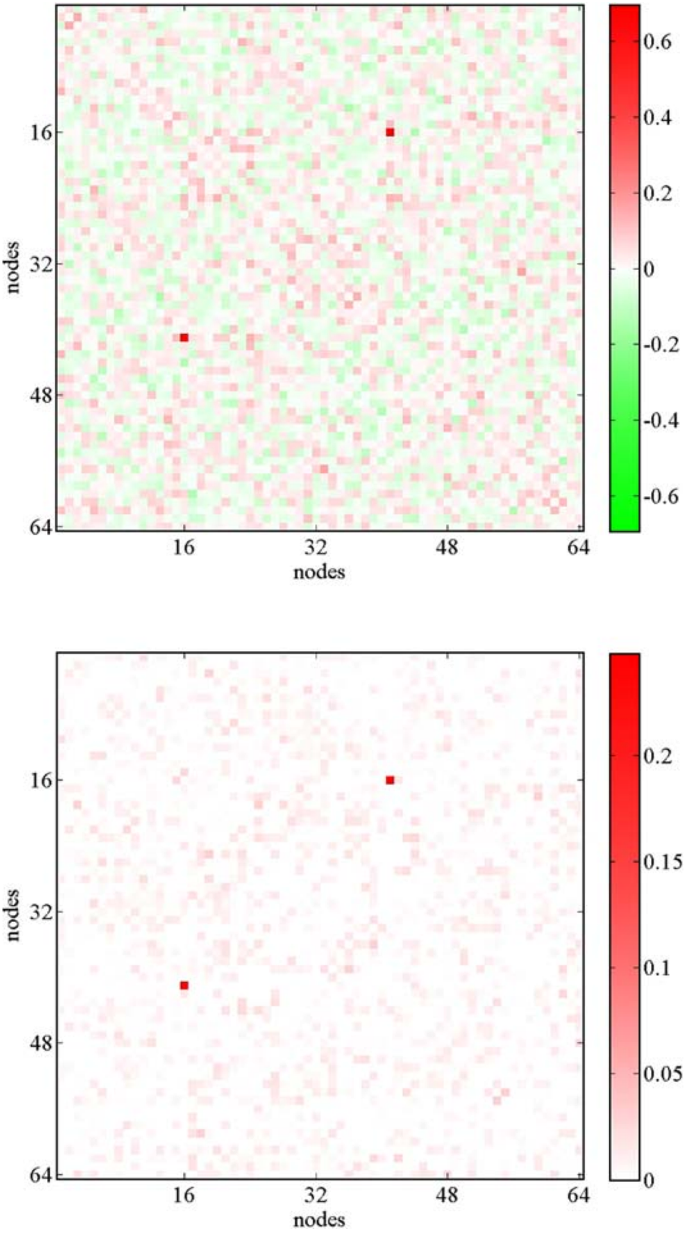

Fig. 4. Linear correlation matrix (upper panel) and mutual information matrix (lower panel) for all pairs of nodes, estimated from the innovations of the multivariate time series shown in Fig. 1. Mutual information was estimated by a histogram estimator. In both panels values on the diagonal have been omitted.

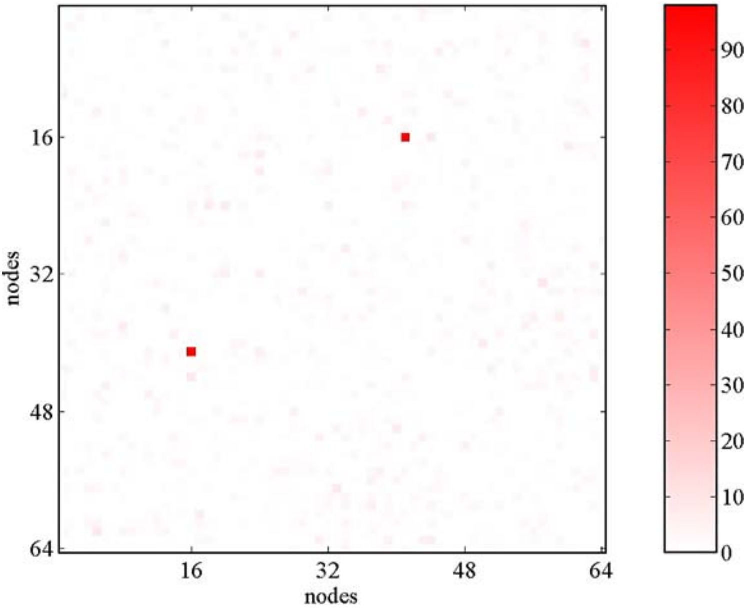

Fig. 5. Mutual information matrix for all pairs of nodes, estimated by the parametric estimator based on the difference of log-likelihood of pairwise model comparisons from the multivariate time series shown in Fig. 1. Values on the diagonal have been omitted.

detected, assuming a value of $I(16, 41) = 98.1332$, while the second largest value is 5.7633. Obviously the scaling of this estimate of mutual information is different from that based on histograms, but the relative behaviour is essentially the same. The values for a distribution of 1000 realisations are $m\,(I(16, 41)) = 88.7080$ and $s\,(I(16, 41)) = 21.3805$.

Finally we investigate how the performance of the whitening approach in correctly identifying intrinsically correlated voxel pairs depends on time series length N_t and standard deviation of observational noise χ_n (the latter averaged over nodes). For this purpose we formulate the null hypothesis "there is no intrinsic correlation for the pair (16, 41)" and try to reject it based on simple ranking of the values of r and I for all pairs of voxels: If $r(16, 41) > r(v, w)$ for all pairs $(v, w) \neq (16, 41)$ the null hypothesis is rejected. Instead of the coefficient of linear correlation r also the mutual information I, estimated by either a histogram estimator or by Eq. (35), may be used. While repeating this procedure 500 times (each time choosing new parameters for the simulated system according to the distributions given above, keeping parameters of the distributions fixed at the values given above, and choosing a new set of observation noise covariances $\sigma_{\text{obs}}^{(v)}$ from $\mathcal{N}(\chi_n, \sigma_n)$, where $\sigma_n = \chi_n/10$) for each pair of values of N_t and χ_n, we

Table I. Estimate of time series length N_t required for achieving probability of type-II error $p_{\mathrm{II}} = 0.0, \ldots, 0.75$ for given (average) observation noise standard deviation $\chi_n = 0.02, 0.1, 0.2, 0.4$, using correlation coefficient, histogram estimate of mutual information or parametric estimate of mutual information (by Eq. (35))

$p_{\mathrm{II}} =$	0.0	0.25	0.50	0.75
	correlation coefficient r			
$\chi_n = 0.02$	50	30	25	20
0.1	120	45	40	30
0.2	300	135	100	70
0.4	>700	>700	650	375
	mut. inf. I (histogram estimate)			
$\chi_n = 0.02$	105	60	50	40
0.1	240	130	100	80
0.2	700	430	330	240
0.4	>700	>700	>700	>700
	mut. inf. I (parametric estimate)			
$\chi_n = 0.02$	80	30	25	20
0.1	150	55	45	30
0.2	480	180	120	80
0.4	>700	>700	>700	690

Note. Values have been estimated by simulation (see text for details) and rounded to multiples of 5.

record the number of failures to reject the null hypothesis, i.e. we measure the rate of type-II errors p_{II}.

When applying this analysis to estimates of linear correlation and of mutual information obtained from original (unwhitened) data, p_{II} is always larger than 0.9 and in most cases 1.0, in agreement with the results presented so far; for estimates obtained from the innovations we obtain the results shown in Table 1. The table gives the approximate length of time series N_t which is, at a given value of χ_n, required to reduce the probability of type-II error p_{II} to a given value; values for N_t have been rounded to multiples of 5. Note that according to Eq. (37) the distribution of the simulated dynamics is essentially bounded by $-1 \leq y_t^{(v)} \leq 1$; in fact we find that the sample standard deviations of the simulated data sets *before* adding observational noise scatter around a value of 0.6. This value is to be compared with the values for χ_n given in the table.

From the table it can be seen that the parametric estimate for mutual information provides better detection of intrinsic correlation than the histogram estimate, while the linear correlation coefficient displays the best performance. In summary it can be stated that after whitening intrinsic correlation can be detected reliably in time series of a few hundred points length even in the presence of strong observational noise.

7.3. Application of Full MAR Analysis

For the purpose of comparison, we shall consider the case of fitting a full MAR model, according to Eq. (24), to the same data as studied so far, instead of a parsimonious model. In case of the smallest model order $p = 1$, as in POP analysis, we find from Eq. (25) that the number of model parameters N_{par} is 18.85% of the total number of available data values $N_v N_t$, while $p = 2$ (which is the "true" model order of this simulation) yields 31.35%, so both values exceed the 10%-rule mentioned in Sec. 4.1; the parsimonious model Eq. (32) gives a much more favourable ratio of 0.98%.

Ignoring this problem for a moment and fitting a MAR(2) model (i.e. a MAR model with model order $p = 2$) by Whittle's algorithm,[39] we obtain transition matrices A_1 and A_2; as should be expected we find that these matrices appear "blurred" as compared to the correct matrices employed in the simulation, i.e. the non-vanishing values are systematically underestimated (in terms of their absolute values; diagonal elements of A_2 which should always be negative, are also sometimes estimated as positive), while the vanishing values (in A_1 all elements referring to non-neighbouring pairs of nodes, and in A_2 all off-diagonal elements) assume non-zero values, scattering around zero. A similar remark applies to the covariance matrix of the driving noise, $\mathsf{S}_{\epsilon(\mathbf{x})}$. The intrinsic correlation between nodes 16 and 41 can be detected, also by this analysis: it appears now partly in the corresponding off-diagonal elements of the estimate of $\mathsf{S}_{\epsilon(\mathbf{x})}$ (which in the parsimonious model, due to the Laplacian transformation, corresponds to a diagonal matrix), and partly in the remaining correlations of the innovations of the full MAR model fit. Naturally, due to the overparametrisation, this intrinsic correlation will easily be lost in noise, resulting either from estimation errors or being present in the data itself.

In this simulation the data was created by a parsimonious MAR model, so it is not surprising that a similar model shows better performance than the full MAR model; in the analysis of real-world data it will depend on the properties of the data and the underlying system which model is more appropriate. In systems for which the assumption of fast local dynamics is invalid, the full MAR model may indeed be superior. Whether the much larger amount of model parameters is justified, can be determined by information criteria such as AIC and BIC;[36] e.g. for the example of the simulation presented above, we find that the AIC of the MAR models of first and second order assume values of 15158.61 and 17255.94, respectively, while the parsimonious model assumes a much lower value of 10044.47. We remark that the pure likelihood of the MAR models is *larger* than that of the parsimonious model; however, due to their large numbers of parameters, these models are heavily punished by AIC for overfitting. But since in this case the number of parameters is too large for reliable estimation of AR models anyway, compared to the number of available data values, the parsimonious model provides a convenient alternative.

7.4. Application of ICA

It is possible to apply ICA algorithms to the simulated data shown in Fig. 1; however, in doing so we impose the constraint of the existence of a set of non-Gaussian independent signals, which are related to the data by a mixing matrix, as shown in Eq. (36). The only independent signals which were involved in generating the data, are the driving noises $\boldsymbol{H}_t = (\eta_t^{(1)}, \ldots, \eta_t^{(N_v)})$, but their relationship to the data is of more complicated nature than described by Eq. (36), therefore these noises cannot be retrieved by ICA; the estimation of the driving noises is the goal of the whitening approach (assuming that observation noise can be neglected) and requires a dynamical model of the process generating the data. Furthermore, the available knowledge about neighbourhood relationships between nodes is ignored by standard ICA algorithms.

If we nevertheless perform an ICA decomposition of the simulated data, employing the "FastICA" algorithm introduced by Hyvärinen[46] (using "symmetric" estimation, i.e. estimating all independent components in parallel) we obtain the results shown in Fig. 6. The algorithm produces a set of 64 components, but there is no preferred ordering of these components, since the spatial neighbourhood relationships of the data have been lost; for this reason the plot shows no equivalent to spatial structures. In the lower panel of Fig. 6 two arbitrarily chosen components are shown explicitly; it can be seen that they display little temporal structure, except for one sharp spike-like maximum of each component. Upon closer inspection it is found that such sharp extrema are present for most of the independent components; since they never occur at the same time point for two components, such spikes contribute to reducing the residual dependencies between components.

Indeed, the residual correlations between the components provided by FastICA are virtually zero: the maximum value of linear correlation among all pairs of components is $r = 5.79 \times 10^{-16}$; for the residual mutual information, using the same histogram-based estimator as before, a value of $I = 0.0115$ is found. These values show that this decomposition was successful, with respect to the underlying assumption of the ICA approach; nevertheless, the decomposition provides no useful information about the actual dynamics of this particular system and about its spatial correlation structures, such as the existence of one pair of intrinsically correlated non-neighbouring nodes.

Finally, we remark that some ICA algorithms have been proposed which are based on maximising a (log-)likelihood, as it is also the case in time series

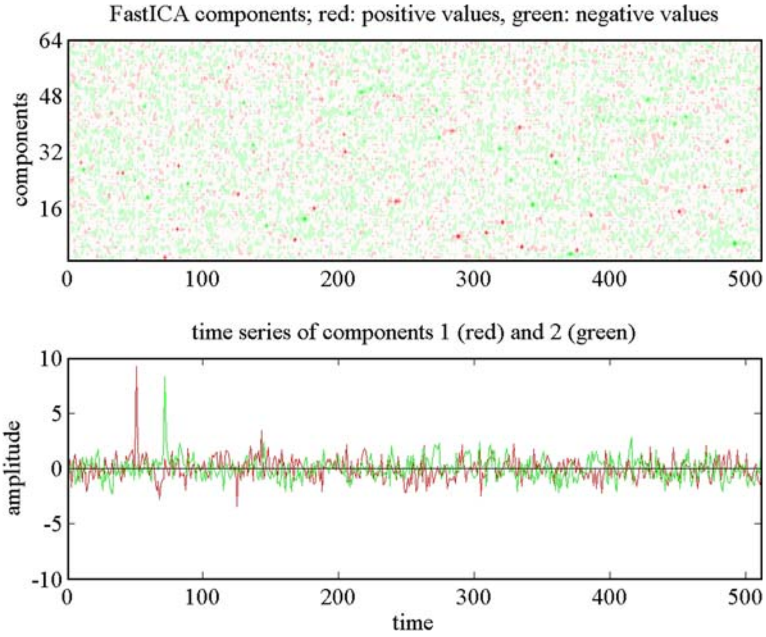

Fig. 6. FastICA decomposition of the data shown in Fig. (1); upper panel: amplitude (colour-coded) vs. component number (vertical axis) and time (horizontal axis); ordering of components is arbitrary. Lower panel: time series of amplitudes of components 1 (red) and 2 (green) vs. time.

modelling by dynamical models; since likelihood (after the bias correction against overfitting, i.e. employing AIC or BIC, see above) represents a global measure of the quality of models for given data, the performance of these methods can be compared directly to time series modelling. As examples we mention the work of Wu and Principe(49) and of Choi *et al.*,(50) both of which are based on the "generalised Gaussian distribution," a class of distributions which contains the Gaussian distribution as a particular case. Fitting ICA models based on this class of distributions to data bears close similarity to the likelihood-based version of classical Factor Analysis,(51) but without the constraint of Gaussianity. By applying an ICA algorithm based on these ideas to the same simulated data as before, we obtain a value of AIC=24986.08. The fact that this value is much larger than the AIC of parsimonious MAR modelling, AIC=10044.47 (thereby demonstrating a much inferior description of the data, as compared to parsimonious MAR), reflects partly the benefit of including dynamics into the model, i.e. of temporal whitening, and partly the penalty for the much larger number of parameters of the ICA model (which in this case is dominated by the N_v^2 parameters in the mixing matrix C); but in this case even the pure likelihood of the parsimonious MAR model is considerably larger than the likelihood of the ICA model.

8. DISCUSSION AND CONCLUSION

In this paper we have explored a new viewpoint concerning the issue of definition and estimation of mutual information between time series within multivariate data sets sampled from spatially extended dynamical systems; the link between mutual information and log-likelihood which we have demonstrated, applies to the estimation of mutual information between any pair of temporally correlated time series. We have presented a conceptually simple approach for parsimonious spatiotemporal modelling of data, which is based on the notion of whitening with respect to both space and time. The whitening approach proposed in this paper can be summarised by the following steps:

1. For each time point the data vector is multiplied by the appropriate Laplacian matrix (see Eqs. (27) and (28)); thereby the data is spatially whitened.
2. The resulting transformed data is modelled by a suitable MAR model (see Eq. (26)); thereby temporal whitening is provided. For high-dimensional time series (representing a large number of grid points in space) parsimonious modelling is achieved by limiting direct interactions (i.e. non-zero model parameters in the transition matrices) to neighbours.
3. The resulting innovations are analysed for remaining correlations; such correlations will usually be instantaneous spatial correlations, as demonstrated in the simulation example, but they could also involve time lags between different spatial locations.
4. Finally there is the option of refitting the autoregressive model (i.e. repeating step 2), but augmented by the additional correlations identified in step 3; an example, limited to one pair of grid points, is given in Eq. (34).

Provided the model is sufficiently successful in whitening the data, the innovations can be assumed to have a multivariate Gaussian distribution without temporal or spatial correlations, therefore they represent a convenient basis for the estimation of mutual information, and hence connectivity, as has been shown in this paper. By exploiting this property of innovations we have derived a parametric estimator of mutual information, which may replace the commonly used non-parametric estimators. But it should be stressed that the crucial point is the removal of spatiotemporal correlations, i.e. the twofold whitening step; once this has been done, also non-parametric estimators provide considerably improved results.

It represents a typical situation in present-day scientific research that spatially extended dynamical systems are investigated by sampling at discrete spatial positions and time points. It may be argued that bivariate mutual information was insufficient for such situations since it does not properly address the various potential sources for spatial and temporal correlations. One possible extension of the standard definition is given by conditional mutual information;(12) whitening by explicit modelling of the dynamics with respect to both space and time provides an alternative approach. As we have demonstrated in the simulation example, it is not required that the correct dynamical model be employed; we expect that for a large fraction of cases a linear low-order autoregressive model with nearest-neighbour interactions may prove to be a useful approximation.

As a natural consequence of the methodology discussed in this paper, the estimate of mutual information becomes dependent on the dynamical model used for whitening; as an example, compare the lower panels of Figs. 2 and 4: Whereas in Fig. 2 all correlations contribute to the mutual information, in Fig. 4 most correlations have been removed by whitening, and only those remain which are not captured by the model. By this method it becomes possible to decompose correlations into different layers, such that each refinement of the model renders it possible to describe more correlations through the model, and delegate less correlations to the class of "unexplained" correlations remaining in the innovations.

Whether explicitly nonlinear model classes need to be employed, depends on the data; generally, strong deviations from Gaussianity will indicate the necessity of employing nonlinear elements in the modelling of the dynamics or the observation process. The case of data generated by a sigmoid nonlinearity is, to some degree, a well-behaved case, since it corresponds to thin-tail deviations from Gaussianity. On the other hand, it has been demonstrated in this paper that also pronounced asymmetries can efficiently be mapped back to approximately Gaussian distributions by employing a simplified linear model. We do not yet know how this model class would perform in the presence of heavy-tail distributions; further work is required in order to obtain more experience with respect to these issues.

While in this paper we have relied upon using a fairly simple model class for the dynamics, it has to be admitted that the true dynamics of many spatially extended systems occurring in Nature may be considerably more complex, and therefore it may in some cases be very difficult to find efficient models for whitening the data. In such situations it will be necessary to explore more sophisticated model classes and to spend considerable effort on choice and adaptation of the models, possibly employing prior external knowledge about the dynamical properties of the system in question.

Also in simulations it is possible to find systems for which the approach as described in this paper fails. As an example we mention the logistic map lattice used in Ref. 48; a stochastic version of this lattice could be defined by (again using a one-dimensional chain of points with periodic boundary conditions)

$$y_t^{(v)} = a_1^{(v)} f\left(y_{t-1}^{(v)}\right) + b_1^{(v,v-1)} f\left(y_{t-1}^{(v-1)}\right) + b_1^{(v,v+1)} f\left(y_{t-1}^{(v+1)}\right) + \eta_t^{(v)}, \quad (44)$$

where $f(.)$ denotes the logistic map

$$f(y) = c - y^2; \tag{45}$$

here c denotes a constant parameter. Obviously Eq. (44) represents a strongly nonlinear system that cannot be easily approximated by linear models. When repeating all numerical experiments of this paper with data generated by Eq. (44) (using $c = 1.99$) instead of Eq. (37), again driving two nodes by a common noise input, none of the matrices of correlation and mutual information, whether for raw data or innovations, detects this intrinsically correlated pair. This would probably require modelling the data by employing the correct model class, as given by Eqs. (44) and (45). For real data the correct model class remains unknown; in this case the performance of linear and nonlinear models needs to be compared by their corresponding likelihoods, or preferably by their values of AIC or BIC.

As has already been mentioned, it is possible to generalise the analysis discussed in this paper to the case of lagged correlations between different points in space. While here we have confined our attention to instantaneous correlations in the innovations, it would be easily possible to analyse more general cross-correlations within the innovations, and thereby investigate directed connectivity and *causal* relationships between pairs of points in space. Again we presume that it may be useful and even necessary to first remove the layers of easily explainable correlations, before the deeper causality relationships can be uncovered.

Finally we mention that, not surprisingly, the results of modelling and of estimation of mutual information will be the more reliable, the longer the available time series are. By using a time series of $N_t = 512$ points length in the simulation we have chosen an intermediate case; when using longer time series, in the raw data the intrinsic correlation between pairs of nodes will gradually become completely hidden, whereas in Fig. 2 it is still possible to discern positive values of $r(16, 41)$ and $I(16, 41)$ which weakly stand out against the background. In such cases whitening approaches to uncovering intrinsic correlations become even more useful. On the other hand, in various fields, such as biomedical data analysis or palaeoclimatic research, it is not uncommon that the length of the available time series is of the order of only 100 time points or even shorter. We have found in additional simulations that also in such unfavourable cases the whitening approach succeeds in providing results similar to those shown in Figs. 4 and 5; however, in such cases the actual appearance of the matrices $r(v, w)$ and $I(v, w)$ will strongly depend on the particular realisation sampled from the system.

We are currently applying the methodology developed in this paper to spatiotemporal data sets obtained by functional magnetic resonance imaging (fMRI) from the brains of volunteers participating in stimulation experiments; the detailed results of these investigations will be presented and discussed elsewhere. However, we expect that the theory and methodology of whitening through parsimonious

MAR modelling will bear relevance for the analysis of spatiotemporal data sets arising in numerous fields of science.

APPENDIX A: MUTUAL INFORMATION OF BIVARIATE INNOVATION TIME SERIES

In order to evaluate Eq. (8), the two main contributions from the log-likelihoods, Eqs. (17) and (18), need to be evaluated, while the constant terms (those including $\log(2\pi)$) cancel out. The first contribution is given by (omitting the factor (-1/2))

$$N_t \log |\hat{\mathsf{S}}_{\epsilon(x,y|x,y)}| - N_t \log \hat{\sigma}^2_{\epsilon(x|x)} - N_t \log \hat{\sigma}^2_{\epsilon(y|y)}; \tag{46}$$

by inserting the determinant of Eq. (19) and rearranging, this contribution yields readily

$$\begin{aligned} & N_t \log(1 - \hat{r}^2(\epsilon(x), \epsilon(y))) + N_t \left(\log \hat{\sigma}^2_{\epsilon(x|x,y)} - \log \hat{\sigma}^2_{\epsilon(x|x)} \right) \\ & + N_t \left(\log \hat{\sigma}^2_{\epsilon(y|x,y)} - \log \hat{\sigma}^2_{\epsilon(y|y)} \right); \end{aligned} \tag{47}$$

here the "hat" notation denotes the maximum-likelihood (ML) estimates of the corresponding parameters.

The second contribution is given by (again omitting $(-1/2)$)

$$\begin{aligned} & \sum_t (\epsilon_t(x|x,y), \epsilon_t(y|x,y)) \, \hat{\mathsf{S}}^{-1}_{\epsilon(x,y|x,y)} \, (\epsilon_t(x|x,y), \epsilon_t(y|x,y))^\dagger \\ & - \sum_t \frac{\epsilon_t^2(x|x)}{\hat{\sigma}^2_{\epsilon(x|x)}} - \sum_t \frac{\epsilon_t^2(y|y)}{\hat{\sigma}^2_{\epsilon(y|y)}}. \end{aligned} \tag{48}$$

In order to evaluate this contribution the appropriate ML estimates need to be inserted explicitly. From a standard theorem of multivariate statistics (see e.g. Theorem 4.3.1 in Ref. 52) it follows that the ML estimate of $\mathsf{S}_{\epsilon(x,y|x,y)}$ is given by

$$\hat{\mathsf{S}}_{\epsilon(x,y|x,y)} = \begin{pmatrix} N_t^{-1} \sum_t \epsilon^2(x|x,y) & N_t^{-1} \sum_t \epsilon(x|x,y)\epsilon(y|x,y) \\ N_t^{-1} \sum_t \epsilon(x|x,y)\epsilon(y|x,y) & N_t^{-1} \sum_t \epsilon^2(y|x,y) \end{pmatrix}. \tag{49}$$

By comparing Eqs. (19) and (49) the ML estimators of the variances and the linear correlation coefficient result as

$$\hat{\sigma}^2_{\epsilon(.|.)} = \frac{1}{N_t} \sum_t \epsilon^2(.|.) \quad \text{and} \quad \hat{r}\,(\epsilon(x), \epsilon(y)) = \frac{\sum_t \epsilon(x|x,y)\epsilon(y|x,y)}{N_t \hat{\sigma}(x|x,y) \hat{\sigma}(y|x,y)}, \tag{50}$$

respectively. Here $\epsilon(.|.)$ stands for any of the combinations $\epsilon(x|x)$, $\epsilon(y|y)$, $\epsilon(x|x,y)$ and $\epsilon(y|x,y)$. By inserting these estimates, the second and third sum in expression (48) each reduce to N_t, while the first sum, after inserting the inverse of Eq. (19) and some further transformations, is found to reduce to $2N_t$, such that expression (48), i.e. the second contribution, altogether vanishes.

Therefore the mutual information is given by taking expression (47) times $(-1/2)$, as claimed in Eq. (20).

APPENDIX B: LOG-LIKELIHOOD FOR BIVARIATE TIME SERIES WITH INSTANTANEOUS COUPLING

For simplicity, in this appendix we shall not use different notation for a statistical quantity and its estimate. We start by rewriting Eq. (34) as

$$\begin{pmatrix} x_t^{(v)} \\ x_t^{(w)} \end{pmatrix} = \mathsf{C}^{-1} \begin{pmatrix} \mu^{(v)} + \sum_{t'} a_{t'}^{(v)} x_{t-t'}^{(v)} + \sum_{u} b_1^{(v,u)} x_{t-1}^{(u)} \\ \mu^{(w)} + \sum_{t'} a_{t'}^{(w)} x_{t-t'}^{(w)} + \sum_{u} b_1^{(w,u)} x_{t-1}^{(u)} \end{pmatrix} + \mathsf{C}^{-1} \begin{pmatrix} \epsilon_t(v) \\ \epsilon_t(w|v) \end{pmatrix}, \tag{51}$$

where we have defined $\mathsf{C} = \begin{pmatrix} 1 & 0 \\ -c_{vw} & 1 \end{pmatrix}$. The innovation term in Eq. (51) is given by $\mathsf{C}^{-1}\,(\epsilon_t(v), \epsilon_t(w|v))^\dagger$, and the corresponding covariance matrix follows as

$$\mathsf{S}_{\epsilon(v,w|v)} = \mathsf{C}^{-1} \begin{pmatrix} \sigma^2_{\epsilon(v)} & 0 \\ 0 & \sigma^2_{\epsilon(w|v)} \end{pmatrix} (\mathsf{C}^{-1})^\dagger = \begin{pmatrix} \sigma^2_{\epsilon(v)} & c_{vw}\sigma^2_{\epsilon(v)} \\ c_{vw}\sigma^2_{\epsilon(v)} & c_{vw}^2\sigma^2_{\epsilon(v)} + \sigma^2_{\epsilon(w|v)} \end{pmatrix}. \tag{52}$$

Note that here we are using the fact that the covariance matrix of the innovation term in Eq. (34), $(\epsilon_t(v), \epsilon_t(w|v))^\dagger$, is known to be diagonal, since all instantaneous correlations are captured by the coupling term $c_{vw}\, x_t^{(v)}$. The log-likelihood of model (34) is given by

$$\mathcal{L}(v, w|v) = -\frac{1}{2} \sum_{t=p+1}^{N_t} \Big(\log\big|\mathsf{S}_{\epsilon(v,w|v)}\big| + (\epsilon_t(v), \epsilon_t(w|v))\, \mathsf{S}^{-1}_{\epsilon(v,w|v)} \\ \times\; (\epsilon_t(v), \epsilon_t(w|v))^\dagger + 2\log(2\pi) \Big), \tag{53}$$

where $|.|$ denotes matrix determinant. After inserting Eq. (52) and some further transformations this expression becomes

$$\mathcal{L}(v, w|v) = -\frac{1}{2}(N_t - p)\big(\log\sigma^2_{\epsilon(v)} + \log\sigma^2_{\epsilon(w|v)}\big) - (N_t - p)\,(1 + \log(2\pi)) \\ + (N_t - p)\frac{2c_{vw}\,\sigma^2_{\epsilon(v),\epsilon(w|v)} - c_{vw}^2\,\sigma^2_{\epsilon(v)}}{2\sigma^2_{\epsilon(w|v)}}, \tag{54}$$

where we have defined $\sigma^2_{\epsilon(v),\epsilon(w|v)} = \mathcal{E}\,(\epsilon_t(v)\epsilon_t(w|v))$. Since for the first p data points no predictions can be performed, $(N_t - p)$ arises in this expression instead of N_t. Typically p will be small, therefore the missing likelihood contribution of the first p points can be neglected.

Note that the log-likelihood of the uncoupled model Eq. (33) is given by

$$\mathcal{L}(v, w) = -\frac{1}{2}(N_t - p)\big(\log \sigma^2_{\epsilon(v)} + \log \sigma^2_{\epsilon(w)}\big) - (N_t - p)\,(1 + \log(2\pi))\,. \quad (55)$$

From Eqs. (9) and (14)–(16) it follows that the mutual information of the time series at grid points v and w is given by (identifying $x_t^{(v)}$ with y_t and $x_t^{(w)}$ with x_t)

$$I\big(x^{(v)}, x^{(w)}\big) = \mathcal{L}(v, w|v) - \mathcal{L}(v, w). \quad (56)$$

Thereby we obtain the estimator for mutual information

$$I\left(x^{(v)}, x^{(w)}\right) = (N_t - p)\frac{2c_{vw}\,\sigma^2_{\epsilon(v),\epsilon(w|v)} - c^2_{vw}\,\sigma^2_{\epsilon(v)}}{2\sigma^2_{\epsilon(w|v)}}, \quad (57)$$

as claimed in Eq. (35).

Although the majority of values provided by this estimator are positive, it has to be mentioned that frequently also negative estimates occur, although mutual information cannot be negative. From the viewpoint of Likelihood Ratio Testing, this behaviour corresponds to the case of the more general model performing worse than the special model. While closer examination may be required in order to fully understand this surprising behaviour, we interprete this effect as a mainly numerical problem resulting from replacing the correct symmetric bivariate model, Eq. (23), by the approximative asymmetric model, Eq. (34). For the numerical results presented in this paper we have taken the following approach: Since for each pair of grid points (v, w) two asymmetrical models can be formulated, only the model corresponding to the higher likelihood is accepted; and in the case that both models yield negative estimates of mutual information, a value of zero is assumed.

Finally we briefly consider once again $\mathsf{S}_{\epsilon(v,w|v)}$, as given by Eq. (52), and compare with $\mathsf{S}_{\epsilon(v,w|v,w)}$, as given by Eq. (19). Each of these two expressions contains three independent parameters, such that they can be interpreted as providing two different parametrisations for the covariance matrix of bivariate time series; but it is obvious that Eq. (19) represents the general case, while Eq. (52) represents an approximation. By comparing both expressions it can be seen that the coupling parameter c_{vw} corresponds roughly (but not precisely) to the coefficient of linear correlation $r\,(\epsilon(v), \epsilon(w))$. Further study may be required for understanding the differences and similarities between these two parametrisations in full detail.

ACKNOWLEDGMENTS

A. Galka gratefully acknowledges support by the Deutsche Forschungsgemeinschaft (DFG) through project GA 673/1-1 and by the Japanese Society for the Promotion of Science (JSPS) through fellowship ID No. P 03059. T. Ozaki gratefully acknowledges support by (JSPS) through grants KIBAN 13654075 and KIBAN 15500193. The comments of the anonymous referees have contributed to improving this paper considerably.

REFERENCES

1. P. E. Cladis and P. Palffy-Muhoray (eds.), *Spatio-Temporal Patterns in Nonequilibrium Complex Systems*. SFI Studies in the Sciences of Complexity (Addison Wesley Longman, Reading, MA, 1995).
2. L. B. Almeida, F. Lopes da Silva and J. C. Principe (eds.), *Spatiotemporal Models in Biological and Artificial Systems*, volume 37 of *Frontiers in Artificial Intelligence and Applications* (IOS Press, Amsterdam, 1997).
3. D. Walgraef, *Spatio-temporal Pattern Formation* (Springer, Berlin, Heidelberg, New York, 1997).
4. E. Gehrig and O. Hess, *Spatio-Temporal Dynamics and Quantum Fluctuations in Semiconductor Lasers*. Springer Tracts in Modern Physics, Vol. 189 (Springer, Berlin, Heidelberg, New York, 2003).
5. G. Christakos, *Modern Spatiotemporal Geostatistics*. Studies in Mathematical Geology, Vol. 6 (Oxford University Press, Oxford, 2000).
6. J. Bascompte and R. V. Solé (eds.), *Modelling Spatiotemporal Dynamics in Ecology* (Springer, Berlin, Heidelberg, New York, 1998).
7. E. Yago, C. Escera, K. Alho, M. H. Giard and J. M. Serra-Grabulosa, Spatiotemporal dynamics of the auditory novelty-P3 event-related brain potential. *Cogn. Brain Res.* **16**:383–390, 2003.
8. C. Lukas, J. Falck, J. Bartkova, J. Bartek and J. Lukas, Distinct spatiotemporal dynamics of mammalian checkpoint regulators induced by DNA damage. *Nature Cell Biol.* **5**:255–260, 2003.
9. M. Mori, M. Kaino, S. Kanemoto, M. Enomoto, S. Ebata and S. Tsunoyama, Development of advanced core noise monitoring system for BWRS. *Prog. Nucl. Energy* **43**:201–207, 2003.
10. W. Li, Mutual information functions versus correlation functions. *J. Stat. Phys.* **60**:823–837, 1990.
11. P. Hall and S. C. Morton, On the estimation of entropy. *Ann. Inst. Statist. Math.* **45**:69–88, 1993.
12. D. R. Brillinger, Second-order moments and mutual information in the analysis of time series. In Y. P. Chaubey (ed.), *Recent Advances in Statistical Methods*, pp. 64–76 (Imperial College Press, London, 2002).
13. D. R. Brillinger, Some data analyses using mutual information. *Brazilian J. Prob. Statist.* **18**:163–183, 2004.
14. K. Kaneko, Spatiotemporal chaos in one- and two-dimensional coupled map lattices. *Physica D* **37**:60–82, 1989.
15. C. E. Shannon, A mathematical theory of communication. *Bell Syst. Tech. J.* **27**:379–423, 623–656, 1948.
16. L. Boltzmann, Über die Beziehung zwischen dem zweiten Hauptsatze der mechanischen Wärmetheorie und der Wahrscheinlichkeitsrechnung respektive den Sätzen über das Wärmegleichgewicht. *Wiener Berichte* **76**:373–435, 1877.
17. H. Akaike, On the likelihood of a time series model. *The Statistician* **27**:217–235, 1978.
18. T. M. Cover and J. A. Thomas, *Elements of Information Theory* (Wiley, New York, 1991).
19. Y. Moon, B. Rajagopalan and U. Lall, Estimation of mutual information using kernel density estimators. *Phys. Rev. E* **52**:2318–2321, 1995.
20. D. Prichard and J. Theiler, Generalized redundancies for time series analysis. *Physica D* **84**:476–493, 1995.

21. A. Kraskov, Harald Stögbauer and P. Grassberger, Estimating mutual information. *Phys. Rev. E* **69**:066138, 2004.
22. R. Moddemeijer, A statistic to estimate the variance of the histogram based mutual information estimator based on dependent pairs of observations. *Signal Proc.* **75**:51–63, 1999.
23. H. Stögbauer, A. Kraskov, S. A. Astakhov and P. Grassberger, Least-dependent-component analysis based on mutual information. *Phys. Rev. E* **70**:066123, 2004.
24. K. Ito, On stochastic differential equations. *Mem. Am. Math. Soc.* **4**:1–51, 1951.
25. J. L. Doob, *Stochastic Processes* (Wiley, New York, 1953).
26. A. H. Jazwinski, *Stochastic Processes and Filtering Theory* (Academic Press, San Diego, 1970).
27. T. Kailath, A view of three decades of linear filtering theory. *IEEE Trans. Inf. Theory* **20**:146–181, 1974.
28. P. Lévy, Sur une classe de courbes de l'espace de Hilbert et sur une équation intégrale non linéaire. *Ann. Sci. École Norm. Sup.* **73**:121–156, 1956.
29. P. Protter, *Stochastic Integration and Differential Equations* (Springer-Verlag, Berlin, 1990).
30. P. A. Frost and T. Kailath, An innovation approach to least squares estimation—part III: Nonlinear estimation in white gaussian noise. *IEEE Trans. Autom. Contr.* **16**:217–226, 1971.
31. B. V. Gnedenko, *The Theory of Probability* (Mir Publishers, Moscow, 1969).
32. G. E. P. Box and G. M. Jenkins, *Time Series Analysis, Forecasting and Control* 2nd edition (Holden-Day, San Francisco, 1976).
33. R. Moddemeijer, On estimation of entropy and mutual information of continuous distributions. *Signal Proc.* **16**:233–248, 1989.
34. A. Galka and T. Ozaki, Testing for nonlinearity in high-dimensional time series from continuous dynamics. *Physica D* **158**:32–44, 2001.
35. H. Akaike, Prediction and entropy. In A. C. Atkinson and S. E. Fienberg (eds.), *A Celebration of Statistics*, pp. 1–24 (Springer, Berlin, Heidelberg, New York, 1985).
36. J. Kuha, AIC and BIC: Comparisons of assumptions and performance. *Sociol. Methods Res.* **33**:188–229, 2004.
37. M. Stone, An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *J. Roy. Statist. Soc. B* **39**:44–47, 1977.
38. N. Levinson, The Wiener rms error criterion in filter design and prediction. *J. Math. Phys.* **25**:261–278, 1947.
39. P. Whittle, On the fitting of multivariate autoregressions, and the approximate canonical factorization of a spectral density matrix. *Biometrika* **50**:129–134, 1963.
40. A. Neumaier and T. Schneider, Estimation of parameters and eigenmodes of multivariate autoregressive models. *ACM Trans. Math. Softw.* **27**:27–57, 2001.
41. C. Penland, Random forcing and forecasting using principal oscillation pattern analysis. *Monthly Weather Rev.* **117**:2165–2185, 1989.
42. A. Galka, O. Yamashita, T. Ozaki, R. Biscay and P. A. Valdés-Sosa, A solution to the dynamical inverse problem of EEG generation using spatiotemporal Kalman filtering. *NeuroImage* **23**:435–453, 2004.
43. A. Galka, O. Yamashita and T. Ozaki, GARCH modelling of covariance in dynamical estimation of inverse solutions. *Phys. Lett. A* **333**:261–268, 2004.
44. J. Geweke, Inference and causality in economic time series models. In Z. Grilliches and M. Intriligator (eds.), *Handbook of Econometrics*, pp. 1101–1144 (North-Holland, Amsterdam, 1984).
45. A. Buse, The likelihood ratio, Wald, and Lagrange multiplier test: An expository note. *Am. Stat.* **36**:153–157, 1982.
46. A. Hyvärinen, J. Karhunen and E. Oja, *Independent Component Analysis* (Wiley, New York, 2001).
47. L. Molgedey and H. G. Schuster, Separation of a mixture of independent signals using time delayed correlations. *Phys. Rev. Lett.* **72**:3634–3637, 1994.

48. T. Schreiber, Spatio-temporal structure in coupled map lattices: two-point correlations versus mutual information. *J. Phys. A: Math. Gen.* **23**:L393–L398, 1990.
49. H. Wu and J. Principe, Generalized anti-Hebbian learning for source separation. In *Proceedings of ICASSP'99*, pp. 1073–1076 (IEEE Press, Piscataway, NJ, 1999).
50. S. Choi, A. Cichocki and S. Amari, Flexible independent component analysis. *J. VLSI Signal Process* **26**:25–38, 2000.
51. K. G. Jöreskog, Some contributions to maximum likelihood Factor Analysis. *Psychometrika* **32**:443–482, 1967.
52. B. Flury, *A First Course in Multivariate Statistics* (Springer, New York, Berlin, Heidelberg, 1997).