UNIVERSITÄT OSNABRÜCK

University of Osnabrück

Doctoral Dissertation

Time Series Analysis informed by Dynamical Systems Theory

by

Johannes Schumacher

from

Siegen

A thesis submitted in fulfilment of the requirements
for the degree of Dr. rer. nat.

Neuroinformatics Department
Institute of Cognitive Science
Faculty of Human Sciences

May 16, 2015

Dedicated to the loving memory of Karlhorst Dickel (1928 – 1997) whose calm
and thoughtful ways have inspired me to become an analytical thinker.

ABSTRACT

This thesis investigates time series analysis tools for prediction, as well as detection and characterization of dependencies, informed by dynamical systems theory. Emphasis is placed on the role of delays with respect to information processing in dynamical systems, as well as with respect to their effect in causal interactions between systems.

The three main features that characterize this work are, first, the assumption that time series are measurements of complex deterministic systems. As a result, functional mappings for statistical models in all methods are justified by concepts from dynamical systems theory. To bridge the gap between dynamical systems theory and data, differential topology is employed in the analysis. Second, the Bayesian paradigm of statistical inference is used to formalize uncertainty by means of a consistent theoretical apparatus with axiomatic foundation. Third, the statistical models are strongly informed by modern nonlinear concepts from machine learning and nonparametric modeling approaches, such as Gaussian process theory. Consequently, unbiased approximations of the functional mappings implied by the prior system level analysis can be achieved.

Applications are considered foremost with respect to computational neuroscience but extend to generic time series measurements.

PUBLICATIONS

In the following, the publications that form the main body of this thesis are listed with corresponding chapters.

Chapter 5:

> Johannes Schumacher, Hazem Toutounji and Gordon Pipa (2014). *An introduction to delay-coupled reservoir computing*. Springer Series in Bio-/Neuroinformatics 4, Artificial Neural Networks – Methods and Applications, P. Koprinkova-Hristova et al. (eds.). Springer International Publishing Switzerland 2015, `10.1007/978-3-319-09903-3_4`
>
> Johannes Schumacher, Hazem Toutounji and Gordon Pipa (2013). *An analytical approach to single-node delay-coupled reservoir computing*. Artificial Neural Networks and Machine Learning – ICANN 2013, Lecture Notes in Computer Science Volume 8131, Springer, pp 26-33.

Chapter 6:

> Hazem Toutounji, Johannes Schumacher and Gordon Pipa (2015). *Homeostatic plasticity for single node delay-coupled reservoir computing*. Neural Computation, `10.1162/NECO_a_00737`.

Chapter 7:

> Johannes Schumacher, Robert Haslinger and Gordon Pipa (2012). *A statistical modeling approach for detecting generalized synchronization*. Physical Review E, `85.5(2012):056215`.

Chapter 8:

> Johannes Schumacher, Thomas Wunderle, Pascal Fries and Gordon Pipa (2015). *A statistical framework to infer delay and direction of information flow from measurements of complex systems*. Accepted for publication in Neural Computation (MIT Press Journals).

NOTHING.

A monk asked, “What about it when I don’t understand at all?”
The master said, “I don’t understand even more so.”
The monk said, “Do you know that or not?”
The master said, “I’m not wooden-headed, what don’t I know?”
The monk said, “That’s a fine ’not understanding’.”
The master clapped his hands and laughed.

The Recorded Sayings of Zen Master Joshu [Shi and Green, 1998]

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

DCR	Delay-coupled reservoir
RC	Reservoir computing
REG	Reconstruction error graph
GS	Generalized synchronization
DDE	Delay differential equation
LFP	Local field potential

Part I

INTRODUCTION

This part motivates the general problem, states the scientific goals and provides an introduction to the theory of statistical inference, as well as to the reconstruction of dynamical systems from observed measurements.

1 TIME SERIES AND COMPLEX SYSTEMS

This thesis documents methodological research in the area of time series analysis, with applications in neuroscience. The objectives are prediction, as well as detection and characterization of dependencies, with an emphasis on delayed interactions. Three main features characterize this work. First, it is assumed that time series are measurements of complex deterministic systems. Accordingly, concepts from differential topology and dynamical systems theory are invoked to derive the existence of functional relationships that can be employed for inference. Data analysis is thus complemented by insight from research of coupled dynamical systems, in particular chaos theory. This is in contrast to classical methods of time series analysis which, by and large, derive from a theory of stochastic processes that does not consider how the data was generated. Second, it is attempted to employ, as rigorously as practical necessity allows it, the *Bayesian paradigm of statistical inference*. This enables one to incorporate more realistic assumptions about the uncertainty pertaining to the measurements of complex systems and allows for consistent scientific reasoning under uncertainty. Third, the resulting statistical models make use of advanced nonlinear concepts from a modern machine learning perspective to reduce modeling bias, in contrast to many classical methods that are inherently linear and still pervade large communities of practitioners in different areas.

In summary, the work presented here contributes to the growing body of nonlinear methods in time series analysis, with an emphasis on the role of delayed interactions and statistical modeling using Bayesian inference calculus.

The remainder of this thesis is organized as follows. The present chapter continuous with a more detailed account of the problem statement and the motivation for this work. In a next step, the necessity and possibility of an axiomatic theory of statistical inference is discussed in some detail in the context of the Bayesian paradigm. This is followed by an account of the body of theory from differential topology which is referred to as *embedding theory*. Its invaluable contribution as interface between data and underlying dynamical system is highlighted with regard to its use in Part ii. The introduction concludes with a chapter that relates and outlines the published work representing the cumulative content of this thesis. The latter is then documented in Part ii, while Part iii contains a general discussion and conclusion.

1.1 MOTIVATION AND PROBLEM STATEMENT

Our world can be structured into complex dynamical systems that are in constant interaction. Their system states evolve continuously in time such that subsequent states causally depend on the preceding ones. Complexity can be characterized intuitively by the amount of information necessary to predict a system's evolution accurately. System behavior may be complex for different reasons. In a low-dimensional system with a simple but nonlinear temporal evolution rule, the latter may give rise to an intrinsically complex, even fractal, state space manifold on which the system evolves chaotically. Such systems are only truly predictable if their states are known with arbitrary precision. A different type of complexity arises in very high-dimensional systems that are already characterized by a complex temporal evolution rule. Imagine a large hall containing thousands of soundproofed cubicles in each of which a person claps his hands at a particular frequency. To account for the collective behavior of these oscillators, frequency and phase of each person's clapping would have to be known.

Inference in natural science with respect to such systems is necessarily based on empirical measurements. Measurements always cause a substantial loss of information: Individual samples of the otherwise continuously evolving system states are gathered in discrete sets, the states are measured with finite precision, and measurements often map high-dimensional system states non-injectively into low-dimensional sample values. The sample sets are indexed by time and called *time series*. Inference based on time series is thus usually subject to a high degree of uncertainty.

Consider for example a time series $x = (x_i)_{i=1}^n, x_i \in \mathbb{N}$, the samples of which represent the number of apples person X has eaten in his life, as polled once a year. Although it is clear that there is a strong dependency $x_{i+1} \geq x_i$, the time series is near void of information regarding the temporal evolution of any underlying system whose dynamics lead to apples being eaten by X. With respect to predicting the increment $dx_{k+1} = x_{k+1} - x_k$ while lacking knowledge other than $(x_i)_{i=1}^k$, dx_{k+1} is practically a *random event* without discernible cause. A probability distribution over possible values for dx_{k+1} has to summarize this uncertainty and may possibly be inferred from historical samples. Consequently, x is essentially a stochastic black-box, described by a *random process*.

The most prominent continuous random process is the so-called *Wiener process*, which is the mathematical formalization of the phenomenon of Brownian motion. The latter pertains to the random movement of particles suspended in fluids, which was discovered in the early 19^{th} century by Robert Brown, a Scottish botanist. A corresponding time-discrete realization is called a white noise process. The Wiener process can be used to generate more general stochastic processes via Itō's theory of a stochastic calculus. They are used ubiquitously in areas of mathematical finance and econometrics and form a body of statistical methods to which the term *time series analysis* commonly refers (see Box et al. [2013]; Neusser [2011]; Kreiß and Neuhaus [2006]). These methods are often characterized by linearity in functional dependencies. An example of the latter are autoregressive mappings which formalize the statistical dependence of a process state at time index i on its past values at indeces $j < i$. Such autoregressive processes are rather the result of mathematical considerations and derivations from simpler processes than the product of an informed modeling attempt. Similar to the example discussed before, these models are stochastic black-boxes that do not consider how the data was generated. Although this level of abstraction is suitable in the example above, time series may be substantially more informative about the underlying dynamical systems. In this situation, inference may strongly benefit from exploiting the additional information.

In this regard, a growing body of nonlinear methods is emerging that adopts the dynamical systems view and employs forms of nonlinear methodology (see Kantz and Schreiber [2004] and, in particular, Mees [2001]). Informed mainly by corresponding deterministic concepts, it is not uncommon that these approaches, too, refrain from considering more advanced formalizations of uncertainty, as outlined in Chapter 2. However, if the complex systems view is adopted, additional theory can inform the practitioner in analyses. An elaborate theoretical branch of differential topology, *embedding theory*, allows one to reconstruct an underlying dynamical system from its observed measurements, including its topological invariants and the flow describing the temporal evolution, even in the presence of actual measurement noise. This yields, amongst other things, an informed justification for the existence of nonlinear autoregressive moving average models (NARMA) [Stark et al., 2003] for prediction, although this important result often appears to be underappreciated. Moreover, conditions are defined under which further hidden drivers may be reconstructed from measurements of a driven system alone. As will be shown, these and other insights from differential topology can be used to estimate delays, as well as the direction of information flow between the systems underlying measurement time series.

The conditions for *reconstructibility* of the underlying systems pertain largely to their dimensionality with respect to the amount of available time series data. Consider again the system of people clapping in soundproofed cubicles. This is a high-dimensional system of uncoupled oscillators that do not interact. If local acoustic coupling is introduced, as given in an applauding audience, it is a well-known phenomenon that people tend to spontaneously synchronize their clapping under such conditions. Synchrony is a form of self-organization ubiquitous in complex natural systems, be it groups of blinking fireflies or the concerted actions of neuronal populations that process information in the brain. In this particular example, the originally high-dimensional dynamics collapse onto a a low-dimensional synchronization manifold. In a synchronized state, the collective clapping amounts to a single oscillator that is described by a single phase-frequency pair. In general, a high-dimensional or even infinite-dimensional system may exhibit bounded attractor-dynamics that are intrinsically low-dimensional and thus reconstructible from data. These reductions in dimensionality typically arise from information exchange between subsystems, mediated via the network coupling structures of the global system. In the brain, for example, such concerted dynamics lead to extremely well-structured forms of information processing. During epilepsy, on the other hand, physiological malformations cause a pathological form of mass-synchrony in the cortex. The ensuing catastrophic loss of dimensionality is tantamount to a complete loss of information processing. Such catastrophic events are prone to occur if individual subsystems can exert hub-like strong global influence on the rest of the system. With respect to time series of stock prices, individual trading participants of the system have the capacity to induce herd dynamics and cause crashes, which may also be seen as the result of abundant global information exchange in a strongly interconnected network.

In the stock market example, it is clear that information propagates with varying delay but never actually instantaneous. Past events and trends in a time series will be picked up by traders and acted upon such that future states are delay-coupled to past states. Delay-coupled systems are not time-invertible (the inverse system would be acausal) and the semi-flow that describes their temporal evolution operates on a state space of functions. States characterized by functions in this manner can be thought of as overcountably infinite-dimensional vectors and, consequently, allow for arbitrary forms of complexity in the system. This situation is found abundantly in natural systems, which often consist of spatially distributed interconnected subsystems where delayed interactions are the rule. The brain again represents a particular example. It is therefore important to understand the effect of delays in dynamical systems, both, with regard to reconstructibility from measurements as well as with regard to information processing in general. As will be shown, accounting for delays also creates further opportunities for inference in time series analysis, for example in the context of detecting causal interactions.

This thesis documents work in nonlinear time series analysis, informed by dynamical systems theory, and with particular regard to the role of delays in interactions. It is guided by the assumption that more interesting phenomena and dependencies encountered in data are the result of complex dynamics in the underlying systems. Studying complex systems on a theoretical level, in particular the effects of interactions and information exchange due to coupling, may therefore yield important insight for data analysis. A prominent example is the host of synchronization phenomena that have been investigated since the early 1990s in coupled chaotic systems. Chaotic systems are particularly interesting in this context because they can be low-dimensional enough to allow analytical understanding, yet portrait a level of complexity that causes non-trivial dynamic features. With respect to applications in neuroscience, chaotic systems also often exhibit nonlinear oscillatory behavior that is similar in appearance to measurements from neural systems and therefore provide reasonable test data in the development of new methodology.

The two main tasks that have been considered here are *prediction* of time series, as well as *detection and characterization of dependencies*. Due to the high level of uncertainty in time series data, these tasks have to be treated within a proper *theory of statistical inference* to assure that reasoning under uncertainty is consistent. A discussion of this particular subject is given in Chapter 2. In this context, the statistical model always formalizes uncertainty pertaining to a particular functional dependency for which a parametric form has to be chosen. Two types of models are considered. The model that is considered canonically is referred to as *discrete Volterra series operator* and an illustrative derivation is discussed in the appendix of Chapter 8. The second model is in itself a complex system, a so-called delay-coupled reservoir, and has been studied for its own sake as part of this thesis. An introduction to the topic will be given in Chapter 5. Delay-coupled reservoirs afford the investigation of delays in information-processing. Moreover, they can be implemented fully optically and electronically, which holds a large potential for automated hardware realizations of statistical inference in nonlinear time series problems.

Prediction will be considered in two different scenarios. The classical scenario pertains to prediction in an autoregressive model. The existence of such a functional dependence is discussed in Chapter 3 and amounts to estimating the flow of the underlying system. In Chapter 5, exemplary data from a far-infrared laser operating in a chaotic regime is considered. The corresponding time series feature non-stationarities in mean and variance, as well as catastrophic jumps, which are usually considered at a purely stochastic level. Such features pose severe problems for many classical stochastic analysis methods but, as will be demonstrated, often can be absorbed and accounted for already by the nonlinear deterministic dynamics of the underlying systems. The second scenario considered for prediction arises in the context of *generalized synchronization* [Rulkov et al., 1995]. In cases where the coupling between two subsystems causes their dynamics to collapse onto a common synchronization manifold, the latter is reconstructible from measurements of both subsystems. By definition, one subsystem is then fully predictable given knowledge of the other. The corresponding functional relationship can be estimated in a statistical model. Examples have been studied in Chapter 7.

Detection and characterization of dependencies was studied in the context of generalized synchronization, as well as in situations characterized by interactions of spatially distributed lower-dimensional subsystems, weakly coupled to a high-dimensional global system. A particular application is found in neuroscience, where local field potential measurements are of this type. At the core of this thesis stood the development of a method, documented in Chapter 8, which estimates delay, as well as direction of information flow here. In this regard, a causal dependency between two time series is understood to represent directed information flow between the underlying dynamical systems as the result of their directional coupling. Causal interactions of this type can be measured in terms of reconstructibility of the time series. The latter is a result of certain functional dependencies the existence of which can be derived by embedding theory.

Chapter 4 will provide a more detailed outline of the different studies that have been conducted and highlight their relationship in the context of the framework described here. Beforehand, Chapter 2 discusses the methodological approach to uncertainty and statistical inference that is adopted throughout this work, and Chapter 3 provides a summary and discussion of selected topics from differential topology that will bridge the gap between dynamical systems theory and data analysis.

2 STATISTICAL INFERENCE

Normatively speaking, a theory of statistical inference has the purpose of formalizing *reasoning under uncertainty* in a consistent framework. Such a theory represents the fundamental basis of all natural scientific inference which is always based on a set of measurements and thus subject to uncertainty. First, there is *epistemic uncertainty*, pertaining to finite measurement accuracy, as well as to a finite number of samples from which the scientist has to generalize. Scientific inference is therefore often *inductive* in nature. In addition, there is the notion of *aleatoric uncertainty* pertaining to unknowns, included in measurement, that differ each time the measurements are taken in the same experimental situation. Most methdos that deal with uncertainty employ probability theory. The latter provides an important axiomatic calculus but does not address the issue of formalizing uncertainty nor the consistency of inference. Probability theory therefore does not characterize a theory of statistical inference. Indeed, it is not at all clear how probability theory and statistics are related in the first place. In the remainder of this chapter, I will attempt to outline this relationship.

A schism exists among modern practitioners and theoreticians alike which, on the surface, appears to arise from a difference in the philosophical interpretation of what a probability represents. On the one hand, there is the *frequentist* perspective which maintains a strictly aleatoric approach to uncertainty: Probabilities are the limiting ratios of long-run sampling frequencies. As a result, parameters in inference problems are not random variates that can be associated with probabilities. On the other hand, there is the *subjective* perspective which views probabilities directly as a measure of *subjective uncertainty* in some quantity of interest, including parameters. The statistical theory which arises from the subjective perspective usually employs Bayes rule as basic inference mechanism and is therefore referred to as the *Bayesian paradigm*. The subjective view on statistical inference, however, should historically rather be attributed to Ramsey, de Finetti and Savage. In a complementary fashion, Daniel Bernoulli's, Laplace's or Jeffreys' work stress the inductive nature of statistics [Stigler, 1986].

The two approaches differ most obviously with regard to estimating unknown parameters of interest in a statistical model. While the subjective approach allows formalizing epistemic uncertainty directly by means of a probability distribution on the parameter space, the frequentist approach has to employ additional concepts both, for estimating parameters, as well as characterizing the variability of these estimates. Uncertainty in such point estimates is treated by means of frequentist *confidence intervals* [Pawitan, 2013]. In this conceptualization, uncertainty pertains to the interval boundaries and not to the parameter itself. In addition, the conceptual view on uncertainty in these boundaries is concerned with their long-run sampling behavior which is purely hypothetical.

This example already alludes to the fact that the aforementioned schism in statistical practice goes much deeper than mere differences in philosophical interpretation. Methodology developed from a frequentist perspective often fails to address the problem of reasoning under uncertainty at a foundational level. Examples include Fisher's framework of likelihood-based inference for parameter estimation (see Pawitan [2013]; Fisher et al. [1990]) and the framework of decision rules for hypothesis testing developed by Neyman and Pearson [1933]. The former gave rise to the more general framework of *generalized linear models* (GLM) [McCullagh and Nelder, 2000] which is found in standard toolboxes of many areas, including genetics. In time series analysis, it is a common design to have the data described in terms of the probability theory of stochastic processes, paired with a simple device of calculus such as *least squares* [Kreiß and Neuhaus, 2006] to determine model parameters. Such an approach is also representative for objective function based "learning procedures" in some branches of machine learning, such as neural network models. The aforementioned methods are all based on "ad hoc" ideas, as Lindley calls them [Lindley, 1990], that are not representative for a consistent theoretical framework. In particular, with regard to the foundations of statistical inference these approaches are rudimentary and neglect a growing body of theory that has been trying to remedy this situation since the early 1930s.

Generally speaking, the lack of a theoretical foundation gives rise to inconsistencies. The examples of inconsistencies in ad hoc methods are numerous and of great variety, I therefore refer to the extensive body of literature and discussions on this topic elsewhere (see e.g. Lindley [1972]; Jaynes and Bretthorst [2003] or Berger [1985]). The question that frequentists leave in principle unanswered is by what theory of reference one can ultimately judge "goodness" of a statistical model in a mathematically rigorous and consistent fashion. I am of the strong opinion that only the Bayesian paradigm truly recognizes the problem of statistical inference and attempts to formalize it in a consistent axiomatic theory. At the same time, it is a deplorable fact that these attempts do not yet actually form a coherent body of theory that could in good conscience be called *a complete theory of statistics*, although they all point to the same set of operational tools. The following sections will therefore review and discuss these theoretical attempts to the extent permitted in the context of this thesis. The goal is to arrive at an operational form of the Bayesian paradigm that consolidates the work documented in Part ii.

2.1 THE BAYESIAN PARADIGM

The basic problem of inference is perhaps best described by the following statement.

> "Those of us who are concerned with our job prospects and publication lists avoid carefully the conceptually difficult problems associated with the foundations of our subject" (Lavis and Milligan [1985]; found in Lindley [1990]).

This may explain why after roughly a century of dedicated research in this area a "unified theory of statistical inference" has yet to emerge in a single complete treatise. In the remainder of this chapter, the operational form of the Bayesian paradigm will be established, accompanied by a brief discussion of its theoretical foundations and axiomatization. While the operational form is more or less unanimously agreed upon, the axiomatic foundations are numerous and of great variety. Savage's axiom system will be discussed here in an exemplary fashion since it is most well-known and has the broadest scope of application. Along the way, open questions and discordances related to this approach will be highlighted.

In a first step, the inference problem, as opposed to the decision problem, is discussed. Inference could denote here induction, the transition from past observed data to statements about unobserved future data, or abduction, the transition from observed data to an explanatory hypothesis. In this context, so-called *dutch book arguments* will be invoked to show that in order to avoid a particular type of inconsistency, uncertainty has to be summarized by a probability measure. As a consequence, Bayes formula obtains as the sole allowed manipulation during the inference step, which is thus carried out completely within the calculus of probability. In a second step, the more general decision problem is considered. It is a natural extension of the inference problem. For example, in light of certain observed data, the scientist has to decide whether to accept or refute a particular hypothesis. Decision theory has large expressive power in terms of formalizing problems and qualifies therefore as a framework for a theory of statistical inference. Moreover, it affords an axiomatization that yields consistent inference via a preference relation on the space of possible decisions. The axiom system implies the existence of a unique probability measure that formalizes subjective uncertainty pertaining to the decision problem and yields a numerical representation of preference in terms of *expected utility*. In combination with the dutch book arguments, Savage's theory of expected utility yields the operational form of the Bayesian paradigm and consolidates the statistical methods employed in this thesis.

2.1.1 *The Inference Problem*

In light of the previous discussion, many authors liken reasoning under uncertainty to a form of *weak logic* (see Jeffreys [1998]; Jaynes and Bretthorst [2003]; Lindley [1990]). A scientist is charged with the task of obtaining a general result from a finite set of data and to quantify the degree of uncertainty pertaining to this result. In general, one is interested in statements like "given B, A becomes more plausible", together with an arithmetization of this plausibility.

We will, for the moment, assume the notion of probability quantifying our degree of uncertainty in some unknown parameter or unobserved data $\theta \in \Theta$. Observed data is denoted by $x \in \mathcal{X}$, and corresponding random variables will be denoted by T and X with values in Θ and X respectively. X and Θ are assumed to be Borel spaces, in particular instances of $\mathbb{R}^d$. The Bayesian solution to formalizing statements like the one above is by using conditional probability distributions, e.g. $\mathbf{P}(T \in A|X = x)$ to express the plausibility of $A \subset \Theta$ given data $x \in \mathcal{X}$. If there is no danger of confusion, big letters will label distributions with symbolic arguments for reference in a less formal manner, such as $P(\theta|x)$ or $P(T|x)$. In the remainder, we are mainly interested in continuous probability distributions where a *posterior distribution* $\mathbf{P}(T \in A|X = x)$ is defined as a *regular conditional distribution* by the relationship

$$\mathbf{P}(T \in A, X \in B) = \int_B \mathbf{P}(T \in A|X = x) d\mathbf{P}_X(x), \quad B \subset \mathcal{X}, \tag{1}$$

in terms of the marginal distribution $\mathbf{P}_X$ for almost all $x \in X$. As shown below, the posterior is unique if continuity in x applies to the stochastic kernel $\mathbf{P}(T \in A|X = x)$ in this situation. For details, see Klenke [2007]. If $p(\theta, x)$ denotes the density of the joint distribution, the marginal distribution is given by

$$\mathbf{P}_X(X \in B) = \int_B \underbrace{\left(\int p(\theta, x) d\lambda(\theta)\right)}_{:=p(x)} d\lambda(x), \tag{2}$$

where λ denotes the Lebesgue measure. Now extend equation 1 by

$$\begin{aligned} \mathbf{P}(T \in A, X \in B) &= \int_{A\times B} p(\theta, x) d\lambda(\theta, x) \\ &= \int_B \int_A p(\theta, x) d\lambda(\theta) d\lambda(x) \\ &= \int_B \int_A p(\theta, x) d\lambda(\theta) p(x)^{-1} d\mathbf{P}_X(x) \\ &= \int_B \underbrace{\int_A p(\theta|x) d\lambda(\theta)}_{=\mathbf{P}(T\in A|X=x)} p(x) d\lambda(x) \end{aligned} \tag{3}$$

where $p(x)^{-1}$ corresponds in the third equality to the *Radon-Nikodym density* of the Lebesgue measure with respect to P_X. We have thus defined the conditional density

$$p(\theta|x) := \frac{p(\theta, x)}{p(x)}.$$

Likewise, $p(x|\theta)$ can be obtained. If the densities are continuous, it follows that

$$\begin{aligned} p(\theta|x)p(x) &= p(x|\theta)p(\theta) \\ p(\theta|x) &= \frac{p(x|\theta)p(\theta)}{p(x)}. \end{aligned} \tag{4}$$

This is Bayes formula for densities which explicitly formalizes inductive inference: The density $p(x|\theta)$ of the sampling distribution $P_\theta(x)$ (*likelihood*) incorporates knowledge of the observed data, while $p(\theta)$ represents our *prior* state of knowledge regarding θ. The posterior density $p(\theta|x)$ combines prior information with information contained in data x, hereby realizing the inference step from prior to posterior state of knowledge. The belief in θ is updated in light of new evidence. Thus, if probability theory obtains as formalization of uncertainty, all manipulations pertaining to inference can be carried out completely within the calculus of probability theory and are therefore consistent, while integration of information is always coherent. This chapter will continue to explore in how far probability theory and Bayes formula obtain from axiom systems that try to characterize consistency in a rational, idealized person's attitude towards uncertainty.

2.1.2 *The Dutch Book Approach to Consistency*

The first approach to consistent statistical inference that will be discussed in this chapter is based on so-called *dutch book arguments* and is essentially attributed to

De Finetti [1937]. Both, de Finetti and Ramsey [1931], argued for a subjectivistic interpretation of probabilities as epistemic *degrees of belief*. They discussed their ideas in the context of gambling or betting games since these present natural environments for *eliciting* subjective beliefs in the outcome of a set of events. Intuitively speaking, person B's subjective degree of belief in an event A is associated with the amount B is willing to bet on the outcome A in light of the *odds* B assigns to A obtaining. For a bookkeeper who accepts bets on the events $A \subset \Omega$ from many different gamblers it is most important to assign odds *consistently* across events to avoid the possibility of incurring *certain loss*. That is, the odds have to *cohere* in such a way that no gambler can place bets that assure profit irrespective of the actual outcomes. The latter is called a *dutch book*. To prevent a dutch book, the assignment of odds have to make the game fair in the sense that the *expected payoff* is always 0. In turn, odds are associated with prior degrees of belief since this peculiar setup was chosen to ensure truthful elicitation of beliefs. Thus, if odds are coherent and reflect beliefs, the set of beliefs will be in this sense consistent.

Although the ideas of gambling and games of chance are not very appealing as foundations of statistical inference, the dutch book approach affords a clear formalization of a particular notion of *consistency*. The latter implies a concise characterization of the mathematical form of belief assignment. In particular, the assignment of beliefs has to correspond to a probability measure in order to exclude the possibility of dutch book inconsistency. Freedman [2003] gives a very elegant modern formulation of de Finetti's result as follows.

Let Ω be a finite set with $\text{card}(\Omega) > 1$. On every proper $A \subset \Omega$, a bookkeeper assigns finite, positive odds λ_A. A gambler having bet stakes $b_A \in \mathbb{R}$, $|b_A| < \infty$, on A wins b_A/λ_A if A occurs and $-b_A$ otherwise. The net payoff for A is given by

$$\phi_A = \mathbf{1}_A \frac{b_A}{\lambda_A} - (1 - \mathbf{1}_A)\, b_A, \tag{5}$$

where $\mathbf{1}_A$ denotes the indicator function. Corresponding to each set of stakes $\{b_A | A \subset \Omega\}$ there is a payoff function,

$$\phi = \sum_{A \subset \Omega} \phi_A. \tag{6}$$

For fixed odds, each gambler generates such a payoff function. A bookkeeper is called a *Bayesian* with prior beliefs π if π is a *probability measure* on Ω that reflects the betting quotient

$$\pi(A) = \frac{\lambda_A}{1 + \lambda_A}. \tag{7}$$

Consequently, $\lambda_A = \pi(A)/(1 - \pi(A))$. In this case, all possible payoff functions have expectation 0 relative to the prior:

$$\sum_{\omega \in \Omega} \pi(\omega)\phi(\omega) = 0. \tag{8}$$

Freedman proves in particular the following equivalences:

- The bookie is a Bayesian $\Leftrightarrow$ Dutch book cannot be made against the bookie

- The bookie is not a Bayesian $\Leftrightarrow$ Dutch book can be made against the bookie

In other words, consistency implies that uncertainty in an event has to be formalized by a probability measure.

Freedman and Purves [1969] have extended de Finetti's result to the principal situation in statistical inference in the following way. They considered a finite set of parametric models $\{P(\bullet|\theta)|\theta \in \Theta\}$ specifying probability distributions on a finite set $\mathcal{X}$. In addition, $Q(\bullet|x)$ defines an *estimating probability* on Θ for each $x \in \mathcal{X}$. After seeing an observation x drawn from $Q(\bullet|\theta)$, the bookkeeper has to post odds on subsets $C_i \subset \Theta$ with $i = 1, ..., k$, on the basis of his uncertainty assigned by $Q(\bullet|x)$. A gambler is now allowed to bet $b_i(x)Q(C_i|x)$ for any bounded $b_i(x)$, and wins $b_i(x)$ if $\theta \in C_i$ obtains. The net payoff is now given by the function

$$\begin{aligned} &\phi : \Theta \times \mathcal{X} \to \mathbb{R}, \\ &\phi(\theta, x) = \sum_{i=1}^{k} b_i(x) \left(\mathbf{1}_{C_i}(\theta) - Q(C_i|x)\right). \end{aligned} \tag{9}$$

Accordingly, the *expected payoff* is defined as a function of θ by

$$\mathbb{E}_\theta[\phi] = \sum_{x \in \mathcal{X}} \phi(\theta, x) P(x|\theta). \tag{10}$$

A dutch book can be made against the estimating probability $Q(\bullet|x)$ if $\exists \epsilon > 0 : \forall \theta \in \Theta : \mathbb{E}_\theta[\phi] > \epsilon$. That is, there is a gambling system with uniformly positive expected payoff causing certain loss for the bookkeeper which thus defines an *incoherent* assignment of odds or beliefs. For a probability distribution π on Θ, the bookkeeper is once more called a *Bayesian* with prior π if

$$Q(\theta|x) = \frac{P(x|\theta)\pi(\theta)}{\sum_{\theta \in \Theta} P(x|\theta)\pi(\theta)}. \tag{11}$$

Freedman and Purves [1969] have shown that for the Bayesian bookkeeper with prior π, the set of expected payoff functions (as functions of θ given x) have expectation 0 relative to π. As a result, dutch book cannot be made against a Bayesian bookie, and the same two equivalences hold as before.

Freedman [2003] and Williamson [1999] show, in addition, that for infinite Ω and if events A generate a σ-algebra $\mathcal{A}$, prior π has to be a countably additive measure. In particular, Williamson argues against de Finetti's rejections of countable additivity by demonstrating that coherency of odds only obtains on σ-algebras of events if the measure assigning degrees of belief is countably additive. However, the subject of countable versus finite additivity is a source of much controversy (see also Schervish et al. [2008]).

As a final consideration in this section, a formulation of inconsistency is provided explicitly for the *prediction problem* on infinite spaces that is also at the core of all time series analysis considered in this thesis. I follow the formulation of Eaton [2008] who extended Stone's concept of *strong inconsistency* [Stone, 1976]. The prediction problem was originally stated already by Laplace [Stigler, 1986] and pertains to the situation where a random variable Y with values in $\mathcal{Y}$ is to be predicted from observations X with values in $\mathcal{X}$, on the basis of a joint parametric probability model with distribution $P(X, Y|\theta)$ and $\theta \in \Theta$. As before, Θ is a parameter space and θ unknown. Of interest is now the *predictive distribution* $Q(Y|x)$

which summarizes uncertainty in Y given $X = x$. In a Bayesian inference scheme, predictive distributions often arise as marginal distributions where θ is integrated out after assuming a prior distribution $\pi(\theta)$.

Eaton defines a predictive distribution $Q(Y|x)$ to be *strongly inconsistent with the model* $\{P(X, Y|\theta) : \theta \in \Theta\}$ if there exists a measurable function $f(x, y)$ with values in $[-1, 1]$ and an $\epsilon > 0$ such that

$$\sup_x \int_{\mathcal{Y}} f(x,y)dQ(y|x) + \epsilon \leq \inf_\theta \int_{\mathcal{X}\times\mathcal{Y}} f(x,y)dP(x,y|\theta). \tag{12}$$

The intuition is, as before, that when inequality 12 holds, irrespective of the distribution of X, choose e.g. $m(X)$ arbitrarily,

$$\forall\theta \in \Theta : \quad \int_{\mathcal{X}} \int_{\mathcal{Y}} f(x,y)dQ(y|x)dm(x) + \epsilon \leq \int_{\mathcal{X}\times\mathcal{Y}} f(x,y)dP(x,y|\theta). \tag{13}$$

This means that under all models for (X, Y) consistent with $Q(Y|x)$, the expectation of f is at least ϵ less than any expectation of f under the assumed joint probability model and therefore strongly inconsistent with the assumption. Stone [1976] and Eaton [2008] show that strong inconsistencies can arise as a consequence of using *improper* prior distributions in Bayesian inference schemes, where “improper” means the measure does not satisfy countable additivity. This is therefore a second argument that suggests countable additivity as a necessity for consistency.

As Eaton points out, for the prediction problem, strong inconsistency is equivalent to incoherence, as discussed in the preceding situation of Freedman and Purves. Accordingly, problem statement 9 can be modified for the predictive problem as follows. Let $C \subset \mathcal{X} \times \mathcal{Y}$, and $C_x := \{y|(x, y) \in C\} \subset Y$. An *inferrer* (the bookie before) uses $Q(Y|x)$ as a predictive distribution, given observed $X = x$. As a result, the function

$$\Psi(x,y) = \mathbf{1}_C(x,y) - Q(C_x|x) \tag{14}$$

has $Q(\bullet|x)$-expectation zero:

$$\begin{aligned}\mathbb{E}_{Y|X}[\psi(x,y)] &= \int \mathbf{1}_C(x,y) - Q(C_x|x)dQ(y|x) \\ &= \int_{C_x} dQ(y|x) - Q(C_x|x)\int dQ(y|x) \\ &= Q(C_x|x) - Q(C_x|x) = 0.\end{aligned} \tag{15}$$

Ψ denotes the former payoff function where a gambler pays $Q(C_x|x)$ dollars for the chance to win 1 dollar if $y \in C_x$ obtains. As before, in a more complicated betting scenario involving subsets $C_1, ..., C_k \in \mathcal{X} \times \mathcal{Y}$ there is a net payoff function

$$\Psi(x,y) = \sum_{i=1}^{k} b_i(x)\left(\mathbf{1}_{C_i}(x,y) - Q(C_{i,x}|x)\right) \tag{16}$$

which again has expectation zero relative to $Q(\bullet|x)$. The inferrer therefore regards the gambler’s scheme as fair. In this situation, Eaton calls the predictive distribution $Q(\bullet|x)$ incoherent if the gambler has nonetheless a uniformly positive expected gain over θ under the joint parametric probability model. That is,

$$\exists\epsilon > 0 : \forall\theta \in \Theta : \mathbb{E}_\theta[\Psi(X,Y)] = \int_{\mathcal{X}\times\mathcal{Y}} \Psi(x,y)dP(x,y|\theta) \geq \epsilon, \tag{17}$$

in which case the predictive distribution is strongly inconsistent with the model. If the inferrer is Bayesian, on the other hand, he chooses a proper prior distribution $\pi(\theta)$ and a well-defined marginal

$$m(X \in B) := \int_{\Theta} \int_{\mathcal{Y}} P(X \in B|y,\theta) dP(y|\theta) d\pi(\theta),$$

such that

$$\begin{aligned} Q(Y \in A|X) m(X \in B) &:= \int_{B} \int_{A} dQ(y|x) dm(x) \\ &= \int_{\Theta} P(X \in B, Y \in A|\theta) d\pi(\theta), \end{aligned} \tag{18}$$

for $A \subset \mathcal{Y}$, $B \subset \mathcal{X}$. Analogous to the proof for the parameter inference problem by Freedman and Purves, consider inequality 13 in expectation relative to π. Consistency now obtains from

$$\int_{\Theta} \int_{\mathcal{X}\times\mathcal{Y}} f(x,y) dP(x,y|\theta) d\pi(\theta) = \int_{\mathcal{X}} \int_{\mathcal{Y}} f(x,y) dQ(y|x) dm(x), \tag{19}$$

which is another way to say that dutch book cannot be made against a Bayesian bookie in the prediction problem. The latter is important at a conceptual level since it formalizes the inductive inference problem faced by science. That is, *learning from experience* is subject to dutch book arguments of consistency which require the inferrer to formalize uncertainty in terms of probability distributions and to carry out all manipulations pertaining to inference within the calculus of probability theory.

In summary, the dutch book approach provides a clever formal criterion of inconsistency in assigning uncertainty as subjective belief to events. As was discussed, for consistency to obtain, uncertainty must be assigned by a probability measure. The individual properties that define a probability measure can therefore be seen as "dutch book axioms of consistency" and could also be derived in a constructive fashion (see for example Jeffrey [2004]). Moreover, it is rather natural to define conditional probabilities in terms of conditional bets: You can bet on an event H conditional on D. That is, if D does not obtain, the bet on H is called off and you are refunded. In the same constructive fashion it is easily seen that for dutch book consistency (i.e. coherence) to obtain, the product rule

$$\pi(H \cap D) = \pi(H|D)\pi(D)$$

must hold, from which one can obtain now the usual definition of conditional probability. This is very appealing since the conditional probability that lies at the heart of Bayesian inference is *constructed* at a conceptual level as opposed to *defined* after probability calculus has obtained from dutch book arguments.

The dutch book arguments are built around a particular type of inconsistency applicable only in restricted contexts. In the following section, the insights gained so far will be complemented by an axiomatic foundation that affords a constructive approach to consistency irrespective of the context. To this end, the framework of decision theory will now be introduced.

2.1.3 *The Decision Problem*

Following the expositions of Lindley [1972], Berger [1985] or Kadane [2011], the operational form of the Bayesian paradigm will be characterized in terms of *decision problems*. Decision theory is often associated rather with economic theory than statistics and some authors, such as Jaynes and Bretthorst [2003] and MacKay [2003], prefer to disentangle the inference problem from the decision problem. However, every inference problem can in principle be stated as a decision problem. The latter also provides a natural description in the context of scientific inference where a practitioner has to decide whether to accept or refute a hypothesis in light of the evidence given by the data. In addition, decision theory introduces a *loss function* that can be thought of as an external *objective* on which decisions of the inferrer are based. As a result, the expressive power of decision theory in terms of problem statements can in principle account for a diversity of theoretical branches that are usually formulated independently, such as *robust estimation theory* [Wilcox, 2012], *regularization* or other objective function based methods sometimes referred to as *unsupervised learning*. It also allows to interpret frequentist approaches such as maximum likelihood estimation and the notion of minimum variance unbiased estimators, with surprising insights (see [Jaynes and Bretthorst, 2003, ch. 13] for a discussion). Moreover, in estimation problems it clearly disentangles loss functions, such as mean-squared or absolute error functions, from the statistical inference step and the choice of sampling distributions which are often subject to confusion.

We will now establish the basic concepts of decision theory and turn to the axiomatic foundations of statistical inference. The foundations of decision theory as described in this section are attributed to Savage [1954] who developed his theory from a standpoint of subjective probabilities, and who contributed major results regarding the axiomatization of statistical inference. A similar theory was developed independently by Ramsey [1931]. Savage references and contrasts the work of Wald [1949, 1945] who developed decision theory from a frequentist point of view. This section is structured likewise to derive the operational form of decision theory, as found in modern text books. The subsequent section will then concentrate on the axiomatic foundations with focus on Savage's contributions.

Decision theory assumes that there are unknowns $\theta \in \Theta$ in the world that are subject to uncertainty. These may be parameters in question but also data $x \in X$. They form the *universal set* S that contains "*all states of the (small) world*" under consideration. In general, $S = \Theta \times X$. The carriers of uncertainty are now subsets of S, called *events*, that belong to a set of sets $\mathcal{S}$, usually assumed to be a σ-algebra. In addition, there is a set of *acts* (also referred to as *decisions* or *actions*) $a \in \mathcal{A}$ which are under evaluation. Savage introduces acts as functions of states. Having decided on an action a while obtaining $s \in S$ yields a consequence $c = a(s)$. In most modern literature actions are not functions but other abstract entities that form consequences as tuples $c = (a, s)$. In Wald's frequentist statement of the decision problem, $S = X$, since the frequentist notion of uncertainty does not extend to parameters. A consequence with respect to an action is assigned a real-valued *loss* $L(\theta, a)$ in case action a is taken. The loss function as such is an arbitrary element in Wald's theory, its existence is simply assumed. It determines a penalty for a

when θ is the *true state of the world* under consideration. In contrast, the work of Ramsey and Savage regarding the axiomatic foundations supply conditions for the existence of a loss function via *utilities*, as will be discussed in the next section. The decision maker chooses an action $a = d(x)$, in this context called a *decision function*, which is evaluated through the expected loss or *frequentist risk*

$$R_d(\theta) = \int_X L(\theta, d(x)) dP_\theta(x), \tag{20}$$

where $P_\theta(x)$ once more denotes the parametrized sampling distribution defined on the set of sets $\mathcal{S}$. A decision function d is called *admissible* in Wald's theory if

$$\forall d' \in \mathcal{A} : \forall \theta \in \Theta : R_d(\theta) \leq R_{d'}(\theta).$$

Consequently, the frequentist needs some further external principle, such as *maximum likelihood*, to get an estimator for θ.

In contrast, the Bayesian evaluates the problem in its *extensive form* (due to Raiffa and Schlaifer [1961]) where $S = X \times \Theta$ and chooses

$$\min_a L^*(a) = \min_a \int_\Theta L(\theta, a) dP(\theta|x), \tag{21}$$

where $P(\theta|x)$ denotes the posterior distribution *after* x has been observed, as defined by the likelihood $P_\theta(x) = P(x|\theta)$ and a prior distribution $\pi(\theta)$. Inference problems, as discussed before, are naturally included by identifying $\mathcal{A} = \Theta$. In the extensive form, once x is observed and included via the likelihood it is irrelevant. However, if the loss function is bounded and π a proper probability distribution, the Bayesian may equivalently minimize the average risk

$$R^*(d) = \int_\Theta R_d(\theta) d\pi(\theta). \tag{22}$$

The equivalence of (21) and (22) follows from

$$\begin{aligned} \min_d R^*(d) &= \min_d \int_\Theta \int_X L(\theta, d(x)) dP(x|\theta) d\pi(\theta) \\ &= \int_X \min_d \int_\Theta L(\theta, d(x)) dP(\theta|x)\, dP(x). \end{aligned} \tag{23}$$

This is referred to as the *normal form* and considers the decision problem *before* data x is observed. Like the Waldean risk, it employs a hypothetical sampling space X, the choice of which is in principle just as arbitrary as the choice of the prior distribution π for which the Bayesian paradigm is often criticized by frequentists. The extensive form is much simpler to evaluate, as will also be shown in section 2.1.5. Wald's most remarkable result was to show that the class of decision rules d that are *admissible* in his theory are in fact Bayesian rules resulting from normal form 22 for some prior distribution $\pi(\theta)$. However, the result is rigorous only if *improper* priors are included that don't satisfy countable additivity.

The extensive form 21 will be considered as operational realization of the Bayesian paradigm and the next section explores its axiomatic foundations.

2.1.4 *The Axiomatic Approach of Subjective Probability*

Decision theory generalizes the objective of inference by introducing additional concepts such as actions the inferrer, now decision maker, can carry out, and the relative merit or *utility* of their consequences, as measured by a loss function in the previous section. It therefore provides a rich environment for reasoning under uncertainty which many authors believe to serve as a proper theory of statistics. Every proper theory needs an axiomatic foundation and numerous attempts have been made to establish the latter. Although good review articles exist (e.g. Fishburn [1981, 1986], Lindley [1972]), the number of different axiomatic theories is rather confusing. In particular, up to this day apparently no consensus has been reached regarding a *unified theory of statistical inference*. At the same time, it is heartening to see that the various different approaches essentially share the same implications, part of which were already established by the dutch book arguments. The axiom systems differ mainly in the scope of their assumptions regarding technical aspects of primitives such as the sets S, $\mathcal{S}$ and $\mathcal{A}$. The theory of Savage [1954] is one of the oldest, most well-known and most general in scope of application, with certain appealing aspects to the formalization of its primitives, as outlined in the previous section. It will be adopted in this thesis and presented exemplarily.

As evident from [Savage, 1961], Savage was a firm advocate of the Bayesian paradigm of statistics in which Bayes theorem supplies the main inference mechanism, the latter being carried out to full extent within the calculus of probabilities. The problem he tried to solve in his foundational work [1954] was to derive probabilities in the decision-theoretic framework as subjective degrees of belief (or uncertainty). This was in strong contrast to the prevalent strictly aleatoric interpretation of probabilities as limiting ratios of relative frequencies on which Wald's decision theory was based. The shortcomings of this interpretation were already discussed at the beginning of this chapter. In Savage's own words [1961],

> "Once a frequentist position is adopted, the most important uncertainties that affect science and other domains of application of statistics can no longer be measured by probabilities. A frequentist can admit that he does not know whether whisky does more harm than good in the treatment of snake bite, but he can never, no matter how much evidence accumulates, join me in saying that it *probably* does more harm than good."

As will become clear shortly, Savage's axiomatic foundations imply a theory of *expected utility*, later to be identified with expected loss (see eq. 21), that extends results from von Neumann and Morgenstern [2007]. It also complements and generalizes the dutch book arguments from section 2.1.2. In this context, the dutch book arguments can be thought of as *minimally consistent requirements* for a theory. They approach the problem by focusing on a particular type of inconsistency, the dutch book as a failure of coherence in de Finetti's sense, and derive constraints on a formalization of subjective uncertainty. Although the implications of coherence were substantial, their scope is rather narrow and originally only focused on gambles.

Notions like gambles and lotteries also play an important role in the derivation of a theory of expected utility, however, they do not yet establish a general framework in which arbitrary decision problems can be formalized. In Savage's theory, gambles are a particular type of simple acts that are measurable and the domain of which is empty for all but a finite number of consequences. To be able to account for general types of acts, further concepts have to be introduced. Recall that an act f maps a state s to a consequence c. For $c \in \mathcal{C}$, a utility is a function $U : \mathcal{C} \rightarrow \mathbb{R}$ that assigns a numerical value to a consequence. For bounded utility U, a loss function can be defined equivalently as $L(c) := \max_{c^*} U(c^*) - U(c)$. As before, the utility assigns a reward (or penalty in terms of the loss) for the decision maker to a particular consequence.

Now recall that Ramsey and de Finetti originally considered the gambling scenario as a way to elicit a person's subjective degree of belief truthfully by placing bets on events under the threat of financial loss. This means, the subjective degrees of belief are measured only indirectly here by a person's willingness to place particular bets. That is, it is inferred by the person's *preferences* among possible gambles. Like the gambler who has to decide for a particular bet, the decision maker has to decide for a particular act. Axiomatic approaches to decision theory pick up on the notion of preference in a more general scope and use it to define relations among acts. For example, for $f, g \in \mathcal{A}$, $f \prec g$ if the inferrer generally prefers act g over f. Such a qualitative preference of one act over another is influenced both by the assignment of consequences to certain states, as well as the inferrer's belief that corresponding events will obtain. Theories of expected utility therefore are defined by axiom systems that imply a qualitative preference relation together with a particular *numerical representation* that allows for an arithmetization of the decision process. In Savage's theory,

$$f \prec g \quad \Leftrightarrow \quad \mathbb{E}_{P^*}[U(f)] < \mathbb{E}_{P^*}[U(g)], \tag{24}$$

where

$$\mathbb{E}_{P^*}[U(f)] = \int_{s \in \mathcal{S}} U(f(s))dP^*.$$

Axioms have to be found such that they imply a unique probability measure P^* and a real-valued utility function U that is unique up to affine transformations. The uniqueness of P^* is necessary to yield proper conditional probabilities. In the following, a first outline of the theory is given without resorting to the technical details of the actual axiom system.

Savage's primitives involve an uncountable set S, the power set $\mathcal{S} = 2^S$ and $\mathcal{A} = \mathcal{C}^S$. He stresses, however, that there are no technical difficulties in choosing a smaller σ-algebra $\mathcal{S}$. Richter [1975] considered the case for finite $\mathcal{C}$ and S. The theory proceeds to establish a further relation on the set of events $\mathcal{S}$, called *qualitative probability* and denoted by "$\prec^*$". For $A, B \in \mathcal{S}$, $A \prec^* B$ means A is subjectively *not more probable* than B. The qualitative probability relation is derived from preference among acts by considering special acts

$$f_A(s) = \begin{cases} c & \text{if } s \in A \\ c' & \text{else,} \end{cases} \quad , \quad f_B(s) = \begin{cases} c & \text{if } s \in B \\ c' & \text{else,} \end{cases} \tag{25}$$

and defining

$$A \prec^* B \quad \Leftrightarrow \quad f_A \prec f_B \quad \wedge \quad c' \prec c. \tag{26}$$

The preference $c' \prec c$ is a preference among acts by considering consequences as special constant acts. Intuitively speaking, since f_A and f_B yield equivalent "reward" c, the preference can only arise from the fact that the decision maker considers A less probable than B. The first arithmetization in Savage's theory is already achieved at this point by the definition of *agreement* between the qualitative probability relation and a corresponding numerical probability measure P,

$$\forall A, B \in \mathcal{S}: \quad A \prec^* B \quad \Leftrightarrow \quad P(A) < P(B). \tag{27}$$

The axiom system implies the existence of a unique measure P^* that fulfills this criterion. In a next step, Savage uses P^* to construct *lotteries* from gambles (simple acts) and derives the von Neumann-Morgenstern axioms of *linear utility*. Lotteries are defined as simple probability distributions that are nonzero only for a finite set of mutually exclusive events. For example, $p(A)$ would denote the probability that event A will occur if lottery p is played. The von Neumann-Morgenstern theory thus formalizes the betting scenario that was discussed in the context of dutch book arguments earlier, although care has to be taken not to confuse payoff and utility. The theory establishes numerical representation (24) for preferences among simple acts and is subsequently generalized to all of $\mathcal{A}$ by invoking further axioms.

Savage's result was conveniently summarized by [Fishburn, 1970, ch. 14] in a single theorem, as stated in appendix A. In light of the technical nature of the seven axioms Savage developed, this section will continue with a qualitative description only, informed by Fishburn [1981]. For details, the reader is referred to appendix A. Axiom P1 establishes that $\prec$ on $\mathcal{A}$ is *asymmetric* and *transitive*, and thus a *weak order*. Axioms P2 and P3 realize Savages *sure-thing principle* which states that the weak ordering of two acts is independent of states that have identical consequences. P4 pertains to the definition of the qualitative probability relation (26) and expresses the assumption that the ordering does not depend on the "reward" c itself. For the qualitative probability to be defined, P5 demands that at least two consequences exist that can be ordered. Axiom P6 expresses a continuity condition that establishes an important but rather technical partitioning feature of $\mathcal{S}$. It also prohibits consequences from being, in a manner of speaking, infinitely desirable.

Axioms P1–P6 ensure that the qualitative probability relation $\prec^*$ on $\mathcal{S}$ is a weak ordering, as well, and consequently allow to derive the existence of a unique probability measure P^* that agrees with $\prec^*$. P^* is non-atomic for uncountable S, that is, on each partition $B \subset S$ it takes on a continuum of values. Note that P^* is not necessarily countably additive. Savage [1954] maintained that countable additivity should not be assumed axiomatically, due to its nature of mere technical expedience which was already critically remarked upon by Kolmogorov himself [Kolmogoroff, 1973]. In particular, Savage stated that countable additivity should only be included in the list of axioms if we feel that its violation deserves to be called inconsistent. However, in light of the dutch book arguments for countable additivity advocated by Williamson [1999] and Freedman [2003], as discussed in section 2.1.2, I conclude that not to demand it leads to dutchbook incoherence in case S is infinite and $\mathcal{S}$ a corresponding σ-algebra. As reviewed by Fishburn [1986], in the context of the present axiomatization that implies the existence of a unique P^* which agrees with the qualitative probability $\prec^*$, P^* is countably additive *if and only if* $\prec^*$ is *monotonely continuous*. The latter therefore has to be accepted as eighth postulate in the list of axioms:

Definition 1 Monotone continuity
For all $A, B, A_1, A_2, ... \in \mathcal{S}$, if $A_1 \subset A_2 \subset, \dots, A = \bigcup_i A_i$ and $A_i \prec^* B$ for all i, then $A \prec^* B$.

Monotone continuity thus demands that in the limit of nondecreasing A_i converging on event A, the ordering $A_i \prec^* B$ that holds for all i cannot suddenly jump to $B \prec^* A$ and reverse in this limit. This demand is intuitively appealing because it ensures reasonable limiting behavior in the infinite, a subject which in general rather defies the human mind.

The last axiom P7 has a similar continuity effect for utilities and ensures in particular that the utility function is bounded. As such, it allows the final generalization of the numerical representation (24) to the full set of acts $\mathcal{A}$.

As a final point, it is interesting to discuss the notion of conditional probability that arises from Savage's theory. At the level of qualitative probability, he defines for $B, C, D \in \mathcal{S}$

$$B \prec^* C \text{ given } D \quad \Leftrightarrow \quad B \cap D \prec^* C \cap D,$$

and shows that if $\prec^*$ is a qualitative probability, then so is $\prec^*$ given D. Furthermore, there is exactly one probability measure $P(B|D)$ that almost agrees with $\prec^*$ as a function of B for fixed D and it can be represented by

$$P(B|D) = \frac{P(B \cap D)}{P(D)}.$$

The interpretation of the comparison among events given D is in temporal terms, that is, $P(C|D)$ is the probability a person would assign to C after having observed D. Savage stresses that it is conditional probability that gives expression in the theory of qualitative probability to the phenomenon of *learning by experience*. Some authors criticize this fact for lacking constructiveness and wish to include comparisons of the kind $A|D \prec^* C|F$. A more detailed discussion and further references can be found in [Fishburn, 1986] and comments. For the remainder of this thesis, Savage's theory combined with the dutch book arguments of coherence will be deemed sufficient to support the modern operational form of the Bayesian paradigm, as stated in section 2.1.3.

In summary, this section gave a brief introduction to a prominent axiomatic foundation that supports the decision theoretic operationalization of the Bayesian paradigm of statistical inference. The axioms ensure the existence of a weak order among the actions a decision maker can carry out, and yield a numerical representation in terms of expected utilities. The preference relation among acts implies the existence of a unique probability measure on the set of events, and the existence of a bounded utility function (and hence also a loss function) that is unique modulo affine transformations. Savage's main contribution is the subjective interpretability of this probability measure which allows one to formalize the full range of uncertainty pertaining to a decision problem. This makes it possible to realize learning from experience by Bayes theorem and to formalize situations where an inferrer has to make a decision after observing certain evidence. Furthermore, the dutch book arguments show that violation of this principle leads to inconsistency in the decision process. The decision maker thus has to choose an action that maximizes expected utility or, equivalently, minimizes expected loss. Lindley [1990] provides additional arguments and justification for using the criterion of *expected utility* in the decision process. I do not think this is necessary. It is clear that in the context of Savage's theory a functional on acts is needed to establish real values that allow comparisons and summary statistics of acts as functions. The particular functional $\mathbb{E}_{P^*}[U(f(s))]$ contains the consistent formalization of uncertainty in the decision problem via P^*, and an otherwise arbitrary degree of freedom in the nonlinear "kernel mapping" U. This lends numerical representation (24) a canonical and intuitive appeal.

2.1.5 *Decision Theory of Predictive Inference*

As a final step in this chapter, the decision problem for the analysis pertaining to time series will be formulated. In essence, the problem is always one of predictive inference, as was already introduced in the context of equation (12) and can be stated dually as parametric inference in a regression analysis (see Chapter 5), as well as directly in nonparametric form in the context of Gaussian process regression. A brief introduction to the latter is provided in the appendices of both, Chapter 5, as well as Chapter 8. The task is always to predict or, equivalently, reconstruct a target time series $y \in \mathbb{R}^N$ using covariate time series $x \in \mathbb{R}^n$ which may coincide with y. In particular, the modeling assumption is usually an extension of the following,

$$\forall i \in \{1, ..., N\}, y_i \in y : \exists x_{\mathbf{k}} = (x_k, ..., x_{k+m}) \subset x : \quad y_i = f(x_{\mathbf{k}}) + \epsilon_i, \tag{28}$$

where $\epsilon_i \sim \mathcal{N}(0, \sigma_i^2)$ and $f : \mathbb{R}^m \to \mathbb{R}$ a particular form of a model for the targeted interaction. The latter is either a Volterra series expansion, as introduced in detail in section 8.6.2, or the delay-coupled reservoir architecture derived in Chapter 5. Both admit regression with nonlinear basis functions. The a priori assumption of normality for the residuals represents, in general, the best informed choice that can be made: Residuals with respect to a "suitably true" f are never actually observable and the normal distribution fulfils the criterion of *maximizing entropy* (see [Jaynes and Bretthorst, 2003, ch. 7]). In particular, Jaynes argues that, from an information theoretic point of view, the assumption of normality can only be improved upon by knowledge of higher moments of the residual distribution.

To be able to always use the full extend of available information, the preferred approach is to compute for each target data point $y_* \in \mathbb{R}$ with covariates $x_* \in \mathbb{R}^m$ the *leave-one-out* predictive distribution $P(y_*|x_*, D_*)$, conditional on all remaining data $D_* = (y \backslash \{y_*\}, x \backslash x_*)$. Note that the predictive distribution is either computed as marginal distribution with respect to model parameters, the latter being integrated out after obtaining their posterior distribution from the observed data, or directly as a Gaussian process posterior given a prior process assumption with constant expected value 0 (see chapters 8 and 5 for details). In order for this particular inductive setup to make sense we will assume that the individual data points are *exchangeable*.

We can now state the inference problem in extensive form (21). Having observed x_* and D_*, an estimator for y_* has to be found, denoted by $\hat{y}_*$. The loss function for the corresponding decision task will always be the squared error loss $L(\hat{y}, y) = (\hat{y} - y)^2$. Punishing larger errors stronger seems prudent. For some time series analysis tasks the squared error can also be motivated by other extrinsic arguments, as given in Chapter 8. The Bayesian loss can now be stated as

$$L^*(\hat{y}_*|x_*) = \int_{\mathbf{R}} (\hat{y}_* - y_*)^2 p(y_*|x_*, D_*) dy_*. \tag{29}$$

Minimizing the expected loss yields

$$\begin{aligned} \frac{d}{dy_*} L^*(\hat{y}_*|x_*) &= 0 \\ \hat{y}_* \int_{\mathbf{R}} p(y_*|x_*, D_*) dy_* - \int_{\mathbf{R}} y_* p(y_*|x_*, D_*) dy_* &= 0 \\ \hat{y}_* = \int_{\mathbf{R}} y_* p(y_*|x_*, D_*) dy_* &= \mathbb{E}[y_*|x_*]. \end{aligned} \tag{30}$$

This result is no surprise since the inference step is concluded by calculating the predictive distribution as a summary of the inferrer's uncertainty. Its expected value presents an intuitive choice for an estimator. However, different loss functions, for example the absolute loss $|\hat{y} - y|$, would yield different estimators, in this case the *median* of the predictive distribution. For consistency it is therefore worth noting that the choice of the expected value as estimator corresponds to a choice of squared error loss. Consider also the following miscellaneous fact: The maximum likelihood estimation procedure can be interpreted in the decision theoretic framework and would correspond to the choice of a binary loss function given a constant prior on model parameters. The latter is also referred to as an *improper prior distribution* since it does not satisfy countable additivity. Invoking further external constraints such as unbiasedness of the resulting estimator represents an immediate violation of the *likelihood principle* (a direct consequence of the Bayesian paradigm, see Lindley [1972] for details) and therefore leads to inconsistent inference.

As shown in appendix B, stating the decision problem in normal form (22), as opposed to the extensive form above, yields the same result, its optimization is, however, a lot more involved and requires variational calculus, as well as additional knowledge and constraints regarding the domain of the time series. With the statement of optimization problem (30) the discussion of foundations and operational form of a theory of statistical inference is deemed sufficient to consolidate the methods employed in the remainder of this thesis. However, in light of the inhomogeneity of the literature and the sheer extent of the topic, said discussion can only be called rudimentary.

3 DYNAMICAL SYSTEMS, MEASUREMENTS AND EMBEDDINGS

I treat time series as data with auto structure that typically represent measurements from dynamical systems. As such, statistical models that realize functional mappings on the measurements can be justified theoretically by considering mappings on the underlying systems and their geometrical information. The latter may also imply the existence of functional mappings *between* time series, in case their underlying systems are coupled. These theoretical considerations can provide a rich foundation and interpretation for statistical models in time series analysis. The crucial step in practice is therefore to explicate the geometric information of the underlying systems that is implicit in the measurements. By reconstructing the underlying systems, results from dynamical systems theory become available and can inform the data analysis process. This chapter will give a brief overview of a theoretical branch from differential topology that solves this problem for the practitioner. Primary application will be with respect to prediction of individual time series, as well as detection of causal interactions between time series. The latter requires additional conceptualization and intuitions which are provided in the following section.

3.1 DIRECTED INTERACTION IN COUPLED SYSTEMS

In the context of dynamical systems theory, causality can be conceptualized as the direction of interaction between coupled systems. A system X that is coupled to another system Y influences Y's temporal evolution by injecting its own state information over time into Y's internal state. The additional information of the driver causes alterations in the driven system's state space: Y encodes information about X geometrically. To illustrate this, consider a unidirectionally coupled Rössler-Lorenz System. The Rössler driver is given by

$$\begin{aligned}
\dot{x}_1 &= -6(x_2 + x_3),\\
\dot{x}_2 &= 6(x_1 + 0.2x_2),\\
\dot{x}_3 &= 6(0.2 + x_3(x_1 - 5.7)),
\end{aligned} \tag{31}$$

while the Lorenz response system is given by

$$\begin{aligned}
\dot{y}_1 &= \sigma(y_2 - y_1),\\
\dot{y}_2 &= ry1 - y_2 - y_1 y_3 + \mu x_1,\\
\dot{y}_3 &= y_1 y_2 - b y_3 + \mu x_1,
\end{aligned} \tag{32}$$

where μx_1 denotes the interaction term by which X influences Y. With $\sigma = 10$, $r = 28$, $b = \frac{8}{3}$ and $\mu = 0$, both systems are uncoupled and feature a stable chaotic attractor in their three dimensional state spaces (upper part of figure 1). The lower part of figure 1 depicts the case where $\mu = 10$, such that the Rössler driver injects its own state information into the Lorenz system. One can see how the interaction

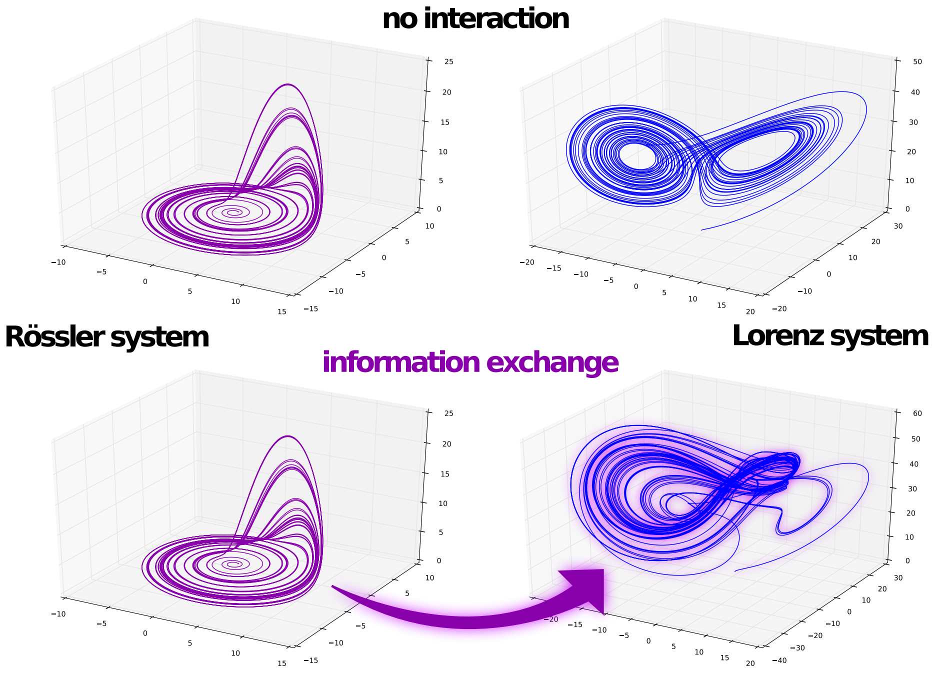

Figure 1: Rössler system driving a Lorenz system. Upper part: Uncoupled state. Lower part: A particular coupling term causes state information of the Rössler system to flow into the Lorenz system (see eq. 32). This leads to a smooth warping of the driven attractor manifold to account for the additional information injected by the driver.

causes the attractor manifold of the driven Lorenz system to smoothly warp, hereby encoding information about the Rössler driver geometrically.

When discussing causality in this context, we are thus interested in the direction of information flow between dynamical systems, realized by interaction terms in the temporal evolution of the systems. As a result of such interactions, a driven system may encode geometrically information about the driver. To determine the coupling scenario, one therefore has to quantify in how far one system is geometrically informative about another. The problem is further complicated by the fact that one usually has no direct access to the full systems and their geometry. Instead, the inference has to be carried out on time series data, which represent down-sampled, down-projected, noisy measurements of the underlying systems. Accessibility to relevant information via such measurements is provided by embedding theory, which is discussed in the following section.

3.2 EMBEDDING THEORY

Dependencies between time series may be reflections of geometrically encoded information resulting from interactions due to coupling, as discussed in the previous section. To infer the causal structure of these interactions it is thus necessary to unfold the geometric information of the systems from the measurement data. Likewise, in prediction tasks the flow of the underlying system (which may be the solution to a differential equation) has to be approximated by a functional map-

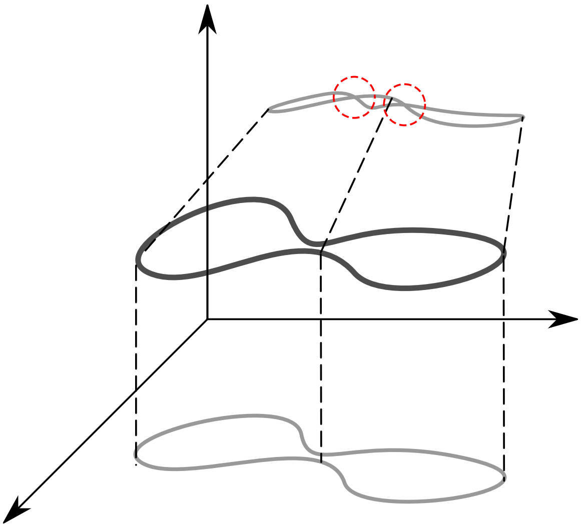

Figure 2: Embedding a one-dimensional manifold in two or three-dimensional Euclidean space. The two circles indicate intersections in the projection into two-dimensional space, which therefore fails to be an embedding.

ping that operates on the time series directly. These ideas have been formalized in a branch of differential topology which may be referred to as embedding theory. The groundwork was supplied by Whitney [1936], who showed that the definition of an abstract manifold by some intrinsic coordinate system is equivalent to an extrinsic definition as a submanifold in higher dimensional Euclidean space. Consider the example depicted in figure 2: The “rubber-band” manifold is intrinsically one-dimensional. Its projections (*measurements*) into the coordinates of the surrounding two or three-dimensional Euclidean space are called an embedding if the resulting map is bijective and preserves the manifold’s differential structure. The upper two-dimensional projection is not bijective due to the intersections, while the lower one is. It is intuitively reasonable that “almost all” projections into three-dimensional space will yield proper embeddings without intersections. In general, any continuous mapping from a smooth m-dimensional manifold into $\mathbb{R}^d$ can be approximated by a proper embedding if $d > 2m$.

Data acquisition in natural science can be compared to the coordinate projections into Euclidean space in the example above. Let a measurement be a real-valued function $\phi : M \to \mathbb{R}$, the domain of which is a manifold M. If the phenomenon of interest to the scientist is a dynamical system, the state space in which it evolves temporally may be comprised by such a manifold M. Examples of such manifolds are already given in figure 1, which shows chaotic attractor manifolds embedded in three-dimensional Euclidean space.

The problem differential topology solves for the practitioner is that of reconstructing a system that is observed only indirectly via real-valued measurements. Consider, for example, local field potentials (LFPs) from electrode recordings in cortex. These yield a time series measurement of the unobserved neuronal network activity contributing to the LFPs. Aeyels [1981] was one of the first to work on this topic and provides the most intuitive access. He considered time-continuous dynamical systems given by vector fields defined on a differentiable manifold M with m dimensions. Each vector field admits a flow $f : M \times \mathbb{R} \to M$ which describes the temporal evolution of a dynamical system by mapping some initial state $x_0 \in M$ forward in time by a factor of t to the state $x(t)$. Thus, f defines a temporal evolution of the dynamical system and corresponding trajectories on M. *Measurements*, such as LFPs, are defined as continuous functions $\phi : M \to \mathbb{R}$. As a function of time, the system f is observed only indirectly via the measurements $\phi(f(x,t))$ which constitute the observed time series. Suppose the measurements were sampled at a set of d points $t_i \in [0, T]$ along an interval of length T. This set is called a *sample program* $\mathcal{P}$.

Definition 2 *A system (f, ϕ) is called $\mathcal{P}$-observable if for each pair $x, y \in M$ with $x \neq y$ there is a $t_i \in \mathcal{P}$, such that $\phi(f(x,t_i)) \neq \phi(f(y,t_i))$.*

In other words, if a system is observable, the mapping of an initial condition x into the set of measurements defined by $\mathcal{P}$,

$$\mathrm{Rec}_d(x) = (\phi(x), \phi(f(x,t_1)), ..., \phi(f(x,t_{d-1}))$$

is bijective. $\mathrm{Rec}_d : M \to \mathbb{R}^d$ is called a *reconstruction map*. If $x \neq y$, $\mathrm{Rec}_d(x)$ and $\mathrm{Rec}_d(y)$ differ in at least one coordinate, hereby allowing one to distinguish between x and y in measurement. Aeyels showed that, given an almost arbitrary vector field, it is a *generic* property of measurement functions ϕ that the associated reconstruction map Rec_d is bijective if $d > 2m$. Genericness is defined here in terms of topological concepts (*open and dense subsets* of function spaces) as provided in theorem 1. As a result, the temporal evolution of f on M becomes accessible via the reconstruction vectors corresponding in time.

For purposes of statistical modeling, this level of description is quite sufficient. In general, however, it is natural to also demand differentiability of Rec_d such that its image is a submanifold in $\mathbb{R}^d$. In this case, the reconstruction map is called an *embedding* and also preserves the smoothness properties of M. In turn, an embedding affords the investigation of topological invariants and further properties of the dynamical system in measurement. Takens [1981] showed in a contemporaneous piece of work that Rec_d is generically an embedding if $d > 2m$, together with stronger statements of genericness. To this end, he considered diffeomorphisms $F : M \to M$, which may be given by $F := f(\bullet, \Delta t)$, and showed that the reconstruction map

$$\begin{aligned} &\Phi_{F,\phi} : M \to \mathbb{R}^d, \\ &\Phi_{F,\phi}(x) = (\phi(x), \phi(F(x)), ..., \phi(F^{d-1}(x)))^T \end{aligned} \tag{33}$$

is an embedding for *generic* F and ϕ if $d > 2m$. In this case, the reconstruction map is also called a *delay embedding*. Here, F^d denotes the d-fold composition $F \circ F \circ ... \circ F$ of functions. More formally, following [Stark, 1999], denote by $\mathcal{C}^r(M, \mathbb{R})$ the space of all r times differentiable real-valued functions on M, and by $\mathcal{D}^r(M)$ the space of all r times differentiable diffeomorphisms on M. The following theorem now holds.

Theorem 1 Takens 1980
Let M be a compact m-dimensional manifold on which a smooth (r times differentiable) diffeomorphism $F \in \mathcal{D}^r(M)$ is defined, and $\phi : M \to \mathbb{R} \in \mathcal{C}^r(M, \mathbb{R})$ a smooth real-valued measurement function. Then if $d > 2m$, the set of (F, ϕ) for

which the map $\Phi_{F,\phi}$ is an embedding is open and dense in $\mathcal{D}^r(M) \times \mathcal{C}^r(M, \mathbb{R})$ for $r \geq 1$.

Note that for diffeomorphisms this includes $F := f(\bullet, -\Delta t)$. Sauer et al. [1991] extended this result in several ways. First, by a new concept called *prevalence* the genericity of the embedding theorem was extended in a measure-theoretic sense. Second, it was remarked that the theorem holds even if the delay embedding map is composed of different measurement functions, similar to Whitney's original embedding theory. A formal proof was given by Deyle and Sugihara [2011]. Furthermore, it was proven that an application of linear time invariant filters on the measurements preserves the embedding. The latter is quite important since for example electrode-recordings in neuroscience are automatically filtered in most hardware setups. Extending the neuroscience example, one may also combine recordings from different electrodes to yield an embedding if they contain overlapping measurements, which creates interesting opportunities for multi-channel recordings even on short time intervals and in the presence of strong background noise. Finally, it was shown that the embedding dimension d may be much smaller than the dimension of M, if, for example, the dynamical system is restricted to an attractor submanifold with low box-counting dimension.

Furthermore, in 2002, Takens proved an important generalization regarding the dynamical system from which measurements are taken [Takens, 2002]. The generalization pertains to weaker assumptions about the temporal evolution of a dynamical system. In the previous theorem, the latter was provided by a diffeomorphism F, which is time-invertible. If F is not invertible, it is called an *endomorphism* and we denote by $\text{End}^1(M)$ the space of all continuous endomorphisms on M. The weak extension to endomorphisms is important, for example, when dealing with retarded systems that feature delayed interaction terms, such as the Mackey-Glass system [Glass and Mackey, 2010]. This is owed to the fact that the temporal evolution of a retarded system is described only by a semi-flow which is not invertible in time (the inverse would be acausal).

Theorem 2 Takens 2002
Let $F : M \to M$ be an endomorphism under the conditions of theorem 1 and $X_d \subset \mathbb{R}^n$ be the image of $\Phi_{F,\phi}$. Then there is an open and dense subset $U \subset$ End$^1(M) \times C^1(M)$, where End$^1(M)$ denotes the space of C^1-endomorphisms on M, such that, whenever $(F, \phi) \in U$ and $d > 2m$, there exists a unique map $\pi_d : X_d \to M$ with $\pi_d \circ \Phi_{F,\phi} = F^{d-1}$, which is differentiable in a neighborhood of X_d. As a result, a sequence of d successive measurements from a system determines the system state at the end of the sequence of measurements.

For both, endomorphisms and diffeomorphisms, the delay embedding $\Phi_{F,\phi}$ constitutes a dynamical system since there exists a smooth function H, such that

$$H \circ \Phi_{F,\phi} = \Phi_{F,\phi} \circ F. \tag{34}$$

H therefore describes a temporal evolution on the state space X_d and commutes with F, thus defining a conjugacy under the coordinate change $\Phi_{F,\phi}$ (see figure 2).

In order to approach the problem of causal interaction structures between coupled systems, the theory needs further refinement. One would want to treat the

$$\begin{tikzcd}
x_i \arrow[r, "F"] \arrow[d, "\Phi_{F,\phi}"'] & x_{i+1} \\
z_i \arrow[r, "H"] & z_{i+1} \arrow[u, "\Phi^{-1}_{F,\phi}"']
\end{tikzcd}$$

Figure 3: Equivalence of the dynamical system given by a diffeomorphism F and its delay embedded counterpart H under coordinate change Φ. F is defined on the manifold M, with states $x_j \in M$. Conversely, $z_j \in \Phi_{F,\phi}(M) \subset \mathbb{R}^d$ denotes the states of the delay embedding. H is thus defined through $H = \Phi_{F,\phi} \circ F \circ \Phi^{-1}_{F,\phi}$. For endomorphisms the situation is slightly more complicated.

case where different time series represent non-overlapping measurements of different dynamical systems which putatively exchange information via some specific coupling. Although the full system of coupled subsystems may be viewed as autonomous, the subsystems from which measurements are taken must be treated as non-autonomous, since they receive dynamic input from other subsystems.

This problem was tackled by Stark [1999], who formalized this situation using skew product systems. Let M and N be m- and n-dimensional manifolds, and $G : N \to N$ a diffeomorphism on N that represents the evolution of a dynamical system with temporally indexed state $y_{i+1} = G(y_i) \in N$. Furthermore, let $F : M \times N \to M$ be a diffeomorphism, such that F_y is defined on M *for every* $y \in N$ by $F_y(x) = F(x, y)$ with $x \in M$. This leads to a skew product system on $M \times N$ of the form

$$\begin{aligned} x_{i+1} &= F(x_i, y_i) \\ y_{i+1} &= G(y_i), \end{aligned} \tag{35}$$

where y represents the driver and x the non-autonomous driven system. According to Muldoon et al. [1998], N may be embedded in M if $n < m$, which would be the most obvious correspondence with the idea of *geometric information encoding*, as described in the previous section. More generally, Stark proved that delay maps $\Phi_{F,G,\phi}$, composed of measurements from the driven system $\phi : M \to \mathbb{R}$, are generically embeddings:

Theorem 3 Forced Takens Theorem, Stark 1999
Let M and N be compact manifolds of dimension $m \geq 1$ and n, respectively. Suppose that the periodic orbits of period $< 2d$ of $G \in \mathcal{D}^r(N)$ are isolated and have distinct eigenvalues, where $d \geq 2(m+n)+1$. Then for $r \geq 1$, there exists an open and dense set of $(F, \phi) \in \mathcal{D}^r(M \times N, M) \times C^r(M, \mathbb{R})$ for which the map $\Phi_{F,G,\phi} : M \times N \to \mathbb{R}^d$ is an embedding.

Intuitively speaking, the theorem asserts that given local measurements of a driven system, these measurements allow under certain conditions the reconstruction of the full skew product system consisting of both, driver and driven system. The driver is indirectly reconstructible via its geometric encoding in the driven system. Stark et al. [2003] extended this result to stochastic driver systems. Stochastic systems are usually modeled through the use of shift spaces. Let $\mathcal{X}$ be a topological space and $\Sigma := \mathcal{X}^{\mathbb{Z}}$ the space of bi-infinite sequences $\omega = (\omega_i)_{i=-\infty}^{\infty}$ of elements in $\mathcal{X}$ with the product topology. The *shift map* $\sigma : \Sigma \to \Sigma$ is defined by $[\sigma(\omega)]_i = [\omega]_{i+1}$, where $[\omega]_i$ is the i^{th} component of $\omega \in \Sigma$. For $F \in \mathcal{D}^r(M \times \mathcal{X}, M)$, a skew product system can be defined by

$$\begin{aligned} x &\mapsto F(x, \omega_0) \\ \omega &\mapsto \sigma(\omega). \end{aligned} \tag{36}$$

In addition, assume a noisy measurement function $\phi : M \times \mathcal{X}' \to \mathbb{R}$, where $\phi_{\eta_i}(x) = \phi(x, \eta_i)$ for $\eta \in \Sigma' = (\mathcal{X}')^{\mathbb{Z}}$. Define $F_{\omega_0} = F(\bullet, \omega_0)$ and $F_{\omega_k \dots \omega_0} = F_{\omega_k} \circ \dots \circ F_{\omega_0}$. A stochastic delay map with noisy measurements is then given by

$$\Phi_{F,\phi,\omega,\eta}(x) = (\phi_{\eta_0}(x), \phi_{\eta_1}(F_{\omega_0}(x)), \dots, \phi_{\eta_{d-1}}(F_{\omega_{d-2}\dots\omega_0}(x))). \tag{37}$$

A problem now arises because the shift map on Σ and Σ' represents infinite dimensional systems so that there is no hope of $\Phi_{F,\phi}$ embedding $M \times \Sigma \times \Sigma'$. Stark therefore defines the notion of a *bundle embedding* by requiring that the map $\Phi_{F,\phi,\omega,\eta} = \Phi_{F,\phi}(x, \omega, \eta)$ embeds M *given* typical $(\omega, \eta) \in \Sigma \times \Sigma'$. Typical is defined here in terms of product measures $\mu_\Sigma, \mu'_{\Sigma'}$ on Σ and Σ' respectively.

Theorem 4 Bundle Embeddings for stochastic Systems with noisy Observations, Stark 2002
Let $M, \mathcal{X}$ *and* $\mathcal{X}'$ *be compact manifolds, with* $m = \dim M > 0$. *Suppose that* $d \geq 2m + 1$. *Then for* $r \geq 1$, *there exists a residual set of* $(F, \phi) \in \mathcal{D}^r(M \times \mathcal{X}, M) \times \mathcal{C}^r(M \times \mathcal{X}', \mathbb{R})$ *such that for any* (F, ϕ) *in this set there is an open dense set* $\Sigma_{F,\phi} \subset \Sigma \times \Sigma'$ *such that* $\Phi_{F,\phi,\omega,\eta}$ *is an embedding for all* $(\omega, \eta) \in \Sigma_{F,\phi}$. *If* μ_Σ *and* $\mu'_{\Sigma'}$ *are invariant probability measures on* Σ *and* Σ' *respectively, such that* μ_{d-1} *and* μ'_d *are absolutely continuous with respect to Lebesgue measure on* $\mathcal{X}^{d-1}$ *and* $(\mathcal{X}')^d$ *respectively, then we can choose* $\Sigma_{F,\phi}$ *such that* $\mu_\Sigma \times \mu'_{\Sigma'}(\Sigma_{F,\phi}) = 1$.

Consequently, one can define an embedding from noisy measurements for a stochastically driven system. However, in this scenario the stochastic driver is not explicitly reconstructible and rather represents a parametrization of the driven system in terms of random variables. This random parametrization extends in particular to any mapping defined on the measurements, thus leading to uncertainty, as will be discussed shortly. Note that such bundle embeddings of the driven system on M exist naturally in the deterministic case of theorem 3 and can be used in situations where the driver is already known, for example due to the experimental setup.

Stark et al. [2003] also point out the following important consequence. Consider once more the commutative diagram in figure 3 and note that two states $z_i, z_{i+1} \in \mathbb{R}^d$ of the embedded system H only differ meaningfully in the last coordinate,

$$\begin{aligned} \Phi_{F,\phi}(x_i) = z_i &= (\phi(x_i), \phi(F(x_i)), \dots, \phi(F^{d-1}(x_i)))^T \\ \Phi_{F,\phi}(F(x_i)) = z_{i+1} &= (\phi(F(x_i)), \phi(F^2(x_i)), \dots, \phi(F^d(x_i)))^T. \end{aligned} \tag{38}$$

The existence of H implies the existence of its smooth coordinate functions. Denote the last coordinate function by $h := \mathrm{pr}_d \circ H$, which realizes the autoregressive mapping

$$h : (\phi(x_i), \phi(x_{i+1}), \dots, \phi(x_{i+d-1}))^T \mapsto \phi(x_{i+d}). \tag{39}$$

This continuous mapping provides a theoretical justification for the existence of any predictive model common in time series analysis. Predictive models therefore are always partial approximations of the underlying system flow (with fixed time argument). Mapping (39) can also be derived from a *bundle conjugacy* implied by theorem 4. Assume for illustration a stochastic dynamical system with noise-free deterministic measurement function. For almost all $\omega \in \Sigma$, $\Phi_{F,\phi,\omega}$ and $\Phi_{F,\phi,\sigma(\omega)}$ are both embeddings of the manifold M on which the dynamical system F is defined. A bundle conjugacy is then given by varying coordinate changes, such that $H_\omega = \Phi_{F,\phi,\sigma(\omega)} \circ F_{\omega_0} \circ \Phi^{-1}_{F,\phi,\omega}$ is a well defined diffeomorphism between $\Phi_{F,\phi,\omega} \subset \mathbb{R}^d$ and $\Phi_{F,\phi,\sigma(\omega)} \subset \mathbb{R}^d$ (see again definitions (37),(36)). As a result, for an orbit $(x_i, \sigma^i(\omega))$ where $x_{i+1} = F(x_i, \omega_i)$, it holds that

$$\begin{aligned} z_{i+1} &= \Phi_{F,\phi,\sigma^{i+1}(\omega)}(x_{i+1}) \\ &= \Phi_{F,\phi,\sigma^{i+1}(\omega)}(F_{\omega_i}(x_i)) \\ &= \Phi_{F,\phi,\sigma(\sigma^i(\omega))}(F_{\omega_i}((\Phi_{F,\phi,\sigma^i(\omega)})^{-1}(z_i))) \\ &= H_{\sigma^i(\omega)}(z_i). \end{aligned} \tag{40}$$

Once more, denote the last component of H_ω by $h_\omega : \Phi_{F,\phi,\omega}(M) \to \mathbb{R}$. This results in the mapping

$$\begin{aligned} \phi(x_{i+d}) &= h_{\sigma^i(\omega)}(\phi(x_i), \phi(x_{i+1}), ..., \phi(x_{i+d-1})) \\ &= h(\phi(x_i), \phi(x_{i+1}), ..., \phi(x_{i+d-1}), \omega_i, \omega_{i+1}, ..., \omega_{i+d-1}), \end{aligned} \tag{41}$$

where the last equation makes explicit the dependence of h on the d stochastic terms $\omega_i, ..., \omega_{i+d-1}$. Note that ω_{i+d-1} corresponds to new uncertainty entering the system between time steps $i + d - 1$ and $i + d$ and is therefore inaccessible at the time the prediction is made. As a result, the bundle conjugacy implies the existence of a general autoregressive statistical model, albeit only for almost all ω, of which the well-known nonlinear autoregressive moving average models (NARMA, see Billings [2013]) are special cases.

An important point to be reconsidered here in the context of prediction is that of forcing from finite dimensional systems. In this case, theorem 3 is applicable and allows the reconstruction not only of the driven system under observation, but also of its reconstructible drivers. Depending on the accessibility of information in the data, uncertainty in mapping 41 may be reduced dramatically as a result. Consequently, the statistical model is more than just *auto*regressive and can in addition draw on information of reconstructible drivers. This theoretical insight is of high significance since for many natural systems under observation it will be the rule rather than the exception that attempted measurements are local and therefore can only capture a local fraction of the full system.

In summary, embedding theory allows one to reconstruct the geometric information of the phase space of a dynamical system in different measurement scenarios by constructing delay maps of real-valued measurement functions. It was shown that these reconstructions provide the basis for predictive statistical models. Moreover, Stark's embedding theorems for skew products give a formal justification for the intuition developed in figure 1 regarding the fact that a driven system geometrically encodes information about its driver. The skew product embedding theorems also point to situations where the reconstruction of certain systems is not possible. This admits the possibility of exploiting asymmetries in reconstructibility to arrive at causal interaction structures that will be considered in Chapter 8.

4 OUTLINE AND SCIENTIFIC GOALS

The scientific goals of this thesis are the investigation of time series analysis tools for prediction and causal analysis, informed by dynamical systems theory. Emphasis is placed on the role of delays with respect to information processing in dynamical systems, as well as the effect on causal interactions between systems. The importance of investigating the role of delays is obvious from the fact that the majority of physical systems under investigation is comprised of spatially distributed subsystems that have to interact over a distance. Applications are considered foremost with respect to computational neuroscience but extend to generic time series measurements if the respective prerequisites apply.

In terms of the individual research projects that form the body of this thesis, the remainder is organized as follows. Chapter 5 and Chapter 6 correspond to papers documenting work on delay-coupled reservoirs. With respect to time series analysis the task of prediction is considered foremost. In addition, delay-coupled reservoirs provide a natural framework to study the role of delays for information processing in dynamical systems. In the context of information exchange between different systems, Chapter 7 and Chapter 8 explore dependencies between time series and approach the problem of causal analysis. In particular, Chapter 8 considers the effect of delays in interactions between coupled dynamical systems. The documented work aims at both, estimating these delays and exploiting them for causal inference.

I will now discuss the individual contents in more detail and provide additional background information. First, an introduction to delay-coupled reservoir computing, as developed by Appeltant et al. [2011], is provided. The corresponding Chapter 5 motivates the approach, introduces all mathematical theory involved, presents a numerical solution scheme for the model based on an analytical approximation, and evaluates predictive capabilities of the model both, on synthetic data and on time series measurements from a chaotic far-infrared laser system. Delay-coupled reservoirs provide nonlinear functional mappings in statistical models but represent by themselves already dynamical systems delay-coupled to themselves. The reservoir computing paradigm therefore affords investigating the capacity of delay-coupled systems to process information. This is of high relevance for the area of computational neuroscience since neural networks in practice always feature retarded interactions.

With respect to prediction, the delay-coupled reservoir is placed within a statistical model that employs Bayesian inference calculus. The latter allows for a proper treatment of uncertainty usually neglected in studies of the reservoir computing paradigm. A realization via Gaussian process regression is discussed, too. The laser time series targeted for prediction shows features known as nonstationarities. These comprise trends in mean and variance, as well as catastrophic events like sudden breakdowns of the amplitudes in the system's irregular oscillation. In contrast to the conventional purely stochastic treatment of nonstationarities, it will be shown that nonstationarities in the laser time series can be fully accounted for in terms of deterministic nonlinear modeling of the system's flow describing its temporal evolution (see Chapter 3).

Open questions pertain to the determination and practical treatment of system hyperparameters. The latter are buried within the dynamics of a retarded functional differential equation that constitutes the delay-coupled reservoir. In addition, some of these parameters are inconveniently entangled with the temporal grid to which approximate and numerical solutions of the system are confined. Given the context of computational neuroscience, Chapter 6 presents an investigation of biologically inspired optimization procedures of the reservoir informed by principles of homeostatic plasticity, as well as information theoretic considerations. This approach tries to evade the aforementioned difficulties by allowing the system to self-organize. In addition, it represents a first attempt at exploring unsupervised computational paradigms that explicitly target temporal aspects of the delay-coupling.

The two subsequent chapters document work focused on the analysis of measurements from interacting systems. Statistical models used to estimate functional interactions in this context will always be based on Volterra series operators, which I consider to be a canonical model for continuous functional relationships. A detailed motivation and derivation of this statement is provided in the supplementary material of Chapter 8. Data analysis here benefits greatly from theoretical studies of coupled dynamical systems. Particularly important examples are coupled chaotic systems the interaction of which was studied mainly in the 1990s (see e.g. Rulkov et al. [1995]; Rosenblum et al. [1997]; Kocarev and Parlitz [1996]; Senthilkumar et al. [2008]). A large portion of the possible regimes of interaction can be accounted for in terms of synchronization. It is typically understood that one of the weakest forms is phase synchrony in systems that feature a discernible main frequency component where it makes sense to define a mean-phase. As the coupling strength between systems is increased, the interaction becomes stronger and involves also the system amplitudes. This regime is called generalized synchronization. In contrast to phase synchrony, generalized synchronization is not limited to systems with well-defined mean-phase. Its main feature is complete predictability of one system given the other, although their interaction may be highly nonlinear. As the coupling strength increases further, in systems with high similarity lag synchronization and complete synchronization can be observed as special forms of generalized synchronization. The aforementioned synchronization phenomena are well-understood analytically and can inform data analysis. Chapter 7 presents a method that uses a particular analytical definition of generalized synchronization to justify the existence of a predictive functional relationship between time series, based on a simple reconstruction of the underlying systems. The latter is provided by the theory of Chapter 3 and not explicitly discussed in the short format of the method paper corresponding to Chapter 7. A statistical model based on Volterra series operators is used to estimate the targeted functional relationships. In a final step, the applicability of the method to electrode recordings from monkey visual cortex is demonstrated.

Chapter 8 is the central piece of this thesis and continuous the investigation of the previous method into weak coupling regimes that are not dominated by synchrony. Delayed interactions are explicitly accounted for and an emphasis is placed on estimating the delays. Picking up on theoretical aspects of existing work by Sugihara et al. [2012], a method is developed that exploits asymmetries in the reconstructibility of one time series from another, due to directedness of information flow between the underlying systems. The method draws heavily on various concepts from differential topology, as detailed in Chapter 3 and the supplementary material of Chapter 8, to justify the unidirectional existence of functional relationships between the time series. In the presence of delayed interactions the delays can be utilized to achieve inference even in strong coupling regimes. Moreover, the method uses an analytical criterion to estimate the delays. At the heart of this procedure is a statistical model that employs Gaussian process regression with Volterra series operators to reconstruct one time series from another and to quantify a difference in reconstructibility. The model makes exemplary use of Bayesian inference calculus to formalize and analytically treat uncertainty pertaining to the measurements that was previously unconsidered. In a supplementary material section both, the formalization of uncertainty, as well as a derivation of the Volterra series operator are provided. The main application of the resulting method is to electrode recordings from visual areas of cat cortex. It is shown that delay estimation is possible on raw data and leads to a biologically plausible connectivity diagram between the individual recording sites in different layers and areas.

Part II

PUBLICATIONS

This part contains the unaltered manuscripts of the publications that form the body of the thesis.

5 AN INTRODUCTION TO DELAY-COUPLED RESERVOIR COMPUTING

5.1 ABSTRACT

Reservoir computing has been successfully applied in difficult time series prediction tasks by injecting an input signal into a spatially extended reservoir of nonlinear subunits to perform history-dependent nonlinear computation. Recently, the network was replaced by a single nonlinear node, delay-coupled to itself. Instead of a spatial topology, subunits are arrayed in time along one delay span of the system. As a result, the reservoir exists only implicitly in a single delay differential equation, the numerical solving of which is costly. We give here a brief introduction to the general topic of delay-coupled reservoir computing and derive approximate analytical equations for the reservoir by solving the underlying system explicitly. The analytical approximation represents the system accurately and yields comparable performance in reservoir benchmark tasks, while reducing computational costs practically by several orders of magnitude. This has important implications with respect to electronic realizations of the reservoir and opens up new possibilities for optimization and theoretical investigation.

5.2 INTRODUCTION TO RESERVOIR COMPUTATION

Predicting future behavior and learning temporal dependencies in time series of complex natural systems remains a major goal in many disciplines. In Reservoir Computing, the issue is tackled by projecting input time series into a recurrent network of nonlinear subunits [Jaeger, 2001; Maass et al., 2002]: Recurrency provides memory of past inputs, while the large number of nonlinear subunits expand their informational features. History-dependent nonlinear computations are then achieved by simple linear readouts of the network activity.

In a recent advancement, the recurrent network was replaced by a single nonlinear node, delay-coupled to itself [Appeltant et al., 2011]. Such a setup is formalized by a delay differential equation which can be interpreted as an *infinite-dimensional* dynamical system. Whereas classical reservoirs have an explicit spatial representation, a delay-coupled reservoir (DCR) uses temporally extended sampling points across the span of its delayed feedback, termed *virtual nodes*. The main advantage of such a setup is that it allows for easy realization in optical and electronic hardware [Soriano et al., 2013] which has great potential for industrial application.

A drawback of this approach is the fact that the actual reservoir is always only implicit in a single delay differential equation. Consequently, in many implementations the underlying system has to be solved numerically. This leads to a computational bottleneck and creates practical limitations for reservoir size and utility. The lack of reservoir equations also presents problems for applying optimization procedures.

To overcome this, we present a recursive analytical solution used to derive approximate virtual node equations. The solution is assessed in its computational capabilities as a DCR, and compared against numerical solvers in nonlinear benchmark tasks. While computational performance is comparable, the analytical approximation leads to considerable savings in computation time, allowing the exploration of exceedingly large setups. The latter allows us to stir the system away from toy examples to the realm of application.

Moreover, we provide in this chapter a general introduction to the topic of delay-coupled reservoir computing to familiarize the reader with the concept, and to give some practical guidelines for setting up a DCR. To this end, we first discuss some theory regarding solvability and dynamics of the delay differential equation underlying a DCR. In a next step, we use this insight to derive the approximate analytical equations for the reservoir and explore their accuracy. A computationally efficient numerical solution scheme arises that allows the investigation of large reservoirs, the performance of which is evaluated on classical benchmark tasks. These will also serve to illustrate a practical implementation. Finally, we present an application to an experimental recording of a far-infrared laser operating in a chaotic regime and show how a DCR can be embedded into Gaussian process regression to deal with uncertainty and to be able to optimize hyperparameters.

5.3 SINGLE NODE DELAY-COUPLED RESERVOIRS

In this section, we discuss the theory of simple retarded functional differential equations, of which the delay differential equations underlying a DCR are a particular instance, and derive approximate expressions for the virtual nodes. These will serve as the basis for a practical implementation in later parts of this chapter.

5.3.1 *Computation via Delayed Feedback*

In a DCR, past and present information of a covariate time series undergo nonlinear mixing via injection into a dynamically evolving "node" with delayed feedback. Formally, these dynamics can be modeled by a delay differential equation of the type

$$\frac{dx(t)}{dt} = -x(t) + f(x(t-\tau), J(t)) \in \mathbb{R}, \tag{42}$$

where τ is the delay time, $J(t)$ is a weighted and temporally multiplexed transformation of some input signal $u(t)$ driving the system, and f is a sufficiently smooth real-valued nonlinear function. The nonlinearity is necessary to provide a rich feature expansion and separability of the information present in the input time series. Although a solution to system (42) will in general not be obtainable analytically, it can often be fully optically or electronically realized, for example in an all-optical laser system with nonlinear interference given by its own delayed feedback [Larger et al., 2012b].

The Mackey-Glass system represents a possible choice of nonlinearity which also admits a hardware implementation of the system. The corresponding differential equation is given by

$$\frac{dx(t)}{dt} = -x(t) + \frac{\eta(x(t-\tau) + \gamma J(t))}{1 + (x(t-\tau) + \gamma J(t))^p} \in \mathbb{R}. \tag{43}$$

The parameters $\gamma, \eta, p \in \mathbb{R}$ determine in which dynamical regime the system operates. Although the Mackey-Glass system can exhibit even chaotic dynamics for $p > 9$, a fixed point regime appears to have the most suitable properties with respect to memory capacity of the resulting reservoir. In a chaotic regime, the system would have an excellent separability of input features, due to its sensitive dependence on initial conditions. However, presumably as a result of the intrinsic entropy production and exponential decay of auto-correlation in strange attractors, the chaotic system essentially lacks memory capacity with respect to the input signal. Furthermore, the required precision for computing the trajectories in a numerical solver would result in prohibitive computational costs with increasing chaoticity of the system. Other choices of nonlinearities are possible and have been investigated [Appeltant et al., 2011]. For purposes of illustration, we will use the Mackey-Glass system throughout this chapter.

Injecting a signal $u(t)$ into the reservoir is achieved by multiplexing it in time: The DCR receives a single constant input $u(\bar{t}) \in \mathbb{R}$ in each reservoir time step $\bar{t} = \lceil \frac{t}{\tau} \rceil$, corresponding to one τ-cycle of the system. For example, $(i-1)\tau \leq t \leq i\tau$ denotes the ith τ-cycle and is considered to be a single reservoir time step during which $u(t) = u_i = const$. This scheme is easily extendable to vector-valued input signals. The dynamics of x during a single delay span τ are to be seen as the temporally extended analogon of an artificial neural network, where the nonlinear subunits are arrayed not in space, but in time. Since one is interested in operating $x(t)$ in a fixed point regime, the system would at this point simply saturate and converge in the course of a τ-cycle during which u_i is constant, e.g. $\lim_{t\to\infty} x(t) = 0$ for $u_i = 0$ and suitable initial conditions (see section 5.3.2). To add perturbation and create a rich *feature expansion* of the input signal in time (analogous to neural network activity in space), the delay line τ is shattered into N subintervals of length θ_j, for $j = 1, ..., N$. On these subintervals an additional mask function reweights the otherwise constant input value u_i, such that the saturating system x is frequently perturbed and prevented from converging. That is, the mask is a real-valued function on an interval of length τ, e.g. $[-\tau, 0]$, which is piecewise constant:

$$m(t) = m_j \in \mathbb{R} \qquad \text{for} \quad \theta_{j-1} < t \leq \theta_j, \quad \sum_{j=1}^{N} \theta_j = \tau. \tag{44}$$

The input to the reservoir $x(t)$ during the ith τ-cycle is thus given by $J(t) = m(t)u_i$ (compare eq. (42)).

In the original approach, the m_j were simply random samples from $\{-1, 1\}$, meant to perturb the system and to create transient trajectories. The impact of certain binary sequences on computation, as well as multi-valued masks, have also been studied and may lead to a task-specific improvement of performance. General principles for the choice of optimal $m(t)$ are, however, not yet well understood,

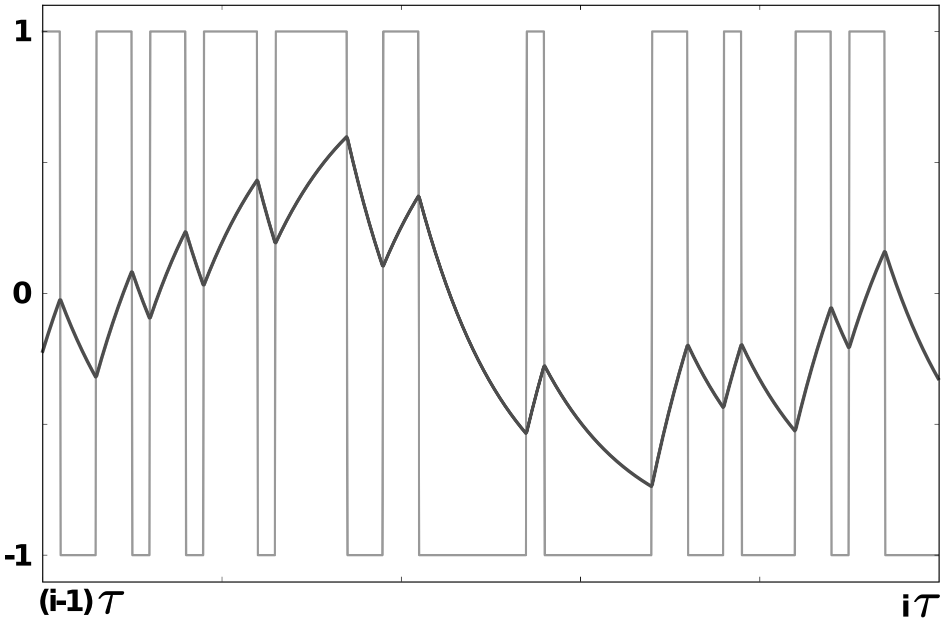

Figure 4: Exemplary trajectory (dark gray) of system (43) during one τ-cycle of length 10, with equidistant $\theta_j = 0.2$ for $j = 1, ..., 50$, $\eta = 0.04$, $\gamma = 0.005$, $p = 7$, while $u_i = 0.385$ constant during the entire τ-cycle. In light gray, the piecewise constant mask (eq. (44)) with $m_j \in \{-1, 1\}$ for $j = 1, ..., 50$ is shown. The system is in a fixed point regime (parametrized by $J(t)$) and starts converging on intervals where $J(t) = u_i m(t)$ is constant.

mainly due to problems in solving the system efficiently and due to an inconvenient dependence on the θ_j. For optimal information processing, θ_j has to be short enough to prevent convergence of the system trajectory (so as to retain the influence of past states), but long enough for the system to act upon the masked input signal and expand information. In the case of equidistant $\theta_j = \theta$, it can be determined experimentally that choosing θ to be one fifth of the system's intrinsic time scale yields good computational performance in benchmark tasks [Appeltant et al., 2011]. In system (42) the intrinsic timescale is the multiplicative factor of the derivative on the left-hand side, which is 1, so that $\theta = 0.2$ accordingly. Figure (4) illustrates how the resulting dynamics of $x(t)$ may look like during a single τ-cycle, using a short delay span $\tau = 10$, shattered into 50 subintervals.

To obtain a statistical model, a sample is read out at the end of each θ_j, thus yielding N predictor variables at each τ-cycle (i.e. reservoir time step $\bar{t} = \lceil \frac{t}{\tau} \rceil$). These are termed "*virtual nodes*" in analogy to the spatially extended nodes of a neural network. During the ith τ-cycle, virtual node j is sampled as

$$x_j(u_i) := x((i-1)\tau + \sum_{k=1}^{j} \theta_k)$$

and used in a linear functional mapping

$$\hat{y}_i = \sum_{j=1}^{N} \alpha_j x_j(u_i) \approx g(u_i, ..., u_{i-M}) \tag{45}$$

to predict some scalar target signal y by the estimator $\hat{y}$. The latter can be seen as a function g of covariate time series u, where the finite fading memory of the

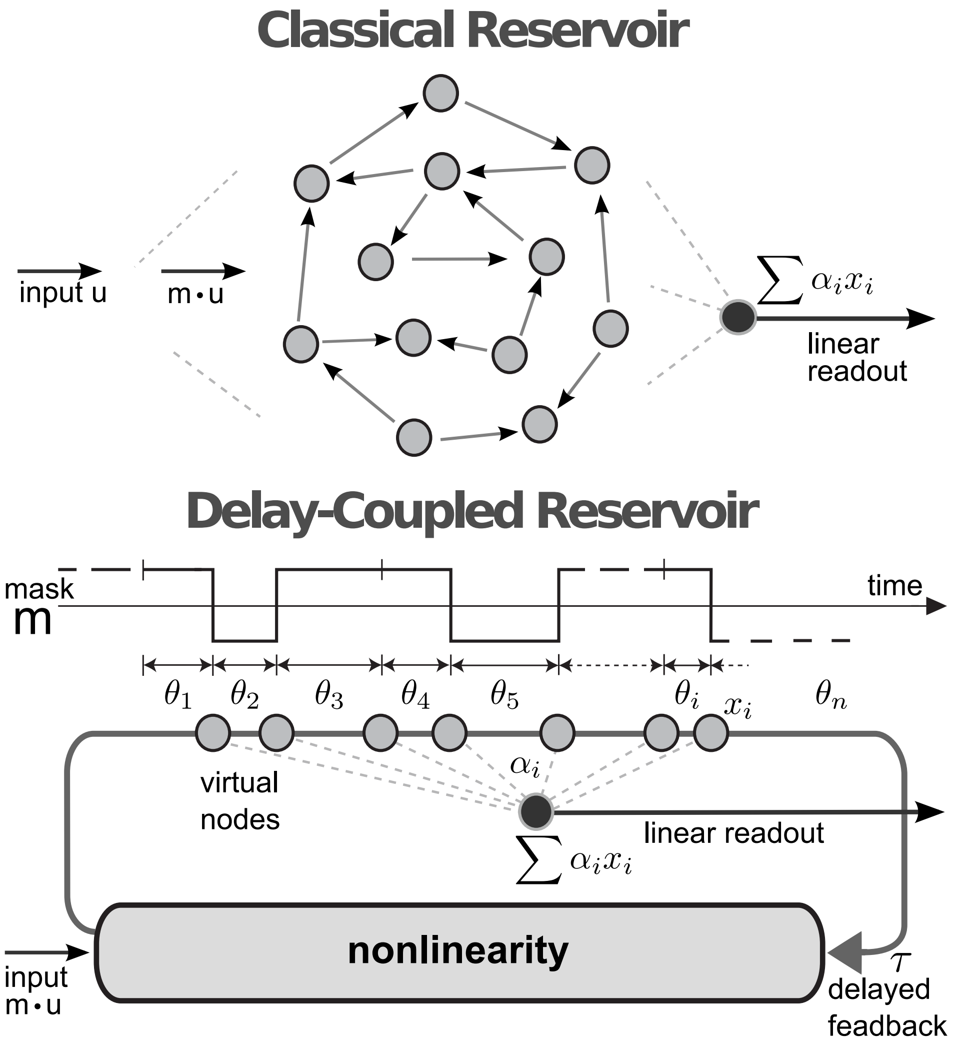

Figure 5: Schematic illustration of a DCR (bottom), where nonlinear units (*virtual nodes*) are arrayed in time along the delay span τ, with temporal distance θ_j between virtual nodes $j-1$ and j. In contrast, a classical reservoir (top) has nonlinear units extended in space according to a certain network topology.

underlying system x causes g to be a function of at most $M+1$ predictor variables $u_i, ..., u_{i-M}$, where $\mathbb{N} \ni M \ll \infty$. The memory capacity of the system, indicated by M, represents the availability for computation of past input values u_i across τ-cycles. For details, the reader is referred to Appeltant et al. [2011]. A schematic illustration of such a DCR is given in figure (5) and compared to a classical neural network with spatially extended nodes. The α_j are free parameters of the statistical model. The simplest way to determine the coefficients is by linear regression, i.e. using the *least squares solution* minimizing the sum of squared errors, $\sum_i (y_i - \hat{y}_i)^2$. However, in general, this approach will only be feasible in case of a noise-free target y and on large data sets. For a more rigorous statistical treatment of uncertainty, the problem can be formalized, for example, within the framework of Bayesian statistics or Gaussian process regression (see 5.4.5 and appendix). The delay span τ has, due to the recursive nature of (1), no impact on the memory capacity and can simply be chosen to harbor an adequate number N of virtual nodes along the delay line. Typically, $N = 400$ strikes a good balance between computational cost and reservoir performance in benchmark tasks.

In order to study the DCR model (45), system (1) has to be solved and virtual nodes sampled accordingly. However, (1) can neither be solved directly, due to the recursive terms, nor exists, to the best of our knowledge, a series expansion of the solution (for example of *Peano-Baker* type), due to the nonlinearity of f. Typically, (1) is therefore solved numerically using a lower order Runge-Kutta method, such as Heun's method. Although for the simple fixed point regime, the system operating in a numerical step size of $\theta/2$ is sufficient, the computational cost for large numbers of virtual nodes $N \gg 500$ can still be quite excessive if the hyperparameters are unknown and have to be determined.

From a modeling perspective, the hyperparameters θ_j, τ, N, $m(t)$, γ, η are shrouded in the nonlinear non-solvable retarded functional differential equation (1), hardly accessible to optimization. Furthermore, due to the piecewise sampling procedure of the virtual nodes, the shape of m is inconveniently restricted and entangled with the sampling points θ_j, as well as the numerical simulation grid. If the latter is chosen too fine-grained, the computational costs become prohibitive in the scheme described above. Task-specific optimization of the hyperparameters is, however, crucial prior to any hardware implementation. In an optical implementation, for example, τ may directly determine the length of the glass-fibre cable that projects the delayed feedback back into the system. Determining optimal hyperparameters therefore has to be done in advance by means of numerical simulation of the system and with respect to rigorous statistical optimality criteria that will allow for proper confidence in the resulting permanent hardware setup.

To address these issues it is necessary to study system (1) in some detail. The goal is to gain insight into the dynamics of the system underlying the DCR, and to understand its theoretical solution as a semi-flow in an infinite-dimensional state space. These insights will form the basis for an approximate analytical solution scheme, alleviating access to the system as a functional statistical model and its hyperparameters. In addition, the approximate solution gives rise to a fast simulation algorithm, which will be key to the analysis and optimization of the DCR.

5.3.2 *Retarded Functional Differential Equations*

Following Guo and Wu [2013b], let $C_\tau := C([-\tau, 0], \mathbb{R})$ denote the Banach space of continuous mappings from $[-\tau, 0]$ into $\mathbb{R}$, equipped with the supremum norm. If $t_0 \in \mathbb{R}$, $A \geq 0$ and $x : [t_0 - \tau, t_0 + A] \to \mathbb{R}$ a continuous mapping, then $\forall t \in [t_0, t_0 + A]$, one can define $C_\tau \ni x_t : C_\tau \to C_\tau$ by $x_t(\sigma) = x(t + \sigma)$ for $\sigma \in [-\tau, 0]$. Furthermore, let $H : C_\tau \to \mathbb{R}$ be a mapping such that

$$\frac{dx(t)}{dt} = H(x_t), \tag{46}$$

then (46) is called a *retarded functional differential equation*. A solution to (46) is a differentiable function x satisfying (46) on $[t_0, t_0 + A]$ such that $x \in C([t_0 - \tau, t_0 + A), \mathbb{R})$. If H is locally Lipschitz continuous then x is unique, given an initial condition $(t_0, \phi) \in \mathbb{R} \times C_\tau$.

To illustrate this, consider a solution $x(t)$ of system (43) for $t \geq 0$, where

$$h : \mathbb{R} \times \mathbb{R} \to \mathbb{R} \tag{47}$$

such that

$$h : (x(t), x(t-\tau)) \mapsto \frac{\eta(x(t-\tau) + \gamma m(t) u(\bar{t})}{1 + (x(t-\tau) + \gamma m(t) u(\bar{t}))^p} - x(t)$$

as given in (43), where mask m and input u are known (recall $m(t)u(\bar{t}) = J(t)$) and $\bar{t} = \lceil \frac{t}{\tau} \rceil$. Assume $p = 1$ for illustration in this section. The system will depend on its history during $[-\tau, 0]$ as specified by

$$\phi : [-\tau, 0] \to \mathbb{R}.$$

If ϕ is continuous, then h is continuous and locally Lipschitz in $x(t)$, since f is differentiable and $\sup |\frac{d}{dx} h(x, \phi)| = 1$. As a result, for $t \in [0, \tau]$

$$\frac{dx(t)}{dt} = h(x(t), \phi(t-\tau)), \quad t \in [0, \tau], \quad x(0) = \phi_0(0)$$

specifies an initial value problem that is solvable. Denote this solution on $t \in [0, \tau]$ by ϕ_1. Then

$$\frac{dx(t)}{dt} = h(x(t), \phi_1(t-\tau)), \quad t \in [\tau, 2\tau]$$

becomes solvable, too, since $x(\tau) = \phi_1(\tau)$ is already given. One can iterate this procedure to yield solutions to (43) on all intervals $[(i-1)\tau, i\tau]$, subject to some initial condition $\phi_0 = x|_{[-\tau,0]}$. This procedure is know as the *method of steps* [Guo and Wu, 2013b].

The function $x_t : C_\tau \to C_\tau$, defined above as

$$x_t(\sigma) = x(t + \sigma), \quad \sigma \in [-\tau, 0],$$

specifies a translation of the segment of x on $[t-\tau, t]$ back to the initial interval $[-\tau, 0]$. Together with $x_0 = \phi_0$, the solution to system (57) (and, more generally, to (42)) can be shown to yield a semiflow $[0, \infty] \ni t \mapsto x_t \in C_\tau$.

It is now apparent that $\sigma \in [-\tau, 0]$ corresponds to a parameterization of *coordinates* for an "*infinite vector*" $x(\sigma)$ in state space C_τ, which is why (42) really constitutes an infinite-dimensional dynamical system. Delay-coupled reservoir computation can thus be thought of as expanding an input time series nonlinearly into an infinite-dimensional feature state space. However, given the sampling scheme of *virtual nodes*, and the piecewise constant mask, these properties will be effectively reduced to the number of samples N that constitute the covariates of the statistical model (45).

To study the stability properties of system (57), consider the autonomous case with constant input $J(t) = const$, and recall the temporal evolution h of system x as given by (47). For a fixed point x^* it holds that

$$\frac{dx}{dt} = h(x^*, x^*) = 0$$

Setting $p = \gamma = 1$ for illustration, solving for x^* evaluates to

$$\begin{aligned} 0 &= \frac{\eta(x^* + J)}{1 + (x^* + J)} - x^* \\ x^* &= \frac{\eta - 1 - J}{2} \pm \sqrt{\left(\frac{1 + J - \eta}{2}\right)^2 + \eta J}. \end{aligned} \tag{48}$$

To simplify the expression, let $J = 0$, in which case one obtains

$$x^* = \frac{\eta - 1}{2} \pm \frac{1 - \eta}{2}.$$

Of interest is now the central solution $x^* = 0$, around which the reservoir will be operated later on by suitable choice of η and γ. To determine its stability, one linearizes the system in a small neighborhood of x^* by dropping all higher order terms in the Taylor series expansion of h, which gives

$$\frac{dx}{dt} = D_x[h](x^*, x^*)x(t) + D_y[h](x^*, x^*)x(t - \tau) \tag{49}$$

where

$$\begin{aligned} D_x[h](x^*, x^*) &= \frac{\partial}{\partial x} h(x, y)|_{x=y=x^*} = -1 \\ D_y[h](x^*, x^*) &= \frac{\partial}{\partial y} h(x, y)|_{x=y=x^*} = \frac{\eta}{1 + x^*}. \end{aligned} \tag{50}$$

If $D_x[h](x^*, x^*) + D_y[h](x^*, x^*) < 0$ and $D_y[h](x^*, x^*) \geq D_x[h](x^*, x^*)$, $x^* = 0$ is *asymptotically stable* (Theorem 4.7, Smith [2010]), which is the case for $\eta \in [-1, 1)$. In general, the analysis of eq. (49) may be quite involved and can result in an infinite number of linearly independent solutions $x(t) = Ce^{\lambda t}$, since the corresponding characteristic equation $\lambda = D_x[h](x^*, x^*) + D_y[h](x^*, x^*)e^{-\lambda\tau}$ gives rise to an analytic function that may define roots on the entire complex plane.

Equations (48) and (50) already indicate that, given suitable η and γ, for $J(t) = m_j u_i$ a fixed point regime may still exist between two virtual node sampling points $\theta_{j-1} < t \leq \theta_j$, in which the reservoir computer can be operated while driven by input time series u. However, for $p > 1$ bifurcations can occur (even a period doubling route to chaos), if the feedback term is weighted too strongly. In this case, the virtual nodes would not yield a consistent covariate feature expansion and, in all likelihood, lead to a bad performance of the statistical model (45). The above considerations also show that the nonlinear feature expansion desired in reservoir computation will depend completely on the mask $m(t)$, since the input u_i is constant most of the time and $x(t)$ practically converges in a fraction of τ. The changes in m are in fact the only source of perturbation that prevents the semiflow $x(t)$ from converging to its fixed point. It is not at all clear in this situation that a piecewise constant binary (or even n-valued) mask is the most interesting choice to be made. More research is needed to determine optimal mask functions.

5.3.3 *Approximate virtual node equations*

In the following, we discuss a recursive analytical solution to equation (42), employing the method of steps. The resulting formulas are used to derive a piecewise solution scheme for sampling points across τ that correspond to the reservoir's virtual nodes. Finally, we use the *trapezoidal rule* for further simplification, hereby deriving approximate virtual node equations, the temporal dependencies of which only consist of other virtual nodes. As will be shown in the remainder of this article, the resulting closed-form solutions allow reservoir computation without significant loss of performance as compared to a system obtained by explicit numerical solutions, e.g. *Heun's method* ((1,2) Runge-Kutta).

First, we discuss a simple application of the method of steps. System (42) is to be evaluated for the span of one τ during the ith τ-cycle, so $(i-1)\tau \leq t \leq i\tau$. Let a continuous function $\phi_{i-1}(\sigma) \in C_{[(i-2)\tau,(i-1)\tau]}$ be the solution for $x(t)$ on the previous τ-interval. We can now replace the unknown $x(t-\tau)$ by the known $\phi_{i-1}(t-\tau)$ in equation (42). Consequently, (42) can be solved as an ODE where the *variation of constants* [Heuser, 2009] is directly applicable. The variation of constants theorem states that a real valued differential equation of type

$$\frac{dy}{dt} = a(t)y + b(t)$$

with initial value $y(t_0) = c \in \mathbb{R}$ has exactly one solution, given by

$$y(t) = y_h(t)\left(c + \int_{t_0}^{t} \frac{b(s)}{y_h(s)} ds\right), \tag{51}$$

where

$$y_h(t) = exp\left(\int_{t_0}^{t} a(s)ds\right)$$

is a solution of the corresponding homogeneous differential equation $\frac{dy}{dt} = a(t)y$. In system (42), we identify $a(t) = -1$ and $b(t) = f(\phi_{i-1}(t-\tau), J(t))$. Applying (51) now yields immediately the solution to the initial value problem (42) on the interval $t_{i-1} = (i-1)\tau \leq t \leq i\tau$, with initial value $x(t_{i-1}) = \phi_{i-1}((i-1)\tau)$, given by

$$x(t) = \phi_{i-1}(t_{i-1})e^{t_{i-1}-t} + e^{t_{i-1}-t}\int_{(i-1)\tau}^{t} f(\phi_i(s-\tau), J(s))e^{s-t_{i-1}}ds. \tag{52}$$

Recall that the semi-flow corresponding to $x(t)$ is determined by the mapping $x_t : C_\tau \to C_\tau$, defined above as $x_t(\sigma) = x(t+\sigma)$ with $\sigma \in [-\tau, 0]$. This specifies a translation of the segment of x on $[t-\tau, t]$ back to the initial interval $[-\tau, 0]$. Accordingly, we can reparametrize the solution for $x(t)$ in terms of $\sigma \in [-\tau, 0]$. Let x_i denote the solution on the ith τ-interval, then

$$x_i(\sigma) = x_{i-1}(0)e^{-(\tau+\sigma)} + e^{-(\tau+\sigma)}\int_{-\tau}^{\sigma} f(x_{i-1}(s), m(s)u_i)e^{s+\tau}ds, \tag{53}$$

where u_i denotes the constant input in the i^{th} reservoir time step, and we assume $m(\sigma)$ only has a finite number of discontinuities (see (44)). Accordingly, $x_{i-1} \in C_{[-\tau,0]}$.

Due to the recursion in the nonlinear $x_{i-1} = \phi_{i-1}$, the integral in (53) cannot be solved analytically. To approximate the integral, the recursion requires repeated evaluation at the same sampling points in each τ-cycle. The family of *Newton-Cotes formulas* for numerical integration is appropriate in this situation. We use the *cumulative trapezoidal rule* [Quarteroni et al., 2006], which is 2^{nd} order accurate and the simplest in that family. It is given by

$$\int_a^b g(x)dx \approx \frac{1}{2}\sum_{j=1}^{N}(\chi_j - \chi_{j-1})(g(\chi_j) + g(\chi_{j-1})), \quad \text{with} \quad \chi_0 = a, \chi_N = b. \tag{54}$$

To approximate the integral in (53), consider a non-uniform grid
$-\tau = \chi_0 < ... < \chi_N = 0$, where $\chi_j - \chi_{j-1} = \theta_j$. This yields the approximation

$$\begin{aligned} x_i(\chi_k) \approx{} & x_{i-1}(\chi_N)\exp(-\sum_{i=1}^{k}\theta_i) \\ & + \exp(-\sum_{i=1}^{k}\theta_i)[\frac{1}{2}\sum_{j=1}^{k}\theta_j \exp(\sum_{i=1}^{j-1}\theta_i)(f[x_{i-1}(\chi_j), m(\chi_j)u_i]e^{\theta_j} \\ & + f[x_{i-1}(\chi_{j-1}), m(\chi_{j-1})u_i]], \end{aligned} \tag{55}$$

which can be computed as cumulative sum in one shot per τ-cycle for all approximation steps χ_j simultaneously.

We are now interested in equations for $1 \leq k \leq N$ single virtual nodes $x_{i\,k}$, during reservoir time step i. At this point we will choose an equidistant sampling grid of virtual nodes. This is not crucial but simplifies the notation considerably for the purpose of illustration. Assuming equidistant virtual nodes, it holds that $\tau = N\theta$ where N is the number of virtual nodes. For application of formula (54), the numerical sampling grid will be uniform and chosen directly as the sampling points of virtual nodes, such that $\chi_j = -\tau + j\theta$ with $j = 0, ..., N$. To get an expression for $x_{i\,k}$, we now have to evaluate equation (53) at the sampling point $t = -\tau + k\theta$, which results in

$$\begin{aligned} x_{i\,k} &= x_i(k\theta) = x((i-1)\tau + k\theta) \\ &\approx e^{-k\theta}x_{(i-1)\,N} + \frac{\theta}{2}e^{-k\theta}f[x_{(i-2)\,N}, J_N(i-1)] \\ &\quad + \frac{\theta}{2}f[x_{(i-1)\,k}, J_k(i)] + \sum_{j=1}^{k-1}\underbrace{\theta e^{(j-k)\theta}}_{c_{k\,j}}f[x_{(i-1)\,j}, J_j(i)]. \end{aligned} \tag{56}$$

Here $J_k(i) = m_k u_i$ denotes the masked input to node k at reservoir time step i, given a mask that is piecewise constant on each θ-interval (see eq. 44).

Note that equation (56) only has dependencies on sampling points corresponding to other virtual nodes. An exemplary coupling coefficient is indicated by c_{kj}, weighting a nonlinear coupling from node j to node k. In addition, each node receives exponentially weighted input from virtual node N (first term eq. (56)). In analogy to a classical reservoir with a spatially extended network topology, we can derive a corresponding weight matrix. Figure (6) shows an exemplary DCR weight matrix for a temporal network of 20 virtual nodes, equidistantly spaced along a delay span $\tau = 4$ with distance $\theta = 0.2$. The lower triangular shape highlights the

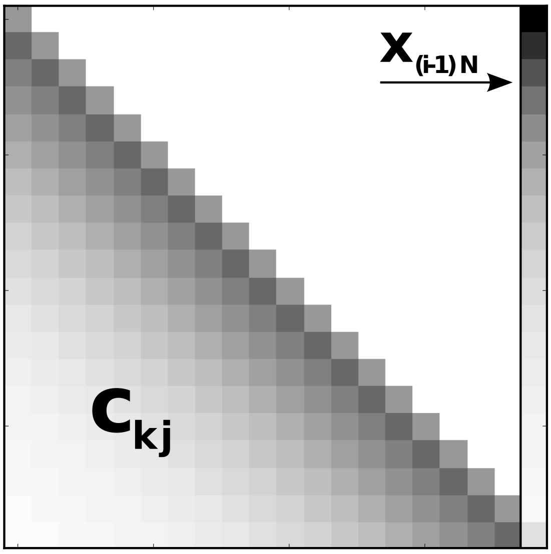

Figure 6: Illustration of a *temporal* weight matrix for a DCR comprised of 20 virtual nodes, arrayed on a delay line of length $\tau = 4$ with equidistant spacing $\theta = 0.2$. The lower triangular part of the matrix corresponds to the c_{kj} of formula (56), where darker colour indicates stronger coupling. The last column indicates the dependence on node $x_{(i-1)N}$, as given by the first term in expression (56).

fact that a DCR corresponds to a classical reservoir with ring topology. A ring topology features the longest possible feed-forward structure in a given network. Since the memory capacity of a recurrent network is mainly dependent on the length of its internal feed-forward structures, the ring topology is most advantageous with respect to the network's finite fading memory [Ganguli et al., 2008].

Formula (56) allows simultaneous computation of all nodes in one reservoir time step (τ-cycle) by a single vector operation, hereby dramatically reducing the computation time of simulating the DCR by several orders of magnitude in practice, as compared to an explicit second order numerical ODE solver.

5.4 IMPLEMENTATION AND PERFORMANCE OF THE DCR

We compare the analytical approximation of the Mackey-Glass DCR, derived in the previous section, to a numerical solution obtained using Heun's method with a stepsize of 0.1. The latter is chosen due to the relatively low computational cost and provides sufficient accuracy in the context of DCR computing. As a reference for absolute accuracy, we use numerical solutions obtained with *dde23* [Shampine and Thompson, 2001], an adaptive (2,3) Runge-Kutta based method for delay differential equations. The nonlinearity f is chosen according to the Mackey-Glass equation with $p = 1$ for the remainder of this chapter, such that the system is given by

$$\frac{dx}{dt} = -x(t) + \frac{\eta(x(t-\tau) + \gamma J(t))}{1 + x(t-\tau) + \gamma J(t)}, \tag{57}$$

where η, γ and p are metaparameters, τ the delay length, and $J(t)$ is the temporally stretched input $u(\bar{t})$, $\bar{t} = \lceil \frac{t}{\tau} \rceil$, multiplexed with a binary mask m (see eq. 44).

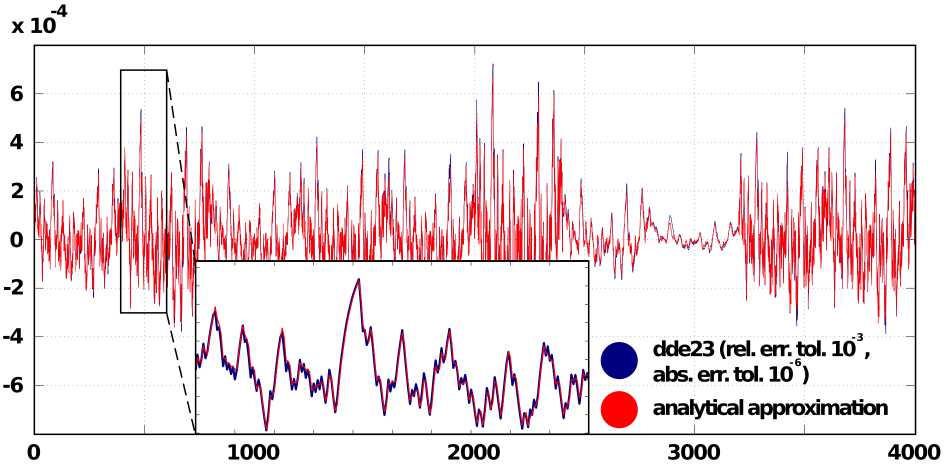

Figure 7: Comparison between analytical approximation and numerical solution for an input-driven Mackey-Glass system with parameters $\eta = 0.4$, $\gamma = 0.005$ and $p = 1$, sampled at the temporal positions of virtual nodes, with a distance $\theta = 0.2$.

Note that the trapezoidal rule used in the analytical approximation, as well as Heun's method, are both second order numerical methods that should yield a global truncation error of the same complexity class. As a result, discrepancies originating from different step sizes employed in the two approaches (e.g. 0.2 in the analytical approximation and 0.1 in the numerical solution) may be remedied by simply decreasing θ in the analytical approximation, for example by increasing N while keeping a fixed τ (see sec. 5.4.4).

5.4.1 *Trajectory Comparison*

In a first step, we wish to establish the general accuracy of the analytical approximation in a DCR relevant setup. Figure 7 shows a comparison of reservoir trajectories computed with equation (56) (red) against trajectories computed numerically using *dde23* (blue) with relative error tolerance 10^{-3} and absolute error tolerance 10^{-6}. The system received uniformly distributed input $u(\bar{t}) \sim \mathcal{U}_{[0,0.5]}$. The sample points correspond to the activities of $N = 400$ virtual nodes with a temporal distance of $\theta = 0.2$, and $\tau = 80$ accordingly. Given 4000 samples (corresponding to 10 reservoir time steps $\bar{t}$), the mean squared error between the trajectories is *MSE* $= 5.4 \times 10^{-10}$. As can be seen in the figure, the trajectories agree very well in the fixed point regime of the system (autonomous case). Although it is expected that the *MSE* would increase in more complex dynamic regimes (e.g. chaos), the latter are usually not very suitable for a DCR for various reasons. The following results also show a high task performance of the analytical approximation when used for DCR computing.

5.4.2 *NARMA-10*

A widely used benchmark in reservoir computing is the capacity of the DCR to model a nonlinear autoregressive moving average system y in response to uniformly distributed scalar input $u(k) \sim \mathcal{U}_{[0,0.5]}$. The NARMA-10 task requires the DCR to compute at each time step k a response

$$y(k+1) = 0.3y(k) + 0.05y(k)\sum_{i=0}^{9} y(k-i) + 1.5u(k)u(k-9) + 0.1.$$

Thus, NARMA-10 requires modeling of quadratic nonlinearities and shows a strong history dependence that challenges the DCR's memory capacity. We measure performance in this task using the correlation coefficient $r(y,\hat{y}) \in [-1,1]$ between the target time series y and the DCR output $\hat{y}$ in response to u. Here, the DCR is trained (see sec. 5.3.1) on 3000 data samples, while $r(y,\hat{y})$ is computed on an independent validation data set of size 1000. Figure 8A summarizes the performance of 50 different trials for a DCR computed using the analytical approximation (see eq. 56), shown in red, as compared to a DCR simulated with Heun's method, shown in blue. Both reservoirs consist of $N = 400$ virtual nodes, evenly spaced with a distance $\theta = 0.2$ along a delay line $\tau = 80$. Both systems show a comparable performance across the 50 trials, with a median correlation coefficient between $r(y,\hat{y}) = 0.96$ and 0.97, respectively.

5.4.3 *5-Bit Parity*

As a second benchmark, we chose the delayed 5-bit parity task [Schrauwen et al., 2008a], requiring the DCR to handle binary input sequences on which strong nonlinear computations have to be performed with arbitrary history dependence. Given a random input sequence u with $u(k) \in \{-1,1\}$, the DCR has to compute at each time step k the parity $p_m^\delta(k) = \prod_{i=0}^{m} u(k-i-\delta) \in \{-1,1\}$, for $\delta = 0,...,\infty$. The *performance* ϕ_m is then calculated on n data points as $\phi_m = \sum_{\delta=0}^{\infty} \kappa_m^\delta$, where *Cohen's Kappa*

$$\kappa_m^\delta = \frac{\frac{1}{n}\sum_{k=1}^{n} \max(0, p_m^\delta(k)\hat{y}(k)) - p_c}{1 - p_c} \in \{0,1\}$$

normalizes the average number of correct DCR output parities $\hat{y}$ by the chance level $p_c = 0.5$. We used 3000/1000 data points in training and validation set respectively. To compare performance between analytical approximation and numerical solution of the DCR, we chose $m = 5$ and truncated ϕ_m at $\delta = 7$, so that $\phi_5 \in [0,7]$. For parameters $\eta = 0.24$, $\gamma = 0.032$ and $p = 1$, and a DCR comprised of 400 neurons ($\tau = 80$), figure 8B shows that performance ϕ_5 is comparable for both versions of the DCR, with median performances between 4.3 and 4.5. across 50 different trials of this task. As the performance is far from the ideal value of 7 and the model suffers slightly from overfitting (not shown), it is clear that the delayed 5-bit parity task is a hard problem which leaves much space for improvement.

5.4.4 *Large Setups*

We repeated the tasks in larger network setups where the computational cost of the numerical solver becomes prohibitive. In addition to increasing the number of virtual nodes N one can also decrease the node distance θ, thus fitting more

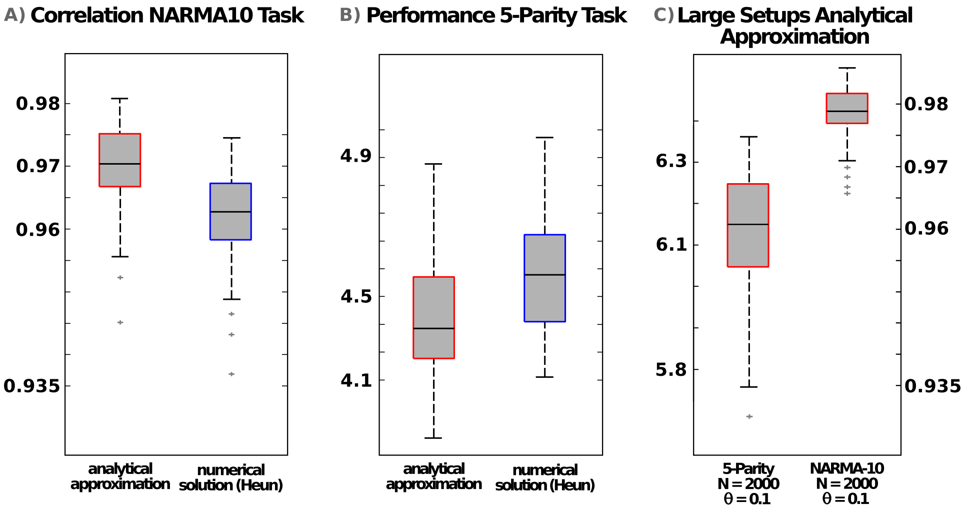

Figure 8: Comparison on nonlinear tasks between analytical approximation and numerical solution for an input-driven Mackey-Glass system, sampled at the temporal positions of virtual nodes with a distance $\theta = 0.2$. Mackey-Glass parameters are $\eta = 0.4$, $\gamma = 0.005$ and $p = 1$ (NARMA-10) and $\eta = 0.24$, $\gamma = 0.032$ and $p = 1$ (5-bit parity), respectively. Results are reported for 400 neurons ($\tau = 80$) on data sets of size $3000/1000$ (training/validation) in figures 8A and 8B, size $3000/1000$ in 8C (right plot), as well as for data sets of size $10000/10000$ in figure 8C (left plot). Each plot is generated from 50 different trials. The plots show median (black horizontal bar), $25^{th}/75^{th}$ percentiles (boxes), and most extreme data points not considered outliers (whiskers).

nodes into the same delay span τ. Although too small θ may affect a virtual node's computation negatively, decreasing θ increases the accuracy of the analytical approximation.

5.4.4.1 *NARMA-10*

We illustrate this by repeating the NARMA-10 task with $N = 2000$ virtual nodes and $\tau = 200$. This results in $\theta = 0.1$, corresponding to the step size used in the numerical solution before. Note that this hardly increases the computational cost of the analytical approximation since the main simulation loop along reservoir time steps $\bar{t}$ (τ-cycles) remains unchanged. The results are summarized for 50 trials in figure 8C (right boxplot). The median correlation coefficient increased significantly to nearly 0.98 while the variance across trials is notably decreased (compare fig. 8A).

5.4.4.2 *5-Bit Parity*

For the 5-bit parity task, we addressed the task complexity by increasing both, training and validation sets, to a size of 10000. Second, we increased once more the virtual network size to $N = 2000$ virtual nodes and $\tau = 200$. The performance of the resulting DCR setup, computed across 50 trials using the analytical approximation, is summarized in figure 8C (left boxplot). The model no longer suffers as much from overfitting and the performance on the validation set increased dramatically to a median value of 6.15, which is now close to the theoretical limit of 7.

5.4.5 *Application to Experimental Data*

When analyzing experimental data that is subject to measurement noise and uncertainty, a more sophisticated statistical model than presented so far is required. In this section, we embed the output of a DCR into a proper Bayesian statistical framework and discuss practical difficulties that may arise. A part of this discussion and further motivation for the statistical model can be found in the appendix. As exemplary task we chose the one-step ahead prediction of the Santa Fe laser time series, which is an experimental recording of a far-infrared laser operating in a chaotic regime [Huebner et al., 1989]. These data consist of 9000 samples as supplied in the Santa Fe time-series competition [Weigend and Gershenfeld, 1993]. Employing Bayesian model selection strategies with a Volterra series model [Rugh, 1981] (which can be thought of as a Taylor series expansion of the functional (45)), we found that the prediction of a time series sample $y(t+1)$ required at least 8 covariates $y(t), y(t-1), ..., y(t-7)$ to capture the auto-structure, and a model with 4^{th} order nonlinearities, which makes the one-step ahead prediction of the chaotic Santa Fe laser data an interesting and challenging task for a DCR.

When predicting experimental data, in addition to computing a point estimator it is also important to quantify the model confidence in some way, as a result of the potential variability of the estimator (across data sets) and one's ignorance about certain aspects of the target quantity. In a Bayesian approach, this uncertainty is summarized in a probability distribution. If the task is to predict a target time series y in terms of a covariate time series u, the uncertainty in the prediction of unseen data y_* given covariate time series u_* is summarized in a predictive distribution $P(y_*|u_*, u, y, H)$. The predictive distribution incorporates knowledge of the given data $\mathcal{D} = (y, u)$ to infer y_* given u_*. We denote by H the particular modeling assumption that has to be made and the corresponding hyperparameters. The covariance structure of this distribution represents the model uncertainty and supplies confidence intervals.

To derive a predictive distribution, first recall the DCR functional model (45),

$$\hat{y}_i := g(\bar{u}_i) = g(u_i, u_{i-1}, ..., u_{i-M}) = \sum_{j=1}^{N} \alpha_j x_{ij},$$

where x_{ij} denotes the j^{th} virtual node (56) in reservoir time step $\bar{t} = i$, and the dependence on covariates $\bar{u}_i = \{u_{i-1}, ..., u_{i-M}\}$ is given implicitly in each x_{ij} via the temporal evolution of $x(t)$ (52). Let

$$X = \begin{pmatrix} x_{11} & \dots & x_{1N} \\ \vdots & \ddots & \vdots \\ x_{n1} & \dots & x_{nN} \end{pmatrix} \in \mathbb{R}^{n \times N}, \quad \alpha \in \mathbb{R}^N, \tag{58}$$

then

$$\hat{y} = X\alpha \in \mathbb{R}^n \tag{59}$$

shall denote the estimator of target time series y, given covariate time series u.

We can now state a statistical model. First, define a multivariate isotropic normal prior distribution for α as

$$\alpha \sim \mathcal{N}(0, \lambda^2 I), \tag{60}$$

where $I \in \mathbb{R}^{N \times N}$ is the identity matrix. We denote the corresponding density by $p(\alpha|\lambda^2)$. This choice of isotropic prior results effectively in an L_2-regularized model fit [Hoerl and Kennard, 1970].

Given data $y, u \in \mathbb{R}^n$, the standard modeling assumption about the target time series y is that of noisy measurements of an underlying process $g(\bar{u}_i)$, stated as

$$y_i = g(u_i, ..., u_{i-M}) + \epsilon_i. \tag{61}$$

Here, $\forall i = 1, ..., n : \epsilon_i \sim \mathcal{N}(0, \sigma_\epsilon^2)$, and it is usually assumed that ϵ_i is independent of ϵ_j for $j \neq i$. Note that the assumption of normality is also implicitly present in any model fit given by the *least squares* solution as the minimizer of the *sum-of-squared-errors* objective (compare section 5.3). It follows that the sampling distribution for $y \in \mathbb{R}^n$ is given by

$$y \sim \mathcal{N}(\mathbf{g}(u), \sigma_\epsilon^2 I), \tag{62}$$

where $\mathbf{g}(u) = (g(\bar{u}_1), ..., g(\bar{u}_n))^T \in \mathbb{R}^n$ and $I \in \mathbb{R}^{n \times n}$. We denote the density of this distribution by $p(y|u, \alpha, \lambda^2, \sigma_\epsilon^2)$.

One can now derive all distributions necessary to formalize the uncertainty associated with data (y, u). The details of these derivations can be found, for example, in Bishop [2006]. First, Bayes formula supplies an expression for the posterior distribution that summarizes our uncertainty about α. The corresponding density is given as

$$p(\alpha|y, u, \lambda^2, \sigma_\epsilon^2) = \frac{p(y|u, \alpha, \lambda^2, \sigma_\epsilon^2) p(\alpha|\lambda^2)}{p(y|u, \lambda^2, \sigma_\epsilon^2)}, \tag{63}$$

where the normalizing constant $p(y|u, \lambda^2, \sigma_\epsilon^2)$ will be referred to as *marginal likelihood*. The posterior distribution of the model coefficients α can be explicitly computed as a normal distribution in this case. Given this posterior, one can compute the predictive distribution as a convolution of the sampling distribution (62) with the posterior of model coefficients, which will again be normal with density

$$p(y_*|u_*, u, y, \lambda^2, \sigma_\epsilon^2) = \int_{\mathbb{R}^N} p(y_*|u_*, \alpha, \sigma_\epsilon^2) p(\alpha|y, u, \lambda^2, \sigma_\epsilon^2) d\alpha. \tag{64}$$

The full distribution can be derived as

$$P(y_*|u_*, u, y, \lambda^2, \sigma_\epsilon^2) = \mathcal{N}(m_*, S_*), \tag{65}$$

with

$$\begin{aligned} m_* &:= \mathbb{E}[y_*] = X_*(X^T X + \frac{\sigma_\epsilon^2}{\lambda^2} I)^{-1} X^T y \\ S_* &:= \text{Cov}[y_*] = \sigma_\epsilon^2 I + X_*(X^T X \frac{1}{\sigma_\epsilon^2} + \frac{1}{\lambda^2} I)^{-1} X_*^T. \end{aligned} \tag{66}$$

	n=600, N=100	n=600, N=250	n=600, N=500	n=600, N=1000	n=5000, N=250
$r^2(y_t, \hat{y}_t)$	0.952	0.953	0.975	0.975	0.997
$r^2(y_t, \hat{y}_{cv})$	0.932	0.934	0.947	0.943	0.995
$r^2(y_v, \hat{y}_v)$	0.922	0.928	0.945	0.946	0.997

Table 1: Results on the Santa Fe data set: N denotes the number of virtual nodes used in the DCR model and n is the number of data points used in the training set (that is, the number of data on which the predictive distribution is conditioned). The goodness-of-fit is measured by the squared correlation coefficient $r^2 \in [0,1]$ and evaluated for an estimate $\hat{y} = m_*$ from equation (66) and the corresponding actual data y. First, the training set estimate $\hat{y}_t \in \mathbb{R}^n$ conditional on the whole training set y_t is considered, then the *leave-one-out* cross-validated estimator $\hat{y}_{cv} \in \mathbb{R}^n$ (see text above), and finally an estimate $\hat{y}_v \in \mathbb{R}^{600}$ for a validation set of size 600, conditional on the preceding training set y_t.

These terms are numerically computable. The predictive distribution (65) summarizes all uncertainty related to our forecasting of y_* and provides m_* as point estimator $\hat{y} = X_* \hat{\alpha}$, where $\hat{\alpha} = (X^T X + \frac{\sigma_\epsilon^2}{\lambda^2} I)^{-1} X^T y$ denotes the expected value of the coefficient posterior (63). From a *decision theoretic* point of view, it can also be shown that m_* minimizes the *expected* squared error loss $(y_* - \hat{y})^2$ [Berger, 1985]. For an explanation of how to derive estimators $\hat{\sigma}_\epsilon^2, \hat{\lambda}^2$ and a short discussion on how to deal with hyperparameters in a proper Bayesian model, the interested reader is referred to the appendix.

To evaluate the model, one can compute $\hat{y}_v = m_v(\hat{\sigma}_\epsilon^2, \hat{\lambda}^2)$ from (66) on a separate validation set y_v, conditional on training data (u_t, y_t). However, not only would one always want to condition the prediction on all available data, and not just on a fixed training data set y_t, but a lower number of conditional data samples in a prediction may also lead to a reduced performance and a higher susceptibility to overfitting. It may therefore seem more natural to compute the *leave-one-out* (LOO) predictive distribution $P(y_i | u_i, u_{\setminus(i)}, y_{\setminus(i)}, \hat{\sigma}_\epsilon^2, \hat{\lambda}^2)$, where $(u_{\setminus(i)}, y_{\setminus(i)})$ denotes the full data set with the i^{th} data point removed. The joint LOO predictive distribution for all data points i supplies its mean vector $\hat{y}_{cv} \in \mathbb{R}^n$ as optimal predictor for the given data y. Accordingly, this yields a cross-validated performance measure by e.g. the correlation coefficient $r(y, \hat{y}_{cv})$.

To yield a confidence interval for the correlation coefficient and somehow quantify our uncertainty in this statistic, note that all uncertainty pertaining to r is governed by the predictive distribution (65). Using the LOO predictive distribution accordingly, one can perform a parametric resampling of the data, yielding new data sets $y^* \in \mathbb{R}^n$ with mean vector $\hat{y}_{cv}$. With these, it is possible to recompute $r(\mathbb{E}[y] = \hat{y}_{cv}, y^*)$ for each resampled time series y^* to get an empirical distribution for r and quantify its expected spread on data sets of similar size. The resulting empirical distribution over r can be used to determine confidence bounds as indicators for the variability of the goodness of fit. Alternatively, although time-consuming, non-parametric bootstrapping could be employed by resampling design matrix rows (58) and used with the *BCa* method [Efron and Tibshirani, 1993] to yield confidence intervals that are corrected for deviation from normality in several ways.

Applying the above scheme to the Santa Fe laser time series, we set up an auto-regressive model with $u_i = y(i-1)$ to predict $y(i)$. We will first use only $n = 600$ samples in the training data set to illustrate the importance of formalizing uncertainty and test the DCR model with a varying number of virtual nodes. As goodness-of-fit measure the squared correlation coefficient $r^2 \in [0,1]$ will be computed either on the leave-one-out cross-validated prediction of the training data set y_t $(r^2(y_t, \hat{y}_{cv}))$, or as actual prediction of the 600 consecutive data points y_v given y_t $(r^2(y_v, \hat{y}_v))$. The data will be mean corrected and rescaled using the (5,95) percentiles. As such, the DCR parameters can be chosen again as $\gamma = 0.005, \eta = 0.4$ to maintain a proper dynamical range of the underlying Mackey-Glass system given the normalized input. Furthermore, $p = 1$ and $\tau = 100$ are set and operated with a randomly sampled binary mask m. Results will be reported for $N = 100, 250, 500, 1000$ virtual nodes respectively, leading to different uniform sampling positions $\theta \in \mathbb{R}^N$ in the τ-cycle with fixed length 100. In this section, we use equation (55) as an approximate solution scheme with step-size $k := \chi_j - \chi_{j-1} = 0.01$ on an equidistant approximation grid $\chi_0 = 0, ..., \chi_{10000} = \tau = 100$. For example, the DCR with 500 virtual nodes will have an equidistant sampling grid according to $\theta_j = \tau/500 = 0.2$. In this case, the activity of a virtual node during θ_j is computed using $\theta_j/k = 20$ samples from the underlying approximation grid.

The results on the Santa Fe data are summarized in table 1. It can be seen that all predictions conditional on merely $n = 600$ data points suffer from overfitting, as there is a noteworthy difference between predictions on the training set data and the two types of cross-validation. The models with $N = 500$ and $N = 1000$ virtual nodes appear to have no noteworthy difference in performance. In figure (9), 500 points of the validation set are shown for the DCR model with 500 virtual nodes. The predictive distribution is computed conditional on the first 600 data points of the time series and confidence intervals for the individual data points are derived as two standard deviations ($\sqrt{S_*}$) above and below the estimated expected value (m_*, see eq. (66)).

The two main characteristics of the Santa Fe time series are its irregular oscillatory behavior, as well as several sudden breakdowns of the amplitude, occurring infrequently in intervals of several hundred data points. As highlighted in the magnification of figure (9), around these rare events the confidence intervals are less accurate and don't always contain the sample. Presumably, the predictions conditional on merely $n = 600$ data points do not contain enough information to account for the rare events. In contrast, using $n = 5000$ training data may increase the accuracy of the prediction around these rare events, as more information is available to the model. As can be seen in the last column of table 1, the performance is much improved and shows no longer signs of overfitting.

Figure 10 shows in addition for each model the variability associated with the r^2 goodness-of-fit measure, as computed using the parametric resampling strategy described earlier. The $N = 500$ model shows the smallest estimated variance and it becomes obvious from this figure that an increase of virtual nodes to $N = 1000$ leads to overfitting. In general, all predictions conditional on merely $n = 600$ data

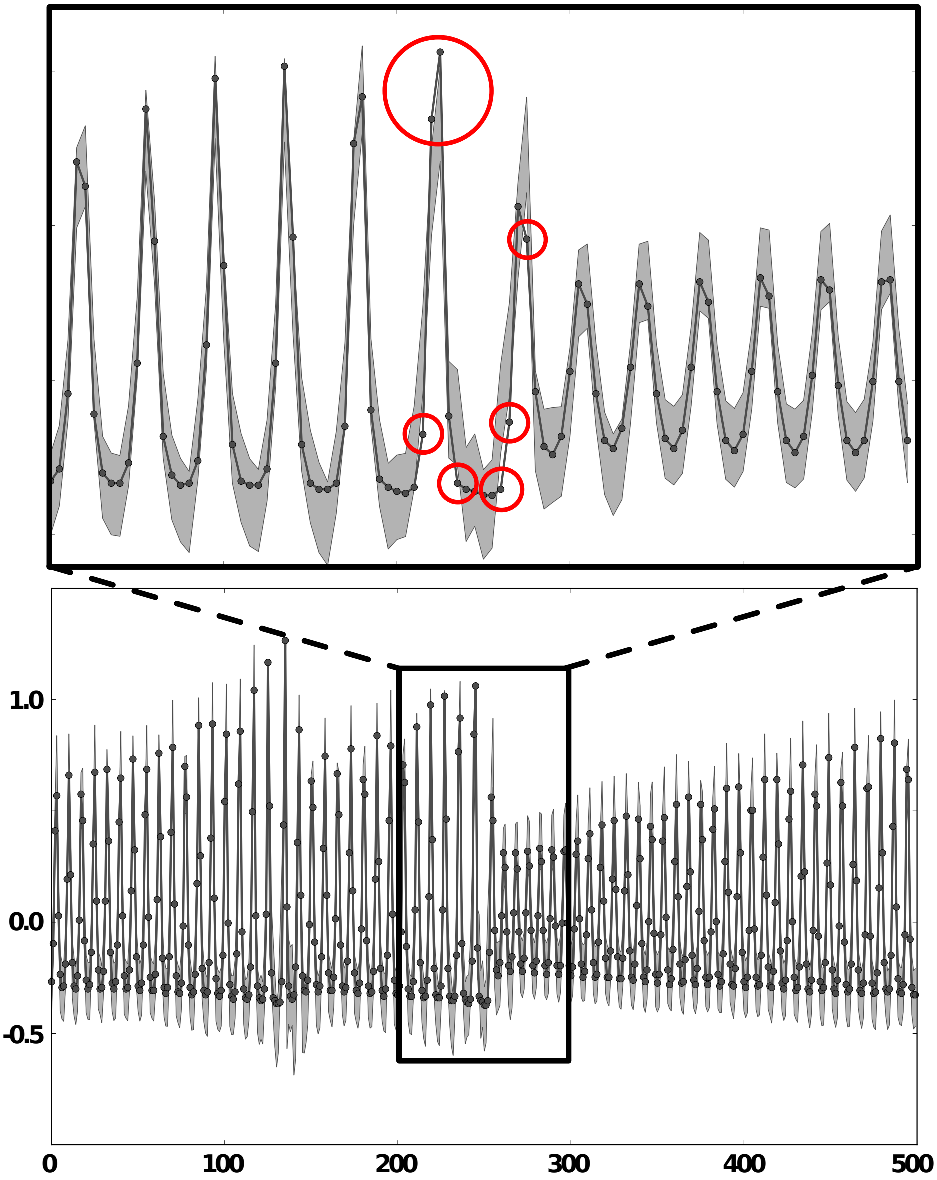

Figure 9: Normalized data points 600 to 1100 of the Santa Fe data, corresponding to a validation set. In gray, a confidence interval of 2 standard deviations of the predictive distribution is shown, as computed using a model with 500 virtual nodes, trained on the first 600 samples of the time series. It can be seen that the confidence intervals lose accuracy in a neighborhood around the sudden amplitude change after the 250^{th} data point.

points show a substantial variability in accuracy, which highlights the necessity for a proper quantification of this variability in any report of performance. In contrast, the predictions conditional on $n = 5000$ data points have a very low estimated variance and high accuracy, which suggests that the data set is in fact not very noisy. The goodness-of-fit is elevated to state-of-the-art levels with a squared correlation coefficient of $r^2(y_t, \hat{y}_{cv}) = 0.99$, as can be seen in the last figure on the right.

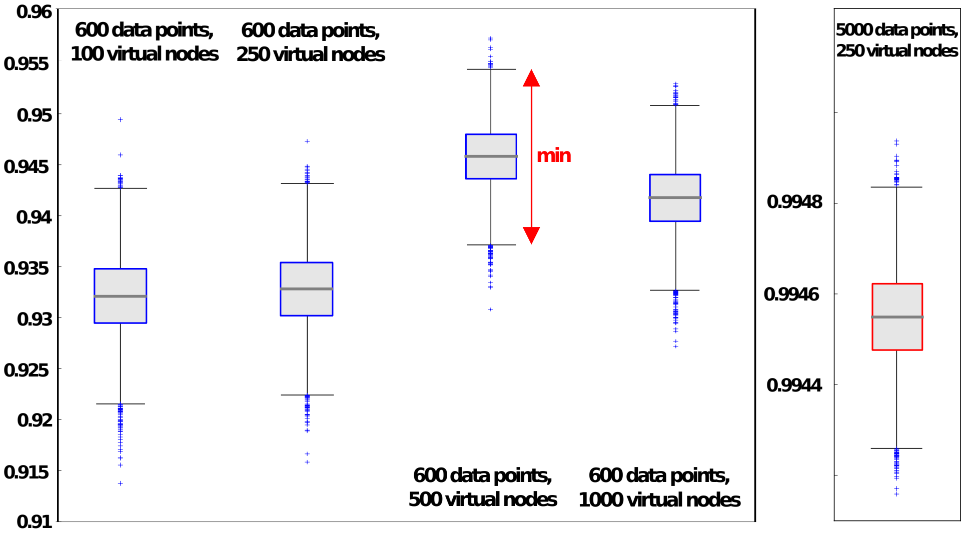

Figure 10: Squared correlation coefficient r^2 of *leave-one-out* cross-validated prediction $\hat{y}_{cv}$ with parametrically resampled training data sets y_t^* (see text). The boxes denote quartiles, whiskers 95% confidence intervals, as given by the empirical distribution upon resampling 10000 times $r^2(\hat{y}_{cv} = \mathbb{E}[y_t]_{cv}, y_t^*)$. The model with $N = 500$ virtual nodes shows the lowest variance among predictions conditional on $n = 600$ training data points.

5.5 DISCUSSION

In summary, we provided in this chapter a general introduction to the emerging field of delay-coupled reservoir-computing. To this end, a brief introduction to the theory of delay differential equations was discussed. Based on these insights, we have developed analytical approaches to evaluate and approximate solutions of delay differential equations that can be used for delay-coupled reservoir computing. In particular, we derived approximate closed-form equations for the virtual nodes of a DCR. It has been shown that the resulting update equations in principle lose neither accuracy with respect to the system dynamics nor computational power in DCR benchmark tasks. Using the analytical approximation reduces computational costs considerably. This enabled us to study larger networks of delay-coupled nodes, yielding a dramatic increase in nonlinear benchmark performance that was not accessible before. These results can lead to serious improvement regarding the implementation of DCRs on electronic boards.

Moreover, the approach yields an explicit handle on the DCR components which are otherwise implicit in equation (42). This creates new possibilities to investigate delay-coupled reservoirs and provides the basis for optimization schemes, a crucial necessity prior to any hardware implementation. Together with the reduction in computation time, this makes the use of supervised batch-update algorithms feasible to directly optimize model hyperparameters (see eq. (57) and appendix). A few of these possibilities were illustrated in a practical application to an experimental recording of a far-infrared laser operating in a chaotic regime, where the DCR model was embedded in a fully Bayesian statistical model. The relation to a potential application in Gaussian process regression is discussed in the appendix, in light of numerical difficulties that may arise with a DCR.

Future research will be focused on optimal hyperparameters of the DCR and improved adaptability to input signals. Foremost, optimal mask functions have to be identified systematically. To this end, the inconvenient coupling to the sampling grid of virtual nodes has to be overcome, so as to make an independent evaluation of mask functions possible. Accounting for and optimizing non-uniform virtual node sampling grids could be an interesting next step (compare Toutounji et al. [2012]). In addition, optimization procedures may include unsupervised gradient descent schemes on DCR parameters (e.g. θ, τ, N) with respect to other information theoretic objectives. Continuing this line of thought, one may even try to modify the update equations directly according to self-organizing homeostatic principles, inspired, for example, by neuronal plasticity mechanisms (e.g. Lazar et al. [2009]). We intend to explore these possibilities further in future work to maximize the system's computational power and render it adaptive to information content in task-specific setups.

ACKNOWLEDGEMENT

The authors acknowledge the financial support of the State Lower Saxony, Germany via the University of Osnabrück, and the European project PHOCUS in the Framework 'Information and Communication Technologies' (FP7-ICT-2009-C/Propopsal Nr. 240763). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

5.6 APPENDIX

In this section, we expand and motivate the statistical model employed in section 5.4.5. A proper statistical model allows one to treat uncertainty associated with the data in a formally optimal way. The authors believe that the greatest *theoretical* rigor in this regard is achieved with Bayesian statistics (see for example Lindley [1990]). In a first step, we therefore choose to formalize the model accordingly. In a second step, we try to implement the theory as far as possible while dealing with practical issues such as numerical accuracy, data size and computability.

Recall the DCR functional model (45),

$$\hat{y}_i := g(\bar{u}_i) = g(u_i, u_{i-1}, ..., u_{i-M}) = \sum_{j=1}^{N} \alpha_j x_{ij},$$

where x_{ij} denotes the j^{th} virtual node (56) in reservoir time step $\bar{t} = i$, and the dependence on covariates $u_{i-1}, ..., u_{i-M}$ is given implicitly in each x_{ij} via the temporal evolution of $x(t)$ (52). We chose an isotropic prior $\alpha \sim \mathcal{N}(0, \lambda^2 I)$, which corresponds effectively to an L_2 regularization. The regularization term plays an important part in safeguarding the model from overfitting. Further arguments (see Jaynes and Bretthorst [2003],Berger [1985]) suggest this prior represents our state of ignorance optimally, since (60) maximizes entropy of α given mean and variance $(0, \lambda^2)$, while being invariant under a certain set of relevant transformations. It is thus the *least informative* choice of a prior distribution, in addition to the assumptions following from the role of α in model (59).

The likelihood function (62) is also normal, as a result of the normal choice for the distribution of residuals ϵ_i. Although we may have no reason to believe that ϵ_i is actually normally distributed, one can again make strong points that the normal distribution nonetheless represents our state of knowledge optimally, unless information about higher moments of the sampling distribution is available [Jaynes and Bretthorst, 2003]. In practice, the latter will often not be the case, since there is no explicit access to the residuals ϵ_i.

From these considerations, the predictive distribution (66) can be derived analytically, as was suggested in an earlier section. Note that, equivalently, we could have derived these formulas in a framework of *Gaussian Process Regression*. The resulting expressions, though equivalent, would look slightly different. The Gaussian Process framework has many interesting advantages and allows for elegant derivations of the necessary distributions in terms of Bayesian statistics, in particular with regard to the hyperparameters $(\sigma_\epsilon^2, \lambda^2)$. It has thus become an important theoretical and practical tool in modern machine learning approaches and would have been our first choice for a statistical model. In the following, we will therefore discuss the predictive distribution in light of Gaussian processes, allude to difficulties in applying this framework to DCR models, and present a practical alternative for deriving estimates for the hyperparameters $(\sigma_\epsilon^2, \lambda^2)$.

A Gaussian Process is a system of random variables indexed by a linearly ordered set, such that any finite number of samples are jointly normally distributed [Hida and Hitsuda, 2007]. A Gaussian Process thus defines a distribution over functions and is completely specified by a *mean function* and a *covariance function*. In terms of a Gaussian process, we can define the DCR functional model $g \sim \mathcal{GP}$ as

$$\begin{aligned} \mathbb{E}[g] &= 0, \\ \mathrm{Cov}[g(\bar{u})] &= \mathbb{E}[g(\bar{u}_i)g(\bar{u}_j)] \\ &= X\mathbb{E}[\alpha\alpha^T]X^T = \lambda^2 XX^T =: [K(u,u)]_{ij}, \end{aligned} \tag{67}$$

where $[K]_{ij}$ denotes coordinate i, j of matrix K.

One can now derive all distributions necessary to formalize the uncertainty associated with data (y, u). The details of these derivations can be found, for example, in Rasmussen and Williams [2006]. The covariance matrix K_y corresponding to y is given by

$$K_y := \mathrm{Cov}[g(\bar{u}) + \epsilon] = \mathrm{Cov}[g(\bar{u})] + \mathrm{Cov}[\epsilon] = K(u,u) + \sigma_\epsilon^2 I.$$

Accordingly, for data (y_*, u_*) (which may be the same as (y, u)),

$$\begin{pmatrix} y \\ y_* \end{pmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} K(u,u) + \sigma_\epsilon^2 I & K(u,u_*) \\ K(u_*,u) & K(u_*,u_*) \end{bmatrix}\right). \tag{68}$$

If one is interested in predicting the noisy measurement of y_*, adding σ_ϵ^2 to $K(u_*, u_*)$ is appropriate and was presumed in the earlier derivations of section 5.4.5. From (68), the marginal likelihood can be derived as $y|u \sim \mathcal{N}(0, K_y)$ with log-density

$$\log[p(y|u)] = -\frac{1}{2}y^T K_y^{-1} y - \frac{1}{2}\log[|K_y|] - \frac{n}{2}\log[2\pi]. \tag{69}$$

Furthermore, the predictive distribution can be derived as

$$P(y_*|u_*,u,y) = \mathcal{N}(m_*, S_*), \tag{70}$$

with

$$\begin{aligned} m_* &:= \mathbb{E}[y_*] = K(u_*,u)K_y^{-1}y \\ S_* &:= \mathrm{Cov}[y_*] = K(u_*,u_*) - K(u_*,u)K_y^{-1}K(u,u_*). \end{aligned} \tag{71}$$

To see that the predictive distribution (66) derived earlier and (71) are in fact equivalent, consider the pseudo-inverse Φ^+ of a matrix Φ,

$$\Phi^+ = \lim_{\delta \downarrow 0}(\Phi^T\Phi + \delta I)^{-1}\Phi^T = \lim_{\delta \downarrow 0}\Phi^T(\Phi\Phi^T + \delta I)^{-1}.$$

These limits exists even if $\Phi^T\Phi$ or $\Phi\Phi^T$ are not invertible. In addition, the equality also holds for $\delta > 0$ [Albert, 1971]. This allows us to rewrite (71) as

$$\begin{aligned} m_* := \mathbb{E}[y_*] &= K(u_*,u)K_y^{-1}y \\ &= X_*X^T(XX^T + \frac{\sigma_\epsilon^2}{\lambda^2}I)^{-1}y \\ &= X_*(X^TX + \frac{\sigma_\epsilon^2}{\lambda^2}I)^{-1}X^Ty \\ S_* := \mathrm{Cov}[y_*] &= K(u_*,u_*) - K(u_*,u)K_y^{-1}K(u,u_*) \\ &= K(u_*,u_*) - X_*(X^TX + \frac{\sigma_\epsilon^2}{\lambda^2}I)^{-1}X^TXX_*^T\lambda^2 \\ &= \lambda^2X_*(I - (X^TX + \frac{\sigma_\epsilon^2}{\lambda^2}I)^{-1}X^TX)X_*^T \\ &= \lambda^2X_*(\frac{\sigma_\epsilon^2}{\lambda^2}(X^TX + \frac{\sigma_\epsilon^2}{\lambda^2}I)^{-1})X_*^T \\ &= \sigma_\epsilon^2X_*(X^TX + \frac{\sigma_\epsilon^2}{\lambda^2}I)^{-1}X_*^T. \end{aligned} \tag{72}$$

Unfortunately, K_y has deplorable numerical properties regarding inversion, as necessary in (71). This is most likely owed to the fact that the individual virtual nodes have a very high pairwise correlation across time, as can be expected of samples from a smooth system. Recall that $K(u,u) = \lambda^2X^TX \in \mathbb{R}^{n\times n}$. Since typically there will be much less virtual nodes than data points, $N < n$, so that $K(u,u)$ is usually rank deficient. Although in theory K_y should always be invertible, numerically this fact tends to hold only if $\sigma_\epsilon^2/\lambda^2 \geq 10^{-12}$. Note that this ratio corresponds to the L_2-regularization parameter. While this is a common problem in Gaussian process regression and poses for many functional models no severe difficulties, the covariance matrix built from the reservoir samples is very fickle in this regard: Setting $\sigma_\epsilon^2/\lambda^2 \geq 10^{-12}$ can already lead to severe loss of performance in non-noisy reservoir benchmark tasks such as NARMA-10. Using the standard formulation of the predictive distribution (66) allows one to sidestep these numerical issues. In both frameworks, however, the marginal likelihood in (63) is given by $y|u \sim \mathcal{N}(0, K_y)$ with log-density

$$\log[p(y|u)] = -\frac{1}{2}y^TK_y^{-1}y - \frac{1}{2}\log[|K_y|] - \frac{n}{2}\log[2\pi].$$

The marginal likelihood is important for Bayesian model selection, e.g. in computing a *Bayes Factor* or *Posterior Odds*. Given the numerical troubles discussed above, a straight-forward application of Bayesian model selection therefore seems unavailable to DCR models.

In a fully Bayesian treatment, the hyperparameters $\sigma_\epsilon^2, \lambda^2$ would have to be assigned (non-informative) priors and be integrated out of the distributions relevant for inference or model selection. However, in general it is not possible to get rid of dependence on both, hyperparameters and α. Instead, explicit values for the hyperparameters may be estimated by maximizing, for example, the marginal likelihood (69), or the predictive distribution (65) in a leave-one-out cross-validation scheme [Sundararajan and Keerthi, 2001]. With respect to computability of the involved terms, given the particular numerical difficulties arising in the DCR setup, a compromise between theory and realizability is needed. We found a good practical performance to be achieved by the following method. Looking at the marginal likelihood (69), one notes that it contains essentially a term that reflects how well the model fits the data, and another term that measures the complexity of the model as a function of K_y. A similar approach is given by the information criterion AIC_M [Konishi and Kitagawa, 2008], which can be stated as

$$\text{AIC}_M := n(\log(2\pi) + 1) + n\log(\hat{\sigma}_\epsilon^2) + 2[\text{tr}(G) + 1]. \tag{73}$$

Here, G denotes a *smoother matrix*, i.e. a matrix that is multiplied with the data y to arrive at a prediction. From the first equation in (66), which represents our optimal predictor, we can infer

$$G(\lambda^2) = X_*(X^T X + \frac{\hat{\sigma}_\epsilon^2}{\lambda^2} I)^{-1} X^T. \tag{74}$$

The term $\text{tr}(G)$ is called *effective number of parameters* and was proposed by Hastie and Tibshirani [1990] to control the complexity of a model. The model fit to the data in (73) is given by $\log(\hat{\sigma}_\epsilon^2)$ as a function of an estimator of the residual variance. Although one could estimate $\hat{\sigma}_\epsilon^2$ along with λ^2, it is usually possible to express one as a function of the other, thus simplifying the optimization problem. To account for generalization of the model with respect to prediction performance, we choose

$$\hat{\sigma}_\epsilon^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \mu^{(i)})^2. \tag{75}$$

The term $\mu^{(i)}$ denotes the predictive estimator of y_i, obtained in a *leave-one-out cross-validation* scheme by computing $m_* = m_i$ in equation (66) with the ith data point removed from X, y. This can be efficiently done in one shot for all i [Konishi and Kitagawa, 2008], yielding

$$\mu^{(i)} = \frac{[Gy]_i - [G]_{ii} y_i}{1 - [G]_{ii}}. \tag{76}$$

One can now determine

$$\hat{\lambda}^2 = \underset{\lambda^2}{\text{argmax}}\ \text{AIC}_M(\lambda^2)$$

and compute $\hat{\sigma}_\epsilon^2$ as a function of $\hat{\lambda}^2$ accordingly. In addition, the AIC$_M$ score can be used to perform a selection between models of varying complexity, for example to compare DCRs employing different numbers of virtual nodes.

6

HOMEOSTATIC PLASTICITY FOR SINGLE NODE DELAY-COUPLED RESERVOIR COMPUTING

6.1 ABSTRACT

Supplementing a differential equation with delays results in an infinite dimensional dynamical system. This property provides the basis for a reservoir computing architecture, where the recurrent neural network is replaced by a single nonlinear node, delay-coupled to itself. Instead of the spatial topology of a network, subunits in the delay-coupled reservoir are multiplexed in time along one delay span of the system. The computational power of the reservoir is contingent on this temporal multiplexing. Here, we learn optimal temporal multiplexing by means of a biologically-inspired homeostatic plasticity mechanism. Plasticity acts locally and changes the distances between the subunits along the delay, depending on how responsive these subunits are to the input. After analytically deriving the learning mechanism, we illustrate its role in improving the reservoir's computational power. To this end, we investigate, firstly, the increase of the reservoir's memory capacity. Secondly, we predict a NARMA-10 time series, showing that plasticity reduces the normalized root-mean-square error by more than 20%. Thirdly, we discuss plasticity's influence on the reservoir's input-information capacity, the coupling strength between subunits, and the distribution of the readout coefficients.

6.2 INTRODUCTION

***R**eservoir **c**omputing*, or RC for short [Jaeger, 2001; Maass et al., 2002; Buonomano and Maass, 2009; Lukoševičius and Jaeger, 2009], is a computational paradigm that provides both a model for neural information processing [Häusler and Maass, 2007; Karmarkar and Buonomano, 2007; Yamazaki and Tanaka, 2007; Nikolić et al., 2009], and powerful tools to carry out a variety of spatiotemporal computations. This includes time series forecasting [Jaeger and Haas, 2004], signal generation [Jaeger et al., 2007], pattern recognition [Verstraeten et al., 2006], and information storage [Pascanu and Jaeger, 2011]. RC also affords a framework for advancing and refining our understanding of neuronal plasticity and self-organization in recurrent neural networks [Lazar et al., 2007, 2009; Toutounji and Pipa, 2014].

This article presents a biologically inspired neuronal plasticity rule to boost the computational power of a novel RC architecture that is called a *single node **d**elay-**c**oupled **r**eservoir*, or DCR for short. The DCR realizes the same RC concepts using a single nonlinear node with *delayed feedback* [Appeltant et al., 2011]. This simplicity makes the DCR particularly appealing for physical implementations, which has already been demonstrated on electronic [Appeltant et al., 2011], optoelectronic [Larger et al., 2012a; Paquot et al., 2012], and all-optical hardware [Brunner et al., 2013]. The optoelectronic and all-optical implementations utilizes a semiconductor laser diode as the nonlinear node, and an optical fiber as a delay line, allowing them to maintain high sampling rates. They are also shown to compare in performance to standard RC architectures in benchmark computational tasks.

The DCR operates as follows. Different nonlinear transformations and mixing of stimuli from the past and the present are achieved by sampling the DCR's activity at *virtual nodes*, or *v-nodes*, along the delay line. While neurons of a recurrent network are mixing stimuli via their synaptic coupling, which forms a network topology, the v-nodes of a DCR are mixing signals via their (nonlinear) temporal interdependence. Therefore, the v-nodes' temporal distances from one-another, henceforth termed *v-delays*, are made shorter than the characteristic time scale of the nonlinear node. Thus, v-nodes become analogous to the connections of a recurrent network, providing the DCR with a certain network-like topology. In analogy to the *spatial distribution* of input in a classical reservoir, stimuli in a DCR are *temporally multiplexed* (see Figure 11). To process information, the external stimuli are applied to the dynamical system, thereby perturbing the reservoir dynamics. Here, we operate the DCR in an asymptotically stable fixed point regime. To render the response of the DCR transient, i.e., reflecting nonlinear combinations of past and present inputs, the reservoir dynamics must not converge to the fixed point, where it becomes dominated by the current stimulus. To ensure this, a random piecewise constant masking sequence is applied to the stimulus before injecting the latter to the reservoir [Appeltant et al., 2011]. The positions where this *mask* may switch value match the positions of the v-nodes, which are initially chosen *equidistant*. However, given the fact that the v-delays directly influence the interdependence of the corresponding v-nodes states, and therefore the nonlinear mixing of the stimuli, it is immediately evident that v-delays are important parameters that may significantly influence the performance of the DCR.

To optimize the computational properties of the DCR, we employ neuroscientific principles using biologically-inspired *homeostatic plasticity* [Davis and Goodman, 1998; Zhang and Linden, 2003; Turrigiano and Nelson, 2004] for adjusting the v-delays. Biologically speaking, homeostatic plasticity does not refer to a single particular process. It is rather a generic term for a family of adaptation mechanisms that regulate different components of the neural machinery, bringing these components to a functionally desirable operating regime. The choice of the operating regime depends on the functionality a model of homeostatic plasticity aims to achieve. This resulted in many flavors of homeostatic plasticity for regulating recurrent neural networks in computational neuroscience [Somers et al., 1995; Soto-Treviño et al., 2001; Renart et al., 2003; Lazar et al., 2007, 2009; Marković and Gros, 2012; Remme and Wadman, 2012; Naudé et al., 2013; Zheng et al., 2013; Toutounji and Pipa, 2014], neurorobotics [Williams and Noble, 2007; Vargas et al., 2009; Hoinville et al., 2011; Dasgupta et al., 2013; Toutounji and Pasemann, 2014], and reservoir computing [Schrauwen et al., 2008b; Dasgupta et al., 2013]. Here, we use a homeostatic plasticity mechanism to regulate the v-delays so as to balance responsiveness to the input and its history on the one hand, against optimal expansion of its informational features into the high dimensional phase space of the system on the other hand. Furthermore, we show that this process can be understood as a competition between the v-nodes' *sensitivity* and their *entropy*, resulting in a functional specialization of the v-nodes. This leads to a high increase in the DCR's memory capacity, and to a significant improvement in its ability to carry out nonlinear spatiotemporal computations. We discuss the implications of the plasticity mechanism with respect to the DCR's entropy, as well as the virtual network topology, and the resulting regression coefficients.

6.3 MODEL

In this section, we describe the RC architecture that is based on a single nonlinear node with delayed feedback. We then formulate this architecture using concepts from neural networks.

6.3.1 *Single Node Delay-Coupled Reservoir*

Generally speaking, RC comprises a set of models where a large dynamical system called a reservoir, a recurrent neural network for example, nonlinearly maps a set of varying stimuli to a high-dimensional space [Jaeger, 2001; Maass et al., 2002]. The recurrency allows a damped trace of the stimuli to travel within the reservoir for a certain period of time. This phenomenon is termed *fading memory* [Boyd and Chua, 1985]. Then, random nonlinear motifs within the reservoir nonlinearly mix past and present inputs, allowing a desired output to be *linearly* combined from the activity of the reservoir using a linear regression operation. As the desired output is usually a particular transformation of the temporal and spatial aspects of the stimuli, the operations that a RC architecture are trained to carry out are termed *spatiotemporal computations*.

In a classical RC architecture, past and present inputs $\delta \in \mathbb{R}^m$ undergo nonlinear mixing via injection into a ***r**ecurrent **n**eural **n**etwork* (RNN) of n nonlinear units. This spatial distribution of the input is a mapping $f : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}^n$. The dynamics are modeled by a difference equation for discrete time

$$x(t+1) = f\big(x(t), \delta(t)\big), \tag{77}$$

or an ***o**rdinary **d**ifferential **e**quation* (ODE) for continuous time

$$\dot{x}(t) = f\big(x(t), \delta(t)\big), \tag{78}$$

where $x(t) \in \mathbb{R}^n$ is the network activity, and $\dot{x}(t)$ the activity's time derivative.

In a *single node **d**elay-**c**oupled **r**eservoir* (DCR), the recurrent neural network is replaced by a single nonlinear node with delayed feedback. Formally, the dynamics can be modeled by a forced (or driven) ***d**elay **d**ifferential **e**quation* (DDE) of the form

$$\dot{x}(t) = -x(t) + f\big(x(t-\tau), \delta(t)\big) \tag{79}$$

where τ is the delay time, and $x(t), x(t-\tau) \in \mathbb{R}$ are the current and delayed DCR activities. Figure 11 illustrates the DCR architecture and compares it to the standard RNN approach to reservoir computing.

Solving the system (79) for $t \geq 0$ requires specifying an appropriate initial value function $\phi_0 : [-\tau, 0] \to \mathbb{R}$. As is already suggested by the initial conditions,

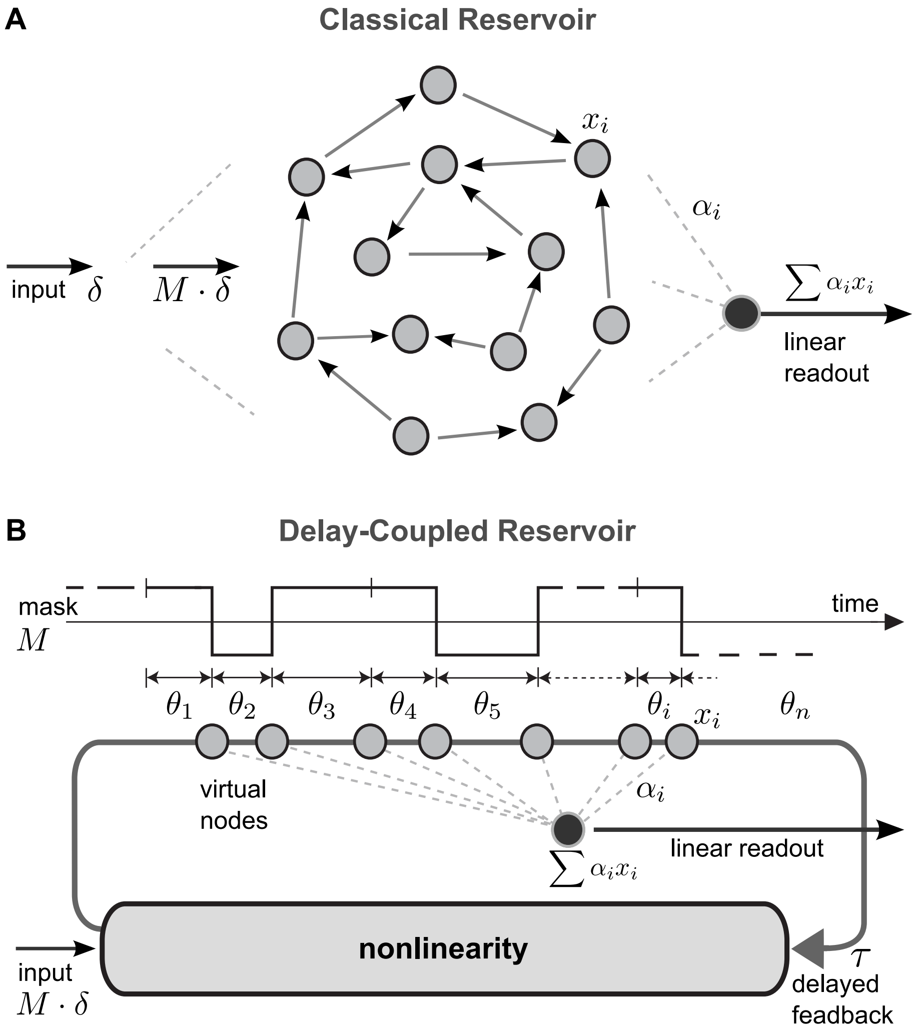

Figure 11: Comparing classical and single node delay-coupled reservoir computing architectures. **(A)** A classical RC architecture. The input δ is *spatially distributed* by input weights M to a RNN of size n. The activity of the RNN is then linearly readout. **(B)** A single node delay-coupled reservoir. The input δ is *temporally multiplexed* across a delay line of length τ by using a random binary mask M of n bits. Each mask bit M_i is held constant for a short v-delay θ_i such that the sum of these delays is the length of the delay line τ. The masked input is then nonlinearly transformed and mixed with past input by a nonlinear node with delayed feedback. At the end of each v-delay θ_i resides a v-node from which linear readouts learn to extract information and perform spatiotemporal computations through linear regression.

the phase space of system (79) is a *Banach space* $C_{1,\tau} = C\big([-\tau,0],\mathbb{R}\big) \ni \phi_0$ which is *infinite dimensional* [Guo and Wu, 2013a]. Using a DDE as a reservoir, this phase space thus provides a high-dimensional feature expansion for the input signal, which is usually achieved by using a RNN with more neurons than input channels.

To inject a signal into the reservoir, it is multiplexed in time: The DCR receives a single constant input $u(\bar{t}) \in \mathbb{R}^m$ in each reservoir time step $\bar{t} = \lceil \frac{t}{\tau} \rceil$, corresponding to one τ-cycle of the system. During each τ-cycle, the input is again linearly transformed by a mask $M \in [0, \tau]^m$ that is piecewise constant for short periods θ_i, representing the temporal spacing, or *v-delays*, between sampling points of $i = 1, \ldots, n$ virtual nodes, or *v-nodes*, along the delay line. Accordingly, the v-delays satisfy $\sum_{i=1}^{n} \theta_i = \tau$, where n is the effective dimensionality of the DCR. Here, the mask M is chosen to be binary with random mask bits $M_i \in \{-\mu, +\mu\}^m$, so that the v-node i receives a weighted input $M_i u(\bar{t})$. In order to assure that the DCR possesses fading memory of the input, the system (79) is set to operate in a regime governed by a single fixed point in case the input is constant. Thus, the masking procedure effectively prevents the driven dynamics of the underlying system from saturating to the fixed point.

A sample is read out at the end of each θ_i, yielding n predictor variables (v-nodes) $x_i(\bar{t})$ per time step $\bar{t}$. Computations are performed on the predictors using a linear regression model for some scalar target time series y, given by $\hat{y}(\bar{t}) = \sum_{i=1}^{n} \alpha_i x_i(\bar{t})$ where x_i with $i = 1, \ldots, n$ denote the DCR's v-nodes (see equation (82)), and α_i are the coefficients determined by regression, e.g. using the *least squares solution* minimizing the sum of squared errors $\sum_{\bar{t}} \big(y(\bar{t}) - \hat{y}(\bar{t})\big)^2$.

In what follows, our model of choice for the DCR nonlinearity is an input-driven *Mackey-Glass* system [Glass and Mackey, 2010] that is operating, when not driven by input, at a fixed point regime:

$$\dot{x}(t) = -x(t) + \frac{\eta\big(x(t-\tau) + \gamma M \delta(t)\big)}{1 + \big(x(t-\tau) + \gamma M \delta(t)\big)} \tag{80}$$

where γ and η are model parameters. In addition to favorable analytical properties that are to be stated in turn, the current choice of nonlinearity is motivated by the superior performance it achieves on spatiotemporal computations. It can also be approximated by electronic circuits [Appeltant et al., 2011]. Figure 12A shows the response of the DCR governed by equation (80) to a single-channel input.

6.3.2 *The DCR as a Virtual Network*

The goal is to optimize the computational properties of the DCR as a network, given a vector of v-delays $\Theta = (\theta_1, \ldots, \theta_i, \ldots, \theta_n)$ of its n v-nodes. In the case of equidistant v-delays, approximate v-node equations were already derived by [Appeltant et al., 2011], who also conceptualized the DCR with equidistant v-delays as a network. We extend this result to account for arbitrary v-node spacings on which our plasticity rule can operate. To that end, we first need to define the activity $x(t)$ of the DCR given θ_i for $i = 1, \ldots, n$.

First, we solve the DDE (79) by applying the *method of steps* (see Appendix A for details on solving and simulating the DCR). If the system (79) is evaluated at $(\nu - 1)\tau \leq t \leq \nu\tau$, where a continuous function $\phi_\nu \in C_{[(\nu-2)\tau,(\nu-1)\tau]}$ is the solution for $x(t)$ on the previous τ-interval, we can replace $x(t - \tau)$ by $\phi_\nu(t - \tau)$.

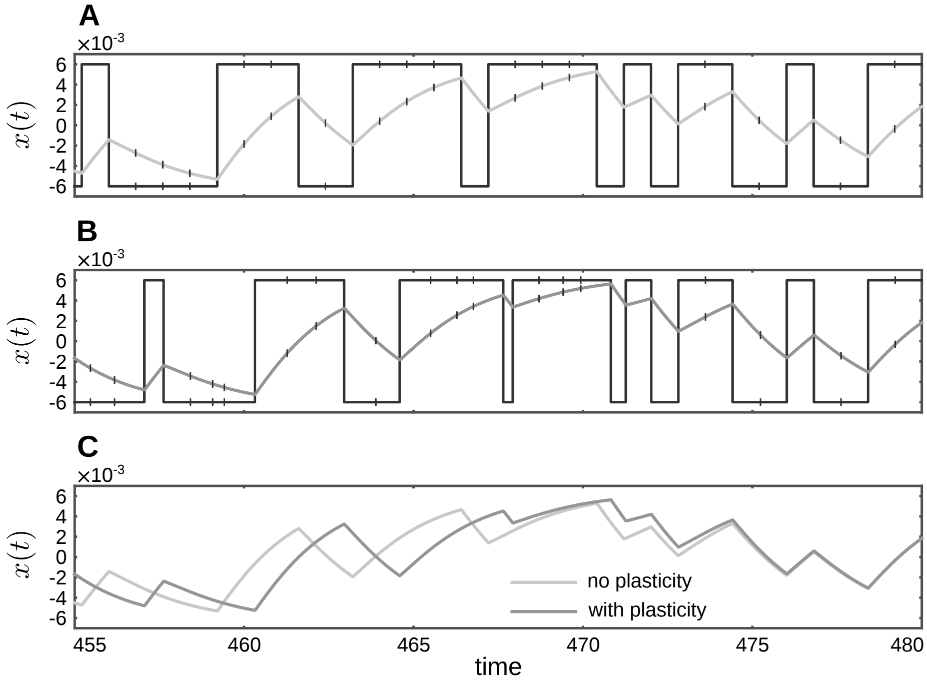

Figure 12: DCR activity superimposed on the corresponding mask **(A)** before and **(B)** after plasticity. **(C)** Comparison between the DCR's activity before and after plasticity.

Consequently, the solution to (79) subject to $x\big((\nu-1)\tau\big) = \phi_\nu\big((\nu-1)\tau\big)$ is given by

$$\begin{aligned} x(t) = {} & \phi_\nu\big((\nu-1)\tau\big)e^{(\nu-1)\tau-t} \\ & + e^{(\nu-1)\tau-t}\int_{(\nu-2)\tau}^{t-\tau} f\big(\phi_\nu(s),\delta(s)\big)e^{s-(\nu-2)\tau}ds. \end{aligned} \tag{81}$$

Let the the DCR activity at a particular v-node $x_i(\bar{t}) = x\big((\nu-1)\tau + \sum_{j=1}^{i}\theta_j\big)$, its nonlinearity $f_i(\bar{t}) = f\big(x_i(\bar{t}-1), M_i \cdot u(\bar{t})\big)$, and the DCR time step $\bar{t} = \lceil \frac{t}{\tau} \rceil = \nu$. As shown in Appendix A, the solution mapping (81) to the DCR can be approximated by assuming $f(\cdot)$ to be piecewise constant at each θ_i. This is a valid approximation since $\theta_i \ll \tau$, and it yields the following expression of the DCR activity at a v-node i as a function of $\{\theta_1, \ldots, \theta_i\}$:

$$x_i(\bar{t}) = e^{-\sum_{j=1}^{i}\theta_j}x_n(\bar{t}-1) + \sum_{j=1}^{i}(1-e^{-\theta_j})e^{-\sum_{k=j+1}^{i}\theta_k} \cdot f_j(\bar{t}) \tag{82}$$

Equation (82) suggests that the activity of v-node i is a weighted sum of the nonlinear component of the preceding v-nodes' activity, down to the last v-node n in the cyclic network, the activity of which is carried over from the previous reservoir time step. The first conceptualization of this type of weight matrix for equidistant v-nodes The resulting directed network topology is shown as a virtual weight matrix for equidistant v-nodes ($\theta_i = \tau/n$) in Figure 13A.

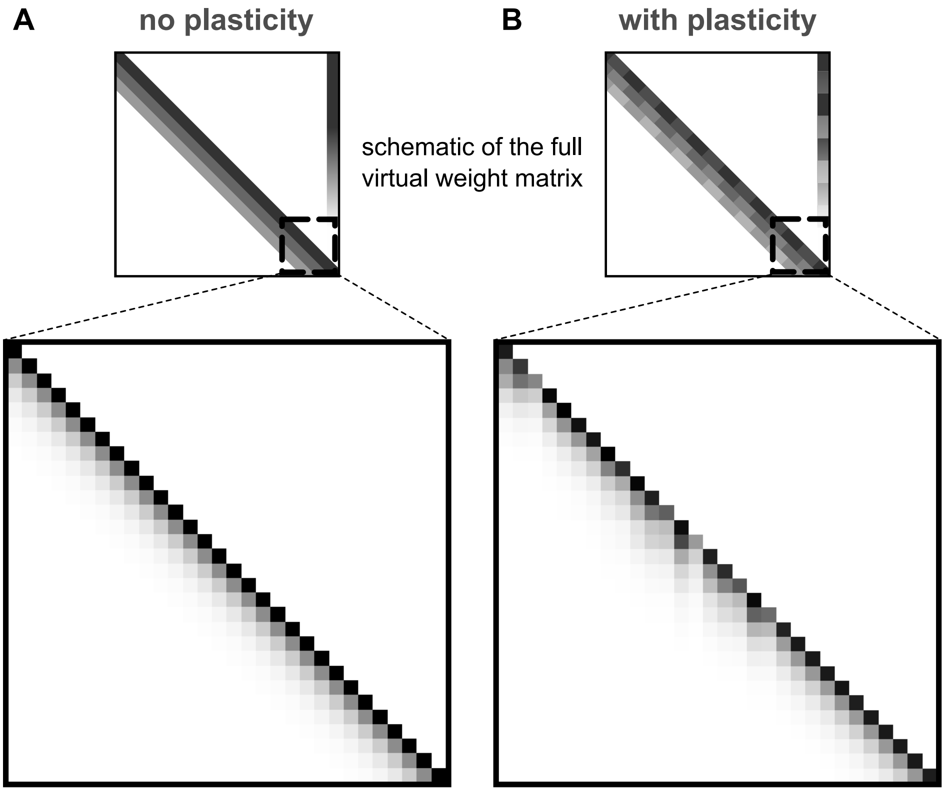

Figure 13: Virtual weight matrix of a DCR **(A)** before and **(B)** after plasticity. The magnified section corresponds roughly to connectivity within part of the delay span τ shown in Figure 12.

6.4 PLASTICITY

An important role of the randomly alternating mask M is to prevent the DCR dynamics from saturating and thus losing history dependence and sensitivity to input. However, the random choice of the mask values and the equal v-delays do not guarantee an optimal choice of masking. A simple example which already illustrates this point is given by the occurrence of sequences of equal valued mask bits, as shown in Figure 12A, which leads to unwanted saturation. In general, many more factors exist that determine optimal computation in the reservoir and that need balancing.

Our goal in this section therefore is to develop a plasticity mechanism that optimizes the resulting v-delays with respect to *sensitivity*, while retaining a suitable nonlinear feature expansion into the DCR's phase space. As will be shown in Section 6.6.1, this results in a trade-off between sensitivity and *entropy* of the v-nodes. Entropy and sensitivity counteract each other, thus forcing v-nodes to specialize. In a first step (Section 6.4.1), we develop a partial plasticity mechanism that maximizes solely the *sensitivity* of individual v-nodes. In a second step (Section 6.4.2), the mechanism will be augmented by a counteracting regulatory term which tries to retain diverse feature expansion of the input. The delay τ together with the number n of v-nodes the mask M, and the parameters γ and η of the delayed nonlinearity are treated as fixed and given hyperparameters which determine the particular DCR that is the subject of the optimization process.

6.4.1 *Sensitivity Maximization*

We measure a v-node's sensitivity by the slope of its activity at the readout point, i.e., the end point of the θ_i interval, where bigger slope corresponds to less saturation. The objective is to maximize the overall sensitivity of the DCR for all v-nodes simultaneously. First, we use the approximate solution mapping of a v-node's dynamics from equation (82) to derive a formula of a v-node's activity as a function of the v-delay θ_i from the previous v-node alone:

$$x_i(\bar{t}) = e^{-\theta_i} x_{i-1}(\bar{t}) + (1 - e^{-\theta_i}) f_i(\bar{t}), \qquad i = 2, \ldots, n \tag{83}$$

$$x_1(\bar{t}) = e^{-\theta_1} x_n(\bar{t} - 1) + (1 - e^{-\theta_i}) f_n(\bar{t} - 1). \tag{84}$$

In addition, the dynamics of the DCR at a particular v-node i in units of reservoir time steps $\bar{t}$ is given by

$$\dot{x}_i(\bar{t}) = -x_i(\bar{t}) + f_i(\bar{t}). \tag{85}$$

Substituting equation (83) into (85) yields the following expression for the sensitivity of a v-node i as a function of θ_i:

$$S_i(\bar{t}) = \dot{x}_i(\bar{t}) = \big(-x_{i-1}(\bar{t}) + f_i(\bar{t})\big) e^{-\theta_i} \tag{86}$$

From equation (86), we define a sensitivity vector $\mathbf{S} \in \mathbb{R}^n$. To optimize the overall sensitivity of the DCR, we maximize an objective function under the constraint that the sum of the v-delays stays equal to the overall delay τ:

$$\arg\max_{\Theta \geq 0} \big\{ \|\mathbf{S}\|_2^2 \big\} \quad \text{subject to} \quad \sum_{i=1}^{n} \theta_i = \tau, \tag{87}$$

where $\|\cdot\|_2$ is the Euclidean norm.

To find the vector Θ that solves the constrained optimization problem (87), we follow the direction of the steepest ascent which is the gradient of the objective function, and we project the outcome to the simplex $\sum_{i=1}^{n} \theta_i = \tau$. The element-wise gradient is given by:

$$\nabla_i \|\mathbf{S}\|_2^2 = \frac{\partial \|\mathbf{S}\|_2^2}{\partial \theta_i} \tag{88}$$

By iteratively inserting expression (83) into the sensitivity formula (86), and eliminating the iteration with (84), we can show that the sensitivity of a v-node i depends on the v-delays θ_j of all the preceding v-nodes $j \leq i$:

$$S_i(\bar{t}) = e^{-\sum_{k=j+1}^{i} \theta_k} \cdot S_j(\bar{t}) + \Gamma(\theta_{j+1}, \cdots, \theta_i), \tag{89}$$

where $\Gamma(\cdot)$ is a term independent of θ_j. However, since the term $e^{-\sum_{k=j+1}^{i} \theta_k}$ decays exponentially the further the v-node i is from the v-node j, one can ignore the contribution of θ_j to the sensitivity of the v-node i for $i > j$. This simplifies the element-wise gradient to

$$\begin{aligned}\nabla_i \|\mathbf{S}\|_2^2 &= \frac{\partial S_i^2}{\partial \theta_i} \\ &= -2\big(-x_{i-1}(\bar{t}\,) + f_i(\bar{t}\,)\big)^2 e^{-2\theta_i}. \end{aligned} \tag{90}$$

6.4.2 *Homeostatic Plasticity*

The optimization problem (87) maximizes the sensitivity of a v-node i by decreasing θ_i, its temporal distance from the previous v-node, as is suggested by the element-wise gradient (90). As a result, the v-node becomes more sensitive to the input history delivered from its predecessor. This however leads to a loss of diversity in expanding informational features of the present input, since the smaller the time alloted to a v-node is, the less excitable by present input it becomes. In addition, the optimization objective prefers small θ_i, many of which may even go to 0, despite the constraint $\sum_{i=1}^{n} \theta_i = \tau$, which leads to a reduction of the reservoir's effective dimensionality.

We hypothesize that good spatiotemporal computational performance is achieved when diversity and sensitivity are balanced. To this end, we introduce a regulatory term into the sensitivity measure that punishes small v-delays, thus counteracting sensitivity by enforcing an increase in a v-node's distance from its predecessor. The choice of the regulatory term is motivated by favorable analytical properties (mentioned later in the current section), and by allowing flexibility in the choice of regulation between diversity and sensitivity. As *entropy* is a natural measure of informational diversity, we later support the current intuitions behind our choice of the regulatory term by a rigorous mathematical argumentation. Namely, we show in Section 6.6.1 how a plasticity mechanism that solely maximizes entropy of the v-nodes leads to an unbounded increase of v-delays and therefore presents a proper counteract to sensitivity.

The sensitivity measure with regulatory term has the form

$$S_{i\rho}(\bar{t}\,) = \theta_i^{\rho} \cdot \dot{x}_i(\bar{t}\,) = \theta_i^{\rho}\big(-x_{i-1}(\bar{t}\,) + f_i(\bar{t}\,)\big)e^{-\theta_i}, \tag{91}$$

where $\rho > 0$ is a *regulating parameter* that modulates the penalty afflicted on the decrease in θ_i. Lower ρ leads the objective to favor smaller v-delays and vice versa.

From equation (91), we define a homeostatic sensitivity vector $\mathbf{S}_\rho \in \mathbb{R}^n$, and an optimization problem

$$\underset{\Theta \geq 0}{\arg\max} \big\{ \mathcal{O}(\Theta) = \|\mathbf{S}_\rho\|_2^2 \big\} \quad \text{subject to} \quad \sum_{i=1}^{n} \theta_i = \tau, \tag{92}$$

and we maximize $\mathcal{O}$ by following the direction of the steepest ascent. Since the contribution of θ_j to the sensitivity of a v-node i for $i > j$ is ignorable, following the argumentation from equation (89), the element-wise gradient is simplified to

$$\begin{aligned}\nabla_i \, \mathcal{O}(\Theta) &= \frac{\partial S_{i\rho}^2}{\partial \theta_i} \\ &= -2\theta_i^{2\rho-1}(\theta_i - \rho)\big(-x_{i-1}(\bar{t}\,) + f_i(\bar{t}\,)\big)^2 e^{-2\theta_i}. \end{aligned} \tag{93}$$

Defining a v-node's scaling factor $\sigma_i(\bar{t}) = \big(-x_{i-1}(\bar{t}) + f_i(\bar{t})\big)^2 \geq 0$, the maximized function $S^2_{i\rho} = \sigma_i \theta_i^{2\rho} e^{-2\theta_i}$ is *unimodal*, which entails the existence of a *global maximum* $\theta_i = \rho$, despite $S^2_{i\rho}$ not being *convex* (For nonnegative v-delays, $S^2_{i\rho}$ has one inflection point when $\rho \leq 0.5$, and two inflection points otherwise). This assures convergence to the global maximum of the unconstrained optimization problem. the homeostatic plasticity learning rule for a single v-node i then reads

$$\theta_i(\bar{t}+1) = \theta_i(\bar{t}) - 2\alpha\sigma_i(\bar{t})\big(\theta_i(\bar{t}) - \rho\big)\theta_i^{2\rho-1}(\bar{t})e^{-2\theta_i(\bar{t})}, \tag{94}$$

where the term $\big(\theta_i(\bar{t}) - \rho\big)$ *homeostatically* balances between the v-delay's increase and decrease, depending on the choice of the regulating term ρ.

Given the above, the update rule of the vector $\Theta = (\theta_1, \cdots, \theta_n)$ is given by

$$\Theta \leftarrow \pi_V\big(\Theta + \alpha \cdot \mathcal{J}_{\mathcal{O}}(\Theta)\big), \tag{95}$$

where α is a scalar learning rate, $\mathcal{J}_{\mathcal{O}} = \nabla\, \mathcal{O}(\Theta)$ is the Jacobian matrix of $\mathcal{O}$ with respect to Θ, and π_V is an orthogonal projection which assures that Θ remains on the constraint simplex V defined by $\sum_{i=1}^n \theta_i = \tau$ (see Appendix B for details of the constraint satisfaction). The global maximum belongs to V, only when $\rho = \tau/n$, which leads to the convergence to equidistant v-nodes. Otherwise, the constrained gradient leads to the point on V, closest to the global maximum.

6.5 COMPUTATIONAL PERFORMANCE

In the following, we test the effect of the homeostatic plasticity mechanism (94) on the performance of the DCR. Simulations are carried on 100 DCRs, the activity of each is sampled at 600 v-nodes that are initially equidistant with $\theta = 0.8$. Each DCR is completely distinguishable from the other by its binary mask M, and the 600 mask values are randomly chosen from the set $\{-0.1, +0.1\}$. Simulation starts with a short initial period for stabilizing the dynamics, followed by a plasticity phase of $n_p = 500$ time steps, each corresponding to one τ. The learning rate α is set to 0.01 and the regulating parameter ρ to 1.0. Afterwards, readouts are trained on $n_t = 5000$ samples for both the original and modified v-delays θ_i, and validated on another $n_v = 1000$ samples. The model parameters of the Mackey-Glass nonlinearity (see equation (80)) are set to $\gamma = 0.05$ and $\eta = 0.4$. The DCR is subject to uniformly distributed scalar input $u(\bar{t}) \sim \mathcal{U}_{[0,0.5]}$. At this positive input range, the DCR dynamics resulting from the Mackey-Glass nonlinearity is *saturating*, as illustrated in Figure 12A. This condition assures that the approximation (82) is accurate enough, such that a decrease in a v-delay does increase a v-node's sensitivity.

Given a task-dependent target time series y and a linear regression estimate $\hat{y}(\bar{t}) = \sum_{i=1}^n \alpha_i x_i(\bar{t})$ (x_i being the DCR's v-nodes response to the input u), we measure the performance using the *normalized root-mean-square error*

$$\text{nrmse}(y, \hat{y}) = \sqrt{\frac{\sum_{n_v}(y-\hat{y})^2}{n_v \text{var}(y)}}. \tag{96}$$

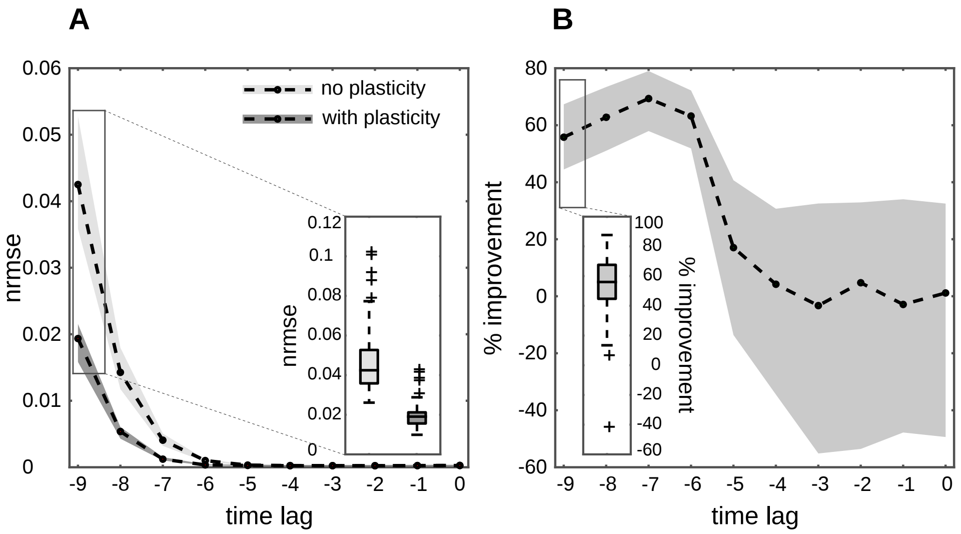

Figure 14: Memory capacity before and after plasticity. **(A)** Performance on memory construction before and after plasticity for different time lags. The inset shows performance on memory construction ten time steps in the past ($\ell = -9$) before and after plasticity. **(B)** Relative improvement, measured by the decrease in nrmse, after applying homeostatic plasticity. The inset shows the improvement on memory construction ten time steps in the past ($\ell = -9$). **(A,B)** The dotted lines are the medians of the corresponding plots, while the shaded areas mark the first and third quartiles. In addition to marking the quartiles, the insets show whiskers that extend to include data points within 1.5 times the interquartile range (the difference between the third and first quartiles). The crosses specify data points outside this range and correspond to outliers.

6.5.1 *Memory Capacity*

The memory capacity of a reservoir is a measure of its ability to retain in its activity a trace of its input history. Optimal linear classifiers are trained for reconstructing the uniformly distributed scalar input $u(\bar{t}) \sim \mathcal{U}_{[0,0.5]}$ at different time lags ℓ. Figure 14 compares the memory capacity of DCRs before and after plasticity. For time lags $|\ell| > 5$, where the ability to reconstruct the input history starts to diverge from optimal (see Figure 14A), the increase of the DCR's memory capacity can reach up to 70%. The improvement is measured as the relative change in nrmse at each time lag, due to plasticity. Only one of the 100 DCRs showed $\sim$ 20% deterioration in memory capacity after plasticity for the largest time lag (see inset in Figure 14B).

6.5.2 *Nonlinear Spatiotemporal Computations*

A widely used benchmark in reservoir computing is the capacity to model a nonlinear autoregressive moving average system y in response to the uniformly dis-

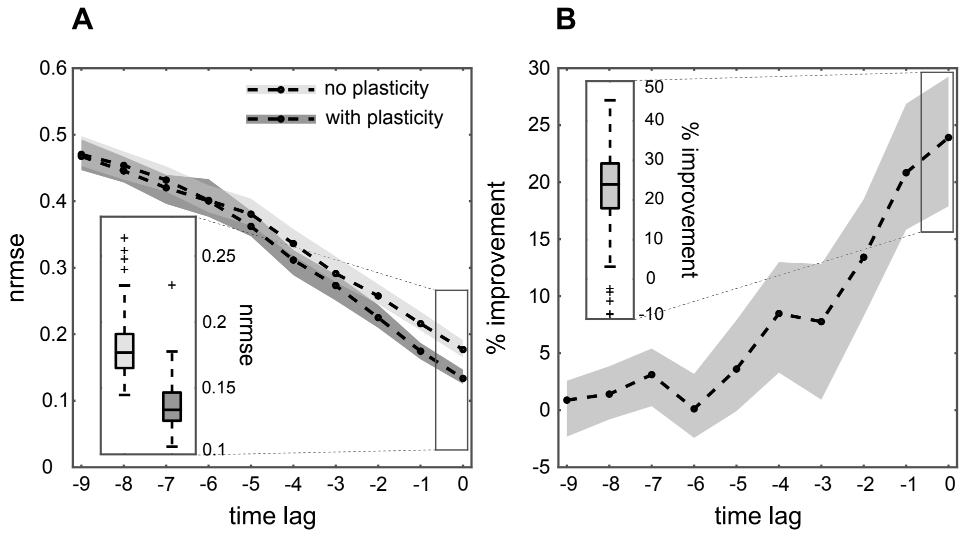

Figure 15: Spatiotemporal computational power before and after plasticity. **(A)** Performance on the NARMA-10 task before and after plasticity for different time lags. The inset shows performance at zero time lag before and after plasticity. **(B)** Relative improvement, measured by the decrease in nrmse, after applying homeostatic plasticity. The inset shows the improvement at zero time lag. **(A,B)** The dotted lines are the medians of the corresponding plots, while the shaded areas mark the first and third quartiles. In addition to marking the quartiles, the insets show whiskers that extend to include data points within 1.5 times the interquartile range. The crosses specify data points outside this range and correspond to outliers.

tributed scalar input $u(\bar{t}) \sim \mathcal{U}_{[0,0.5]}$. The NARMA-10 task requires the DCR to compute at each time step $\bar{t}$ a response

$$y(\bar{t}) = 0.3y(\bar{t}-1) + 0.05y(\bar{t}-1)\sum_{\bar{s}=1}^{10} y(\bar{t}-\bar{s}) + 1.5u(\bar{t}-1)u(\bar{t}-10) + 0.1. \tag{97}$$

Thus, NARMA-10 requires modeling quadratic nonlinearities and shows a strong history dependence that challenges the DCR's memory capacity. Figure 15 compares the performance in nrmse of DCRs before and after plasticity for different time lags. Even with no time lag $|\ell| = 0$, the task still requires the DCR to retain fading memory. This is in order to account for the dependence on inputs and outputs 10 time steps in the past. The plasticity mechanism achieves $\sim$ 22.8% improvement in performance on average, surpassing state-of-the-art values in both classical [Verstraeten et al., 2006] and delay-coupled reservoirs [Appeltant et al., 2011] with an average nrmse of 0.138 ± 0.02std. Only in five trials did the performance deteriorate (see inset in Figure 15B). The improvement decreases for larger time lags due to the deterioration in the DCR's memory capacity observed in Figure 14, but remains significant for $|\ell| < 5$.

6.6 DISCUSSION: EFFECTS OF PLASTICITY

In order to explain the observed results, we analyze and discuss the effects of the homeostatic plasticity mechanism (94) on the system's entropy $\mathcal{H}(x)$, virtual network topology, and the readout weights distribution $p(\alpha)$. We also discuss the role of the regulating parameter ρ.

6.6.1 *Entropy*

In Section 6.4.2, we stated that expanding the informational features of present input requires a mechanism that counteracts the reduction of a v-delay due to the maximization of the v-node's sensitivity. To prove this hypothesis, we derive a learning mechanism that explicitly maximizes the *mutual information* between the DCR's response and its present input. Again, we assume the v-nodes are independent, and for a particular v-node i, we maximize the quantity

$$\mathcal{I}(x_i; u) = \mathcal{H}(x_i) - \mathcal{H}(x_i|u), \tag{98}$$

where $\mathcal{H}(x_i)$ is the entropy of the v-node's response, while $\mathcal{H}(x_i|u)$ is the entropy of the v-node's response *conditioned* on the input. In other words, $\mathcal{H}(x_i|u)$ is the entropy of the response that does not result from the input. Bell and Sejnowski [1995] argued that maximizing (98) with respect to some parameter θ is equivalent to maximizing $\mathcal{H}(x_i)$, since the conditional entropy $\mathcal{H}(x_i|u)$ does not depend on θ, i.e., maximizing a v-node's input-information capacity is equivalent to maximizing its self-information capacity or entropy.

The entropy of x_i is given by $\mathcal{H}(x_i) = -E\left[\ln p_x(x_i)\right]$, where $p_x(x_i)$ is the ***p**robability **d**ensity **f**unction* (PDF) of the v-node's response. Since x_i is an invertible function of the Mackey-Glass nonlinearity f_i (see equation (83)) that is itself an invertible function of the input u (if the nonlinearity is chosen appropriately such as in equation (80)), the PDF of x_i can be written as a function of the PDF of f_i:

$$p_x(x_i) = \frac{p_f(f_i)}{|\frac{\partial x_i}{\partial f_i}|} \tag{99}$$

The entropy of the v-node's response is then given by

$$\mathcal{H}(x_i) = E\left[\ln\left|\frac{\partial x_i}{\partial f_i}\right|\right] - E\left[\ln p_f(f_i)\right] \tag{100}$$

The term $-E\left[\ln p_f(f_i)\right]$ measures the entropy of the nonlinearity f_i and is independent of θ_i. From equation (100), and taking into account equation (83), we can derive a learning rule that maximizes the entropy of the response by applying stochastic gradient ascent:

$$\begin{aligned}\Delta\theta_i \propto \frac{\partial \mathcal{H}(x_i)}{\partial \theta_i} &= \frac{\partial}{\partial \theta_i}\left(\ln\left|\frac{\partial x_i}{\partial f_i}\right|\right) \\ &= \left(\frac{\partial x_i}{\partial f_i}\right)^{-1}\frac{\partial}{\partial \theta_i}\left(\frac{\partial x_i}{\partial f_i}\right) \\ &= (1-e^{-\theta_i})^{-1}\frac{\partial}{\partial \theta_i}(1-e^{-\theta_i}).\end{aligned} \tag{101}$$

This leads to the following learning rule:

$$\Delta\theta_i = \alpha \frac{e^{-\theta_i}}{1 - e^{-\theta_i}}, \tag{102}$$

where $0 < \alpha \ll 1$ is a learning rate.

The update term (102) is a strictly positive monotonic function of the v-delay θ_i. This entails that, when unconstrained, maximizing a v-node's informational feature expansion results in an unbounded increase in its v-delay, i.e., $\theta_i \to +\infty$. On the other hand, the plasticity rule (94) can be rewritten as

$$\Delta\theta_i = \alpha\varsigma_i\rho - \alpha\varsigma_i\theta_i, \tag{103}$$

where $\varsigma_i = 2\sigma_i\theta_i^{2\rho-1}e^{-2\theta_i} > 0$. The term $\alpha\varsigma_i\rho$ in the plasticity mechanism (103) is also positive. This entails that it results, similar to (102), in an unbounded increase in the v-delay, and as a corollary, in an increase in the v-node's informational feature expansion.

Given the above, the homeostatic plasticity mechanism (94), for a particular DCR with delay τ, improves spatiotemporal computations by leading v-nodes to *specialize* in function. This is mediated by a competition between the v-nodes' sensitivity and their entropy. Some v-nodes become more sensitive to small fluctuations in input history, while others are brought closer to saturation where their entropy is higher, and as such, their ability for expanding informational features.

6.6.2 *Virtual Network Topology*

The effects of the homeostatic plasticity mechanism (94) on the DCR's network topology can be deduced from equation (82), according to which self-weights are given by $w_{ii} = (1 - e^{-\theta_i})$, and the weights the v-node i receives from the preceding v-node $j = i - 1$ is $w_{ij} = (1 - e^{-\theta_j})e^{-\theta_i}$.

When θ_i decreases, so does the v-node's self-excitation w_{ii}, which is consistent with less saturation of the v-node's activity. In addition, the choice of the regulating parameter ρ describes the tendency of the v-node i to converge towards a particular self-excitation level $w_{ii} = (1 - e^{-\rho})$. This entails that for higher ρ, the v-node's target activity level increases, which also corresponds to higher entropy, as discussed in Section 6.6.1.

The decrease in θ_i also leads the corresponding v-node's afferent w_{ij} to increase. This in turn increases the v-node j's influence on the activity of the v-node i, which results in higher *correlation* between the two (or higher *anti-correlation*, depending on the signs of the corresponding mask values M_j and M_i). The increase of correlation is in agreement with simulation results and in accord with the decrease of the v-node's entropy as its v-delay decreases. This is the case since the influence of the current input is overshadowed by information from the input history that is delivered from the preceding v-node j, which now drives the v-node i. Figure 13B shows an exemplary virtual weight matrix following plasticity, which illustrates these changes in network topology due to the repositioning of v-nodes on the delay line.

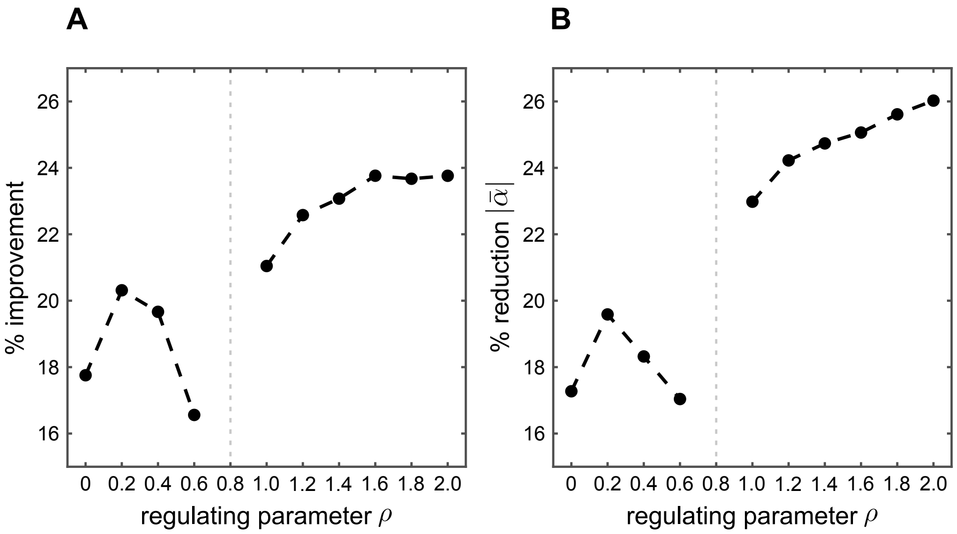

Figure 16: **(A)** Average improvement in performance and **(B)** reduction in average absolute values of the readout coefficients $|\bar{\alpha}|$ for different values of the regulating parameter ρ in comparison to the equidistant v-nodes case $\rho = 0.8$.

6.6.3 *Homeostatic Regulation Level*

Introducing the parameter ρ is necessary to regulate the trade-off between sensitivity and entropy, i.e., increasing and decreasing θ_i, as discussed analytically in Section 6.6.1. It is also the defining factor in the v-node's tendency to collapse, as is evident from the form of the plasticity function (103). The collapse of v-nodes is tantamount to a reduction in the DCR's dimensionality, which may be unfavorable with regard to the DCR's computational performance.

We test the latter hypothesis, and the choice of the regulating parameter, by running 1000 NARMA-10 trials for different ρ values that ranged between 0 and 2. Each trial shares the same mask M and the same NARMA-10 time series. As shown in Figure 16, the average improvement in performance in comparison to the reference equidistant case $\rho = \tau/n = 0.8$ increases for smaller ρ values, but drops again for $\rho = 0$. An increase in $\rho > 0.8$ also increases the improvement of performance but this increase saturates at $\rho = 1.6$. This is the case since the increase in v-delays favored by high ρ values, makes the collapse of other v-delays inevitable, in order to preserve the DCR's constant delay τ.

In a more detailed analysis, for each of the 1000 trials, we ranked different ρ values according to the resulting improvement of performance in reference to the equidistant case $\rho = \tau/n = 0.8$. We then calculated the percentage of trials that achieved the highest improvement in performance (1$^{\text{st}}$ rank) for some ρ value, compared to all other ρ values. We carried the same procedure for the 2$^{\text{nd}}$ and 3$^{\text{rd}}$ ranks as well. Figure 17 confirms the previous results as it shows that for $\rho = 0$, it is still possible to achieve the best improvement in performance, but it is less likely than other values. Figure 17 also illustrates a striking result. For none of the trials was the equidistant case, where no plasticity took place, the best choice regarding the computational power of the DCR. Only in 0.3% of the trials, did the nonplastic

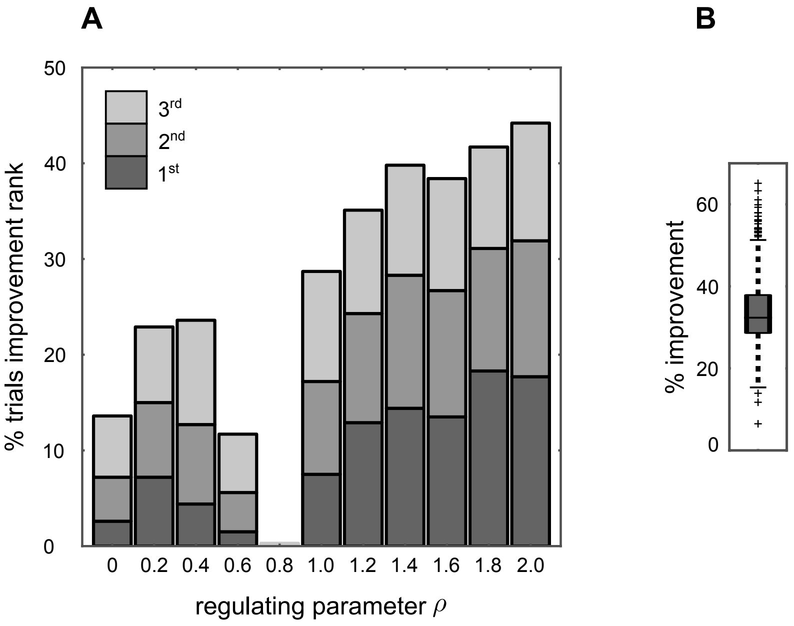

Figure 17: Performance of 1000 NARMA-10 trials for regulating parameter ρ values between 0 and 2. **(A)** Percentage of trials that achieved the 1st, 2nd, and 3rd highest improvement in performance for each ρ value. **(B)** Relative improvement, measured by the decrease in nrmse, after applying homeostatic plasticity, given for each trial the best choice in ρ. The box plot marks the median and the first and third quartiles. Whiskers extend to include data points within 1.5 times the interquartile range. The crosses specify data points outside this range and correspond to outliers.

equidistant case rank 3rd. As a result, for a given DCR setup there always exists a choice of ρ that results in nonequidistant v-nodes where spatiotemporal computations are enhanced. This is also summarized in Figure 17B, which shows the improvement in performance given the best choice in the regulating parameter ρ for each trial. The nrmse is reduced by $\sim 33.7\%$, with an average performance that reaches an unprecedented value of nrmse $\sim 0.117 \pm 0.01$std.

We point out that the homeostatic plasticity mechanism (94) also reduces the average absolute values of the readout coefficients $|\bar{\alpha}|$ (see Figure 16), which is similar in effect to an L_2-regularized model fit. This is not only advantageous with respect to numerical stability, but L_2-regularization also allows for a lower mean-square error on the validation set as compared to an unregularized fit [Hoerl and Kennard, 1970].

We now briefly discuss the effects of the homeostatic regulation level ρ on the virtual network topology. As expected, and due to the simplex constraint, both smaller and larger values of ρ lead to a more uniform distribution of v-delays. However, most of the distribution's mass remains concentrated at $\theta_i = \tau/n$, i.e., most v-delays remain unchanged or change only slightly. This has no effect on the qualitative features of the virtual network topology as outlined in Section 6.6.2, but quantitatively, more weights approach the extremes of the range $[0, 1]$.

6.7 COMMENTARY ON PHYSICAL REALIZABILITY

We demonstrated that the suggested plasticity mechanism (94) leads to spatiotemporal computational performances that surpass those of state-of-the-art results. An intuitive alternative to the plasticity mechanism would be to increase the number n of v-nodes within the constant full delay τ of the DCR. This solution suffers, however, from major drawbacks, particularly in regard to its realizability on physical hardware. Namely, there exists a physical constraint on the sampling rate of the DCR's activity, below which the speed and the feasibility of physical implementation is jeopardized. This imposes a minimal, admissible v-delay within the full delay line, and as such, represents an upper bound on the number of equidistant v-nodes. This constraint is accounted for in the current approach by restricting the updates of v-delays due to plasticity to discrete step sizes κ. The parameter κ then corresponds to the minimal admissible v-delay (different from 0, which results in *pruning* the DCR). This is the case since κ is chosen such that $\tau/(\kappa n)$ is an integer, where this integer refers to the number of minimal v-delays that fit in one τ/n, the v-delay in the equidistant v-nodes case. In the current results, κ was chosen such that $\tau/(\kappa n) = 100$, in order for the discretization to present a good approximation of continuous v-delay values. Nevertheless, simulations show that improved computational power persists even for $\tau/(\kappa n) = 8$, which corresponds to κ that is an order of magnitude larger than the minimal, experimentally viable v-delay. A stringent comparison between the results for different values of κ is problematic, since rougher quantization of v-delays, resulting from higher κ values, leads to less predictable effects on the behavior of the optimization problem (92), particularly of how the discretized v-delay grid relates to the global maximum, which itself depends on the choice of the regulating term ρ. Nevertheless, the persistent improvement in performance stands in favor of the method's applicability in physical realizations.

Furthermore, increasing the number of v-nodes poses a practical limitation, even when v-delays remain within the constraints of physical implementation. As expected, the average computational performance does increase for larger number of v-nodes, but saturates at some point. The plasticity mechanism improves the computational performance, and most importantly, reaches the saturation point in performance with smaller number of v-nodes than the equidistant case. Beyond the performance saturation point, plasticity becomes ineffective on average, That is, it leads some trials to an increase, and others to deterioration in computational performance. However, the redundancy resulting from increasing the number of v-nodes, even within the constraint of physical implementation, is disadvantageous in regard to the computational resources of the DCR: The linear readout mechanism remains a bottleneck, since increasing the number of regressors by sampling more v-nodes demands storing and inverting larger matrices, which is a serious challenge for both simulation and physical implementations. Again, the comparison between the results for different numbers of v-nodes is problematic, since changing the number of v-nodes modifies the statistics of the mask pattern, which may affect the proper choice of the regulating term ρ. These considerations in mind, the plasticity mechanism is suitable for physical realization since it is resources-saving by keeping the number of v-nodes smaller (and possibly pruning by leading some v-delays to collapse), and is computationally beneficial within the constraint of physical implementation, since it approaches the saturation point of computational performance using smaller number of virtual nodes. Nevertheless, further detailed investigation remains necessary for addressing boundary conditions and applicability on physical implementation of the suggested plasticity mechanism.

6.8 CONCLUSION

We have introduced a plasticity mechanism for improving the computational capabilities of a DCR, a novel RC architecture where a single nonlinear node is delay-coupled to itself. The homeostatic nature of the derived plasticity mechanism (94) relates directly to the information processing properties of the DCR in that it balances between sensitivity and informational expansion of input (see Section 6.6.1). While the role of homeostasis in information processing and computation has only been discussed more recently, its function as a stabilization process of neural dynamics has acquired earlier attention [von der Malsburg, 1973; Bienenstock et al., 1982]. From the perspective of the nervous system, pure Hebbian potentiation or anti-Hebbian depression would lead to destabilization of synaptic efficacies by generating amplifying feedback loops [Miller, 1996; Song et al., 2000], necessitating a homeostatic mechanism for stabilization [Davis and Goodman, 1998; Zhang and Linden, 2003; Turrigiano and Nelson, 2004]. Similarly, as suggested by the effects of the plasticity mechanism (103) on the virtual network topology (see Section 6.6.2), the facilitating sensitivity term $-\alpha\varsigma\theta$ is counteracted by the depressive entropy term $+\alpha\varsigma\rho$, which prevents synaptic efficacies from overpotentiating or collapsing.

In addition, rewriting (94) as $\Delta\theta \sim (\theta - \rho)$ strongly relates the derived plasticity mechanism to *normalization models* of neural homeostatic plasticity. Normalization models consider plasticity rules that regulate the activity of the neuron towards a *target firing rate*. They are usually of the form $\Delta q \sim (r - r_{\mathrm{tr}})$, where q is some quantity of relevance for learning, such as synaptic weights or the neuron's intrinsic excitability, r is an estimate of the neuron's output firing rate, and r_{tr} is the target firing rate [Kempter et al., 2001; Renart et al., 2003; Lazar et al., 2007, 2009; Zheng et al., 2013; Toutounji and Pipa, 2014]. In analogy, the v-delay estimates the v-node's activity, since a larger θ_i results in higher self-excitation w_{ii}, while ρ defines the target activity of the v-node (see Section 6.6.2). Furthermore, entropy of a neuron's output increases with its firing rate. As such, the increase of the v-delay θ in response to higher regulatory term ρ also increases the v-node's entropy, as confirmed analytically in Section 6.6.1.

Currently, and similar to the target firing rate r_{tr} which is usually chosen according to biological constraints, the regulating parameter ρ is left as a free parameter, and its optimal choice for a praticular DCR configuration is decided by brute force (see Section 6.6.3). However, the statistics in Figure 16 and Figure 17 conclusively show that any choice of ρ within the tested range leads to average and dominant improvement in computational performance in comparison to the equidistant case $\rho = \tau/n = 0.8$. Nevertheless, it is reasonable to assume that heuristics exist for the optimal choice of ρ, given a particular mask structure M, since the alterations in the mask values influence a v-node's sensitivity and entropy. A possible heuristic may relate the value of ρ to properties of *maximum length sequences*, by which Appeltant et al. [2014] constructed mask sequences with equidistant v-nodes. Similarly, we speculate that the direction and amplitude of a v-delay's change that are computationally advantageous depend on the corresponding and preceding v-node's mask values M_j for $j \leq i$. The main difficulty arises from the fact that, within the current formulation of the DCR in equations (81) and (82), no terms exist for relating different mask values to one another, and to corresponding v-delays. This is also the main obstacle facing the derivation of plasticity mechanisms for updating the mask M beyond the binary pattern $\pm\mu$. The appropriate choice of ρ is complicated further by its dependence on the demands of the executed task in terms of memory, nonlinear computations, and entropy. Finding criteria that connect these aspects to the optimal choice of ρ requires an extensive research that is a subject of current endeavors.

Enhancing the temporal multiplexing of input to the nonlinear node was the main goal of this article. We speculate that similar multiplexing may suggest a further important functionality of the extensive dendritic trees in some neuron types. On the one hand, Izhikevich [2006] discussed the infinite dimensionality dendritic propagation delays offer to recurrent neural networks. On the other hand, several studies investigated the computational role of the spatial distribution of active dendrites [Rumsey et al., 2006; Gollo et al., 2009; Graupner and Brunel, 2012]. In this article, we advocate a unified computational account that may integrate both the temporal and spatial aspects of dendritic computations. In particular, the spatial location of dendritic arbors may be optimized to achieve computationally favorable temporal multiplexing of the soma's input, in the fashion suggested by the DCR architecture. Consolidating this speculation would be the subject of future studies.

Acknowledgments

The contributions of Marcel Nonnenmacher and Anna-Birga Ostendorf to early stages of this work are gratefully acknowledged, so are the fruitful discussions with the members of the PHOCUS consortium. The authors acknowledge the financial support of the State of Lower Saxony, Germany via the University of Osnabrück, and the European project PHOCUS in the Framework 'Information and Communication Technologies' (FP7-ICT-2009-C / Proposal Nr. 240763).

6.9 APPENDIX A: SOLVING AND SIMULATING THE DCR

In this section, we derive equations (81) and (82). We would like to solve system (79) for $x(t)$, with $(\nu-1)\tau \leq t \leq \nu\tau$. Due to the recurrent dependency $x(t-\tau)$, this is not possible right away. However, if we assume a continuous function $\phi_\nu \in C_{[(\nu-2)\tau,(\nu-1)\tau]}$ is the solution for $x(t)$ on the previous τ-interval, we can replace $x(t-\tau)$ by $\phi_\nu(t-\tau)$. After the substitution, system (79) becomes solvable by the elementary method of *variation of constants* [Heuser, 2009]. The latter provides a solution to an equation of type $\dot{x}(t) = a(t)x(t) + b(t)$ with initial condition $x(t_0) = c$. The general solution on the interval I to the inhomogeneous equation is then given by

$$x(t) = x_h(t) \left(c + \int_{t_0}^{t} x_h(t)^{-1} b(t) dt \right), \quad t \in I,$$

where

$$x_h(t) = \exp \left(\int_{t_0}^{t} a(t) dt \right)$$

denotes a solution to the associated homogeneous differential equation. Consequently, for $a(t) = -1$ and $b(t) = f\big(\phi_\nu(t-\tau), \delta(t)\big)$, the solution to

$$\dot{x}(t) = -x(t) + f\big(\phi_\nu(t-\tau), \delta(t)\big),$$

subject to $x\big((\nu-1)\tau\big) = \phi_\nu\big((\nu-1)\tau\big)$, is given by

$$x(t) = e^{(\nu-1)\tau - t} \left(\phi_\nu\big((\nu-1)\tau\big) + \int_{(\nu-2)\tau}^{t-\tau} f\big(\phi_\nu(s), \delta(s)\big) e^{s-(\nu-2)\tau} ds \right). \quad (104)$$

This expression can be used right away in a numerical solution scheme, where the integral is solved using the *cumulative trapezoidal rule*. The resulting simulation of the DCR has been shown to be comparable in its accuracy and computational capabilities to adaptive numerical solutions, while considerably saving computation time [Schumacher et al., 2013].

Recall that $t_i = (\nu-1)\tau + \sum_{j=1}^{i} \theta_j$, with θ_j the temporal distances between consecutive virtual nodes. To arrive at a manageable analytical expression of the above solution for the sampling point t_i of virtual node i during the ν^{th} τ-cycle, we make the following approximation:

Let the DCR activity at a particular v-node $x_i(\bar{t}) = x(t_i)$, its nonlinearity $f_i(\bar{t}) = f\big(x_i(\bar{t}-1), M_i \cdot u(\bar{t})\big)$, and the DCR time step $\bar{t} = \lceil \frac{t}{\tau} \rceil = \nu$. If we assume that $f(\cdot)$ is piecewise constant at each θ_i, which is a valid approximation since $\theta_i \ll \tau$, expression (104) simplifies further to

$$\begin{aligned} x_i(\bar{t}) &= e^{-\sum_{j=1}^{i} \theta_j} \left(x_n(\bar{t}-1) + \sum_{j=1}^{i} f_i(\bar{t}) \int_0^{\theta_j} e^s ds \right) \\ &= e^{-\sum_{j=1}^{i} \theta_j} \left(x_n(\bar{t}-1) + \sum_{j=1}^{i} f_i(\bar{t}) (e^{\theta_j} - 1) \right) \\ &= e^{-\sum_{j=1}^{i} \theta_j} x_n(\bar{t}-1) + \sum_{j=1}^{i} (1 - e^{-\theta_j}) e^{-\sum_{k=j+1}^{i} \theta_k} \cdot f_j(\bar{t}). \end{aligned}$$

6.10 APPENDIX B: CONSTRAINT SATISFACTION

The sensitivity update rule of the virtual node distances θ_j has to satisfy the constraint $\sum_j \theta_j = \tau$. This describes a constraint manifold for valid virtual node distance vectors $\Theta \in \mathbb{R}^n$ during learning. The manifold has the structure of a simplex

$$V := \{x | x = (x_1, \dots, x_n)^T, \sum_{i=1}^{n} x_i = \tau\}$$

with $\dim V = n-1$ and simplex corners given by τe_i $(i = 1,\dots,n)$, where $(e_i)_{i=1}^n$ is the standard orthonormal basis of $\mathbb{R}^n$. We implemented the constraint optimization problem by first computing an unconstrained update for Θ, followed by an orthogonal projection onto V. Due to the simple linear structure of V, this strategy will converge onto the constrained optimum for Θ.

Denote by $n_V = \frac{\tau}{n}\sum_i e_i$ the central point of the constraint simplex, and let $(v_i)_{i=1}^{n-1}$, $v_i \in \mathbb{R}^n$, be an orthonormal basis for V. The latter is computed from an orthogonal basis, which can be constructed by simple geometrical considerations from the simplex corner point vectors as

$$\begin{pmatrix} -\tau & \cdots & -\tau \\ & \tau I_{n-1} & \end{pmatrix} \in \mathbb{R}^{n\times n-1}, \tag{105}$$

where I_{n-1} denotes the $(n-1)$-dimensional unit matrix. It is easily verified that this basis spans V and is indeed orthogonal. In conjunction with the inhomogeneity n_V, a normal vector with respect to V, any point on V can be expressed via the v_i. For some $x \in \mathbb{R}^n$ being the result of an unconstrained sensitivity update step, the constraint can be met by projecting x orthogonally onto V via the mapping

$$\begin{aligned} \pi_V(x) &= n_V + \sum_{i=1}^{n-1} \left((x-n_V)^T v_i\right) v_i \\ &= n_V + \sum_{i=1}^{n-1} \left(v_i^T (x-n_V)\right) v_i \\ &= n_V + \underbrace{\left(\sum_{i=1}^{n-1} v_i v_i^T\right)}_{:=M\in\mathbb{R}^{n\times n}} (x-n_V) \\ &= n_V + M(x-n_V). \end{aligned} \tag{106}$$

The addition and subtraction of n_V take care of the fact that V as a hyperplane is translated out of the origin by the inhomogeneity n_V. If the V-plane was centered in the origin,

$$\underbrace{(v_i^T x)}_{\in\mathbb{R}} v_i$$

would denote the orthogonal projection of x onto the i^{th} orthonormal basis vector. Accordingly, the linear combination of these projections yields the representation of $\pi_V(x)$ with respect to the basis $(v_i)_{i=1}^{n-1}$.

7 A STATISTICAL MODELING APPROACH FOR DETECTING GENERALIZED SYNCHRONIZATION

7.1 ABSTRACT

Detecting nonlinear correlations between time series presents a hard problem for data analysis. We present a generative statistical modeling method for detecting nonlinear generalized synchronization. Truncated Volterra series are used to approximate functional interactions. The Volterra kernels are modeled as linear combinations of basis splines, whose coefficients are estimated via l1 and l2 regularized maximum likelihood regression. The regularization manages the high number of kernel coefficients and allows feature selection strategies yielding sparse models. The method's performance is evaluated on different coupled chaotic systems in various synchronization regimes and analytical results for detecting m:n phase synchrony are presented. Experimental applicability is demonstrated by detecting nonlinear interactions between neuronal local field potentials recorded in different parts of macaque visual cortex.

7.2 INTRODUCTION

Many natural systems generate complex collective dynamics through interactions between their component parts. A prominent example is the transient neural dynamics of the brain which presumably involve strong functional couplings between cortical regions. Determining the nature of such interactions is not easy. At the most general level, the problem is one of detecting *generalized synchronization* Rulkov et al. [1995] between time series $x(t)$ and $y(t)$. That is, detecting the existence of a functional, potentially nonlinear, time delayed or other stable relationship such that $y(t) = F[x](t)$ is predictable. Strictly speaking, generalized synchronization results from interactions between systems that create stable attractors in their total phase spaces, i.e. given $x(t)$ the response system y has to be stable. Lag and other forms of synchronization are subsets of this problem, and systems may transition from phase, via lag, to complete synchronization as coupling strengths increase Rosenblum et al. [1997].

When the interactions are nonlinear, or the coupled systems themselves complex or chaotic Kocarev and Parlitz [1996]; Senthilkumar et al. [2008], standard linear methods, such as cross correlation or coherency, may not be able to detect an interaction. Nonlinear methods are therefore necessary. Existing approaches are usually based on reconstructing the phase space of the underlying system by finding an appropriate time-delay embedding Takens [1981]. Recent methodologies include the *Joint Probability of Recurrence* (*JPR*) method Marwan et al. [2007]. JPR is based on the evaluation of trajectory recurrence probabilities in small neighborhoods of the reconstructed phase space. The *JPR* is mathematically similar to another technique, the *Synchronization Likelihood* Stam [2002], which is derived from generalized mutual information concepts and popular in neuroscientific research areas. Although *JPR* and *SL* can detect nonlinear synchronization in many data sets (see e.g. Kreuz et al. [2007]; Sakkalis et al. [2009]; Romano et al. [2005]), it can be hard to determine the appropriateness of the embedding space. Further, such methods do not yield information about the functional form (nonlinearity) of the interaction.

Here we propose a different approach, directly estimating a functional which describes nonlinear interactions between two time series $x(t)$ and $y(t)$. In particular, we predict time series $y(t)$ from $x(t)$ using a Volterra series operator F on $x(t)$. The kernels of F are expanded using a set of basis functions, the coefficients of which fit using maximum posteriori regression. After obtaining an estimated signal $y_E = F[x]$ the degree to which y can be predicted from x is determined by computing the correlation coefficient $r(y_E, y)$ on an independent validation data set. Modeling F using a Volterra series is a canonical choice, since Volterra series are well-known for their versatility in nonlinear system identification (see e.g. Rugh [1981],Boyd et al. [1984]). They allow F to approximate arbitrary continuous functionals and flows of many non-autonomous dynamical systems, in particular systems with memory. The existence of non-zero second order or higher terms indicates nonlinear interactions. Furthermore, in agreement with the stability condition of generalized synchronization, a *steady-state theorem* for Volterra series (see Boyd et al. [1984]) asserts that for $x(t) \rightarrow x_s(t)$ within the radius of convergence of F, the response system is stable, i.e. $F[x](t) \rightarrow F[x_s(t)](t)$ as $t \rightarrow \infty$.

We call F the *Functional Synchrony Model* (*FSM*) and apply our method to several coupled chaotic systems for which generalized nonlinear synchronization is known to exist. We recover the nonlinear interactions with much greater accuracy than with either linear approaches, or the JPR method. We also demonstrate the existence of nonlinear coupling between local field potentials recorded in macaque visual cortex during stimulation by natural scenes movies.

Interactions between time series $x, y \in \mathbb{R}^N$ are modeled using a truncated Volterra series operator of order n with a history dependence (memory) of K time steps:

$$y_E(t) = F[x](t) = \sum_{j=0}^{n} Y_{j,K}(t), \tag{107}$$

where $Y_{j,K}$ is the jth order Volterra functional

$$Y_{j,K}(t) = \sum_{k_1=0}^{K} \cdots \sum_{k_j=0}^{K} h_j(k_1, ..., k_j) x(t-k_1) \cdots x(t-k_j). \tag{108}$$

Restrictions of this model form, particularly for modeling m:n phase synchronization are discussed below. To flexibly capture a wide variety of interactions, we expand the Volterra kernels h_j in a set of basis functions $B = \{b_m(k) \mid m = 1, ..., M\}$ as

$$\begin{aligned} & h_j(k_1, ..., k_j) = \\ & \sum_{m_1=1}^{M} \cdots \sum_{m_j=1}^{M} \tilde{a}_j(m_1, ..., m_j) b_{m_1}(k_1) \cdots b_{m_j}(k_j), \end{aligned} \tag{109}$$

with parameters $\tilde{a}_j(m_1, ..., m_j) \in \mathbb{R}$. Inserting eq. (109) into (108), we yield

$$Y_{j,K} = \sum_{m_1=1}^{M} \cdots \sum_{m_j=1}^{M} \tilde{a}_j(m_1, ..., m_j)\phi_{m_1,...,m_j}. \tag{110}$$

Denoting $\tilde{x}_n = \{x(n-K), x(n-(K-1)), ..., x(n)\}$, the $\phi_{m_1,...,m_j}$ are nonlinear basis functions in $\tilde{x}_n$ that constitute the covariates of our model, given by

$$\phi_{m_1,...,m_j} = \sum_{k_1=0}^{K} \cdots \sum_{k_j=0}^{K} b_{m_1} \cdots b_{m_j} x(n-k_1) \cdots x(n-k_j). \tag{111}$$

The covariates are symmetric in $\{m_1, .., m_j\}$, i.e. for all permutations $\pi(m_1, ..., m_j)$, $\phi_{\pi(m_1,...,m_j)}$ represents the same covariate and can be factored out in the model, yielding new coefficients a_j (as sums of the former $\tilde{a}_j$) and a corresponding reduction in summation indices

$$Y_{j,K} = \sum_{m_1=1}^{M} \sum_{m_2 \geq m_1}^{M} \cdots \sum_{m_j \geq m_{j-1}}^{M} a_j(m_1, ..., m_j)\phi_{m_1,...,m_j}. \tag{112}$$

Furthermore, the covariates can be factored out into products of simple convolutions,

$$\begin{aligned} \phi_{m_1,...,m_j}(\tilde{x}_n) &= \left(\sum_{k_1=0}^{K} b_{m_1}(k_1)x(n-k_1)\right) \\ &\cdots \left(\sum_{k_j=0}^{K} b_{m_j}(k_j)x(n-k_j)\right) \\ &= \phi_{m_1} \cdots \phi_{m_j}. \end{aligned} \tag{113}$$

Consequently, all higher order covariates are simply products of 1^{st} order covariates ϕ_{m_i}.

In this paper we expand the kernels using cubic basis splines. This basis spans a vector space of piecewise polynomial functions with smooth nonlinearities, and is uniquely determined by a knot sequence τ_K on the memory interval $[0, K]$. Using the *de Bohr algorithm* Boor [2001] on τ_K, all basis splines are fully specified and can be constructed recursively. The first order functional is thus given by a linear combination of basis splines, corresponding to a piecewise polynomial operating on $x(t)$ as a finite impulse response filter. Higher order kernels weight monomials of x, e.g. $x(t-k_1)x(t-k_2)$, which intuitively represent interactions between different points $t-k_j$ in time. Other bases, for example wavelets, could of course have been used.

Regardless of the basis chosen, the final model in eq. (112) is linear with respect to the coefficients a_j. Thus the coefficients can easily be determined by maximum likelihood based linear regression. Indexing all covariates and coefficients in eq. (112) and eq. (107) with the set $[1, ..., A]$, we define a design matrix for time series $x, y \in \mathbb{R}^N$ as

$$\Phi(x) = \begin{pmatrix} \phi_1(\tilde{x}_1) & \phi_2(\tilde{x}_1) & \cdots & \phi_A(\tilde{x}_1) \\ \phi_1(\tilde{x}_2) & \phi_2(\tilde{x}_2) & \cdots & \phi_A(\tilde{x}_2) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_1(\tilde{x}_N) & \phi_2(\tilde{x}_N) & \cdots & \phi_A(\tilde{x}_N) \end{pmatrix} \in \mathbb{R}^{N \times A} \tag{114}$$

and a vector of coefficients $\mathbf{a} \in \mathbb{R}^A$. We can now state a linear regression problem with nonlinear basis functions as

$$\Phi(x)\mathbf{a} = y.$$

To select a sparse set of relevant coefficients and ensure the model generalizes to validation data, we use *Elastic Net* regularization, interpolating $l_1 - l_2$ norm with a hyperparameter β Friedman et al. [2010]. Interpreted in a Bayesian maximum posteriori framework, changing the interpolation and regularization effectively changes the assumed prior distribution of model coefficients. While the l_1 norm corresponds to an isometric Laplace prior, the l_2 norm is normally distributed. As a result, the l_1 norm promotes sparse coefficient vectors, assuming few independent covariates carry most of the information, whereas the l_2 norm is known to foster clusters of correlated covariates. After fitting the model to training data, we test its generalizability by using it to predict an independent validation data set. Model accuracy is judged using the correlation coefficient between the signal and the prediction. Our statistical framework would also allow other goodness of fit measures, such as *Akaike information criterion* or likelihood based cross validation, to be used.

7.3 RÖSSLER-LORENZ SYSTEM

To study the performance of our method in a setup of two unidirectionally coupled nonidentical systems, we first consider a Rössler system driving a Lorenz system, which is a standard benchmark in the literature. We will also use this example to walk through the fitting procedure in detail. The equations of the drive system are

$$\begin{aligned} \dot{x}_1 &= 2 + x_1(x_2 - 4), \\ \dot{x}_2 &= -x_1 - x_3, \\ \dot{x}_3 &= x_2 + 0.45x_3, \end{aligned} \tag{115}$$

while the response system is given by

$$\begin{aligned} \dot{y}_1 &= -\sigma(y_1 - y_2), \\ \dot{y}_2 &= ru(t) - y_2 - u(t)y_3, \\ \dot{y}_3 &= u(t)y_2 - by_3, \end{aligned} \tag{116}$$

where $u(t) = x_1 + x_2 + x_3$. With $\sigma = 10$, $r = 28$, $b = \frac{8}{3}$, the driven Lorenz system is asymptotically stable Kocarev and Parlitz [1996] and thus in a regime of generalized synchronization with the Rössler system.

The systems' third coordinates x_3, y_3 are chosen as time series $x(t), y(t)$ respectively, with 10000 data points sampled at $\Delta t = 0.02$. The linear correlation coefficient is $r(y, x) = -0.168$, corresponding to the projection of the complex generalized synchronization manifold onto (x_3, y_3), shown in figure 18a_1. We try to predict $y(t)$ as $y_E(t) = F[x](t)$ with a 2^{nd} order Volterra series model F. To fully specify the model, we merely need to choose a knot sequence τ_K over a memory interval of K time steps. By visual inspection of the time series, $K = 350$ is chosen to span at least a full period of both systems. Accordingly, τ_K is chosen to

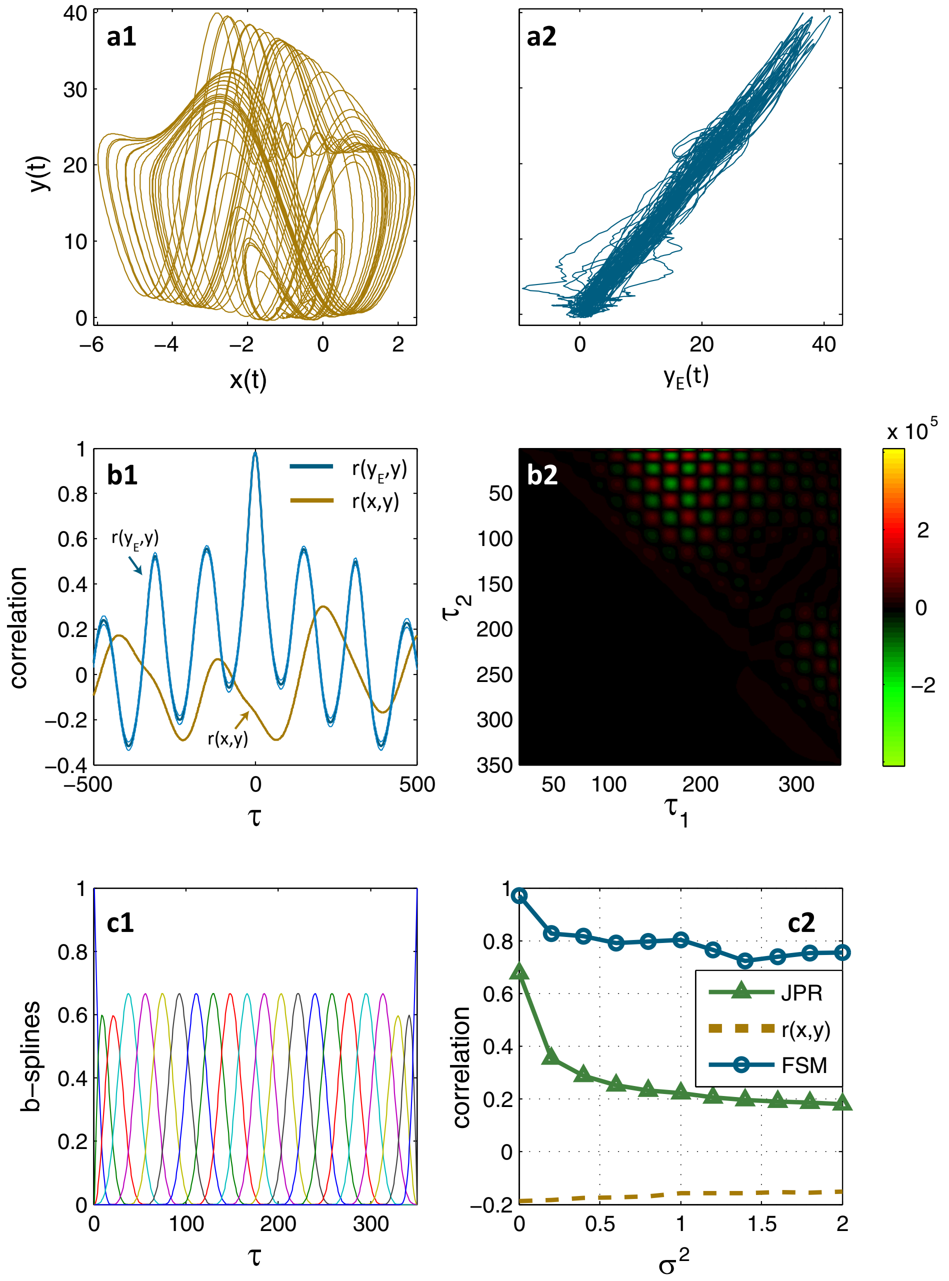

Figure 18: Identification of nonlinear interaction in a coupled Rössler-Lorenz system. **a1)**: Nonlinear synchronization manifold between original sampled data x and y (the systems' 3^{rd} coordinates) in generalized synchronization with correlation $r(x,y) = -0.168$. **a2**: Linearized manifold between y_E and y, where $y_E(t) = F[x](t)$ is the output of a 2^{nd} order Volterra model, yielding $r(y_E,y) = 0.98$. **b1)** Delay-shifted (by τ) correlation coefficients. **b2)** 2^{nd} order kernel corresponding to a2). **c1)** Set of cubic b-splines corresponding to b2), used in eq. (109). **c2)** Performance of the method (FSM, $r(y_E,y) \in [-1,1]$) for Rössler-Lorenz system with additive white noise over increasing variance σ^2, compared against correlation $r(x,y)$, as well as the $JPR \in [0,1]$.

cover the interval $[0,350]$ with 22 equidistantly spaced knots, each corresponding to the onset of the nonzero compact carrier of a particular cubic basis spline. A density of 22 splines is deemed sufficient for our model to capture the variations in the signals $x(t), y(t)$. The resulting set of basis splines is shown in figure 18c_1. We can now construct a design matrix (eq. 114) with 10000 × 276 entries, where $A = 276$ denotes the number of covariates, consisting of a 0 order constant, as well as 22 1^{st} order and 253 2^{nd} order covariates, as given by eq. (112). Using an isometric normally distributed prior distribution of coefficients ($\beta = 0.01$), we assume all covariates share a similar amount of information. Accordingly, using a mild regularization parameter $\lambda = 0.001$ the feature selection procedure finds 275 covariates to be constitutive for our model $y_E = F[x]$.

The model fit yields a correlation coefficient $r(y, y_E) = 0.98$ on an independent validation set of size 10000. Thus, generalized synchronization is detected with perfect accuracy. Moreover, the resulting model is fully predictive with respect to $y(t)$. Figure 18a_2 shows that our method "linearized" the synchronization manifold. The lag correlation plot in figure 18b_1 shows the correlation of the two signals as a function of varying delay shift τ between the signals, where $\tau = 0$ corresponds to $r(y, y_E) = 0.98$. The periodic relationship between the two chaotic oscillators is apparent. Figure 18b_2 depicts the 2^{nd} order Volterra kernel, i.e. the nonlinear aspects of the model that are necessary to capture the interaction. Here, the periodicity is also present, in form of alternations across the diagonal. While the regularization produced only two local clusters of covariates as main constituents of the model, the very regular weighting within the clusters reflects the assumptions encoded in the coefficient prior. Note that due to the symmetry of the kernels (see eq. (112)) only the "upper triangular" part of (τ_1, τ_2) space is populated by model covariates. Adding additional white noise to the data, our method also shows a strong noise robustness across an increasing variance σ^2 (fig. 18c_2).

To compare our method against the *JPR*, we chose embedding space parameters producing results on this data set comparable to Marwan et al. [2007]. The *JPR* is clearly outperformed and suffers greatly from the additive noise (fig. 18c_2). These effects may be countered by increasing the ϵ-neighborhoods in which the recurrence probabilities are evaluated, however, lacking any goodness-of-fit measure for the parameter set this may also increase the number of false positives and render the results meaningless.

7.4 MACKEY-GLASS NODES

Our second example involves generalized synchronization between delay-coupled Mackey-Glass nodes described by the equation

$$\dot{x}_i(t) = \frac{2x_{i-1}(t - \tau_d/n)}{1 + x_{i-1}(t - \tau_d/n)^9} - x_i(t), \quad \tau_d = 300 \tag{117}$$

The data is sampled from a ring containing up to $n = 16$ Mackey-Glass nodes, displaying chaotic dynamics, where node i receives delay-coupled input from node $i - 1$, with a total delay of $\tau_d = 300$ in the whole ring. The existence of generalized synchronization for the case of x_i driving $x_{i-n/2}$ can be demonstrated using the auxiliary systems approach Abarbanel et al. [1996].

Figure 19a shows the delay-embedded chaotic attractor (nonlinear synchronization manifold, brown) of two coupled Mackey-Glass nodes, where driving time

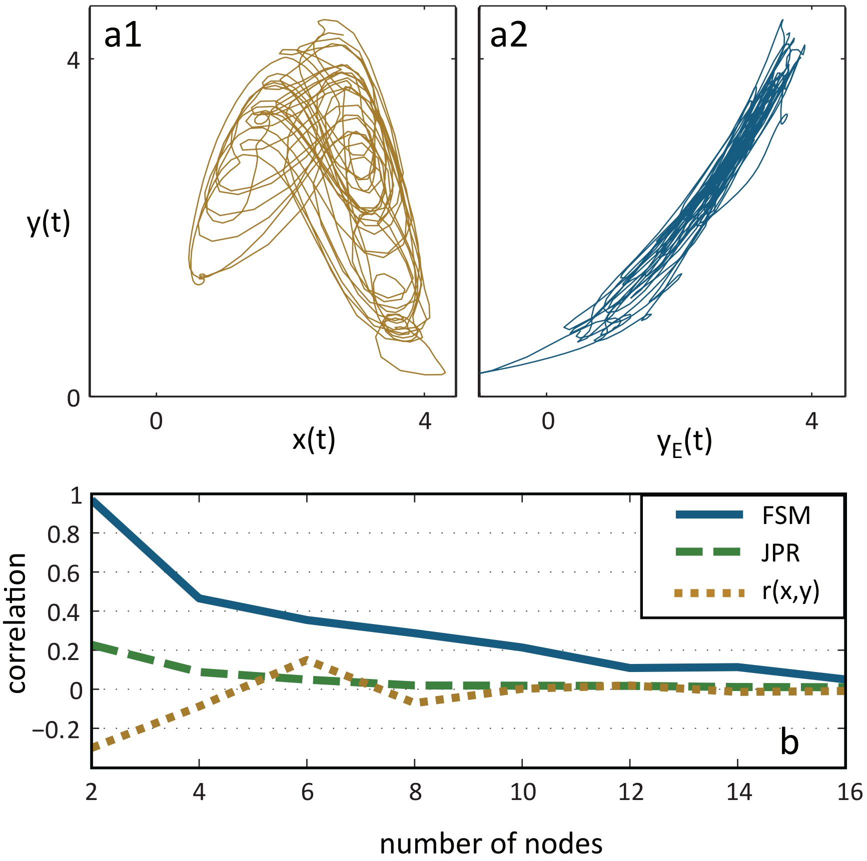

Figure 19: Performance on generalized synchronized Mackey-Glass delay rings. **a1)**: Nonlinear time-embedded GS manifold of ring with two nodes x and y. **a2)**: Linearized synchronization manifold between y_E (2^{nd} order model) and y. **b)** Results for Mackey-Glass rings of varying size. Shown are prior correlation in data ($r(x,y)$, dashed), *JPR* (light dashed) and our method (FSM, solid) using Volterra series models up to order 3.

series $x(t)$ correspondes to node x_i, and target $y(t)$ corresponds to $x_{i-n/2}$. The blue graph shows the transformation to a linear manifold after application of our method. We use a 2^{nd} order model, with non-uniform knot sequence τ_K supporting local maxima in the autocorrelation function of the Mackey-Glass ring that occur due to the system's delay-feedback. In total, 28 b-splines are used to cover the interval $[0, 350]$, encompassing the total delay-time τ_d in the ring. With $\beta = 0.99$, feature selection yields a sparse set of 158 predictive covariates that we apply to data sets of size 30000 or higher. While detection is possible with less than 10000 data points, yielding a fully predictive model on this complex data set needs more data to generalize and capture the strong nonlinear components of the interaction.

Figure 19b summarizes the resulting correlation $r(y_E, y)$ (blue) for functional Volterra series models of order $j \leq 3$. Performance is plotted against an increasing number n of nodes in the ring. A fully predictive model is found for $n = 2$, while detection of significant nonlinear interaction (significance determined using bootstrapped confidence intervals) is still possible for $n \leq 16$, where no linear correlation $r(x, y)$ is measurable in the data. In comparison, the *JPR* (dashed green line) failed to detect the interaction in rings larger than $n = 2$ for all tested embedding space and recurrence parameters which we chose manually as well as automatically using mutual information and *false nearest neighbours* criteria. Note, however, that much less data points (up to 10000) could be used for the recurrence based method, which draws heavily on computational resources since it has to compute an $N \times N$ recurrence matrix (where N is the number of delay-embedded data points) for both time series.

7.5 COUPLED RÖSSLER SYSTEMS

Our third example application is to two identical coupled Rössler systems, described by the equations

$$\begin{aligned} \dot{x}_{1,2} &= -\omega_{1,2} y_{1,2} - z_{1,2}, \\ \dot{y}_{1,2} &= \omega_{1,2} x_{1,2} + 0.16 y_{1,2} + \mu(y_{2,1} - y_{1,2}), \\ \dot{z}_{1,2} &= 0.1 + z_{1,2}(x_{1,2} - 8.5). \end{aligned} \tag{118}$$

We use $\omega_1 = 0.98, \omega_2 = 1.02$ corresponding to a phase coherent regime of the two slightly dissimilar chaotic oscillators. These coupled three dimensional systems exhibit a wide range of synchronization dynamics as a function of the coupling strength μ Osipov et al. [2003], transitioning from an unsynchronized regime to complete synchronization via (1:1) phase synchronization as μ is increased from 0 to 0.15.

Using the first coordinates (x_1, x_2) as the driving (x) and target time series (y) respectively with 15000 data points sampled at $\Delta t = 0.02$, we can detect nonlinear interaction even for very weak coupling ($\mu = 0.034$) with a 2^{nd} order model and a memory of 500 time steps, encompassing a full period of the nonlinear oscillators. Lacking further information about the interaction, we choose a dense equidistant knot sequence for 52 cubic b-splines. Consequently, many covariates will contribute only little information to the model. This is accounted for by imposing strong regularization and choosing a sparse prior for feature selection ($\beta = 0.99$), resulting in a total of 109 informative covariates for the model.

At $\mu = 0.034$, x and y lie on a highly complex manifold (fig. 20a_1) and the correlation coefficient between x and y is zero. Our Volterra series approach "linearizes" the synchronization manifold between the model prediction and the data (fig. 20a_2) and accurately describes the functional interaction, yielding $r(y_E, y) = 0.97$. Figure 20b shows the corresponding first and second order Volterra kernels. Both kernels are highly sparse, and strong quadratic interactions between $x(t)$ at different times during the memory period prove necessary to predict $y(t)$. The interaction can, in fact, be described over a broad range of coupling strengths, as demonstrated in Figure 20c. The method yielded fully predictive models for nearly all μ as indicated by correlation coefficients $r(y_E, y)$ near 1 for $\mu \in [0, 0.15]$.

7.6 PHASE SYNCHRONY

A drawback of the current formalism is that Volterra series impose restrictions for modeling phase synchrony. By definition, two nonlinear oscillators x, y are phase synchronized if for their phases ϕ_i it holds that $|n\phi_x - m\phi_y| < \epsilon$, with

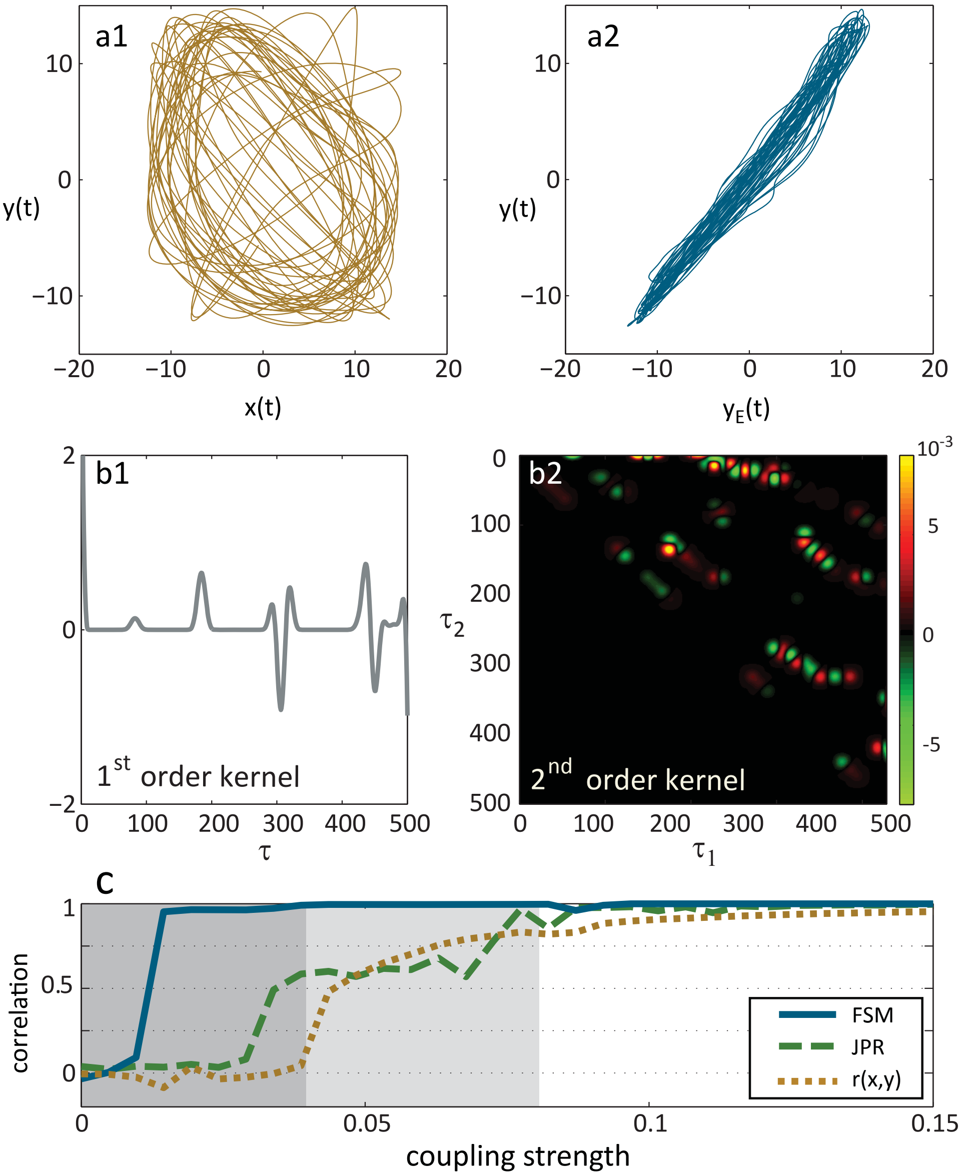

Figure 20: Identification of nonlinear interaction between coupled Rössler systems. **a1)**: Nonlinear synchronization manifold between original sampled data x and y (the systems' 1^{st} coordinates) at onset of phase synchronization ($\mu = 0.034$) with correlation $r(x, y) \approx 0$. **a2)**: Linearized manifold between y_E and y, where $y_E(t) = F[x](t)$ is the output of a 2^{nd} order Volterra model, yielding $r(y_E, y) = 0.97$. **b1)** 1^{st} order kernel corresponding to a2). **b2)** 2^{nd} order kernel corresponding to a2). **c)** As μ increases, the system transitions from unsynchronized, via phase ($\mu > 0.04$) to generalized chaotic synchronization ($\mu > 0.08$). Performance of the method (*FSM*, $r(y_E, y) \in [-1, 1]$) for various coupling strengths μ is compared to correlation of the raw data $r(x, y)$, as well as the $JPR \in [0, 1]$.

$n, m \in \mathbb{Z}, \epsilon \in \mathbb{R}$. The generative model may thus have to scale ϕ_x by a fraction to yield ϕ_y. In theory, Volterra series cannot achieve this, as a result of the *periodic steady state theorem* Boyd et al. [1984]: Periodicity present in $x(t)$ must reoccur in the Volterra series $F[x](t)$. The case of n:1, however, is possible by increasing the frequency of the input signal by a factor n, retaining the orginal slower periodicity in the resulting faster signal.

To illustrate the Volterra series response to a single frequency component of an oscillatory signal, consider for example the harmonic complex oscillation $u(t) = \alpha_k e^{i\omega kt}$. The truncated Volterra series response breaks down into the components of the kernel functions, given by the covariates specified in eq. (113). Higher order covariates are products of 1^{st} order covariates ϕ_m which constitute linear time-invariant systems such that $u(t)$ is an eigenfunction. Consequently, $\phi_m[u](t) = e^{i\omega kt}\alpha_k H_m(i\omega k)$, where $H_m(i\omega k)$ is the frequency response of ϕ_m given by the discrete Laplace transform of the corresponding 1^{st} order kernel basis function b_m. For an n^{th} order covariate $\Phi^{(n)}$ it follows that

$$\begin{aligned}\Phi^{(n)}[u](t) &= \phi_{m_1}[u](t)\phi_{m_2}[u](t)\cdots\phi_{m_n}[u](t)\\ &= \underbrace{\alpha_k^n H_{m_1}(i\omega k)\cdots H_{m_n}(i\omega k)}_{\tilde{H}(\omega)}\, e^{(i\omega kt)^n}\\ &= \tilde{H}(\omega)e^{in\omega kt}.\end{aligned}\tag{119}$$

Hence, the phase dynamics of $u(t)$ are scaled by a factor n, which suggests that an n^{th} order Volterra series operator can account for n:1 phase synchronization.

We confirmed this hypothesis using white noise jittered cosines ($\sigma^2 = 0.4$) with n:1 phase relationships for $n \leq 5$. All models were fully predictive with $r(y_E, y) \approx 1$. Following Chen et al. [2001], we also applied the method to two identical Rössler systems coupled in a drive-response scenario and locked in 4:1 phase synchronization. The drive oscillator is described by

$$\begin{aligned}\dot{x}_1 &= -y_1 - z_1,\\ \dot{y}_1 &= x_1 + 0.15y_1,\\ \dot{z}_1 &= 0.2 + z_1(x_1 - 10).\end{aligned}\tag{120}$$

The response oscillator is governed by

$$\begin{aligned}\dot{x}_2 &= -y_2 - z_2 + 80(r_2\cos(\frac{n}{m}\phi_1) - x_2),\\ \dot{y}_2 &= x_2 + 0.15y_2 + 80(r_2\sin(\frac{n}{m}\phi_1) - y_2),\\ \dot{z}_2 &= 0.2 + z_2(x_2 - 10),\end{aligned}\tag{121}$$

with phase and amplitude defined as

$$\begin{aligned}\phi_1 &= \arctan\left(\frac{y_1}{x_1}\right),\\ r_2 &= (x_2^2 + y_2^2)^{1/2}.\end{aligned}\tag{122}$$

The phase synchronization was verified for $m = 4, n = 1$ by checking the frequency locking condition $\Delta\Omega_{4:1} = 4\Omega_1 - \Omega_2 < 10^{-6}$, where $\Omega_i = \langle\dot{\phi}_i\rangle$ for $i = 1, 2$ the mean frequency averaged over 80.000 data points sampled at $\Delta t = 0.01$.

Using the first coordinates x_1, x_2 as time series $x(t), y(t)$ with 30000 data points each (fig. 21a_1) we fit a 4^{th} order model $F[x](t) = y_E(t)$. We set $\beta = 0.95$ to enforce sparse solutions since it is expected that a few 4^{th} order features are most

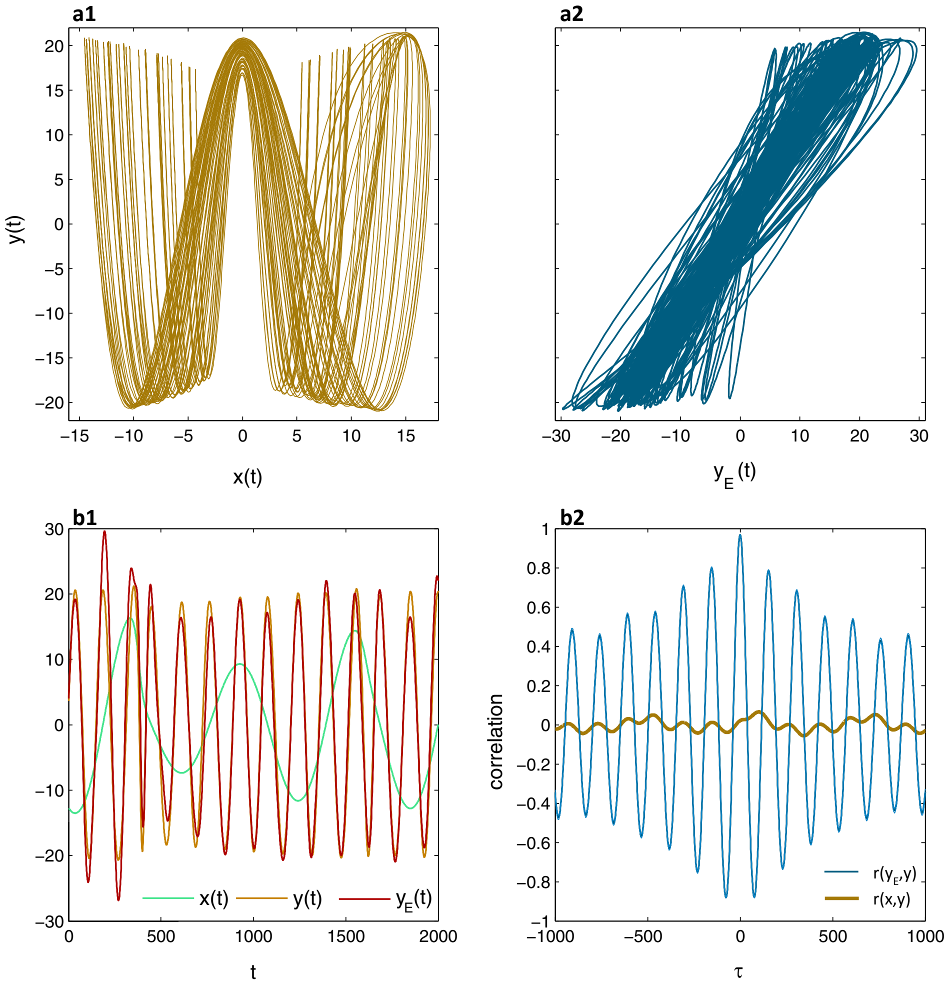

Figure 21: Identification of interaction between unidirectionally coupled Rössler systems in 4:1 phase synchronization (eq. 121). **a1)**: Nonlinear synchronization manifold projected onto first coordinates x_1, x_2 of the two systems (brown). **a2)**: Linearized synchronization manifold after application of a 4^{th} order Volterra series operator (blue). **b1)**: Time domain plots of original signals x(t) (green (light gray)) and y(t) (orange (gray)) compared to the 4^{th} order model prediction $y_E(t)$ (red (dark gray)). **b2)**: Delay-shifted (by τ) correlation plots of original signals $r(x,y)$ (brown, thick) and model performance $r(y_E, y)$ (blue, thin). Bootstrapped confidence intervals are shown as dashed lines in light blue.

informative. An equidistant knot sequence with 14 knots in $[0, 1000]$ is chosen to cover at least one full amplitude of each system. The feature selection process yields 117 mostly 4^{th} order covariates. The resulting model is fully predictive with $r(y_E, y) = 0.97$, as compared to $r(x, y) = 0.02$ in the original signals, and clearly captures the periodicity, as can be seen in the delay-shifted correlation coefficient plot (fig. $21b_2$). Figure $21b_1$ shows original time series $x(t), y(t)$ in comparison to the prediction $y_E(t)$ plotted against time t. We compare this result against the recurrence based phase synchronization index $CPR \in [0, 1]$ Marwan et al. [2007], which essentially quantifies the coincidence of maxima in two generalized autocorrelation functions for x and y and represents a complimentary tool to the JPR. Our best result for a particular choice of parameters yields $CPR = 0.5$ on a corresponding data set of size 5000. The low index is explained by the fact that for phase synchronization with $m, n \neq 1$, fewer coincidences of maxima in the generalized autocorrelation functions of x, y occur.

7.7 LOCAL FIELD POTENTIALS IN MACAQUE VISUAL CORTEX

Finally, we demonstrate the applicability of our method to noisy and unprocessed data from biological systems. To this end, we apply our method to LFP data recorded from electrodes located in macaque primary visual cortex (V1).

The monkey was watching a short (2.8 sec) natural scenes movie with 600 repetitions (for details about the experimental setup, see Gerhard et al. [2011]). V1 is retinotopically organized, so the different electrodes recorded signals generated by neuronal populations receiving input from distinct parts of the visual field. However, it has been hypothesized that there are strong lateral interactions between different parts of V1 which combine information about different parts of the visual stimulus. We use our methodology to detect nonlinear interactions between electrode signals with near zero linear correlation coefficient. In particular, recordings of pairs of analyzed channels were made from the opercular region of V1 (receptive field centers 2.0° to 3.0° eccentricity) and from the superior bank of the calcarine sulcus (10.0° to 13.0° eccentricity), respectively. The distance regarding the receptive field position is therefore of the order of 7° eccentricity and thus much larger than the receptive field sizes of the projection neurons. Therefore, the populations recorded by both channels have no common bottom-up input.

No significant interactions could be detected prior to stimulus onset. Post stimlus onset we analyzed both the induced potential (IP, unaltered LFP recordings) and the evoked potential (EP, the signal average across all trials). Here, the EP signals contained 2800 data points (the length of one experimental trial) in both, validation and training set. These were obtained by randomly selecting subsets of several hundred trials for averaging. IP data sets were substantially larger as time series from individual experimental trials were chosen randomly to be concatenated and used as a single data set.

In Figure $22b_1$ we use the LFP of one electrode (x), to predict the LFP of another (y) at various time lags, and show the resulting performance of our method. The data shown has close to zero linear correlation between the two LFPs (lag 0) for both EP ($r_{EP}(x, y)$) and IP (not shown). In contrast, the correlation coefficient between the model prediction and LFP is substantial, for both the IP ($r_{IP2}(y_E, y)$) and the EP ($r_{EP2}(y_E, y)$). Performance was substantially improved when second order models were used, indicating significant nonlinear interactions. This can be seen by comparing the performance of the 2^{nd} order model for predicting the EP ($r_{EP2}(y_E, y) \approx 0.89$) with a first order model ($r_{EP1}(y_E, y) \approx 0.53$). The second order interactions (Volterra kernel) are visualized in figure $22b_2$. Figure 22a shows the corresponding interaction manifolds of the EP tetrode signals x and y (brown) which is clearly linearized by the method (blue). Similar results were obtained using other LFPs from both this, and a different monkey. Although it is known that the neuronal populations generating the two LFPs are directly stimulated by different parts of the visual field, our result that there are strong nonlinear interactions between the populations suggests that V1 neurons may combine information from

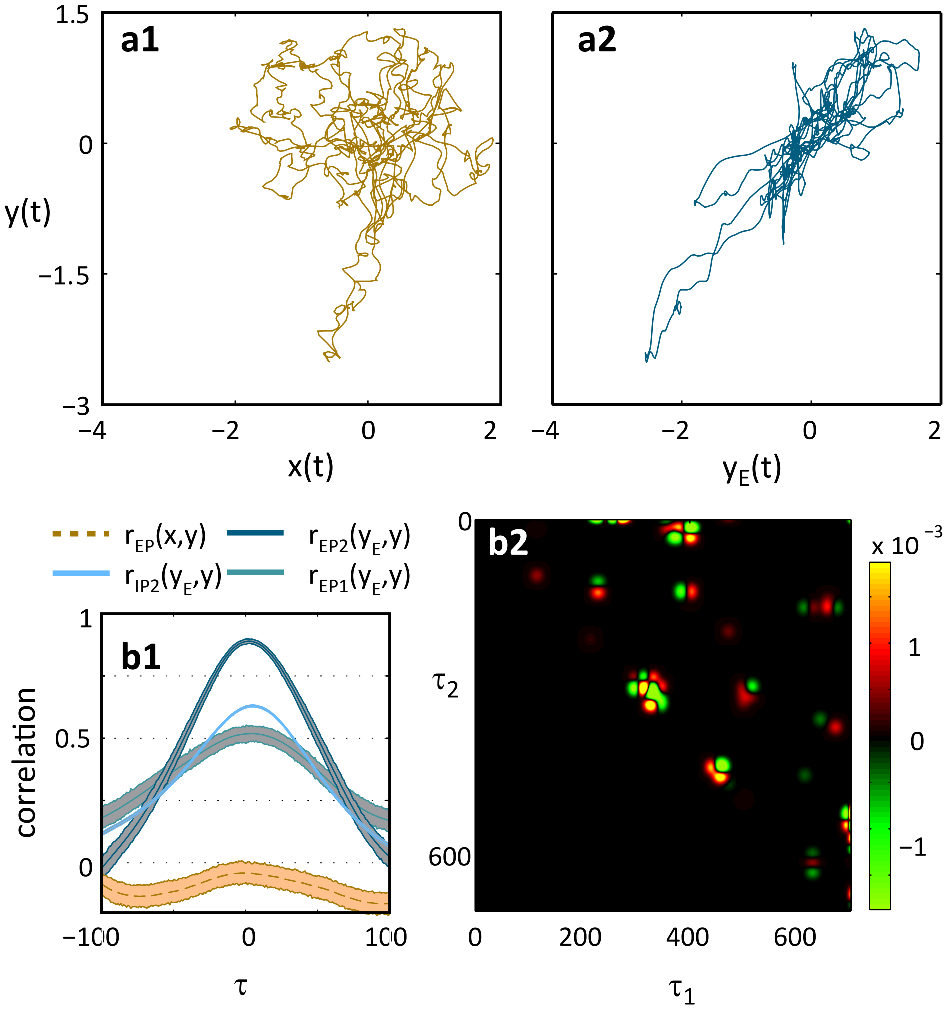

Figure 22: Two macaque V1 LFP recordings x and y recorded from electrodes with different retinotopy. **a)** Interaction manifolds of the EPs. a1: Nonlinear manifold between x and y. a2: Linearized manifold corresponding to $r_{EP2}(y_E, y)$ in b1). **b1)** Lagged correlation coefficient between EPs of x and y ($r_{EP}(x, y)$), and between a 1^{st} order ($r_{EP1}(y_E, y)$) and 2^{nd} order ($r_{EP2}(y_E, y)$) model $y_E = F[x]$ and predicted tetrode y. For the IPs correlations are shown between a 2^{nd} order model y_E and y ($r_{IP2}(y_E, y)$). Lighter coloured areas show the bootstrapped confidence intervals of the respective models. **b2)** Shows the 2^{nd} order kernel.

different parts of the visual field. While the possibility of spatial correlations in the natural scene stimulus causing the synchronization (due to a common factor) is not directly discernable in this setup, we have nonetheless shown that our method could present a powerful tool to investigate these phenomena, as the result would not have been detectable by linear methods.

7.8 CONCLUSION

In summary, we have presented a statistical modeling framework for the detection of nonlinear interactions between time series. Interactions are modeled as Volterra series expanded in basis functions and fit using l1 and l2 regularized maximum likelihood. The method is computationally efficient and yields sparse analytic models of the interaction which generalize to new data. When compared to the *Joint Probability of Recurrence* method (*CPR* respectively) our approach showed higher detection capabilities (often close to fully predictive) for all tested data and synchronization regimes. This was despite our carefully evaluating different JPR (CPR) embedding-space parameters, both manually and algorithmically selected (false nearest neighbours, mutual information criteria) and only comparing the best results with our method. While our main goal is the detection of generalized synchronization, we showed analytically and experimentally how the method generalizes to m:n phase synchronization, the detection of which represents a hard problem in nonlinear data analysis.

One drawback of the current formalism is that it does not capture auto-structure from the target signal $y(t)$. Perhaps more critically, the Volterra series operator cannot model m:n phase synchronization in rare cases of both $m, n > 1$. Both auto-structure and full m:n phase synchronization could be captured by also fitting a second Volterra functional $G[y]$, so that $F[x](t) - G[y](t) = 0$. Using nonlinear synchronization as a formalization of complex interactions is intriguing with respect to information processing in the brain where oscillatory and synchronization phenomena are frequently reported Uhlhaas et al. [2009]. Theoretical studies Pasemann and Wennekers [2000] also show the existence of generalized partial synchronization in a variety of artificial neural networks. In this context, Volterra series could be a natural model of neural transient interactions Friston [2001].

8

D^2IF – A STATISTICAL FRAMEWORK TO INFER DELAY AND DIRECTION OF INFORMATION FLOW FROM MEASUREMENTS OF COMPLEX SYSTEMS

8.1 ABSTRACT

In neuroscience, data is typically generated from neural network activity. The resulting time series represent measurements from spatially distributed subsystems with complex interactions, weakly coupled to a high-dimensional global system. We present here a statistical modeling framework, D^2IF (Delay & Direction of Information Flow), to estimate the direction of information flow and its delay in measurements from systems of this type. Informed by differential topology, Gaussian process regression is employed to reconstruct measurements of putative driving systems from measurements of the driven systems. These reconstructions serve to estimate the delay of the interaction by means of an analytical criterion developed for this purpose. The model accounts for a range of possible sources of uncertainty, including temporally evolving intrinsic noise, while assuming complex nonlinear dependencies. Furthermore, we show that if information flow is delayed, this approach also allows for inference in strong coupling scenarios of systems exhibiting synchronization phenomena. The general validity of the method is established on a variety of delay-coupled chaotic oscillators. In addition, we show that these results seamlessly transfer to real data generated from local field potentials in cat visual areas.

8.2 INTRODUCTION

Dependencies in time series typically arise from interactions of underlying complex systems, which may be delayed and nonlinear. Determining the direction of information flow, as well as the delay of these interactions, is of high interest for many practitioners and presents a difficult task in data analysis.

Different methodological approaches exist to determine the causality of interactions hidden in time series data. The oldest approach is Granger causality [Granger, 1969] which formalizes the problem in terms of stochastic processes. Although originally not intended for dynamical systems, applications of Granger causality in neuroscience have become increasingly popular. While the method can be extended to estimate delays (directed coherence phase [Witham et al., 2010]), it is limited to linear statistical models. More severely, Sugihara for example pointed out that Granger causality may report false negatives when applied to coupled dynamical systems [Sugihara et al., 2012]. In a similar approach, Nolte et al. [2004] investigate delayed interactions of EEG signals using imaginary coherency. A further method established in neuroscience is transfer entropy [Vicente et al., 2011]. Employing information theoretical concepts, the approach does not presume linearity. A more recent application [Wibral et al., 2013] allows for estimation of delays and can be understood as an information-theoretic variant of Granger causality.

Sugihara et al. [2012] introduces in the context of species population dynamics a neighborhood method based on reconstruction. Although not explicitly mentioned by the author, the theoretical foundation of this approach is given by Stark's *skew-product embedding theory* [Stark, 1999], which we also use as basis for the method presented in this article. The neighborhood method, however, is not designed to account for delays, which play a minor role in population dynamics, and is mainly applicable to weakly coupled systems away from regimes of synchronization.

In neuroscience, on the other hand, data is usually composed of local measurements from spatially distributed subsystems, such as local field potentials from electrode recordings. It is clear that distant brain regions will communicate only with significant delay. The latter is therefore highly informative regarding distribution of information and functional connection of the underlying neural network and can be exploited for inference on data. The interactions between different neuronal populations may be highly nonlinear. Furthermore, synchronization and phase-locking phenomena play an important role in many analyses and are often the focus of investigation. One is interested, for example, in determining the direction of interactions between different cortical areas to identify and differentiate top-down modulatory influence from bottom-up signal propagation. At the same time, the measured subpopulations can never be considered as autonomous systems. Instead, one always has to assume possibly very high-dimensional influence from unobserved brain areas as a source of intrinsic noise, dynamically altering the temporal evolution of its targets. In addition, recordings suffer strongly from neural mass background activity and other sources of measurement noise.

We meet these challenges in a statistical modeling approach designed to infer the direction of information flow and its delay in this specific situation. We formalize the detectability of information flow in terms of *reconstructibility* from measurements by considering skew-product embedding theory [Stark, 1999]. The latter suggests that under certain conditions a driver may be reconstructed from measurements of the driven system alone. This gives rise to a statistical model which aims to recover the measurements of a putative driver from measurements of the driven system. We use Gaussian process regression with Volterra series operators modeling functional interactions. The model also accounts for additional non-reconstructible drivers from surrounding parts of the brain that may lead to temporally evolving intrinsic noise in the measured signals. This is achieved by considering the additional theory of *bundle embeddings* [Stark et al., 2003], which formalizes a parametrized version of reconstructibility in the presence of stochastic driver systems. As a result, additional random variables enter the model nonlinearly, which we treat in an exemplary application of Bayesian inference calculus. The stochastic model provides reconstructions of the putative driver at different delays, leading to reconstruction-error graphs for which the theory predicts the lowest error in certain reconstructible areas relative to the delay of the interaction. An analytical criterion allows us to determine the onset of reconstructibility in time, hereby providing a point estimator for the interaction delay, together with an area of confidence. The criterion is tailored to work in situations where the dynamical system is not given by a diffeomorphism, which is usually a standard assumption, but by an endomorphism. This entails dynamical systems that are not invertible in time, such as delay-coupled systems given by *retarded functional differential equations*, which we have to consider as default scenario in our domain of application.

The outline of this article is given as follows. First, we give a gentle introduction to the theoretical concepts from differential topology on which the method is based. This is followed by a step-by-step introduction of the different aspects of the model, including Gaussian process regression and the Volterra series operator which we use to model functional interactions.

In the second part, we first aim to establish the validity of the method by application to synthetic data generated from different delay-coupled chaotic systems with intrinsic noise given by Wiener processes. As measurement time series we use linear combinations of all coordinate functions of the subsystems, which appears to be closest to the case of local field potentials or other types of recordings encountered in neuroscience. The synthetic data sets include standard examples such as coupled logistic maps with intrinsic noise, which we use to practically demonstrate the workings of *bundle reconstructions* in a situation where we can have full knowledge of the stochastic driver. A non-standard example is given by a weakly coupled Lorenz-Rössler system, which illustrates the applicability of the method in case of systems operating on very different time scales. This is followed by a Rössler-Lorenz system in generalized synchronization. Here we show that the special consideration of delays in the interaction makes inference possible even if the systems are very strongly coupled. Coupled Mackey-Glass systems provide examples of endomorphisms where each subsystem is in itself delay-coupled.

The final application of the method is to real data given by local field potentials from anaesthetized cat visual cortices stimulated by visual gratings. We show that the resulting reconstruction-error graphs across delays feature the same interpretable patterns that were previously introduced on the synthetic data sets. Furthermore, our method demonstrates a surprisingly rich insight into the different types of interactions, as well as the most informative parts of the trials in relation to stimulus onset. We present a resulting connectivity graph for the electrodes which appears to be physiologically plausible, both, with respect to the estimated delays, as well as with respect to type and direction of inter- and intra-area interactions.

8.3 METHODS

In this part, we derive the method to estimate delays and directedness of information flow. First, an introduction to the concepts from differential topology that present the theoretical foundation is provided. Second, the statistical model is introduced as Gaussian process with Volterra series operator and adapted to the problem derived beforehand. Finally, reconstruction-error graphs across delays, as supplied by the statistical model, are analyzed and a criterion is developed to estimate the interaction delays.

8.3.1 *Embedding Theory*

In this section we provide a brief summary of the theoretical considerations that underly the presented method. A detailed derivation is given in a corresponding section of the supplementary information. At the heart of the method lies a certain functional mapping F between time series, which we try to estimate directly from the data. The existence of this functional is suggested by theorems from differential topology which we refer to as embedding theory and motivate here as follows.

The problem differential topology solves for the practitioner is that of reconstructing a system that is observed only indirectly via real-valued measurements. Consider, for example, local field potentials (LFPs) from electrode recordings in cortex. These yield a time series measurement of the unobserved neuronal network activity contributing to the LFPs. In general, one considers time-continuous dynamical systems with states x defined on a differentiable manifold M. A system's temporal evolution is defined by a map $\phi_t : M \to M$ which takes an initial condition $x \in M$ and maps it forward in time to $\phi_t(x) \in M$. The continuous change of the system state over time thus defines trajectories on M. *Measurements*, such as LFPs, are defined as continuous functions $f : M \to \mathbb{R}$ that map a system state, as given by $x \in M$, to a real number.

As a function of time, the system ϕ_t is observed only indirectly via the measurements $f(\phi_t(x))$ which constitute the observed time series, indexed by t here. Theorems from differential topology now assert the following. Define a group of measurements by $\text{Rec}_d : M \to \mathbb{R}^d$, called a *reconstruction map*, with image

$$\text{Rec}_d(x) = (f(x), f(\phi_{t_1}(x)), ..., f(\phi_{t_{d-1}}(x))), \tag{123}$$

where the set of time indeces $t_i \in \mathcal{P}$ will be called a *sampling program* $\mathcal{P}$, using the language of Aeyels [1981]. This vector is a part of the full time series and corresponds to system state x. Moreover, this correspondence will usually constitute a one-to-one relationship between reconstruction vector 123 and the underlying system state on M if the reconstruction dimension d is chosen large enough. As a result, the temporal evolution ϕ on M becomes explicitly accessible via the reconstruction vectors. That is, the sequence of reconstruction vectors in $\mathbb{R}^d$, where each vector is associated with a particular sampling point of the time series, is equivalent to the corresponding temporal evolution of system states on M and conveys basically the same geometric information. This was first established by Aeyels [1981]; Takens [1981].

These results have been extended by Stark [1999; 2003] to deal with two more general problems. First, in a situation where the measurements are taken from a non-autonomous system that is parametrized by an additional unobserved driver, a reconstruction map Rec_d can be found that allows the reconstruction of the full system including the driver. That is, the driver is reconstructible from observations of the driven system alone. Second, if the drivers are already known or too high-dimensional to be reconstructed from the data, a parametrized version of Rec_d can be found in principle that still provides a one-to-one relationship with the underlying states of the driven system. These are called *bundle embeddings* and exist also in case the driver is stochastic, i.e. infinite-dimensional. Reconstructions via bundle embeddings can only be realized explicitly, however, if the driver is observed and supplied as parametrization of Rec_d. If the drivers are unknown, the theory still serves to provide the existence of functional dependencies, albeit subject to uncertainty.

With applications in neuroscience in mind, Ørstavik and Stark [1998] have proposed such a framework of *stochastic forcing* to formalize in approximation situations where the measurements are presumably taken from a lower-dimensional local subsystem, weakly coupled to a high-dimensional global system that is unobserved. If the global system, such as the neural network in the brain, is high-dimensional enough, one can view it in this context as practically stochastic. In turn, any locally measured neuronal subpopulation may be regarded as a stochastically forced low-dimensional deterministic system and treated by means of bundle embeddings.

We implement this framework as part of our method in the following way, as visualized schematically by figure 23. Assume finite measurement time series $(f_i)_{i=0}^{D}$ and $(g_i)_{i=0}^{D}$ are given, defined by $g_i = g(\psi^{(i)}(x_0))$. Here, ψ describes the temporal evolution of the initial state $x_0 \in N$ of a driver system, and ϕ describes the temporal evolution of a driven system with initial state $y_0 \in M$. M and N denote differentiable manifolds that represent state spaces of the two systems. Being the driven system, the forward-mapping ϕ is parametrized by states of the driver. Furthermore, we assume both systems are subject to additional stochastic forcing ω and ω', respectively, which further parametrize the forward-mappings. The stochastic forcing $\omega, \omega' \in \mathbb{R}^D$ play here the role of non-reconstructible high-dimensional input, for example from other brain areas in case of LFP recordings.

Given a reconstruction vector Rec_d of the driven system, Stark's skew-product and bundle embedding theory now suggest the existence of a functional mapping $F : \mathbb{R}^d \to \mathbb{R}$ that maps covariate measurement samples f_j of the driven system (Rec_d) to measurement samples g_i of the driver (see figure 23 for illustration). This mapping is realized by first using Rec_d to reconstruct the full system of driver and slave, i.e. the *product manifold* $N \times M$, projecting down into the driver system N (denoted pr_N), and finally mapping into the measurement g_i of the corresponding driver state. Here i, j index the points in time corresponding to the measurement samples. As will be illustrated in more detail in a later section and in the supplementary information, F will typically exist for a number of $i \leq j$. The mappings involved in the realization of F are in addition parametrized by the stochastic forcing ω, ω', as a result of the bundle reconstruction map Rec_d.

The dependencies on the unknown drivers can be made explicit, analogous to Stark's discussion of the NARMA model in [2003]. This yields the *bundle reconstruction*

$$\begin{aligned} F : \mathbb{R}^d &\to \mathbb{R}, \\ \text{Rec}_{d,\omega,\omega'}(y_j) &\mapsto g(x_i), \\ F(\text{Rec}_{d,\omega,\omega'}) &= F(f_j, ..., f_{j+d-1}, \omega'_j, ..., \omega'_{j+d-2}, \omega_{i+1}, ..., \omega_{i+d-2}), \end{aligned} \tag{124}$$

as illustrated in figure 23. A more detailed mathematical derivation is given in the corresponding section of the supplementary information. At the heart of the method proposed here, we estimate the functional mapping F directly from the data. The resulting candidate reconstructions will be used to estimate delay and direction of the interaction underlying two measurement time series, as explained in a later section. If ω_i and ω'_j are taken to be one-dimensional, we can immediately include them as random variables in a statistical model. A proper treatment of this source of

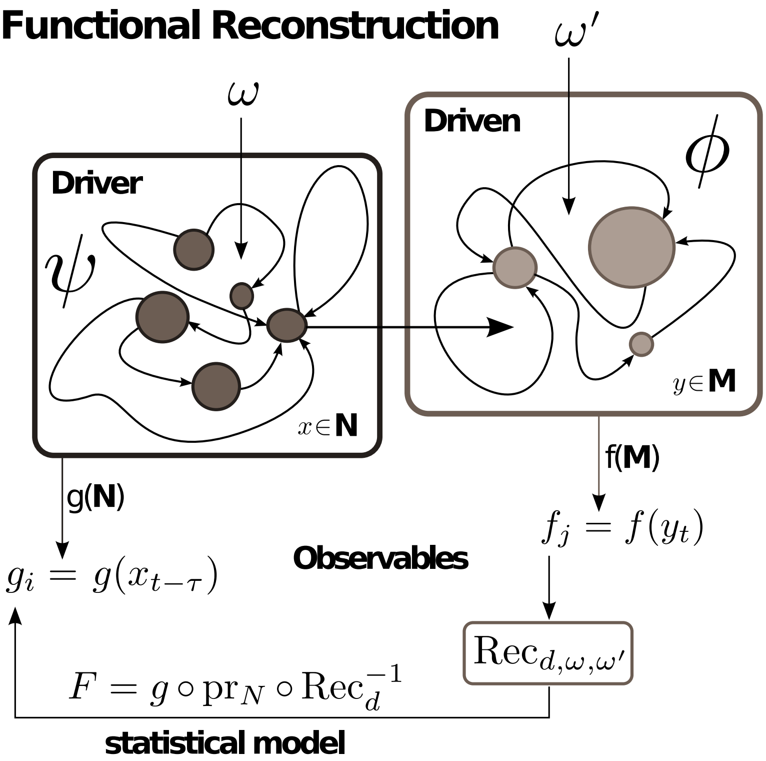

Figure 23: Functional reconstruction mapping. The figure illustrates mapping F (124) between time series $(f_i)_{i=0}^D$ and $(g_i)_{i=0}^D$ which represent real-valued measurements of dynamical systems ϕ and ψ, evolving on manifolds M and N, respectively. ω and ω' denote stochastic input that parametrizes the forward mapping and simulates high-dimensional non-reconstructible input. F is the composition of the inverse reconstruction map (123) which yields a point on the product manifold $M \times N$. This is followed by the canonical projection $\mathrm{pr}_N : M \times N \to N$, the image of which is the domain of measurement g.

uncertainty with Bayesian techniques will be discussed after the statistical model has been introduced.

With regard to inferring the direction of the interaction, the crucial point is an asymmetry in reconstructibility. Under the hypothesis $x \to y$ (x drives y), the driver measurements $g(x)$ are reconstructible via F from measurements $f(y)$. However, F only exists if $x \to y$ obtains, since measurements of the driver do not contain information about the driven system and will therefore fail to reconstruct the autonomous dynamics of the latter. The general idea of formalizing causal dependencies in terms of reconstructibility was, to the best of our knowledge, first proposed by Sugihara et al. [2012], albeit not in the context of a functional model. In case of stronger forms of synchronization, also called *generalized synchronization* (GS), this approach becomes problematic, since knowledge of the driver would afford by definition (see Kocarev and Parlitz [1996]) a perfect prediction of the driven system. The latter succumbs in this situation to being a mere transformation of the driver. However, as will be shown later, if the interaction is delayed, inference may yet be possible.

In the results section, we practically demonstrate the soundness of the mapping F using coupled chaotic logistic maps with intrinsic noise. It is shown that in case the noise sequences ω and ω' are observed and included in F, the time series reconstruction becomes exact.

8.3.2 *Statistical Model*

In this section we derive the statistical model that uses *bundle reconstruction* 124 to compute estimators of the measurement time series of putative drivers, given measurement time series of the driven system as covariates. To this end, we first discuss the realization of F (124) by Volterra series operators. Formalized by Gaussian process regression, the model is treated with Bayesian inference calculus, axiomatically rooted in decision theory (see [Lindley, 1972; Berger, 1985]). In this statistical framework we are able to elegantly deal with the uncertainty introduced into equation 124 by ω and ω', as well as additional measurement noise.

8.3.2.1 *Volterra Series Operator*

In order to derive point estimators for the measurement time series, the function $F : \mathbb{R}^d \to \mathbb{R}$ (124) has to be approximated by a model. F being continuous exhausts prior knowledge about its functional form. A choice that may be called canonical in this situation is a *discrete Volterra series operator*. The term Volterra series was originally coined in the context of functional analysis [Volterra, 1915]. In general, it can loosely be thought of as Taylor series expansion of F whose polynomial terms are rewritten in a certain coordinate representation. The result is a linear form with respect to its parametrization, which is a desirable property in a statistical model because it simplifies computation. For the discrete case, the model can be stated in the form

$$\begin{aligned} p(x) = & h_0 + \sum_{k_1=0}^{d-1} h_1(k_1)x(k_1) + \\ & \sum_{k_1=0}^{d-1}\sum_{k_2=0}^{d-1} h_2(k_1,k_2)x(k_1)x(k_2) + \ldots, \end{aligned} \tag{125}$$

where $x = (x_0, \ldots, x_{d-1})^T \in X \subset \mathbb{R}^d$ and $x(i) = x_i$.

Although we cannot assume that F is differentiable, it is easy to show that the Volterra series operator can approximate arbitrary continuous F of the required type. We give a proof of this fact and a detailed derivation of the operator in the corresponding supporting information. Although a host of literature exists on the topic of Volterra series in general (see for example [Rugh, 1981] and [Franz and Schölkopf, 2006]), we feel that a problem-based derivation of the discrete operator in the context of statistical modeling is often lacking, despite the fact that the discrete setting is natural in the context of data analysis. In this context, the derivation of the discrete Volterra series expansion is comparably easy and largely based upon theory from commutative algebra.

In summary, the choice of Volterra series is quite natural in the present context: The functional form of the model has an intuitive derivation via Taylor series expansions and, employed in a statistical model, it will allow us to approximate arbitrary continuous F, as given by the desired functional mapping 124, without any modeling bias. This is very important for the method being developed here in order to be able to interpret a *lack of reconstructibility* as a true *lack of information*. Furthermore, the series expansion can be truncated at a particular order (equation 125 shows summands up to order 2). This enables one to control model complexity via the nonlinearity of the series, which is important to be able to deal with overfitting in case relevant information is represented less than optimally in the data. The h_i in equation 125 are kernel functions, or rather coefficients, which are subject to uncertainty in a finite parametric statistical model, where the series expansion has to be truncated at a certain order. When formulated as Gaussian process, the model can also be stated in nonparametric form with infinite series expansion, as will be discussed in the next section.

8.3.2.2 *Gaussian Process Regression*

So far, we have identified a number of sources of uncertainty that arise when trying to reconstruct a putative driver measurement time series. These include measurement noise, hidden stochastic drivers, as well as the particular functional form of the bundle reconstruction F in 124. In a Bayesian approach, this uncertainty is summarized in a probability distribution. If the task is to predict a target time series $y \in \mathbb{R}^D$ in terms of a covariate time series $x \in \mathbb{R}^D$, the uncertainty in the prediction of unseen data y_* given covariate time series x_* is summarized in a predictive distribution $P(y_*|x_*, x, y, H)$. The predictive distribution incorporates knowledge of the given data $\mathcal{D} = (y, x)$ to infer y_* given x_*. We denote by H the particular modeling assumption that has to be made and the corresponding hyperparameters. The covariance structure of this distribution represents the model uncertainty and supplies confidence intervals.

We formalize the statistical model as

$$\begin{aligned} y_j &= F(x_i, ..., x_{i+d-1}, \omega_0, ..., \omega_h) + \epsilon_j \\ &= F(z_j) + \epsilon_j, \end{aligned} \tag{126}$$

where $z_j \in \mathbb{R}^{d+h}$, $\epsilon_j \sim \mathcal{N}(0, \sigma_\epsilon^2)$ and $\omega_k \sim \mathcal{N}(0, \sigma_\omega^2)$. While ϵ represents measurement noise of the target time series, ω pertains to uncertainty from hidden stochastic drivers and covariate measurement noise, as implied by embedding theory. Although we may have no reason to believe that the ϵ_j are actually normally distributed, one can put forward strong arguments that the normal distribution nonetheless represents our state of knowledge optimally, unless information about higher moments of the sampling distribution is available [Jaynes and Bretthorst, 2003]. In practice, the latter will usually not be the case, since there is no explicit access to the residuals. Regarding the ω_k, we can safely assume in the present context that they are bounded, as indicated by a finite variance. Beyond that no sensible assumption can be made, so that the normal distribution is again a natural choice given our limited prior knowledge (it is in fact the distribution that maximizes *entropy* in this situation [Hida and Hitsuda, 2007], see also [Jaynes, 1968]).

By formalizing this model as a Gaussian process, we can treat uncertainty with regard to F as uncertainty in the function itself rather than withdrawing to uncertainty in a finite parametrization of F. A Gaussian Process is a system of random variables indexed by a linearly ordered set, such that any finite number of samples are jointly normally distributed Hida and Hitsuda [2007]. A Gaussian Process thus defines a distribution over functions and is completely specified by a *mean function* and a *covariance function*. In terms of a Gaussian process, we can define the functional model $F \sim \mathcal{GP}$ as

$$\begin{aligned} \mathbb{E}[F] &= 0, \\ \mathrm{Cov}[F(z_j), F(z_k)] &= \mathbb{E}[F(z_j)F(z_k)] =: [K(z,z)]_{jk}. \end{aligned} \tag{127}$$

where $[K]_{jk}$ denotes coordinate (j,k) of matrix K.

One can now derive all distributions necessary to formalize the uncertainty associated with data $\mathcal{D}$. The details of these derivations can be found, for example, in Rasmussen and Williams [2006]. The covariance matrix K_y corresponding to y is given by

$$K_y := \mathrm{Cov}[y_j, y_k] = K(z,z) + \sigma_\epsilon^2 I.$$

Accordingly, for data (y_*, x_*) (which may be the same as (y, x)),

$$\begin{pmatrix} y \\ y_* \end{pmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} K(z,z) + \sigma_\epsilon^2 I & K(z,z_*) \\ K(z_*,z) & K(z_*,z_*) \end{bmatrix}\right). \tag{128}$$

In the Gaussian process framework, this distribution specifies a prior over bounded functions. From (128), the marginal likelihood can be derived as $y|z \sim \mathcal{N}(0, K_y)$ with log-density

$$\log[p(y|z)] = -\frac{1}{2} y^T K_y^{-1} y - \frac{1}{2}\log[|K_y|] - \frac{D}{2}\log[2\pi]. \tag{129}$$

The marginal likelihood is important for Bayesian model selection, e.g. in computing a *Bayes Factor* or *Posterior Odds*. Furthermore, the desired predictive distribution which summarizes our uncertainty in the reconstruction of putative drivers can be derived as

$$P(y_*|z_*, z, y) = \mathcal{N}(m_*, S_*), \tag{130}$$

with

$$\begin{aligned} m_* &:= \mathbb{E}[y_*] = K(z_*,z)K_y^{-1}y \\ s_* &:= \mathrm{Cov}[[y_*]_j, [y_*]_k] = K(z_*,z_*) - K(z_*,z)K_y^{-1}K(z,z_*). \end{aligned} \tag{131}$$

From a *decision theoretic* point of view, it can be shown that m_* provides a point estimator that minimizes the *expected* squared-error loss to target time series y_* Berger [1985], which we therefore employ for reconstruction.

To compute estimators of the target time series, we now only need to specify $K(z,z)$ corresponding to the Volterra series model $p(z_j)$ in equation 125. This is a straightforward derivation using equations 125 and 127. For details and a more general discussion of Volterra theory in polynomial regression models, see [Franz and Schölkopf, 2006]. If the series expansion is truncated at order n, the covariance matrix is essentially constructed as

$$[K^{(n)}(z,z)]_{ij} = (1 + z_i^T z_j)^n = \sum_{k=0}^{n} \binom{n}{k} (z_i^T z_j)^k, \tag{132}$$

but may have to be extended by certain hyperparameters depending on the prior assumptions on the h_k in equation 125. For the nonparametric model, the kernel functions h_k have to be reparametrized by a scaling factor of $1/\sqrt{k!}$ (see [Steinwart, 2002]). As a result, it is possible to compute the limit

$$[K^{(\infty)}(z,z)]_{i,j} = \lim_{n\to\infty} \sum_{k=0}^{n} \frac{1}{k!}(z_i^T z_j)^k = \exp(z_i^T z_j). \tag{133}$$

Next, we have to deal with the unknown ω_k from equation 126. Denote by $\mathbf{w}_j = (\omega_0, ..., \omega_h)^T \in \mathbb{R}^h$ the vector of unknowns, and by $\mathbf{x}_j(x_i, ..., x_{i+d-1})^T \in \mathbb{R}^d$ the vector of known covariates in equation 126, such that $z_j = (\mathbf{x}_j, \mathbf{w}_j)$. The obvious strategy is to integrate $\mathbf{w}$ already out of the prior process 128, and perform the inference step in the predictive distribution with the resulting process, independent of $\mathbf{w}$. However, the resulting process is no longer Gaussian. Although the ω_k have been assigned a normal prior, they enter via K_y nonlinearly into the prior process likelihood 129. As a result, the analytic derivation of the predictive distribution in 131 no longer applies.

In the context of noisy covariates in Regression, however, Girard et al. [2003] suggest a *Gaussian approximation*. This is done by computing expectation and covariance of the noise-driven process explicitly while still assuming a normal distribution. They show that although the actual distribution (as approximated by MCMC methods) is far from normal, the expected value of the Gaussian approximation predictive distribution yields a comparable predictor, and the predictive variance captures the model uncertainty quite well, allowing for reasonable posterior density credible sets.

We will adopt the same approach here and compute by the *tower property of conditional expectation* and the *law of total covariance* [Klenke, 2007]

$$\begin{aligned} m(\mathbf{x}_i) &= \mathbb{E}[y_i|\mathbf{x}_i] = \mathbb{E}_w[\mathbb{E}[y_i|\mathbf{x}_i, \mathbf{w}_i]] = 0, \\ s(\mathbf{x}_i, \mathbf{x}_j) &= \operatorname{Cov}[y_i, y_j|\mathbf{x}_i, \mathbf{x}_j] \\ &= \mathbb{E}_w[\operatorname{Cov}[y_i, y_j|\mathbf{x}_i, \mathbf{x}_j, \mathbf{w}_i, \mathbf{w}_j]] + \operatorname{Cov}[\mathbb{E}[y_i|\mathbf{x}_i, \mathbf{w}_i]\mathbb{E}[y_j|\mathbf{x}_j, \mathbf{w}_j]] \\ &= \mathbb{E}_w[\operatorname{Cov}[y_i, y_j|\mathbf{x}_i, \mathbf{x}_j, \mathbf{w}_i, \mathbf{w}_j]] = \sigma'[K(x,x)]_{ij} + \sigma^2\delta_{ij}. \end{aligned} \tag{134}$$

This defines a new prior process which will be treated as approximately Gaussian. The new prior no longer depends on $\mathbf{w}$, which has been integrated out using the normal priors of the ω_k. To determine $s(\mathbf{x}_i, \mathbf{x}_j)$, the integrals occurring in the expectation with respect to $\mathbf{w}$ have to be solved. We show the necessary calculations exemplary for the nonparametric model 133 in a corresponding section of the supplementary information. These calculations yield a new hyperparameter, σ'. Interestingly, this hyperparameter has an effect equivalent to L_2-regularization (see [Rasmussen and Williams, 2006; Hoerl and Kennard, 1970]) and helps to prevent overfitting. With this newly defined prior process, one computes the predictive distribution $p(y_*|x_*, \mathcal{D}, H)$ in Gaussian approximation in the same way described before, according to equation 131.

In a fully Bayesian treatment, the hyperparameters $\sigma_\epsilon^2, \sigma'$ would have to be assigned (non-informative) priors and be integrated out of the distributions relevant for inference or model selection, too. However, in general it is not possible to get rid of dependence on both, hyperparameters and model parametrization. Instead, explicit values for the hyperparameters may be estimated by maximizing, for example, the marginal likelihood (129), or the predictive distribution in a leave-one-out cross-validation scheme Sundararajan and Keerthi [2001]. We employ the latter, which is less prone to overfitting and less dependent on model specifications.

The remaining free parameters of our statistical model, indicated by H in the conditioning part of the predictive distribution, pertain to d, the size of the reconstruction space, and the order of the Volterra series expansion. These have to be determined in a model selection process, where for each choice of H the hyperparameters are estimated individually.

As a final note, we discuss a preprocessing step necessary in time series analysis to avoid overfitting. In time series that represent measurements of systems evolving continuously in time, subsequent samples tend to be highly correlated. The higher the sampling resolution, the smaller the difference between neighboring sample points. If the time series are in this sense oversampled, there is a strong possibility for overfitting, even in the leave-one-out cross-validation. This is a result of equation 131 which provides the point estimator m_*. This point estimator is essentially only a product of a so-called *smoother matrix* with the data vector y. The latter contains in this situation very redundant information in the individual samples on a short timescale, whereas more complex dynamic features on longer timescales may be underrepresented in y. In particular, for each target y_* a near identical covariate in close temporal proximity may be found in y. Modeling the y_* on a short timescale therefore becomes trivial and does not reflect functional dependencies on longer more relevant timescales that one is really interested in.

To counter this problem, the time series have to be down-sampled at a relevant timescale. A criterion for establishing what a relevant timescale is, is to compute the auto mutual information of the time series. This is also a standard procedure to determine proper lags between samples in reconstruction vectors for embeddings (compare with sampling program $\mathcal{P}$ in definition 3). In practice, jointly down-sampling the time series until mutual information between neighboring samples is less than 1 is desirable. If this is not possible, one seeks for convergence of mutual information as a function of the lag between subsequent samples. If the two time series evolve on very different timescales, it may happen that the covariate time series is still oversampled as compared to the target time series. In this case one may yield a bad model performance if the dimensionality d of the covariate vector Rec_d is chosen too low. The elements of Rec_d need to span a relevant part of the time series with respect to the timescale, otherwise geometric information is "squashed" in the reconstruction. We discuss such a situation in the results section. The solution is to either increase d, or to introduce additional lags between the samples in Rec_d.

Since we assume that interaction between driver and driven system may occur with a substantial but unknown delay τ, it is not yet clear, however, how to pair the temporal indeces of a sampling program $\mathcal{P}$ for covariate vectors Rec_d with corresponding targets in the putative driver measurements (compare with definition 3 and functional mapping 124). Since cause and effect can only propagate forward in time, it is clear that, given a covariate vector $\mathrm{Rec}_d = (x_{i-d-1}, ..., x_i)$, only targets y_j with $j \leq i$ are candidates for reconstruction.

8.3.3 *Estimation of Interaction Delays*

At this point, a statistical model has been devised which allows the reconstruction of a putative driver measurement time series out of another time series representing the driven system. The model is defined under the assumption that both measured systems are lower-dimensional local subsystems, which are weakly coupled to a higher-dimensional global system that is not reconstructible from the data. The statistical model accounts for uncertainty pertaining to measurement noise, as well as dynamically evolving intrinsic noise resulting from the hidden global drive. This formalizes the situation which one encounters in neuroscientific data, such as local field potentials, which represent measurements from local neuronal subpopulations, embedded in a global brain-network which can be treated as practically infinite dimensional.

In a first step towards estimating the delay, for a particular choice of d and $\mathcal{P}$, we create reconstructions of the putative measurement time series for a candidate set of negative delays $[..., -2, -1, 0]$ by which we shift the target series. Figure 24A shows the situation for a particular covariate reconstruction vector Rec_d corresponding to time t. If the true interaction delay is τ, the earliest reconstructible target sample is y_{j^*}, where $j^* = \text{argmax}_j \, j \leq t - \tau$. We can assume y_{j*} is always reconstructible, since it relates to the last temporal index in the sampling program $\mathcal{P}$. This fact is discussed in detail in the supplementary information section in the context of definition 3.

As a brief summary, we note that the dynamical systems that are usually considered in embedding theory admit a temporal evolution that is invertible in time. Given a certain initial condition, past and future of the system are therefore uniquely determined. As a result, given a reconstruction vector Rec_d, system states corresponding to all individual sampling points in Rec_d are in principle reconstructible. This is owed to the fact that for a time-invertible system, these states can be uniquely related forward and backward in time to all samples in Rec_d. In case of systems that are not invertible in time, this is no longer true. This case is important to consider for us since delay-coupled systems, such as spatially distributed neural networks, are in general not time-invertible (the time-reversed system would be acausal). In later work, Takens [2002] showed that if the forward-mapping ϕ of a dynamical system is not invertible, it may not even admit embedding. However, he proved that the system state corresponding to the last sample in the reconstruction vector Rec_d is in general reconstructible. We stress this fact once more in summary: The system state corresponding to the end of the reconstruction vector is always reconstructible and we will exploit this fact here. We must note, however, that this is a conjecture insofar as the corresponding proof in [Takens, 2002] does not explicitly treat the skew-product scenario we consider here. Nevertheless, the results presented in the next section strongly suggest that this is a valid generalization.

We associate with j^* the *onset of reconstructibility* which is informative about the delay of the interaction. Note that if $d > d^*$, where d^* denotes the true intrinsic dimensionality of the underlying product manifold, the reconstructibility of j^* will extend to the $d - d^*$ preceding samples of the target series. The reconstructibility of y_j for $j < j^*$ depends on how uniquely the inverse temporal mapping of the underlying system can be inferred from the data. In any case, the last reconstructible

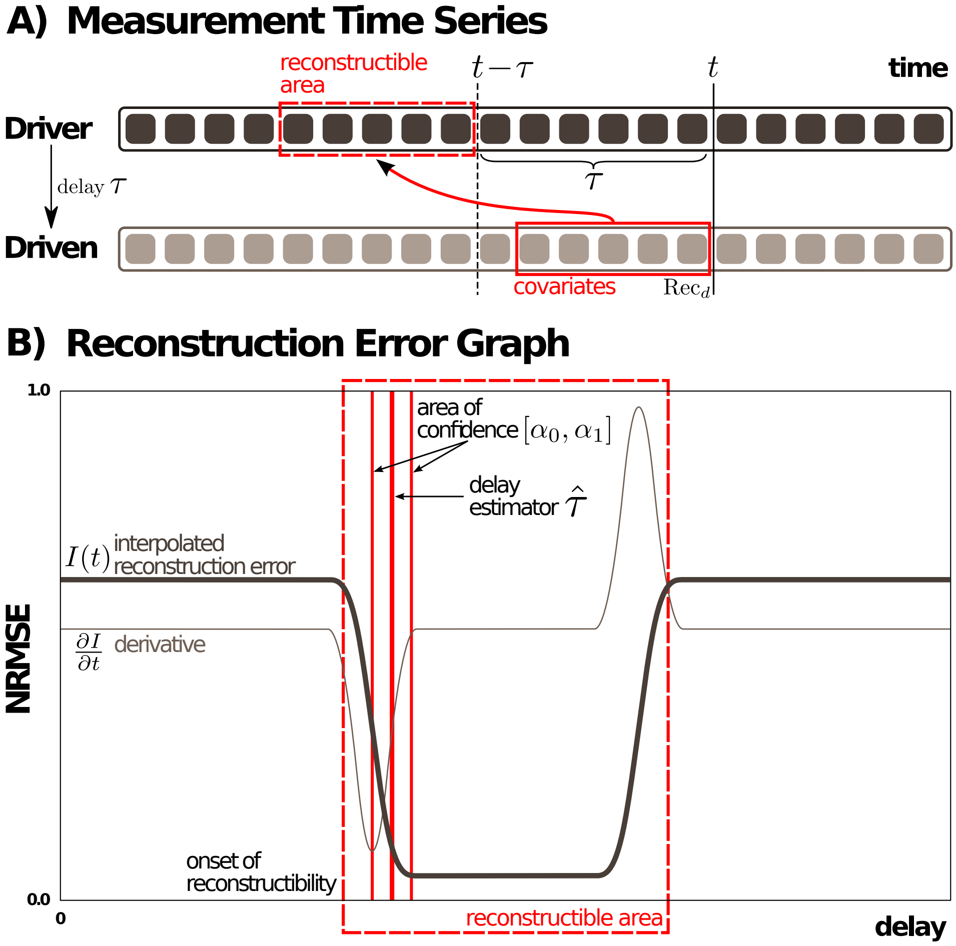
A) Measurement Time Series
reconstructible area
$t-\tau$
t
time
Driver
delay τ
τ
Driven
covariates
Rec_d
B) Reconstruction Error Graph
1.0
0.0
0
NRMSE
delay
area of confidence $[\alpha_0, \alpha_1]$
delay estimator $\hat{\tau}$
$I(t)$ interpolated reconstruction error
$\frac{\partial I}{\partial t}$ derivative
onset of reconstructibility
reconstructible area

Figure 24: Delay Estimation. **A)** The time series of the putative driven system provides covariate vector Rec_d (123) as input for the statistical model realizing functional mapping F (124). Targets of the mapping are candidate measurements of the driver taken before time t. If the interaction has a delay of τ, the set of points in time before $t - \tau$ index reconstructible driver measurements. **B)** The interpolation $I(t)$ of an idealized reconstruction error graph (135) is shown, together with its first derivative. $I(t)$ is plotted against increasing candidate delays the order of which is mirrored in comparison to subfigure A. The dotted red rectangle corresponds to the same reconstructible area already marked in subfigure A, with left boundary, called the *onset of reconstructibility*, in close correspondence to the point $t - \tau$. The wider vertical bar represents delay estimator $\hat{\tau}$ (139). Lighter vertical bars indicate the associated *area of confidence* for the delay with boundaries α_0, α_1 given by formulas 137 and 138, respectively.

point is y_{j^*-d-1}, associated with the first element of $\mathcal{P}$. We call the temporal indeces $\{j|j^* - d - 1 \leq j \leq j^*\}$ the *reconstructible area*, marked by the dashed-blue rectangle in figure 24A.

The reconstructible area corresponds to the information about the driver present in the covariate measurements of the driven system by way of actual information exchange via the delayed coupling. Outside of the reconstructible area, the candidate reconstructions of the statistical model become predictions backward or forward in time, respectively, which are necessarily based upon the closest temporal state reconstructible with the given covariate vector Rec_d. Due to the global stochastic drivers, predictions will be subject to further uncertainty, in addition to the uncertainty entering Rec_d. We can thus expect to have the lowest variance in candidate reconstructions within the reconstructible area. Besides, predictability is likely to decline rapidly away from the closest system state that is actually reconstructible. The variance will therefore be our criterion to evaluate reconstructibility.

We estimate the variance of the reconstruction at a particular candidate delay in two steps. First, for each y_j in the delay-shifted target time series the leave-one-out (LOO) point estimator $\hat{y}_j := m_{\setminus(j)}$ from equation 131 is computed conditional on the data set $\mathcal{D}\setminus\{x_j, y_j\}$ where the j^{th} sample is removed. Second, we compute the mean-squared error between the candidate reconstruction and the target. This yields a LOO cross-validation estimator of the variance which is asymptotically correct [Konishi and Kitagawa, 2008] and will be lowest for candidate delays within the reconstructible area. To be able to compare the reconstructibility in one causal direction against another, we report the normalized root-mean-squared error (NRMSE), given by

$$\text{NRMSE}(x, y) = \sqrt{\frac{\sum_{i=1}^{D}(x_i - y_i)^2}{D\sigma_y^2}}, \tag{135}$$

where σ_y denotes the standard deviation of target time series y. Note that

$$\text{NRMSE}(\langle y \rangle, y) = \sigma_y / \sigma_y = 1,$$

where $\langle y \rangle$ denotes the mean of time series y. We therefore choose 1.0 as the upper bound for reconstructibility, in which case the reconstruction is as good as fitting a horizontal line to y, indicating a lack of interesting functional dependencies.

In addition, it is important to establish confidence intervals for the NRMSE which incorporate the uncertainty of the underlying statistical model. The latter is expressed through the covariance matrix of the process in equation 131, which is readily computable. The LOO point estimator yields a cross-validated estimator $\hat{y} \in \mathbb{R}^D$ of the target time series y, for which we report $\text{NRMSE}(\hat{y}, y)$. Since the $y_j \in y$ are by assumption 126

$$y_j = \hat{y}_j + \epsilon_j, \tag{136}$$

one way to quantify the variability of the NRMSE is to resample the full predictive process as $\hat{y}^{(b)} \in \mathbb{R}^D$ using the covariance structure resulting from 131, for $b = 1, ..., B$. For each b, we can compute $\text{NRMSE}(\hat{y}, \hat{y}^{(b)})$ and derive percentiles for this statistic from its resulting empirical distribution. However, recall from the previous section that the assumption of normality in the predictive process is only approximate. While this does not pose difficulties for confidence intervals of the individual $\hat{y}_j$, the nonlinear NRMSE statistic presents a problem and we found that the empirical distribution of $\mathrm{NRMSE}(\hat{y}, \hat{y}^{(b)})$ often yields confidence intervals which do not contain $\mathrm{NRMSE}(\hat{y}, y)$. To counter this problem, we compute the non-normal distribution of residuals in equation 136 empirically from the data and employ bootstrapping [Efron and Tibshirani, 1994] to generate $B = 50000$ sample sets $\hat{y}^{(b)}$ to yield a distribution for $\mathrm{NRMSE}(\hat{y}, \hat{y}^{(b)})$. Confidence intervals are computed conservatively from the 99 percentiles. Where computationally feasible, the resulting confidence intervals may be improved by the $BC\alpha$-method [Efron, 2003] which employs additional correction for deviations from normality.

Ideally, one would expect a reconstruction-error graph (REG) across candidate delays as shown by the dark brown graph in figure 24B, where the reconstructible area corresponds to a *sink*. The lowest NRMSE need not be either at the beginning or the end of the reconstructible area and is therefore not a good indicator for the delay. In fact, the sink need not even be a level plain but may contain additional structure, in particular if the timescales of the unobserved systems differ substantially. Furthermore, as was discussed before, we cannot in general expect full reconstructibility for all candidate delays in the reconstructible area. The only reliable indicator of the true interaction delay τ therefore is the onset of reconstructibility in the REG.

A further problem is the fact that the time series have been sampled down substantially. In order to achieve sub-sampling accuracy in the delay estimation, the REG has to be interpolated in a first step. In a second step, *curvature* and *slope* of the resulting interpolated graph will be used to determine the onset of the sink. Consequently, the model used for interpolation has to exhibit a certain smoothness and be at least twice differentiable. It is important to avoid overfitting in order to yield a simple sink shape, where possible. In our case, samples of the REG will be coarsely spaced across delays. Overfitting leads here to extreme slope and curvature between samples that would interfere with the delay estimation.

The best interpolation results were achieved by fitting a cubic spline function [Boor, 2001]] as Gaussian process parametrized by time (see equation 127). The spline function is a linear combination of basis splines and fully determined in terms of a knot sequence which defines the starting points of the basis splines' compact carriers. As knot sequence, we choose the given samples of the REG. For interpolation, we compute m_y (see equation 131) on a very fine grid across delays, conditional on the *linear interpolation* of the REG samples on this grid. The latter enforces a type of regularization between the actual samples which asserts a certain well-behavedness of the graph in these underspecified areas. This dramatically reduces overfitting and non-informative curvature in the interpolation. We interpolate both, confidence intervals and the REG itself in this fashion. If, on the other hand, the REG is already sampled at a fine resolution, a better approach would be to use a Gaussian process with stationary covariance function $s(x_i, x_j) = \exp((x_i - x_j)^2/(2l))$, where the length-scale l controls how strongly the model interpolates.

Denote the interpolated REG by $I(t)$ with $t \geq 0$. I is a smooth function defined on an arbitrarily fine timescale, as given by the dark brown graph in figure 24B. To determine the onset of reconstructibility and an area of confidence for the true delay

τ, we first compute numerically the derivative of I and choose as left boundary for the area of confidence its global minimum

$$\alpha_0 = \text{argmin}_t \left(\frac{\partial I}{\partial t}\right). \tag{137}$$

This corresponds to the highest rate of negative change in the REG and reliably marks the onset of the sink if one exists. To check for existence, it is sufficient to verify that the first NRMSE significantly larger than the minimum of the REG is to the right of α_0.

As right boundary α_1 of the area of confidence for τ we target the hypothetical point where the sink becomes level ($\frac{\partial I}{\partial t} = 0$), right after its left corner (see figure 24B). We can have very high confidence that the true τ resides within $[\alpha_0, \alpha_1]$, since it would mark the corner of the sink in case sampling precision and reconstruction were perfect. We found the most reliable way to compute α_1 in practice is by first detecting the corner following α_0 and then choosing the closest point where the *curvature* has relaxed *sufficiently*. Following Wang and Brady [1995], the corner is found by, first, computing a curvature score as

$$C(t) = \frac{\partial^2 I(t)}{\partial t^2} - c\left(\frac{\partial I(t)}{\partial t}\right)^2,$$

and, second, finding $t_C := \text{argmax}_t\, C(t)$ for $t > \alpha_0$. The parameter c determines the sensitivity to curvature, which we set to 20 to be able to deal with shallow sinks. Consequently,

$$\alpha_1 = \text{argmax}_t\, C(t) \leq C(t_C) - \epsilon, \tag{138}$$

where $\epsilon = 0.2\,\text{std}(C)$ achieves robustness against uninformative structure of I within the sink. Finally, a simple point estimator for the true delay τ is obtained as

$$\hat{\tau} = \frac{\alpha_0 + \alpha_1}{2}. \tag{139}$$

This estimator works very well in practice and allows for fully automatic determination of interaction delays. If the reconstructible area is marked by a shallow sink, the area of confidence will be larger, the point estimator 139, however, may still yield accurate results. If the sink has a steep onset, the method is very precise. Most problematic are cases where the corners of the sink are less well pronounced and the curvature relaxes only very slowly. In this case and for all computations described above, $\frac{\partial I(t)}{\partial t} \approx 0$ provides an upper bound for the t under consideration. Figure 24B summarizes and illustrates the approach schematically for the hypothetical ideal case.

To determine the preferred direction of information flow, we compute the REG in both possible directions of information flow (either $x \to y$ or $y \to x$), determine the minimum NRMSE across delays and check whether they are significantly different by means of the bootstrapped 99% confidence intervals. Further details will be discussed using concrete examples in the next part of this article.

8.3.4 *Experimental Procedures*

Experimental data presented here were obtained from one adult cat. All procedures in this study were approved by the ethics committee of the state of Hessen in accordance with the guidelines of the German law for the protection of animals. Details about the experimental protocol can be found in [Wunderle et al., 2013]. In brief, the animal was initially anesthetized with a mixture of Ketamin/Medetomidine (10mg/kg and 0.02mg/kg) and subsequently maintained by artificial ventilation of O2/N2O (30%/70%) supplemented with isoflurane (1.5 - 0.7%). Analgesia was ensured by continuous intravenous infusion of Sufentanil (4μg/kg/h) together with a ringer solution. Depth of anesthesia was monitored by regular inspection of heart rate, body temperature and expiratory CO2. Two small craniotomies were made over the visual areas 17 and 21a according to stereotactic coordinates. Through each craniotomy, a single shank multi-contact laminar probe was implanted into the brain tissue (1MΩ, Neuronexus, MI, USA) in order to record from all cortical layers. The recorded signals were band-pass filtered (0.7 – 300Hz) and downsampled (1024Hz) to obtain the LFP signal. After the surgical procedures, the animal was stimulated with gratings drifting in 8 directions (45° steps) presented with a 22" LCD monitor (Samsung SyncMaster 2233RZ) in front of the cat. The gratings were at full contrast and the temporal (1.8Hz) and spatial frequency (0.4cy/deg) was chosen to evoke responses in both areas, 17 and 21a.

8.4 RESULTS

We have established the validity of the method in practice on different delay-coupled chaotic systems. These involve systems that are time discrete or continuous, irreversible in time, governed by retarded functional differential equations, exhibit generalized synchronization, and total systems where the coupled subsystems operate on very different timescales. The synthetic data thus provide a rich test bed of potential difficulties in application. All of these systems are designed according to figure 23, with additional stochastic driver input to both subsystems. For time continuous systems, the latter is given by a Wiener process with additional volatility which we tried to set as high as possible before catastrophically altering the system manifolds in the numerical solution.

In a final step, we have applied the method to real data composed of LFPs from cat visual areas. Here we show that interpretable patterns very similar to the synthetic case are produced by the method which suggest a connectivity structure and interaction delays that are physiologically plausible.

8.4.1 *Logistic Maps*

In a first step, we consider a simple discrete system to practically demonstrate the soundness of the functional mapping established in 124. To this end, we present results where the stochastic drivers ω, ω' are unknown, and treated with Bayesian statistics according to 134, as well as results where ω and ω' were supplied for the mapping as covariates according to equation 124.

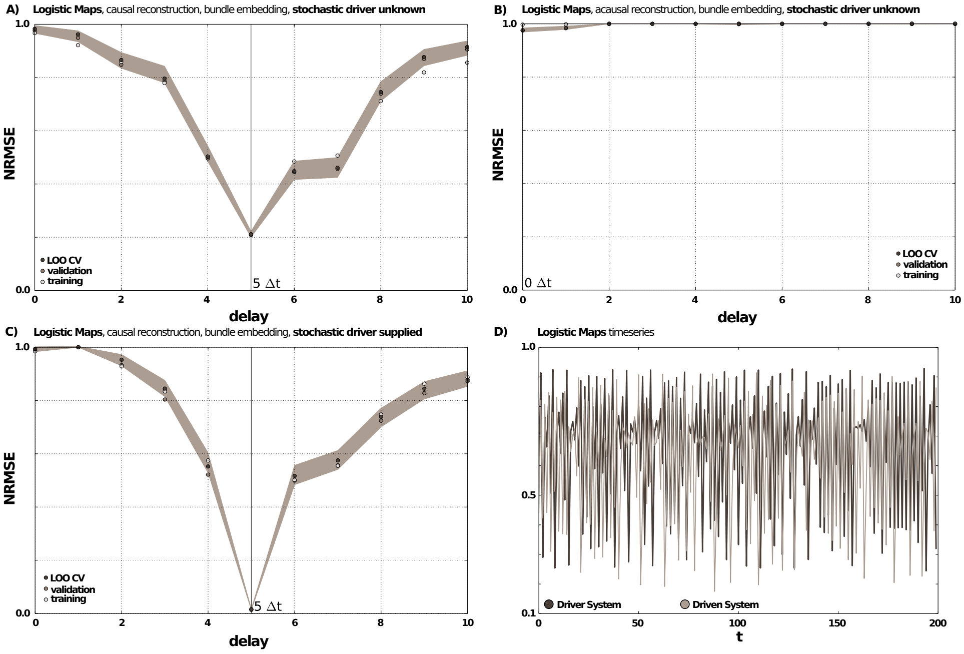

Figure 25: Delay-coupled logistic maps. Analysis of measurements from two coupled logistic maps (140). **A-C)** Interpolated reconstruction error graphs (135) plotted against candidate delays under true causal hypothesis (A,C) and acausal hypothesis (B), respectively. In subfigure A, the stochastic driver (ω in model 126) is unknown and integrated out of the statistical model, whereas in subfigure C stochastic time series ω is explicitly supplied as covariate in mapping 126. **D)** Logistic map time series example.

The system under consideration is a skew-product of logistic maps, defined by

$$\begin{aligned} y(t) &= 3.7y(t)(1-y(t)) - 0.4\omega_t \\ x(t) &= 3.7x(t)(1-x(t)) - 0.2\omega'_t - 0.2y(t-5)^2, \end{aligned} \tag{140}$$

where $\omega_t, \omega'_t \sim \mathcal{N}(0, 0.05)$. Subsystem y is driving x via a coupling term in the definition of x. Both subsystems are chaotic and show sensitive dependence on ω and ω', respectively, which are part of the intrinsic temporal evolution of the system. Although the stochastic forcing is given by a linear term, its effect is nonlinearly amplified through the forward mapping. Examples of the resulting time series are shown in figure 25D. Logistic maps are examples of endomorphisms since the forward mapping is in general not invertible. As was discussed before, reconstructibility may therefore not extend to the full reconstructible area. We fit the full nonparametric statistical model according to 131 with prior covariance matrix 133 under both hypotheses, $x \rightarrow y$ and $y \rightarrow x$, and computed REGs as described in the previous section. Note that under hypothesis $x \rightarrow y$, x is assumed to be the driver and thus the target of the reconstruction, whereas y provides covariates. In this simple example, a reconstruction dimension $d = 2$ is sufficient for the covariate vector Rec_d. The REGs are reported for the LOO predictive distribution (dark brown, LOO CV) and the predictive distribution conditional on all available data (light brown, training) to indicate overfitting. In addition, we also report the REG for a separate validation set of the same size (brown, validation), which should alleviate any concerns regarding the ability of the complex nonparametric model to generalize. Confidence intervals (shaded area) were computed for the LOO estimator only. Figure 25 shows the results for a small data set where each time series consisted of 1000 samples.

In subfigure A, the reconstructibility is shown under the correct hypothesis $y \to x$ for the case where the stochastic driver is unknown and integrated out of the model. The lowest reconstruction rate is achieved at the true delay $\Delta t = 5$. Reconstructibility is confined to this delay, which corresponds to the last coordinate of covariate vector Rec_2, in agreement with theoretical reconstructibility of endomorphisms discussed beforehand. Subfigure B shows reconstructibility of the same model under the wrong assumption $x \to y$, which hardly ever deviates from the baseline NRMSE= 1.0. The direction of interaction, as well as the delay, are thus clearly discernible on this data set.

In subfigure C, reconstruction was repeated as in A, however, this time the stochastic drivers ω, ω' were supplied explicitly as covariates for functional mapping 124. The results are highly interesting. As can be seen, reconstructibility only improved within the reconstructible area. At candidate delays where information about the corresponding driver state is not actually available through the covariate vector Rec_2, the inclusion of the noise terms into the functional model may even lead to a deterioration of performance. At the true delay $\Delta t = 5$, on the other hand, the additional driver information leads to a practically perfect reconstruction of y, which beautifully demonstrates the validity of the theoretical approach behind mapping 124 in practice.

8.4.2 *Lorenz-Rössler System*

A weakly coupled Lorenz-Rössler system is the first continuous system we consider. It is given by

$$\begin{aligned}
\dot{x}_1 &= 10(x_2 - x_1)\\
\dot{x}_2 &= 28x_1 - x_2 - x_1x_3 + \mu\omega_x\\
\dot{x}_3 &= x_2x_1 - \frac{8}{3}x_3\\
\dot{y}_1 &= y_3 - y_2\\
\dot{y}_2 &= y_1 + 0.2y_2 + \mu\left(\frac{1}{3}\sum_i x_i(t-\tau) + \omega_y - y_2\right)\\
\dot{y}_3 &= 0.2 + y_3(y_1 - 5.7),
\end{aligned} \tag{141}$$

where the coupling coefficient $\mu = 1$ and the delay $\tau = 2$. The stochastic forcing ω_x, ω_y is realized by two Wiener processes with additional volatility 10^5, whereas the system is numerically solved using a fourth-order Runge-Kutta method with fixed stepsize 0.0001 which parametrizes the actual variance of the stochastic processes. This numerical solution scheme is necessary to include the stochastic processes, otherwise Shampine's *dde23* [Shampine and Thompson, 2001] would be a much better choice.

As time series we consider the linear combinations $\sum_i x_i(t)$ and $\sum_j y_j(t)$, reminiscent of the contributions from different spatially distributed components to a LFP, with sampling rate $dt = 0.1$, as shown in figure 26D. It can be seen that

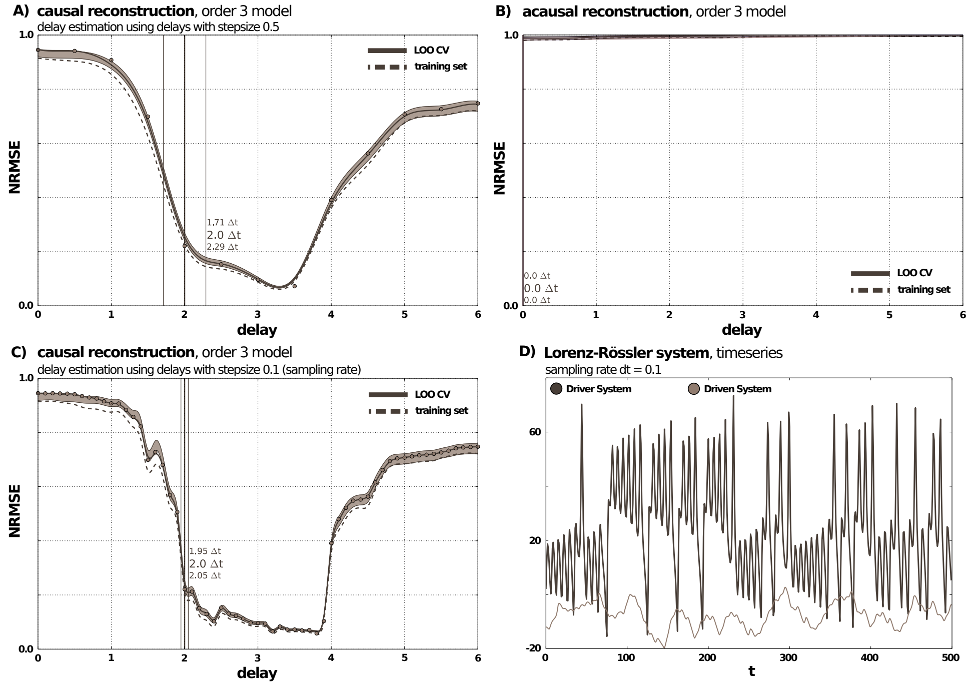

Figure 26: Delay-coupled Lorenz-Rössler System. Analysis of measurements from a Lorenz system driving a Rössler system (141). **A-C)** Interpolated reconstruction error graphs (135) plotted against candidate delays under true causal hypothesis (A,C) and acausal hypothesis (B), respectively. Subfigure A shows the interpolated REG against candidate delays sampled with a stepsize of 0.5. In contrast, subfigure C shows the same interpolation given steps of size 0.1 between candidate delays. **D)** Lorenz-Rössler time series example.

the natural oscillations of the driven Rössler system (y) are much simpler and occur at a much slower timescale than those of the Lorenz driver (x) with its two-lobed attractor geometry. Analysis of time series evolving on different timescales is challenging and we have designed this particular example to demonstrate the reconstructibility of a fast complex signal from a slower oscillatory response.

Optimal results are reported for a third-order model with reconstruction dimension $d = 20$. Although the intrinsic dimensionality of the skew-product manifold described by equation 141 is likely less than 6, a larger d is necessary to capture enough geometric information of response system y with covariate vector Rec_d. We use this data set to show how different lags between candidate delays of the REG can affect the delay estimation process. In figure 26A we show the interpolated REG under the correct hypothesis $x \to y$ with candidate delay reconstructions given at a stepsize of 0.5, marked by dots. In comparison, figure 26C shows the REG sampled with smaller stepsize 0.1. Although both point estimators for the delay yield the correct result $\hat{\tau} = 2.0$, the analytical area of confidence is much wider in A, indicating a larger uncertainty associated with the estimator as a result of a less well-pronounced sink in the interpolated REG. The finer stepsize of candidate reconstructions in subfigure C provides more accurate information for the delay estimation process and yields consequently a very narrow area of confidence. The reconstructible area corresponds to $[2, 4]$, as a result of the time series sampling rate $dt = 0.1$ and the model choice $d = 20$. For this system, the sink in the interpolated

REG fully covers the reconstructible area and has sharply defined corners. Despite the fact that there is substantial oscillatory and trend-like uninformative structure within the sink, the analytical criteria for establishing the area of confidence prove to be robust and capture the informative geometric features, as is readily verified by visual inspection. In contrast, figure 26B shows that under the acausal hypothesis $y \rightarrow x$ no significant reconstruction is possible at any of the considered candidate delays.

8.4.3 *Rössler-Lorenz System*

The scenario we discuss in this section is an adaptation of a standard Rössler-Lorenz system [GSPAPER] which we extend by a delay of $\tau = 2$ in the coupling,

$$\begin{aligned}
\dot{x}_1 &= a(x_3 - x_2) \\
\dot{x}_2 &= a(x_1 + 0.2x_2) + \mu\omega_x \\
\dot{x}_3 &= a(0.2 + x_3(x_1 - 5.7)), \\
\dot{y}_1 &= 10(y_2 - y_1) \\
\dot{y}_2 &= 28y_1 - y_2 - y_1y_3 + \mu(x_2(t-\tau) + \omega_y) \\
\dot{y}_3 &= y_2y_1 - \frac{8}{3}y_3.
\end{aligned} \tag{142}$$

This time, the Rössler system x is the driver and its intrinsic timescale is set to $a = 6$ such that it oscillates with a similar frequency as the driven Lorenz system y. The stochastic drive is once again a Wiener process with additional volatility of 10^5 at a numerical solution stepsize of 0.0001.

In [Pyragas, 1996] it was shown for the case without stochastic forcing and delay that for $\mu > 6.66$ the two systems enter a regime of *generalized synchronization* (GS) [Kocarev and Parlitz, 1996]. Adding a delay is compatible with the definition of GS, and the stochastic forcing does not catastrophically alter the system manifolds but acts merely as a perturbation. We set $\mu = 8$ such that the systems will be in a perturbed regime of GS. Time series are once more sampled as linear combinations of the coordinate functions of the subsystems, as shown in figure 27D.

This data set poses a twofold challenge. First, the two subsystems show very similar oscillatory behavior. As a result of the (perturbed) GS regime, we can assume the oscillators will exhibit more or less rigid phase-locking behavior. This means a certain functional dependency will be present across any candidate delay in the analysis. Second, by definition of GS, the driven system is fully predictable from knowledge of the driver alone, although this transformation may be very complex. The dynamics of the full skew-product system have collapsed onto a common synchronization manifold. As a result, the causal structure of the interaction may be masked since predictability in this sense cannot be distinguished from reconstructibility. If the interaction is delayed, however, one may still be able to discern a sink in the interpolated REG that is informative. Furthermore, prediction only works forward in time whereas reconstruction works backwards. As a result, among the set of candidate delays in the REG, an optimally predictable state (that is, in acausal reconstruction direction) of the driven system can only ever be found

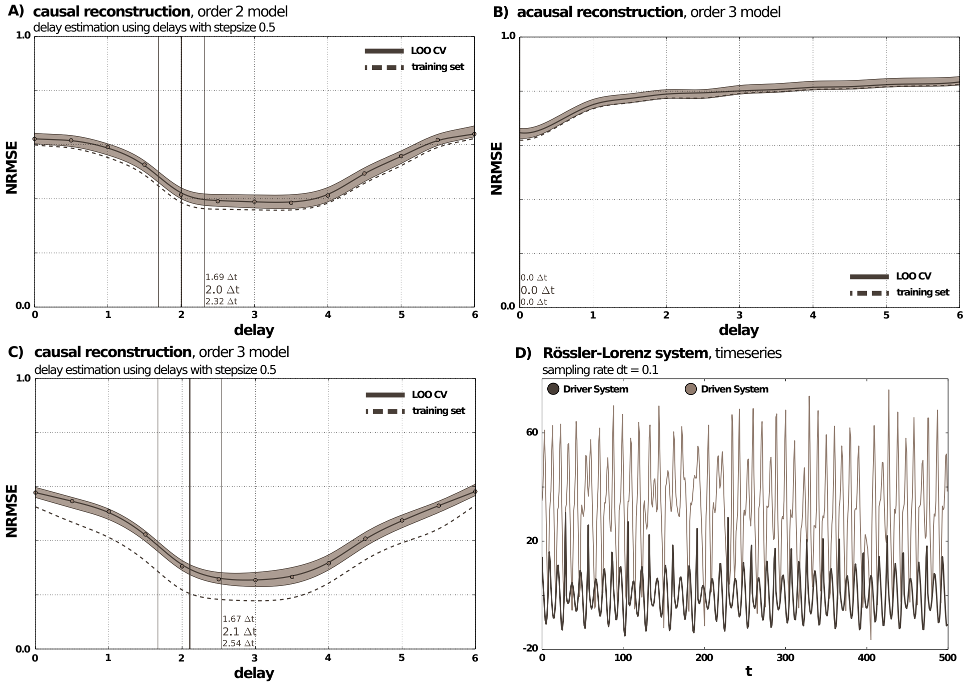

Figure 27: Delay-coupled Rössler-Lorenz System in Generalized Synchronization. Analysis of measurements from a Rössler system driving a Lorenz system (142). **A-C)** Interpolated reconstruction error graphs (135) plotted against candidate delays under true causal hypothesis (A,C) and acausal hypothesis (B), respectively. Subfigure A shows the interpolated REG for a second order Volterra series model ($n = 2$ in equation 132). In contrast, subfigure C shows the same interpolation given a third order model. **D)** Rössler-Lorenz time series example.

at 0, and only if the true interaction delay $\tau = 0$. The larger τ, the better the interaction and its delay will be distinguishable. It is also our impression that a statistical model provides in general better functional reconstructions than predictions, since in reconstructions driver information is actually present in the temporal evolution of the driven time series.

In figure 27 we present results for models with reconstruction dimension $d = 25$. We consider the true hypothesis $x \to y$ first, shown in subfigures A and C. Here, overfitting already becomes apparent at model order 3, as can be seen in subfigure C. The shape of the resulting sink is not very regular, the corners are not well-pronounced and curvature hardly declines inside the sink. As a result, the delay estimation is slightly off. Subfigure A shows, in contrast, a second order model, where the sink is very regular with less curvature and the model does not suffer from overfitting. Consequently, the delay estimation of the second order model is more accurate and yields, in fact, the correct value of $\hat{\tau} = 2.0$. Note that in this case the addition of further candidate delays to the REG at a finer resolution will not improve the shape of the sink. Its shallowness is rather owed to the more simple oscillatory behavior of the two systems and their phase-locking which improve predictability of the systems. This also leads to a much lower *baseline reconstructibility* which is now apparent even under the acausal hypothesis in subfigure B. Here, the baseline predictability at candidate delay $\Delta t = 0$ is comparable to the baseline reconstructibility outside of the sink in subfigure A. Note that the mere presence of a sink in the REG with significant temporal offset under one coupling hypothesis is highly informative about the direction of interaction.

We conclude that if delays are involved in the interaction, our method may still be able to reliably discern coupling scenarios shrouded in weaker forms of synchrony and phase-locking.

8.4.4 *Mackey-Glass System*

The final synthetic data set we present is generated by two coupled chaotic Mackey-Glass oscillators, given by the system

$$\begin{aligned} \dot{x} &= -x + 2\frac{x(t-\tau)+\omega_x}{1+(x(t-\tau)+\omega_x)^{9.65}} \\ \dot{y} &= -0.95y + 2\frac{0.9y(t-\tau)+0.3x(t-\tau)+\omega_y}{1+(0.9y(t-\tau)+0.3x(t-\tau)+\omega_y)^{10}}, \end{aligned} \tag{143}$$

where once more $\tau = 2$. Each subsystem is in itself delay-coupled and therefore belongs to a class of *functional differential equations*. Although the defining equation is one-dimensional, the underlying semi-flow that governs the temporal evolution of the system operates on a state space of real-valued functions with domain $[-\tau, 0]$ and is therefore infinite-dimensional [Guo and Wu, 2013a]. The system's chaotic attractor manifold, however, is intrinsically low-dimensional and can be reconstructed from the resulting time series. However, retarded systems are not time-invertible since the temporal dependencies of the delay-coupling would render the inverted system acausal. The volatility of the Wiener processes ω_x, ω_y was set to 100, whereas the stepsize of the numerical solver was set to 0.0001. To manage mutual information between neighboring samples, the time series are provided at a sampling rate $dt = 0.5$ only and shown in figure 28D.

The two systems oscillate on very similar timescales and may have entered a weak form of synchronization, albeit perturbed by the stochastic forcing, as a result of the coupling. We therefore expect again a certain baseline reconstructibility across candidate delays. Figure 28 shows the results for an optimal reconstruction dimension $d = 10$. At the sampling rate of $dt = 0.5$, the hypothetical reconstructible area would have size 5. Under the correct hypothesis $x \rightarrow y$ reconstructibility of a third-order model in subfigure A is compared against a model employing the full nonparametric series expansion in subfigure C. Both agree in their delay estimate. The REG sink resulting from the nonparametric model, however, has less curvature and more pronounced corners, leading to a smaller area of confidence and a correct point estimator $\hat{\tau} = 2.0$. As expected, a certain baseline reconstructibility is apparent under the acausal hypothesis $y \rightarrow x$ in subfigure B but does not impede inference regarding the direction of the interaction. It is noteworthy that the sinks in the REGs of subfigures A and C barely cover half of the reconstructible area, which would be in this case $[2, 7]$, as a result of the temporal dependencies in equation 143. The onset of reconstructibility nonetheless affords reliable delay estimation, which shows that functional mapping 124 extends to retarded systems if the intrinsic dimensionality of the attractor manifolds allows reconstruction.

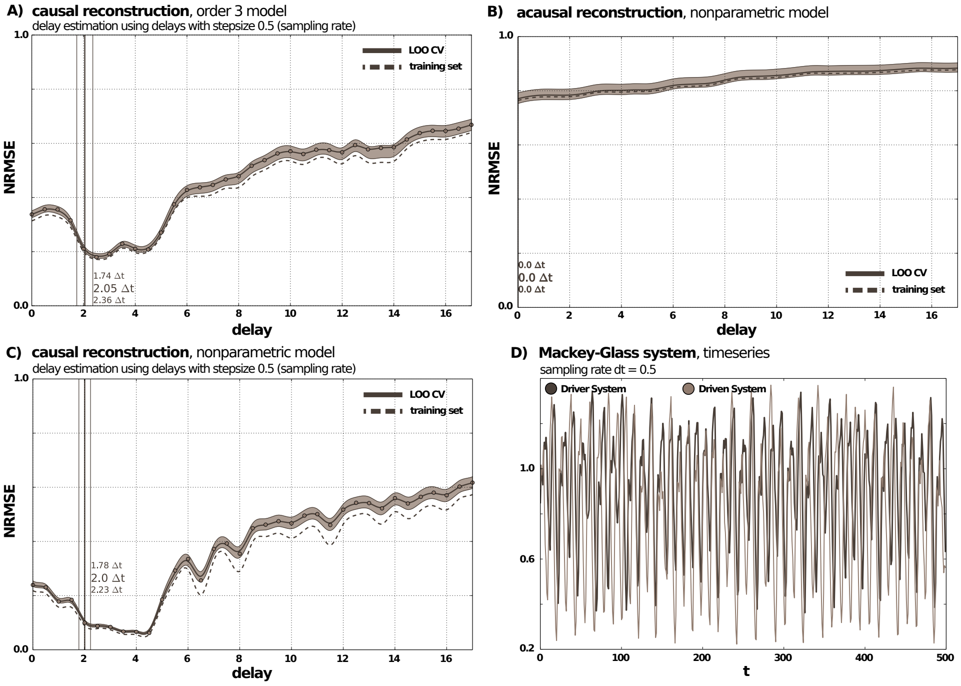

Figure 28: Delay-coupled Mackey-Glass Oscillators. Analysis of measurements from two coupled Mackey-Glass systems (143). **A-C)** Interpolated reconstruction error graphs (135) plotted against candidate delays under true causal hypothesis (A,C) and acausal hypothesis (B), respectively. Subfigure A shows the interpolated REG for a third order Volterra series model ($n = 3$ in equation 132). In contrast, subfigure C shows the same interpolation given a nonparametric model with covariance matrix specified by formula 133. **D)** Mackey-Glass time series example.

8.4.5 *Local Field Potentials of Cat Visual Areas*

In this section we report results our method yielded on an actual neuroscientific data set. The data consist of local field potential recordings measured at eight different electrodes in cat visual areas. Four electrodes were placed in in area 17, homologue to V1 in primates, another four were placed in area 21a, a homologue to V4 [Peters and Payne, 1993]. In each area, the electrodes were chosen to correspond to a particular layer of the cortex, as indicated in figure 30. The laminar position of the recording electrode is an important aspect, because information flow within and between areas follows a specific pattern along the layers of the cortex [Douglas and Martin, 2004]. At the time of the analysis, the cortical layers from which individual electrodes had been recording were unknown to the theoreticians. The cats were under general anesthesia during the recordings and visual stimulation was performed by drifting gratings of different orientations and directions. Our goal was to test the method on an actual data set, compare results with the patterns yielded on the synthetic data, and, finally, to create a connectivity graph of the eight electrodes. We were not interested in answering a particular neuroscientific question at this point.

The data was recorded at a sampling rate of 1024 Hz. Time series appeared not to show substantial differences in response to different stimulus orientations,

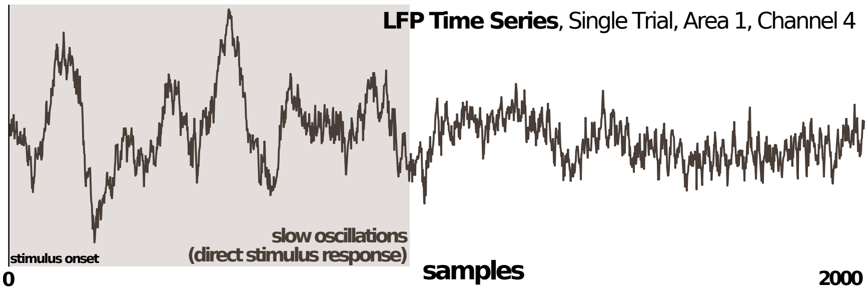

Figure 29: Exemplary LFP Time Series from Cat Visual Cortex Area 17. The time series corresponds to a single trial and is exemplary for other trials. In particular, the trials show a direct stimulus response in the first half of the samples that is characterized by slow oscillations with large amplitudes.

so we picked one orientation at random for analysis. This resulted in 50 trials, during each of which the stimulus presentation lasted for 2035 samples. In terms of preprocessing, filtering the time series with a Parzen window, the size of which was subject to model selection, improved reconstructibility slightly. Furthermore, the data had to be sampled down substantially, by a stepsize of 15, for mutual information between neighboring samples in the time series to reduce to a level adequate for statistical modeling. The statistical model was always conditioned on all 50 trials, the covariate vectors Rec_d, however, were constructed respecting trial boundaries for each trial individually. As a result, if the reconstruction dimension is set to $d = 20$, for example, Rec_d spans an area of more than 300ms. Since the covariate vectors are generated for each trial individually, this means the first 300ms of each trial directly after stimulus onset are lost for analysis, since the regression target is always associated with the end of the covariate vector for sure reconstructibility, as discussed earlier.

This turned out to be a problem. Figure 29 shows the first 2000 samples after stimulus onset of an exemplary LFP time series of a single trial from the data set. The time series is obviously nonstationary. In particular, in the first half the direct stimulus response is characterized by slow oscillations with large amplitudes which are markedly different from the statistics in the second half of the series. They also grossly distort the NRMSE statistic which is normalized by the time series standard deviation and assumes stationarity. The direct stimulus response is therefore often routinely discarded in analyses and deemed to be the result of uninformative transient dynamics.

We checked this assumption by performing pairwise reconstructions for all channels, in total 56 REGs, based upon different temporal windows within the individual trials. The analysis revealed that the samples directly after stimulus onset are by far the most informative, whereas REGs conditional upon samples from the second half of the stimulus presentation revealed no information about interaction delays at all. As a result, we based all subsequent analyses on the first 1000 samples after stimulus onset only, such that the statistics were approximately stationary. Furthermore, in light of this discovery, the loss of several hundred milliseconds of information right after stimulus onset due to the regression setup became unacceptable. To improve this situation, given a model with reconstruction dimension d, we added a zero-padding of length $d - 1$ to the beginning of each trial time series.

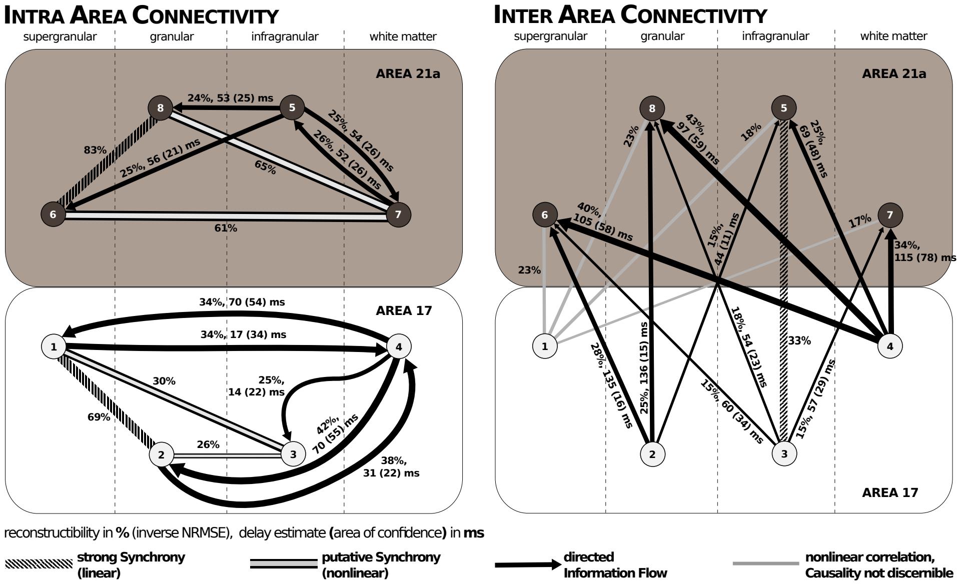

Figure 30: Connectivity Diagram of the LFP Recording Sites. *Intra* and *inter* area couplings have been separated for better visualization. On the left, intra area connectivity is shown. The darker shaded box on top corresponds to recordings in area 21a, while the white box contains recordings from area 17. Four probes were placed in each area, respectively, and are represented by the nodes in the graph. The probes were placed into cortical layers with varying depth, as indicated by the labels above the area-boxes and the corresponding dotted vertical bars. Depth increases from left to right with *supergranular* being the most superficial, corresponding to cortical layers 2 and 3. *Granular* corresponds to cortical layer 4, *infragranular* to layers 5 and 6. *White matter* denotes the deepest measurements corresponding to afferent axons. Connectivity is visualized by the graph's edges. According to the legend above, the edges distinguish between four different types of interaction. Directed edges indicate a clearly discernible direction of information flow. Edge labels inform, first, about the degree of reconstructibility, as measured in percent of the inverse NRMSE $\in [0, 1]$ (135). Second, the delay estimate $\hat{\tau}$ (139) is provided in milliseconds, together with its symmetric area of confidence (in brackets). A larger area of confidence indicates shallower reconstruction error graphs and thus higher uncertainty associated with the delay estimator (see text accompanying equation 139). A corresponding diagram is shown on the right for inter area connectivity.

Consequently, the first i samples in each trial where $i < d$ are now included in the model, albeit with suboptimal reconstruction dimension $d_i = i < d$. Nonetheless, this yielded more pronounced sinks in the REGs and improved the delay estimation process substantially.

We report our final results for a model of order 2 with $d = 20$, on time series sampled down to steps of 15.36ms. Higher order models were too susceptible to overfitting. Figure 30 summarizes the results in a connectivity diagram. The diagram looks convincing and the estimated delays are physiologically plausible. Intra-area one is likely to find strongly synchronized populations, putative mutual coupling motifs, as well as short interaction delays. In contrast, the inter-area connectivity is characterized by long interaction delays, up to an estimate of 136ms, and the network motif appears to correspond to a feed-forward structure from lower to higher visual areas.

The statistical model is highly informative about the underlying interaction. In figure 30, we distinguish between 4 different types. The first one we labeled *strong synchrony* with linear dependency. This interaction was found between channels 1 and 2 in area 17 and between channels 6 and 8 in area 21a. Interestingly, both pairs consist of one channel located in the superficial and the other in the granular layer. Corresponding REGs are shown in figure 31. The left plot shows the reconstruction under the hypothesis $1 \rightarrow 2$, that is, channel 1 is the target of the reconstruction, channel 2 provides covariates. Reconstructibility under both hypotheses does not differ significantly. In both directions there is no temporal offset for the sink in the REG which fully extends to the maximally reconstructible area and has sharply pronounced corners. Moreover, reconstructibility with a second order model is not significantly better than a first order model. The first order model is basically only a linear time-invariant filter. In addition, only for $d < 3$ does the NRMSE increase significantly. This suggests that the underlying systems are strongly synchronized and most likely merely locked to oscillations induced by the visual gratings. The latter hypothesis is further supported by the abrupt increase of the NRMSE beyond the reconstructible area, indicating a lack of predictability from covariate information. The measurements therefore provide no information about interactions that could be exploited for further inference.

The second type of interaction we discovered appears to be a slightly weaker kind of synchrony characterized by nonlinear dependencies. Channels 7 and 8 provide an example, their REGs are shown in figure 31. Although similar at first glance to the situation of channels 1 and 2, reconstructibility declines significantly upon reducing both order and reconstruction dimension d of the model, and the sink is slightly shallower. This indicates a nontrivial kind of information exchange between the underlying neuronal populations, albeit without a discernible delay.

The third type of interaction that can be found in the data actually admits a discernible direction of information flow. We show exemplary REGs of channels 4 and 8 in figure 31 which are the first to exhibit an asymmetry in reconstructibility. On the left hand side, under the hypothesis that information flows from lower to higher visual area, the full extent of a sink is apparent with a large offset. The estimated delay is about 100ms. Moreover, reconstructibility is significantly better than under the counter hypothesis on the right hand side. Both, shape and asymmetry of the REGs are strongly reminiscent of the situation that arises in a regime of generalized synchronization resulting from uni-directional coupling, which we discussed before in the context of the Rössler-Lorenz system (compare with figure 27). The example channel 4 was located in the white matter below area 17, whereas channel 8 was located in the granular layer of the downstream area 21a. The white matter consists of fiber tracts projecting to and from other areas and evoked potentials can be easily recorded. Accordingly, our method reveals a strong feed-foreward flow of information from the lower area 17 to the major recipient layer (granular layer 4) in the downstream area 21a.

The fourth type of interaction depicted in figure 30 is labeled *nonlinear correlation*. These are cases where ghosts of sinks appear in the corresponding REGs which are, however, not statistically significant. It may be indicative for a situation where the statistical model cannot capture enough information from the data to yield a more pronounced sink. Furthermore, it is worth noting that this kind of interaction was found for the channel located in the superficial layers of area 17 with all the other channels in the downstream area 21a. The superficial layers exhibit a high degree of recurrent connections. It may be that the computations done there are more complex and are not captured by the initial evoked response used here for the reconstruction. We distinguish this from scenarios where the REGs show no patterns at all, such that their deviation from 1.0 may rather be explained as a baseline due to oscillatory activity. These cases are left unmarked in the connectivity diagram.

A last finding we report here pertains to putative mutual coupling scenarios and occurred only intra-area. An example is given by channels 2 and 4, with REGs shown in the last row of figure 31. Reconstructibility is not actually significantly better in one direction over the other, although this is a close call. Both REGs exhibit, however, significantly pronounced sinks, clearly indicating delayed information flow in both directions. We hypothesize that this pattern might correspond to a situation of weak mutual coupling. This is a scenario we have not discussed so far for reasons that will be explained at a later point in the discussion.

8.5 DISCUSSION

In summary, we have presented a statistical method that estimates the direction of information flow and its delay in situations motivated by neuroscientific application where data consists of local measurement from spatially distributed subsystems which are part of a larger global system. The situation is formalized by assuming the local subsystems are low-dimensional and deterministic, delay-coupled to each other, and receive additional stochastic forcing from the surrounding global system. The latter is by assumption rendered practically infinite-dimensional. This corresponds for example to the situation encountered with local field potentials, which provide recordings of local neuronal subpopulations embedded in an extremely high-dimensional global brain-network.

The method formalizes directed information flow in terms of reconstructibility, where the latter is based on theory from differential topology that is concerned with embeddings of manifolds, similar to the approach by Sugihara et al. [2012] in the context of population dynamics. Given the particular domain of application in neuroscience, a statistical model formalizes the different sources of uncertainty

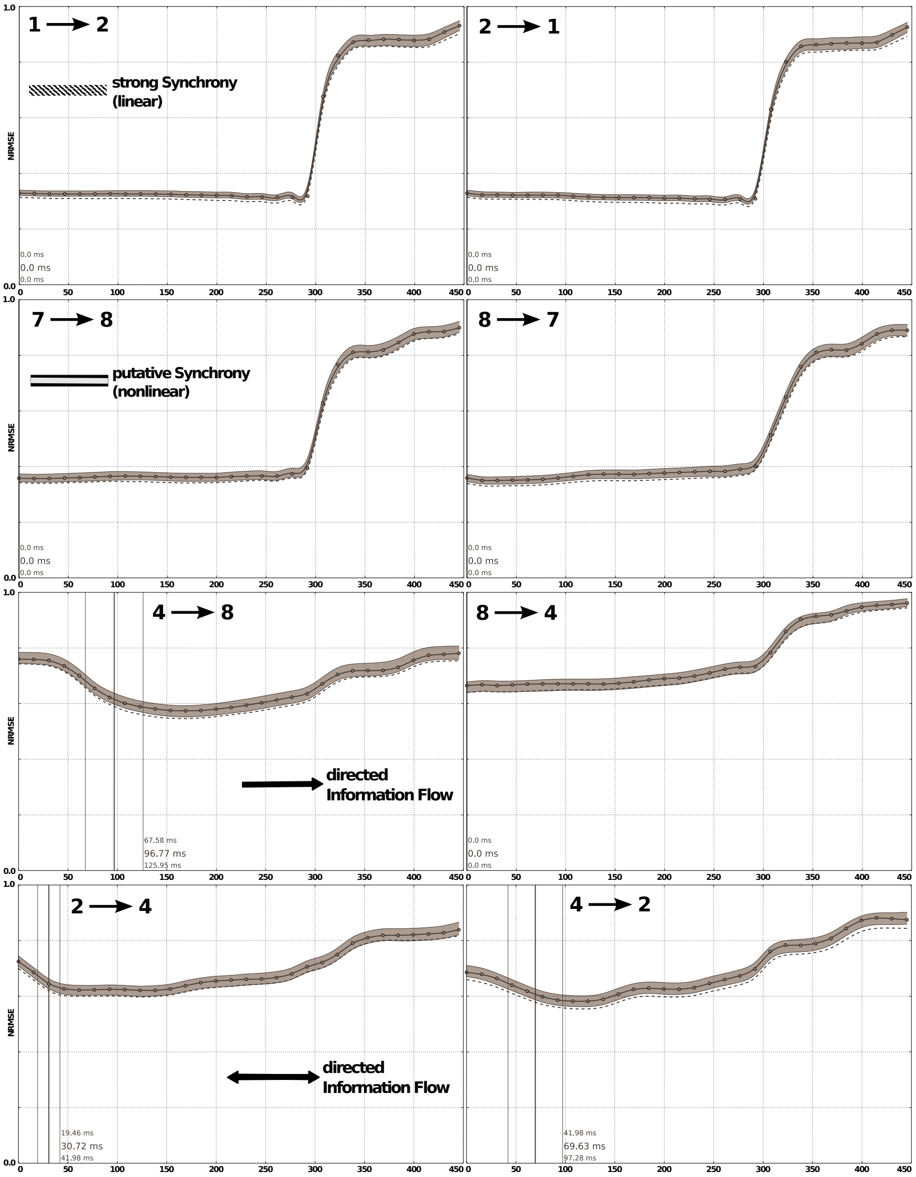

Figure 31: Reconstruction Error Graphs for Cat LFPs. REGs corresponding to selected edges of the connectivity graph in figure 30. Causal hypotheses are indicated in terms of the respective node labels.

pertaining to the reconstruction. While the functional dependencies, resulting from embedding theoretic considerations, are realized via discrete Volterra series operators, the full statistical model is given by an approximate Gaussian process. Under different hypotheses regarding the coupling direction, the statistical model yields interpolated reconstruction-error graphs across candidate delays which are informative about the true delay of the interaction. We propose an analytical criterion, informed by differential topology, which provides a point estimator for the delay, together with an area of confidence.

The validity of the method was established on a variety of delay-coupled chaotic systems which exhibit both nonlinear oscillatory behavior, as well as synchronization and phase-locking phenomena that relate well to the character of neuroscientific data. While complexity in actual neural networks likely arises in the first place from the sheer number of interacting subsystems, chaotic systems achieve a similar level of complexity due to their intrinsic chaoticity. The latter admits a low-dimensional deterministic description that retains analytical tractability of the systems, making them ideal candidates for testing purposes. Our method proved to be versatile in the analysis of systems operating on very different timescales, was able to carry out inference in the presence of phase-locking and generalized synchronization phenomena, and allowed inference on data from retarded functional systems that are likely to be encountered in the neuroscientific context.

Furthermore, we demonstrated that the applicability and interpretability of the method seamlessly transfered to a real data set of local field potentials from cat visual areas. As was shown, the method yielded both, delay estimates and network coupling motifs that are physiologically highly plausible. In addition, the statistical model allowed us to differentiate between different types of interaction, accounting for putative synchronization phenomena as well as scenarios of delayed and asymmetric information flow. Our results were reported conditional on 50 trials. The method as such, however, supports in principle also single trial analysis. In addition, it can be limited to particular temporal windows in a time series to deal with nonstationarities and to achieve a temporal resolution of the analysis. In this context, our results revealed that the slow oscillatory activity directly after stimulus onset, which is often discarded as uninformative transient response, is in fact highly informative about the interaction as well as its delay.

The final results on the LFP data set were reported for a model with reconstruction dimension $d = 20$. Given that LFPs may record activity from more than 10000 neurons, this number seems to be very low. Although higher d did not substantially improve reconstructibility in this case, the maximal dimension that can be investigated is severely limited by the intricate temporal dependencies on the stimulus within individual trials. It is worth pointing out here that one definite advantage of the statistical modeling approach over other methodologies is that it could in principle afford a dimensionality of several 1000. The strongest restriction in this regard actually comes from the data itself. Future interdisciplinary research could address this issue by fostering a stronger collaboration between theoreticians and experimentalists to develop experimental designs that could yield data sets that afford investigating higher-dimensional phenomena in appropriate application domains.

On the other hand, we have shown that substantial information is already reconstructible at $d = 20$ and comparable to the results on low-dimensional synthetic data sets. This is most likely owed to the fact that the dynamics of neuronal subpopulations are largely characterized by concerted oscillatory activity and an abundance of reported synchronization and phase-locking phenomena [Harris, 2005; Fries, 2009]. The latter arises as a result of the high interconnectedness within populations. The more concerted the joint activity, the more collapses the intrinsic dimensionality of the joint state space manifold on which the neuronal population evolves temporally. This, in turn, renders the dynamics more susceptible to reconstruction. At the same time, however, as was discussed earlier, synchronization obstructs inference based on asymmetry in reconstructibility. It is therefore important that the methodology can account for this to a certain degree and that a trade-off at a spatial scale is found.

We would like to address an important issue concerning the formalization of causality in the proposed method. The informative criterion regarding a directed causal interaction is the (inverse) reconstructibility of the putative driver. The theory asserts the existence of functional mapping 124 in this case which we try to estimate from the data by means of a statistical model. However, the existence of a functional dependency between two time series is in itself not necessarily indicative for an underlying dynamic interaction. A simple example that comes to mind are a sine and a cosine time series which are related by a trivial phase-offset. This may give rise to concerns regarding *false positives*. The complete lack of reconstructibility under the wrong hypothesis in the synthetic examples provided earlier, in particular for the simple discrete system of logistic maps, should alleviate these concerns. The biggest challenge in this regard are indeed phase-locked or synchronized oscillatory signals, as discussed by example of the Rössler-Lorenz and Mackey-Glass systems. We have shown that the immediate consequence of phase-locking is a baseline reconstructibility across delays, which is easily established and does not derogate the existence of informative sink-structures in the REGs. Since we have defined a concise domain of application where measurements are taken from subsystems embedded in high-dimensional global systems, we can always expect a certain complexity in the measurement time series. This could be, for example, a variation in the amplitudes of oscillatory signals which we also encountered in a preliminary analysis of epilepsy data where the LFPs supposedly document extremely synchronized cortical activity. We therefore have no qualms of false positives on neuroscientific data sets.

The reconstruction approach does, however, require a degree of autonomous intrinsic activity of the driven system to establish directedness. In strongly synchronized states this is no longer possible. As was discussed by example of channels 1 and 2 (see figure 31) of the cat LFP data, such states are easily identified via properties of the functional dependency supplied by the statistical model. We assume here that intrinsic dynamics are absent and the two signals are strongly locked to the visual stimulus. In general, if the coupling is weaker such that autonomous intrinsic activity is still present in the driven subsystem, the method will be robust against common drive and yield symmetric REGs with high NRMSE values.

In application to neuroscientific data such as EEG, the biggest concern is with regard to overlapping measurements. If the domains of different measurement functions overlap, the same underlying system will be reconstructible. As a trivial result, both measurements are mutually reconstructible. This is a problem that has to be addressed at the level of experimental design and great care has to be taken to exclude this possibility. Countering intuition, an application of preprocessing techniques such as *independent component analysis* (ICA), which is popular in neuroscience, would lead to a deterioration of the situation beyond repair. Since each IC is a linear combination of the original measurement signals, it can be understood as a *virtual measurement* of a dynamical system constituted by the original signals. Consequently, each IC has perfect knowledge of the full underlying measurement system such that it should be possible to mutually reconstruct any pair of ICs. Preliminary results on ICs from EEG data support this hypothesis and yielded pairwise reconstructibility that can only be described as suspicious.

The final point we would like to address is the evidence of mutually delayed coupling scenarios in the cat LFP data, as reported with respect to channels 2 and 4 in figure 31. For several reasons, we have not discussed mutual coupling scenarios beforehand. First, embedding theory does not provide at this point a rigorous examination of the skew-product scenario in terms of mutually coupled systems, as opposed to the uni-directional coupling case. Stark conjectured, however, that the theory could in principle be extended to account for mutual coupling [Stark, 1999]. Second, our own preliminary results on mutually delay-coupled systems indicate that the delay estimation approach works in principle if the coupling is weak enough. Given the shape of the sinks involved in the reconstructions of channels 2 and 4, we feel confident that the delay estimate is correct. In stronger coupling scenarios that are most likely characterized by generalized synchronization, however, we came across less expected results where delays vanished in one direction. We therefore decided that this scenario warrants further studying and should be accompanied by proper theoretical research in terms of differential topology.

8.6 SUPPORTING INFORMATION

8.6.1 *Embedding Theory*

The method presented here is built on a theoretical framework given by concepts from differential topology. As is also explicitly stated by practitioners in this field (see [Huke, 2006]), these concepts are complicated and usually accompanied by a long tail of other theoretical dependencies (an introduction can be found in [Hirsch, 2012] and the appendix of [Broer and Takens, 2010]), which makes them a challenge even for seasoned theoreticians. Our goal in this section is, nonetheless, to give an introduction to some key concepts which are necessary to understand later parts of the proposed method and to explain the red line by which they are connected.

The problem differential topology solves for the practitioner is that of reconstructing a system that is observed only indirectly via real-valued measurements. Consider, for example, local field potentials (LFPs) from electrode recordings in cortex. These yield a time series measurement of the unobserved neuronal network activity contributing to the LFPs. Aeyels [1981] was one of the first to work on this topic and provides the most intuitive access. He considered time-continuous dynamical systems given by vector fields defined on a differentiable manifold M with m dimensions. Each vector field admits a diffeomorphism $\phi_t : M \to M$

which takes an initial condition $x \in M$ and maps it forward in time to $\phi_t(x) \in M$. Thus, ϕ defines a temporal evolution of the dynamical system and corresponding trajectories on M. *Measurements*, such as LFPs, are defined as continuous functions $f : M \to \mathbb{R}$. As a function of time, the system ϕ_t is observed only indirectly via the measurements $f(\phi_t(x))$ which constitute the observed time series. Suppose the measurements were sampled at a set of d points $t_i \in [0, T]$ along an interval of length T. This set is called a *sample program* $\mathcal{P}$.

Definition 3 *A system (ϕ, f) is called $\mathcal{P}$-observable if for each pair $x, y \in M$ with $x \neq y$ there is a $t_i \in \mathcal{P}$, such that $f(\phi_{t_i}(x)) \neq f(\phi_{t_i}(y))$.*

In other words, if a system is observable, the mapping of an initial condition x into the set of measurements defined by $\mathcal{P}$,

$$\mathrm{Rec}_d(x) = (f(x), f(\phi_{t_1}(x)), ..., f(\phi_{t_{d-1}}(x))),$$

is bijective. $\mathrm{Rec}_d : M \to \mathbb{R}^d$ is called a *reconstruction map*. If $x \neq y$, $\mathrm{Rec}_d(x)$ and $\mathrm{Rec}_d(y)$ differ in at least one coordinate, hereby allowing one to distinguish between x and y in measurement. Aeyels showed that, given an almost arbitrary vector field, it is a *generic* property of measurement functions f that the associated reconstruction map Rec_d is bijective if $d > 2m$. Genericness is defined here in terms of topological concepts that lie outside of the scope of this introduction. As a result, the temporal evolution of ϕ on M becomes accessible via the reconstruction vectors corresponding in time.

For purposes of statistical modeling, this level of description is quite sufficient. In general, however, it is natural to also demand differentiability of Rec_d such that its image is a submanifold in $\mathbb{R}^d$. In this case, the reconstruction map is called an *embedding* and also preserves the smoothness properties of M. In turn, an embedding affords the investigation of topological invariants and further properties of the dynamical system in measurement. Takens [1981] showed in a contemporaneous piece of work that Rec_d is generically an embedding if $d > 2m$, together with stronger statements of genericness.

With respect to our final goal of estimating interaction delays, the following is important to note. A diffeomorphism is invertible and defines the temporal evolution of a dynamical system both, forward and backward in time. As a result, given a reconstruction vector $\mathrm{Rec}_d(x)$, all system states $x_{\mathcal{P}}$ corresponding to the sample program $\mathcal{P}$ are in principle reconstructible (the relative position of the $t_i \in \mathcal{P}$ to $x_{\mathcal{P}}$ are not material for Aeyels' proof as long as they can be uniquely associated with x via ϕ). In case of endomorphisms, that is, systems not invertible in time, this is no longer true. Endomorphisms are important to consider for us since delay-coupled systems, such as spatially distributed neural networks, are in general not time-invertible (the time-reversed system would be acausal). In later work, Takens [2002] showed that endomorphisms may not even admit embeddings. However, in terms of the language introduced here beforehand, he proved the genericness of the $\mathcal{P}$-observability of system state $\phi_{t_{d-1}}(x)$ given the reconstruction vector $\mathrm{Rec}_d(x) = (f(x), f(\phi_{t_1}(x)), ..., f(\phi_{t_{d-1}}(x))))$ in case $d > 2m$. Hence, the system state corresponding to the end of the sample program is always reconstructible and we will exploit this fact for later stages of the method proposed here.

At the same time, for the purpose of statistical modeling, the aforementioned restrictions may be less severe in practice: A loss of differentiability in some points or even a finite number of intersections in $\mathrm{Rec}_d(M) \subset \mathbb{R}^d$ may still admit substantial reconstructibility with regard to certain functional mappings that will be introduced later. Also, Sauer et al. [1991] prove that application of linear time-invariant filters to the measurements does not destroy embeddings. This result is highly important for neuroscientific data which is often already filtered within the recording device. Sauer et al. also conjectured that different measurement functions may be combined in a reconstruction map, which was later proven by Deyle and Sugihara [2011].

While these results provide a very rich apparatus for autonomous systems, a common problem in practice is that only parts of a larger system are observed. In particular with respect to time series prediction, the objective seems hopeless if one has to assume the existence of hidden drivers that are unobserved and therefore cannot be explicitly accounted for in a statistical model. In this regard, the most important result for time series analysis was provided by Stark [1999] who showed that under certain conditions, a hidden driver may be reconstructed from measurements of the non-autonomous driven system alone. He considered skew-product systems as a formal framework, consisting of a driver system given by a diffeomorphism ψ operating on a differentiable manifold N with dimension n, and a non-autonomous driven system ϕ operating on M, where

$$\begin{aligned} \psi_t &: N \to N, \\ \phi_t &: N \times M \to M. \end{aligned} \tag{144}$$

Stark showed that, given real-valued measurements $f : M \to \mathbb{R}$ of states $x \in M$ of the driven system alone, the corresponding reconstruction map $\mathrm{Rec}_d(x)$ generically embeds the full product manifold $N \times M$ in case $d > 2(m+n)$. This means N is reconstructible via measurements of M alone. While this result dramatically improves chances at predicting a single time series, causal analyses in the Granger framework may yield distorted outcomes (see [Sugihara et al., 2012] for a concrete example). However, it also gives rise to an asymmetry in reconstructibility which can be exploited to formalize causal interactions. Unless the subsystems are generalized synchronized, in which case the dynamics of the product system collapse onto a synchronization manifold, measurements of the driver manifold N will not allow a reconstruction of M since there is no information about M present in the temporal evolution of ψ. In case of synchronization this approach becomes problematic, since knowledge of N would allow by definition a perfect prediction of ϕ [Kocarev and Parlitz, 1996].

With respect to our purpose of delay estimation, we note that the individual time scales of the two systems are not material for the skew-product embedding, which is formulated in terms of indexed sequences of points on the manifolds. Also, it is not required that the temporal evolution of the full product system 144 is a diffeomorphism, this restriction only pertains to ϕ and ψ individually. As a result, the theory can in principle account for delay-coupling scenarios. Denote by $x_t := \psi_t(x_0) \in N$ the driver system state at time t, given an initial condition $x_0 \in N$. A coupling with delay τ would induce forward dynamics on M with M-intrinsic time index t' as $\phi_{t'}(x_{t-\tau}, y_t) = y_{t'} \in M$. Conversely, for a given measurement function $f : M \to \mathbb{R}$ and a corresponding reconstruction vector $\text{Rec}_d(y_{t'})$, we would expect to be able to reconstruct the mixed temporal state $(x_{t-\tau}, y_t)$ of the full product system on $N \times M$ if $d > 2(m+n)$.

In particular, if $g : N \to \mathbb{R}$ is a continuous, real-valued measurement function of the driver system, a functional mapping $F : \mathbb{R}^d \to \mathbb{R}$ exists, defined by $F := g \circ \text{pr}_N \circ \text{Rec}_d^{-1}$. Here, pr_N denotes the natural projection from $N \times M$ into N. This mapping could yield for example

$$F\left(f(y_t), f(\phi_{t'_1}(y_t)), ..., f(\phi_{t'_{d-1}}(y_t))\right) = g(x_{t-\tau}). \tag{145}$$

Note that F is a mapping between the explicitly observed measurement time series, allowing one to reconstruct the indexed measurements g_i using reconstruction vectors composed of f_i. Furthermore, the reconstruction is backwards in time if delays are involved, which employs the strongest possible constraint causal interactions can exhibit. That is, a cause-effect relationship can only extend forward in time, and, consequently, reconstruction is only possible backwards in time.

F is a continuous functional mapping if it exists and our goal will be to estimate it with a statistical model directly from any given measurement time series. It is not sufficient, however, to assume the data stems from a single autonomous skew-product system. When considering e.g. LFP recordings, measurement time series corresponding to different recording sites will contain information from local neuronal subpopulations that may interact with each other. In any case, each subpopulation is likely to receive massive modulatory influence from other areas of the brain. Such additional drivers may be too high-dimensional for practical purposes to be reconstructed from the data. Nonetheless, they will dynamically alter the temporal evolution of any measured subpopulation. At the very least, one has to consider this as a form of *intrinsic noise* which, in contrast to simple measurement noise, is part of the system dynamics.

Stark showed that in situations where the driver is not reconstructible or already known, a reconstruction of the driven system can in principle still be defined. Simplifying the temporal indices, suppose the temporal evolution of the driven system is given by

$$\begin{aligned} \phi &: N \times M \to M, \\ &(x_i, y_i) \mapsto y_{i+1}. \end{aligned}$$

Rather than requiring the reconstruction map Rec_d to embed $N \times M$, it is possible to embed $\{x\} \times M$ for *each* $x \in N$. These are called *bundle embeddings* and denote a family of embeddings, parametrized by N. Redefining the reconstruction map $\text{Rec}_d : M \times \{x\} \to \mathbb{R}^d$ and writing $\text{Rec}_{d,x}(y) = \text{Rec}_d(y, x)$, Stark proved that $\text{Rec}_{d,x}$ is generically an embedding for *typical* x if $d > 2m$. This result also holds if the driver system is stochastic, with shift map dynamics on bi-infinite sequences [Stark et al., 2003; Huke, 2006], in which case the driver is effectively infinite-dimensional. Measurement noise can be accounted for analogously. With applications in neuroscience in mind, Ørstavik and Stark [1998] have proposed to use such a framework of *stochastic forcing* in situations where the measurements are presumably taken from a lower-dimensional local subsystem, weakly coupled to a high-dimensional global system. If the global system, such as the neural network in the brain, is high-dimensional enough, one can view it in this context as practically stochastic. In turn, any locally measured neuronal subpopulation may be regarded as a stochastically forced low-dimensional deterministic system.

We thus include additional stochastic forcing into the model and extend the functional mapping F (145) in the following way. Given system 144, assume finite measurement time series $(f_i)_{i=0}^{D}$ and $(g_i)_{i=0}^{D}$ defined as $g_i = g(\psi^{(i)}(x_0))$, with $g : N \to \mathbb{R}$ and $\psi^{(i)}(x_0) = \psi \circ \cdots \circ \psi(x_0) = x_i$. Now extend

$$\begin{aligned} \psi : N \times \mathcal{X} &\to N, \\ (x_i, \omega_i) &\mapsto x_{i+1}, \end{aligned}$$

where $\mathcal{X}$ is a topological space and $\omega_i \in \mathcal{X}$. The latter can be treated as a stochastic dynamical system by introducing shift map dynamics on the bi-infinite sequence $\omega = (..., \omega_{-1}, \omega_0, \omega_1, ...) \in \Sigma$, with $\Sigma := \mathcal{X}^{\mathbb{Z}}$. The *shift map* $\sigma : \Sigma \to \Sigma$ is now defined by $[\sigma(\omega)]_i = [\omega]_{i+1} = \omega_{i+1}$. Likewise, extend

$$\begin{aligned} \phi : N \times M \times \mathcal{X}' &\to M, \\ (x_i, y_j, \omega'_j) &\mapsto y_{j+1}. \end{aligned}$$

We have disentangled the temporal indices of the two deterministic systems here and assume $i \leq j$ to account for interaction delays. The stochastically forced systems can be viewed as parametrized forward mappings, e.g. $\psi_{\omega_i}(x_i) = \psi(x_i, \omega_i)$ and $\psi_{\omega_i \dots \omega_0} = \psi_{\omega_i} \circ \cdots \circ \psi_{\omega_0}$. In case of ϕ, we adopt the notation

$$\phi^{\omega_i}_{\omega'_{j+1}\omega'_j}(x_i, y_j) = \phi_{\omega'_{j+1}}(\psi_{\omega_i}(x_i), \phi_{\omega'_j}(x_i, y_j)).$$

The reconstruction map on M is now twice parametrized,

$$\begin{aligned} \mathrm{Rec}_d(y, \omega, \omega') &= \Big(f(y), f(\phi_{\omega'_0}(x, y)), ..., f(\phi^{\omega_{d-2} \dots \omega_1}_{\omega'_{d-2} \dots \omega'_0}(x, y)) \Big) \\ &= \mathrm{Rec}_{d,\omega,\omega'}(y). \end{aligned} \tag{146}$$

For functional mapping (145), these dependencies on the unknown drivers can be made explicit, analogous to Stark's discussion of the NARMA model in [Stark et al., 2003]. This yields the *bundle reconstruction*

$$\begin{aligned} F : \mathbb{R}^d &\to \mathbb{R}, \\ \mathrm{Rec}_{d,\omega,\omega'}(y_j) &\mapsto g(x_i), \\ F(\mathrm{Rec}_{d,\omega,\omega'}) &= F(f_j, ..., f_{j+d-1}, \omega'_j, ..., \omega'_{j+d-2}, \omega_{i+1}, ..., \omega_{i+d-2}). \end{aligned} \tag{147}$$

If ω_i and ω'_j are taken to be one-dimensional, we can immediately include them as random variables in a statistical model. A proper treatment of this source of uncertainty with Bayesian techniques will be discussed after the statistical model has been introduced. In the results section, we practically demonstrate the soundness of the mapping F using coupled chaotic logistic maps with intrinsic noise. It is shown that in case the noise sequences ω and ω' are observed and included in F, the time series reconstruction becomes exact. Figure 23 summarizes and illustrates the approach graphically.

8.6.2 *Discrete Volterra series operator*

In order to derive point estimators for the measurement time series, the function $F : \mathbb{R}^d \to \mathbb{R}$ has to be approximated by a model. F being continuous exhausts prior knowledge about its functional form. A choice that may be called canonical in this situation is a *discrete Volterra series operator*. The term Volterra series was originally coined in the context of functional analysis, due to Volterra [1915]. In general, it can loosely be thought of as Taylor series expansion of F whose polynomial terms are rewritten in a certain coordinate representation. The result is a linear form with respect to its parametrization, which is a desirable property in a statistical model because it simplifies computation. We now show a possible derivation of the discrete Volterra series operator to illustrate the statistical model and its approximation properties. Theory regarding vector spaces and algebra is covered by [Roman, 2007].

As a well-known result from *Taylor's Theorem*, a $k+1$ times differentiable function $f \in C^{k+1}(X, \mathbb{R})$, defined on $X \in \mathbb{R}^d$ open, may be expanded around a point $x_0 \in X$ by

$$f(x) = \sum_{\substack{\alpha \in \mathbb{N}_0^n \\ |\alpha| \leq k}} \frac{D^\alpha f(x_0)}{\alpha!}(x - x_0)^\alpha + \mathbf{o}(\|x - x_0\|^k), \tag{148}$$

where $\alpha = (\alpha_1, ..., \alpha_n) \in \mathbb{N}_0^n$ is a *multi-index* with $|\alpha| := \sum_i \alpha_i = k$ and $x^\alpha := x_1^{\alpha_1} \dots x_n^{\alpha_n}$. One can choose for example $x_0 = 0$ in which case the Taylor series is also called a Maclaurin series, which denotes a very basic class of polynomial series expansions. This motivates very generally the possibility to approximate a function by a particular kind of series expansion.

With respect to multi-index α, higher order differential operators are denoted by

$$D^\alpha f = \frac{\partial^k}{\partial x_1^{\alpha_1} \dots \partial x_n^{\alpha_n}} f. \tag{149}$$

We now wish to characterize this operator as multilinear map. Let V, W be vector spaces over $\mathbb{R}$ and denote by $L(V, W)$ the space of all linear maps between V and W. In addition, $X_f \subset V$ open. If $f : X_f \to W$ is differentiable, then its derivative is defined via $Df : X_f \to L(V, W)$. Consequently, at a particular point $x \in X_f$,

$$Df(x) \in L(V, W)$$

represents a linear map with domain V. Intuitively, given $v \in V$, $(Df(x))(v)$ yields the derivative in the direction of v at point x. If f is twice differentiable, it holds for the second derivative $D^2 f := D(Df) : X_f \to L(V, L(V, W))$ that

$$D^2 f(x) \in L(V, L(V, W)).$$

In general, for a k-times differentiable function at $x \in X_f$,

$$D^k f(x) \in L(V, L(V, L(..., L(V, W)...).$$

This represents the k-fold concatenation of directional derivatives.

If $V_1, ..., V_n, W$ denote vector spaces over $\mathbb{R}$, a map $g : V_1 \times \cdots \times V_n \to W$ is called *multilinear* if it is linear in each coordinate separately, such that for all $k = 1, .., n$ with $r, s \in \mathbb{R}$, $v, v' \in V_k$ and $u_i \in V_i$, $i \neq k$,

$$\begin{aligned} &g(u_1, ..., u_{k-1}, rv + sv', u_{k+1}, ..., u_n) \\ &\quad = rg(u_1, ..., u_{k-1}, v, u_{k+1}, ..., u_n) + sg(u_1, ..., u_{k-1}, v', u_{k+1}, ..., u_n). \end{aligned} \tag{150}$$

It is easy to show that the space of multilinear maps $L(V_1, ..., V_k; W)$ is topologically isomorphic to $L(V_1, L(V_2, L(..., L(V_k, W)...)$. As a result, higher order derivatives are multilinear maps. For example, the second order derivative is a bilinear map $D^2 f(x) : X_f \times X_f \to W$ by virtue of

$$D^2 f(x)(a, b) = (D^2 f(x)a)b, \quad a, b \in X_f. \tag{151}$$

This multilinear map is also symmetric in its arguments since the order of directions with respect to which the derivatives are taken does not matter. For $X_f \subset \mathbb{R}^n$, $x \in X_f$ and $a = (a_i)_{i=1}^n \in \mathbb{R}^n$, the tangent space in x is spanned by the partial derivatives. The directional derivatives correspond to vectors in this space. As a result, the derivative with respect to direction a can be expressed as

$$Df(x)a = \sum_{i=1}^{n} a_i \frac{\partial f}{\partial x_i}(x).$$

Consequently, equation 151 becomes

$$(D^2 f(x)a)b = \sum_{j=1}^{n} \sum_{i=1}^{n} a_i b_j \frac{\partial^2 f}{\partial x_i \partial x_j}(x). \tag{152}$$

This represents in principle already the type of functional form we are looking for and also serves to illustrate the differential operator 149 in the full series expansion 148. However, at a later time it will be necessary to discard the analytical interpretation. Instead, we would like to give a full characterization in terms of algebraic properties of the multilinear map alone.

To this end, consider the following. Analogous to the matrix representation of linear maps, multilinear maps can always be expressed in terms of *coordinate represenations* with respect to the bases of domain and range. We will illustrate this using the example of a bilinear map $g : V \times V \to W$, where $v = (v_1, ..., v_n), v_i \in V$ and $w = (w_1, ..., w_m), w_i \in W$ are bases that span V and W, respectively. Let $x, y \in V$ be given by

$$x = \sum_i \alpha_i v_i, \qquad y = \sum_j \beta_j v_j,$$

where β and α are the coordinate vectors of x and y with respect to v. Then, by definition of the multilinear map,

$$g(x, y) = g(\sum_i \alpha_i v_i, \sum_j \beta_j v_j) = \sum_i \sum_j \alpha_i \beta_j g(v_i, v_j). \tag{153}$$

Since $g(v_i, v_j) \in W$, there exists a coordinate representation $g(v_i, v_j) = \sum_k \gamma_{k,ij} w_k$ with respect to basis w, such that

$$g(x, y) = \sum_i \sum_j \sum_k \gamma_{k,ij} w_k \alpha_i \beta_j. \tag{154}$$

Now consider once more the special case $V = \mathbb{R}^n$ and $W = \mathbb{R}$. We assume the standard basis for V given by the identity matrix $I = (e_1, ..., e_n) \in \mathbb{R}^{n \times n}$, and the trivial basis 1 for W. In this case

$$x = \sum_i x_i e_i, \qquad y = \sum_j y_j e_j.$$

Consequently, coordinate representation 154 simplifies to

$$g(x, y) = \sum_i \sum_j x_i y_j g(e_i, e_j) = \sum_i \sum_j \gamma_{ij} x_i y_j, \tag{155}$$

and

$$g(x, x) = \sum_i \sum_j \gamma_{ij} x_i x_j. \tag{156}$$

The procedure generalizes to yield coordinate representations for multilinear maps in k arguments. Note that in the motivating Maclaurin series in 148, where $x_0 = 0$, the higher order derivatives are the result of sequential application of the differential operator, each time with respect to the same direction, which in this case is simply x. The corresponding multilinear map receives thus k times the same argument x, as illustrated in the two-dimensional example above. This insight allows one to rewrite the Maclaurin series explicitly as a sum of multilinear maps in the argument $x = (x_i)_{i=1}^d \in X \subset \mathbb{R}^d$ to yield the polynomial form

$$\begin{aligned} p(x) = \gamma_0 + &\sum_{k_1=1}^{d} \gamma_1(k_1) x_{k_1} + \\ &\sum_{k_1=1}^{d} \sum_{k_2=1}^{d} \gamma_2(k_1, k_2) x_{k_1} x_{k_2} + \dots . \end{aligned} \tag{157}$$

This is called a *discrete Volterra series operator*. The $\gamma_i : \mathbb{N}_1^d \times \cdots \times \mathbb{N}_1^d \to \mathbb{R}$ provide in this notation real-valued coefficients weighting the monomials in the x-coordinates. The n^{th} summand of this series represents the former n^{th} order derivative in the Maclaurin series (148), which was shown to be a real-valued multilinear map in n arguments, or *n-form* in short.

Recall our original interest to find a functional form for $F : \mathbb{R}^d \to \mathbb{R}$. Although we cannot assume that F is differentiable and can be expressed by a Taylor series expansion, one can show that the discrete Volterra series operator nonetheless can approximate arbitrary continuous F of the required type. This can be seen as follows. The domain of F is a subset $X \subset \mathbb{R}^d$, defined by the image of the *reconstruction map* Rec_d (see main article). X is compact, since Rec_d is continuous and its domain is assumed to be a compact manifold.

A result from functional analysis, the Stone-Weierstrass Theorem [Werner, 2011], now states the following. Suppose Φ is an algebra of continuous, real-valued functions on the compact Hausdorff space X that separates points of X and contains the constant functions. Then for all $\epsilon > 0$ and any continuous real-valued function F on X, a function $p \in \Phi$ exists such that $|F(x) - p(x)| < \epsilon$ for all $x \in X$. This means a polynomial p can be found that approximates the desired function F arbitrarily exactly.

Via γ_0, polynomials of type 157 contain the constant functions. The separation property demands that for $x, y \in X$ with $x \neq y$ there exists a $p' \in \Phi$ such that $p'(x) \neq p'(y)$. Without loss of generality, assume x and y differ in the j^{th} coordinate. Then choose

$$p'(x) = \sum_{k_1=1}^{d} \gamma_1(k_1) x_{k_1}.$$

That is, $\gamma_i = 0$ for $i \neq 1$. If we let $\gamma_1(k) = 1$ for $k = j$, and 0 otherwise, p' will have the desired property. Furthermore, it is a standard result from commutative algebra (see Roman [2007]; Artin and A'Campo [1998]) that the elements of the form 157 generate an algebra. It is in fact *the graded algebra of homogeneous polynomials*, the elements of which are equipped with an infinite dimensional vector space structure and the standard *product of polynomials* defined on elements of this space.

As a result, we may invoke the Stone-Weierstrass Theorem and the desired approximation properties of the discrete Volterra series operator obtain.

8.6.3 *Treatment of stochastic driver input in the statistical model under the Bayesian paradigm*

To deal with the unknown ω_k from equation 126, denote by $\mathbf{w}_j = (\omega_0, ..., \omega_h)^T \in \mathbb{R}^h$ the vector of unknowns, and by $\mathbf{x}_j(x_i, ..., x_{i+d-1})^T \in \mathbb{R}^d$ the vector of known covariates in equation 126, such that $z_j = (\mathbf{x}_j, \mathbf{w}_j)$. The obvious strategy is to integrate $\mathbf{w}$ already out of the prior process 128, and perform the inference step in the predictive distribution with the resulting process independent of $\mathbf{w}$. To this end, we adopt the Gaussian approximation approach suggested by Girard et al. [2003], as outlined in the main article, and compute

$$\begin{aligned} m(\mathbf{x}_i) &= \mathbb{E}[y_i|\mathbf{x}_i] = \mathbb{E}_w[\mathbb{E}[y_i|\mathbf{x}_i, \mathbf{w}_i]] = 0, \\ s(\mathbf{x}_i, \mathbf{x}_j) &= \text{Cov}[y_i, y_j|\mathbf{x}_i, \mathbf{x}_j] \\ &= \mathbb{E}_w[\text{Cov}[y_i, y_j|\mathbf{x}_i, \mathbf{x}_j, \mathbf{w}_i, \mathbf{w}_j]] + \text{Cov}[\mathbb{E}[y_i|\mathbf{x}_i, \mathbf{w}_i]\mathbb{E}[y_j|\mathbf{x}_j, \mathbf{w}_j]] \\ &= \mathbb{E}_w[\text{Cov}[y_i, y_j|\mathbf{x}_i, \mathbf{x}_j, \mathbf{w}_i, \mathbf{w}_j]]. \end{aligned} \tag{158}$$

We will show exemplary derivations for the nonparametric model defined in equation 133, i.e. the full series expansion. The following elementary identities will be used during the derivations:

$$\int_{-\infty}^{\infty} e^{-ax^2} dx = \sqrt{\frac{\pi}{a}} \qquad (a > 0), \tag{159}$$

and

$$\begin{aligned}\int_{-\infty}^{\infty} e^{-ax^2+bx}dx &= \int_{-\infty}^{\infty} e^{-a(x-\frac{b}{2a})^2+\frac{b^2}{4a}}dx \\ &= \exp\left(\frac{b^2}{4a}\right)\int_{-\infty}^{\infty} e^{-ay^2}dy \\ &= \exp\left(\frac{b^2}{4a}\right)\sqrt{\frac{\pi}{a}} \qquad (a > 0).\end{aligned} \tag{160}$$

Without loss of generality, we assume that $\mathbf{w}_i$ and $\mathbf{w}_j$ do not have common or overlapping coordinates, otherwise the expressions simplify even further.

$$\begin{aligned}s(\mathbf{x}_i, \mathbf{x}_j) &= \int\int (K(z_i, z_j) + \sigma^2\delta_{ij})p(\mathbf{w}_i)p(\mathbf{w}_j)d\mathbf{w}_i d\mathbf{w}_j \\ &= \frac{1}{2\pi\sigma_w^{2h}}\int\int \exp(z_i^T z_j)\exp\left(-\frac{1}{2}\frac{\mathbf{w}_i^T\mathbf{w}_i}{\sigma_w^2} - \frac{1}{2}\frac{\mathbf{w}_j^T\mathbf{w}_j}{\sigma_w^2}\right)d\mathbf{w}_i d\mathbf{w}_j \\ &\quad + \sigma^2\delta_{ij}\int\int p(\mathbf{w}_i)p(\mathbf{w}_j)d\mathbf{w}_i d\mathbf{w}_j \\ &= \sigma^2\delta_{ij} + \frac{1}{2\pi\sigma_w^{2h}}\exp(\mathbf{x}_i^T\mathbf{x}_j) \\ &\quad \times \int \underbrace{\int \exp\left(\mathbf{w}_i^T\mathbf{w}_j - \frac{1}{2\sigma_w^2}\mathbf{w}_i^T\mathbf{w}_i\right)d\mathbf{w}_i}_{:=I_{\mathbf{w}_i}}\exp\left(-\frac{1}{2\sigma_w^2}\mathbf{w}_j^T\mathbf{w}_j\right)d\mathbf{w}_j.\end{aligned} \tag{161}$$

Since $\mathbf{w}_i \in \mathbb{R}^h$, with $d\mathbf{w}_i = dw(i)\dots dw(i+h-1)$, $I_{\mathbf{w}_i}$ evaluates to

$$\begin{aligned}I_{\mathbf{w}_i} &= \int\cdots\int \exp\left(\sum_{k=0}^{h-1} w(i+k)w(j+k) - \frac{1}{2\sigma_w^2}\sum_{k=0}^{h-1} w(i+k)^2\right)d\mathbf{w}_i \\ &= \int \exp\left(w(i)w(j) - \frac{1}{2\sigma_w^2}w(i)^2\right) \\ &\quad \dots\int \exp\left(w(i+h-1)w(j+h-1) - \frac{1}{2\sigma_w^2}w(i+h-1)^2\right)d\mathbf{w}_i \\ &\overset{(160)}{=} \left(\sqrt{2\pi\sigma_w^2}\right)^h \exp\left(\frac{\sigma_w^2}{2}\mathbf{w}_j^T\mathbf{w}_j\right).\end{aligned} \tag{162}$$

Substitution into (161) yields

$$\begin{aligned}s(\mathbf{x}_i, \mathbf{x}_j) &= \frac{\left(\sqrt{2\pi\sigma_w^2}\right)^h}{2\pi\sigma_w^{2h}}\exp(\mathbf{x}_i^T\mathbf{x}_j) \\ &\quad \times \int \exp\left(\frac{\sigma_w^2}{2}\mathbf{w}_j^T\mathbf{w}_j - \frac{1}{2\sigma_w^2}\mathbf{w}_j^T\mathbf{w}_j\right)d\mathbf{w}_j + \sigma^2\delta_{ij} \\ &= \frac{\left(\sqrt{2\pi}\right)^{h-2}}{\sigma_w^h}\exp(\mathbf{x}_i^T\mathbf{x}_j) \\ &\quad \times \int \exp\left(\frac{1}{2}\left(\sigma_w^2 - \frac{1}{\sigma_w^2}\right)\mathbf{w}_j^T\mathbf{w}_j\right)d\mathbf{w}_j + \sigma^2\delta_{ij}.\end{aligned} \tag{163}$$

The integral decomposes in a fashion analogous to (162) and, by (159), converges in case

$$\sigma_w^2 - \frac{1}{\sigma_w^2} < 0,$$

which holds if $\sigma_w^2 < 1$. For now, we will therefore assume the latter and discuss the implications later. As a result,

$$\begin{aligned} s(\mathbf{x}_i, \mathbf{x}_j) &= \underbrace{\frac{\left(\sqrt{2\pi}\right)^{h-2}}{\sigma_w^h} \left(\sqrt{\frac{2\pi}{|\sigma_w^2 - \frac{1}{\sigma_w^2}|}}\right)^h}_{:=\sigma'} \exp(\mathbf{x}_i^T \mathbf{x}_j) + \sigma^2 \delta_{ij} \\ &= \sigma' [K(x,x)]_{ij} + \sigma^2 \delta_{ij}. \end{aligned} \tag{164}$$

In practice, σ' is treated as single hyperparameter of the model and estimated accordingly. The restriction $\sigma_w^2 < 1$ is no great cause of concern, since it can be explained away by rescaling the incoming stochastic driver signal in the deterministic part of the dynamics. If treated in this fashion, the effect of the stochastic input on the inference is thus an increase of uncertainty by a multiplicative factor, which is intuitively plausible.

Part III

DISCUSSION

9 DISCUSSION

The scientific goals of this thesis were the investigation of time series analysis tools for prediction, as well as detection and characterization of dependencies, informed by dynamical systems theory. Emphasis was placed on the role of delays with respect to information processing in dynamical systems, as well as with respect to their effect in causal interactions between systems. The three main features that characterize this work are, first, the assumption that time series are measurements of complex deterministic systems. As a result, functional mappings for statistical models in all methods are justified by concepts from dynamical systems theory. To bridge the gap between dynamical systems theory and data, differential topology, as outlined in Chapter 3, was employed in the analysis. Second, the Bayesian paradigm of statistical inference was employed to formalize uncertainty by means of a consistent theoretical apparatus with axiomatic foundation, as discussed in Chapter 2. Third, the statistical models are strongly informed by modern nonlinear concepts from machine learning and nonparametric modeling approaches, such as Gaussian process theory. Consequently, high modeling power is achieved in terms of unbiased approximations of the functional mappings implied by the prior systems level analysis.

The main body of this thesis is comprised of four methodological studies, corresponding to chapters 5 to 8, which address different points in the aforementioned framework. Chapters 5 and 6 document work in the area of delay-coupled reservoir computing. The main area of application is predictive inference on time series. Furthermore, the delay-coupled reservoir framework affords investigating the role of delays in information processing of dynamical systems. In contrast, chapters 7 and 8 document methodological work on detecting and characterizing dependencies in measurements from interacting systems. While Chapter 7 focuses on systems in regimes of strong interaction, i.e. generalized synchronization, Chapter 8 considers a more general setting that accounts for weaker, as well as delayed interaction scenarios. Providing an estimator for the interaction delay is one of the main objectives in this approach. In the following, a discussion of the results is provided in this context.

9.1 PREDICTION

In this section, the results of chapters 5 and 6 pertaining to delay-coupled reservoir computing will be discussed. Summarizing Chapter 5, a general introduction to the emerging field of delay-coupled reservoir computing was given. In particular, we derived approximate closed-form equations for the virtual nodes of a DCR. Using the analytical approximation reduced computational costs considerably and enabled us to study larger networks of delay-coupled nodes, yielding a dramatic increase in nonlinear benchmark performance that was not accessible before. Moreover, the approximation supplied an explicit handle on DCR components, in particular certain hyperparameters, which are otherwise shrouded in a nonlinear functional differential equation. A few of these possibilities were already illustrated in Chapter 5 in a practical application to an experimental recording of a far-infrared laser operating in a chaotic regime. An interesting feature of these data are nonstationarities and catastrophic events which provide hard problems in conventional time series analysis. The DCR model was embedded in a statistical model corresponding to the Bayesian paradigm with a proper quantization of model confidence. This allowed us to study and compare in some depth the occurrence of overfitting with respect to varying model complexity and conditional on varying information content. Most importantly, the confidence intervals arising from the resulting predictive distribution allowed us to pinpoint rare catastrophic events in the time series as sources of prediction error. However, the problem of modeling these rare events could be overcome by conditioning on more data.

The nonstationarities and catastrophic events present in the laser time series are the result of deterministic phenomena of the underlying chaotic systems. As a result, they can be dealt with at the level of functional modeling. In the case of prediction, the objective is to approximate the chaotic flow of the underlying systems in its conjugate form on the reconstructions via the measurement time series. The corresponding theory was discussed in Chapter 3 and amounts to approximating functional mapping 41 on the data. This example clearly motivates the necessity to distinguish between notions of aleatoric uncertainty, such as measurement noise, on the one hand, and epistemic uncertainty. As evident from figure 10, uncertainty pertaining to the nonstationarities is epistemic in nature in this example, since the inclusion of more data points alleviated prediction errors and variance considerably. Alternative model specifications common in time series analysis, such as the family of (generalized) *autoregressive conditional heteroscedasticity* (ARCH) models [Engle, 1982], may also have appeared suitable for this particular dataset, i.e. the latter may test positive for heteroscedasticity. However, this would rather be an artifact of the inherent model linearity. In general, this example demonstrates that analysis may greatly benefit from considering the systemic origin of the data and its features to arrive at a particular functional form of the model. If for some reasons it is explicitly desired to choose the latter as a mixture of processes, Rasmussen (section 2.7, [2006]) provides an interesting discussion in the context of Gaussian process regression which affords the treatment of heteroscedastic notions in terms of nonlinear process priors.

Open questions remain with regard to the choice of optimal hyperparameters of the DCR and, possibly, adaptability to input signals. Foremost, optimal mask functions have to be identified systematically. To this end, the inconvenient coupling to the sampling grid of virtual nodes has to be overcome, so as to make an independent evaluation of mask functions possible. Accounting for and optimizing non-uniform virtual node sampling grids presents a first step towards this goal. In addition, it is a canonical starting point for investigating the role of delays in information processing.

Consequently, the work documented in Chapter 6 focused on enhancing the temporal multiplexing of input to the DCR's nonlinear node by means of an unsupervised plasticity mechanism. The homeostatic nature of the latter relates directly to the information processing properties of the DCR in that it balances between sensitivity to and informational expansion of the input. The study demonstrates that this is possible merely by manipulating sampling points of virtual nodes across the delay span of the system. This key insight presents a first step towards understanding and investigating the role of delays in more general scenarios of information processing in dynamical systems. We speculate that similar multiplexing strategies may be implemented by the extensive dendritic trees in some neuron types and advocate a unified computational account that may integrate both the temporal and spatial aspects of dendritic computations.

The mask and the number of virtual nodes were assumed to be fixed and given in the studies presented so far. The mask is the central concept in a DCR as it completely determines the nonlinear shape of the otherwise convergent reservoir response during one delay span. In principle, the mask is a function on $[-\tau, 0]$. The restriction to discrete domains and piecewise constant functions therefore seems to be a very limited approach. The mask should at least be subject to optimization. However, it's evaluation is inconveniently entangled, both, with the step size of any numerical solver, as well as with the sampling points of the virtual nodes. Varying the latter greatly affects reservoir performance and makes it therefore hard to compare reservoir behavior for masks at different sampling resolutions.

To address this important issue and to lead the way for future research, I present in Appendix C a tentative functional approach to delay-coupled reservoirs. Its main benefit is the elimination of the virtual node concept from the DCR by incorporating the full delay span $[-\tau, 0]$ in the statistical model in a nonparametric fashion. As a result, arbitrary masks can be considered as elements of the Hilbert space $L^2([-\tau, 0], \mathbb{R})$ without restriction to virtual node sampling grids. This line of investigation should help to foster understanding of the fundamental role of the mask for computation and promises to tap the reservoir's full potential as functional, i.e. infinite dimensional, dynamical system. Despite the theoretical nature of this framework, it seems particularly appealing in combination with a physical realization of the DCR that can provide an explicit solution to the delay differential equation. The functional approach also suggests different sampling strategies for the physical system, e.g. by choice of a Fourier basis for L^2, which would in theory allow one to read out the reservoir in the frequency domain and only once in each delay-cycle. This overcomes practical limitations with respect to sampling frequencies of the system in the time domain. More research in this regard may also lead to DCRs that yield covariance matrices for fully nonparametric Gaussian process regression.

Finally, it remains an open problem to investigate the approximation properties of the functional mappings resulting from DCRs. In particular, it would be desirable to characterize a class of DCRs that can approximate arbitrary continuous functions with domain and range given by subsets of $\mathbb{R}^n$ and $\mathbb{R}$ respectively. The functional approach appears to be once more most promising in this regard. With respect to physical realization in light of the Bayesian paradigm of statistical inference, the possibility of inverting matrices fully optically is very intriguing since this would allow an optical implementation of an inference machine that can in each time step update its own conditioning on data, always incorporating most recent input from some data stream into its online predictions. In machine learning terms, the model would be "retrained" conditional on most recent data in each time step.

In this sense, computation would be adaptive to slow changes in the input stream and thus free of maintenance.

9.2 DETECTION AND CHARACTERIZATION OF DEPENDENCIES

Chapters 7 and 8 presented studies of methods to detect and characterize dependencies between time series measurements. These methods were essentially guided by theoretical studies of interactions between chaotic dynamical systems. As was discussed at an earlier point, chaotic systems have dynamics that are complex enough to be interesting but low-dimensional enough to allow statistical models to fully capture their dynamics. In addition, they are often characterized by irregular oscillatory activity, reminiscent of features observed in neuronal time series measurements, which makes them good candidates for testing and developing statistical methods for application in neuroscience. In particular, they allow the statistical modeling to be complemented by insights from theoretical studies of coupled systems. The latter allows for various analytical characterizations of interaction regimes which are for example parametrized by the coupling strength.

In essence, it makes sense to distinguish between two main regimes of interactions. The first is a weak coupling scenario where in particular driven systems retain autonomous dynamics in the presence of driver influence. The second characterizes stronger coupling scenarios which are described by the concept of generalized synchronization. Synchronization becomes in this context a synonym for predictability since a driven system loses autonomous dynamics, i.e. driver and driven system collapse onto a common synchronization manifold. The full state space is thus accessible by measurements from both, driver and driven system, which means knowledge of one affords prediction of the respective other.

This implies the existence of a particular predictive functional mapping between time series that was employed in a statistical modeling approach, as documented in Chapter 7. The model worked with an interesting variant of the standard Volterra series operator by expanding the kernel parts as functions instead of individual scalar coefficients. A parametric form was given by splines, which reduces parameters considerably and is therefore of high interest for reconstructing high-dimensional dynamical systems. It is in principle no problem to obtain reconstruction dimensions with magnitude of 10^3 or higher, in so far as the data permits it. Apart from selected coupled chaotic systems, the method was evaluated on local field potential data recorded from electrodes in macaque primary visual cortex. The results clearly indicate the applicability of functional statistical models to raw biological data. A particular qualm often arises in this regard with respect to the amount of data necessary to obtain sufficient "statistical power". Typically, the issue here will be to reduce epistemic uncertainty in the reconstruction of the underlying dynamical systems in case they are high-dimensional. Although the latter is to be expected in neuronal data where electrodes might record form more than 10000 cells at the same time, our studies indicate that even single trial analysis can be an option with functional statistical modeling. This is owed to the fact that, apparently, lower dimensional reconstructions of the involved neuronal dynamics already capture substantial information for inference.

The latter remark may no longer hold in situations where the dynamics are not characterized by synchrony. Chapter 8 documents a study that generalizes the findings from Chapter 7 in this regard. First and foremost, a different functional formalization was considered to extend to weaker coupling scenarios in the underlying dynamical systems. This allowed us to exploit asymmetries in reconstructibility of time series to arrive at conclusions regarding the direction of information flow. Furthermore, the method considered the case of delayed coupling which has to be acknowledged as the standard scenario in studies of spatially distributed physical systems, such as the brain.

The approach featured in the second study benefits greatly from the combination of different theoretical branches. First, the existence of a particular functional mapping is rigorously argued for in terms of dynamical systems theory and differential topology. Second, given a corresponding mapping between time series, an unbiased functional form can be chosen accordingly in terms of approximation properties from analytic considerations. In a final step, a statistical model is established that properly formalizes uncertainty and allows for consistent inference. Most importantly, given the subjective and epistemic interpretation of uncertainty adopted in this thesis, the prior theoretical considerations directly allow for a reduction of uncertainty by providing explicit conditions and forms of the targeted functional dependencies. Those parts of these dependencies that enter the statistical model as random variates are clearly delineated. This formalization grants the practitioner more certainty in interpreting the findings and may prevent unwarranted discarding of information. Moreover, it allows an exploitation of systemic properties of data beyond mere statistical statements.

In particular, as was shown in Chapter 8, not only can the analysis in terms of functional models serve to estimate the delay of the interactions, the delay may also help in determining direction of information flow in strong coupling scenarios, similar to those considered in Chapter 7. The main application of the method was to local field potential recordings from cat visual areas where it obtained a highly plausible diagram of connectivity and corresponding delay estimates between different layers and areas in visual cortex. Surprisingly, as was previously found in Chapter 7, low-dimensional reconstructions yielded enough information to arrive at the desired inference. At the same time, however, large parts of the targeted signals could not be accounted for. We speculate that this relates to more complex and high-dimensional dynamics in unsynchronized interaction regimes over greater spatial distances. Even more surprising, most informative with regard to the targeted interactions were parts of the time series that many practitioners in the field discard routinely.

In this context, it may also be interesting to take a critical look at the related popular method of *Granger causality*, as discussed in Appendix D. Developed in a field dominated by white-noise based stochastic process models, at a time before the theoretical tools of differential topology, as outlined in Chapter 3, were available, Granger causality necessarily disregards insights from dynamical systems theory. As explained in detail in Appendix D, this disregard may lead to wrong inference with respect to causal relationships between time series. In a related note, the discussion of Chapter 8 considered the application of *independent component analysis*, ubiquitously used in the neurosciences, to partially overlapping measurement time series. As already pointed out, instead of "disentangling" shared information between channels, ICA may potentially obfuscate causal analysis by creating *pseudo-measurements*, the independent components, which all contain the same complete information about the original measurement signals and are thus pairwise reconstructible by arguments from differential topology. That is, while the notion of *statistical independence* between independent components is addressed by ICA, the systemic aspects of shared information are not. The above examples thus show that not only can data analysis greatly benefit if informed by dynamical systems theory but that practitioners ignore the latter at their own peril.

Regarding the directions of future research, the reduction of epistemic uncertainty must be the foremost goal of analysis. While it is encouraging how much insight can be gained with low reconstruction dimensions, many aspects of the data remain unaccounted for and defy reconstruction. Data analysis should be further augmented by domain specific simulations and theoretical models that incorporate plausible coupling and network structures to study in how far and under what conditions the dimensionality of the observed system collapses. This should lead to better prior expectations regarding reconstruction dimensions and the application of differential topology to data. Furthermore, the limiting factor in the latter are not the computational methods. As discussed in the context of Chapter 7, there are no real technical problems in using models that can account for a dimensionality of several thousand, which should in principle be sufficient to reconstruct neural dynamics underlying local field potentials. Most of the restrictions arise directly from the data, in particular if intricate temporal dependencies on stimuli apply whose time span produces only short signals at relevant time scales. Therefore, theoretical models have to inform experimental designs in order to produce time series which afford reconstruction of the measured dynamical systems in relevant dimensionality.

Part IV

APPENDIX

A

THE SAVAGE REPRESENTATION THEOREM

The following theorem is taken from Fishburn [1970], with some of the definitions modified for better readability by the original form in [Savage, 1954]. Let S be the set of states, $\mathcal{C}$ the set of consequences and $\mathcal{A}$ the set of all acts as functions on S into $\mathcal{C}$. In addition, $\mathcal{S} = 2^S$.

Theorem 5 Savage Representation Theorem
Suppose that the following seven conditions hold for all $f, g, h, f', g' \in \mathcal{A}$; $A, B \in \mathcal{S}$; $c, d, c', d' \in \mathcal{C}$:

P1 $\prec$ *on $\mathcal{A}$ is a weak order.*

- *a) Either $f \prec g$ or $g \prec f$.*
- *b) If $f \prec g$ and $g \prec h$ then $f \prec h$.*

P2 If f, g and f', g' are such that:

- *a) in B^c, f agrees with g and f' agrees with g',*
- *b) in B, f agrees with f' and g agrees with g',*
- *c) $f \prec g$*

then $f' \prec g'$.

P3 If $f \equiv c$, $f' \equiv c'$ and B is not null, then $f \prec f'$ given B, if and only if $c \prec c'$.

P4 If c, c', d, d'; A, B; f_A, f_B, g_A, g_B are such that

- *a) $c' \prec c, \quad d' \prec d$,*
- *b) $f_A(s) = c, \quad g_A(s) = d \quad$ for $s \in A$,*
  $f_A(s) = c', \quad g_A(s) = d' \quad$ for $s \in A^c$,
- *c) $f_B(s) = c, \quad g_B(s) = d \quad$ for $s \in B$,*
  $f_B(s) = c', \quad g_B(s) = d' \quad$ for $s \in B^c$,
- *d) $f_A \prec f_B$,*

then $g_B \prec g_B$.

P5 There is at least one pair of consequences c, c' such that $c \prec c'$.

P6 If $f \prec g$ and c is any consequence then there exists a finite partition of S such that, if A is any event in the partition, it holds that

- *a) ($f' = c$ on A, $f' = f$ on A^c) $\Rightarrow f' \prec g$,*
- *b) ($g' = c$ on A, $g' = g$ on A^c) $\Rightarrow f \prec g'$.*

P7 If $f \prec g(s)$ given B for every $s \in B$ then $f \prec g$ given B.

Then, with $\prec^*$ defined on $\mathcal{S}$ by

$$A \prec^* B \Leftrightarrow f \prec g$$

whenever

$$c \prec d \quad \wedge \quad f(A) = d,\ f(A^c) = c,\ g(B) = d,\ g(B^c) = c,$$

there is a unique *probability measure P^* on $\mathcal{S}$ that satisfies*

$$\forall A, B \in \mathcal{S}: \quad A \prec^* B \Leftrightarrow P^*(A) < P^*(B)$$

and that is non-atomic*,*

$$\forall B \in \mathcal{S} : \exists C \subset B, \rho \in [0,1] : \quad P^*(C) = \rho P^*(B),$$

and, with P^* as given, there is a real-valued function U on $\mathcal{C}$ that is bounded, unique up to affine transformation, and for which

$$\forall f, g \in \mathcal{A} : \quad f \prec g \quad \Leftrightarrow \quad \mathbb{E}_{P^*}[U(f(s))] < \mathbb{E}_{P^*}[U(g(s))].$$

B

NORMAL FORM OF THE PREDICTIVE INFERENCE DECISION PROBLEM

In this section we state the normal form of decision problem (29). The main difference is the fact that the data is not assumed to be given before a decision has to be made. Instead, assumptions about the sampling space have to be made. In the context of time series analysis, this is possible because the data samples are assumed to be measurements from dynamical systems with bounded dynamics on a compact manifold. Thus, $y_* \in Y \subset \mathbb{R}$ and $x_* \in X \subset \mathbb{R}^m$. As a matter of fact, X may even be a submanifold. Estimator $\hat{y}_* = f(x_*)$ is now a mapping of covariate vector x_* which is not given, and the corresponding *Bayes risk* is stated as

$$L^*(y_*, x_*)(f) = \int_Y \int_X (f(x_*) - y_*)^2 p(y_*, x_* | D_*) dx_* dy_*. \tag{165}$$

Note that the L^* optimization problem is now in the function $f : X \to Y$, $f \in Q$, and can no longer be treated by ordinary calculus. Instead, the *Gâteaux derivative* of functional $L^* : Q \to \mathbb{R}$ has to be considered as generalization of the concept of directional derivative. The required theory pertains to results from nonlinear functional analysis and is covered by Werner [2011].

The first obstacle lies with the requirement that domain and image of functional L^* have to be normed spaces. In the context of time series as measurements from complex systems we can make the assumption that $Q := C^r(\mathbb{R}^m)$, the space of continuous, $0 \leq r < \infty$ times differentiable functions. This space carries naturally a weak topology on compacta $K \subset \mathbb{R}^m$, generated by the subbase consisting of neighborhoods

$$U_{K,\epsilon,g,r} := \left\{ f \in C^r(\mathbb{R}^m) : \sup_{x \in K} |D^r(f(x) - g(x))| < \epsilon,\ \epsilon > 0, g \in C^r(\mathbb{R}^m) \right\}.$$

It can be shown that this space is complete metrisable and one can define $\|x\| := d(x, 0)$ in terms of the metric d. As a result, it is sensible to define the Gâteaux differentiability of L^* on an open subset $U \subset Q$, $x_0 \in U$, if there exists a continuous linear operator $T : Q \to \mathbb{R}$ such that for $h \in \mathbb{R}$

$$\lim_{h \to 0} \frac{L^*(x_0 + hv) - L^*(x_0)}{h} = Tv, \quad \forall v \in Q, \tag{166}$$

where the derivative is denoted by $Df(x_0) := T$. Denote by ϕ_{x_0} the helper function $\phi_{x_0}(h) := L^*(x_0 + hv)$, which exists for $|h| < \alpha / \|v\|$ in case $\{x : \|x - x_0\| \leq \alpha\} \subset U$.

We can now write the Gâteaux differential quotient (166) for our optimization problem in equation 165 as

$$\begin{aligned}\frac{d}{dh}\,\phi_f(h)\Big|_{h=0} &= \frac{d}{dh}\int_Y\int_X (f(x_*)+hv(x_*)-y_*)^2 p(y_*,x_*|D_*)dx_*dy_*\Bigg|_{h=0}\\ &= \int_Y\int_X \frac{d}{dh}(f(x_*)+hv(x_*)-y_*)^2 p(y_*,x_*|D_*)dx_*dy_*\Bigg|_{h=0} \qquad (*)\\ &= \int_X \underbrace{2\int_Y (f(x_*)-y_*)p(y_*,x_*|D_*)dy_*}_{:=\frac{\delta L^*}{\delta f}} v(x_*)dx_*.\end{aligned} \tag{167}$$

Exchanging integrals and derivative in line $(*)$ can be legitimized by corollary A.3.3 in [Werner, 2011]. Furthermore, it is easy to prove the following theorem (III.5.6 in [Werner, 2011]),

Theorem 6 *If $U \subset X$ open, $f : U \to \mathbb{R}$ Gâteaux-differentiable with local extremum at $x_0 \in U$, then $Df(x_0) = 0$.*

Since the domain of f outside of X is of no interest, we can weaken assumptions pertaining to the direction of derivation by

$$v \in \left\{ v \in C^r(X) : v^{(k)}\Big|_{\partial X} = 0, k = 0, 1, ..., r \right\}$$

and conclude $(DL^*(f) = 0) \Rightarrow \left(\frac{\delta L^*}{\delta f} = 0\right)$ on X (see lemma 4.4 and exercise 4.5 in [Troutman, 2012]), where ∂X denotes the boundary of compactum X and $v^{(r)}$ the r^{th} derivative. Different assumptions could be made here to simplify the optimization problem in a similar fashion, but the given one serves to make the point below. With the above restriction on v, the solution to the decision problem can be calculated as

$$\begin{aligned} 0 &= \frac{\delta L^*}{\delta f}\\ 0 &= f(x_*)p(x_*) - \int_Y y_* p(y_*|x_*, D_*)p(x_*)dy_* \\ f(x_*) &= \int_Y y_* p(y_*|x_*, D_*)dy_* = \mathbb{E}[y_*|x_*]. \end{aligned} \tag{168}$$

The choice of $P(x_*)$ is arbitrary, however, the restriction of x_* to a compact submanifold $X \subset \mathbb{R}^m$ would allow the choice of a non-informative proper prior uniform distribution on X. The result of the decision problem in normal form is thus indeed the same as in chapter 2 but requires a substantial amount of theory and additional assumptions for a correct treatment.

C

RESERVOIR LEGERDEMAIN

The original delay-coupled reservoir approach consisted of the idea to implement a nonlinear regression for time series analysis by means of a functional model given as the solution of a retarded functional differential equation, the latter being fully optically or electronically realized, for example as a laser system with nonlinear inteference given by its own delayed feedback.

Using the Mackey-Glass nonlinearity to operate the retarded system in a simple fixed point regime (given constant input), the resulting differential equation is given by

$$\frac{dx(t)}{dt} = \gamma g[x(t-\tau), m(t)u(t)] - x(t) \tag{169}$$

where

$$g(y) = \frac{y}{1+y} = \sum_{n=1}^{\infty} (-1)^{n+1} y^n. \tag{170}$$

For $(i-1)\tau \leq t \leq i\tau$, the ith τ-cycle is considered as a single reservoir time step during which $u(t) = u_i = const.$ As such, $x(t)$ would simply saturate and converge onto a fixed point determined by u_i, for example $\lim_{t\to\infty} x(t) = 0$ for $u_i = 0$. To add perturbation and create a "*feature expansion*" of the input signal useful for computation, the delayline τ is shattered into N subintervals of length θ, on which the mask is constant. That is, the mask is a function on $[-\tau, 0]$ which is piecewise constant,

$$m(t) = m_j \qquad \text{for} \quad (j-1)\theta < t \leq j\theta.$$

The m_j were originally simply random samples from $\{-1, 1\}$ meant to perturb the system and create transient trajectories chasing the fixed point, determined by m_j, as it "*jumps around*" in phase space. In this setup, θ is experimentally determined to be short enough to prevent practical convergence of the system trajectory during the span of θ, but long enough for the system to act upon the masked input signal. It was determined that for T being the intrinsic time scale of the system, $\theta = T/5$ yields good computational performance. Since we chose $T = 1$, $\theta = 0.2$ accordingly.

In the optical implementation, one then simply samples $x(t)$ at the end of each θ-interval, calls the corresponding jth sample during the ith τ-cycle $x((i-1)\tau + j\theta) = x_j(u_i)$ a "*virtual node*", and uses a functional mapping

$$\hat{y}_i = \sum_{j=1}^{N} a_j x_j(u_i) \approx f(u_i, ..., u_{i-M}) \tag{171}$$

to predict or model some target signal y as a function of covariate time series u. The memory capacity of the system, denoted by M can be experimentally determined to be in the order of $M = 10$. The $\tau = N\theta$ has no impact on the memory capacity and can simply be chosen to harbor an adequate number N of virtual nodes along the delay line. Typically, $N = 400$ is chosen as a good balance of computational cost and reservoir performance in benchmark tasks.

In order to study the reservoir computer (171), system 169 has to be solved and virtual nodes sampled accordingly. However, since (169) is a nonlinear DDE it can neither be solved analytically, due to the recursive terms, nor exists a Peano-Baker series, due to the nonlinearity of g. In approximation, the system always has to be sampled at a grid which must encompass the virtual node sampling points as a subset.

From a modeling perspective, the hyperparameters of the system are thus buried in a nonlinear non-solvable retarded functional differential equation, hardly accessible to optimization. Furthermore, due to the piecewise sampling procedure of the virtual nodes, the shape of m is strongly restricted and inconveniently entangled with the sampling points θ_j, as well as the numerical simulation grid. At the same time, the mask function is the most important constituent of the DCR since it provides the only source of perturbation in the span of one delay. The mask therefore completely specifies the computational properties of the reservoir. In order to optimize this important parameter, it has to be disentangled from the virtual node samples.

To address these issues, we start by solving system (169) theoretically. Denote by $x_i(\sigma)$ for $\sigma \in [-\tau, 0]$ our system state on the i^{th} τ-interval. Given x_{i-1}, we can use the *Variation of Constants* to solve (169) as

$$x_i(\sigma) = x_{i-1}(0)e^{-(\tau+\sigma)} + e^{-(\tau+\sigma)} \int_{-\tau}^{\sigma} g[x_{i-1}(s), u_i m(s)]e^{s+\tau} ds, \tag{172}$$

where u_i denotes the constant in the i^{th} reservoir time step, corresponding to τ-interval i, and we assume $m(\sigma) \in \mathcal{L}_2([-\tau, 0])$. Due to the recursion in x_{i-1} the integral cannot be solved analytically in this situation, and because g is nonlinear, no Peano-Baker series expansion has, to the best of my knowledge, been found for this type of delay differential equation.

Given the exact $x_i(\sigma)$ from equation (172), we are now interested in evaluating a good choice for the mask $m(\sigma)$. $\mathcal{L}_2([-1, 1])$ is an infinite dimensional inner product space that can be spanned by *Legendre* polynomials, which provide an orthogonal basis with respect to the $\mathcal{L}_2$ inner product. They are constructed via the recursive relationship

$$(b+1)P_{b+1}(s) = (2b+1)sP_b(s) - bP_{b-1}(s)$$

anchored by $P_0(s) = 1$,$P_1(s) = s$. It holds that

$$\int_{-1}^{1} L_i(\sigma)L_j(\sigma)d\sigma = \delta_{ij}\frac{2}{2i+1} \tag{173}$$

Using an affine transformation on the input domain, $\tilde{\sigma}(\sigma) = \frac{2\sigma}{\tau} + 1$, we can translate the polynomials back onto $[-\tau, 0]$ without harming the orthogonality properties:

$$\begin{aligned}\int_{-\tau}^{0} P_i(\frac{2\sigma}{\tau}+1)P_j(\frac{2\sigma}{\tau}+1)d\sigma &= \int_{\tilde{\sigma}(-\tau)}^{\tilde{\sigma}(0)} P_i(\tilde{\sigma})P_j(\tilde{\sigma})\frac{\tau}{2}d\tilde{\sigma} \\ = \frac{\tau}{2}\int_{-1}^{1} P_i(\tilde{\sigma})P_j(\tilde{\sigma})d\tilde{\sigma} &= \begin{cases} \frac{\tau}{2i+1} & \text{if } i = j \\ 0 & \text{else} \end{cases} .\end{aligned} \tag{174}$$

An orthonormal basis on $[-\tau, 0]$ is therefore given by

$$L_i(\sigma) = \sqrt{\frac{2i+1}{\tau}} P_i(\sigma).$$

We can now write

$$m(\sigma) = \sum_{n=0}^{N} m_n L_n(\sigma),$$

where we could take $N \to \infty$. In practice, however, the higher N, the faster $m(\sigma)$ changes. Since in practice we are interested in evaluating functions appropriate for physical implementation, we need only consider finite N for which the sum will evaluate to a practically smooth function with "*finite change rate*", which is admittedly an ill-defined concept but may provide sufficient detail for our purposes at this time.

Although m is now continuously defined on $[-\tau, 0]$, we are still forced to evaluate it at best using the finite sampling grid $(\chi_j)_{j=0}^{N}$, which seems to be a rather limiting and unsatisfactory procedure. In order to get rid of the dependence on "virtual node samples" and to exploit the full power of our inifinite-dimensional dynamical system, we could try to rewrite our functional model (171) in the following way. Let

$$f(u_i, ...) = \int_{-\tau}^{0} a(\sigma) x_i(\sigma) d\sigma, \tag{175}$$

where $a \in \mathcal{L}_2([t, 0], \mathbb{R})$ and $a \sim \mathcal{GP}$ a Gaussian process with

$$\begin{aligned} \mathbb{E}[a] &= 0, \\ \mathbb{V}[a(\sigma_i), a(\sigma_j)] &= \mathbb{E}[a(\sigma_i)a(\sigma_j)] = \lambda^2 \delta_{ij}, \end{aligned} \tag{176}$$

where δ_{ij} is the Kronecker delta. Expanding $a(\sigma) = \sum_{h=0}^{\infty} a_h L_h(\sigma)$ in equation (175) yields

$$f(u_i, ...) = \sum_{h=0}^{\infty} a_h \int_{-\tau}^{0} L_h(\sigma) x_i(\sigma) d\sigma. \tag{177}$$

Since $x_i(\sigma) \in C_\tau^\infty$ is smooth (compare eq. (170)) and bounded while operating in the fixed point regime, it can be expanded into a Legendre series as well. Furthermore, since x is convergent given constant input and the only change is induced by $m(\sigma)$, $x_i(\sigma)$ will change on the same time scale as m and is thus well-approximated on $[-\tau, 0]$ by a truncated series

$$x_i(\sigma) \approx \sum_{q=0}^{N} c_q^{(i)} L_q(\sigma). \tag{178}$$

If we now use formula (55) with a fine enough sampling grid $\alpha = \alpha_1, ..., \alpha_S$ (that is, $\alpha_{j+1} - \alpha_j$ should be smaller than the time scale at which the mask changes), we can readily find an excellent approximate sampling vector

$$\vec{x}_i = \begin{pmatrix} x_i(\alpha_0) \\ \vdots \\ x_i(\alpha_S) \end{pmatrix} \in \mathbb{R}^{S+1}.$$

This approximation will serve for illustration in the remainder of this chapter. However, the sampling vector can of course be chosen exactly by reading out a physical implementation of the system, e.g. an optical self-coupled laser system that physically realizes a solution to equation 169. This makes the following considerations particularly interesting for practice.

Since no fast changes in between α_j are expected and we know the system evolves smoothly between α_j and α_{j+1}, $x_i(\sigma)$ will be well approximated between sampling points by the Legendre expansion above. Determining the $c_q^{(i)}$ is now reduced to an interpolation problem which can be solved by maximum likelihood regression as

$$c^{(i)} = \begin{pmatrix} c_1^{(i)} \\ \vdots \\ c_N^{(i)} \end{pmatrix} = (L^T L)^{-1} L^T \vec{x}_i \in \mathbb{R}^{N+1}, \tag{179}$$

where L denotes the matrix of Legendre polynomials evaluated on the grid α,

$$L = \begin{pmatrix} L_0(\alpha_0) & \cdots & L_N(\alpha_0) \\ \vdots & \ddots & \vdots \\ L_0(\alpha_S) & \cdots & L_N(\alpha_S) \end{pmatrix} \in \mathbb{R}^{(S+1)\times(N+1)}. \tag{180}$$

We can now rewrite functional equation (175) and, by a *slight of hand*, arrive at

$$\begin{aligned} f(u_i, ...) &\approx \sum_{h=0}^{\infty} a_h \int_{-\tau}^{0} L_h(\sigma) \sum_{q=0}^{N} c_q^{(i)} L_q(\sigma) d\sigma \\ &= \sum_{h=0}^{\infty} \sum_{q=0}^{N} a_h c_q^{(i)} \int_{-\tau}^{0} L_h(\sigma) L_q(\sigma) d\sigma \\ &\overset{(174)}{=} \sum_{q=0}^{N} a_q c_q^{(i)} = a^T \underbrace{(L^T L)^{-1} L^T}_{:=\Lambda} \vec{x}_i = a^T \Lambda \vec{x}_i. \end{aligned} \tag{181}$$

This expression is practically exact for the right α and, for all practical intents and purposes, as good as the unobtainable continuous analytical solution.

We evaluate the *Reservoir Legerdemain* on 1000 samples of NARMA-10 in both, training and validation set, with $\tau = 100, \gamma = 0.4$ and $\sigma_\epsilon^2 = 0$. The $u(k)$ are projected as input into the reservoir and are being held constant over the span of one τ-cycle. The reservoir is sampled using the approximate equation (55) with a grid $\alpha \in \mathbb{R}^{10001}$, which yields a practically exact solution of the system operating in the simple fixed point regime. The high number of samples is mainly chosen to allow the corresponding series expansion in higher orders, since for sampling sizes small with respect to the series order N, the Legendre matrix L (180) may be rank deficient. At the same time, series orders as low as 200 or less may still provide an approximation of the system suitable for reservoir computation without loss of performance. Figure (32) shows the 100^{th} τ-interval for the presentation of the training data set to the system. Dark blue is the reservoir trajectory, cyan the resulting legendre approximation on this interval. Red shows the mask, expanded into a Legendre series of order 500, where the coefficients are randomly chosen from $\{-1, 1\}$. As

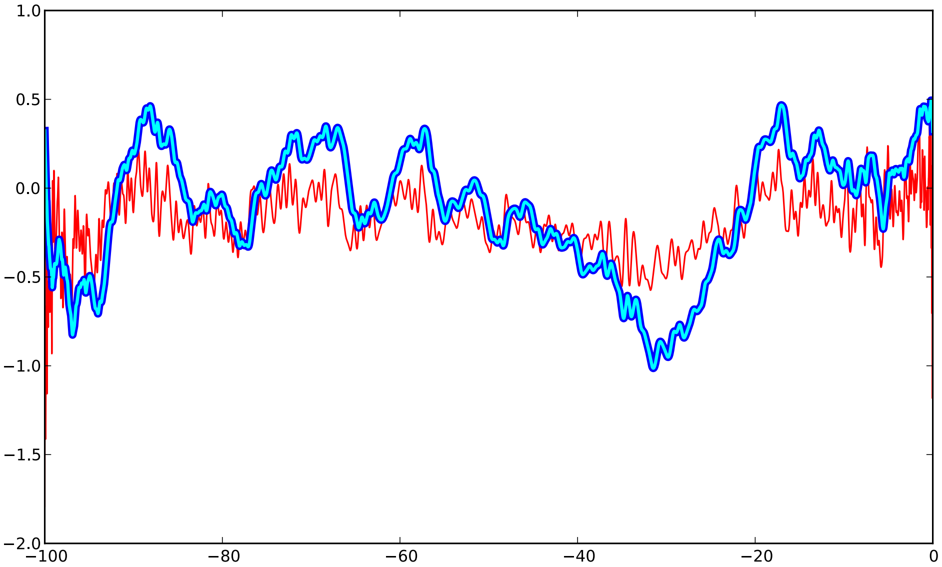

Figure 32: Reservoir Legerdemain trajectory and mask. Red: mask, blue: reservoir trajectory $x_i(\sigma)$, cyan: Legendre fit to $x_i(\sigma)$

can be seen for this exemplary case, the Legendre curve fit is indeed perfect. We can therefore assume that, given the high resolution α, expression (181) operates with a practically exact solution of the Mackey-Glass system here. We predict the target y of the NARMA-10 task using the optimal predictor $\hat{y}$ given by expression (66), where the pseudo-inverse in the expression is computed by singular value decomposition. On the validation set, given the random smooth mask, we reach a performance of $r^2 \approx 0.98$ in squared correlation. Across data sets, the variance of the estimator is extremely low, although different random smooth masks differ substantially in performance.

To exploit the full potential of the continuous *Reservoir Legerdemain*, a next step is to use further Bayesian statistics to start searching for optimal masks, which can now be developed as arbitrary functions in $\mathcal{L}_2$ that are continuously defined on $[-\tau, 0]$ without the limitations formerly enforced by the virtual node sampling points. In fact, the "virtual nodes concept" has been completely eliminated from the DCR, which was the purpose of this chapter.

As a final remark, we note that the choice of Legendre basis for $\mathcal{L}_2$ served as an illustration but is not imperative. The same properties would be shared, for example, by the *Fourier basis*, i.e. the set of functions $\{e_n = e^{2\pi i n x} : n \in \mathbb{Z}\}$ which is an orthonormal basis for $\mathcal{L}_2([-\pi, \pi])$. This creates exciting opportunities for more advanced readout mechanisms of the physical system: Instead of sampling the system in time, one could sample it approximately in the frequency domain by a number of hardware components that supply frequency coefficients $c_q^{(i)}$ for a truncated series expansion analogous to (178).

D

A REMARK ON GRANGER CAUSALITY

Granger Causality is a method developed in the 1960s by Granger [1969] to analyze econometric time series. In more recent years, Granger Causality has become increasingly popular also in the neurosciences to analyze data from EEG, MEG, or electrode recordings [Seth, 2007]. Granger Causality usually assumes two (or more) time series X and Y can be approximated by linear stochastic processes. For illustration, consider scalar time series with corresponding stochastic process model

$$\begin{aligned} X(t) &= \sum_{j=1}^{p} A_j X(t-j) + \sum_{j=1}^{p} B_j Y(t-j) + \epsilon_X(t) \\ Y(t) &= \sum_{j=1}^{p} C_j Y(t-j) + \sum_{j=1}^{p} D_j X(t-j) + \epsilon_Y(t), \end{aligned} \tag{182}$$

where $\epsilon(t)$ denotes the models' residuals, usually taken to be a white-noise random vector. If, for example, the variance of $\epsilon_X(t)$ is reduced by the inclusion of the Y terms in the first equation (that is, $\forall j : B_j > 0$), then it is said that Y Granger-causes X.

Note that, although stochastic process models have been successfully employed in econometric time series forecast, usually no further justification for the existence of such autoregressive models is given, nor are the implications for the underlying systems that generated the time series discussed (and vice versa). Furthermore, when considering highly nonlinear complex systems (compare figure 1), the linearity of the model may pose a severe restriction. Although nonlinear variants of Granger Causality have been developed, their use appears to be less common, since they can be more difficult to use in practice and it is sometimes stated [Seth, 2007] that their statistical properties are less well understood. Justifying the particular choice of a nonlinear model poses a further problem that practitioners may be less inclined to tackle.

Sugihara et al. [2012] alludes to the following problematic case. If the driving system is finite dimensional, the embedding can extend to the driver, as a result of theorem 3. Consequently, the reliability of the inference with Granger causality breaks down, because an autoregressive model can be found which already contains all relevant information about the driver from an observable of the driven system alone. Sugihara therefore sees the main area of application for Granger causality in purely stochastic processes. This has devastating implications for data analysis. If $X(t)$ and $Y(t)$ are time series as assumed in eq. 41 and X is driven by a purely stochastic process, the measurements $Y(t)$ must directly represent this stochastic driver for inference in the Granger framework, since it could not be reconstructed from a delay embedding. Furthermore, most time series data in natural science will probably not represent samples from an actual random process, which we think of rather as a tool to formalize epistemic uncertainty than a truly random phenomenon to be frequently encountered in a natural environment. In particular in neuroscience one would hope time series data are measurements from systems that are, by and large, deterministic, for if the neural dynamics of the brain were dominated mainly by random events, no free will or even coherent behavior would be possible.

Note that Granger causality could in principle be applied to deterministic weak coupling scenarios if special care is taken to involve the reconstruction dimension in the inference procedure. Suppose $X(t) = \phi_x(x(t))$ and $Y(t) = \phi_y(y(t))$ are measurements from systems given by diffeomorphisms F and G on m and n dimensional manifolds M and N respectively, as in Chapter 3, and assume X is driven by Y. Recall that a bundle embedding (see theorem 4)

$$\Phi_{F,G,\phi_x,y} : M \times \{y\} \to \mathbb{R}^d$$

is an embedding defined by $\Phi_{F,G,\phi_x,y}(x) = \Phi_{F,G,\phi_x}(x,y)$ for typical $y \in N$. Thus, if one has independent knowledge of y_i, and Φ_{F,G,ϕ_x,y_i} and $\Phi_{F,G,\phi_x,y_{i+1}}$ are embeddings,

$$z_{i+1} = \Phi_{F,G,\phi_x,y_{i+1}} \circ F_{y_i} \circ (\Phi_{F,G,\phi_x,y_i})^{-1}(z_i) \tag{183}$$

holds and defines a temporal evolution function H_{y_i} as in equation 40 and figure 3. Denote again its last component by $h_{y_i} : \mathbb{R}^d \to \mathbb{R}$, such that

$$\phi(x_{i+d}) = h_{y_i}(\phi_x(x_i), \phi_x(x_{i+1}), ..., \phi_x(x_{i+d-1})). \tag{184}$$

Then h_y can be estimated as a function with domain $\mathbb{R}^d \times N$. To estimate the covariates from N, measurements ϕ_y of the driving system can be used to reconstruct y_i via delay embeddings $(\phi_y(y_i), \phi_y(y_{i+1}), ..., \phi_y(y_{i+d-1}))^T$, corresponding to the cross terms $Y(t-j)$ in the first line of equation 182. Importantly, for the bundle embedding it is enough to have $d \geq 2m+1$ in the embeddings $\Phi_{F,G,\phi_x,y}(x)$, since the driver is reconstructed separately via the $Y(t-j)$. In contrast, without the additional covariates $Y(t-j)$, theorem 3 applies and the embedding has to reconstruct the driver as well. In this case, we need $d \geq 2(m+n)+1$. If appropriate model selection strategies are employed to arrive at a model with minimal number of covariates, a substantial increase of d should be necessary to retain model performance if covariates from the putative driver are removed from the model. This may yet serve for inference regarding the direction of information flow. If $n \neq m$, however, p should be different for covariates $X(t-j)$ and $Y(t-j)$, respectively. A selection strategy where the full model 182 with cross terms is fitted, followed by tests to determine significant deviation from zero for the model coefficients A_j, B_j may not be reliable.

As a side remark, this line of reasoning appears to suggest that in cases where $n > m$, it could be helpful to include “acausal” cross terms in equation 182.

BIBLIOGRAPHY

H. Abarbanel, N. Rulkov, and M. Sushchik. Generalized synchronization of chaos: The auxiliary system approach. *Physical Review E*, 53(5):4528–4535, May 1996. ISSN 1063-651X. URL http://www.ncbi.nlm.nih.gov/pubmed/9964787.

D. Aeyels. Generic observability of differentiable systems. *SIAM Journal on Control and Optimization*, 19(5):595–603, 1981.

A. Albert. *The Penrose-Moore Pseudo Inverse with Diverse Statistical Applications. Part I. The General Theory and Computational Methods*. Defense Technical Information Center, 1971. URL http://books.google.de/books?id=Web4tgAACAAJ.

L. Appeltant, M. C. Soriano, G. Van der Sande, J. Danckaert, S. Massar, J. Dambre, B. Schrauwen, C. R. Mirasso, and I. Fischer. Information processing using a single dynamical node as complex system. *Nat. Commun.*, 2:468, 2011. doi: 10.1038/ncomms1476.

L. Appeltant, G. Van der Sande, J. Danckaert, and I. Fischer. Constructing optimized binary masks for reservoir computing with delay systems. *Sci. Rep.*, 4, 2014. doi: 10.1038/srep03629.

M. Artin and A. A'Campo. *Algebra*. Birkhäuser advanced texts. Birkhäuser Basel, 1998. ISBN 9783764359386. URL http://books.google.de/books?id=3AKfV7g2CiAC.

A. J. Bell and T. J. Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural comput.*, 7(6):1129–1159, 1995. doi: 10.1162/neco.1995.7.6.1129.

J. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer Series in Statistics. Springer, 1985. ISBN 9783540960980.

E. L. Bienenstock, L. N. Cooper, and P. W. Munro. Theory for the development of neuron selectivity: orientation specificity and binocular interaction in visual cortex. *J. Neurosci.*, 2(1):32–48, 1982.

S. Billings. *Nonlinear System Identification: NARMAX Methods in the Time, Frequency, and Spatio-Temporal Domains*. Wiley, 2013. ISBN 9781118535554. URL http://books.google.de/books?id=SaQ2AAAAQBAJ.

C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006. ISBN 0387310738.

C. D. Boor. *A practical guide to splines*. Springer, New York, 2001.

G. Box, G. Jenkins, and G. Reinsel. *Time Series Analysis: Forecasting and Control*. Wiley Series in Probability and Statistics. Wiley, 2013. ISBN 9781118619063. URL http://books.google.de/books?id=jyrCqMBW_owC.

S. Boyd and L. O. Chua. Fading memory and the problem of approximating nonlinear operators with volterra series. *IEEE Trans. Circuits Syst.*, 32(11):1150–1161, 1985. doi: 10.1109/TCS.1985.1085649.

S. Boyd, L. Chua, and C. Desoer. Analytical foundations of Volterra series. *IMA Journal of Mathematical Control and Information*, 1(3):243, 1984. ISSN 0265-0754. doi: 10.1093/imamci/1.3.243. URL http://imamci.oxfordjournals.org/content/1/3/243.abstract.

H. Broer and F. Takens. *Dynamical Systems and Chaos*. Applied Mathematical Sciences. Springer, 2010. ISBN 9781441968708. URL http://books.google.de/books?id=yaov2qvj5YQC.

D. Brunner, M. C. Soriano, C. R. Mirasso, and I. Fischer. Parallel photonic information processing at gigabyte per second data rates using transient states. *Nature commun.*, 4:1364, 2013. doi: 10.1038/ncomms2368.

D. V. Buonomano and W. Maass. State-dependent computations: spatiotemporal processing in cortical networks. *Nat. Rev. Neurosci.*, 10(2):113–125, 2009. doi: 10.1038/nrn2558.

J. Y. Chen, K. W. Wong, H. Y. Zheng, and J. W. Shuai. Phase signal coupling induced $n : m$ phase synchronization in drive-response oscillators. *Phys. Rev. E*, 63:036214, Feb 2001. doi: 10.1103/PhysRevE.63.036214. URL http://link.aps.org/doi/10.1103/PhysRevE.63.036214.

S. Dasgupta, F. Wörgötter, and P. Manoonpong. Information dynamics based self-adaptive reservoir for delay temporal memory tasks. *Evolving Systems*, 4(4): 235–249, 2013. doi: 10.1007/s12530-013-9080-y.

G. W. Davis and C. S. Goodman. Synapse-specific control of synaptic efficacy at the terminals of a single neuron. *Nature*, 392(6671):82–86, 1998. doi: 10.1038/32176.

B. De Finetti. La prévision: ses lois logiques, ses sources subjectives. In *Annales de l'institut Henri Poincaré*, volume 7, pages 1–68. Presses universitaires de France, 1937.

E. R. Deyle and G. Sugihara. Generalized theorems for nonlinear state space reconstruction. *PLoS One*, 6(3):e18295, 2011.

R. J. Douglas and K. Martin. Neuronal circuits of the neocortex. *Annu. Rev. Neurosci.*, 27:419–451, 2004.

M. L. Eaton. Dutch book in simple multivariate normal prediction: another look. In *Probability and Statistics: Essays in Honor of David A. Freedman*, pages 12–23. Institute of Mathematical Statistics, 2008.

B. Efron. Second thoughts on the bootstrap. *Statistical Science*, 18(2):135–140, 2003.

B. Efron and R. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis, 1994. ISBN 9780412042317. URL http://books.google.de/books?id=gLlpIUxRntoC.

B. Efron and R. J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall, New York, 1993.

R. F. Engle. Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. *Econometrica: Journal of the Econometric Society*, pages 987–1007, 1982.

P. Fishburn. *Utility theory for decision making*. Publications in operations research. Wiley, 1970. URL http://books.google.de/books?id=lyUoAQAAMAAJ.

P. Fishburn. Subjective expected utility: A review of normative theories. *Theory and Decision*, 13(2):139–199, 1981.

P. Fishburn. The axioms of subjective probability. *Statistical Science*, pages 335–345, 1986.

R. Fisher, J. Bennett, and F. Yates. *Statistical Methods, Experimental Design, and Scientific Inference*. Oxford science publications. OUP Oxford, 1990. ISBN 9780198522294. URL http://books.google.de/books?id=e6XOngEACAAJ.

M. Franz and B. Schölkopf. A unifying view of wiener and volterra theory and polynomial kernel regression. *Neural Computation*, 18(12):3097–3118, 2006.

D. A. Freedman. Notes on the dutch book argument. *Lecture Notes, Department of Statistics, University of Berkley at Berkley*, 2003. URL http://www.stat.berkeley.edu/~{}census/dutchdef.pdf.

D. A. Freedman and R. A. Purves. Bayes' method for bookies. *The Annals of Mathematical Statistics*, pages 1177–1186, 1969.

J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2010. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2929880/.

P. Fries. Neuronal gamma-band synchronization as a fundamental process in cortical computation. *Annual review of neuroscience*, 32:209–224, 2009.

K. J. Friston. Brain function, nonlinear coupling, and neuronal transients. *The Neuroscientist*, 7(5):406–18, Oct. 2001. ISSN 1073-8584. URL http://www.ncbi.nlm.nih.gov/pubmed/11597100.

S. Ganguli, D. Huh, and H. Sompolinsky. Memory traces in dynamical systems. *Proceedings of the National Academy of Sciences*, 105(48):18970–18975, 2008. ISSN 1091-6490. doi: 10.1073/pnas.0804451105.

F. Gerhard, G. Pipa, B. Lima, S. Neuenschwander, and W. Gerstner. Extraction of network topology from multi-electrode recordings: Is there a small-world effect? *Frontiers in Computational Neuroscience*, 5 (00004), 2011. ISSN 1662-5188. doi: 10.3389/fncom.2011.00004. URL http://www.frontiersin.org/Journal/Abstract.aspx?s=237&name=computational_neuroscience&ART_DOI=10.3389/fncom.2011.00004.

A. Girard, C. E. Rasmussen, J. Quinonero-Candela, and R. Murray-Smith. Gaussian process priors with uncertain inputs? application to multiple-step ahead time series forecasting. 2003.

L. Glass and M. Mackey. Mackey-Glass equation. *Scholarpedia*, 5(3):6908, 2010. doi: 10.4249/scholarpedia.6908.

L. L. Gollo, O. Kinouchi, and M. Copelli. Active dendrites enhance neuronal dynamic range. *PLoS Comput. Biol.*, 5(6):e1000402, 2009. doi: 10.1371/journal.pcbi.1000402.

C. W. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society*, pages 424–438, 1969.

M. Graupner and N. Brunel. Calcium-based plasticity model explains sensitivity of synaptic changes to spike pattern, rate, and dendritic location. *Proc. Natl. Acad. Sci. U.S.A.*, 109(10):3991–3996, 2012. doi: 10.1073/pnas.1109359109.

S. Guo and J. Wu. *Bifurcation theory of functional differential equations*. Springer New York, 2013a. doi: 10.1007/978-1-4614-6992-6.

S. Guo and J. Wu. *Bifurcation Theory of Functional Differential Equations*. Applied Mathematical Sciences. Springer London, Limited, 2013b. ISBN 9781461469919. URL http://books.google.de/books?id=ZM2CmAEACAAJ.

K. D. Harris. Neural signatures of cell assembly organization. *Nature Reviews Neuroscience*, 6(5):399–407, 2005.

T. Hastie and R. Tibshirani. *Generalized Additive Models*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis, 1990. ISBN 9780412343902. URL http://books.google.de/books?id=qa29r1Ze1coC.

S. Häusler and W. Maass. A statistical analysis of information-processing properties of lamina-specific cortical microcircuit models. *Cereb. Cortex*, 17(1):149–162, 2007. doi: 10.1093/cercor/bhj132.

H. Heuser. *Lehrbuch der Analysis*. Number pt. 1 in Mathematische Leitfäden. Teubner Verlag, 2009. ISBN 9783834807779. URL http://books.google.de/books?id=CQ_wc67PkFQC.

T. Hida and M. Hitsuda. *Gaussian Processes*. Translations of Mathematical Monographs. American Mathematical Society, 2007. ISBN 9780821843581. URL http://books.google.de/books?id=eHC9_wbxSE0C.

M. Hirsch. *Differential Topology*. Graduate Texts in Mathematics. Springer New York, 2012. ISBN 9781468494518. URL http://books.google.de/books?id=W3ftnQEACAAJ.

A. E. Hoerl and R. W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970. doi: 10.1080/00401706.1970.10488634.

T. Hoinville, C. T. Siles, and P. Hénaff. Flexible and multistable pattern generation by evolving constrained plastic neurocontrollers. *Adapt. Behav.*, 19(3):187–207, 2011. doi: 10.1177/1059712311403631.

Huebner, Abraham, and Weiss. Dimensions and entropies of chaotic intensity pulsations in a single-mode far-infrared nh3 laser. *Phys Rev A*, 40(11):6354–6365, 1989. ISSN 1050-2947.

J. Huke. Embedding nonlinear dynamical systems: A guide to takens' theorem. 2006. URL http://eprints.ma.man.ac.uk/175/01/covered/MIMS_ep2006_26.pdf.

E. M. Izhikevich. Polychronization: computation with spikes. *Neural Comput.*, 18 (2):245–282, 2006. doi: 10.1162/089976606775093882.

H. Jaeger. The "echo state" approach to analysing and training recurrent neural networks. Techn. rep. gmd 148, Bremen: German National Research Center for Information Technology, 2001.

H. Jaeger and H. Haas. Harnessing nonlinearity: Predicting chaotic systems and saving energy in wireless communication. *Science*, 304(5667):78–80, 2004. doi: 10.1126/science.1091277.

H. Jaeger, M. Lukoševičius, D. Popovici, and U. Siewert. Optimization and applications of echo state networks with leaky-integrator neurons. *Neural Netw.*, 20 (3):335–352, 2007. doi: 10.1016/j.neunet.2007.04.016.

E. Jaynes. Prior probabilities. *Systems Science and Cybernetics, IEEE Transactions on*, 4(3):227–241, 1968.

E. Jaynes and G. Bretthorst. *Probability Theory: The Logic of Science*. Cambridge University Press, 2003. ISBN 9780521592710. URL http://books.google.de/books?id=tTN4HuUNXjgC.

R. Jeffrey. *Subjective Probability: The Real Thing*. Cambridge University Press, 2004. ISBN 9780521536684. URL http://books.google.de/books?id=0DPEOiagafMC.

H. Jeffreys. *The Theory of Probability*. OUP Oxford, 1998. ISBN 9780191589676. URL http://books.google.de/books?id=vh9Act9rtzQC.

J. Kadane. *Principles of Uncertainty*. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis, 2011. ISBN 9781439861615. URL http://books.google.de/books?id=uZ53AtZl-dAC.

H. Kantz and T. Schreiber. *Nonlinear Time Series Analysis*. Cambridge nonlinear science series. Cambridge University Press, 2004. ISBN 9780521529020. URL http://books.google.de/books?id=RfQjAG2pKMUC.

U. R. Karmarkar and D. V. Buonomano. Timing in the absence of clocks: encoding time in neural network states. *Neuron*, 53(3):427–438, 2007. doi: 10.1016/j.neuron.2007.01.006.

R. Kempter, W. Gerstner, and J. L. Van Hemmen. Intrinsic stabilization of output rates by spike-based hebbian learning. *Neural Computation*, 13(12):2709–2741, 2001.

A. Klenke. *Probability Theory: A Comprehensive Course*. Universitext. Springer, 2007. ISBN 9781848000483. URL http://books.google.de/books?id=tcm3y5UJxDsC.

L. Kocarev and U. Parlitz. Generalized synchronization, predictability, and equivalence of unidirectionally coupled dynamical systems. *Physical Review Letters*, 76(11):1816–1819, Mar. 1996. ISSN 0031-9007. URL http://www.ncbi.nlm.nih.gov/pubmed/10060528.

A. Kolmogoroff. *Grundbegriffe der Wahrscheinlichkeitsrechnung*. Ergebnisse der Mathematik und ihrer Grenzgebiete. Springer, 1973. URL http://books.google.de/books?id=7xg8nQEACAAJ.

S. Konishi and G. Kitagawa. *Information Criteria and Statistical Modeling*. Springer Series in Statistics. Springer, 2008. ISBN 9780387718866.

J. Kreiß and G. Neuhaus. *Einführung in Die Zeitreihenanalyse*. Statistik und ihre Anwendungen. Springer, 2006. ISBN 9783540335719. URL http://books.google.de/books?id=NGUkBAAAQBAJ.

T. Kreuz, F. Mormann, R. Andrzejak, A. Kraskov, K. Lehnertz, and P. Grassberger. Measuring synchronization in coupled model systems: A comparison of different approaches. *Physica D: Nonlinear Phenomena*, 225(1):29–42, Jan. 2007. ISSN 01672789. doi: 10.1016/j.physd.2006.09.039. URL http://linkinghub.elsevier.com/retrieve/pii/S0167278906003836.

L. Larger, M. Soriano, D. Brunner, L. Appeltant, J. M. Gutiérrez, L. Pesquera, C. R. Mirasso, and I. Fischer. Photonic information processing beyond turing: an optoelectronic implementation of reservoir computing. *Opt. Express*, 20(3):3241–3249, 2012a. doi: 10.1364/OE.20.003241.

L. Larger, M. C. Soriano, D. Brunner, L. Appeltant, J. M. Gutierrez, L. Pesquera, C. R. Mirasso, and I. Fischer. Photonic information processing beyond turing: an optoelectronic implementation of reservoir computing. *Opt. Express*, 20(3):3241–3249, Jan 2012b. doi: 10.1364/OE.20.003241. URL http://www.opticsexpress.org/abstract.cfm?URI=oe-20-3-3241.

D. Lavis and P. Milligan. The work of e.t. jaynes on probability, statistics and statistical physics. *The British Journal for the Philosophy of Science*, 36(2):193–210, 1985.

A. Lazar, G. Pipa, and J. Triesch. Fading memory and time series prediction in recurrent networks with different forms of plasticity. *Neural Netw.*, 20(3):312–322, 2007. doi: 10.1016/j.neunet.2007.04.020.

A. Lazar, G. Pipa, and J. Triesch. SORN: a self-organizing recurrent neural network. *Frontiers in computational neuroscience*, 3, 2009. ISSN 1662-5188. doi: 10.3389/neuro.10.023.2009. URL http://dx.doi.org/10.3389/neuro.10.023.2009.

D. Lindley. *Bayesian Statistics, A Review*. CBMS-NSF Regional Conference Series in Applied Mathematics. Society for Industrial and Applied Mathematics, 1972. ISBN 9780898710021. URL http://books.google.de/books?id=m0p32RNci_wC.

D. Lindley. The 1988 wald memorial lectures: The present position in bayesian statistics. *Statistical Science*, 5(1):44–65, 02 1990. doi: 10.1214/ss/1177012253.

M. Lukoševičius and H. Jaeger. Reservoir computing approaches to recurrent neural network training. *Computer Science Review*, 3(3):127–149, 2009. doi: 10.1016/j.cosrev.2009.03.005.

W. Maass, T. Natschläger, and H. Markram. Real-time computing without stable states: a new framework for neural computation based on perturbations. *Neural computation*, 14(11):2531–60, 2002. ISSN 0899-7667. doi: 10.1162/089976602760407955. URL http://www.ncbi.nlm.nih.gov/pubmed/12433288.

D. MacKay. *Information Theory, Inference and Learning Algorithms*. Cambridge University Press, 2003. ISBN 9780521642989. URL http://books.google.de/books?id=i0XJngEACAAJ.

D. Marković and C. Gros. Intrinsic adaptation in autonomous recurrent neural networks. *Neural Comput.*, 24(2):523–540, 2012. doi: 10.1162/NECO_a_00232.

N. Marwan, M. C. Romano, M. Thiel, and J. Kurths. Recurrence plots for the analysis of complex systems. *Physics Reports*, 438(5-6):237–329, Jan. 2007. ISSN 03701573. doi: 10.1016/j.physrep.2006.11.001. URL http://linkinghub.elsevier.com/retrieve/pii/S0370157306004066.

P. McCullagh and J. Nelder. *Generalized Linear Models*. Monographs on statistics and applied probability. Champman and Hall/CRC, 2000. URL http://books.google.de/books?id=sGmcmQEACAAJ.

A. Mees. *Nonlinear Dynamics and Statistics*. Birkhäuser Boston, 2001. ISBN 9780817641634. URL http://books.google.de/books?id=pH_OmkD4ZaQC.

K. D. Miller. Synaptic economics: competition and cooperation in synaptic plasticity. *Neuron*, 17(3):371–374, 1996. doi: 10.1016/S0896-6273(00)80169-5.

M. Muldoon, D. Broomhead, J. Huke, and R. Hegger. Delay embedding in the presence of dynamical noise. *Dynamics and Stability of Systems*, 13(2):175–186, 1998.

J. Naudé, B. Cessac, H. Berry, and B. Delord. Effects of cellular homeostatic intrinsic plasticity on dynamical and computational properties of biological recurrent neural networks. *J. Neurosci.*, 33(38):15032–15043, 2013. doi: 10.1523/JNEUROSCI.0870-13.2013.

K. Neusser. *Zeitreihenanalyse in Den Wirtschaftswissenschaften*. Studienbücher Wirtschaftsmathematik. Vieweg Verlag, Friedr, & Sohn Verlagsgesellschaft mbH, 2011. ISBN 9783834886538. URL http://books.google.de/books?id=lT4pBAAAQBAJ.

J. Neyman and E. S. Pearson. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231:289–337, 1933. ISSN 02643952. URL http://www.jstor.org/stable/91247.

D. Nikolić, S. Häusler, W. Singer, and W. Maass. Distributed fading memory for stimulus properties in the primary visual cortex. *PLoS Biol.*, 7(12):e1000260, 2009. doi: 10.1371/journal.pbio.1000260.

G. Nolte, O. Bai, L. Wheaton, Z. Mari, S. Vorbach, and M. Hallett. Identifying true brain interaction from eeg data using the imaginary part of coherency. *Clinical Neurophysiology*, 115(10):2292–2307, 2004.

S. Ørstavik and J. Stark. Reconstruction and cross-prediction in coupled map lattices using spatio-temporal embedding techniques. *Physics Letters A*, 247(1): 145–160, 1998.

G. Osipov, B. Hu, C. Zhou, M. Ivanchenko, and J. Kurths. Three Types of Transitions to Phase Synchronization in Coupled Chaotic Oscillators. *Physical Review Letters*, 91(2):1–4, July 2003. ISSN 0031-9007. doi: 10.1103/PhysRevLett.91.024101. URL http://link.aps.org/doi/10.1103/PhysRevLett.91.024101.

Y. Paquot, F. Duport, A. Smerieri, J. Dambre, B. Schrauwen, M. Haelterman, and S. Massar. Optoelectronic reservoir computing. *Sci. Rep.*, 2, 2012. doi: 10.1038/srep00287.

R. Pascanu and H. Jaeger. A neurodynamical model for working memory. *Neural Netw.*, 24(2):199–207, 2011. doi: 10.1016/j.neunet.2010.10.003.

F. Pasemann and T. Wennekers. Generalized and partial synchronization of coupled neural networks. *Network: Computation in Neural Systems*, 11(1):41–61, Feb. 2000. ISSN 0954-898X. URL http://www.ncbi.nlm.nih.gov/pubmed/10735528.

Y. Pawitan. *In All Likelihood: Statistical Modelling and Inference Using Likelihood*. OUP Oxford, 2013. ISBN 9780199671229. URL http://books.google.de/books?id=8T8fAQAAQBAJ.

A. Peters and B. Payne. Numerical relationships between geniculocortical afferents and pyramidal cell modules in cat primary visual cortex. *Cerebral Cortex*, 3(1): 69–78, 1993.

K. Pyragas. Weak and strong synchronization of chaos. *Physical Review E*, 54(5): R4508, 1996.

A. Quarteroni, R. Sacco, and F. Saleri. *Numerical Mathematics*. Texts in Applied Mathematics 37. Springer, Berlin, 2. edition, 2006. ISBN 978-3-540-43616-4. URL http://www.springer.com/math/cse/book/978-3-540-43616-4.

H. Raiffa and R. Schlaifer. *Applied Statistical Decision Theory [by] Howard Raiffa and Robert Schlaifer*. Studies in managerial economics. Division of Research, Graduate School of Business Administration, Harvard University, 1961. URL http://books.google.de/books?id=kyh7kgAACAAJ.

F. P. Ramsey. Truth and probability (1926). *The foundations of mathematics and other logical essays*, pages 156–198, 1931.

C. Rasmussen and C. Williams. *Gaussian Processes for Machine Learning*. Adaptative computation and machine learning series. University Press Group Limited, 2006. ISBN 9780262182539. URL http://books.google.de/books?id=vWtwQgAACAAJ.

M. Remme and W. J. Wadman. Homeostatic scaling of excitability in recurrent neural networks. *PLoS Comput. Biol.*, 8(5):e1002494, 2012. doi: 10.1371/journal.pcbi.1002494.

A. Renart, P. Song, and X.-J. Wang. Robust spatial working memory through homeostatic synaptic scaling in heterogeneous cortical networks. *Neuron*, 38(3):473–485, 2003. doi: 10.1016/S0896-6273(03)00255-1.

M. Richter. Rational choice and polynomial measurement models. *Journal of Mathematical Psychology*, 12(1):99–113, 1975.

S. Roman. *Advanced Linear Algebra*. Graduate Texts in Mathematics. Springer, 2007. ISBN 9780387728315. URL http://books.google.de/books?id=bSyQr-wUys8C.

M. C. Romano, M. Thiel, J. Kurths, I. Z. Kiss, and J. Hudson. Detection of synchronization for non-phase-coherent and non-stationary data. *EPL (Europhysics Letters)*, 71(3):466–472, 2005. URL http://www.iop.org/EJ/abstract/0295-5075/71/3/466.

M. Rosenblum, A. Pikovsky, and J. Kurths. From phase to lag synchronization in coupled chaotic oscillators. *Physical Review Letters*, 78(22):4193–4196, 1997. ISSN 1079-7114. URL http://link.aps.org/doi/10.1103/PhysRevLett.78.4193.

W. Rugh. *Nonlinear system theory: the Volterra/Wiener approach*. Johns Hopkins series in information sciences and systems. Johns Hopkins University Press, 1981. URL http://books.google.co.in/books?id=XvRQAAAAMAAJ.

N. F. Rulkov, M. M. Sushchik, L. S. Tsimring, and H. D. I. Abarbanel. Generalized synchronization of chaos in directionally coupled chaotic systems. *Phys. Rev. E*, 51:980–994, Feb 1995. doi: 10.1103/PhysRevE.51.980. URL http://link.aps.org/doi/10.1103/PhysRevE.51.980.

C. C. Rumsey, L. F. Abbott, et al. Synaptic democracy in active dendrites. *J. Neurophysiol.*, 96(5):2307–2318, 2006. doi: 10.1152/jn.00149.2006.

V. Sakkalis, C. D. Giurcaneanu, P. Xanthopoulos, M. E. Zervakis, V. Tsiaras, Y. Yang, E. Karakonstantaki, and S. Micheloyannis. Assessment of linear and nonlinear synchronization measures for analyzing eeg in a mild epileptic paradigm. *Trans. Info. Tech. Biomed.*, 13(4):433–441, July 2009. ISSN 1089-7771. doi: 10.1109/TITB.2008.923141. URL http://dx.doi.org/10.1109/TITB.2008.923141.

T. Sauer, J. Yorke, and M. Casdagli. Embedology. *Journal of statistical Physics*, 65(3-4):579–616, 1991.

L. Savage. *The Foundations of Statistics*. Dover Books on Mathematics. Dover Publications, 1954. ISBN 9780486137100. URL http://books.google.de/books?id=N_bBAgAAQBAJ.

L. Savage. *The foundations of statistics reconsidered*. University of Calif Press, 1961. URL http://cs.ru.nl/~peterl/teaching/CI/savage.pdf.

M. J. Schervish, T. Seidenfeld, and J. B. Kadane. On the equivalence of conglomerability and disintegrability for unbounded random variables. 2008.

B. Schrauwen, L. Buesing, and R. A. Legenstein. On computational power and the order-chaos phase transition in reservoir computing. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *NIPS*, pages 1425–1432. Curran Associates, Inc., 2008a. URL http://dblp.uni-trier.de/db/conf/nips/nips2008.html#SchrauwenBL08.

B. Schrauwen, M. Wardermann, D. Verstraeten, J. J. Steil, and D. Stroobandt. Improving reservoirs using intrinsic plasticity. *Neurocomputing*, 71(7):1159–1171, 2008b. doi: 10.1016/j.neucom.2007.12.020.

J. Schumacher, H. Toutounji, and G. Pipa. An analytical approach to single node delay-coupled reservoir computing. In P. Mladenov, V. Koprinkova-Hristova, G. Palm, A. E. P. Villa, B. Appollini, and N. Kasabov, editors, *Artificial Neural Networks and Machine Learning–ICANN 2013*, volume 8131 of *Lecture Notes in Computer Science*, pages 26–33. Springer Berlin Heidelberg, 2013. doi: 10.1007/978-3-642-40728-4_4.

D. V. Senthilkumar, M. Lakshmanan, and J. Kurths. Transition from phase to generalized synchronization in time-delay systems. *Chaos (Woodbury, N.Y.)*, 18(2):023118, June 2008. ISSN 1089-7682. doi: 10.1063/1.2911541. URL http://www.ncbi.nlm.nih.gov/pubmed/18601485.

A. Seth. Granger causality. *Scholarpedia*, 2(7):1667, 2007. revision #91329.

L. F. Shampine and S. Thompson. Solving ddes in matlab. In *Applied Numerical Mathematics*, volume 37, pages 441–458, 2001.

Z. Shi and J. Green. *The Recorded Sayings of Zen Master Joshu*. International Sacred Literature Trust Series. AltaMira Press, 1998. ISBN 9780761989851. URL http://books.google.de/books?id=T4sJ5fK6_vYC.

H. Smith. *An Introduction to Delay Differential Equations with Applications to the Life Sciences*. Texts in Applied Mathematics. Springer, 2010. ISBN 9781441976468. URL http://books.google.de/books?id=EonZt2KRhPMC.

D. C. Somers, S. B. Nelson, and M. Sur. An emergent model of orientation selectivity in cat visual cortical simple cells. *J. Neurosci.*, 15(8):5448–5465, 1995.

S. Song, K. D. Miller, and L. F. Abbott. Competitive hebbian learning through spike-timing-dependent synaptic plasticity. *Nat Neurosci.*, 3(9):919–926, 2000. doi: 10.1038/78829.

M. C. Soriano, S. Ortín, D. Brunner, L. Larger, C. R. Mirasso, I. Fischer, and L. Pesquera. Optoelectronic reservoir computing: tackling noise-induced performance degradation. *Optics express*, 21(1):12–20, Jan. 2013. ISSN 1094-4087. URL http://www.ncbi.nlm.nih.gov/pubmed/23388891.

C. Soto-Treviño, K. A. Thoroughman, E. Marder, and L. Abbott. Activity-dependent modification of inhibitory synapses in models of rhythmic neural networks. *Nat. Neurosci.*, 4(3):297–303, 2001. doi: 10.1038/85147.

C. Stam. Synchronization likelihood: An unbiased measure of generalized synchronization in multivariate data sets. *Physica D: Nonlinear Phenomena*, 163(3-4):236–251, Mar. 2002. ISSN 01672789. doi: 10.1016/S0167-2789(01)00386-4. URL http://linkinghub.elsevier.com/retrieve/pii/S0167278901003864.

J. Stark. Delay embeddings for forced systems. i. deterministic forcing. *Journal of Nonlinear Science*, 9(3):255–332, 1999.

J. Stark, D. S. Broomhead, M. Davies, and J. Huke. Delay embeddings for forced systems. ii. stochastic forcing. *Journal of Nonlinear Science*, 13(6):519–577, 2003.

I. Steinwart. On the influence of the kernel on the consistency of support vector machines. *The Journal of Machine Learning Research*, 2:67–93, 2002.

S. Stigler. *The History of Statistics: The Measurement of Uncertainty Before 1900*. Belknap Series. Belknap Press of Harvard University Press, 1986. ISBN 9780674403413. URL http://books.google.de/books?id=M7yvkERHIIMC.

M. Stone. Strong inconsistency from uniform priors. *Journal of the American Statistical Association*, 71(353):114–116, 1976.

G. Sugihara, R. May, H. Ye, C.-h. Hsieh, E. Deyle, M. Fogarty, and S. Munch. Detecting causality in complex ecosystems. *science*, 338(6106):496–500, 2012.

S. Sundararajan and S. S. Keerthi. Predictive approaches for choosing hyperparameters in gaussian processes. *Neural Computation*, 13(5):1103–1118, 2001.

F. Takens. Dynamical systems and turbulence, warwick 1980: Detecting strange attractors in turbulence. In *Lecture Notes in Mathematics*, volume 898/1981, pages 366–381. Springer, 1981. URL http://www.springerlink.com/index/b254x77553874745.pdf.

F. Takens. The reconstruction theorem for endomorphisms. *Bulletin of the Brazilian Mathematical Society*, 33(2):231–262, 2002.

H. Toutounji and F. Pasemann. Behavior control in the sensorimotor loop with short-term synaptic dynamics induced by self-regulating neurons. *Front. Neurorobot.*, 8:19, 2014. doi: 10.3389/fnbot.2014.00019.

H. Toutounji and G. Pipa. Spatiotemporal computations of an excitable and plastic brain: neuronal plasticity leads to noise-robust and noise-constructive computations. *PLoS Comput. Biol.*, 10(3):e1003512, 2014. doi: 10.1371/journal.pcbi.1003512.

H. Toutounji, J. Schumacher, and G. Pipa. Optimized Temporal Multiplexing for Reservoir Computing with a Single Delay-Coupled Node. In *The 2012 International Symposium on Nonlinear Theory and its Applications (NOLTA 2012)*, 2012.

J. Troutman. *Variational Calculus and Optimal Control: Optimization with Elementary Convexity*. Undergraduate Texts in Mathematics. Springer New York, 2012. ISBN 9781461268871. URL http://books.google.de/books?id=cU56kwEACAAJ.

G. G. Turrigiano and S. B. Nelson. Homeostatic plasticity in the developing nervous system. *Nat. Rev. Neurosci.*, 5(2):97–107, 2004. doi: 10.1038/nrn1327.

P. Uhlhaas, G. Pipa, B. Lima, L. Melloni, S. Neuenschwander, D. Nikolic, and W. Singer. Neural synchrony in cortical networks: history, concept and current status. *Frontiers in Integrative Neuroscience*, 3(00017), 2009. ISSN 1662-5145. doi: 10.3389/neuro.07.017.2009. URL http://www.frontiersin.org/Journal/Abstract.aspx?s=571&name=integrative_neuroscience&ART_DOI=10.3389/neuro.07.017.2009.

P. A. Vargas, R. C. Moioli, F. J. Von Zuben, and P. Husbands. Homeostasis and evolution together dealing with novelties and managing disruptions. *Int. J. Intelligent Computing and Cybernetics*, 2(3):435–454, 2009. doi: 10.1108/17563780910982680.

D. Verstraeten, B. Schrauwen, and D. Stroobandt. Reservoir-based techniques for speech recognition. In *International Joint Conference on Neural Networks*, pages 1050–1053. IEEE, 2006. doi: 10.1109/IJCNN.2006.246804.

R. Vicente, M. Wibral, M. Lindner, and G. Pipa. Transfer entropyâĂŤa model-free measure of effective connectivity for the neurosciences. *Journal of computational neuroscience*, 30(1):45–67, 2011.

V. Volterra. *The theory of permutable functions*. Vanuxem lectures. Princeton university press, 1915. URL http://books.google.de/books?id=eGltAAAAMAAJ.

C. von der Malsburg. Self-organization of orientation sensitive cells in the striate cortex. *Kybernetik*, 14(2):85–100, 1973.

J. von Neumann and O. Morgenstern. *Theory of Games and Economic Behavior (60th Anniversary Commemorative Edition)*. Princeton Classic Editions. Princeton University Press, 2007. ISBN 9781400829460. URL http://books.google.de/books?id=jCN5aNJ-n-0C.

A. Wald. Statistical decision functions which minimize the maximum risk. *Annals of Mathematics*, pages 265–280, 1945.

A. Wald. Statistical decision functions. *The Annals of Mathematical Statistics*, pages 165–205, 1949.

H. Wang and M. Brady. Real-time corner detection algorithm for motion estimation. *Image and Vision Computing*, 13(9):695–703, 1995.

A. Weigend and N. Gershenfeld, editors. *Time series prediction: forecasting the future and understanding the past*, 1993. SFI studies in the sciences of complexity, Addison-Wesley.

D. Werner. *Funktionalanalysis*. Springer-Lehrbuch. Springer, 2011. ISBN 9783642210174. URL http://books.google.de/books?id=jCAkBAAAQBAJ.

H. Whitney. Differentiable manifolds. *Annals of Mathematics*, pages 645–680, 1936.

M. Wibral, N. Pampu, V. Priesemann, F. Siebenhühner, H. Seiwert, M. Lindner, J. T. Lizier, and R. Vicente. Measuring information-transfer delays. *PloS one*, 8 (2):e55809, 2013.

R. Wilcox. *Introduction to Robust Estimation and Hypothesis Testing*. Statistical modeling and decision science. Academic Press, 2012. ISBN 9780123869838. URL http://books.google.de/books?id=zZ0snCw9aYMC.

H. Williams and J. Noble. Homeostatic plasticity improves signal propagation in continuous-time recurrent neural networks. *Biosystems*, 87(2):252–259, 2007. doi: 10.1016/j.biosystems.2006.09.020.

J. Williamson. Countable additivity and subjective probability. *The British Journal for the Philosophy of Science*, 50(3):401–416, 1999.

C. Witham, M. Wang, and S. Baker. Corticomuscular coherence between motor cortex, somatosensory areas and forearm muscles in the monkey. *Frontiers in systems neuroscience*, 4, 2010.

T. Wunderle, D. Eriksson, and K. E. Schmidt. Multiplicative mechanism of lateral interactions revealed by controlling interhemispheric input. *Cerebral Cortex*, 23 (4):900–912, 2013.

T. Yamazaki and S. Tanaka. The cerebellum as a liquid state machine. *Neural Netw.*, 20(3):290–297, 2007. doi: 10.1016/j.neunet.2007.04.004.

W. Zhang and D. J. Linden. The other side of the engram: experience-driven changes in neuronal intrinsic excitability. *Nat. Rev. Neurosci.*, 4(11):885–900, 2003. doi: 10.1038/nrn1248.

P. Zheng, C. Dimitrakakis, and J. Triesch. Network self-organization explains the statistics and dynamics of synaptic connection strengths in cortex. *PLoS Comput. Biol.*, 9(1):e1002848, 2013. doi: 10.1371/journal.pcbi.1002848.

COLOPHON

This document was typeset using the typographical look-and-feel `classicthesis` developed by André Miede. The style was inspired by Robert Bringhurst's seminal book on typography "*The Elements of Typographic Style*". `classicthesis` is available for both LaTeX and LyX:

http://code.google.com/p/classicthesis/

Happy users of `classicthesis` usually send a real postcard to the author, a collection of postcards received so far is featured here:

http://postcards.miede.de/

Final Version as of May 16, 2015 (`classicthesis` version 1.0).

DECLARATION

Erklärung über die Eigenständigkeit der erbrachten wissenschaftlichen Leistung

Ich erkläre hiermit, dass ich die vorliegende Arbeit ohne unzulässige Hilfe Dritter und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe. Die aus anderen Quellen direkt oder indirekt übernommenen Daten und Konzepte sind unter Angabe der Quelle gekennzeichnet.

Weitere Personen waren an der inhaltlichen materiellen Erstellung der vorliegenden Arbeit nicht beteiligt. Insbesondere habe ich hierfür nicht die entgeltliche Hilfe von Vermittlungs- bzw. Beratungsdiensten (Promotionsberater oder andere Personen) in Anspruch genommen. Niemand hat von mir unmittelbar oder mittelbar geldwerte Leistungen für Arbeiten erhalten, die im Zusammenhang mit dem Inhalt der vorgelegten Dissertation stehen.

Die Arbeit wurde bisher weder im In- noch im Ausland in gleicher oder ähnlicher Form einer anderen Prüfungssbehörde vorgelegt.

Osnabrück, September 2014

Johannes Schumacher, May 16, 2015