*AppliedMath*

MDPI

*Article*

# Measuring Dependencies between Variables of a Dynamical System Using Fuzzy Affiliations

**Niklas Wulkow** [1,2]

[1] Department of Mathematics and Computer Science, Freie Universität Berlin, 14195 Berlin, Germany; niklas.wulkow@zib.de

[2] Zuse Institute Berlin, 14195 Berlin, Germany

**Abstract:** A statistical, data-driven method is presented that quantifies influences between variables of a dynamical system. The method is based on finding a suitable representation of points by fuzzy affiliations with respect to landmark points using the Scalable Probabilistic Approximation algorithm. This is followed by the construction of a linear mapping between these affiliations for different variables and forward in time. This linear mapping, or matrix, can be directly interpreted in light of unidirectional dependencies, and relevant properties of it are quantified. These quantifications, given by the sum of singular values and the average row variance of the matrix, then serve as measures for the influences between variables of the dynamics. The validity of the method is demonstrated with theoretical results and on several numerical examples, covering deterministic, stochastic, and delayed types of dynamics. Moreover, the method is applied to a non-classical example given by real-world basketball player movement, which exhibits highly random movement and comes without a physical intuition, contrary to many examples from, e.g., life sciences.

**Keywords:** dependency measures; influence detection; data-driven modelling; Scalable Probabilistic Approximation; barycentric coordinates; basketball analytics

**MSC:** 37M10; 62B10; 65C20

check for updates

**Citation:** Wulkow, N. Measuring Dependencies between Variables of a Dynamical System Using Fuzzy Affiliations. *AppliedMath* **2022**, *2*, 284–311. https://doi.org/10.3390/appliedmath2020018

Received: 12 April 2022
Accepted: 19 May 2022
Published: 16 June 2022

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

## 1. Introduction

Over the last few decades, detecting influences between variables of a dynamical system from time series data has shed light on the interplay between quantities in various fields, such as between genes [1,2], between wealth and transportation [3], between marine populations [4], or within the Earth's climate [5,6]. Such analyses can reveal the driving forces behind complex phenomena and enable the understanding of connections that otherwise would have been left uncovered. For example, recently, novel candidates for cancer-causing genes have been detected by systematic dependency analyses using patient data [7–9].

There exist various conceptually different numerical methods that aim to detect relations between variables from data. For example, a prominent method is Convergent Cross-Mapping (CCM) [10], introduced in 2012, which uses the delay-embedding theorem of Takens [11]. Another method is Granger causality [12,13], which was first used in economics in the 1960s and functions based on the intuition that if a variable $X$ forces $Y$, then values of $X$ should help to predict $Y$. Much simpler than these methods is the well-known Pearson correlation coefficient [14], introduced already in the 19th Century and one of many methods that detect linear dependencies [15]. There exist many more, such as the Mutual Information Criterion [16], the Heller–Gorfine test [17], Kendall's $\tau$ [18], the transfer entropy method [19], the Hilbert–Schmidt independence criterion [20], or the Kernel Granger test [21].

Recently, in [22], a summarising library comprising a broad range of statistical methods was presented together with many examples. This illustrates the large number of already existing statistical techniques to detect influences between variables of a dynamical system. However, as also pointed out in [22], such methods have different strengths and shortcomings, making them well-suited for different scenarios while ill-suited for others. For example, CCM requires data coming from dynamics on an attracting manifold and suffers when the data are subject to stochastic effects. Granger causality needs an accurate model formulation for the dynamics, which can be difficult to find. Pearson's correlation coefficient requires statistical assumptions on the data, which are often not met. The authors of [22] therefore suggest to utilise many different methods on the same problem instead of only a single one to optimally find and interpret relations between variables.

In this article, a practical method is presented that aims at complementing the weaknesses of related methods and enriching the range of existing techniques for the detection of influences. It represents the data using probabilistic, or *fuzzy* , affiliations with respect to landmark points and constructs a linear probabilistic model between the affiliations of different variables and forward in time. This linear mapping then admits a direct and intuitive interpretation in regard to the relationship between variables. The landmark points are selected by a data-driven algorithm, and the model formulation is quite general, so that only very little intuition of the dynamics is required. For the fuzzy affiliations and the construction of the linear mapping, the recently introduced method Scalable Probabilistic Approximation (SPA) [23,24] is used, whose capacity to locally approximate highly nonlinear functions accurately was demonstrated in [23]. Then, the properties of this mapping are computed, which serve as the dependency measures. The intuition, which is further elaborated on in the article, now is as follows: if one variable has little influence on another, the columns of the linear mapping should be similar to each other, while if a variable has a strong influence, the columns should be very different. This is quantified using two measures, which extract quantities of the matrix, one inspired by linear algebra, the other a more statistical one. The former computes the sum of singular values of the matrix with the intuition that a matrix consisting of similar columns is close to a low-rank matrix for which many singular values are zero [25]. The latter uses the statistical variance to quantify the difference between the columns. It is shown that they are in line with the above interpretation of the column-stochastic matrix, and the method is applied to several examples to demonstrate its efficacy. Three examples are of a theoretical nature, where built-in influences are reconstructed. One real-world example describes the detection of influences between basketball players during a game.

This article is structured as follows: In Section 2, the SPA method is introduced, including a mathematical formalisation of influences between variables in the context of the SPA-induced model. In Section 3, the two dependency measures that quantify the properties of the SPA model are introduced and connected with an intuitive perception of the SPA model. In Section 4, the dependency measures are applied to examples.

To outline the structure of the method in advance, the three main steps that are proposed to quantify influences between variables of a dynamical system are (also see Figure 1):

1. Representation of points by fuzzy affiliations with respect to landmark points.
2. Estimation of a particular linear mapping between fuzzy affiliations of two variables and forward in time.
3. Computation of the properties of this matrix.

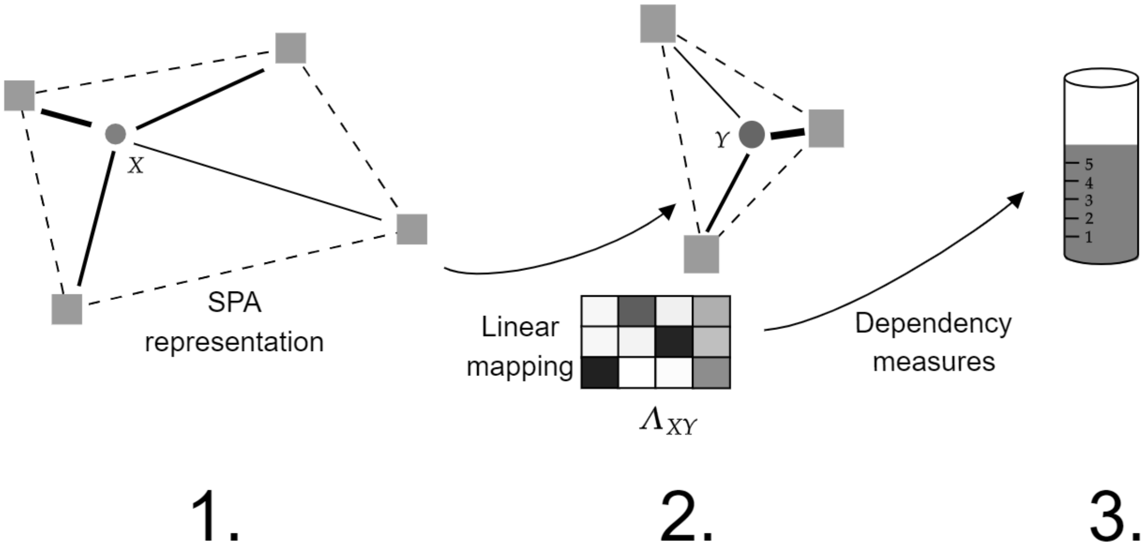

**Figure 1.** Illustration of the three steps of the method presented in this article.

## 2. Scalable Probabilistic Approximation

Scalable Probabilistic Approximation (SPA) [23] is a versatile and low-cost method that transforms points from a $D$-dimensional state space to $K$-dimensional probabilistic, fuzzy, affiliations. If $K$ is less than $D$, SPA serves as a dimension reduction method by representing points as closely as possible. If $K > D$, SPA can be seen as a probabilistic, or fuzzy, clustering method, which assigns data points to *landmark points* in $D$-dimensional space depending on their closeness to them. For two different variables $X$ and $Y$, SPA can furthermore find an optimal linear mapping between the probabilistic representations.

The first step, the transformation to fuzzy affiliations, will be called SPA I, while the construction of the mapping in these coordinates will be called SPA II.

### *2.1. Representation of the Data in Barycentric Coordinates*

The mathematical formulation of SPA I is as follows: Let $\mathbf{X} = [X_1, \dots, X_T] \in \mathbb{R}^{D\times T}$. Then, solve

$$\begin{gathered}[\Sigma,\Gamma] = \arg\min \|\mathbf{X} - \Sigma\Gamma\|_F \\ \text{subject to } \Sigma = [\sigma_1,\dots,\sigma_K] \in \mathbb{R}^{D\times K}, \quad \Gamma = [\gamma_1,\dots,\gamma_T] \in \mathbb{R}^{K\times T}, \\ (\gamma_t)_i \geq 0, \quad \sum_{i=1}^{K} (\gamma_t)_i = 1.\end{gathered} \tag{SPA 1}$$

It was discussed in [26] that for $K \leq D$, the representation of points in this way is the orthogonal projection onto a convex polytope with vertices given by the columns of $\Sigma$. The coordinates $\gamma$ then specify the position of this projection with respect to the vertices of the polytope and are called *Barycentric Coordinates* (BCs). A high entry in such a coordinate then signals the closeness of the projected point to the vertex.

**Remark 1.** *A similar representation of points has already been introduced in PCCA+ [27].*

For $K > D$, however, in [23], the interpretation of a probabilistic clustering was introduced. According to the authors, the entries of a $K$-dimensional coordinate of a point denote the probabilities to be inside a certain box around a landmark point. One can generalise this interpretation to *fuzzy affiliations* to these landmark points, again in the sense of closeness. A BC $\gamma$ then denotes the distribution of affiliations to each landmark point, and (SPA 1) can be solved without loss, i.e., so that

$$X_t = \Sigma \gamma_t \tag{1}$$

holds for all data points. Figure 2 shows the representation of a point in $\mathbb{R}^2$ with respect to four landmark points.

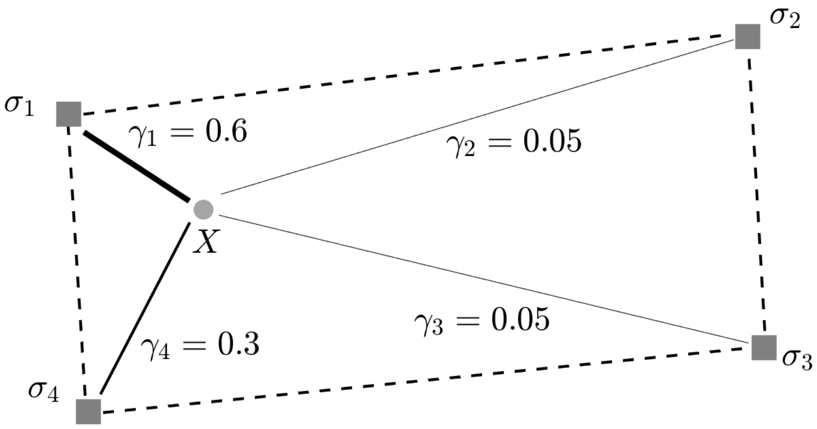

**Figure 2.** Representation of a point $X$ by barycentric coordinates $\gamma_1, \ldots, \gamma_4$ with respect to the vertices of a polytope $\sigma_1, \ldots, \sigma_4$.

**Remark 2.** *From now on, regardless of the relation between K and D, mostly the term BC will be used instead of "fuzzy affiliations" for the $\gamma_t$ to emphasise that they are new coordinates of the data, which can always be mapped back to the original data as long as Equation (1) is fulfilled. Note that while commonly in the literature, the term "barycentric coordinate" refers to coordinates with respect to the vertices of a simplex (i.e., is $(K-1)$-dimensional with K vertices, contrary to the assumption $K > D + 1$), in various publications, e.g., [28], the term generalised barycentric coordinates is used if the polytope is not a simplex. In any case, the term "generalised" will be omitted and only "BC" will be written.*

For $K > D + 1$ and given landmark points, the representation with barycentric coordinates is generally not unique (while the set of points that can be represented by $\Sigma\gamma$ is a convex polytope, some landmark points can even lie inside this polytope if $K > D + 1$). Therefore, let us define the representation of a point $X$ analogously to [26] in the following way:

$$\begin{aligned} \rho_\Sigma(X, \gamma) &:= \arg\min_{\gamma^*} \|\gamma - \gamma^*\|_2 \\ \text{s.t. } \gamma^* &= \arg\min_{\gamma'} \|X - \Sigma\gamma'\|_2 \text{ with } \gamma'_\bullet, \gamma^*_\bullet \geq 0 \text{ and } \|\gamma'\|_1, \|\gamma^*\|_1 = 1. \end{aligned} \tag{2}$$

$\gamma^*$ should be selected among all barycentric coordinates that represent $X$ without loss so that it is closest to the *reference coordinate* $\gamma$.

**Remark 3.** *Note that the solution for the landmark points determined by SPA I is never unique if $K > 1$ [23]. In order to solve (SPA 1), its objective function is iteratively minimised by separately solving for $\Sigma$ and $\Gamma$. For this, initial values are randomly drawn. Therefore, for $K > D + 1$, an exact solution can be achieved by placing $D + 1$ points so that all data points lie inside their convex hull (all convex combinations of them) while the remaining $K - (D + 1)$ landmark points can be chosen arbitrarily. As a consequence, the placement of landmark points depends strongly on the randomly chosen initial values of the optimisation process. In simple cases, e.g., when the state space of the data is one-dimensional, one could even manually place landmark points across an interval containing the data and solve only for $\Gamma$.*

**Remark 4.** *Note that with a fixed $\Sigma$, the computations of all columns of $\Gamma$ are independent of each other so that this step can easily be parallelised, making SPA a very efficient method for long time series and justifying the term "Scalable". Moreover, the time complexity of the full (SPA 1) problem grows linearly with the number of data points [23], which is comparable to the K-means method [29], while still giving a much more precise, for $K \geq D+1$ even lossless, representation of points.*

Let now a dynamical system be given by

$$X_t = F(X_{t-1}). \tag{3}$$

In this case, let us select $\gamma_{t-1}$ as the reference coordinate for the point $X_t$. Using that Equation (1) holds for $K > D$, this gives

$$\gamma_t = \rho_\Sigma(X_t, \gamma_{t-1}) = \rho_\Sigma(F(X_{t-1}), \gamma_{t-1}) = \rho_\Sigma(F(\Sigma\gamma_{t-1}), \gamma_{t-1}) =: \mathbf{v}(\gamma_{t-1}). \tag{4}$$

With this, $\gamma_t$ solely depends on $\gamma_{t-1}$. This establishes a time-discrete dynamical system in the barycentric coordinates. By choosing the reference coordinate as $\gamma_{t-1}$, it is asserted that the steps taken are as short as possible.

### *2.2. Model Estimation between Two Variables*

Assuming that data from two different variables $X \in \mathbb{R}^D$ and $Y \in \mathbb{R}^E$ is considered and one strives to investigate the relationship between them, one can solve (SPA 1) for both of them, finding landmark points denoted by the columns of $\Sigma^X \in \mathbb{R}^{D\times K_X}, \Sigma^Y \in \mathbb{R}^{E\times K_Y}$ and BCs $\gamma_t^X$ and $\gamma_t^Y$ for $t = 1, \ldots, T$ and solve Equation (SPA 2), given by

$$\begin{aligned}\Lambda &= \underset{\Lambda^* \in \mathbb{R}^{K\times K}}{\arg\min} \|[\gamma_1^Y|\cdots|\gamma_T^Y] - \Lambda^*[\gamma_1^X|\cdots|\gamma_T^X]\|_F, \\ &\text{subject to } \Lambda \geq 0 \text{ and } \sum_{k=1}^{K} \Lambda_{k,\bullet} = 1.\end{aligned} \tag{SPA 2}$$

$\Lambda$ is a column-stochastic matrix and, therefore, guarantees that BCs are mapped to BCs again. It is an optimal linear mapping with this property that it connects $X$ to $Y$ on the level of the BCs. One can easily transform back to the original state space by

$$Y_t = \Sigma\gamma_t^Y \approx \Sigma\Lambda\gamma_t^X. \tag{5}$$

Since, by construction, the prediction must lie inside the polytope again, it is a weighted linear interpolation with respect to the landmark points, similar as, e.g., in Kriging interpolation [30]. Note that is was demonstrated in [23] that this linear mapping applied to the BCs can accurately approximate functions of various complexity and is not restricted to linear relationships.

**Remark 5.** *In [23], a way to combine SPA I and SPA II into a single SPA I problem was shown. The landmark points are then selected so that the training error of the SPA II problem can be set to 0. In other words, optimal discretisations of the state spaces are determined where optimal means with respect to the exactness of the ensuing (SPA 2) model.*

#### Estimation of Dynamics

A special case of (SPA 2) is the estimation of dynamics, i.e., solving

$$\begin{aligned}\Lambda &= \underset{\Lambda^* \in \mathbb{R}^{K\times K}}{\arg\min} \|[\gamma_2^X|\cdots|\gamma_T^X] - \Lambda^*[\gamma_1^X|\cdots|\gamma_{T-1}^X]\|_F, \\ &\text{subject to } \Lambda \geq 0 \text{ and } \sum_{k=1}^{K} \Lambda_{k,\bullet} = 1.\end{aligned} \tag{6}$$

$\Lambda$ is therefore a linear, column-stochastic approximation of the function $\mathbf{v}$ from Equation (4). Such a matrix is typically used in so-called Markov State Models [31,32]. With $\Lambda$, one can construct dynamics in the BCs with the possibility to transform back to the original state space by multiplication with $\Sigma$, since

$$X_t = \Sigma\gamma_t = \Sigma\mathbf{v}(\gamma_{t-1}) \approx \Sigma\Lambda\gamma_{t-1}. \tag{7}$$

### *2.3. Forward Model Estimation between Two Processes*

Given two dynamical systems:

$$X_t = F(X_{t-1}) \in \mathbb{R}^D \text{ and } Y_t = G(Y_{t-1}) \in \mathbb{R}^E, \tag{8}$$

let us now determine a column-stochastic matrix that propagates barycentric coordinates from one variable to the other and forward in time. This mapping will be used to quantify the effect of one variable on future states of the other.

With landmark points in $\Sigma^X \in \mathbb{R}^{D\times K_X}, \Sigma^Y \in \mathbb{R}^{E\times K_Y}$ and BCs $\gamma_t^X$ and $\gamma_t^Y$ for $t = 1, \dots, T$, let us find a column-stochastic matrix $\Lambda_{XY}$ that fulfils

$$\Lambda_{XY} = \underset{\Lambda^*\in\mathbb{R}^{K_Y\times K_X}}{\arg\min} \|[\gamma_2^Y|\cdots|\gamma_T^Y] - \Lambda^*[\gamma_1^X|\cdots|\gamma_{T-1}^X]\|_F. \tag{9}$$

$\Lambda_{XY}$ represents a model from $X_{t-1}$ to $Y_t$ on the level of the BCs and tries to predict subsequent values of $Y$ using only $X$.

Now, let us assume that $X$ in fact has a direct influence on $Y$, meaning that there exists a function:

$$H(Y_{t-1}, X_{t-1}) = Y_t. \tag{10}$$

Then, similar as when constructing the dynamical system in the BCs previously in Equation (4), we can observe, using Equations (1) and (10),

$$\begin{aligned} \gamma_t^Y &= \rho_{\Sigma^Y}(Y_t, \gamma_{t-1}^Y) \\ &= \rho_{\Sigma^Y}(H(Y_{t-1}, X_{t-1}), \gamma_{t-1}^Y) \\ &= \rho_{\Sigma^Y}(H(\Sigma^X\gamma_{t-1}^X, \Sigma^Y\gamma_{t-1}^Y), \gamma_{t-1}^Y) \\ \text{defined as } &=: \mathbf{w}(\gamma_{t-1}^X, \gamma_{t-1}^Y). \end{aligned} \tag{11}$$

$\gamma_t^Y$ therefore directly depends on $\gamma_{t-1}^X$ and $\gamma_{t-1}^Y$, while $\Lambda_{XY}$ attempts to predict $\gamma_t^Y$ using only $\gamma_{t-1}^X$. Assuming that the approximation:

$$\Lambda_{XY}\gamma_{t-1}^X \approx \gamma_t^Y \tag{12}$$

is close for each pair of $\gamma_{t-1}^X, \gamma_t^Y$, one can assert

$$\gamma_t^Y \approx \Lambda_{XY}\gamma_{t-1}^X = \sum_{j=1}^{K_X} (\Lambda_{XY})_{|j}(\gamma_{t-1}^X)_j, \tag{13}$$

where $(\Lambda_{XY})_{|j}$ is the $j$th column of $\Lambda_{XY}$. A prediction for $\gamma_t^Y$ is therefore constructed using a weighted average of the columns of $\Lambda_{XY}$. The weights are the entries of $\gamma_{t-1}^X$.

**Remark 6.** *Note that the same argumentation starting in Equation* (9) *holds if one chooses a time shift of length $\tau > 0$ and considers information of $X_{t-\tau}$ about $Y_t$. If $\tau = 1$, it will simply be written $\Lambda_{XY}$, but otherwise, $\Lambda_{XY}^{(\tau)}$.*

**Remark 7.** *Since $\Lambda_{XY}\gamma^X_{t-1}$ estimates $\gamma^Y_t$ using only $\gamma^X_{t-1}$, although it additionally depends on $\gamma^Y_{t-1}$, one can interpret $\Lambda_{XY}$ as an approximation to the conditional expectation of $\gamma^Y_t$ given $\gamma^X_{t-1}$, i.e., assuming that $\gamma^Y_{t-1}$ is distributed by a function $\mu_Y$,*

$$\Lambda_{XY}\gamma \approx \mathbb{E}_{\mu_Y}[\gamma^Y_t | \gamma^X_{t-1} = \gamma]. \tag{14}$$

*In Appendix A.1, this intuition is formalised further.*

*In the original SPA paper [23], the $\gamma_t$ were interpreted as probabilities to "belong" to states $\sigma$ that describe a discretisation of the state space. $\Lambda$ then uses the law of total probability and the ij-entry denotes $(\Lambda_{XY})_{ij} = \mathbb{P}[Y_t = \sigma^Y_i | X_{t-1} = \sigma^X_j]$. Together with the combined solution of the (SPA 1) and (SPA 2) problems outlined in Remark 5, the authors of [23] described their method as finding an optimal discretisation of the state space to generate an exact model on the probabilities, thereby satisfying desired physical conservation laws. In this light, SPA then directly competes with the before-mentioned Markov State Models (MSMs) with the significant advantage that for MSMs with equidistant discretisations, i.e., grids, the number of "grid boxes" increases exponentially with the dimension and box size. This is further elaborated in Appendix A.3. While the probabilistic view is a sensible interpretation of the objects involved in SPA, one is not restricted to it, and the view taken in this article focuses on the exact point representations, which the barycentric coordinates yield. However, also acknowledging the probabilistic view point, it should be possible to make the arguments following in the next sections in a satisfying and intuitive way.*

## 3. Quantification of Dependencies between Variables

In the following, two methods that quantify the strength of dependence of $Y_t$ on $X_{t-1}$ by directly computing the properties of $\Lambda_{XY}$ will be defined. The intuition can be illustrated as follows: If a variable $X$ is important for the future state of another variable, $Y$, then the multiplication of $\Lambda_{XY}$ with $\gamma^X_{t-1}$ should give very different results depending on which landmark point $X_{t-1}$ is close to, i.e., which weight in $\gamma^X_{t-1}$ is high. Since $\gamma^Y_t$ is composed by a weighted average of the columns of $\Lambda_{XY}$ by Equation (13), this means that the columns of $\Lambda_{XY}$ should be very different from each other. In turn, if $X$ has no influence and carries no information for future states of $Y$, the columns of $\Lambda_{XY}$ should be very similar to each other. In Appendix A.1, it is shown that given the independence of $\gamma^Y_t$ from $\gamma^X_{t-1}$, all columns of $\Lambda_{XY}$ should be given by the mean of the $\gamma^Y$ in the data. There, this is also connected to conditional expectations and the intuition given in Equation (14).

In the extreme case that the columns are actually equal to each other, $\Lambda_{XY}$ would be a rank-1 matrix. If the columns are not equal, but similar, $\Lambda_{XY}$ is at least *close to* a rank-1 matrix. One should therefore be able to deduce from the similarity of the columns of $\Lambda_{XY}$ if $X$ could have an influence on $Y$. This is the main idea behind the dependency measures proposed in this section. The intuition is similar to the notion of the predictability of a stochastic model introduced in [33].

**Remark 8.** *Note that if there is an intermediate variable $Z$ that is forced by $X$ and forces $Y$ while $X$ does not directly force $Y$, then it is generally difficult to distinguish between the direct and indirect influences. In Section 4.2, an example for such a case is investigated.*

### *3.1. The Dependency Measures*

Now, the two measures are introduced that will be used for the quantification of dependencies between variables. Note that these are not "measures" in the true mathematical sense, but the term is rather used as synonymous with "quantifications".

#### 3.1.1. Schatten-1 Norm

For the first measure, let us consider the *Singular Value Decomposition* (SVD) [34] of a matrix $\Lambda$, given by $\Lambda = USV^T \in \mathbb{R}^{K_Y \times K_X}$. $S \in \mathbb{R}^{K_Y \times K_X}$ is a matrix that is only nonzero in the entries $(i,i)$ for $i = 1, \ldots, min(K_X, K_Y)$, which are denoted by $s_1, \ldots, s_{min(K_X,K_Y)} \geq 0$.

$U \in \mathbb{R}^{K_Y \times K_Y}$ and $V \in \mathbb{R}^{K_X \times K_X}$ fulfil certain orthogonality properties and consist of columns $u_i, v_i$. One can thus write $\Lambda$ as

$$\Lambda = \sum_{i=1}^{r} u_i v_i^T s_i \tag{15}$$

A classic linear algebra result asserts that $rank(\Lambda) = \#\{s_i > 0\}$. As a consequence, if some of the $s_i$ are close to 0, then Equation (24) means that only a small perturbation is sufficient to make $\Lambda$ a matrix of lower rank. Therefore, the sum of singular values, the so-called *Schatten-1 norm* [35], will be used as a continuous measure of the rank and, thus, of the difference in the rows of $\Lambda$.

**Definition 1** (Schatten-1 norm)**.** *Let the SVD of a matrix $\Lambda \in \mathbb{R}^{K_Y \times K_X}$ be given by $\Lambda = USV^T$ with singular values $s_1, \ldots, s_{min(K_X,K_Y)}$. Then, the Schatten-1 norm of $\Lambda$ is defined as*

$$\|\Lambda\|_1 := \sum_{i=1}^{min(K_X,K_Y)} s_i. \tag{16}$$

### 3.1.2. Average Row Variance

As the second dependency measure, the difference of the columns of a matrix $\Lambda$ is directly quantified using the mean statistical variance per row. Therefore, every row is considered, and the variance between its entries is computed, thereby comparing the columns with respect to this particular row. Then, the mean of these variances is taken across all rows.

**Definition 2** (Average row variance)**.** *For a matrix $\Lambda \in \mathbb{R}^{K_Y \times K_X}$, let $\bar{\Lambda}_{i-}$ denote the mean of the ith row of $\Lambda$. Let*

$$\nu_i := \frac{1}{K_X - 1} \sum_{j=1}^{K_X} (\Lambda_{ij} - \bar{\Lambda}_{i-})^2$$

*be the statistical variance of the entries of the ith row. Then, the average row variance is defined as*

$$\nu(\Lambda) := \frac{1}{K_Y} \sum_{i=1}^{K_Y} \nu_i. \tag{17}$$

The calculated values for $\|\cdot\|_1$ and $\nu$ will be stored in tables, respectively, matrices of the form

$$M_{\|\cdot\|_1} = \begin{pmatrix} \|\Lambda_{XX}\|_1 & \|\Lambda_{XY}\|_1 \\ \|\Lambda_{YX}\|_1 & \|\Lambda_{YY}\|_1 \end{pmatrix}, M_v = \begin{pmatrix} \nu(\Lambda_{XX}) & \nu(\Lambda_{XY}) \\ \nu(\Lambda_{YX}) & \nu(\Lambda_{YY}) \end{pmatrix}. \tag{18}$$

Then, for each of these matrices, the property $M - M^T$ should be interesting for us, because they contain the differences between dependency measures, stating how strongly $X$ depends on $Y$ compared to $Y$ depending on $X$. Let us therefore define the following:

**Definition 3** (Relative difference between dependencies)**.** *Let $M$ be one of the matrices from Equation (18). The relative difference between dependencies in both directions is defined as*

$$\delta(M)_{ij} = \frac{M_{ij} - M_{ji}}{max(M_{ij}, M_{ji})}. \tag{19}$$

**Remark 9.** *In Appendix A.3, it is explained how this method differs from linear correlations, Granger causality, and the same approach using boxes instead of fuzzy affiliations.*

When using the dependency measures to analyse which of two variables more strongly depends on the other, it is unclear at this point whether the dimension of the variables and the number of landmark points affects the outcome of the analysis. Hence, for all numerical examples in the next section, pairs of variables that have the same dimension are considered, and the same number of landmark points for them is used, i.e., it holds $K_X = K_Y$ in each example to make the comparison as fair as possible. In this light, the following theoretical results on the Schatten-1 norm and the average row variance are directly helpful.

### *3.2. Maximisers and Minimisers of the Dependency Measures*

About $\|\cdot\|_1$ and $\nu$, one can prove properties that validate why they represent sensible measures for the strengths of dependency between two processes. For this, let us make the following definition.

**Definition 4** (Permutation matrix)**.** *As a permutation matrix, a matrix $A \in \{0,1\}^{n\times n}$ is defined with the property that every row and column contains exactly one* 1*.*

Then, one can prove the following results on the maximisers and minimisers of the Schatten-1 norm and average row variance (half of them only for $K_Y \geq K_X$ or $K_Y = K_X$), whose proofs can be found in Appendix A.2.

**Proposition 1** (Maximal Schatten-1 norm, $K_Y \geq K_X$)**.** *Let $\Lambda \in \mathbb{R}^{K_Y\times K_X}$ with $K_Y \geq K_X$. Then, the Schatten-1 norm of $\Lambda$ obtains the maximal value $K_X$ if deletion of $K_Y - K_X$ rows of $\Lambda$ yields a $K_X \times K_X$ permutation matrix.*

**Proposition 2** (Minimal Schatten-1 norm)**.** *The Schatten-1 norm of a column stochastic $(K_Y \times K_X)$-matrix $A$ is minimal if and only if $A_{ij} \equiv \frac{1}{n}$, and its minimal value is equal to* 1*.*

For the average row variance, the following results can be derived:

**Proposition 3** (Maximal average row variance, $K_Y = K_X$)**.** *The average row variance of a column-stochastic $(K_Y \times K_X)$-matrix $\Lambda$ with $K_Y = K_X$ obtains the maximal value $\frac{1}{K_Y}$ if it is a $K_X \times K_X$ permutation matrix.*

It seems likely that for $K_Y > K_X$, the maximisers of the Schatten-1 norm from Proposition 1 also maximise the average row variance with maximal value $\frac{1}{K_Y}$.

**Proposition 4** (Minimal average row variance)**.** *The average row variance of a column-stochastic $(K_Y \times K_X)$-matrix $\Lambda$ obtains the minimal value* 0 *if and only if all columns are equal to each other.*

In order to analyse the dependencies between two variables for which one uses different numbers of landmarks, i.e., $K_X \neq K_Y$, it would be desirable if similar results as above could be inferred for the case $K_Y < K_X$ so that one could make valid interpretations of both $\Lambda_{XY}$ and $\Lambda_{YX}$. However, it was more difficult to prove them, so that only the following conjectures are made for the case $K_Y < K_X$:

**Conjecture 1** (Maximal Schatten-1 norm, $K_Y < K_X$)**.** *The Schatten-1 norm of a column-stochastic $(K_Y \times K_X)$-matrix $\Lambda$ with $K_Y < K_X$ is maximal if and only if $\Lambda$ contains a $K_Y \times K_Y$ permutation matrix and the matrix of the remaining $K_X - K_Y$ columns can be extended by $K_Y$ columns to a permutation matrix.*

**Conjecture 2** (Maximal average row variance, $K_Y < K_X$)**.** *The average row variance of a column-stochastic $(K_Y \times K_X)$-matrix $\Lambda$ with $K_Y < K_X$ is maximal if and only if $\Lambda$ contains an $K_Y \times K_Y$ permutation matrix and the matrix of the remaining $K_X - K_Y$ columns can be extended by $K_Y$ columns to a permutation matrix.*

In summary, the maximising and minimising matrices of $\|\cdot\|$ and $\nu$ are identical and are of the following forms:

$$\text{Maximal: } \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad \text{Minimal: } \begin{pmatrix} \frac{1}{n} & \cdots & \frac{1}{n} \\ \vdots & & \vdots \\ \frac{1}{n} & \cdots & \frac{1}{n} \end{pmatrix}. \tag{20}$$

These results show that the two dependency measures $\|\cdot\|_1$ and $\nu$ fulfil important intuitions: they are minimal, when information about $X_{t-1}$ gives us no information about $Y_t$ because, in this case, all columns of $\Lambda$ should be identical and even equal to each other. Maximal dependence is detected if the information about $X_{t-1}$ is maximally influential on $Y_t$. This happens when $\Lambda$ is, respectively can be reduced or extended to, a permutation matrix. This is illustrated in Figure 3.

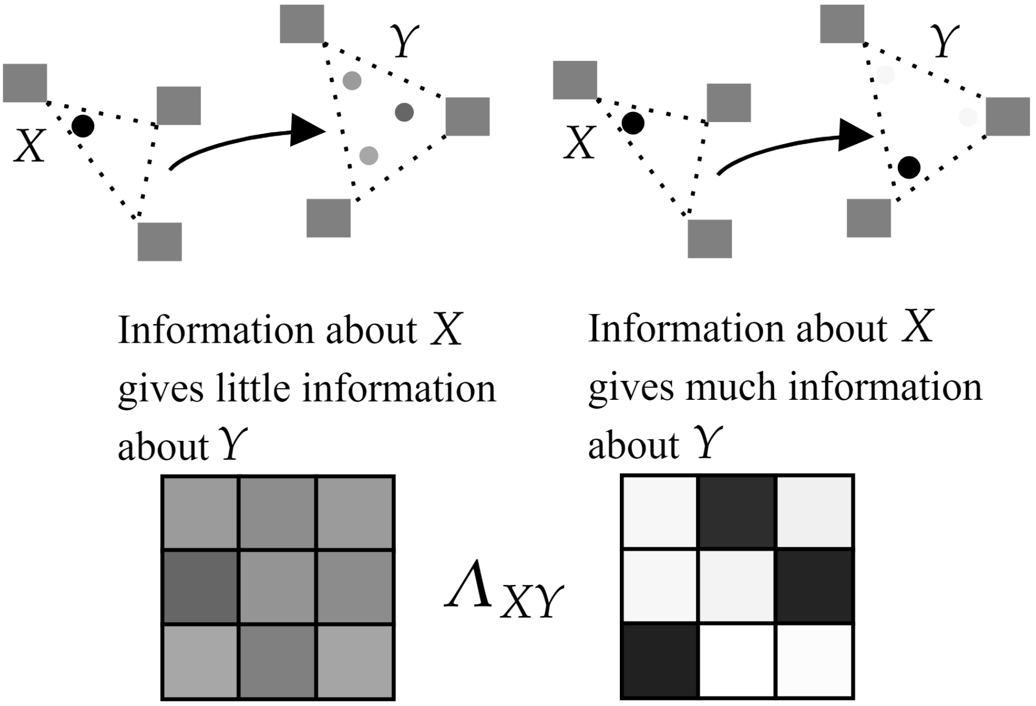
X
Y
X
Y
Information about $X$ gives little information about $Y$
Information about $X$ gives much information about $Y$
$\Lambda_{XY}$

**Figure 3.** Illustration of the intuition behind the dependency measures. If the distribution of $Y$ is independent of $X$, the matrix $\Lambda_{XY}$ will have similar columns (left). If the distribution of $Y$ is strongly dependent on $X$, the columns of $Y$ will be very different from each other.

## 4. Numerical Examples

Now, the two dependency measures are demonstrated on examples of dynamical systems with the aim to cover different structures and properties. In order to assess their efficacy, unidirectional dependencies are explicitly installed in the formulation of the dynamics, and it is investigated whether these are detected. For the usage of the SPA method, the Matlab code provided at https://github.com/SusanneGerber/SPA (accessed 14 August 2019) was used. For two examples, the results are briefly compared with those of Convergent Cross-Mapping (CCM) for which a Matlab implementation provided in [36] was used.

### *4.1. Low-Dimensional and Deterministic: The Hénon System*

The first example is the classical Hénon system [37]. It describes a two-dimensional dynamical system on an attracting set, meaning that all trajectories converge to a specific set of points. Its equations read

$$\begin{aligned} X_t &= 1 - aX_{t-1}^2 + Y_t \\ Y_t &= bX_{t-1}. \end{aligned} \tag{21}$$

For the—typically used—values of $a = 1.4, b = 0.3$, this makes $X$ largely autonomous, while $Y$ is merely a scaled copy of the previous value of $X$. Hence, $Y$ should depend more strongly on the previous value of $X$ than vice versa. Please find an image of a trajectory of this dynamical system in Appendix A.4.

Dependency Analysis

For the dependency analysis, it is investigated whether the higher dependence of $Y$ on $X$ can be reproduced by the dependency measures. A time shift $\tau$ of one time step is chosen for the analysis. Note that for $K > D + 1$, the (SPA 1) solution is not unique in both $\sigma_i$ and $\gamma_t$. As explained before in Remark 3, the solution can depend on the randomly chosen initial values of the optimisation process. Therefore, 20 different solutions for $K_X, K_Y = 2, \ldots, 10$ of length 2000 time steps are computed, the dependency analysis is performed, and the mean over the dependency results is taken for each value of $K_X, K_Y$. The starting values $\gamma_0^X, \gamma_0^Y$ of the trajectories with respect to the landmarks are for both variables always chosen as positive only for the two landmarks directly above and below the initial points $X_0, Y_0$. The result of the analysis is shown in Figure 4.

It can be observed that, consistently, the stronger influence of $X$ on $Y$ is reconstructed since, for all values of $K_X, K_Y$, the green graph corresponding to the direction $X$ to $Y$ lies above the purple one corresponding to the other direction. Surprisingly, the relative difference decreases for increasing values for $K_X, K_Y$ for both measures in this example. Note that with increasing dimension of the $\Lambda$ matrix, the sum of singular values should naturally increase in magnitude. This is not the case for the average row variance for which one divides by the number of columns and rows. It is, however, surprising that for only two landmarks, the difference is that pronounced by $\nu$ compared to a higher number of landmarks. Overall, this example gives a satisfying result for the dependency measures since both were able to detect the higher influence of $X$.

It was checked whether the results were consistent for different (SPA 1) solutions: only in 3 out of 180 cases (9 values for $K_X, K_Y$ and 20 repetitions for each), $Y$ was assigned a higher influence than $X$ by the average row variance, while this did not happen with the Schatten-1 norm. It was also checked whether using equidistant landmarks instead of (SPA 1) solutions, a similar result could be produced. In fact, this led to more strongly emphasised differences and the relative difference being very stable over the number of landmarks (see Appendix A.4). This can be interpreted as an even better result. However, for systems in higher dimensions, this approach entails the problem that the number of landmarks would scale exponentially with the dimension, thereby potentially making it infeasible. This is also discussed in Appendix A.3.

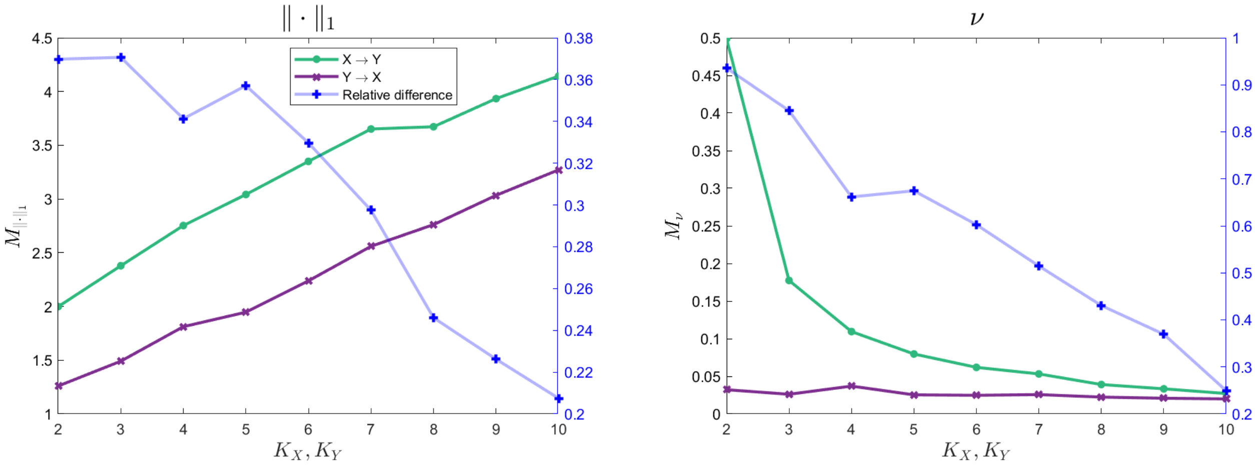

**Figure 4.** Result of the dependency analysis on the Hénon system. The higher influence of $X$ on $Y$ is reconstructed, as can be seen by the green graphs lying above the purple ones. The relative difference surprisingly decreases with increasing number of landmarks.

### *4.2. Low-Dimensional and Stochastic: Continuous Movement by Stochastic Diffusions*

This model describes a continuous evolution of processes $A, B, C$ along solutions of a Stochastic Differential Equation (SDE) [38] where $C$ acts autonomously and $A$ and $B$ hierarchically depend on each other. In short:

$$\begin{tikzcd} C \arrow[r, "\text{influences}"] \arrow[rr, bend left, "\text{influences}"] & B \arrow[r, "\text{influences}"] & A \end{tikzcd}$$

#### 4.2.1. The Dynamics

The SDE is given by

$$\begin{aligned} dC_t &= G(C_t)\,dt + \sigma_C\,dW_{Ct} \\ dB_t &= (\alpha H(B_t) - 10(B_t - C_t))\,dt + \sigma_B Id\,dW_{Bt} \\ dA_t &= (\alpha H(A_t) - 5(A_t - B_t) - 5(A_t - C_t))\,dt + \sigma_A Id\,dW_{At} \end{aligned} \tag{22}$$

where

$$G(x) = \begin{pmatrix} 0 \\ -10(x_2^3 - x_2) \end{pmatrix}, \quad H(x) = \begin{pmatrix} -(x_1^3 - x_1) \\ -1 \end{pmatrix} \tag{23}$$

and $Id$ is the identity matrix.

The choice of the function $G$ ensures a metastable behaviour of $C$ between regions around the values $-1$ and 1 in the $x_2$-coordinate, meaning that it remains around one of the values for some time and, due to the randomness in the SDE, at some point in time, suddenly moves towards the other value. In the $x_1$-direction, the movement of $C$ is determined by noise with covariance matrix $\sigma_C$. The movements of the other processes are partly governed by the function $H$, which gives a metastable behaviour in the $x_1$-coordinates, and by a difference function between themselves and $C$, respectively (Figure 5 for $\sigma_A = \sigma_B = 0.01$). The movement of $A$ depends equally on its difference from $C$ as on its difference from $B$. Since diffusion processes are attracted to low values of their governing potential, this urges $B$ to move into the direction of $C$. Furthermore, it urges $A$ into the directions of both $C$ and $B$. The parameter $\alpha$ therefore governs how autonomous the processes $A$ and $B$ are.

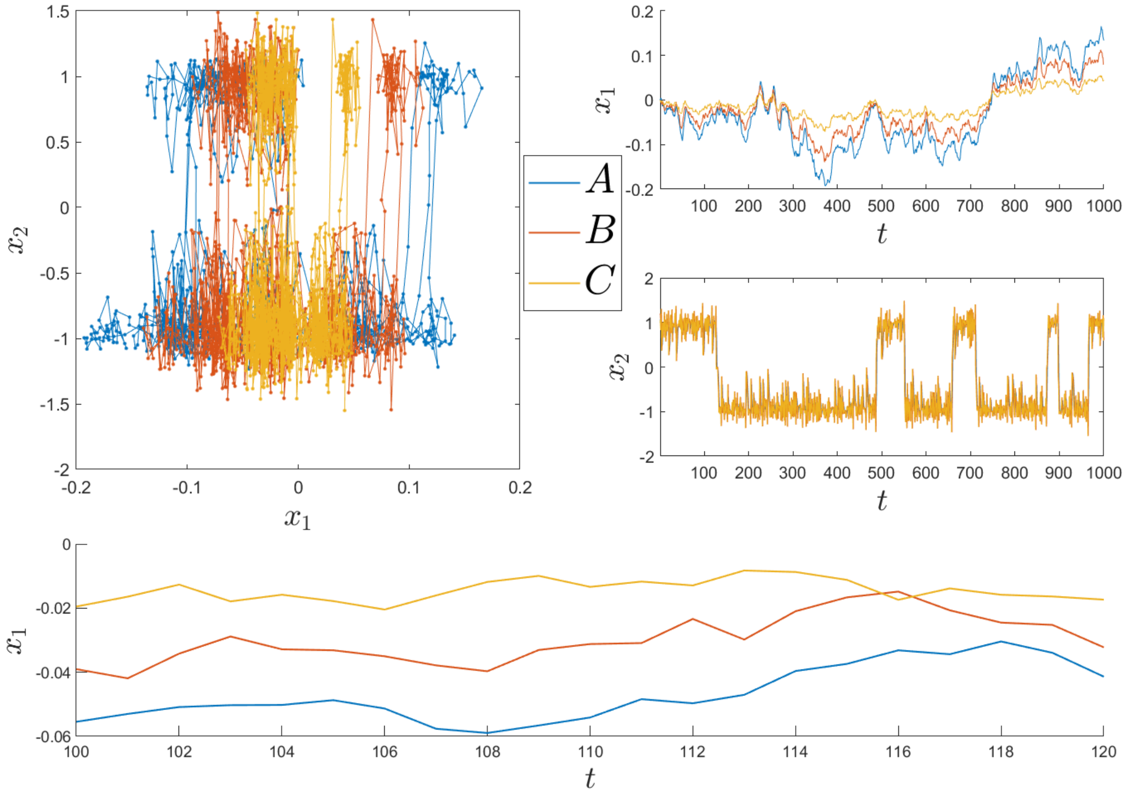

**Figure 5.** Realisation of the system described by Equation (22). $A$ in blue, $B$ in orange, $C$ in yellow. One can see the metastable behaviour in the $x_2$–coordinate in $C$, which the other processes emulate. Parameters: $\alpha = 5$, $\sigma_C = diag((0.01, 0.05))$, $\sigma_B = \sigma_A = 0.01$.

Realisations of the processes are created with the Euler–Maruyama scheme [39] with a time step of size $\Delta t = 0.1$ for 1000 time steps. The parameters were set to $\alpha = 5$ and $\sigma_C = diag((0.01, 0.05))$. For the noise in the evolution of $A$ and $B$, multiple different values $\sigma_B = \sigma_A = 0.01, 0.2, 0.5$ and 1 were used.

### 4.2.2. Dependency Analysis

For SPA I, again, $K_X = K_Y = 10$ was used. Contrary to the previous example, here, the dependencies of the time-differences $\Delta A, \Delta B, \Delta C$ between time steps were computed, meaning that instead of $\Lambda_{XY}^{(\Delta t)}$, $\Lambda_{\Delta X \Delta Y}^{(\Delta t)}$ for $A, B, C = X, Y$ was computed. It was also attempted to reconstruct the dependencies using the process values instead of the time-differences, but, while one could still reconstruct directions of influences, this came with a lower emphasis in the quantitative results. This presumably is the case since the SDE governs not the next state, but rather, the next time-differenced value of each process, making the time-differences a better choice here.

For different values for the noise variances $\sigma_A, \sigma_B$, 50 realisations each of the SDE Equation (22) were created with the same initial conditions, and the analysis was performed for each. The SPA I solution was always computed anew, so that these solutions were generally different from each other, to test the robustness of the method with regard to the SPA I solution.

The results of the dependency analysis well reflected the hierarchical dependencies between the three processes. An exemplary result is given in Equations (24) and (25). The statistics of the overall analysis are given in Table 1. It shows that, in the vast majority of the realisations, the less influential out of two variables was correctly measured as such. The average row variance $\nu$ gives more clear results with the minimal relative difference at least 0.19 for large noise, but generally around 0.4. For the Schatten-1 norm, the relative differences were mostly around 0.2. Note that the results were not strongly influenced by the strength of noise, which loosens the strict dependence of $A$ and $B$ on $C$, except for the dependencies between $A$ and $C$ for $\sigma_A = \sigma_B = 1$ for which often $A$ was falsely measured to have the stronger influence between the two.

It was also checked whether the Pearson correlation coefficient was able to detect the directional influences between the variables. It turned out that for $\sigma_A, \sigma_B = 0.01$, the time-differences between the variables were highly correlated according to the imposed hierarchy, e.g., $\Delta A_{t+\Delta t}$ was highly correlated with $\Delta B_t$, but for $\sigma_A, \sigma_B \geq 0.2$, mostly correlation coefficients below 0.1 could be observed, indicating that the correlation coefficient was not well suited to discover directional dependencies for these dynamics, while the dependency measures could do so. CCM did not manage to correctly identify the direction of dependencies in most realisations of the SDE. By construction, the strengths of CCM rather lied in the domain of deterministic dynamics on an attracting manifold (since it uses the delay-embedding theorem of Takens [11]), but it is less suited for stochastic dynamics, such as this SDE.

$$M_{\|\cdot\|_1}: \begin{array}{|l|l|l|l|}\hline \text{From} \downarrow \text{to} \rightarrow & \Delta A_{t+\Delta t} & \Delta B_{t+\Delta t} & \Delta C_{t+\Delta t} \\ \hline \Delta A_t & 5.51 & 3.86 & 3.46 \\ \hline \Delta B_t & 5.29 & 4.11 & 2.94 \\ \hline \Delta C_t & 4.51 & 7.72 & 4.26 \\ \hline \end{array} \Rightarrow \delta(M_{\|\cdot\|_1}) = \begin{pmatrix} 0 & -0.27 & -0.23 \\ 0.27 & 0 & -0.61 \\ 0.23 & 0.61 & 0 \end{pmatrix} \tag{24}$$

$$M_{\nu}: \begin{array}{|l|l|l|l|}\hline \text{From} \downarrow \text{to} \rightarrow & \Delta A_{t+\Delta t} & \Delta B_{t+\Delta t} & \Delta C_{t+\Delta t} \\ \hline \Delta A_t & 0.039 & 0.021 & 0.018 \\ \hline \Delta B_t & 0.046 & 0.024 & 0.016 \\ \hline \Delta C_t & 0.039 & 0.065 & 0.026 \\ \hline \end{array} \Rightarrow \delta(M_{\nu}) = \begin{pmatrix} 0 & -0.54 & -0.53 \\ 0.54 & 0 & -0.75 \\ 0.53 & 0.75 & 0 \end{pmatrix} \tag{25}$$

**Table 1.** Results of the dependency analysis for the diffusion processes in Equation (22). The third and fifth columns denote the average relative difference between the first variable (in the first row $A$) and the second variable (in the first row $B$) in the Schatten-1 norm and the average row variance. It is always negative, meaning that the first variable (with lower influence) was on average correctly measured as less influential than the second variable. The fourth and sixth columns denote the relative number of occurrences when one variable was falsely identified as more influential, e.g., $A$ as more influential than $B$.

| $\sigma_A, \sigma_B$ | Variables | Average $\delta(M_{\|\cdot\|_1})$ | Incorrect | Average $\delta(M_\nu)$ | Incorrect |
|---|---|---|---|---|---|
| 0.01 | $A, B$ | −0.21 | 0.12 | −0.36 | 0.2 |
| | $A, C$ | −0.22 | 0.08 | −0.56 | 0.08 |
| | $B, C$ | −0.46 | 0 | −0.83 | 0.00 |
| 0.2 | $A, B$ | −0.2 | 0.06 | −0.4 | 0.1 |
| | $A, C$ | −0.17 | 0.16 | −0.51 | 0.14 |
| | $B, C$ | −0.36 | 0 | −0.76 | 0 |
| 0.5 | $A, B$ | −0.2 | 0 | −0.4 | 0.02 |
| | $A, C$ | −0.12 | 0.14 | −0.43 | 0.12 |
| | $B, C$ | −0.30 | 0.02 | −0.71 | 0.04 |
| 1 | $A, B$ | −0.25 | 0.02 | −0.65 | 0 |
| | $A, C$ | −0.04 | 0.28 | −0.19 | 0.30 |
| | $B, C$ | −0.22 | 0.08 | −0.62 | 0.02 |

### 4.3. *Higher-Dimensional, Stochastic, and Delayed: Multidimensional Autoregressive Processes*

In order to demonstrate that influence can be detected for processes whose evolution depends not only on present, but on past terms, as well, realisations of multidimensional linear Autoregressive processes (AR) [40,41] were simulated in which some variables were coupled with others. An $n$-dimensional linear AR($p$) process is a dynamical system of the form

$$X_t = \sum_{i=1}^{p} \phi_i X_{t-1} + \varepsilon_t \tag{26}$$

where $\phi_i \in \mathbb{R}^{n\times n}$ and $\varepsilon_t$ is a stochastic term, which was set to be normally distributed with mean 0 and (positive semi-definite) covariance matrix $C \in \mathbb{R}^{n\times n}$.

### 4.3.1. The Dynamics

Let us now consider AR($p$) processes of the form

$$\begin{pmatrix} X_t \\ Y_t \end{pmatrix} = \sum_{i=1}^{q} \begin{pmatrix} \phi_i^{XX} & \phi_i^{YX} \\ 0 & \phi_i^{YY} \end{pmatrix} \begin{pmatrix} X_{t-i} \\ Y_{t-i} \end{pmatrix} + \varepsilon_t. \tag{27}$$

Specifically, let $X$ and $Y$ be variables in $\mathbb{R}^4$. Thus, $\phi_i^{XX}, \phi_i^{YX}$ and $\phi_i^{YY}$ are matrices in $\mathbb{R}^{4\times 4}$. Through the structure of the coefficient matrices, it was imposed that $X$ is influenced by $Y$, but not vice versa. $p = 3$ was set, and $\phi_i$ were constructed randomly by drawing for each entry a normally distributed value with mean 0 and variance $\sigma_i$, where $\sigma_1 = 0.1$, $\sigma_2 = 0.05, \sigma_3 = 0.03$. This should lead to the coefficients gradually decreasing in magnitude for increasing time lag, so that the most recent process values should have the most influence on future ones. $\varepsilon_t$ is normally distributed with mean 0 and covariance matrix $0.01Id$ (where $Id$ is the identity matrix). Then, a realisation of length $T = 1000$ was created for such a process. This procedure was executed 50 times and the dependency analysis performed on each realisation.

### 4.3.2. Dependency Analysis

$K_X = K_Y = 10$ was chosen again, and $\Lambda_{XY}^{(\tau)}$ and $\Lambda_{YX}^{(\tau)}$ were computed for $\tau = 1, 3, 10, 50$ to investigate how the dependence between the processes evolve with increasing time shift (contrary to the previous example, the process values and not the time differences were used).

One can see in Table 2 that the stronger influence of $Y$ on $X$ is recovered for $\tau = 1$ and $\tau = 3$. For $\tau = 1$, the relative differences were stronger, which is in line with the fact that the AR coefficients $\phi_3$ were selected to be smaller in magnitude than for $\phi_1$, so that $X_t$ should be more strongly influenced by $Y_{t-1}$ than by $Y_{t-3}$. Moreover, not only the relative differences were smaller for $\tau = 3$, but also, the average absolute number of the measures, again correctly indicating a smaller cross-influence with bigger time shift. For $\tau \geq 10$, only negligible differences can be seen. This is consistent with the construction of the processes that include a direct influence up to $\tau = 3$.

Again, it was checked if the Pearson correlation coefficient or CCM could recover the unidirectional dependencies, but this gave negative results: the correlation coefficient was again below 0.1 for most realisations and did not indicate significantly stronger correlation in either direction. CCM indicated only very weak and no directional influence between the variables. By construction, the Granger causality method should perform very well here since the way the dynamics are defined in Equation (27) is perfectly in line with the assumptions of the Granger method. This, of course, cannot be expected for dynamics with an unknown model formulation, such as in the following real-world example.

**Table 2.** Results for dependency measures of $X$ on $Y$ and vice versa for 50 realisations of processes of the form of Equation (27). The third and fifth columns represent the average absolute value of the Schatten-1 norm and average row variance. They decrease with increasing time lag $\tau$, which is consistent with the reconstruction. The fourth and sixth columns denote the average relative differences, which for $\tau = 1, 3$ correctly indicate that $Y$ is more influential on $X$ than vice versa, while for $\tau = 10, 50$, this is no longer the case, which again is consistent with the construction.

| $\tau$ | Direction | Average $\|\cdot\|_1$ | Average $\delta(M_{\|\cdot\|_1})$ | Average $\nu \cdot 10^2$ | Average $\delta(M_\nu)$ |
|---|---|---|---|---|---|
| 1 | $X_{t-1} \to Y_t$ | 1.94 | −0.19 | 1.67 | −0.58 |
| | $Y_{t-1} \to X_t$ | 2.43 | 0.19 | 4.40 | 0.58 |
| 3 | $X_{t-3} \to Y_t$ | 1.88 | −0.15 | 1.43 | −0.49 |
| | $Y_{t-3} \to X_t$ | 2.22 | 0.15 | 3.20 | 0.49 |
| 10 | $X_{t-10} \to Y_t$ | 1.82 | −0.02 | 1.27 | −0.06 |
| | $Y_{t-10} \to X_t$ | 1.86 | 0.02 | 1.38 | 0.06 |
| 50 | $X_{t-50} \to Y_t$ | 1.86 | 0.01 | 1.39 | 0.02 |
| | $Y_{t-50} \to X_t$ | 1.86 | −0.01 | 1.36 | −0.02 |

### 4.4. Real-World Example: Basketball Player Movement

The dependency measures will now be applied to the movement of basketball players during a game and quantify influences between players in the same manner as in the previous examples. For this, player tracking data captured by the SportVU technology (Stats LLC, Chicago, IL, USA) and publicly provided by Neil Johnson [42] from a game of the 2015/16 NBA season between the Dallas Mavericks and the Cleveland Cavaliers, played on 12 January 2016 in Dallas, were used. The data contain the $x$- and $y$-coordinates of each player on the court in 25 time frames per second for most of the 48 min of play. The ball data sometimes seem out of sync with the positions of the players and are not always available; therefore, they were not used in the analysis. The positions were measured in the units feet (ft). Using the same data, although from other games, there are several scientific publications, e.g., [43,44], that focused on different problems than this article does.

In basketball, each team has five players on court at all times. Typically, all five players attack or defend simultaneously so that all ten players are in one half of the court for several seconds, around 10 to 20 s, before moving into the other half. The basketball court has a rectangular shape with a width ($x$ axis) of 94 ft and a length ($y$ axis) of 50 ft. Let us install a coordinate system whose origin is at the mid-point of both axes. If a player is sitting on the bench and not actively participating in the game, the coordinate $(-48.5, -27.5)$ is assigned to him/her, which is slightly outside of the court, but these time frames were excluded from the analysis.

Figure 6 shows the distribution of the positions of Cleveland player LeBron James depending on whether he is attacking or defending and his position on the court over time during the first half of play. One can see that during the attack, James can mostly be found around the three-point-line and, occasionally, closer to the basket, including often at the edge of the small rectangular area around the basket. On defence, he is typically positioned slightly to the left or right of the basket.

#### Applying the Dependency Analysis to the Basketball Data

The dependency analysis on player coordinate data during the first half of the game is now performed, considering only the ten players in the starting lineups of the teams. For the representation of each two-dimensional position of a player, instead of solving (SPA 1), landmark points are chosen in advance, and (SPA 1) is solved only for $\Gamma$, i.e., solely the barycentric coordinates of player coordinates with respect to the landmark points were computed. For the clarity of both visualisation and numerical computations, only the absolute value of the $x$-coordinates were considered, meaning that the coordinates are reflected along the half-court line. For this reason, only the right half of the court has to be taken into account, and the following landmark points were used:

$$\Sigma = \begin{bmatrix} 48.5 & 48.5 & 0 & 0 & 20 & 20 & 40 \\ -27.5 & 27.5 & -27.5 & 27.5 & 15 & -15 & 0 \end{bmatrix} \tag{28}$$

so that $K_X = K_Y = 7$. The landmark points are depicted in Figure 7.

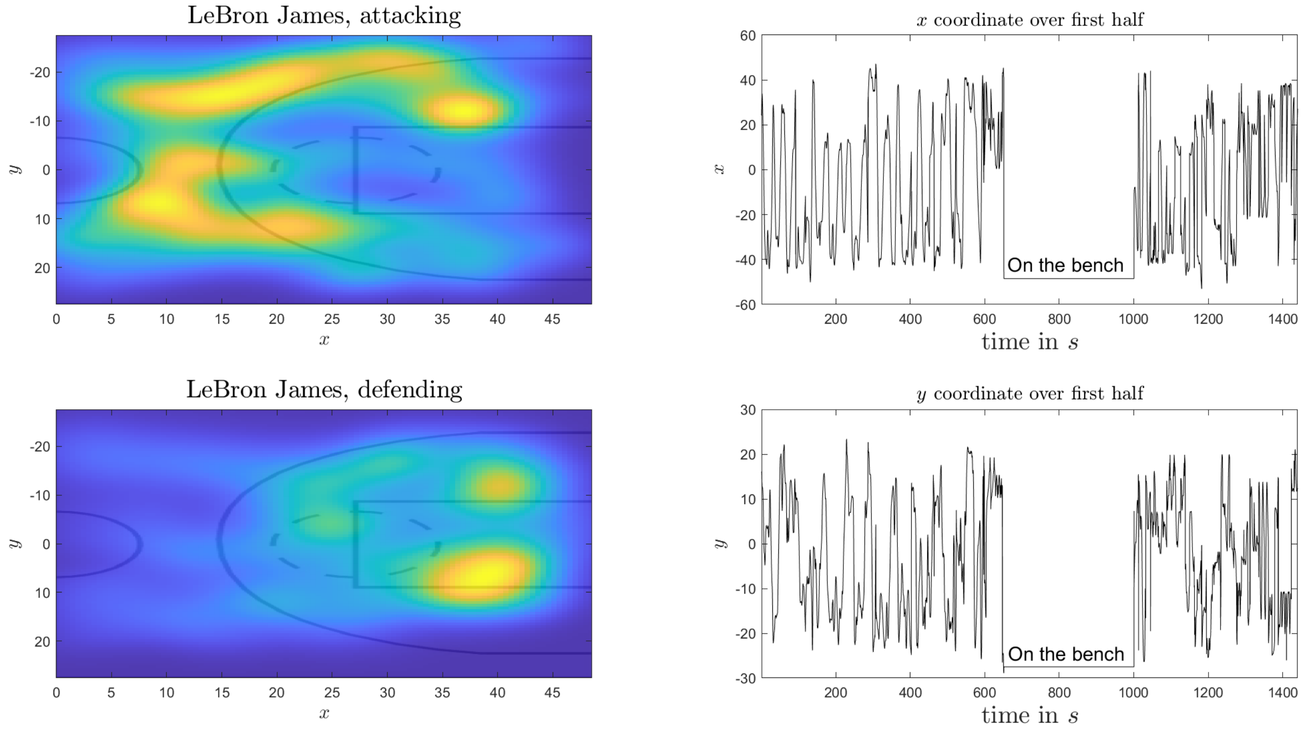

**Figure 6.** (**Left**) Distribution of positions of LeBron James during the first half of the game between the Cleveland Cavaliers and the Dallas Mavericks on 12 January 2016 depending on whether he is in his own team's (defending) or the opponents' half (attacking). The visualisation was derived using Kernel Density Estimation [45]. (**Right**) $x$– and $y$–coordinates over time, measured in ft.

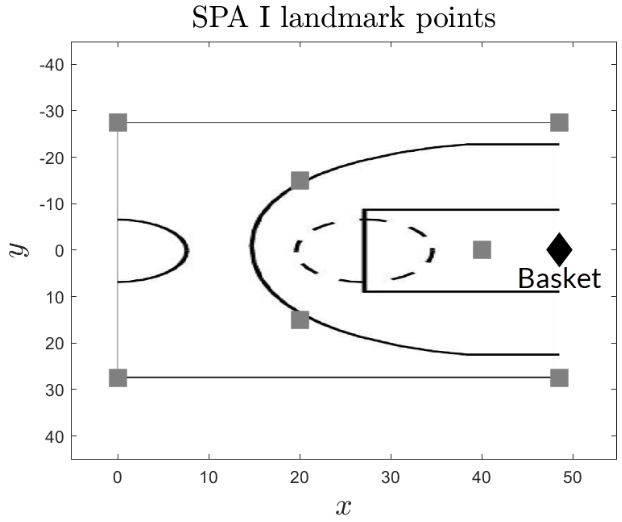

**Figure 7.** Landmark points chosen with respect to which the barycentric coordinates of the player coordinates are computed.

To measure dependencies between each two players, those points in time during the game were used at which both players were on the court, and $\Lambda_{XY}^{(\tau)}$ was computed for each pair of players $X, Y$. The time lag was chosen to be $\tau = 1$ s, so that the influence of the position of a player $X$ for the position of a player $Y$ one second later was investigated.

Note that, as mentioned, basketball games are frequently interrupted for various reason such as fouls or time-outs. This was taken into account by defining an event as the part of play between interruptions and denoting the number of events considered by $L$. Then, the training data were constructed in the form of multiple short time series, i.e., by storing coordinates from the $k$th event, of length $T_k$ seconds, as $\Gamma_k^X = [\gamma_{k,1}^X, \dots, \gamma_{k,T_k-\tau}^X], \Gamma_k^Y = [\gamma_{k,1+\tau}^Y, \dots, \gamma_{k,T_k}^Y]$, and for $\Lambda_{XY}$ minimising $\|[\Gamma_1^Y, \dots, \Gamma_L^Y] - \Lambda[\Gamma_1^X, \dots, \Gamma_L^X]\|_F$ (vice versa for $\Lambda_{YX}$).

Furthermore, it was distinguished between which team was attacking since the decisions of players should be strongly influenced by whether they are attacking or defending. Let us therefore define three different scenarios:

$$\begin{aligned} \text{Dallas attacking:} &\Leftrightarrow \text{at least nine players in Cleveland's half} \\ \text{Cleveland attacking:} &\Leftrightarrow \text{at least nine players in Dallas' half} \\ \text{Transition:} &\Leftrightarrow \text{otherwise} \end{aligned}$$

The analysis on the transition phase is omitted since play is typically rather unstructured during this phase.

The full results of the dependency analysis are, for the sake of visual clarity, shown only in Appendix A.5. Here, in the main text, selected parts are shown, restricted to the average row variance. One can make the following observations:

- As Tables 3 and 4 show, according to the dependency measures, with either team attacking, offensive players seem to have the strong, often the strongest, influence on their direct opponent, i.e., the opposing player playing the same position (Point Guards (PGs) Williams and Irving; Shooting Guards (SGs) Matthews and Smith; Small Forwards (SFs) Parsons and James; Power Forwards (PFs) Nowitzki and Love). Exceptions are Centres (Cs) Pachulia and Thompson. This is intuitive, since in basketball, often times, each offensive player has one direct opponent that follows him.
- One can also see that, typically, the defending Point Guards and Shooting Guards seem to be more strongly influenced by multiple opposing players than the Power Forwards and Centres. The reason for this could be that PFs and Cs usually are more steadily positioned around the basket, while the Guards are more encouraged to actively chase opposing attackers.

  One can see that when Dallas attack, they have a much stronger influence on Cleveland than vice versa, according to the average row variance. The Schatten-1 norm does not strongly emphasise that. When Cleveland attack, the cumulated dependencies are very similar to each other. This can be checked using the full results' data in Appendix A.5. Table 5 also indicates this, as it shows that the relative differences are mostly positive for Dallas' players when Dallas attack.
- When Cleveland attack, Thompson has large positive relative differences over most other players, except for Pachulia (Table 6). This could be explained by the fact that Thompson plays below the basket, giving him a lower position radius, and his position is less dependent on spontaneous or set-up moves by his teammates. Pachulia is his direct opponent from whom he might try to separate himself so that his position is in fact influenced by Pachulia.
- One can check if the players of the depending team orient themselves at the positions of the attacking players. To see this, let us sum over all dependency values between players from opposing teams, i.e., compute $\sum_{X \text{ in team } 1, Y \text{ in team } 2} M_{XY}$ for both dependency measures (see Table 7).

Due to inherent randomness in a sport such as basketball, one must be cautious not to overstate these results. This example is meant to showcase how the method presented in this article can be applied to a complex real-world example. The obtained results, however, seem plausible for the explained reasons. In the future, it would be interesting to see if the method could be tailored more specifically to the basketball context and be used to actually find so-far hidden structures of play.

**Table 3. Dallas attacking: average row variance multiplied by 100** for $\Lambda_{XY}$ for $X$ from attacking players (Dallas) and $Y$ from defending players (Cleveland).

| From ↓ to → | Irving (PG) | Smith (SG) | James (SF) | Love (PF) | Thompson (C) |
|---|---|---|---|---|---|
| Williams (PG) | 2.73 | 1.15 | 1.1 | 0.48 | 0.75 |
| Matthews (SG) | 0.56 | 2.49 | 1.04 | 0.86 | 0.27 |
| Parsons (SF) | 1.25 | 1.83 | 2.45 | 0.57 | 1.21 |
| Nowitzki (PF) | 1.35 | 0.83 | 1.01 | 1.29 | 0.97 |
| Pachulia (C) | 1.2 | 1.77 | 1.21 | 1.09 | 1.02 |

**Table 4. Cleveland attacking: average row variance multiplied by 100** for $\Lambda_{XY}$ for $X$ from attacking players (Cleveland) and $Y$ from defending players (Dallas).

| From ↓ to → | Williams (PG) | Matthews (SG) | Parsons (SF) | Nowitzki (PF) | Pachulia (C) |
|---|---|---|---|---|---|
| Irving (PG) | 2.48 | 1.05 | 0.9 | 1.53 | 0.35 |
| Smith (SG) | 0.67 | 1.8 | 0.84 | 0.43 | 0.33 |
| James (SF) | 1.16 | 0.29 | 1.29 | 0.81 | 0.38 |
| Love (PF) | 0.63 | 0.79 | 0.74 | 1.02 | 0.24 |
| Thompson (C) | 1.01 | 1.56 | 1.2 | 1.49 | 1.04 |

**Table 5. Dallas attacking: relative differences of average row variance** for $\Lambda_{XY}$ for $X$ from attacking players (Dallas) and $Y$ from defending players (Cleveland).

| From ↓ to → | Irving (PG) | Smith (SG) | James (SF) | Love (PF) | Thompson (C) |
|---|---|---|---|---|---|
| Williams (PG) | 0.11 | 0.17 | −0.27 | −0.42 | −0.47 |
| Matthews (SG) | −0.08 | 0.4 | 0.29 | −0.38 | −0.46 |
| Parsons (SF) | −0.12 | 0.34 | 0.1 | −0.28 | −0.37 |
| Nowitzki (PF) | 0.82 | 0.59 | 0.72 | 0.26 | 0.3 |
| Pachulia (C) | 0.55 | 0.62 | 0.57 | 0.16 | 0.21 |

**Table 6. Cleveland attacking: relative differences in the average row variance** for $\Lambda_{XY}$ for $X$ from attacking players (Cleveland) and $Y$ from defending players (Dallas).

| From ↓ to → | Williams (PG) | Matthews (SG) | Parsons (SF) | Nowitzki (PF) | Pachulia (C) |
|---|---|---|---|---|---|
| Irving (PG) | −0.08 | −0.18 | −0.31 | −0.21 | −0.46 |
| Smith (SG) | −0.36 | −0.15 | −0.03 | −0.16 | −0.55 |
| James (SF) | 0.02 | −0.55 | −0.3 | 0.07 | −0.27 |
| Love (PF) | 0.47 | 0.61 | 0.17 | −0.43 | −0.37 |
| Thompson (C) | 0.69 | 0.61 | 0.38 | 0.5 | 0.13 |

**Table 7.** Cumulated sum over the dependency measures considering only one team attacking, the other defending.

| Attacking Team | Measure | $X$ in DAL, $Y$ in CLE | $X$ in CLE, $Y$ in DAL |
|---|---|---|---|
| Dallas | $\|\cdot\|_1$ | 54.4 | 52.2 |
| Dallas | $\nu$ | 0.30 | 0.25 |
| Cleveland | $\|\cdot\|_1$ | 51.4 | 51.3 |
| Cleveland | $\nu$ | 0.25 | 0.24 |

## 5. Conclusions

In this article, a data-driven method for the quantification of influences between variables of a dynamical system was presented. The method deploys the low-cost discretisation algorithm Scalable Probabilistic Approximation (SPA), which represents the data points using fuzzy affiliations, respectively barycentric coordinates with respect to a convex polytope, and estimates a linear mapping between these representations of two variables and forward in time. Two dependency measures were introduced that compute the properties of the mapping and admit a suitable interpretation.

Clearly, many methods for the same aim already exist. However, most of them are suited for specific scenarios and impose assumptions on the relations or the dynamics, which are not always fulfilled. Hence, this method should be a helpful and directly applicable extension to the landscape of already existing methods for the detection of influences.

The advantages of the method lie in the following: it is purely data-driven and, due to the very general structure of the SPA model, requires almost no intuition about the relation between variables and the underlying dynamics. This is in contrast to a method such as Granger causality, which imposes a specific autoregressive model between the dynamics. Furthermore, the presented method is almost parameter-free, since only the numbers of landmark points $K_X$ and $K_Y$ for the representation of points and the time lag $\tau$ have to be specified. Additionally, the dependency measures, the Schatten-1 norm and average row variance, are straightforward to compute from the linear mapping and offer direct interpretation. The capacity of the method to reconstruct influences was demonstrated on multiple examples, including stochastic and memory-exhibiting dynamics.

In the future, it could be worthwhile to find rules for the optimal number of landmark points and their exact positions with respect to the data. Plus, it seems important to investigate how dependencies between variables with differing numbers of landmark points can be compared with the presented dependency measures. Moreover, one could determine additional properties of the matrix $\Lambda_{XY}$ next to the two presented ones. It should also be investigated why in some of the presented examples, the average row variance gave a clearer reconstruction of influences than the Schatten-1 norm and how the latter can be improved. In addition, it should be worthwhile to use the combined solution of the (SPA 1) and (SPA 2) problems as done in [23] and observe whether this improves the performance. Furthermore, constructing a nonlinear SPA model consisting of the multiplication of a linear mapping with a nonlinear function, as done in [26], could give an improved model accuracy and, therefore, a more reliable quantification of influences. Lastly, since in this article, variables were always expressed using a higher number of landmark points, it should be interesting to investigate whether for high-dimensional variables, projecting them to a low-dimensional representation using SPA and performing the same dependency analysis is still sensible. This could be of practical help to shed light on the interplay between variables in high dimensions.

**Funding:** This research was funded by Deutsche Forschungsgemeinschaft, grant CRC 1114 "Scaling Cascades in Complex Systems", Project Number 235221301, Project A01.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The theoretical examples in this article can be reproduced by the details given in the main text. The data or Matlab code can be made available by the author upon reasonable request.

**Acknowledgments:** The publication of this article was funded by Freie Universität Berlin. The author thanks Vikram Sunkara of Zuse Institute Berlin and Illia Horenko from USI Lugano for helpful discussions about the manuscript and the SPA method and Neil Johnson from Monumental Sports & Entertainment for provision of the basketball data.

**Conflicts of Interest:** The author declares no conflict of interest.

## Appendix A

### Appendix A.1. The Dependency Measures in Light of Conditional Expectations

Assume that $X$ does not contain any information for future states of $\Lambda$, then the entries of $\gamma^X_{t-1}$ should have no influence in the composition of the approximation of $\gamma^Y_t$ by $\Lambda_{XY}\gamma^X_{t-1}$. Thus, the columns of $\Lambda_{XY}$ are identical, i.e., $(\Lambda_{XY})_{|1} = \ldots (\Lambda_{XY})_{|K_X} =: \lambda$. It then holds that

$$\Lambda_{XY}\gamma^X \equiv (\Lambda_{XY})_{|1} = \ldots (\Lambda_{XY})_{|K_X} =: \lambda \tag{A1}$$

regardless of $\gamma^X$, since $\Lambda_{XY}$ is column-stochastic and each barycentric coordinate is a stochastic vector.

From classical theory on statistics [46], $\lambda$ should be equal to the mean of the time series $\gamma^Y_2, \ldots, \gamma^Y_T$ in the limit of infinite data:

**Proposition A1.** *Let two time series of length $T$ of barycentric coordinates by given by $\gamma^Y_1, \ldots, \gamma^Y_T$ and $\gamma^X_1, \ldots, \gamma^X_T$. Assume that*

$$\min_{\Lambda^* \in \mathbb{R}^{K_Y \times K_X}} \|[\gamma^Y_2| \cdots |\gamma^Y_T] - \Lambda[\gamma^X_1| \cdots |\gamma^X_{T-1}]\|_F = \min_{\lambda^* \in \mathbb{R}^{K_Y}} \|[\gamma^Y_2| \cdots |\gamma^Y_T] - [\lambda^*, \ldots, \lambda^*]\|_F$$

*Then,*

$$\lim_{T\to\infty} \Lambda_{XY} = \lim_{T\to\infty} [\frac{1}{T}\sum_{i=1}^{T}\gamma^Y_i \ldots \frac{1}{T}\sum_{i=1}^{T}\gamma^Y_i].$$

**Proof.** The minimiser of $\|[\gamma^Y_2| \cdots |\gamma^Y_T] - [\lambda^*, \ldots, \lambda^*]\|_F$ is given by the mean $\lambda^* = \frac{1}{T}\sum_{i=1}^{T}\gamma^Y_i$ for $T \to \infty$. By the assumption of the proposition, the product of $\Lambda_{XY}\gamma^X_t$ should therefore be equal to $\lambda^*$ for each $t$. By Equation (A1), this is achieved by choosing $\Lambda_{XY} = [\lambda^*, \ldots, \lambda^*]$. □

This is consistent with the intuition that if $\gamma^Y_t$ is independent of $\gamma^X_{t-1}$, this means that

$$\mathbb{E}_{\mu_Y}[\gamma^Y_t|\gamma^X_{t-1}] = \mathbb{E}_{\mu_Y}[\gamma^Y_t]. \tag{A2}$$

Therefore, each column of $\Lambda_{XY}$ should be an approximation of $\mathbb{E}_{\mu_Y}[\gamma^Y_t]$. This is naturally given by the statistical mean along a time series.

### Appendix A.2. Proofs of Propositions on the Dependency Measures

In this section, $n$ is written for $K_Y$ and $m$ for $K_X$ to improve the visual clarity.

**Lemma A1.** *The maximal Schatten-1 norm of a column-stochastic $(n \times m)$-matrix is $m$.*

**Proof.** Since $\|\cdot\|_1$ is a norm, the triangle inequality holds and yields

$$\|A + B\|_1 \leq \|A\|_1 + \|B\|_1.$$

Let there be a finite number of matrices $A_i$s that $A = \sum_i A_i$. It then holds that

$$\|A\|_1 \leq \sum_i \|A_i\|_1.$$

Note that $A$ can be written as $A = \sum_{i=1}^{n}\sum_{j=1}^{m} \tilde{A}^{ij}$, where

$$(\tilde{A}^{ij})_{kl} \begin{cases} A_{ij} & (k,l) = (i,j) \\ 0 & \text{else} \end{cases}$$

A matrix with only one nonzero entry $a$ has only one nonzero singular value that is equal to $a$. This means that $\|\tilde{A}^{ij}\|_1 = A_{ij}$. Furthermore, since $A$ is column-stochastic, it holds that $\sum_{i=1}^{n} A_{ij} = 1$. Thus,

$$\|A\|_1 \leq \sum_{i=1}^{n}\sum_{j=1}^{m} \|\tilde{A}^{ij}\| = \sum_{i=1}^{n}\sum_{j=1}^{m} A_{ij} = m.$$

□

**Proposition A2.** *The Schatten-1 norm of a column-stochastic $(n \times m)$-matrix $A$ with $n \geq m$ obtains the maximal value $m$ if deletion of $n-m$ rows of $A$ yields an $m \times m$ permutation matrix (this is Proposition 1).*

**Proof.** Let $A$ be of the form described in the proposition. Then, by the deletion of $n-m$ rows, one can derive a permutation matrix $P$. For those matrices, it holds that

$$PP^T = Id.$$

Thus, the singular values of $P$, which are the square roots of the eigenvalues of $PP^T$, are given by the square roots of the eigenvalues of the $m \times m$ identity matrix. These are given by $\sigma_1 = \cdots = \sigma_m = 1$. Thus, $\|P\|_1 = m$.

All deleted rows must be identical to the zero-vector of length $m$, since the column sums of $A$ have to be equal to 1 and the column sums of $P$ are already equal to 1. Therefore, the singular values of $P$ are equal to the singular values of $A$, and their sum is equal to $m$ because the sum of singular values of a matrix cannot shrink by adding zero rows. This is because

$$AA^T = \begin{pmatrix} P \\ 0 \end{pmatrix} \begin{pmatrix} P^T & 0 \end{pmatrix} = \begin{pmatrix} PP^T & 0 \\ 0 & 0 \end{pmatrix},$$

whose eigenvalues are the eigenvalues of $PP^T$, i.e., the singular values of $P$, and additional zeros. Thus, the sum of singular values does not change. Since $m$ is the maximal value for $\|A\|_1$ by Lemma A1, it holds that $\|A\|_1 = \|P\|_1 = m$. □

**Proposition A3.** *The Schatten-1 norm of a column-stochastic $(n \times m)$-matrix $A$ is minimal if and only if $A_{ij} \equiv \frac{1}{n}$ and in this case is equal to $\sqrt{\frac{m}{n}}$ (this is Proposition 2).*

**Proof.** If all entries of $A$ are given by $\frac{1}{n}$, then

$$(AA^T)_{ij} = \frac{m}{n^2} \quad \text{for all } i,j = 1,\ldots,n. \tag{A3}$$

Then, $AA^T$ has exactly one nonzero eigenvalue, since it is a rank-1 matrix. This is equal to $\frac{m}{n}$ (corresponding to the eigenvector $(1,\ldots,1)^T$), since

$$AA^T \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} = \frac{m}{n^2} \begin{pmatrix} 1 & \ldots & 1 \\ \vdots & & \vdots \\ 1 & \ldots & 1 \end{pmatrix} \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} = \frac{m}{n^2} \begin{pmatrix} n \\ \vdots \\ n \end{pmatrix} = \frac{m}{n} \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}.$$

Singular values of a matrix $A$ are square roots of the eigenvalues of $AA^T$. The square root of the eigenvalue above, which is the only positive singular value, is then $\sqrt{\frac{m}{n}}$. This yields

$$\|A\|_1 = \sqrt{\frac{m}{n}}$$

The reason there cannot be a column-stochastic matrix $B$ with $\|B\|_1 < \sqrt{\frac{m}{n}}$ is the following: It holds that

$$\|B\|_1 \geq \|B\|_2 \quad \text{where } \|B\|_2 := \sqrt{\sum_{i=1}^{min(n,m)} \sigma_i^2} \text{ is the } \textit{Schatten-2-norm.}$$

because it is well-known that the standard 1-norm of a Euclidean vector is bigger or equal to its 2-norm so that $\|(\sigma_1, \ldots, \sigma_{min(n,m)})\|_1 \geq \|(\sigma_1, \ldots, \sigma_{min(n,m)})\|_2$. For the Schatten-2 norm, it holds that

$$\|B\|_2 = \|B\|_F := \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{m} B_{ij}^2}, \tag{A4}$$

which, by a classic linear algebra result, is the Frobenius norm of $B$. Note that, if $B_{ij} = \frac{1}{n}$ for all $i, j$, then

$$\|B\|_2 = \|B\|_F = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{m} \frac{1}{n^2}} = \sqrt{\frac{nm}{n^2}} = \sqrt{\frac{m}{n}}. \tag{A5}$$

Assume that there exists a $B_{ij} < \frac{1}{n}$, e.g., $B_{ij} = \frac{1}{n} - \delta$. Then, since $\sum\limits_{i=1}^{n}\sum\limits_{j=1}^{m} B_{ij} = 1$, there also exists $B_{kj} > \frac{1}{n}$. Immediately, this increases the sum over the squared entries of $B$, simply because $(a+\delta)^2 + (a-\delta)^2 = 2a^2 + 2\delta^2 > 2a^2$ for $\delta > 0$. Therefore, $\|B\|_F$ is minimal only for the choice given above. As a consequence, $\|B\|_2$ is minimal in this case, as well, and thus, this is the only minimiser of $\|B\|_1$. □

**Proposition A4.** *The average row variance of a column-stochastic* $(n \times m)$*-matrix* $A$ *with* $n = m$ *is maximal and equal to* $\frac{1}{m}$ *if it is a permutation matrix (this is Proposition 3).*

**Proof.** For the case $K_Y = K_X$, i.e., $n = m$, this is straightforward: the variance of a row that contains only values between 0 and 1 is maximised if exactly one value is 1 and all other entries are 0. Each column, due to its being stochastic, can only have one 1. For $n \geq m$, one therefore has to distribute the ones into different rows and columns, matching the statement of the proposition. The maximal value is hence equal to the variance of an $m$-dimensional unit vector. Since the mean of this vector is equal to $\frac{1}{m}$, the variance is

$$\begin{aligned}\frac{1}{m-1}((1-\frac{1}{m})^2 + \sum_{i=1}^{m-1}\frac{1}{m^2}) &= \frac{1}{m-1}((\frac{m-1}{m})^2 + \sum_{i=1}^{m-1}\frac{1}{m^2}) \\ &= \frac{1}{m-1}(\frac{m^2-2m+1}{m^2} + \frac{m-1}{m^2}) \\ &= \frac{1}{m-1}\frac{m^2-m}{m^2} \\ &= \frac{1}{m}.\end{aligned} \tag{A6}$$

□

**Proposition A5.** *The average row variance of a column-stochastic matrix* $A$ *obtains the minimal value* 0 *if and only if all columns are equal to each other (this is Proposition 4).*

**Proof.** Trivially, if all columns in $A$ are identical, then the average row variance of $A$ is 0. Since the variance is always non-negative by definition, this is the minimal value. If at least two values differ in a column, then the average row variance immediately becomes positive. □

### *Appendix A.3. Differences from Related Practices*

### Appendix A.3.1. Simple Linear Correlations

The presented measures might seem strongly related to the computation of the linear correlation:

$$C(X,Y) = \frac{1}{T-1}\sum_{t=2}^{T}(X_{t-1} - \bar{X})(Y_t - \bar{Y})^T \tag{A7}$$

(for $\tau = 1$) where $\bar{X}, \bar{Y}$ are the componentwise averages. However, $C$ can only detect *global* linear patterns between $X$ and $Y$. In contrast, points are transformed into a higher-dimensional space by expressing them by barycentric coordinates with $K > D$. While still a linear operator between the variables is determined, given by $\Lambda_{XY}^{(\tau)}$, this is a *local* approximation of a potentially nonlinear function, denoted earlier by $\mathbf{w}$. Furthermore, upon perturbations to the function $\mathbf{w}$, $\Lambda_{XY}^{(\tau)}$ should react in a nonlinear way by construction of the SPA II problem. The dependency measures then are nonlinear functions on $\Lambda_{XY}^{(\tau)}$. In the examples that will follow, using linear correlations could generally not uncover unidirectional dependencies, while the measures presented in this article were able to do so.

### Appendix A.3.2. Granger Causality

A prominent method to measure the influence of one variable on another is by employing the Granger causality framework [12,13]. It functions by determining two models of the form

$$\begin{aligned} Y_t &= f(X_{t-1}, \ldots, X_{t-p}, Y_{t-1}, \ldots, Y_{t-q}) \\ Y_t &= g(Y_{t-1}, \ldots, Y_{t-q}) \end{aligned} \tag{A8}$$

from training data and using them to compute subsequent values of $Y_t$ on testing data that were not used for training. The prediction errors of $f$ and $g$ are then compared. If $f$, which uses the information of $X$, gives a significantly better prediction error, then it is deduced that $X$ influences $Y$.

Typical model forms for $f$ and $g$ are linear autoregressive models [40], which are described in more detail in the next section. It was pointed out in [10] that using past terms of $X$ and $Y$ can constrain the interpretability of the result, since if $Y$ forces $X$, information about $Y$ is stored in past terms of $X$ due to the delay-embedding theorem of Takens [11] (please see [10] including its Supplement for details). Then, if $Y$ can be predicted from past terms of $X$, it actually is a sign that $Y$ forces $X$, not vice versa. This makes the interpretation of the Granger results more difficult. In [10], examples were shown where the Granger causality method failed to detect influences between variables.

This effect does not occur when dispensing of the past terms and instead fitting models $Y_t = f(X_{t-1}, Y_{t-1})$ and $Y_t = g(Y_{t-1})$. However, in systems that are highly stochastic or chaotic, meaning that from similar initial conditions, diverging trajectories emerge, even an accurate model can be prone to give weak prediction errors. In such cases, the prediction error often times has limited meaning.

Furthermore, even if $X$ influences $Y$, one has to select a suitable model family for $f$ and $g$, so that this actually shows. The selection of the model family can be a challenging task of its own.

Nevertheless, Granger causality can be a strong tool for the detection of influences, e.g., as shown for a Gene Regulatory Network in [1].

### Appendix A.3.3. Discretisation by Boxes Instead of Landmark Points

Earlier, the similarity between the model constructed solving (SPA 2) and Markov State Models (MSMs) was mentioned. In MSMs, one discretises the state space into boxes and statistically estimates the transition probabilities of the state of a dynamical system between the boxes, typically by considering the relative frequencies of transitions. One then obtains a column-stochastic transition matrix that contains these relative frequencies. In the same manner, one could compute this matrix for the frequencies that a variable $Y$ is in a certain box at time $t + \tau$ given that a variable $X$ is in a certain box at time $t$ and apply the dependency measures to this transition matrix to assess how meaningful the information about a variable $X$ is for the future value of $Y$. Fixing the edge length of each box, the number of boxes increases exponentially with the dimension of points, while one generally requires an appropriately fine discretisation of the state space to obtain meaningful results, yielding a high number of boxes even for moderate dimensions of the data. One then requires very long time series for a robust estimation of the transition probabilities. The advantage in the SPA I representation of points is that one can derive a column-stochastic matrix, but can maintain a lossless representation of points with $K > D$ as the only prerequisite.

### *Appendix A.4. Dependency Analysis Result for Equidistant Landmarks in the First Example (Hénon System)*

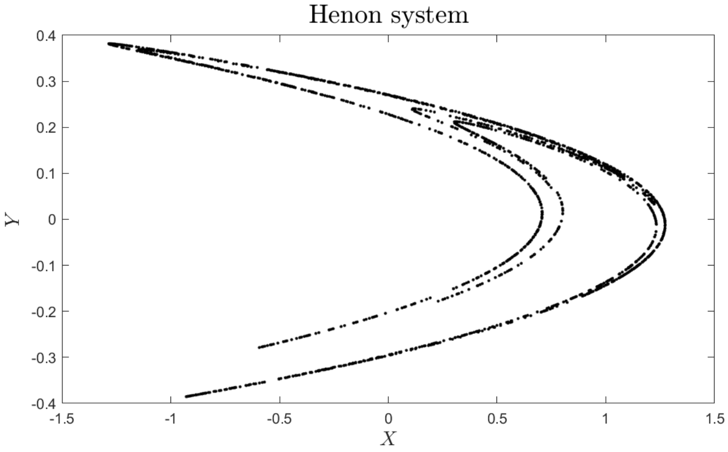

**Figure A1.** Two–dimensional attractor of the Hénon system from Equation (21).

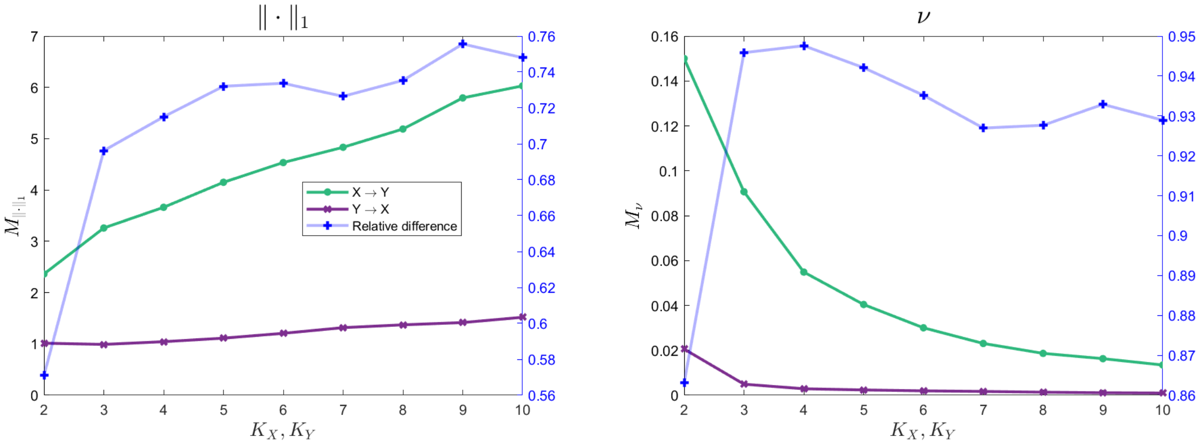

**Figure A2.** Result of the dependency analysis on the Hénon system with equidistant landmarks instead of (SPA 1) solutions. The higher influence of $X$ on $Y$ is reconstructed and pronounced more strongly here as when using the (SPA 1) solutions. Furthermore, the relative difference does not decrease with increasing number of landmarks.

### *Appendix A.5. Full Basketball Dependency Results*

Order of players in all rows and columns: (Dallas) Williams, Matthews, Parsons, Nowitzki, Pachulia, (Cleveland) Irving, Smith, James, Love, Thompson.
Dependencies from one player to himself are omitted; therefore, the zeros are placed on the diagonals.

Dallas attacking

$$M_{\|\cdot\|_1} = \begin{pmatrix} 0 & 2.31 & 2.47 & 1.76 & 1.7 & 2.94 & 2.29 & 2.27 & 1.83 & 1.9 \\ 2.3 & 0 & 2.09 & 1.71 & 2.05 & 1.9 & 2.78 & 2.07 & 2.04 & 1.55 \\ 2.63 & 2.35 & 0 & 1.76 & 2 & 2.23 & 2.52 & 2.77 & 1.85 & 2.13 \\ 2.05 & 1.96 & 1.79 & 0 & 1.76 & 2.05 & 2.02 & 1.99 & 2.17 & 1.94 \\ 2.14 & 2.49 & 2.2 & 1.57 & 0 & 2.27 & 2.47 & 2.2 & 2.18 & 2.08 \\ 2.84 & 1.92 & 2.33 & 1.51 & 1.82 & 0 & 2.25 & 2.23 & 1.96 & 1.99 \\ 2.13 & 2.48 & 2.22 & 1.66 & 1.88 & 2.08 & 0 & 2.39 & 2.29 & 1.75 \\ 2.42 & 1.99 & 2.71 & 1.58 & 1.8 & 2.55 & 2.02 & 0 & 2.13 & 1.87 \\ 2.1 & 2.32 & 1.97 & 2.03 & 2.07 & 2.37 & 2.72 & 2.09 & 0 & 1.73 \\ 2.31 & 1.77 & 2.59 & 1.85 & 1.96 & 2.52 & 2.22 & 2.38 & 1.78 & 0 \end{pmatrix}$$

$$M_\nu = 0.01 \cdot \begin{pmatrix} 0 & 1.51 & 1.71 & 0.52 & 0.49 & 2.73 & 1.15 & 1.1 & 0.48 & 0.75 \\ 1.37 & 0 & 0.87 & 0.53 & 1.24 & 0.56 & 2.49 & 1.04 & 0.86 & 0.27 \\ 1.89 & 1.43 & 0 & 0.66 & 0.86 & 1.25 & 1.83 & 2.45 & 0.57 & 1.21 \\ 0.89 & 0.79 & 0.41 & 0 & 0.57 & 1.35 & 0.83 & 1.01 & 1.29 & 0.97 \\ 1.13 & 1.74 & 1.1 & 0.27 & 0 & 1.2 & 1.77 & 1.21 & 1.09 & 1.02 \\ 2.42 & 0.6 & 1.41 & 0.24 & 0.54 & 0 & 1.08 & 1.23 & 0.69 & 0.76 \\ 0.96 & 1.5 & 1.21 & 0.34 & 0.68 & 0.85 & 0 & 1.57 & 1.27 & 0.62 \\ 1.51 & 0.74 & 2.22 & 0.29 & 0.52 & 1.67 & 0.84 & 0 & 1.38 & 0.57 \\ 0.83 & 1.37 & 0.79 & 0.96 & 0.92 & 1.38 & 2.55 & 1.06 & 0 & 0.56 \\ 1.41 & 0.51 & 1.92 & 0.68 & 0.81 & 1.95 & 1.13 & 1.69 & 0.49 & 0 \end{pmatrix}$$

Cleveland attacking

$$M_{\|\cdot\|_1} = \begin{pmatrix} 0 & 2.04 & 2.71 & 1.99 & 1.78 & 2.89 & 2.19 & 2.11 & 1.62 & 1.62 \\ 2.05 & 0 & 2.26 & 2.22 & 1.9 & 2.28 & 2.54 & 1.81 & 1.64 & 1.82 \\ 2.49 & 2.22 & 0 & 2.08 & 1.77 & 2.21 & 2.1 & 2.53 & 1.89 & 1.96 \\ 2.01 & 2.5 & 1.89 & 0 & 1.97 & 2.51 & 1.81 & 1.97 & 2.49 & 1.99 \\ 2.04 & 2.42 & 2 & 2 & 0 & 1.9 & 1.95 & 1.84 & 1.72 & 2.02 \\ 2.89 & 2.25 & 2.01 & 2.37 & 1.66 & 0 & 2.41 & 2.52 & 2.08 & 1.75 \\ 1.97 & 2.46 & 2.07 & 1.76 & 1.6 & 1.96 & 0 & 2.14 & 2.03 & 1.82 \\ 2.11 & 1.66 & 2.27 & 1.91 & 1.65 & 2.26 & 2.01 & 0 & 1.89 & 1.78 \\ 1.79 & 1.99 & 1.99 & 2.09 & 1.57 & 1.88 & 2.34 & 2.1 & 0 & 1.63 \\ 2.13 & 2.4 & 2.33 & 2.26 & 2.11 & 2.23 & 2.33 & 2.29 & 1.86 & 0 \end{pmatrix}$$

$$M_\nu = 0.01 \cdot \begin{pmatrix} 0 & 0.97 & 2.27 & 0.93 & 0.66 & 2.71 & 1.04 & 1.13 & 0.33 & 0.31 \\ 0.83 & 0 & 1.21 & 1.13 & 0.89 & 1.28 & 2.13 & 0.64 & 0.31 & 0.61 \\ 1.58 & 1.24 & 0 & 1.14 & 0.58 & 1.29 & 0.86 & 1.83 & 0.61 & 0.74 \\ 0.7 & 1.45 & 0.49 & 0 & 0.75 & 1.93 & 0.51 & 0.76 & 1.79 & 0.75 \\ 0.85 & 1.54 & 0.83 & 0.66 & 0 & 0.65 & 0.73 & 0.53 & 0.39 & 0.91 \\ 2.48 & 1.05 & 0.9 & 1.53 & 0.35 & 0 & 1.81 & 1.81 & 0.95 & 0.45 \\ 0.67 & 1.8 & 0.84 & 0.43 & 0.33 & 0.71 & 0 & 1.11 & 0.85 & 0.64 \\ 1.16 & 0.29 & 1.29 & 0.81 & 0.38 & 1.28 & 0.66 & 0 & 0.57 & 0.6 \\ 0.63 & 0.79 & 0.74 & 1.02 & 0.24 & 0.58 & 1.42 & 1.01 & 0 & 0.39 \\ 1.01 & 1.56 & 1.2 & 1.49 & 1.04 & 1.31 & 1.19 & 1.57 & 0.78 & 0 \end{pmatrix}$$

## References

1. Papili Gao, N.; Ud-Dean, S.M.M.; Gandrillon, O.; Gunawan, R. SINCERITIES: Inferring gene regulatory networks from time-stamped single cell transcriptional expression profiles. *Bioinformatics* **2017**, *34*, 258–266. [CrossRef]
2. Shimada, K.; Bachman, J.A.; Muhlich, J.L.; Mitchison, T.J. shinyDepMap, a tool to identify targetable cancer genes and their functional connections from Cancer Dependency Map data. *eLife* **2021**, *10*, e57116. [CrossRef]

3. Yetkiner, H.; Beyzatlar, M.A. The Granger-causality between wealth and transportation: A panel data approach. *Transp. Policy* **2020**, *97*, 19–25. [CrossRef]
4. Nakayama, S.I.; Takasuka, A.; Ichinokawa, M.; Okamura, H. Climate change and interspecific interactions drive species alternations between anchovy and sardine in the western North Pacific: Detection of causality by convergent cross mapping. *Fish. Oceanogr.* **2018**, *27*, 312–322. [CrossRef]
5. Runge, J.; Tibau Alberdi, X.A.; Bruhns, M.; Muñoz, J.; Camps-Valls, G. The Causality for Climate Challenge. In Proceedings of the NeurIPS2019 Competition & Demonstration Track PMLR Post-proceedings, Vancouver, BC, Cananada, 8–14 December 2019.
6. Runge, J.; Bathiany, S.; Bollt, E.; Camps-Valls, G.; Coumou, D.; Deyle, E.; Glymour, C.; Kretschmer, M.; Mahecha, M.; Muñoz, J.; et al. Inferring causation from time series in Earth system sciences. *Nat. Commun.* **2019**, *10*, 2553. [CrossRef]
7. Pacini, C.; Dempster, J.; Boyle, I.; Gonçalves, E.; Najgebauer, H.; Karakoc, E.; Meer, D.; Barthorpe, A.; Lightfoot, H.; Jaaks, P.; et al. Integrated cross-study datasets of genetic dependencies in cancer. *Nat. Commun.* **2021**, *12*, 1661. [CrossRef]
8. Chiu, Y.C.; Zheng, S.; Wang, L.J.; Iskra, B.; Rao, M.; Houghton, P.; Huang, Y.; Chen, Y. Predicting and characterizing a cancer dependency map of tumors with deep learning. *Sci. Adv.* **2021**, *7*, eabh1275. [CrossRef] [PubMed]
9. Tsherniak, A.; Vazquez, F.; Montgomery, P.G.; Weir, B.A.; Kryukov, G.; Cowley, G.S.; Gill, S.; Harrington, W.F.; Pantel, S.; Krill-Burger, J.M.; et al. Defining a Cancer Dependency Map. *Cell* **2017**, *170*, 564–576.e16. [CrossRef] [PubMed]
10. Sugihara, G.; May, R.; Ye, H.; Hsieh, C.; Deyle, E.; Fogarty, M.; Munch, S. Detecting Causality in Complex Ecosystems. *Science* **2012**, *338*, 496–500. [CrossRef] [PubMed]
11. Takens, F. Detecting Strange Attractors in Turbulence. *Lect. Notes Math.* **2006**, *898*, 366–381. [CrossRef]
12. Granger, C. Testing for causality: A personal viewpoint. *J. Econ. Dyn. Control* **1980**, *2*, 329–352. [CrossRef]
13. Kirchgässner, G.; Wolters, J. Granger Causality. In *Introduction to Modern Time Series Analysis*; Springer: Berlin/Heidelberg, Germany, 2007; pp. 93–123. [CrossRef]
14. Pearson, K. Note on Regression and Inheritance in the Case of Two Parents. *Proc. R. Soc. Lond. Ser. I* **1895**, *58*, 240–242.
15. Cliff, O.; Novelli, L.; Fulcher, B.; Shine, J.; Lizier, J. Assessing the significance of directed and multivariate measures of linear dependence between time series. *Phys. Rev. Res.* **2021**, *3*, 013145. [CrossRef]
16. Tourassi, G.; Frederick, E.; Markey, M.; Floyd, C. Application of the mutual information criterion for feature selection in computer-aided diagnosis. *Med. Phys.* **2002**, *28*, 2394–2402. [CrossRef]
17. Heller, R.; Heller, Y.; Gorfine, M. A consistent multivariate test of association based on ranks of distances. *Biometrika* **2013**, *100*, 503–510. [CrossRef]
18. Puka, L. Kendall's Tau. In *International Encyclopedia of Statistical Science*; Lovric, M., Ed.; Springer: Berlin/Heidelberg, Germany, 2011; pp. 713–715. [CrossRef]
19. Ursino, M.; Ricci, G.; Magosso, E. Transfer Entropy as a Measure of Brain Connectivity: A Critical Analysis with the Help of Neural Mass Models. *Front. Comput. Neurosci.* **2020**, *14*, 45. [CrossRef]
20. Gretton, A.; Fukumizu, K.; Teo, C.; Song, L.; Schölkopf, B.; Smola, A. A Kernel Statistical Test of Independence. In Proceedings of the Advances in Neural Information Processing Systems 20, Vancouver, BC, Canada, 3–6 December 2007; Max-Planck-Gesellschaft, Curran: Red Hook, NY, USA, 2008; pp. 585–592.
21. Krakovská, A.; Jakubík, J.; Chvosteková, M.; Coufal, D.; Jajcay, N.; Paluš, M. Comparison of six methods for the detection of causality in a bivariate time series. *Phys. Rev. E* **2018**, *97*, 042207. [CrossRef]
22. Cliff, O.; Lizier, J.; Tsuchiya, N.; Fulcher, B. Unifying Pairwise Interactions in Complex Dynamics. *arXiv* **2022**, arXiv:2201.11941.
23. Gerber, S.; Pospisil, L.; Navandar, M.; Horenko, I. Low-cost scalable discretization, prediction, and feature selection for complex systems. *Sci. Adv.* **2020**, *6*, eaaw0961. [CrossRef]
24. Horenko, I. On a Scalable Entropic Breaching of the Overfitting Barrier for Small Data Problems in Machine Learning. *Neural Comput.* **2020**, *32*, 1563–1579. [CrossRef]
25. Ye, J. Generalized Low Rank Approximations of Matrices. *Mach. Learn.* **2004**, *61*, 112. [CrossRef]
26. Wulkow, N.; Koltai, P.; Sunkara, V.; Schütte, C. Data-driven modelling of nonlinear dynamics by barycentric coordinates and memory. *arXiv* **2021**, arXiv:2112.06742.
27. Weber, M.; Kube, S. Robust Perron Cluster Analysis for Various Applications in Computational Life Science. In Proceedings of the Computational Life Sciences, Konstanz, Germany, 25–27 September 2005; Berthold, R.M., Glen, R.C., Diederichs, K., Kohlbacher, O., Fischer, I., Eds.; Springer: Berlin/Heidelberg, Germany, 2005; pp. 57–66.
28. Anisimov, D.; Deng, C.; Hormann, K. Subdividing barycentric coordinates. *Comput. Aided Geom. Des.* **2016**, *43*, 172–185. [CrossRef]
29. MacQueen, J.B. Some Methods for Classification and Analysis of MultiVariate Observations. In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, the Statistical Laboratory University of California, Berkeley CA, USA, 21 June–18 July 1965, 27 December 1965–7 January 1966; Cam, L.M.L., Neyman, J., Eds.; University of California Press: Berkeley, CA, USA, 1967; Volume 1, pp. 281–297.
30. Shi, G. Chapter 8—Kriging. In *Data Mining and Knowledge Discovery for Geoscientists*; Shi, G., Ed.; Elsevier: Oxford, UK, 2014; pp. 238–274. [CrossRef]
31. Sarich, M. Projected Transfer Operators. Ph.D. Thesis, Freie Universität Berlin, Berlin, Germany, 2011.
32. Husic, B.; Pande, V. Markov State Models: From an Art to a Science. *J. Am. Chem. Soc.* **2018**, *140*, 2386—2396. [CrossRef]
33. Rodrigues, D.R.; Everschor-Sitte, K.; Gerber, S.; Horenko, I. A deeper look into natural sciences with physics-based and data-driven measures. *iScience* **2021**, *24*, 102171. [CrossRef]

34. Golub, G.H.; Loan, C.F.V. *Matrix Computations (Johns Hopkins Studies in the Mathematical Sciences)*; Johns Hopkins University Press: Baltimore, MD, USA, 2013.
35. Lefkimmiatis, S.; Ward, J.P.; Unser, M. Hessian Schatten-Norm Regularization for Linear Inverse Problems. *IEEE Trans. Image Process.* **2013**, *22*, 1873–1888. [CrossRef] [PubMed]
36. Krakovská, A.; Jakubik, J.; Budácoá, H.; Holecyová, M. Causality studied in reconstructed state space. Examples of uni-directionally connected chaotic systems. *arXiv* **2016**, arXiv:1511.00505.
37. Hénon, M. A two-dimensional mapping with a strange attractor. *Commun. Math. Phys.* **1976**, *50*, 69–77. [CrossRef]
38. Øksendal, B. *Stochastic Differential Equations*; Springer: Berlin/Heidelberg, Germany, 2003; ISBN 978-3540047582. [CrossRef]
39. Kloeden, P.; Platen, E. *Numerical Solution of Stochastic Differential Equations*; Springer: Berlin/Heidelberg, Germany, 1992; ISBN 978-3642081071. [CrossRef]
40. Brockwell, P.J.; Davis, R.A. *Time Series: Theory and Methods*; Springer: Berlin/Heidelberg, Germany, 1991.
41. Shumway, R.; Stoffer, D. *Time Series Analysis and Its Applications with R Examples*, 3rd ed.; Springer: New York, NY, USA; Dordrecht/Heidelberg, Germany; London, UK, 2011; ISBN 978-1441978646. [CrossRef]
42. Johnson, N. 2016. Available online: github.com/neilmj/BasketballData (accessed on 19 July 2019).
43. Wu, S.; Bornn, L. Modeling Offensive Player Movement in Professional Basketball. *Am. Stat.* **2018**, *72*, 72–79. [CrossRef]
44. Daly-Grafstein, D.; Bornn, L. Using in-game shot trajectories to better understand defensive impact in the NBA. *J. Sport. Anal.* **2020**, *6*, 235–242. [CrossRef]
45. Parzen, E. On Estimation of a Probability Density Function and Mode. *Ann. Math. Stat.* **1962**, *33*, 1065–1076. [CrossRef]
46. Spokoiny, V.; Dickhaus, T. *Basics of Modern Mathematical Statistics*; Springer: Berlin/Heidelberg, Germany, 2015.