Understanding Transient Network Effects Triggered by Spontaneous Events during Sleep with Biophysical and Statistical Models

Dissertation

zur Erlangung des Grades eines
Doktors der Naturwissenschaften

der Mathematisch-Naturwissenschaftlichen Fakultät
und
der Medizinischen Fakultät
der Eberhard-Karls-Universität Tübingen

vorgelegt
von

Kaidi Shao
aus Henan, China

2021

Erklärung / Declaration: Ich erkläre, dass ich die zur Promotion eingereichte Arbeit mit dem Titel:

"Understanding Transient Network Effects Triggered by Spontaneous Events during Sleep with Biophysical and Statistical Models "

selbständig verfasst, nur die angegebenen Quellen und Hilfsmittel benutzt und wörtlich oder inhaltlich übernommene Stellen als solche gekennzeichnet habe. Ich versichere an Eides statt, dass diese Angaben wahr sind und dass ich nichts verschwiegen habe. Mir ist bekannt, dass die falsche Abgabe einer Versicherung an Eides statt mit Freiheitsstrafe bis zu drei Jahren oder mit Geldstrafe bestraft wird.

I hereby declare that I have produced the work entitled *"**Understanding Transient Network Effects Triggered by Spontaneous Events during Sleep with Biophysical and Statistical Models** "*, submitted for the award of a doctorate, on my own (without external help), have used only the sources and aids indicated and have marked passages included from other works, whether verbatim or in content, as such. I swear upon oath that these statements are true and that I have not concealed anything. I am aware that making a false declaration under oath is punishable by a term of imprisonment of up to three years or by a fine.

Tübingen, den ....

Datum / Date Unterschrift / Signature

Tag der mündlichen Prüfung:	December 15th, 2021
Dekan der Math.-Nat. Fakultät:	Prof. Dr. Thilo Stehle
Dekan der Medizinischen Fakultät:	Prof. Dr. Bernd Picher
1. Berichterstatter:	Prof. Dr. Nikos K. Logothetis
2. Berichterstatter:	Prof. Dr. Martin Giese
Prüfungskommission:	Prof. Dr. Nikos K. Logothetis Prof. Dr. Martin Giese Prof. Dr. Anna Levina Prof. Dr. Masataka Watanabe

CONTENTS

PREFACE

Sleep and sleep stages are well preserved across many species and are accompanied by the regular occurrence of several types of transient events, e.g., slow oscillations, spindles, sharp wave-ripples, theta oscillations and ponto-geniculo-occipital waves. Converging experimental evidence has shown that these transient events are key to the process of memory consolidation and homeostasis during sleep.

In my thesis, I would like to investigate the interplay between events and the underlying network mechanisms. In particular, I am interested in uncovering the mechanism that gives rise to the events and understanding what effects these events exert on the underlying network. Potentially, this involves the characterization of network properties in a state-dependent manner.

My work is focused on addressing these questions with modeling approaches. My projects cover the use of both biophysical models that have the benefits of the interpretability of the underlying network properties and statistical models that can be used to infer the properties of the system based on experimental data. While neither model is perfect, I will discuss at the end how we can have a hybrid model that captures the advantages of both.

SUMMARY

Sleep and its major functions require the precise coordination of transient mechanisms at multiple spatiotemporal scales. These phenomena are reflected in neural signals by the spontaneous occurrence of a variety of oscillatory patterns that we call *neural events*. At the systems level, the network dynamics determine how and when such events are generated. Conversely, the occurrence of events also influences the underlying network properties. We are interested in the interplay between the events and the underlying network mechanisms to address the potential functions of transient activity during sleep. In this thesis, we achieve this goal with both biophysical and statistical modelling approaches.

In Chapter 1, which is designed as a background introduction, we demonstrate the functional significance of spontaneous transient neural events during sleep in memory consolidation and homeostasis. We then provide a review for some mechanistic and functional properties of typical events that occur during different stages of sleep. Afterward, we provide a mind-map of the major contents of this thesis that guides the readers into the subsequent chapters and link them to the reviewed physiological facts.

an introductory chapter for background knowledge and overview

In Chapter 2, we present the project of biophysical modeling of one particular type of neural event, the Ponto-Geniculo-Occipital (PGO) waves, to understand how they influence cortical plasticity. Based on physiological evidence, the model we have built is an acetylcholine-modulated neural mass model of PGO wave propagation through pons, thalamus and cortex, reproducing a broad range of empirical electrophysiological characteristics and their modifications across sleep stages. Using a population model of Spike-Time-Dependent Plasticity (STDP), we show that PGO waves drive recurrent cortical circuits in different transient regimes depending on the sleep stage, leading respectively to the potentiation of cortico-cortical synapses during the pre-REM stage and their depression during REM sleep. Overall, our results provide a new view on how transient sleep events and their associated sleep stage may implement precise control of system-wide plastic changes.

a chapter for biophysical modelling of events and event-triggered plasticity

In Chapter 3, we start the line of statistical modelling of experimental event ensembles to reflect dynamical phenomena emerging in complex systems. We consider the problem of learning accurate models of the peri-event dynamics based only on data gathered by *detecting* these transient events, as a widely-used approach to analyze spontaneous brain activities. We show, however, the event detection procedure entails a selection bias that leads to misrepresentation of the system properties. We analyze the selection bias in the frameworks of dynamical systems and Structural Causal Models (SCMs), and develop the Debiased Snapshot (*DeSnap*) approach to de-bias the time-varying system properties estimated from such peri-event data. As results, we demonstrate the benefits of this de-biasing approach on toy examples and neural time series. In both cases, *DeSnap* reduces artifactual high-frequency peaks caused by the event detection procedure appearing in the spectrograms of the learned systems. Overall, these results suggest that peri-event analysis of spontaneous activities is prone to biases due to event selection, which can be detected and corrected by proper use of time-varying stochastic models.

a chapter for statistical modelling of state-dependent network dynamics based on event ensembles

a chapter for causal investigation between transient events

In Chapter 4, we focus on quantifying brain-wide network interactions based on simultaneous recordings of events in multiple structures. In this context, information-theoretic tools like Kullback-Leibler Divergence allow us to design interpretable measures of causal influence based on principles of causality studies. We review several causality measures based on these tools that are designed for stationary signals and extend them to time-varying versions relying on time-varying Vector Autoregressive (VAR) models. Using the formalism of SCM and their graphical representation, we investigate the theoretical and empirical properties of these time-varying causality measures when they are applied to peri-event data. After showing the limitations of Transfer Entropy and Dynamic Causal Strength defined in the literature, we introduce a novel measure, relative Dynamic Causal Strength, and provide theoretical and empirical support for its benefits. In combination with the *DeSnap* algorithm, these measures are applied to simulated and experimentally recorded neural time series, providing results in agreement with our current understanding of the underlying neural circuits.

an outlook describing the design of linking the results of Chapter 2 and Chapter 3 in a hybrid model

The last chapter is dedicated to an outlook, a potential way to combine the methodology of Chapter 2 and Chapter 3 in order to learn data-driven dynamics and validate our theoretical predictions of event-triggered cortical synaptic rescaling. The problem is that the electrophysiological activities critical to implement the STDP rule are inaccessible through experimental recordings. However, we can design a biophysically meaningful Recurrent Neural Network based on a simplified cortical model (as part of our PGO model). By training such a network with local field potentials recorded experimentally, we might be able to recover these plasticity-related activities as hidden layer responses of the network. Thus this hybrid model is promising to check whether the opposite plasticity effect triggered by two subtypes of PGO waves is consistent with our model prediction.

Part I

BACKGROUND

1 BACKGROUND

1.1 TRANSIENT EVENTS DURING SLEEP

1.1.1 *Sleep and sleep functions*

Sleep, a major component of the life of mammalian species, is essential to their survival. During this physiological process, although external sensory stimuli are shut down, the mammalian brain is able to implement the numerous key functions, which, based on current knwoledge, include energy restoration[12], immunological enhancement [228], homeostasis [56, 236] and cognitive functions like memory consolidation [122] and emotional regulation [238]. Among these functions, homeostasis and memory consolidation are two fundamental functional roles of sleep [122].

Two major functions of sleep

Both functions are represented in the plastic changes of synapses connecting neurons into functional circuits. For homeostasis, as proposed by the Synaptic Homeostasis Hypothesis (SHY), synapses strengthened due to learning experience during wakefulness are weakened during sleep [235, 236, 237]. This is hypothesized to reflect a global homeostatic regulation of synaptic connectivity to avoid unsuitable network behaviors (i.e., network instability). During memory consolidation, synaptic connections are selectively enhanced to encode novel memory traces into existing neural networks. These synaptic modifications can occur in both local circuits, and at the systems level, e.g., memory representations can be transferred from the hippocampus to the neocortex for more stable long-term storage.

Despite early debates, current evidence suggests that these two functions are not mutually exclusive [237, 122, 190, 189]. An overall synaptic downscaling at the population level can be accompanied by upscaling specific synapses hosting important memory traces. Both functions contribute to the stabilized long-term storage of memory engrams, which is critical for the everyday survival.

1.1.2 *Coordination of spontaneous transient mechanism contributes to synaptic rescaling*

Experimental evidence suggests that many brain functions rely on transient network mechanisms that manifest themselves in the multiplicity of neural events that can be observed in brain activity across multiple structures.

Here, transient mechanisms refer to properities of dynamical systems, while events are observations

Such phenomena may occur in response to stimuli, as has been observed for gamma oscillations [232, 74], and may play a role in the dynamic encoding of information. However, key phenomena can also occur spontaneously, as the variety of events occurring during both Rapid-Eye-Movement (REM) sleep and non-Rapid-Eye-Movement (NREM) sleep. Given as examples, the major functionally significant transient sleep events include slow oscillations (SOs), thalamic spindles, hippocampal sharp wave-ripples (SPW-Rs), hippocampal theta oscillations and Ponto-Geniculo-Occipital (PGO) waves [63, 122, 190]. Figure 1.1 illustrates the spread of these events across brain regions and sleep stages. Notably, apart from NREM and REM sleep, a third sleep stage, the pre-REM stage (defined as the transition between NREM and REM stages) is also critcal for the study of sleep events.

It has been hypothesized that the precise coordination of transient mechanisms at multiple spatiotemporal scales ensuring both the synergy between modules contributing to the same task and the non-interference between network activities in charge of different functions [192, 122, 189, 26].

These transient sleep phenomena have received great attention in the last decades, especially the ones occurring during NREM sleep [27, 222, 211, 190, 122, 236, 237, 241]. While local mechanisms of transient event generation are relatively clear nowadays, many aspects of events remain elusive.

two key questions to address in the thesis

Specifically, in this thesis, we attempt to address two key questions:

1. How are different transient events coordinated together?

   Accumulating evidence has shown that multiple types of events occurring in different brain regions appear synchronized with specific phase-locking relationships. At the same time, such co-occurrence is more beneficial to memory consolidation than isolated events (see Section 1.1.3.4, Section 1.1.4.3). It is curious to ask what are the mechanism underlying such coordination. At the systems level, this resorts to exploring the internal dynamics within the inter-regional brain network and the causal interactions that some drive the others. Understanding such mechanisms will help to elucidate how different brain regions activated during event co-occurrence communicate and cooperate to perform the same task of memory processing.

2. How does these transient events contribute to synaptic rescaling?

   Transient events, either detected in electroencephalogram (EEG) or local field potentials (LFPs), are accompanied by (and generated from) specific firing patterns in the hosting neurons. These patterns might exhibit spike-time relationships that favour specific plastic changes in the synapses. Given the consensus that memory promotion during sleep is implemented by the adjustment of synaptic connectivities over night, it is critical to form a mechanistic understanding of how synaptic rescaling is achieved by specific events, both in local neural circuits and distributed brain networks.

These two questions, although addressing different aspects of the system dynamics underlying event occurrence, are closely related (e.g., see Section 1.1.3.4). The coordinated interplay between different event-hosting regions might generate specific circuit dynamics forming the basis of event-triggered plasticity changes. Reversely, synaptic strengthing or weakening might re-organize the connectivity structure, which in turn alters the interactions between different neurons within the circuit.

logic design of next section

The following section will include an introduction to transient events related to the thesis. Considering the emerging functional significance of hippocampal-neocortical-thalamic system in memory consolidation, together with the specialized focus of our lab on PGO waves, the introduction will be confined to the five types of aforementioned transient events, as illustrated in Figure 1.1. With the immense literature exploring these events, it is impossible to provide an overview of all aspects of them. Therefore, the introduction will focus on the following aspects:

- Some electrophysiological characteristics
- The brain regions and species a certain event manifest itself

- Summary of generation mechanism
- Evidence supporting the role of a certain event in memory consolidation
- Evidence to answer the two key questions

The purpose of such a design is the following. Reviewing the first two aspects is to familiarize the readers with some basic facts related to each type of event. Mechanisms underlying event generation are briefly explained to prepare useful mechanistic knowledge for the readers to understand later discussions. Presenting the related evidence in memory consolidation justifies the significance of events we investigated in the thesis. Finally, we discuss the most important evidence that addresses the two key questions. Besides, the contributions of this thesis to each point, if any, will be briefly mentioned. An overview of the links between the results of the thesis and the background facts are established in Section 1.2.

Notably, reviews of events' functional roles often fall into piling up experimental facts without providing a logical chain. To avoid such confusion in the 4^{th} point, the following will present the evidence with the classical rationale to confirm that a particular event contributes causally to memory consolidation. As the validation of any variables in any system, the first step is to find the association (or linearly, correlation) between the occurrence events and certain task performance as a sign of memory consolidation. A causal role is further validated by manipulating the system under study and comparing the outcome before and after manipulation. In the case of transient events, such manipulations include disrupting and further enhancing the events by mediating the event generation circuits. This is possibly implemented by lesions, pharmocological interventions (e.g. injecting the agonists/antagonists that controls the generation of certain events), optogenetic control of the underlying circuits or different types of stimulations. Associations between the change of task performance and the decrease or increase of events after disruption or enhancement suggest a causal role of this event in memory consolidation.

For the last point, synchronized occurrence of events and event-based coordination of large-scale brain networks in Section 1.1.3.4, Section 1.1.4.3 and Section 1.1.4.4. Research status regarding the synaptic changes triggered by transient events will be introduced for each event and for each coupling of different events.

1.1.3 *NREM events*

Three major network rhythms occurring during NREM sleep are SOs, spindles, and SPW-Rs.

for abbreviations check Section 1.1.2

1.1.3.1 *Slow oscillations (SOs)*

One of the most prominent features of NREM sleep is the SOs (<1Hz), appearing as the alternation between synchronized network hyperpolarization (DOWN state) and synchronized depolarization (UP state) of neuronal populations. A transient sharp biphasic SO is also referred to as a K-complex (see Figure 1.1). SOs are generated within the cortex (layer 2/3 and 5) while propagating as travelling waves to a whole range of cortical and subcortical regions, including the hippocampus [152, 174, 245].

Both SOs and K-complexes will be reproduced in a neural mass model in Chapter 2

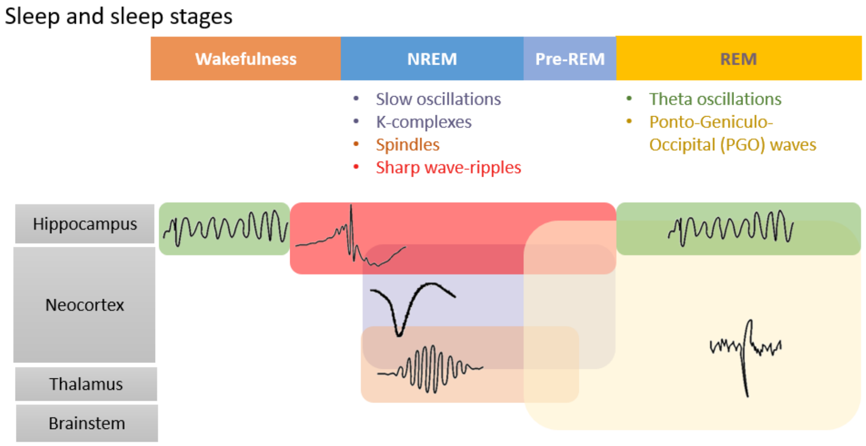

Figure 1.1: Overview of the manifestation of five typical transient neural events during wakefulness and sleep across brain regions. The color-coded half-transparent block, consistent with the corresponding colored names, mark the sleep stages and regions each type of event spans during the transition between wakefulness and sleep.

Both animal and human studies have provided evidence supporting the causal role of SOs in memory consolidation. Correlational correspondence between, on the one hand, the post-learning amount and intensity of SOs, and on the other hand, the performance measures (declarative memory retention/improvement on procedural tasks) has been reported in many species [78, 233, 165, 145]. Reduction of SOs by transcranial stimulation is reported to be correlated with impairment of memory retention after sleep [79]. Furthermore, enhanced SOs via non-invasive brain stimulation techniques contributes to improvement of memory consolidation performance [171, 172, 151, 150, 180, 131].

There is still an ongoing debate on whether SOs promotes consolidation via synaptic weakening, strengthening, or both [189]. SO-triggered depotentiation has been long proposed by SHY, which was confirmed by a series of experiments. For example, an optogenetically controlled Spike-Timing-dependent-Plasticity (STDP) experiment in anesthetized rodents suggests that the DOWN and UP states of SOs modulate the STDP rule implemented by the circuit [87]. Conventional STDP is discovered during the DOWN state. In contrast, another type of STDP rule found during the UP state is biased towards depression, suggesting a gating mechanism that favors overall depression during SOs.

However, other studies also suggest that SOs can induce synaptic strengthening. For example, SOs induced by *in vivo* pre-thalamic stimulation leads to Long-term Potentiation (LTP) in the somatosensory cortex [32]. Besides, SOs appearing together with spindles has been linked to LTP indirectly [173] (see also Section 1.1.3.2 and Section 1.1.3.4). As the experiments are all conducted for different brain regions in different conditions, it is still unclear under which circumstances either these two directions of synaptic rescaling occurs. Our modelling results presented in Chapter 2.3.5 support the role of SOs in synaptic upscaling when SOs (K-complexes) are induced by PGO waves or when they co-occur with spindles.

1.1.3.2 *Spindles*

Another pronounced rhythm characterizing NREM sleep is the spindles, which is prominent in the frequency band of 7-15Hz.

A typical spindle lasts 0.5-2 sec exhibiting a waxing and wining "spindle"-like waveform. Originated in the thalamus, the spindle oscillations are generated by the T-current modulated bursting activities in TRN neurons [220, 219]. TRN neurons pacemakes the T-current controlled bursting of thalamocortical neurons to form the spindle rhythm [221, 222]. The initialization and termination of the spindles are likely controlled by the cortex [35]. These mechanisms form the basis for our reproduction of thalamocortical spindles in a neural mass model in Chapter 2, and the interpretation of causal analysis in Section 4.3.3 and Section 4.3.4.2.

The spindle-generating mechanism will be elaborated in Section 2.2.3.1 and Section 4.3.3

Regarding the correlational role of spindle in memory consolidation, extensive studies in both humans and rats showed that spindle density during NREM sleep after learning was increased [164, 70, 161] and correlated with the subsequent task performance [40, 205, 77, 78, 234, 73, 194]. Furthermore, the causal role of spindles in memory consolidation has been established by manipulating spindle oscillations. For example, enhanced spindles via pharmocological interventions [160], optogenetic stimulation [128] or auditory closed-loop stimulation [171] are all associated with improved task performance reflecting the retention of memory contents.

Spindle-triggered plasticity in the cortex has also been addressed in many studies. One crucial early evidence is the study reported in [197] where stimulating cortical pyramidal neurons with experimentally recorded neuronal firing patterns *in vivo* is able to induce LTP *in vitro*. More recently, an *in vivo* imaging study revealed that spindles in rodents are accompanied by transiently enhanced dendritic calcium influx into the cortical pyramidal neurons [210]. As increased dendritic calcium activities have now been well-accepted as a pre-requisite for LTP induction, this result suggests that the LTP-triggered *in vitro* is likely to exist *in vivo* as well.

To elucidate the effects of potential spindle-triggered LTP in local circuits, another two-photon calcium imaging study shows that isolated spindles and spindle nested in the UP state of an SO modulates the firing rates of three types of cortical neurons in different manners [173]. Specifically, the co-occurrence of SOs and spindles leads to a three-times-higher increase of cortical pyramidal neurons than isolated spindles, while the discharge changes of two other inhibitory neurons remain unaffected by the timing of spindles. Such differentiated modulation of firing rates, which is supported by our modelling results in Section 2.3.5, suggests a functional role of spindle-triggered LTP in re-organizing the local circuits and altering the network dynamics.

Population firing rates can be an indirect sign of plasticity when the synaptic strength are not easily accessible

1.1.3.3 *Sharp Wave-Ripples*

The third type of transient oscillation dominating NREM sleep is the SPW-Rs, which also occur during quiet wakefulness (Figure 1.1).

Phenomenologically, SPW-Rs consists of two components - the sharp waves appearing as large low-frequency deflections in the hippocampal LFPs and high-frequency ripples (140-250 Hz) as the fast-field oscillations. Mechanistically, the SPW-Rs are primarily generated in the CA1 area of the hippocampus [41, 252, 193]. The somas of CA1 pyramidal cells are located in the pyramidal layer (*stratum pyramidale*), while their apical dendritic trees largely occupy the *stratum radiatum*. Driven by strong synchronous excitatory inputs from CA3 neurons, the dendritic trees of CA1 neurons generate post-synaptic activities, corresponding to LFP activities in the low frequencies (0-30 Hz, due to the sharp-wave) and in the gamma band (30-80 Hz) due to CA3 oscillations. Then the dendritic activities propagate to the soma, where recurrent interactions between inhibitory and excitatory cells generate a very fast oscillation, the ripple. These mechanisms forms the basis of causal analysis validation in Section 3.3.3.1.

SPW-Rs have been the focus of extensive work on sleep-dependent memory consolidation over the last two decades. This is mainly due to evidence that SPW-Rs are accompanied with sequential reactivation of neurons encoding memory traces learned during wakefulness, making it a strong candidate for strengthening existing memory traces and embedding new memory traces into the existing neuronal networks [112, 83, 129, 187]. Indeed, a large block of experimental evidence supports the role of SPW-Rs in memory consolidation: increase of SPW-Rs occurrence and frequency correlates with improvement of post-training learning performance both in rodents humans [71, 82, 2]. On the other hand, removing the SPW-R activities is sufficient to impair learning performance, suggesting that SPW-Rs are essential to consolidation processes [67, 81, 109].

Regarding event-triggered synaptic changes, SPW-Rs are assumed to promote hippocampal and cortical plasticity. The rationale behind this assumption is the following: during SPW-Rs, specific hippocampal ensembles are synchronously co-activated, making them likely to optimally impact STDP-based plasticity [245, 136]. Notably, high-frequency bursts occurring during SPW-Rs mimic tetanic stimulation protocols used to induce hippocampal LTP [10, 7, 25, 28].

1.1.3.4 *Coupling between NREM events*

These three NREM events are often found to co-occur with a precise temporal relationship between one another.

As mentioned previously in Section 1.1.3.2, the coupling between SO and spindles is able to trigger stronger enhancement of cortical firing rates of excitatory neurons. Actually, numerous studies have suggested that spindles nested in the UP state of an SO is the central mechanism underlying NREM-based consolidation (see [122] for a review). Correlation evidence includes the study where transcranial stimulation of SO activity in patients with mild cognitive impairment improved both SO-spindle coupling and memory performance after sleep [126]. Causal evidence can be provided by the results in [128] where spindles induced by optogenetic thalamic stimulation enhanced context-conditioned fear memory only when the spindles were induced in phase-lock with spontaneously occurring SO UP state.

Synchronized occurrence of SOs and SPW-Rs are also proposed to be essential for consolidating memory information. For example, [147] shows task performance for recalling is higher if auditorily stimulated SOs are synchronized with SPW-Rs compared to unsynchronized cases. A theoretical framework of synaptic plastic pressure has also been proposed a specific role of the DOWN→UP in [136]: the isolated DOWN→UP transition rescales cortical synapses based on their intrinsic firing rates while promotes the encoding of new memory traces when SPW-Rs co-occur with the DOWN→UP transition.

Recent reviews (e.g.[122, 181]) have also addressed the functional significance of a triplet coupling between SOs, spindles, and SPW-Rs, appearing in both human [36, 218] and rodents [162, 128, 147]. The triplet coupling exhibits a precise temporal relationship where spindles are nested in the UP state of SOs while SPW-Rs, together with the accompanying hippocampal neuronal reactivation, nest in the excitable troughs of spindles [215]. It has been proposed that the precise coordination between three events implement an inter-regional loop within the memory system with both top-down and bottom-up control [122]. In the top-down direction, the DOWN state of cortical SOs provides a general timing window for information transmission and plasticity by suppressing thalamic spindles, hippocampal SPW-Rs, and associated replay. The following UP state then drives the generation of thalamic spindles, which, in turn, act on hippocampal networks to synchronize ripples and ensemble reactivations to their excitable troughs. In other words, spindles control the timing of hippocampal reactivation. In the opposite bottom-up direction, simultaneously, spindles spread to the cortex and reach target networks still during the excitable SO up-state, a condition facilitating synaptic consolidation processes in these networks.

recalling the existence of conventional STDP during DOWN states, as introduced in Section 1.1.3.1

1.1.4 *REM events*

The REM stage is characterized by the occurence of theta oscillations and PGO waves.

1.1.4.1 *Theta oscillations*

Theta rhythms, which oscillate in the band 4-12 Hz, are the most prominent feature of REM sleep. Theta activities appear primarily in the hippocampus and other brain regions, especially those related to emotional memory, e.g., the amygdala, anterior cingulate cortex, and dorsolateral prefrontal cortex. The generation of hippocampal theta oscillations is driven by medial septal input [21, 256, 183], which are later mediated by parvalbumin-expressing (PV+) fast-spiking interneurons within the hippocampus [179, 177].

A growing body of evidence supports the role of theta waves in memory consolidation in humans and animals [175, 188, 179, 177]. An early study found that increases of theta activities immediately following training on a spatial visual discrimination task improved rats' task performance [243]. In rodents, cued fear learning in rats was shown to increase theta coherence between hippocampus and amygdala during subsequent REM sleep, and this increase in coherence predicted the success of associative memory consolidation [188]. Causally, induction of hippocampal theta activities by optogenetic stimulation of PV+ interneurons compensated for the adverse effects of sleep deprivation on the consolidation of fear memories in mice [179, 177].

At the cellular level, theta oscillations are known to support replay of hippocampal place cell sequences, following their sequential activation during experience [142]. Local network effects triggered by theta oscillations has been revealed under the paradigm of contextual fear memory learning, where the consolidation is predicted by the long-term stabilization of spike-timing relationships [178] that favors STDP induction [195]. Electrophysiological data in these experiments shows that the resonance of CA1 pyramidal neurons, in response to the rhythmic activities in local PV+ fast spiking neurons, is associated with the stabilization of spike-timing relationships [179]. The causal role of the resonance in stabilization is validated via manipulating the theta oscillations: pharmocogenetic or optogenetic disruptions of the theta oscillations destabilized CA1 spike-timing relationship and impairs the consolidation of contextual fear memory [179], while optogenetic enhancement of theta oscillations rescues the impairment [177]. Together, available evidence suggests that the highly regular theta-frequency activity that paces hippocampal neurons' firing during REM drives network plasticity and plays a critical role for hippocampally mediated memory consolidation.

1.1.4.2 *Ponto-Geniculo-Occipital Waves*

As another feature of REM sleep, a typical PGO wave often displays biphasic waveforms in LFP activity traces, comprising of a fast-negative component preceding a slower, weaker positive component, possibly followed by more minor fluctuations [88, 30, 43]. Despite the three critical structures - the Pons, the Lateral Geniculate Body of the thalamus, and the Occipital cortex - where PGO waves are most frequently observed and named after [113], PGO waves also manifest themselves in a broad range of brain structures in a variety of species, e.g., cats [113, 97, 43], rodents [120, 55, 58, 57], non-human primates [37, 192] and humans [139, 72]. Interestingly, during pre-REM and REM stages PGO waves exhibit two distinct subtypes [18]: during pre-REM sleep they appear more in high-amplitude singlets [20] whereas REM PGO waves tend to cluster in 3-5 successive weaker deflections [24, 20, 166, 47].

Two subtypes of PGO waves

Electrophysiological evidence has shown that these two subtypes of PGO waves are generated by the same cellular mechanism, which will be extensively elaborated in Chapter 2. Here the mechanisms will be briefly summarized. Triggered by the bursting of a small group of neurons in the pons, the PGO-related activities propagate to the thalamocortical networks and generate the region-specific waveforms with the interactions between several types of neuronal populations related to visual information processing. The differentiation between the two subtypes is implemented by different types of bursting activities (single/clustered) and the cholinergic modulation of cellular activities.

A series of studies from Subimal Datta has demonstrated that PGO waves are potentially involved in the consolidation of emotional memory [45]. In rats, an increase in P-wave density, either occurring spontaneously after learning or induced by injecting an acetylcholine agonist Carbachol, is shown to correlate with effective consolidation performance during post-training REM sleep [44, 51, 153, 53]. As for causal interventions, P-waves induced by injecting the same agonist is able to prevent the learning impairment caused by REM sleep deprivation [52].

In rodents, PGO waves are called P-waves, which propagate to the hippocampus, entorhinal cortex, and amygdala instead of the thalamus and neocortex

Follow-up research in rats by the same group has shown that training-activated P-waves trigger a cascade of plasticity-related gene expressions and protein synthesis in the dorsal hippocampus and amygdala [50, 54]. This causal relationship between P-wave and the genetic process has been validated by disruption experiments [50], suggesting the existence of P-wave-triggered plasticity process in the hippocampus, although the direction of synaptic rescaling is upon further exploration. However, local synaptic effects triggered by PGO waves remain relatively unexplored and thus speculative in other brain regions or species, e.g., cats and monkeys. With a PGO neural mass model based on feline electrophysiological data, we will show in Section 2.3.5 that the two subtypes of PGO waves are able to induce opposite plastic effects in the cortical circuits during different sleep stages.

1.1.4.3 *Coupling between PGO waves and hippocampal events*

PGO-theta coupling

As theta oscillations and REM PGO waves both occur during REM sleep, it is natural to speculate their co-occurrence, which has been indeed supported by early studies [204]. Several studies have demonstrated a correlational relationship between *sustained* theta activities and PGO waves in cats and rodents, either through the phase-locking between extracellular potentials [134, 118, 117] or through the positive correlation between theta-triggered firing rate and PGO densities [119, 169]. A subsequent study revealed that theta oscillations accelerate shortly before the negative peak of PGO waves, suggesting directional interactions between these two events [118]. Outside the pontine-hippocampus system, the intensity of P-waves is also reported to modulate the synchronization between theta oscillations in the hippocampus and amygdala [116].

putative PGO-ripple coupling

In addition to PGO-theta coupling during REM sleep, a recent study in our lab has provided evidence for the putative co-occurrence between pre-REM PGO waves and the other hippocampal events, the SPW-Rs [192]. This study is based on electrophysiological recordings in anesthetized monkeys, exhibiting the alternation between two states resembling the NREM and REM stages, which are accompanied by two subtypes of PGO waves that are similar to pre-REM and REM PGO waves. Interestingly, during the NREM-like state, high-frequency ripple-band and high-gamma-band activities are found to be coupled to pre-REM-like PGOs while low-frequency *transient* theta activities are reported to synchronize with REM-like PGOs. Our data analysis results in Section 3.3.3.2 also confirms that ripple-band and high-gamma band events are associated to one state while theta events to another state. As the alternating states during anesthesia are hypothesized and justified to be an induced NREM-REM transition, this study suggests that it is likely that during natural sleep, pre-REM PGO waves are also coupled to SPW-Rs.

1.1.4.4 *Coupling between PGO waves and thalamocortical events*

As pre-REM PGO waves are present before the beginning of REM sleep, they are mechanistically possible to interact with the other thalamocortical events, i.e., SOs and spindles.

For spindles, both electrophysiological recordings and our modelling work (Section 2.3.1) have shown that pre-REM PGO waves interrupt the occurrence of spindles as a potential mechanism of sleep stage transitions.

There is no clear evidence suggesting a coupling between SOs and pre-REM PGO waves in the literature. However, our modelling work will show that pre-REM PGO waves induce more K-complexes (Section 2.3.1), possibly contributing to the large-amplitude oscillations of pre-REM stage. As PGO waves are found to co-occur with SPW-Rs, the PGO-triggered K-complexes might explain the functional fole of PGO-ripple coupling: the SPW-Rs that occurs following PGO waves are nested in the DOWN→UP transitions of K-complexes such that the SPW-R-associated reactivation can be better consolidated (see also Section 1.1.3.4).

1.2 OVERVIEW OF THE THESIS

This section provides an overview (Figure 1.2) of all the projects to form a mind-map that guides the readers to grasp the main idea quickly. Specifically, it will summarize the links between the reviewed events and the

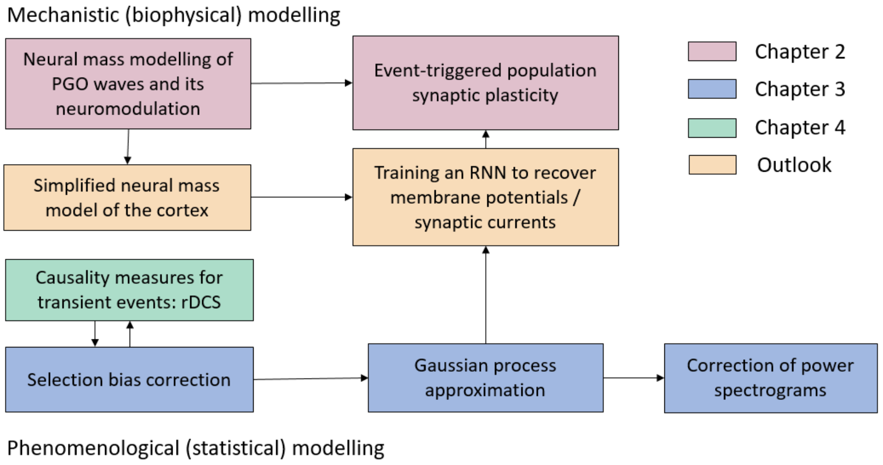
Mechanistic (biophysical) modelling
Neural mass modelling of PGO waves and its neuromodulation
Event-triggered population synaptic plasticity
Simplified neural mass model of the cortex
Training an RNN to recover membrane potentials / synaptic currents
Causality measures for transient events: rDCS
Selection bias correction
Gaussian process approximation
Correction of power spectrograms
Phenomenological (statistical) modelling
Chapter 2
Chapter 3
Chapter 4
Outlook

Figure 1.2: Overview of thesis structure as a mind-map. Colors encode contents for different chapters.

projects such that the readers will have a clearer big picture of how we addressed the two key questions raised in Section 1.1.2.

The thesis starts from two distinct lines of modelling. In one line, from the line of biophysical modeling, we start with a neural mass model of PGO waves (red boxes), which will be presented in Chapter 2. This is an attempt to attack the second question of local synaptic rescaling effects induced by transient events. With the incorporation of several critical neuronal populations, their intrinsic oscillation mechanisms, and neuromodulation mechanism, we successfully reproduced both subtypes of PGO waves in different stages of sleep, as well as cortical SOs (including K-complexes) and thalamocortical spindles during the NREM stage.

This model is an implementation of the generation mechanisms introduced in Section 1.1.4.2 for PGO waves, Section 1.1.3.1 for SOs and Section 1.1.3.2 for spindles. Network dynamics of the model match the PGO-triggered spindle interruptions presented in Section 1.1.4.4. The model also predicts that pre-REM PGO waves induce K-complexes accompanied by a DOWN→UP state transition (Section 1.1.3.4, Section 1.1.4.4), which might explain why pre-REM PGO waves co-occur with SPW-Rs (Section 1.1.4.3).

Combining this model with a population STDP rule, we are able to characterize the effects of cortical plasticity triggered by PGO waves and spindles in the population model. Specifically, the spindle-triggered synaptic changes differentiate with two subtypes of spindles, matching the contrast between the isolated spindles and spindles nested in SO UP states introduced in Section 1.1.3.2. For PGO waves, we found that two PGO subtypes trigger opposite synaptic rescaling effects, which fills the gap of plasticity study related to PGO waves at the circuit level (Section 1.1.4.2). The results can be compared with the rodent experiments introduced in Section 1.1.4.2.

To approach the first question proposed in Section 1.1.2, we start another line of research: statistical modelling of the network mechanism underlying the transient events. Transient mechanisms can be understood as a specific state of network coordination of brain regions, where studying it with observed events might be biased due to the selection procedure of peri-event snapshots. We developed an algorithm - *DeSnap* - to correct for the selection bias on peri-event snapshots, with which we are able to obtain a Gaussian process approximation as a statistical model of the network dynamics. The effectiveness of the bias correction method is validated with simulated toy models. This provides a general methodological framework for appropriate analysis of peri-event data to uncover the network dynamics hosting coordinated event interactions.

By applying the correction algorithm to correcting power spectrograms of hippocampal SPW-R data, we validated in real data that the algorithm can help us recover network dynamics representing specific states underlying the events. The correspondence between events and states is consistent with the PGO-ripple and PGO-theta coupling introduced in Section 1.1.4.3.

To further address the problem of causal interactions between events (as part of the first question), we resort to causality studies of peri-event time series. After comparing several existing causality measures, we proposed a novel measure to characterize causal interactions between transient events. The benefit of the new measure over previous ones is demonstrated with simulation models.

We applied these measures to characterize the interplay between simulated oscillatory events, simulated spindles, and intra-hippocampal recordings and found that the selection bias correction improves the performance of the causality measures. Specifically, the time-varying interactions between spindles and SOs recovered by the combination of bias correction and causality measures is in line with the mechanisms introduced in Section 1.1.3.2 and Section 1.1.3.4. The intra-hippocampal connectivities are in line with the SPW-R generation mechanisms explained in Section 1.1.3.3.

Finally, as a speculative outlook, we propose a way to study the state-dependent event-triggered plasticity with a hybrid model. We simplified the cortical neural mass model and designed a hybrid model that integrates the biophysical mechanisms and the statistical model we corrected so that we can recover the hidden states critical for the plasticity analysis. This is promising to validate our theoretical prediction of PGO-triggered plasticity with experimental data and potentially provides a general framework for studying event-triggered plasticity problem raised in the first question.

Part II

MAIN RESULTS

2 NEURAL MASS MODELLING OF PONTO-GENICULO-OCCIPITAL WAVES

2.1 INTRODUCTION

As briefly introduced in Section 1.1.1, in most mammalian species, the brain undergoes drastic state changes between different sleep stages – REM and NREM sleep – with broad behavioral and functional similarities. There is moreover an emerging consensus on the existence of sleep-induced plastic changes in relation to two hypothesized functions: memory consolidation, and synaptic homeostasis [26, 235, 236]. Considerable evidence supports the hypothesis that such plastic changes in the cortical connections are bidirectional, requiring both LTP to consolidate newly acquired memories, and long-term depression (LTD) to counterbalance increases in synaptic strengths and thereby maintain network stability [91, 241, 136].

The key problem of this chapter is synaptic rescaling

However, the detailed underlying mechanisms remain elusive and their investigation requires assessing the brain-wide impacts on the plasticity of a variety of phenomena happening during sleep. So far, an extensive amount of experimental and computational modelling studies have focused on the neural bases of transient events observed in NREM sleep (e.g., on the hippocampal SPW-Rs, cortical SOs, and thalamic spindles as elaborated in Section 1.1.3). In particular, it has been suggested that these NREM events play a role in long term plastic changes necessary to memory consolidation [222, 211, 190, 122] and synaptic downscaling [236, 237, 241]. However, evidence suggests that key modifications of plasticity not only occur during NREM but also during REM sleep [91, 53, 159, 21, 22, 253].

Interestingly, the transitional stage from NREM sleep to REM sleep (referred to as pre-REM stage) and the subsequent REM sleep stage are associated with the occurrence of another family of phasic events, the PGO waves [224, 88]. As introduced in Section 1.1.4.2, PGO waves manifest themselves in a broad range of brain structures and a variety of species. An important feature of PGO waves is that they exhibit two distinct subtypes during pre-REM sleep and REM sleep.

Given that converging evidence supports different roles played by NREM and REM sleep stages in reorganizing networks across the brain [91, 210], the fact that PGO waves span both sleep stages in the form of two subtypes suggests that these events play a key role in coordinating plastic changes, and their analysis may provide insights into the differences between plasticity promoting mechanisms during NREM and REM sleep. Indeed, experimental evidence reviewed in Section 1.1.4.2 supports a key role of REM-PGO waves in enhancing sleep-dependent learning and memory [44, 51, 153, 52]. In contrast, little is known about the impact of pre-REM PGO waves on plasticity. However, recent electrophysiological evidence for a coupling between hippocampal SPW-Rs and PGO waves provided in Section 1.1.4.3 suggests PGO waves are involved in memory consolidation processes happening during NREM sleep [192], possibly contributing to cortical synaptic rescaling.

In-vivo investigations of PGO-triggered plasticity changes remain challenging because: 1) unlike during repetitive stimulation protocols commonly

difficulties of experimental research

applied in *in vivo* LTP studies [19], where the stimulation amplitude can be manipulated, plastic effects induces by spontaneous activities with uncontrolled strength may be too weak to be observed; 2) With imaging techniques combined with electrophysiological recordings, due to spatial sparsity [138, 251, 253], it is difficult to locate the specific neurons and spines that receive PGO-associated potentials. As an alternative, investigating PGO waves from the computational modelling perspective may provide insights and guide further experimental studies. In particular, building a model of large-scale PGO waves propagation may help elucidate how the phasic potentials originated in the pontine region influence widespread brain regions during sleep and possibly control their plasticity.

In this study, we use a multi-structure neural mass model to reproduce prominent features of PGO waves at a system level and shed light on their possible functions (Section 2.2.1). By including cholinergic neuromodulation in our model (Section 2.2.4), we emphasize reproducing PGO-related phenomena across sleep stages (Section 2.3.1, Section 2.3.2 and Section 2.3.3)and account for the variability of PGO wave subtypes occurring either during pre-REM or REM sleep (Section 2.3.4). Finally, we investigate the putative influence of PGO waves on cortical plasticity through a mesoscopic model of STDP, suggesting that PGO waves may trigger opposite effects in REM and pre-REM sleep (Section 2.3.5). We provide insights on such state-dependent differences are achieved, suggesting a general framework to predict the influence of phasic events on plasticity.

2.2 METHODS

2.2.1 *Overview: Ponto-geniculo-occiptial neural mass model*

We designed a neural mass model to simulate average rate-coded population activities of several groups of homogeneous neurons [247, 248, 110, 141] in three brain structures influenced by PGO waves: the pons, the thalamus (more precisely the lateral geniculate nucleus (LGN), the thalamic reticular nucleus (TRN), and the primary visual cortex. This Methods section will explain the assumptions and mathematical tools we used to build the model.

model type: neural mass model

Section 2.2.2 is designed to be a systematic introduction of the modelling methodology of neural mass models, which represents the state of a population of identical neurons by the average firing rates across this population. The basic elements of such a model are illustrated in Figure 2.1A. Briefly, the population average membrane potential evolves according to its intrinsic dynamics as well as post-synaptic currents it receives from connected populations, and outputs the population firing rate as a non-linear instantaneous function of the membrane potential using a sigmoidal activation. After multiplication with a synaptic strength coefficient, the output firing rate is convolved with the impulse response of the synapses that link the population to downstream neurons. Although neural mass models constitute a strong simplification with respect to single-neuron models, recent work has shown they can reproduce sleep-related phasic patterns, such as spindles and SOs, with a satisfactory degree of realism, by taking into account key intrinsic currents flowing through the cells' membrane [242, 206]. We follow this approach to model the activity of three key structures involved in PGO-wave propagation.

model structure

More specifically, our model involves 6 neuronal populations, as represented in Figure 2.1B. The pontine population, representing low-frequency

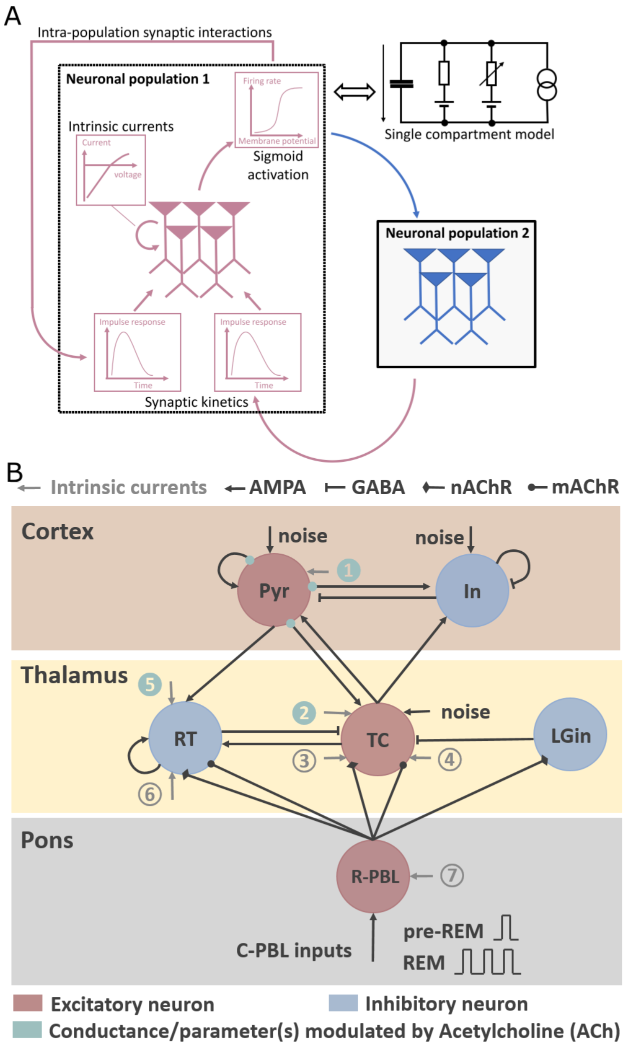

Figure 2.1: Neural mass model of PGO waves. (A)Illustration of the working mechanism of neural mass models. Two neuronal populations modelled by neural masses are illustrated in separate colors, where the dynamic of each depends on the intrinsic currents and synaptic currents it receives. Population dynamics, represented by the population membrane potential, is modelled to trigger firing rate via a sigmoid activation curve. These four elements (marked in the dashed rectangle) describes the activities of a single population, which is mathematically equivalent to a single compartment model. (B) Global view of the model structure. The neural mass model, consisting of biologically plausible neuronal populations and interconnections, receives brief pulses as model inputs to generate PGO-related neuronal activities. The switch of sleep stages is modulated by 4 major parameters (marked in green) associated with the change of Acetylcholine concentration. The numbered conductances in each population, as well as its function in state establishment (see Section 2.3.1) and PGO wave generaton, are listed in Table. 1. Abbreviation for neuronal populations: Pyr: Pyramidal neurons; In: inhibitory neurons; TC: thalamocortical neurons; RT: reticular thalamic neurons; LGin: interneurons in LGN; R-PBL: neurons in the rostal peribrachial nucleus (PGO-transferring neurons); C-PBL: neurons in the caudolateral peribrachial nucleus (PGO-triggering neurons). Abbreviation for receptors: AMPA: α-amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid; GABA: gamma-Aminobutyric acid; nAChR: nicotinic acetylcholine receptor; mAChR: muscarinic acetylcholine receptor.

No.	Population	Intrinsic Currents	Role of intrinsic current
①	Pyr	Na^{+}-dependent K^{+} current I_{KNa}	Induction of K-complexes
②	TC	K^{+} leaky current I^{t}_{LK}	Modulation of membrane potential
③	TC	Ca^{2+} T-current I^{t}_{T}	Spindle generation
④	TC	Inward rectifier K^{+} h-current	Wax and wining of spindle oscillation
⑤	RT	K^{+} leaky current I^{r}_{LK}	Modulation of membrane potential
⑥	RT	Ca^{2+} T-current I^{r}_{T}	Spindle generation
⑦	R-PBL	hyerpolarizating current	Generation of PGO-related bursts

Table 1: List of intrinsic currents and their corresponding roles in each neuronal population

bursting PGO transferring neurons located in the rostral peribrachial area (R-PBL), receives high-frequency bursts from PGO triggering neurons located in the caudolateral peribrachial area (C-PBL) [43, 48, 42, 49]. The thalamus comprises 3 populations that are hypothesized to underlie PGO wave generation: the thalamocortical relay (TC) neurons in the LGN, the reticular thalamic (RT) neurons in the peri-geniculate nucleus (PGN) of TRN [221, 156] and the local intra-geniculate interneurons (LGin) in the LGN [103, 104]. In the cortex, following [206], we model one excitatory population of pyramidal cells (Pyr), and one inhibitory interneuron population (In). The rationale underlying the selection of these neuronal populations, as well as their connectivities, is elaborated in Section 2.2.3.

design of the Method section

Section 2.2.3 also includes other assumptions important to the model construction. Section 2.2.3.3 validates the model input, assumed as bursting activities of C-PBL neurons triggering all the PGO-related activities in the subsequent neuronal populations. Section 2.2.3.1 describes the reasons for incorporating specific key intrinsic currents in each population to model sleep oscillations related to the process of memory consolidation, e.g. spindles and SOs (see Table 1 for a summary).

Section 2.2.4 describes how we implement cholinergic neuromodulation in several neuronal populations that enables the PGO network to switch between different sleep stages. This is the key mechanism in reproducing the stage-dependent characteristic of PGO waves.

Section 2.2.5 presents our work on model validation. The parameter tuning process of the pontine bursts to match experimental data is described in Section 2.2.5.1. Section 2.2.5.2 elaborates on how we validate that the 5 assumed ponto-thalamic connections are all necessary for the reproduction of transient PGO-triggered electrophysiological activities.

Finally, following a mesoscopic approximation of the STDP rule [76, 196], knowing the pre-synaptic current and the post-synaptic firing rate also allows us to make theoretical predictions about the plastic changes at a given type of synapse. This integration of plasticity rule with the neural mass model will be shown in Section 2.2.6.

2.2.2 *Mechanisms of neural mass models*

Neural mass models, with a compromise between model complexity and biophysical mechanisms, reflect average rate-coded population activities of a group of homogeneous neurons [110, 141], where the mechanisms have been briefly discussed in Section 2.2.1 and illustrated in Figure 2.1A.

2.2.2.1 *Population firing rates*

The firing of a given population of neurons is assumed to be elicited when its average membrane potential goes beyond a threshold. Thus in a simplified implementation the firing rate Q_k of a neuron population k is activated through a sigmoidal function of its instantaneous membrane potential V_k :

$$Q_k = \frac{Q_k^{max}}{1+\exp(-(V_k-\theta_k)/\sigma_k)} \tag{2.1}$$

In this function, θ_k denotes the physiological activation threshold which can be obtained from experiments; σ_k is the activation gain that is often influenced by neuromodulators and more generally the brain state.

2.2.2.2 *Intrinsic currents*

Both intrinsic currents and synaptic currents affect the cells' membrane potentials. Modelling intrinsic currents amounts to modelling the membrane's ion channels because they are only driven by the membrane itself instead of being activated by pre-synaptic neurons. The current that passes through a channel i in population k is classically expressed as the product of a fixed maximum conductance $\overline{g}_i$ (representing the conductance when the channel is completely open) with a potential difference between the membrane potential V_k and the reversal potential E_i of the channel (as a driving force). Depending on the electrophysiological characteristics of a channel, sometimes an additional factor is used to model a voltage- (or ion-concentration-) dependent opening probability of the channel. As a consequence, a voltage-dependent intrinsic current from channel i in a population k is modeled in the form:

$$I_i^k = \overline{g}_i \cdot g(V_k) \cdot (V_k - E_i) \tag{2.2}$$

K^+ leaky currents

An important intrinsic current that controls the resting membrane potentials of TC and RT neurons is the K^+ leaky current [124]. The voltage-independence of its opening probability is implied by its name, as 'leaky' refers to the channels with a linear I-V curve. Thus in population k this leaky current (with reversal potential E_k for K^+) is denoted I_{LK}^k and modeled as:

$$I_{LK}^k = \overline{g}_{LK}^k \cdot (V_k - E_K) \tag{2.3}$$

low-threshold calcium T-currents

In addition, both TC and RT neurons possess a low-threshold calcium T-current that generates rebound bursts in their firing patterns. The bursts are generated because the de-inactivation threshold of a T-current is lower than the resting membrane potential - it can only be de-inactivated upon hyperpolarization, so that bursting activities are caused by the overlapping time regime between the voltage-dependent gating of activation and inactivation. The modelling of T-currents in population k follows a Hodgkin-Huxley formulation, i.e. the opening probability is written in the form of products of two voltage-dependent gating variables m_∞^k and h_T^k representing respectively the activation and inactivation variables of this calcium channel (see [62] for more details).

$$I_T^k = \overline{g}_T^k \cdot \left(m_\infty^k\right)^2 \cdot h_T^k \cdot (V_k - E_{Ca}) \tag{2.4}$$

The T-currents in the two neuron types show distinct characteristics in bursting frequency and acceleration. The underlying cellular differences involve (in)activation potentials and time constants of the gating variables, which were already investigated and quantified in earlier studies [61, 60, 62].

Inward rectifying h-current

The h-current in TC neurons, controlling the waxing and waning of spindles is also called an anomalous inward rectifier because its conductance decreases with the rising of membrane [221]. Following what was implemented in the original model, we assume that the opening probability change is due to extracellular calcium (Ca^{2+}) concentration, whereas the time constants were fitted using experimental results [60]:

$$I_h^k = \overline{g}_h \cdot (m_{h1}([Ca^{2+}]_o) + g_{inc} \cdot m_{h2}([Ca^{2+}]_o)) \cdot (V_k - E_h) \qquad (2.5)$$

the sodium-dependent potassiuᴍ current

In the cortex, the generation of SOs (including K-complexes) is mediated by a sodium-dependent potassium current in Pyr neurons, where the conductance depends on sodium concentration $[Na]$:

$$I_{KNa} = g_{KNa} \frac{0.37}{1 + (\frac{38.7}{[Na]})^{3.5}} \cdot (V_k - E_h) \qquad (2.6)$$

$$[\dot{N}a] = (\alpha_{Na} Q_e(V_e)) - Na_{pump}([Na])/\tau_{Na} \qquad (2.7)$$

2.2.2.3 *Postsynaptic currents*

Apart from intrinsic currents, the action potentials of afferent populations of neurons also affect the membrane potential dynamics through a variety of chemical synapses. In the typical case of chemical synapse, neurotransmitters released due to the activation of the pre-synaptic neuron diffuse in the synaptic cleft and lead to the opening of targeted ion channels on the post-synaptic membranes.

Here we assume a monosynaptic connection m from pre-synaptic neuron population k' to post-synaptic population k. Then a synaptic current J_m^k obeys

$$J_m^k = P_m \cdot (V_k - E_m) \qquad (2.8)$$

where P_m is the opening probability of post-synaptic channels. Unlike intrinsic currents, P_m is not only dependent of the target population's own membrane potentials $g(V_k)$, but also on the instantaneous concentration of released neurotransmitters, i.e.

$$P_m = s_{mk} \cdot g(V_k) \qquad (2.9)$$

where s_{mk} represents the opening probability caused by neurotransmitters.

$$s_{mk}(t) = N_{k'k} \cdot Q_{k'}(V_{k'}(t)) \otimes h_m(t) \qquad (2.10)$$

Therefore the complete model for the synapse can be written as

$$J_m^k(s_{mk}) = s_{mk} \cdot g(V_k) \cdot (V_k - E_m) = N_{k'k} \cdot Q_{k'}(V_{k'}(t)) \otimes h_m(t) \cdot g(V_k) \cdot (V_k - E_m) \qquad (2.11)$$

Compared to the general form of intrinsic currents, the product of the synaptic strength and maximum firing rate of pre-synaptic neuron $N_{k'k} \cdot Q_{k'}^{max}$ can be seen as equivalent to the maximum conductance (Eq. 2.1). It is generally be free to adjust $N_{k'k}$ as it depends on many biological parameters of the network for which we do not have a reliable estimate (synaptic strength, number of synapses per cell, etc.). The impulse response h_m can be approximated using the time course of post-synaptic currents (PSCs) in response to brief pulses that can be obtained in voltage-clamp experiments. A common approach to approximate such impulse responses is to model them as an alpha function:

$$h_m = \gamma_m^2 \cdot t \exp(-\gamma_m t) \tag{2.12}$$

The alpha function reaches its peak value at the time point γ_m^{-1}. Such impulse response corresponds to a second order linear dynamical system. As a consequence, the opening probability can be written in the form of a differential equation (for derivation, see Appendix. **??**):

the alpha function

$$\ddot{s}_{mk} = \gamma_m^2 \cdot (N_{k'k} \cdot Q_{k'}(V_{k'}) - s_{mk}) - 2\gamma_m \cdot \dot{s}_{mk} \tag{2.13}$$

which is thus easy to simulate using classical iterative methods.

Most experiments revealing the PSC of an ion channel characterize its kinetics with a rise time and a decay time, which is incompatible with the alpha function. To incorporate these temporal features, an alternative impulse response could be a concatenation of two first-order linear systems with different time constants (referred later as 'two-exponential'). In the time domain, it takes the following form:

$$h_m = B \cdot (\exp(-t/\tau_1) - \exp(-t/\tau_2)) \tag{2.14}$$

where B is the normalization term:

$$B = ((\frac{\tau_2}{\tau_1})^{\tau_{rise}/\tau_1} - (\frac{\tau_2}{\tau_1})^{\tau_{rise}/\tau_2})^{-1} \tag{2.15}$$

In this formula, τ_{rise} denotes the rise time, whereas the decay time is set by τ_1. τ_2 is calculated via the relationship $\tau_{rise} = \tau_1\tau_2/(\tau_2 - \tau_1)$. This impulse response reaches its maximum at time $\log(\tau_1/\tau_2) \cdot \tau_{rise}$. Similarly to the case of alpha function, the convolution by h_m can be written in the form of a differential equation:

$$\ddot{s}_{mk} = B(\tau_2^{-1} - \tau_1^{-1})N_{k'k} \cdot Q_{k'}(V_{k'}) - \tau_2^{-1}\tau_1^{-1}s_{mk} - (\tau_2^{-1} + \tau_1^{-1})\dot{s}_{mk} \tag{2.16}$$

The 'two-exponential' assumption of synaptic kinetics captures more characteristics, but induces one more parameter. Therefore, to limit the number of unknown parameters, we model all synapses with precisely-reported time constants with the 'two-exponential' framework but confine unknown synapses within the alpha function framework.

For synapses with long-distance projections across brain regions (e.g. thalamocortical projection), following the original paper, we deal with the axonal conductance delay by adding another linear filter of the synaptic output, which can be approximated by the convolution of another alpha function:

$$\ddot{\phi}_k = \nu^2 \cdot (Q_k(V_k) - \phi_k) - 2 \cdot \dot{\phi}_k \tag{2.17}$$

2.2.2.4 *Membrane potential and LFPs*

The final step is to establish how the membrane potential evolves with the transmembrane currents. The neuron population is assumed equivalent to a neuron modelled with a single compartment model. As illustrated in Figure 2.1A, the most simplified case is a neuron population k influenced by a single (m-type) synaptic current $J_m^k(s_{mk})$ and a single intrinsic current I_i.

The denoted synaptic conductance follows the representation in Eq. 2.9. With Kirchhoff's current law, we would be able to get the adaptation of membrane potential V_k with currents:

$$C_m\dot{V}_k = -\frac{V_k - E_L}{R_L} - g_{mk}(s_{mk}) \cdot (V_k - E_L) - I_i \tag{2.18}$$

The second term on the right in Eq. 2.18 represents the synaptic current, where $g_{mk}(s_{mk})$ is the real synaptic conductance. Comparing this form with the synaptic current model (Eq. 2.8, 2.9), the synaptic conductance $s_{mk} \cdot g(V_k)$ can be seen as the real conductance $g_{mk}(s_{mk})$ normalized by the leaky conductance $1/R_L$, i.e.

$$s_{mk} \cdot g(V_k) = \frac{g_{mk}(s_{mk})}{1/R_L} \Rightarrow g_{mk}(s_{mk}) = \frac{s_{mk} \cdot g(V_k)}{R_L} \tag{2.19}$$

Thus combining Eq. 2.8, 2.9, Eq. 2.18 can be rewritten as:

$$C_m \dot{V}_k = -\frac{V_k - E_L}{R_L} - \frac{J^k_m(s_{mk})}{R_L} - I_i \tag{2.20}$$

By re-arranging the denominator, we can obtain:

$$R_L \cdot C_m \dot{V}_k = -(V_k - E_L) - J^k_m(s_{mk}) - R_L \cdot I_i \tag{2.21}$$

The first term on the right in Eq. 2.21 can be understood as another equivalent current J^k_L that passes through the membrane. $J^k_L = (V_k - E^k_L)$ is defined as the general leaky current, a linear current unaffected by either pre-synaptic or post-synaptic population. Note the K^+ leaky current is excluded because we specifically design it to be influenced by the concentration of ACh.

In experiments a commonly reported feature is the membrane time constant $\tau_k = R_L \cdot C_m$. With a rearrangement of terms, Eq. 2.21 can be written as:

$$\tau_k \dot{V}_k = -J^k_L - J^k_m(s_{mk}) - C_m^{-1} \cdot \tau_k \cdot I_i \tag{2.22}$$

In more general cases, the numbers of synaptic as well as intrinsic currents are not necessarily confined to one. For multiple currents we can obtain the general form:

$$\tau_k \dot{V}_k = -J^k_L - \sum_{m,k} J^k_m(s_{mk}) - C_m^{-1} \tau_k \sum_i I_i \tag{2.23}$$

In many experimental studies, the thalamic PGO waves are often characterized in LFPs. LFPs reflect some spatially extended measure of the activities of a mass of neurons. While understanding the nature of LFPs is still preliminary, one study attempted to link LFP with the currents flowing through the neuron populations [154]. In this study, the authors approximate the LFP with a weighted sum of synaptic currents, e.g. AMPA and GABA currents, which is to some extent compatible with our model. However, in our model, the intrinsic currents also play important roles. We propose to model the LFP of a neuron population as an instantaneous sum of all currents to take into account the effects of intrinsic current:

$$\mathrm{LFP}_k = -J^k_L - \sum_{m,k} J^k_m(s_{mk}) - \sum_i I_i \tag{2.24}$$

2.2.3 *Assumptions for model structures*

After introducing the modelling methodology for each unit of neuronal population in the neural mass model, we now elaborate on the biological basis forming the assumptions for the model structure. Specifically, this includes the selection of neuron types and deciding on how different neuronal populations are connected.

Our model focuses primarily on the reproduction of electrophysiological characteristics of PGO waves in the thalamus, where experimental PGO wave traces are most prominent and cellular mechanisms are relatively clear after extensive investigations in the field. Well-replicated thalamic PGO waves enable us to speculate on the effect PGO wave triggers in the cortex. Therefore, it is natural that we take as a starting point a neural mass model of thalamocortical network proposed in [206], whose cortex and thalamus modules, as well as the thalamic neuronal types (the TC and RT neurons), match well with our purpose. What we need to do, is to add the LGin neurons and connect them with the pons in a biologically realistic way.

for abbreviations check Section 2.2.1 and Figure 2.1B

2.2.3.1 *Brief description of the thalamocortical module*

First, this section will describe the thalamocortical model we extended ([60, 206], in the following referred to as the *Costa model*), while clarifying the major differences between the thalamocortical module of our model and the Costa model. This Costa model can reproduce signatures of NREM sleep, e.g. thalamic spindles and K-complexes in the cortex, by incorporating various intrinsic currents to generates specific neuronal activities (e.g. bursting). Thus, elaborating the model would also facilitate the investigation of interactions between PGO waves with these NREM events.

The simulated spindles will be used to validate causality measures in Section 4.3.3 and Section 4.3.4.2

The thalamus module consists of TC and RT neurons, as they are the major neuronal types reported in PGO-related studies [224]. Both neurons are modeled as a rated-coded single compartment model described in Eq. 2.24. To model the intrinsic properties of thalamic neurons, we follow the assumptions made in the Costa model , i.e. a K^+ leaky current (Eq. 2.3) and a low-threshold calcium T-current (Eq. 2.4) in both neurons populations, together with a hyperpolarization-activated anomalous rectifier h-current (Eq. 2.5)in TC neurons.

the thalamus module

The cortex is simplified as a population of pyramidal cells interacting with a group of inhibitory neurons. Similarly, for the cortex, we keep the intrinsic current - a sodium-dependent potassium current I_{KNa} (Eq. 2.6 and Eq. 2.7)-in Pyr neurons as the mechanism to maintain the NREM-related cortical oscillations (i.e. SOs and K-complexes).

the cortex module

Following a major hypothesis of thalamic PGO wave generation [103], we added another population representing LGin neurons using the same framework. As for LGin neurons, we don't make any assumptions in their intrinsic currents or their role in spindle generation due to limited knowledge [254].

LGin neurons

2.2.3.2 *Major contributions to adapt the Costa model into a PGO model*

Although the thalamocortical module in our PGO model is based on the Costa model, including the synaptic connections and intrinsic currents, we made major modifications to adapt it to a PGO model where the thalamus receives strong perturbations from the pons. With exploring alternative model settings and exploiting the parameter space (see Section 2.2.5), we found that the following points are indispensable in the reproduction of PGO waves.

1. Critically, for the modelling of neuromodulatory effects (see Section 2.2.4), we differentiate the maximum conductance for the leaky potassium channels in TC and RT, which used to be the same variable in the original model. This change is responsible to modulate the membrane potentials of TC and RT neurons differently during the switch between

pre-REM and REM stages, which is important for the reproduction of PGO-triggered firing patterns in both neuronal populations.

2. Another important modification is the incorporation of the LGin neurons. As briefly addressed in Section 2.2.3.1, this type of neurons are hypothesized to play a role in thalamic PGO wave generation, while our simulation results also show that without this neuronal population the negative peak in PGO waveform cannot be replicated.

3. To avoid additional spindle-like oscillations in the thalamus induced by the pontine PGO inputs, we weakened the connection from Pyr neurons to RT neurons and balanced their potassium leaky conductance accordingly to restore the N2 state for spindle generation and N3 state as the pre-REM state.

4. To modulate the amplitude of REM PGO waves in the LFP, we changed one of the bifurcation parameters of the cortex from the neural gain of the sigmoid function to the maximum firing rate such that the signal-to-noise ratio falls into a reasonable range.

5. Compared to the C++ implementation provided by the authors of the Costa model, we transferred the model to Python and adjusted the noise levels in both the thalamus and the cortex to maintain a preferred frequency of spindle occurrence.

2.2.3.3 *Model assumption for the pons*

After introducing the thalamocortical module, we now focus on the modelling of the pons module as the important input module triggering the thalamocortical network.

Two groups of neurons in the pontine region, termed as the PGO *triggering neurons* and *transferring neurons*, are the executive elements of PGO waves [43]. They are both located in the peribrachial area (PBL) [214], which contains a number of important nuclei that are involved in the regulation of sleep and arousal. Functionally, the PBL can be separated into two parts: the rostal (R-PBL) and the caudolateral (C-PBL) parts [43]. The R-PBL mainly consists of the pedunculopontine tegmentum nucleus and laterodorsal tegmentum nucleus, while the most important nuclei in C-PBL include the parabrachial nucleus.

As their name indicates, experimental evidence showed that triggering neurons are PGO-state-on bursting neurons assumed to initiate the PGO phasic event located in C-PBL [48, 42, 43].

PGO triggering neurons were recorded to burst in high frequency (300-500Hz) 25±7 ms before the thalamic PGO waves. These activities are hypothesized to propagate to the so-called transferring neurons in R-PBL, which are PGO-on low-frequency bursting neurons [43] firing low-frequency (120-180 Hz) bursts with 3-5 spikes 5-15 ms before the thalamic PGO waves [155, 170, 225, 201]. The transferring neurons were presumed to project directly to the thalamus [201] as the last relay station of the local PGO-related circuits in the pons [182].

In the model, C-PBL activities (of the triggering neurons) are assumed as a trigger that initializes the whole network activity. To match the high-frequency bursts pooled across a population, the firing rate of C-PBL neurons should rise rapidly and persist for a short period. It is then natural to model the C-PBL activity in pre-REM stage as a brief pulse lasting

10 ms as an approximation of the bursting duration [48]. Deducing from experimentally-reported peri-PGO spike histograms [48], we assume that REM-related C-PBL bursts can be modelled by three 10-ms pulses with a bursting interval of 200 ms.

R-PBL PGO-transferring neurons

The R-PBL PGO-transferring neuron populations are modelled with the same type of sigmoidal activation function as for the thalamic neurons (Eq. 2.1). The activation threshold remains unchanged because the pontine and thalamic spikings are both associated with fast Na^{2+} spikes and should match each other. We present briefly here the synaptic connection from C-PBL neurons to R-PBL neurons and some intrinsic cellular mechanisms of R-PBL neurons.

the projection from triggering neurons to transferring neurons

The projection from triggering neurons to transferring neurons is presumably glutamatergic [108, 202, 203, 46], indirectly supported by the findings that other neurotransmitters are inhibitory [143, 132, 246]. Electrophysiological studies showed the projection could be mediated by a combination of NMDA and AMPA receptors [203], with a contribution ratio as NMDA: AMPA = 1:5. The same experiment also quantified the decay times for both currents: 8.77 ms for the AMPA channel and 129.4 ms for NMDA. The Nernst potentials were measured to be 16.3 mV for NMDA and 3.4 mV for AMPA (close to the theoretical value of 0 mV).

Conductance of the AMPA channel is invariant to the post-synaptic membrane potential. Non-linearity in NMDA currents has long been reported and well-documented. We followed the classical modelling of the voltage dependence first introduced by Destexhe ([61]):

$$g_{NMDA}(V_l) = \frac{1}{1+\exp(-0.0062V_l)}[Mg^{2+}]_o/3.57 \tag{2.25}$$

where l represents the population of R-PBL transferring neurons, and V_l denotes its membrane potential. $[Mg]_o$ represents the extracellular concentration for Mg^{2+} ions.

T-current in R-PBL neurons

The slow-frequency bursts occurring in transferring neurons are rebound bursts evoked by activating an intrinsic low-threshold calcium T-current under hyperpolarization [115, 114]. As it is similar to the thalamic T-currents discovered in the TC neurons, we modelled it with Eq. 2.4 [62, 61], but refitted the conductance with digitized experimental I-V curves [114]. We constructed a Boltzmann-like function that is able to approximate the activation curve:

$$m^l_\infty(V_l) = \frac{2}{1+\exp(-(V_l+50.6)/0.44)+\exp((V_l+50.6)/17.4})\tag{2.26}$$

In a similar way, we fitted the inactivation gating variable leading to:

$$h^l_\infty(V_l) = \frac{1}{1+\exp((V_l+65)/2.7)}\tag{2.27}$$

the cholinergic hyperpolarizing current in R-PBL neurons

Considering the increasing concentration of ACh during PGO-related sleep stages, together with evidence of the inhibitory effects of ACh on transferring neurons [132], we assume that some cholinergic modulatory inputs cause the hyperpolarization as a prerequisite to de-inactivate the T-current. The cholinergic input goes through a potassium inward rectifier mediated by a muscarinic receptor [132]. The I-V curve of cholinergic influence has already been characterized, from which we defined an approximate mathematical formulation with a procedure similar to the fitting of T-currents.

$$g_{IR}(V_l) = \frac{1}{1+\exp((V_l+35)/10.9)}\tag{2.28}$$

2.2.3.4 *Propagation of pontine PGO waves to the thalamus*

The complete model involving carefully-designed ponto-thalamic synaptic connections is illustrated in Figure 2.1B. After receiving bursting activities from the triggering neurons, the transferring neurons send cholinergic inputs to the three thalamic neurons (TC, RT, and LGin neurons) via 5 cholinergic ponto-thalamic projections. Here we briefly present our assumptions regarding the chemical nature and kinetics of the pontine-thalamic projections and neurophysiological evidence supporting them.

cholinergic projections from R-PBL to TC neurons

The R-PBL neurons send two excitatory projections to the TC neurons via cholinergic projections mediated by both nicotinic acetylcholine receptors (nAChR) and muscarinic acetylcholine receptors (mAChR) receptors. The nAChR-based channel, underlying the generation mechanism of a fast depolarization in diafferated cats [158, 102, 103, 103] and short bursting in naturally sleeping cats [224, 158, 104], let through a mixed cation current with a voltage-independent conductance and a Nernst potential of 18.9±8.9 mV [158]. The mAChR-mediated synapse, generating the prolonged spiking in TC neurons following the initial bursts, is coupled to a leaky K^+ channel via G-protein cascade (reversal potential: -97±6.1 mV, note that "leaky" implies a voltage-independent conductance) [156]. We set the conductance negative as activation of the mAChR-mediated synapse decreases the conductance (i.e. closes the channel). Specifically, we speculate that this potassium leaky channel is the same as the K^+ leaky channel already included in the original model of TC neurons (see Section 2.2.2.2 and Section 2.2.4), whose role is designed to mediate membrane de-/hyperpolarization [158, 17, 16]. We introduced a saturation mechanism: the maximum amount of conductance decrease caused by the mAChR-mediated synapse is equal to its resting conductance, i.e. it cannot go beyond complete closure. This saturation mechanism was not reported but implied in the experimental papers and has been proven useful in replicating the switch between PGO wave subtypes (see Figure 2.6B).

cholinergic projections from R-PBL to RT neurons

Clear evidence suggests that the RT neurons also receive pontine inputs via both nAChR- and mAChR-mediated synapses, corresponding to a PGO-triggered fast depolarization/bursting and slow hyperpolarization observed in RT neurons [103, 130, 176, 230, 11]. The time constants of synatic kinetics were quantified (rise time: 10.8 ms, decay time: 123.6 ms), enabling us to apply the 'two-exponential' form of synaptic model. Following classical models of nAChR-mediated channels, we assume a linear voltage-independent with a reversal potential of -5 mV. On the contrary, the mAChR-regulated channel, associated with a K^+ channel with the Nernst potential of around -93.2±0.6 mV [230], works as a inward rectifier [103, 176, 130, 230, 11], whose conductance decreases with increased membrane potential, with voltage dependence characterized and fitted with asigmoidal function:

$$g_{mAChR}(V_r) = \frac{1}{1 + \exp((V_r + 66.3)/29.1)} \tag{2.29}$$

The corresponding rise time and decay time of the synaptic kinetics are 107.6±8.6 ms and 639.0±102 ms.

cholinergic projections from R-PBL to LGin neurons

The assumption that LGin neurons contribute to thalamic PGO wave generation is supported by the existence of a transient hyperpolarization of TC neurons caused by depolarization in LGin neurons [103, 104, 105, 224]. We model a nAChR-mediated projection from R-PBL neurons to LGin neurons to generate the depolarization [254], with the same model as the corresponding channel in TC neurons with differently-tuned synaptic kinetics.

Conservatively, for simplicity, we omit the other potentially existing intrinsic current, as current experimental evidence is insufficient to support their roles in thalamic PGO wave formation.

The transmission of PGO-related activities from the pons to the thalamus is not instantaneous but delayed by membrane and axonal properties. For the 5 pontine-thalamic connections, the only precisely reported fact is the latency difference between nAChR-mediated and mAChR-mediated currents in RT neurons, as 28 ms [230]. Latency histograms of the nAChR-mediated currents in TC and RT neurons [105] also suggest a constraint in setting the delays. Extrapolating with all taken into consideration, we implement the delays in our model by setting a fixed latency for each projection that is consistent with the experimentally-revealed facts.

2.2.4 *Cholinergic modulation of PGO waves*

After establishing the general form of ponto-thalamo-cortical modules, we now turn to the modelling the state-dependent modulations of the neuronal activities of different subtypes. The transition from pre-REM to REM sleep stages is strongly dependent on the changes of certain neuromodulators [43]. To reproduce the difference between PGO wave subtypes during pre-REM and REM, we need to know how neuromodulation affects the PGO propagating network.

The reason why only cholinergic influence is modelled

As illustrated in Figure 2.2A, modulatory neuron populations associated with ACh and monoamines (serotonin and noradrenaline) are reciprocally interacting to influence the activities of the several neuron types related to PGO wave generation [98]. Wakefulness and NREM sleep is accompanied by a high concentration of monoamines and a low concentration of ACh. In the transitional stage (i.e. pre-REM), the activities of aminergic neurons decrease while the cholinergic neurons are gradually activated. When REM sleep is approached, cholinergic activities remain persistently at a high level whereas aminergic activities are suppressed. In short, aminergic neurons plays a disinhibitory gating role for the cholinergic neurons, i.e. activities of the former are negatively correlated with the latter. Thus for simplicity, we can omit the monoamines and model only the effect of ACh on the network activities.

The levels of acetylcholine concentration in the pons and the thalamus are mainly influenced by the cholinergic tonic firing neurons in the R-PBL which directly project to both thalamic nuclei. From NREM to REM states, these neurons continuously increase their firing rates [225]. We use this observation to build a simple linear approximation of the transition between the two states depending on the normalized ACh concentration $[ACh](t)$ (ranging from 0 to 1).

Our strategies of linking ACh concentration to the activity patterns of PGO-related neurons are as follows:

- First, pick several crucial parameters in the model based on biological plausibility (e.g. maximum conductance for currents or activation threshold of neuronal populations), which were reported to contribute to the switch from NREM sleep to REM sleep.

- Next, adjust and fix these parameters to reproduce the firing modes of neurons in pre-REM and REM states.

- Finally, establish a sigmoidal relationship between the ACh concentration and the crucial parameters based on experimental evidence, i.e. highest ACh concentration during REM and lowest during pre-REM.

For the cortex, the state-reconstruction can be resolved by simplifying the neuromodulated isolated cortical model in [38]: instead of modulating the cortical network with the concentrations of ACh, serotonin, and noradrenaline, we restrict the neuromodulation to ACh. As for the cortical-critical parameters, we followed their choice of the adaptation strength of the sodium-dependent potassium current $\bar{g}_{KNa}$, but changed the other to the maximum firing rate of Pyr neurons Q_p^{max} [6, 66, 217, 226] to ensure a broader range of REM activities.

ACh has been reported *in vitro* and *in vivo* to tonically depolarize TC neurons [157, 59, 223] and hyperpolarize RT neurons , both via a mAChR-mediated K^+ channels [240, 157, 239]. State transitions in the thalamus are implemented by modulating the K^+ leaky conductance present in the model of both TC and RT neurons, as described in Section 2.2.2.2.

For TC neurons, we decreased its potassium leaky conductance from 0.24 to 0.06 to mimic its depolarization. The same conductance in RT neurons was increased from 0.18 to 0.62 to simulate the hyperpolarization. The values of conductance are optimal parameters according to the empirical experience of tuning the dynamic system, whereas various other combinations would also satisfy the requirements for state switches.

The resulting ACh-modulated model parameters are illustrated in Figure 2.2B. While some related studies [157, 59, 255], measuring how much the leaky conductance was changed in response to ACh application, might have provided a quantitative basis for setting the conductance values, we did not follow them because many unknown experimental variables (*in vivo*/*in vitro*, anesthesia, doses of micro-injections, etc.) lead to large uncertainty on the parameter choice.

2.2.5 *Validations for model assumptions*

As introduced in the sections above, the full PGO model seems a complicated structure with massive internal dynamical interactions. However, the key of the model is the perturbations in the thalamocortical module triggered by the pontine inputs which appear in the shape of PGO waves. Therefore, it is critical to make sure that the pontine inputs, as well as the connectivity from the pons to the thalamus, are based on reasonable assumptions. In this section, we validate these two aspects by combining both biophysical mechanisms and exploration of the model.

2.2.5.1 *Pontine parameter tuning*

The firing rate of R-PBL neurons is the final output of the pontine model that is sent to the thalamic neurons. Interestingly, due to the T-current, the bursting duration of R-PBL neurons is highly sensitive to the strength of the cholinergic hyperpolarization. In the model, this strength is represented by the conductance of the cholinergic hyperpolarizing current (g_{ACh}), which requires appropriate adjustment. Therefore, we tuned the parameter of the conductance in the model, where the resulting membrane potential and firing rates are illustrated in Figure 2.3.

The tuning results match the cellular mechanism underlying bursting activities, which are co-activated by a low-threshold calcium T-current [115,

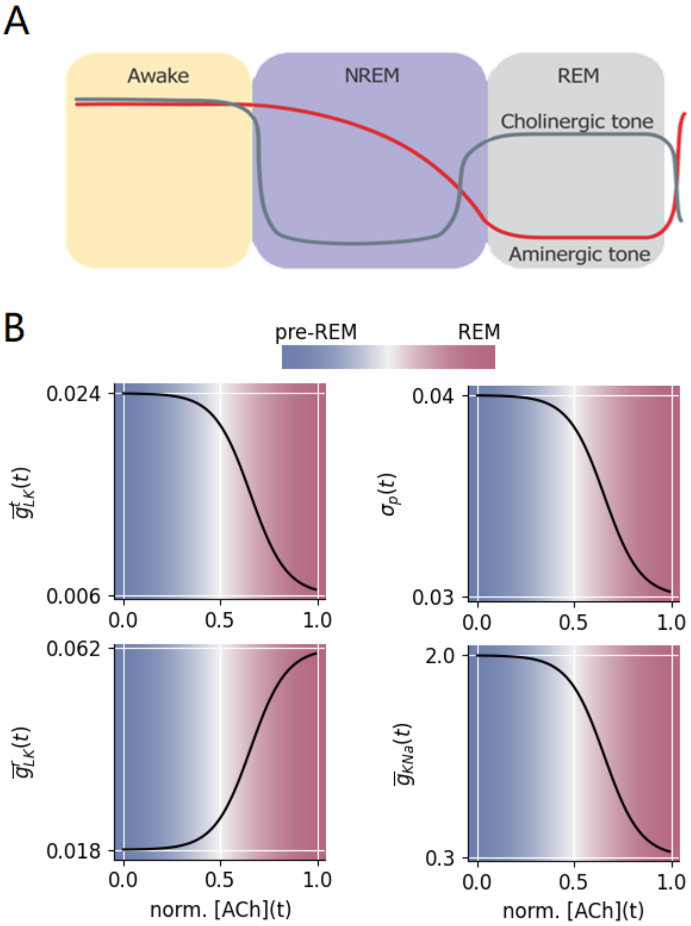

Figure 2.2: ACh-modulated model parameters. (A) State-dependent changes of key neuromodulatory tones (adapted from [98]). The aminergic tone dominates the NREM state while the cholinergic tone is inhibited. When the brain transits from NREM to REM sleep, the aminergic tone decreases, dis-inhibiting the cholinergic tone. During REM sleep, the cholinergic tone dominates while the aminergic tone remains low. (B) 4 critical parameters from the thalamocortical part are selected to reflect the cholinergic influence of the network. Their changes are linked to ACh concentration via a sigmoid-like relationship mimicking smooth stage switches.

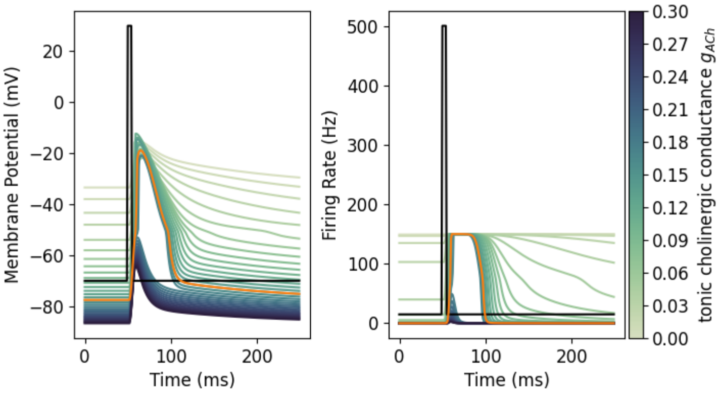

Figure 2.3: ACh-tuned pontine neuronal activities. PGO-triggered membrane potential and firing rate of R-PBL neurons modulated by the conductance of a tonic cholinergic current. The conductance is critical to separate the neuronal activity into three patterns: with small conductance ($g_{ACh} <= 0.04$), the membrane is slightly depolarized; with moderate conductance ($g_{ACh} = [0.04, 0.18]$), a calcium spike rides on the PGO-triggered depolarization; when the conductance is set large ($g_{ACh} > 0.18$), the calcium spike disappears. This effect reflects the nonlinear intrinsic properties of the pontine T-current regulated by the cholinergic input current. The orange line marks the selected value of $g_{ACh} = 0.16$.

Table 2: Quantitative similarities between pontine simulation and experimental results. See [48] for characteristics related to C-PBL and [225] for that of R-PBL neurons.

Electrophysioligical characteristic	simulated (ms)	experimental range (ms)
bursting duration (C-PBL)	10	[6, 16.7]
bursting duration (R-PBL)	27	[16.7, 41.7]
latency to bursting onset (C-PBL → RT)	25	[17.5, 32.5]
latency to bursting onset (R-PBL → RT)	14	[5, 15]
latency to negative peak in LFP (R-PBL → RT)	[35, 39]	[20, 40]
bursting interval (C-PBL)	set to 200 ms	around 20 oms

114, 133] and a hyperpolarizing mAChR-mediated intrinsic cholinergic current [132]. The low-threshold T-current, similar to the standard T-current discovered in the thalamus and fitted on rat electrophysiological data, is activated only upon resting hyperpolarization and a fast depolarizing input (for details see Section 2.2.3.3). The conductance of the cholinergic intrinsic currents regulates the degree of hyperpolarization in the resting potential of R-PBL neurons.

As shown in Figure 2.3, with a fixed strength of synaptic input from C-PBL neurons, only a carefully selected range of cholinergic conductance can trigger a calcium spike with bursts; too strong or too weak resting hyperpolarization only cause a small depolarizing effect but no bursts. Therefore, based on such nonlinear properties, we can find an appropriate value of the cholinergic conductance ($g_{ACh} = 0.16$) to achieve the biologically-based temporal characteristics of pontine activity. This similarity is further reported in Table 2.

2.2.5.2 *Validation of ponto-thalamic projections*

As demonstrated in Section 2.2.3.4, activities of the R-PBL neurons are transmitted to the thalamic module via 5 cholinergic projections. It is then a natural question whether this is a redundant assumption of the ponto-thalamic connectivity and should be validated. Therefore we test the necessity of each of the 5 projections by cutting it and check whether the resulting waveforms are different from before cutting it. The detailed results will be presented in Section 2.3.4, and we will only explain the rationale here.

Importantly, by tuning the strengths of the 5 projections, it is possible to generate highly variable PGO waveforms in the thalamus. Before blocking any of the projections, we already fix an optimized parameter set that generates the PGO waveforms similar to the experimental recordings (see Section 2.3.2). After cutting one of the projections, we still need to account for the variabilities generated by the remaining projections.

Therefore, in practice, together with removing one of the projections, we scan the connectivity strengths of the remaining projections, while each parameter set results in a specific waveform. All these waveforms, as high-dimensional vectors, can be mapped as features in a lower dimension after dimension reduction. Eventually, if the lower-dimensional representation of the optimized waveform does not overlap with the representations of the waveforms after blocking a projection, this suggests that the optimized waveforms are sufficiently distant from the alternative waveforms in the high dimensional space. Therefore, we can conclude that each of the projections is necessary for the modelling.

2.2.6 *Spike-time-dependent plasticity for neural mass models*

motivation of STDP analysis

With the model constructed as presented above, we are able to generate thalamic PGO waveforms resembling their counterparts recorded in the electrophysiological signals (Section 2.3.2). The similarity between simulated and experimentally-reported cortical PGO activities cannot be addressed systematically due to the lack of experimental data of PGO-triggered cortical activities. However, as the thalamocortical interactions in our model are based on biophysical mechanisms, it is reasonable to assume that the PGO-triggered activities in the cortex are a useful approximation of the real signals. We can further assume that the network dynamics underlying the cortical PGO waveform have some biophysical similarities to the real mechanism. Therefore, we use the simulated PGO-triggered activities to explore the plasticity effects that PGO waves may induce in the cortical circuits.

Following a plasticity framework for mean-field models proposed by [196] and [76], we introduce a time-dependent plasticity rule for neural mass models. The change of synaptic strength between an afferent population m and an efferent population k depends on the simultaneous change of synaptic synaptic current J^k_m and post-synaptic firing rate Q_k:

$$\frac{dN}{dt}_{mk} = \int_{-\infty}^{+\infty} \left\langle Q_k(t+\tau)H(\tau)J^k_m(t)\right\rangle_t d\tau = \int_{-\infty}^{+\infty} \left\langle Q_k(t+\tau)J^k_m(t)\right\rangle_t \cdot H(\tau)d\tau \tag{2.30}$$

where $H(\tau)$ is the STDP function shown in Figure 2.4B and the bracket denotes a time averaging. Parameters in the analysis are picked from *in vitro* measurements in hippocampal cell cultures [13].

Simply put, this equation shows that the sign and strength of plastic changes are determined by the similarity between the shape of the classical

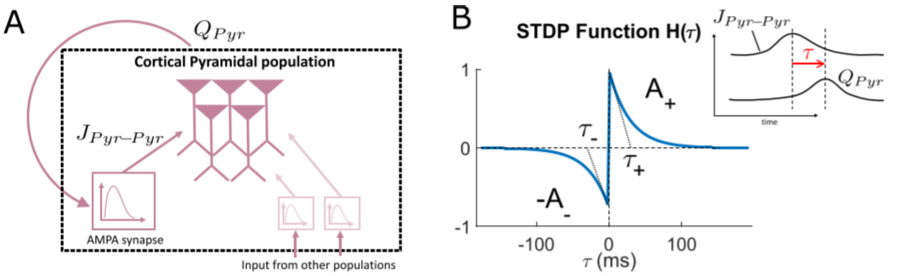

Figure 2.4: Illustration of the population STDP rule. (A) Illustration of the relationship between pyramidal pre-synaptic current $J_{Pyr-Pyr}$ and post-synaptic firing rate Q_{Pyr} controlling cortico-cortical STDP plasticity. The impulse response of the AMPA synapse links the two quantities. (B) Illustration of the STDP window. The horizontal axis τ represents the time difference between pre- and post-synaptic activities (see right inset). The synapse gets strengthened when pre-synaptic activities elicit post-synaptic ones, and vice versa for the opposite sequence.

Intuition underlying the population STDP rule

STDP function (Figure 2.4B) and the shape of the cross-correlation function between the pre- and post-synaptic activities, measured by the integral of their product over time. In the context of this study, we are interested in the plasticity of cortico-cortical excitatory connections, which are expected to be modified by memory consolidation and homeostatic processes during sleep. The considered pre-synaptic current is thus the excitatory AMPA current that the Pyr population sends to itself, and the Pyr post-synaptic firing rate measures the post-synaptic activities (Figure 2.4A), yielding the following equation:

$$\begin{aligned}\frac{dN_{Pyr-Pyr}}{dt} &= \int_{-\infty}^{+\infty} \left\langle Q_{Pyr}(t+\tau)H(\tau)J_{Pyr-Pyr}(t)\right\rangle_t d\tau \\ &= \int_{-\infty}^{+\infty} \left\langle Q_{Pyr}(t+\tau)J_{Pyr-Pyr}(t)\right\rangle_t \cdot H(\tau)d\tau \qquad (2.31)\end{aligned}$$

The STDP rule will be applied to both PGO waves and thalamic spindles to account for their event-triggered plastic changes in the synapse, for details see Section 2.3.5.

2.3 RESULTS

In the Results section, we will illustrate in the following sections how well the model is able to reproduce many of these aspects of the experimental recordings and extrapolate on plasticity effects.

2.3.1 *Establishment of NREM, pre-REM and REM states*

According to Section 2.2.4, we already established a link between ACh concentration and the switch of the model between NREM and REM stages, enabling us to simulate neuronal activities in both sleep stages. Based on this mechanism, we first check whether the model reproduces long-term patterns of ponto-thalamo-cortical activity observed experimentally. Figure 2.5 shows the LFPs in R-PBL, TC and Pyr neurons simulated in the whole network by tuning the ACh-modulated parameters shown in Figure 2.2, illustrating the contrast between the 3 scenarios: NREM (pre-REM without PGO waves), pre-REM (with PGO waves) and REM sleep.

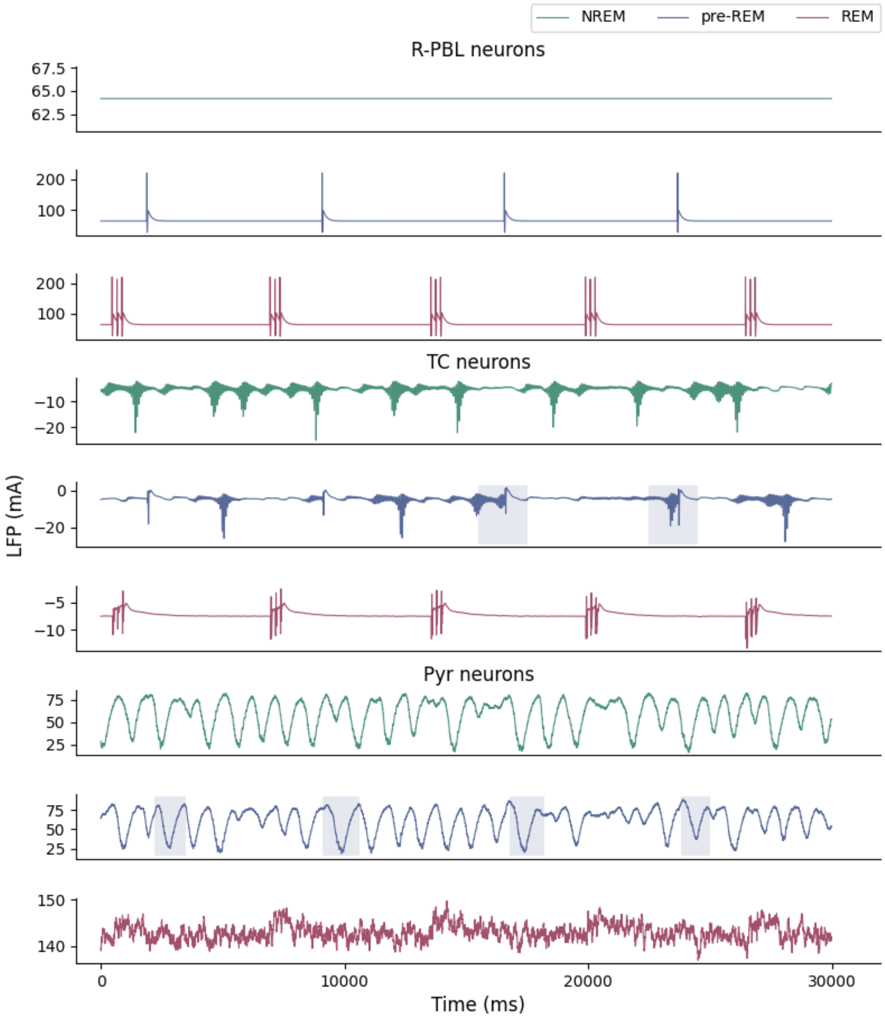

Figure 2.5: Establishment of sleep stages: State comparison of simulated LFPs in R-PBL (top), TC (middle) and Pyr neurons (bottom). LFPs of the three neuronal populations in three states (NREM, pre-REM and REM) are plotted in separate colors for comparison. During the pre-REM state, in the thalamus, contrary to the NREM state, spindles are interrupted by PGO waves, as marked by grey shades of TC neurons in the pre-REM stage. In the cortex, PGO waves trigger more slow oscillations during pre-REM, as marked by grey shades of Pyr neurons in the pre-REM stage.

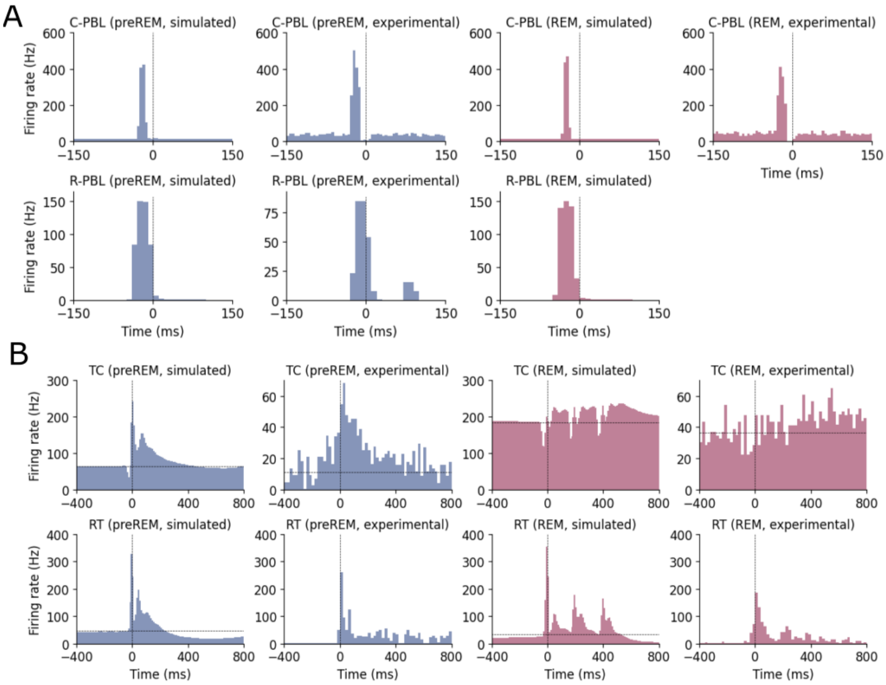

Figure 2.6: Model Reproduction of pontine and thalamic neuronal activities. (A-B) Comparison of simulated and experimental ponto-thalamic peri-PGO histograms during pre-REM and REM states: in all conditions simulated results resemble experimental ones. The histograms are averaged over 1000 trials of simulated events with small variations of the ponto-thalamic projections.

During the simulated NREM stage (Figure 2.5 upper panel), R-PBL neurons do not burst, as shown in early studies [155, 201, 225]. TC neurons show strong spindle oscillations appearing with a frequency of occurrence of 0.1-0.2 Hz in the simulated LFP, accompanied by SOs in the Pyr neurons. In the presence of PGO inputs (Figure 2.5, middle panel), biphasic patterns occur in R-PBL neurons triggered by PGO inputs. In the thalamus, some spindles are blocked by PGO inputs, matching experimentally-recorded spindle interruption by phasic brainstem stimulations [105], thus resulting in stronger SOs in the cortex. During the REM stage (Figure 2.5, lower panel), thalamic spindles disappear with a depolarizing effect of the ACh-modulated current, displaying similar effects for SOs in the cortex. The activities of these simulated stages match the corresponding tonic electrophysiological traces shown in [224], supporting appropriate modelling of these states.

2.3.2 *Model Reproduction of Pontine and Thalamic neuronal activities*

Beyond the sustained activities of pre-REM and REM states, we checked the ability of the model to replicate transient changes in PGO-triggered firing patterns reported by classical literature in cats [224, 48, 225]. Figure 2.6, 2.7 show a comprehensive comparison between the key features of PGO-related neuronal activities in simulations and their experimental counterparts reported in classical electrophysiological studies in both pre-REM and REM stages.

In the thalamus, the similarities are reflected in the following aspects (Figure 2.6A). First, TC and RT neurons show differences in their baseline firing

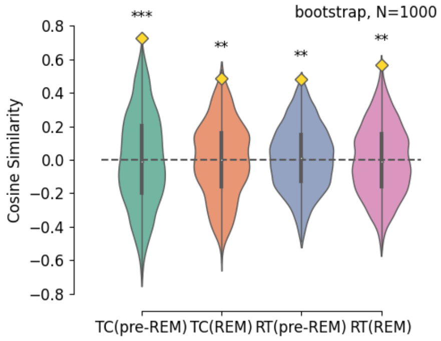

Figure 2.7: Cosine similarity as a measure of similarity between simulated and experimental PGO waves. Yellow diamonds represent the cosine similarity between the simulated and experimental peri-PGO histograms presented in (b); violin-plots show the bootstrapped distribution of cosine similarity calculated with 1000 epochs randomly selected from the original simulation. Stars indicate that the cosine similarity in (b) is significantly different from the bootstrapped null distribution (***:p<0.001; **:p<0.01; *:p<0.05).

rates (indirectly, membrane potentials), reflecting cholinergic modulation of thalamic membrane potentials via a potassium leaky conductance. TC neurons are more depolarized during REM than during pre-REM, while this is the opposite for RT neurons. As depolarization inactivates the T-current of TC cells, state modulation also switches the dynamics of TC neurons from bursting during pre-REM, modeled by the sharp increase in firing rate at PGO wave onset, to non-bursting during the simulated REM state.

In contrast, consistent with experimental studies [224], RT bursting activity spans both sleep stages. Besides, on a finer scale, after bursting RT neurons undergo a slower hyperpolarization induced by the mAChR-receptor-mediated synapse, which is reproduced by the simulated events. Such hyperpolarization is also present during the REM stage, contributing to the decreased response in RT neurons to the second and third pontine pulses. In line with experimental evidence, TC neurons show prolonged increased firing following their initial bursts [157, 156]. Such sustained firing patterns are caused by activation of the cholinergic ponto-thalamic synapse mediated by a mAChR receptor, together with the dis-inhibition effect by RT neurons during their hyperpolarization [104].

In addition, the similarity between experimental and simulation results is significant (permutation test, p<0.05), as measured by a larger cosine similarity between PGO peri-event time windows in comparison with those from randomly selected trials as shown in Figure 2.7. These similarities all validate the reliability of the model in replicating cellular in-vivo activity, supporting it can serve as a steppingstone to investigate plasticity induced by PGO wave activities.

2.3.3 *Model Reproduction of thalamocortical LFP*

Apart from the neuronal firing patterns, the model is also able to reproduce the typical PGO-triggered LFP waveforms in the thalamus and cortex, which were more frequently recorded in early reports of PGO waves (e.g.[30, 224, 170]).

In Figure 2.8, we present a comparison of the LFP waveforms and corresponding spectrograms between two subtypes of PGO waves and two sub-

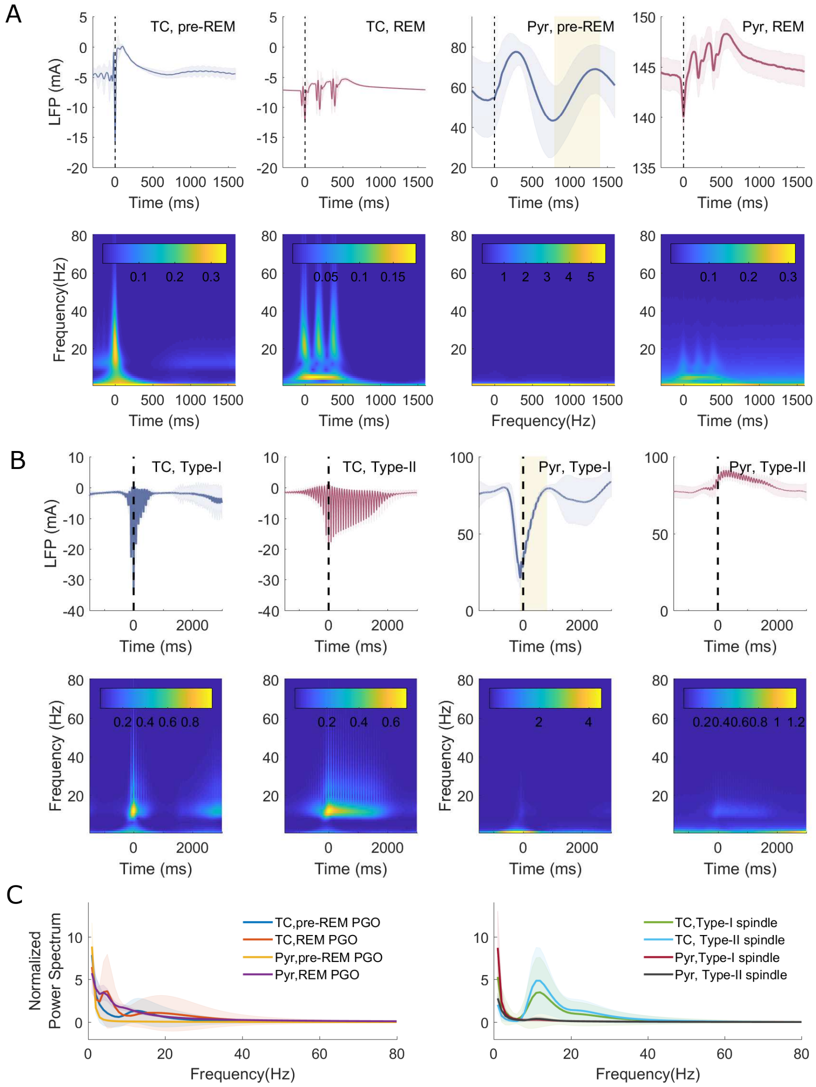

Figure 2.8: Model Reproduction of thalamocortical LFP. (A-B) Averaged events and normalized peri-PGO and peri-spindle spectrograms of TC and Pyr neurons during pre-REM and REM sleep. Half-transparent shades represent the standard deviation of time-varying events across 1000 trials. Yellow shades mark the DOWN→UP state transition in the cortex. (C) Comparison of normalized power spectrum for all the conditions. Power spectrum are computed by an average across time and normalized by frequency-wise standard deviation. Shades reflect variability across 1000 trials (trial-wise standard deviation).

types of spindles presented in the original model paper [38]. The model parameters to generate the spindles are re-tuned (thus slightly different from the original model) to suppress PGO-triggered spindle-like oscillations (see also Section 2.2.3.2).

For the PGO waves, the simulated results in the thalamus can capture many features of the experimental waveforms (compared to traces in e.g. [30]), such as the biphasic pattern consisting of the negative and positive deflections, as well as the stronger amplitude of the negative peak of the pre-REM subtype. Moreover, the duration of the simulated waveforms (approximately 500 ms) matches the experimentally recorded waveforms (e.g. in [30, 224, 170]). Spectrograms of the two subtypes in both the thalamus and the cortex also show that the pre-REM PGO waves are prominent in lower frequency bands compared to REM PGO waves.

In comparison, from both spectrograms and the averaged spectra (Figure 2.8C), spindles in the thalamus still oscillate in a higher frequency band than both subtypes of PGO waves, albeit both pre-REM PGO waves and Type-I spindles as thalamic inputs both trigger a transient DOWN state preceding a Down-to-UP transition in the Pyr neurons of the cortex. This observation, as well as the time scale (200-300 ms), matches well with experimental recordings showing that stimulation of the cortex triggers DOWN→UP transitions [135].

2.3.4 *Validation of Ponto-thalamic projections*

Essentially, reproduction of all the aforementioned electrophysiological characteristics largely depends on the parameter tuning of 5 ponto-thalamic projections (see Figure 2.1B) cooperatively transforming the pulse-like pontine outputs into the subtypes' specific waveforms in the thalamus. As explained in Section 2.3.4, we investigated whether all of these projections play a role in the waveform shapes by comparing our result to simulations resulting from suppressing one projection at a time.

We consider one projection necessary if the PGO waveforms generated with it can not be reproduced in its absence, regardless of the variation of the other projections (see Section 2.2.5.2). In practice, this can be verified by using a dimensionality reduction technique mapping each shape to a point in a 2-dimensional space, and checking whether the waveforms simulated with and without each projection map to distinct regions of this space.

We thus complement the dataset of the original PGO waveforms generated with full ponto-thalamic projections (which underlies the Figure 2.6B and Figure 2.8A) with 5 datasets of 500 substitute PGO waveforms by blocking each of the ponto-thalamic projections. The blockade is implemented by setting the corresponding projection strength to zero, while the strength of unblocked projections is set to randomly deviate maximally 100% from the optimized values. We assume that the range of parameter variations is large enough to cover most of the parameter space for the generation of PGO waveforms.

From these 6 datasets, we take as features the waveforms of population membrane potential, firing rates and LFPs, separately for pre-REM and REM states. We reduce the high-dimensional features into 2 dimensions by applying T-SNE which has been shown to perform well on dimension reduction problems of time series [146]. The point clouds in Figure 2.9 show the REM features of the optimized PGO waveforms with full projections are clearly separated from the 5 blockade conditions, although not during

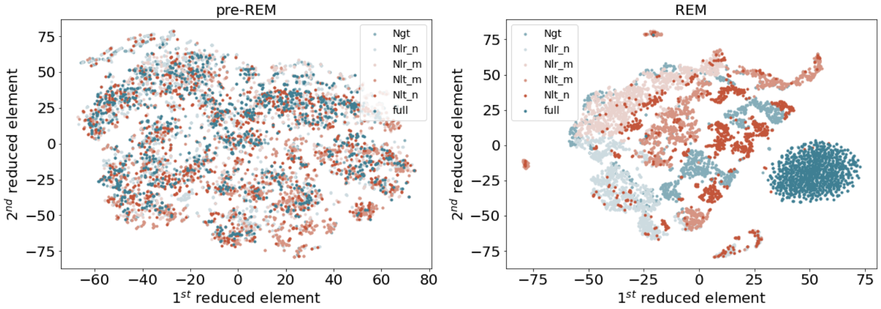

Figure 2.9: Selective blockade of each ponto-thalamic projection alter PGO waveforms. The two subplots compare the separability of dimension-reduced features of optimal PGO waves (dark cyan) and substitute ones obtained by blocking one ponto-thalamic projection at a time (the blocked ones marked in the legend) in the simulated pre-REM (right panel) and REM stages (left panel). Substitute PGO waves are generated with large noise in the projection strengths to cover the large parameter space. Notably, in the REM stage, the isolated cluster of original PGO features shows that the model is not able to generate the optimal PGO waveforms without any of the ponto-thalamic projections, suggesting the necessity of each projection.

pre-REM states, which is enough to indicate that each of the ponto-thalamic projections has a specific role in the generation PGO-related waveforms.

2.3.5 *PGO-triggered cortical plasticity, compared to spindle-triggered plasticity*

After this comprehensive validation of the model's ability to reproduce many electrophysiological aspects, we implement the population STDP rule introduced in Section 2.2.6 on the PGO-triggered cortical activities to investigate potential plastic changes PGO waves trigger in the cortical circuit to shed light on the synaptic rescaling problem.

Different subtypes of PGO waves trigger opposite synaptic effects

The STDP rule characterizes the similarity between the STDP function (Figure 2.4B) and the cross-correlation between the pre-synaptic AMPA-modulating currents and post-synaptic membrane potentials of the Pyr neurons. As results (Figure 2.10A), we find that the synaptic change is different when they are triggered by different PGO wave subtypes. The strength of the intra-cortical excitatory synapse rises sharply during pre-REM PGO waves while increases slowly during no-PGO periods, suggesting that pre-REM PGO waves induce LTP; the opposite effect in the right column implies that REM PGO waves elicit Long-term Depression (LTD) in the same synapse.

To interpret the differentiated plastic effects induced by the two PGO subtypes, the cross-correlation function between the pre- and post-synaptic activities exploited by the STDP rule implementation is shown in Figure 2.10B. The similar location of the maximum cross-correlation lag, achieving a negative value for both subtypes, reflecting that the considered pre-synaptic AMPA current is the post-synaptic firing rate convolved with a causal alpha function modelling the synaptic dynamics, as reflected by both subfigures of Figure 2.10B. However, other characteristics of the cross-correlation differ between the pre-REM and REM PGO waves: the cross-correlation in pre-REM is much flatter than during REM, explaining the signed difference of plasticity. Indeed, the flat pre-REM cross-correlation function implies that Eq. 2.30 is approximately proportional to the integral of the STDP function

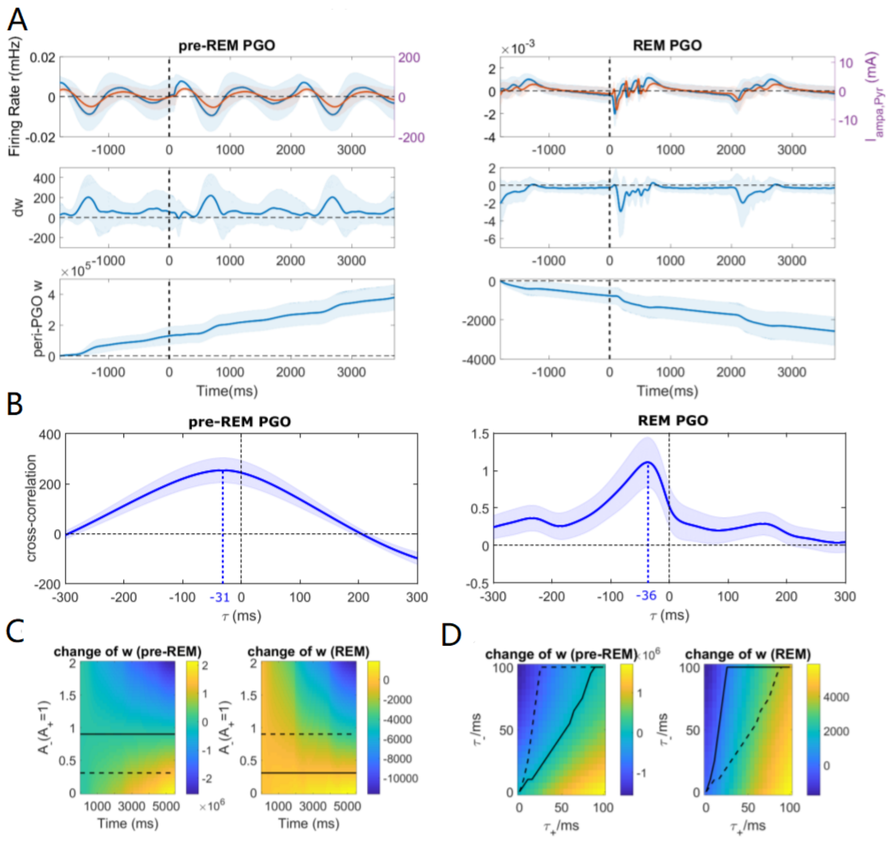

Figure 2.10: PGO-triggered cortical plasticity. (A) Smoothed change of synaptic strength in the intra-cortical excitatory synapse evoked by two subtypes of PGO waves. (top) demeaned waveforms of pre-synaptic current (red) and post-synaptic firing rate (blue). (middle) time-varying change of synaptic strength. (bottom) synaptic strength of the intra-cortical excitatory synapse changing with time. (B) Comparison of cross-correlation between the pre-synaptic current and the post synaptic activities. Light blue shade represents standard deviation at each lag for all the trials calculated. (C) Effect of STDP parameter A^- on the plasticity direction induced by two PGO subtypes. The color bar indicates the relative increase/decrease of synaptic strengths across time. Solid lines mark the critical value of the parameter that switches plasticity direction, i.e. potentiation v.s.de-potentiation; dashed lines correspond to the critical values of other subtypes. (D) Effect of STDP parameters τ_+ and τ_- on the plasticity direction induced by two PGO wave subtypes. Color bars and lines are analogous to (C).

of Figure 2.4B. Because this function has a larger positive area (in the positive lags) than the negative, this leads to an overall potentiating effect. For the REM case, the sharp peak of the cross-correlation function at small negative lags puts more weight in the negative portion of the STDP function of Figure 2.4B in Eq. 2.30, resulting in an overall depressing effect on the synapse.

parameter sensitivity

We next investigated the sensitivity of these results to the parameters of the STDP window: the positive amplitude A^+, the negative amplitude A^- and the time constants τ_+ and τ_-. Considering that the sign and relative magnitude of plastic changes being unchanged by an overall rescaling of the STDP function, we fix the positive amplitude to a constant $A^+ = 1$ for this analysis. Figure 2.10C, 2.10D and 2.11 show how the switch of LTP/LTD induction is modulated by changing these parameters.

Figure 2.10C characterizes how synaptic strength change over time with different values of the negative amplitude A^-. The calculation of the change of synaptic strength in Figure 2.10A is performed for different values of the negative amplitude $A^- = [0, 2]$ with a step of 0.05. The results are directed plotted as rows of the heatmap. The solid line marks the value of A^- which leads to invariant synaptic strength over time for each subtype of PGO waves.

Figure 2.10D plots the change of synaptic strengths along with the time interval [0,1200] ms against the two time constants, whose values are scanned in the range of [0, 100] ms with a step of 5 ms. During the scanning, the negative amplitude A^- is kept constant (at the value of 0.75). Solid lines mark the values of time constants ensuring no change is triggered by one PGO wave of the corresponding type; while dashed lines mark the other type. The parameter region surrounded by the two lines, which we referred to as "common region", are those that guarantee the strengthening of synapses triggered by pre-REM PGOs and weakening by REM PGOs. Interestingly, the biologically-measured STDP function parameters measured by classical studies [13] fall into this region ($A^+ = 1$; $A^- = 0.75$, $\tau_+ = \tau_- = 20$ ms).

Figure 2.11 scans all the three parameters, i.e. the negative amplitude and both time constants. A^- is scanned in the range of [0,2] with a step of 0.25. The time constants are still scanned in the range of [0,100] ms with a step of 5 ms. For each value of the negative amplitude, we obtain the results similar to what are presented in Figure 2.10D, but only plot the common region in the figure to show that there could be a shared parameter set that satisfies both LTP in the pre-REM stage and LTD in the REM stage. This further indicates the robustness of our result, as common sets of biologically meaningful parameters support our conclusion for both PGO wave subtypes.

Type-I spindles is accompanied by the co-occurrence of an SO while Type-II spindles are isolated

In order to interpret the dynamics of other memory-relevant sleep events from a cortical plasticity perspective, we also applied the same STDP framework to the two subtypes of spindles simulated during the NREM stage (Figure 2.12). Type-I spindles induce LTP-like behavior in the cortico-cortical connections, which makes sense as the peri-spindle neurons' activities exhibit a similar pattern as pre-REM PGO waves. More interestingly, Type-II spindles lead to a much weaker (see axis scale) temporal increase of the synaptic strength before it returns to a lower level. This analysis supports that Type-II spindles induce a weaker LTP than type-I. This is in agreement with experimental calcium imaging data (introduced in Section 1.1.3.2) indicating that the co-occurrence of spindles and SOs, corresponding to Type-Ispindles, leads to stronger increases of dendritic calcium in cortical neu-

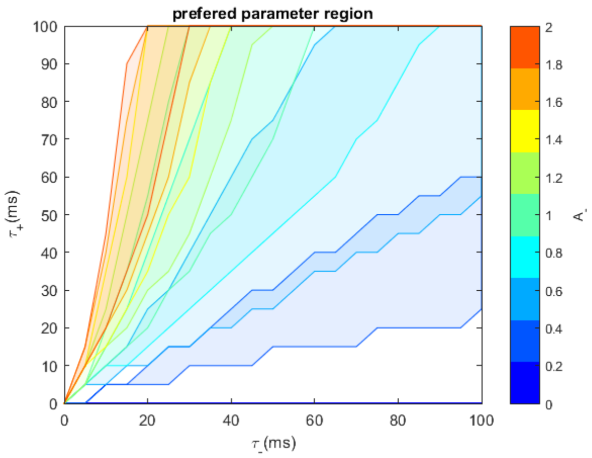

Figure 2.11: Parameter sensitivity of PGO-triggered plasticity. 3 parameters are scanned at the same time (for details see text). The common regions as in Figure 2.10D are plotted in different colors for different values of the negative amplitude A^-.

rons, which in turn are prone to trigger large plastic changes, compared to isolated (Type-II) spindles [173].

2.4 DISCUSSION

In this chapter, we have shown that an acetylcholine-modulated ponto-thalamo-cortical neural mass model spanning three sleep stages - NREM, pre-REM and REM - can reproduce a series of electrophysiological features associated with PGO waves, including the differences in firing rate patterns and LFP waveforms of two PGO wave subtypes. Analysis of cortico-cortical plasticity associated with these events, as well as spindles, suggests a sleep-stage dependent role: NREM and pre-REM sleep events inducing long-term potentiation, while REM events lead to long-term depression.

The choice of a neural mass model helps to maintain the realism of the model by trading off the complexity that would have incurred using individual units. It possesses the ability to account for detailed properties of synaptic and intrinsic currents based on quantitative experimental results. In particular, we show that the acetylcholine-modulated state-switching conductance and the five ponto-thalamic cholinergic projections are the key to the generation of PGO-triggered neural activities. It is worth noting that among these projections, the results support the importance of including a population of local geniculate interneurons receiving pontine input, which is often neglected in previous thalamocortical models [167].

This is to the best of our knowledge the first computational model accounting for the detailed cellular mechanisms of thalamic PGO wave generation, clarifying details in the classical electrophysiological literature [103, 43] as well as paving the way for future single unit models of PGO waves. The STDP framework for neural field theory, which has proved effective in characterizing TMS-induced plasticity [250, 249], is extended here to neural mass models deprived of spatial structure. The STDP rule, computed with parameters estimated experimentally in-vivo [13], reveals an opposite plasticity effect of two subtypes of PGO waves - potentiation for pre-REM PGO waves and depression for REM PGO waves.

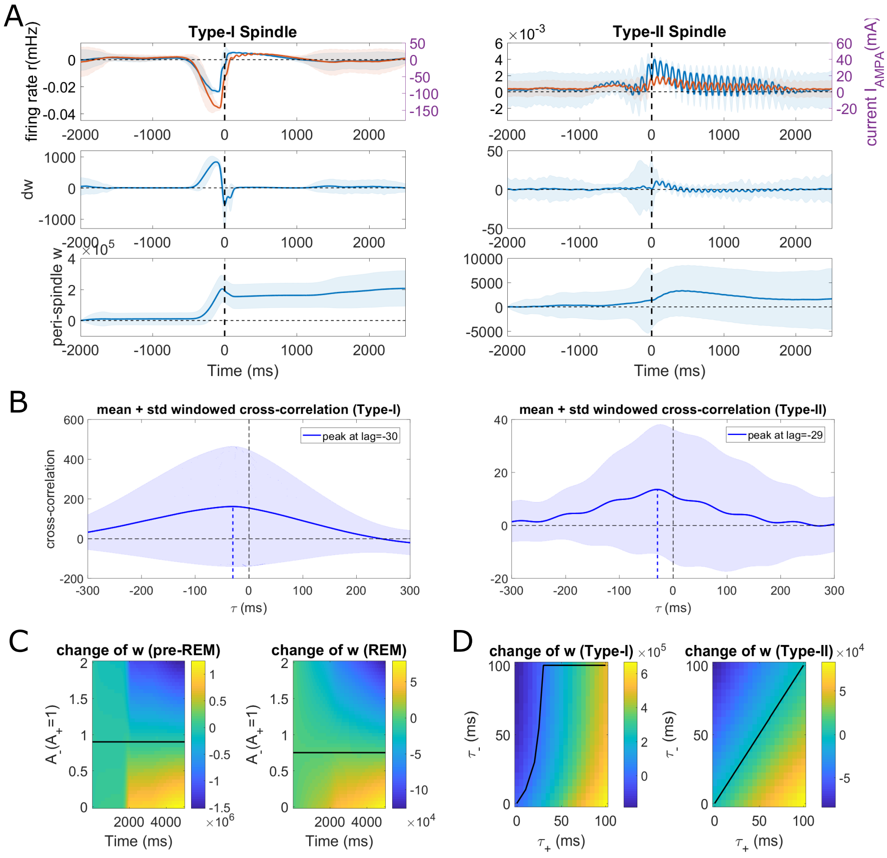

Figure 2.12: Spindle-triggered cortical plasticity. (A) smoothed change of synaptic strength induced by two subtypes of spindles. (top) demeaned waveforms of pre-synaptic current (red) and post-synaptic firing rate (blue). (middle) time-varying change of synaptic strength. (bottom) synaptic strength of the intra-cortical excitatory synapse changing with time. (B) Comparison of cross-correlation between the pre-synaptic current and the post synaptic activities. Light blue shade represents standard deviation at each lag for all the trials calculated. (C) Effect of STDP parameter A^- on the plasticity direction induced by two spindle wave subtypes. The color bar indicates the relative increase/decrease of synaptic strengths across time. Solid lines mark the critical value of the parameter that switches plasticity direction, i.e. potentiation v.s.de-potentiation; dashed lines correspond to the critical values of other subtypes. (D) Effect of STDP parameter τ_+ and τ_- on the plasticity direction induced by two spindle subtypes. Color bars and lines are analogous to (C).

These results support the ability of PGO waves to up- and down-scale cortical synaptic weights in larger proportions than the baseline activity surrounding them. This makes PGO waves the candidates to enforce synaptic homeostasis, a role that spontaneous phasic sleep events are hypothesized to play in line with considerable experimental evidence [173, 90]. Our results for REM PGO waves, for which the contribution to synaptic homeostasis remained elusive (Section 1.1.4.2), is in line with the downscaling effect of REM sleep on cortical spike rates observed in rats in [241], and possibly matches the eliminated spines during REM sleep in the mouse reported by [253]. In contrast, the potentiation effects that we found for pre-REM PGO waves, although apparently contradictory to the classically attributed depotentiating role of NREM sleep, is still consistent with results in [241], as such downscaling may be affected only a specific subgroup of synapses, with the largest synaptic weights in the population, while other subgroups may undergo potentiation to ensure consolidation of newly acquired memories (as predicted in [136]).

Indeed, converging evidence suggests that interacting NREM events may be involved in consolidating specific memory traces into local neural microcircuits, with the SO-ripple coupling a specific example [147, 128, 122]. Interestingly, as reported by Ramirez-Villegas et al. [192] (see also Section 1.1.4.3), pre-REM PGO waves co-occur with SPW-Rs, while our results exhibit a co-occurrence of pre-REM PGO waves and SOs (specifically DOWN→UP state transitions). According to synaptic plastic pressure theory proposed by [136], ripples, associated with off-line reactivation of episodic memories, occurring during DOWN→UP state transitions provide an opportune time window for new memory traces with low average firing rates to get assimilated into the existing network with relatively high firing rates. Pre-REM PGO waves may serve as triggers and facilitators of this procedure.

Overall, our results suggest the variability of neuronal patterns observed during sleep, may serve the purpose of differentially affecting the plasticity of network elements. The multiphasic nature of these events and their grouping in time may put the network in a dynamic state that ensures the right synapses can be targeted for plastic changes despite the strong recurrence of the microcircuit they are embedded in. The idea that specific events are optimal to trigger certain plasticity mechanisms resonates with recent experimental results showing spindles result in a targeted increase of calcium activity in cortical dendrites during NREM sleep, suggesting that these events are able to trigger dendritic depolarization independent of somatic activity [210]. Our results suggest further experimental work should be conducted on the characterization of transient REM activity, and parallel with a detailed modelling of the optimality of transient events for triggering synaptic plasticity processes in recurrent microcircuits.

3 ESTIMATION OF SPONTANEOUS TRANSIENT DYNAMICS BASED ON PERI-EVENT DATA

3.1 INTRODUCTION

As stressed in Section 1.1.2, implementation of synaptic scaling and other network changes related to memory consolidation and homeostasis likely relies on the occurrence of transient mechanisms during sleep. From a complex systems perspective, such a mechanism can be seen as putting the brain network in a transient physiological *state*, where it exhibits specific dynamical behavior and undergoes critical network reorganizations. An essential feature of these states and the underlying mechanisms are their spontaneity, as they are not triggered by an observable external input but instead result from the internal dynamics of the system. Functional recording giving us only a very partial observation of the network dynamics, inferring the properties of the underlying brain state favoring the occurrence of specific neural events is thus a challenging task. One important clue in order to address this question is the transient brain activity surrounding neural events presumed to hallmark the state. These events are typically detected as repetitive patterns of activity using classical filtering (e.g. [193]) or template matching approaches (e.g. [208]).

transient mechanism, transient state, and transient events

Understanding the emergence and dynamics of such transient phenomena in complex systems like the brain is a key challenge in neuroscience, where models are broadly used to investigate the underlying mechanisms. As mentioned in Section 1.2, biophysical modeling can be applied to systems whose operating mechanisms are relatively clear. An example presented in Chapter 2 is the generation of cortical K-complex modelled with a neural mass framework, where bifurcation analysis of the model reveals a canard explosion caused by random perturbation to the system [242, 39]. However, exploiting observation data to inform such modeling in a principled way remains largely elusive.

The critical issue regarding a statistical investigation of transient mechanisms is how to appropriately capture the dynamical properties of a state giving rise to specific events. Specifically, we are interested in learning a dynamical law that determines the future activity of the system based on past values. This state dependent model of the evolution of the system may be used to characterize key properties of the underlying network in a given state, such as causal interactions investigated in Chapter 4.

Classically, the analysis of neural dynamics associated with neural events is based on an empirical detection followed by reporting “event-triggered” averages (see e.g., [140, 229, 144]). This relies on collecting a panel of “peri-event” sequences (i.e., comprising a “peri-event time” dimension, and a “trial” dimension indexing the set of detected events) to perform advanced analysis, such as phase-locking values [4] and Granger causality [89, 212].

However, this approach neglects the fact that the peri-event "trials" accumulated in this way are not from a randomized controlled trial, but instead are selected based on a specific signal detection procedure, and thus potentially subject to selection biases of the state (for elaboration refer to Section 3.2.1).

event selection bias problem

In this sense, other analysis methods developed based on stimulus-triggered activities, e.g., Dynamic Causal Modelling (DCM) in neuroimaging or panel data analysis approaches in econometrics, should not be directly applied to spontaneous activities without considering its specificity [75, 14, 101].

Therefore, it is imperative to attack the problem fundamentally by formalizing mathematically the processes underlying spontaneously emerging neural activity and the limitations they entail from the perspective of statistical data analysis. In this work, we explicitly model the whole event-triggered analysis procedure to emphasize the specific issues to pay attention to when exploiting such data for fitting statistical models (Section 3.2.2). After pointing out identifiability issues related to such an approach in a non-parametric setting (Section 3.2.3), we investigate the linear autoregressive Gaussian case for which classical estimation procedures are shown to be biased (Section 3.2.4). We then develop a bias correction procedure (Section 3.2.5) whose efficiency is illustrated on simulated data (Section 3.3.1 and Section 3.3.2) and further applied on neural recordings (Section 3.3.3). Section 3.3.3.2 shows that the de-biased power spectrograms of SPW-Rs is able to categorize SPW-R events defined in three frequancy bands into two groups associated with different states. Such categorization matches the experimental results introduced in Section 1.1.4.3.

3.2 METHODS

The Method section includes a comprehensive investigation of the event selection bias problem using methodologies ranging from signal processing, dynamical systems, structural causal models and probabilistic modelling. Starting from a motivating example, we will formalize mathematically the state-dependent event detection and dig into the bias problem progressively. Finally, we will propose a bias correction method - the *DeSnap* algorithm in Section 3.2.5.2 and summarize the main idea in Section 3.2.6.

3.2.1 *A motivating example for selection bias in peri-event data*

Assume we want to analyze certain transient network properties of a dynamical system. The true state dynamics is usually unobserved, and we need instead to explore the transient events occurring spontaneously in an observed stochastic process $\tilde{X}_t$ reflecting ongoing activity of the dynamical system. Examples of these transient events are the neural events during sleep, as reviewed in Section 1.1.2. Such investigation starts with an event detection step to find the location of putative transient events in the observed signals. Here we briefly describe the classical detection procedure and point out the potential issues mentioned in Section 3.1.

definition of key terminologies: detection signal, snapshots, reference points

The event detection is typically performed by applying a filter to the original signal to get a *detection signal*. The filtering procedure can be either a bandpass filter that captures frequency-based characteristics or a template-matching approach that extract events exhibiting specific waveforms. Afterward, by setting a threshold for the *detection signal*, one can locate the targeted events in the signal and obtain a multi-trial peri-event dataset by extracting signal sub-sequences in a time window surrounding the locations of each detected event occurrence.

for an intuitive understanding of the naming of "snapshots" refer to Section 3.2.2.1

Such a dataset is also called *panel* data (in econometrics) or *snapshots*. In the context of this thesis, we refer to the two-way multi-trial peri-event dataset as a panel, while data at each peri-event time point from different

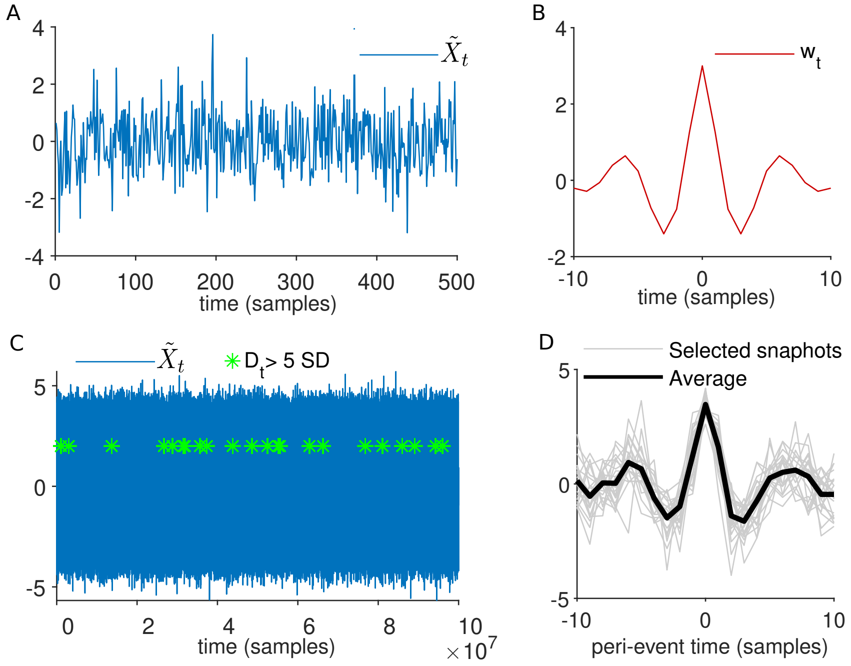

Figure 3.1: Illustration of the snapshot selection procedure on a white noise signal. (A) Time course of one realization of white noise. (B) Template used for detecting events. (C) Detected events for the same realization as (A), based on template matching with a detection threshold of 5SD. (D) Averaged panel of the detected events in the peri-event time course.

trials is called a snapshot. The detected event locations are referred to as *reference points*.

To illustrate the detection procedure, let us consider an example of event detection with a Morlet wavelet-like discrete-time template (exemplifying the detection of some oscillatory event)

$$w_t = \begin{cases} 3\exp(-|t|/4)\cos(t)\,, |t| \leqslant 10\,, \\ 0\,,\ \text{otherwise.} \end{cases}$$

Due to the symmetry of the template (i.e. it is an even function), we can implement the template matching procedure by computing a *detection signal* $\tilde{D}_t$ resulting from the convolution of this template with the observed time series $\tilde{X}_t$

$$\tilde{D}_t = \left(w * \tilde{X}\right)_t \,,$$

Here the tilted non-bold $\tilde{X}_t$ with the tilde refers to uni-variate observed time series (discrete) as shown in Figure 3.1A,C

and extract the peri-event *snapshots* in two steps: first, select the time points t_n $(n = 1\dots N)$ as *reference points* following a thresholding rule of the form

$$\mathcal{T} = \{t_n\} = \{t|\tilde{D}_t \geqslant d_0\}\,,$$

with, for example, d_0 chosen as a multiple of the standard deviation of the realization of $\tilde{D}_t$ computed across time; second, gather the two-way snapshot panel $\{X^{(n)}_{t'}\}(n = 1\dots N)$, representing the peri-event signal on an time window $\mathcal{I} = [-T/2, T/2]$ (with duration T) in the neighborhood of each *reference point* such that

Here the untilted non-bold X_t without the tilde refers to uni-variate peri-event snapshots (discrete) as shown in Figure 3.1D

$$X^{(n)}_{t'} = \tilde{X}_{t'+t_n}, t' \in \mathcal{I}, t_n \in \mathcal{T}\,.$$

As mentioned in Section 3.1, we can investigate the statistical properties of such snapshots datasets, and in particular, how they can be used to accurately infer *state-dependent* properties of the original process $\tilde{X}_t$.

To illustrate a potential problem of such a detection procedure, we applied the above Morlet detector to a white noise signal made of *i.i.d.* normal samples (zero mean, unit variance) using a detection threshold of 5 SD (standard deviation). The original signal, Morlet template and resulting peri-event panel are provided in Figure 3.1A, B and D, showing that the resulting panel contains only peri-event signals very similar to the template.

While such a phenomenon is expected from a template matching approach, this also demonstrates that the selection of snapshots based on such procedure introduces a structure in $X_{t'+t_n}$ that is not related to the properties of the completely unstructured (*i.i.d.*) original time series $\tilde{X}_t$.

In the next section, we will provide a snapshot analysis framework to shed light on the source of selection bias in the theory of dynamical systems.

3.2.2 *Snapshot analysis framework*

We expose here informally a continuous-time framework to justify intuitively the snapshot analysis of transient events and show how it leads to the modelling of peri-event snapshots as a time-varying difference equation.

The main idea of the framework is the following. Transient interactions spontaneously emerging within the system can be modeled by restricting the analysis to particular regions of its state space. We assume that a given type of neural event is associated to a single specific region of the state space favoring their emergence. The dynamics of hidden states in this region is inferred by collecting multiple "trials" that each comprises the sequence of measurements recorded from the system during one occurrence of the targeted type of neural event. We assume these trials correspond to portions of state space trajectories passing through the specific region of the state space where events are prone to emerge.

3.2.2.1 *Continuous time dynamics perspective*

This section describes our approach from a continuous time dynamical perspective that may help the readers more familiar with the investigation of complex systems with such tools. Readers less familiar with those may directly reach the next section describing our discrete time models.

Assume a deterministic continuous-time dynamical system governed by the autonomous differential equation in state space $\mathcal{Z}$

$$\begin{cases} \frac{d\mathbf{z}}{dt}(t) = \mathbf{F}(\mathbf{z}(t))\,, \\ \mathbf{z}(t_0) = \mathbf{z}_0\,, \end{cases} \tag{3.1}$$

where $\mathbf{z}(t)$ represents the state of the system at time t, and $\mathbf{z}_0$ denotes the initial state. Under mild assumptions, the flow of the vector field $\mathbf{F}$ provides the unique solution to this problem

$$\varphi(\mathbf{z}_0, t) = \mathbf{z}(t)\,,\ \mathbf{z}_0 \in \mathcal{Z},\ t \in \mathbb{R}\,,$$

satisfying the property

$$\varphi(\varphi(\mathbf{z}_0, t_1), t_2) = \varphi(\mathbf{z}_0, t_1 + t_2),\ t_1, t_2 \in \mathbb{R}\,.$$

As the states are usually hidden, we assume the observations of the system is denoted as a vector $\tilde{\mathbf{x}}(t)$. For a given event instance, the observations $\tilde{\mathbf{x}}(t)$ are deterministic functions of the current state

$$\tilde{\mathbf{x}}(t) = \tilde{f}(\mathbf{z}(t))\,.$$

Here the bold $\tilde{\mathbf{x}}(t)$ with the tilde refers to multi-variate continuous observations of the system as shown in Figure 3.2

We assume events are prone to occur when the state trajectory crosses a manifold $\mathcal{E}_0$ in the state space illustrated in Figure 3.2. The manifold is the state space representation of the transient mechanism we focus on. We further define $\mathcal{E}_t$ the images of this manifold corresponding to evolution of the system t time steps after crossing $\mathcal{E}_0$ (t can be positive or negative, leading to running the evolution backwards in time in the later case). Given the observation $\tilde{\mathbf{x}}(t)$, the deterministic mapping between two successive states (i.e., $\mathbf{z}(t)$ and $\mathbf{z}(t-1)$) implies that $\tilde{\mathbf{x}}(t)$ is also a deterministic function of the past state $\mathbf{z}(t-1)$.

Following the principle of the *Takens theorem* [231], information about $\mathbf{z}(t-1)$ can be gathered by collecting values of the observations at multiple lags k in the past $\tilde{\mathbf{x}}_{p,t} = \{\tilde{\mathbf{x}}(t-k)\}_{k=1..p}$. However, this information may remain incomplete, especially if the number of lags is small and the dimension of $\mathcal{Z}$ is large, which is likely the case for complex biological systems such as the brain. This would also be the case if we would have considered from the beginning an inherently stochastic dynamical system (governed by stochastic differential equations).

Under *ergodicity* and *mixing* assumptions for our dynamical system (see e.g. [127]), if the event occurs long enough after the initialization of the dynamics, $\mathbf{z}(t-1)$ is approximately distributed according to the invariant measure $\boldsymbol{\mu}$ of the system. As a consequence, it can be modeled as a random vector $\mathbf{Z}_{t-1} \sim \boldsymbol{\mu}$, and the knowledge of the vector of past observations $\tilde{\mathbf{x}}_{p,t}$ up to lag p reduces the uncertainty on the state through the conditional $\mathbf{Z}_{t-1}|\tilde{\mathbf{x}}_{p,t}$.

As shown in Figure 3.2, the deterministic (and invertible) mapping between $\mathcal{E}_{t-1}$ and $\mathcal{E}_t$ through φ leads to a stochastic model for the state $\mathbf{Z}_t|\tilde{\mathbf{x}}_{p,t}$ as well as current observations $\tilde{\boldsymbol{X}}_t|\tilde{\mathbf{x}}_{p,t}$. We can thus parameterize each conditional distribution as

Here the bold tilted $\tilde{\boldsymbol{X}}_t$ with the tilde refers to multi-variate discrete observed time series

$$\tilde{\boldsymbol{X}}_t|\tilde{\mathbf{x}}_{p,t} = \tilde{f}(\varphi(\mathbf{Z}_{t-1}|\tilde{\mathbf{x}}_{p,t}, 1)) = f_t(\tilde{\mathbf{x}}_{p,t}, \boldsymbol{\eta}_t) \tag{3.2}$$

where $\boldsymbol{\eta}_t$ models the randomness of $\mathbf{Z}_{t-1}$ due to the remaining uncertainty on the observations given $\tilde{\mathbf{x}}_{p,t}$, relevant to predict each observed variable and f_t models the time-varying deterministic mappings from the state $\mathbf{z}(t-1) \in \mathcal{E}_{t-1}$ to each observed variable. It is noteworthy that the mapping f_t is assumed time-dependent because the distribution of the random part of the state [1] is only dependent on the current location in the state space. We stress that the above framework remains largely informal as it overlooks many technical requirements for the final Eq. 3.2 to hold. This last equation, however, provides a connection between properties of the continuous time dynamical system and the time varying discrete time models of the observed time series introduced in the next section.

3.2.2.2 *Discrete-time Snapshot Model*

From Section 3.2.2.1, we see the interest of modelling time series data comprising transient events as a *state-dependent* time series, where the focus is

[1] if the state was fully observed, the mapping would be independent of time, because the autonomous differential equation (3.1)

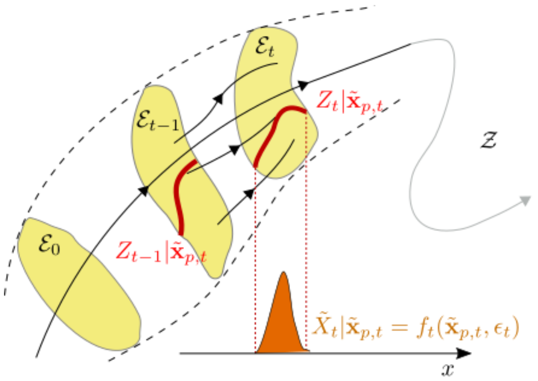
$\mathcal{E}_t$
$\mathcal{E}_{t-1}$
$\mathcal{E}_0$
$Z_t|\tilde{\mathbf{x}}_{p,t}$
$Z_{t-1}|\tilde{\mathbf{x}}_{p,t}$
$\mathcal{Z}$
$\tilde{X}_t|\tilde{\mathbf{x}}_{p,t} = f_t(\tilde{\mathbf{x}}_{p,t}, \epsilon_t)$
x

Figure 3.2: Interpretation of the peri-event analysis using a deterministic continuous time dynamical system.

put on location of the state space where the events emerge. Here we present how to treat the problem in a discrete-time setting, which directly applies to neural time series analysis.

Assuming the time interval between snapshots is sufficiently small, we make a linear approximation of Eq. 3.2, resulting in a linear Vector Autoregressive (VAR) models, for which coefficient estimation procedures are well-established (see Section 3.2.4.2). The VAR model describes time series by systems of difference equations linking future to past values and (potentially) the values of additional exogenous variables.

For the multivariate n-dimensional observation $\tilde{\boldsymbol{X}}_t$, the time-inhomo-geneous linear VAR model of order p takes the form:

$$\tilde{\boldsymbol{X}}_t = A_t\tilde{\boldsymbol{X}}_{p,t} + \boldsymbol{\eta}_t\,, \boldsymbol{\eta}_t \sim \mathcal{N}(\boldsymbol{k}_t, \Sigma_t)\,, \tag{3.3}$$

where $\tilde{\boldsymbol{X}}_{p,t} = \{\tilde{\boldsymbol{X}}_{t-1}, \cdots, \tilde{\boldsymbol{X}}_{t-p}\}$ collects past process values up to lag p as a single column vector, and $\{\boldsymbol{\eta}_t\}_{t\in\mathbb{Z}}$ is called the *innovation* process. Innovations at each time points are assumed to be jointly independent n-dimensional Gaussian random vectors. Moreover, the covariance Σ_t between the components is assumed diagonal. Time-inhomogeneity of this model reflects the time-dependence in Eq. 3.2 and manifests it self both in the time-dependence of the coefficient matrix A_t, as well as in the parameters of the mean and variance innovation distribution (i.e., $\boldsymbol{k}_t$ and Σ_t). Importantly, the VAR model entails additional assumptions allowing the estimation of the parameters from data. Chiefly, the independence assumption between innovations at different time points entails order-p Markovianity of the process, as the distribution of $\tilde{\boldsymbol{X}}_t$ given the whole history of the process up to time $t-1$ depends only on random variable $\tilde{\boldsymbol{X}}_{p,t}$. In relation to the above state space modelling perspective, this implies that the hidden state Z_t does not act as a hidden confounder of the dependency between successive time samples of the observations.

For a similar issue see Section 3.2.3.4

For a simplified representation of the state dependency of the overall dynamics of the system, we use Markov switching models that combine state dependency with VAR dynamics [92]. The Markov switching state Z_t is a discrete Markov chain with m-states and transition matrix M such that

$$p(Z_t = k|Z_{t-1} = j) = M_{k,j} \tag{3.4}$$

and this state controls the time varying parameters of the VAR model for the discrete time series $\tilde{\boldsymbol{X}}_t$ as observation

$$\tilde{\boldsymbol{X}}_t = A_{Z_t}\tilde{\boldsymbol{X}}_{p,t} + \boldsymbol{\eta}_k, \quad \boldsymbol{\eta}_t \sim \mathcal{N}(k_{Z_t}, \Sigma_{Z_t})\,. \tag{3.5}$$

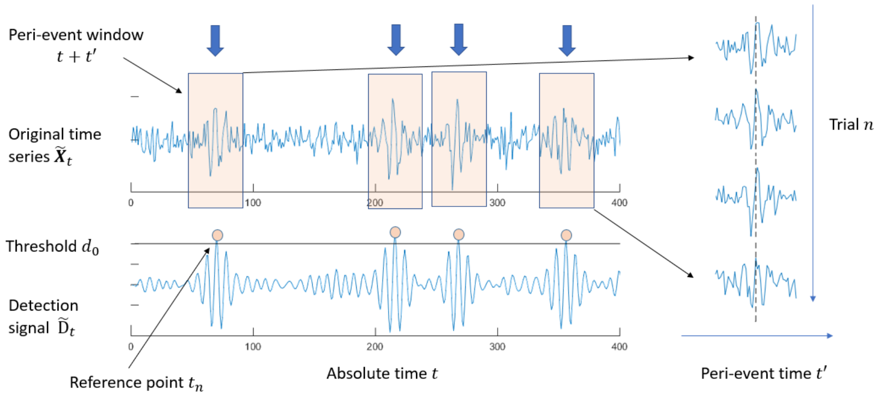

Figure 3.3: Illustration of the event detection procedure for a time series. An original signal $\tilde{\boldsymbol{X}}_t$ with Morlet-shaped events are plotted in blue. The detection signal $\tilde{D}_t$ is obtained by convolving $\tilde{\boldsymbol{X}}_t$ with a Morlet template. The threshold d_0=3SD is marked by the black solid line. *reference points* t_n such that $\tilde{D}_{t_n} \geqslant d_0$ are marked by pink dots. Peri-event data is marked by pink windows in $\tilde{\boldsymbol{X}}_t$ and extracted to form the peri-event panel on the left.

In relation the above discussion on Markovianity of the VAR model, we can see from these equations that the hidden state may affect the dependency between the innovations by influencing the parameters of their distribution at different time points. This violation of markovianity can however be neglected in practice when considering the state changes are small on the considered time intervals. We will make this approximation in our estimation procedure (see also Section 3.2.3.4).

3.2.2.3 *Modelling of peri-event snapshot detection procedure*

Based on this discretized model, we re-state the modelling of the peri-event snapshot detection procedure (as introduced in Section 3.2.1).

As illustrated in Figure 3.3, the detection is typically based on a continuous-value $\tilde{D}_t$ ascribed to each (discrete) time point t. To ease notations, we will consider a causal detector basing its decision on the last N_D samples, where $\tilde{\boldsymbol{X}}_{D,t} = \{\tilde{\boldsymbol{X}}_{t-1}, \cdots, \tilde{\boldsymbol{X}}_{t-N_D}\}$. Snapshots are extracted based on a deterministic detector function that extracts information from N_D past samples

$$\tilde{D}_t = w(\tilde{\boldsymbol{X}}_{D,t})$$

such that only the snapshot satisfying $D_t \geqslant d_0$ are kept, used as *reference points*

$$\mathcal{T} = \{t_n\} = \{t | \tilde{D}_t \geqslant d_0\},$$

The size of *reference points* is denoted as N such $n = 1 \ldots N$, also representing the number of event trials extracted from the time series.

state space understanding of detection

Such a detection procedure is equivalent to reconstructing the state space by delayed embedding of the observations $\tilde{\boldsymbol{X}}_t$ with the embedding dimension of N_D and delay time of $\tau = 1$ and transforming this N_D dimensional space into a 1-dimensional space of detection signal $\tilde{D}_t$. The *reference points* are detected by locating the time points in the embedded manifold passing one side of the embedded space separated by a hyperplane. The 1-dimensional space is orthogonal of the hyperplane such that it is mapped to the point $\tilde{D}_t = d_0$. This intuitive understanding is illustrated in Figure 3.4 as an example when the embedding dimension $N_D = 2$.

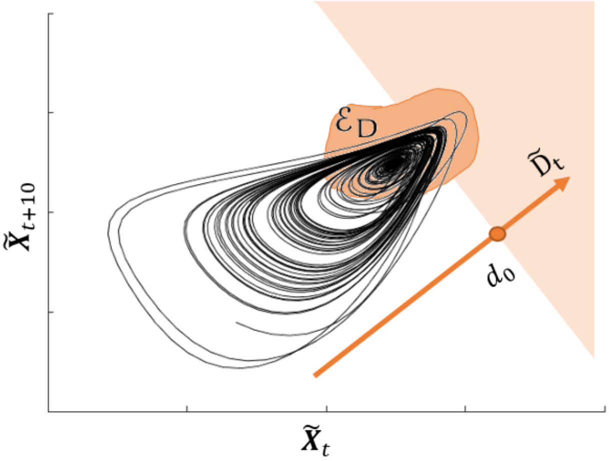

Figure 3.4: Illustration of detection in the embedded state space. Black curves marks the embedded manifold of a non-linear system $\tilde{\boldsymbol{X}}_t$ with the embedding dimension $N_D = 2$ and delay time $\tau = 1$. The orange region represents the embedded mapping of the manifold $\mathcal{E}_0$ in the hidden state space marking $Z_t = 0$. A hyperplaine is marked by the color separation, where the orange side is mapped to the detection criterion $D_t \geqslant d_0$ on the 1-dimensional space orthogonal to the hyperplane.

Detected states are a partial selection of all states in the region.

Consistent with Section 3.2.1, for each *reference point* t_n, the samples from the long time series $\tilde{\boldsymbol{X}}_t$ covering a fixed peri-event time window around t_n, i.e., $\mathcal{I} = [-T/2, T/2]$, are extracted to build a two-way panel $\{\mathbf{X}_{t'}^{(n)}\}$:

$$\mathbf{X}_{t'}^{(n)} = \tilde{\mathbf{X}}_{t'+t_n}, t' \in \mathcal{I}, t_n \in \mathcal{T}. \tag{3.6}$$

key assumption

The following assumption of "perfect detection" is key to our approach.

Assumption 1 (Perfect detection). *We assume that when* $\tilde{D}_t$ *is above a certain known threshold* d_0*, this indicates with probability one that the observed system is in a target state* $Z_t = 0$*, i.e.,* $P(Z_t = 0 | D_t \geqslant d_0) = 1$

With this assumption, for each *reference point* t_n, $P(Z_{t_n} = 0 | D_{t_n} \geqslant d_0) = 1$. Notably, this assumption provides only a sufficient condition to have $Z_t = 0$, but not a necessary one, i.e., $P(D_t \geqslant d_0 | Z_t = 0) \neq 1$. This assumption suggests that such thresholding detection is a partial selection of all the states $Z_t = 0$ thus leading to selection bias (e.g. see Fig. 3.4). As a consequence, collected snapshots at peri-event time t' are distributed according to $\tilde{\boldsymbol{X}}_{t'+t_n} | \tilde{D}_{t_n} \geqslant d_0$ which typically differs from $\tilde{\boldsymbol{X}}_{t'+t} | Z_t = 0$. Recovering this last distribution based on the snapshot panel data is the main goal of this chapter.

$\tilde{\boldsymbol{X}}_{t'+t} | Z_t = 0$ is the unbiased selection of peri-event snapshots

As the interpretation of such detection in the embedded space (e.g., Figure 3.4), a region in the embedded space $\mathcal{E}_D$ (marked by the orange manifold) can be understood as the reconstruction of the manifold $\mathcal{E}_0$ in the hidden state space where $Z_t = 0$ (Section 3.2.2.1). Thresholding detection with the hyperplane only selects the embedded states at one side of the hyperplane without covering the whole region $\mathcal{E}_D$. Thus similarly, the detected states are a partial selection of all desired states in the embedded space. This partial selection typically result from a necessary trade-off: lowering the threshold would cover a larger portion of $\mathcal{E}_D$, but would also cover other regions of the state space that we are not interested in, such that the perfect detection Assumption 1 would be violated.

3.2.3 *Selection bias based on Structural Causal Models*

After establishing the mathematical formalism of peri-event snapshots and their detection procedure, we know that the event selection bias observed in the motivational example (see Section 3.2.1) is due to partial selection of the states $Z_t = 0$ based on thresholding $\tilde{D}_t \geqslant d_0$.

This phenomenon can be stated as a form of sample selection bias, which has long been recognized as both a practical and fundamental issue [94]. It refers to the mechanisms of selection of empirically observed data points, intended or not, that may affect the inference of relevant quantities in statistical or causal models [95, 100].

Notably, this question has been investigated within the framework of Structural Causal Models (SCMs) [184], by using causal graphical models equipped with a special node representing the sampling process. In this section, we will first introduce the key concepts and properties of SCMs to prepare for a theory of SCM-based selection bias. Then we will introduce how to extend the theory to time series models, which includes the VAR models we derived in Section 3.2.2.2.

3.2.3.1 *Basics of Structural Causal Models*

SCMs are generalizations of Bayesian networks that combine Structural Equation Models (SEM) to incorporate directional information for causal analysis [184]. A structural equations takes the form

$$Y \coloneqq f(X_1, \cdots, X_k, \epsilon)$$

where the right hand side determines the assignment of values on the left-hand side. In the most usual case, Y and $\{X_j\}_{j\in\{1,\cdots,k\}}$ represent observed variables and ϵ a variable accounting for (unobserved) exogenous effects.

Based on this, a SCM is defined for a set of random variables $\{V_j\}$ associated to vertices in a graph as the follows.

Definition 1 (Structural Causal Model (SCM) (see e.g. [186])). *A d-dimensional structural causal model is a triplet $(\mathcal{S}, P_N, \mathcal{G})$ consisting of:*

- *a directed acyclic graph $\mathcal{G}$ with d vertices*
- *a set $\mathcal{S}$ of structural equations*

  $$V_j \coloneqq f_j(\boldsymbol{PA}_j, N_j), j = 1, \dots, d,$$

  where $\boldsymbol{PA}_j$ are the variables indexed by the set of parents of vertex j in $\mathcal{G}$
- *a joint distribution P_N over the exogenous variables N_j, which are assumed jointly independent.*

One attractive feature of this formalism is that the SCM's graph entails key properties of the join distribution of the nodes $\{V_j\}$, like the Markov properties and conditional independences (see e.g. [15]).

Proposition 1 (Markov properties). *For a given SCM $(\mathcal{S}, P_N, \mathcal{G})$, the joint distribution P_V is Markovian with respect to $\mathcal{G}$, i.e. it satisfies the following properties:*

1. *(local Markov property) each variable V_j is independent of its non-descendants given its parents $\boldsymbol{PA}_j$,*

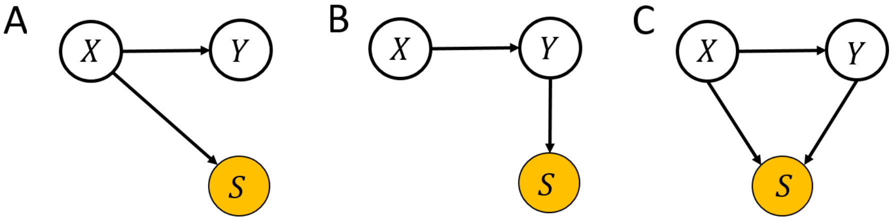

Figure 3.5: SCM, selection bias and recoverability (adapted from [5]). (A) SCM describing sample selection based on X, leading to identifiability of P(Y|X) based on selected data. (B) SCM describing sample selection based on Y, leading to non-identifiability of P(Y|X) based on selected data. (C) SCM describing sample selection based on both X and Y, leading to non-identifiability of P(Y|X) based on selected data.

2. *(Markov factorization property) assume the joint distribution* P_V *has a density, then*

$$p(\boldsymbol{v}) = p(v_1, \ldots, v_d) = \prod_{j=1}^{d} p(v_j|\boldsymbol{pa}_j)$$

With the conditional independence indicated in the local Markov property, the Bayesian network greatly simplifies the calculation of joint probabilities. In addition, the concept of d-separation allows assessing systematically the conditional independences between subsets of nodes in $\mathcal{G}$ based on graphical criteria of d-separation (see e.g. [184]).

Definition 2 (d-separation)**.** *A path* p *in graph* $\mathcal{G}$ *is said to be blocked by a set of nodes* Z *if either: (1)* p *contains a chain* $i \to m \to j$ *or a fork* $i \leftarrow m \to j$ *such that the middle node* m *is in* Z*, or (2)* p *contains a collider* $i \to m \leftarrow j$ *such that the middle node* m *is not in* Z *and such that no descendant of* m *is in* Z*.*

Z *is said to* d*-separate* X *from* Y *in* $\mathcal{G}$ *if and only if* Z *blocks every path from a node in* X *to a node in* Y*. This property is denoted* $X \perp\!\!\!\perp_{\mathcal{G}} Y|Z$*.*

Indeed, d-separation allows stating the *global Markov property* (see e.g. Peters et al. [186]).

Proposition 2 (Global Markov property)**.** *For a given SCM* $(\mathcal{S}, P_N, \mathcal{G})$ *and subsets of nodes* X*,* Y*,* Z *in* $\mathcal{G}$*, then*

$$X \perp\!\!\!\perp_{\mathcal{G}} Y|Z \Rightarrow X \perp\!\!\!\perp_{P_V} Y|Z\,.$$

This proposition indicates that the conditional independences in the graph as defined by d-separation rules also hold for the corresponding random variables of the associated SCM.

The next sections will show how these basic concepts and properties of a SCM would fascilitate the understanding of sampling bias.

3.2.3.2 *Recoverability with Sampling Selection Bias*

In the simplest two-node SCM, the identifiability or recoverability of the effect based on different sampling methods has been investigated in [5].

Figure 3.5A, B, C show three sampling conditions in a two-node SCM consisting of variables X and Y (with X causing Y). Sampling is represented by binary variable S in an additional node designed as descendant for either X or Y. S takes the value 1 when a data point is selected and zero otherwise. In Figure 3.5A, sample selection is a function of X only, while Figure 3.5B describes a sample selection based on Y only. Figure 3.5C presents the condition where sample selection depends on both variables.

In this model, we are interested in estimating the conditional propability of $P(Y|X)$ from sampled data. What is critical is whether $P(Y|X)$ can be recovered from the joint distribution of the selected samples $(X, Y)|S = 1$ given different sampling scenarios. Bareinboim et al. [5] show that, under standard assumptions, a necessary and sufficient condition for recoverability is conditional independence between target variable Y and selection variable S, given conditioning variable X ($Y \perp\!\!\!\perp S|X$)) such that $P(Y|X, S = 1) = P(Y|X)$. For the scenarios of Figure 3.5, this implies that $P(Y|X)$ can be recovered $(X, Y)|S = 1$ in the case of Figure 3.5A, but not in Figure 3.5B and Figure 3.5C.

The rationale is simple according to the d-separation rules (see Section 3.2.3.1 for details). In the condition of Figure 3.5A, conditioning on X corresponds to the "fork" case in the d-separation rules, indicating that the conditional independence $Y \perp\!\!\!\perp S|X$ is satisfied. On the contrary, Figure 3.5B shows the condition where such that $P(Y|X)$ is not recoverable from sample selected data because the above conditional independence requirement (Y independent of S given X) is not satisfied. For Figure 3.5B, detailed proof has been provided in Bareinboim et al. [5]. The case in Figure 3.5C corresponds to the "collider" case of d-separation where a common observed descendant induces extra dependency between the ancestors.

However, it is important to point out that this negative theoretical result corresponds to a non-parametric case. In particular, putting further assumptions on the model that generated X and Y may help identify $P(Y|X)$. We will return to this point at the end of Section 3.2.3.4, followed by the whole Section 3.2.5.2.

3.2.3.3 *SCMs for inhomogeneous VAR models*

The SCM perspective on time series models has been exploited in multiple studies (e.g. [185, 111, 68]) and is potentially helpful for investigating selection bias. This section will show why the VAR model can be treated as an SCM.

This is a general VAR model not specifically related to our above framework.

For a general form of VAR model with the stochastic process $\boldsymbol{X}_t$:

$$\boldsymbol{X}_t \coloneqq A_t \boldsymbol{X}_{p,t} + \boldsymbol{\eta}_t \,, \boldsymbol{\eta}_t \sim \mathcal{N}(\boldsymbol{k}_t, \Sigma_t) \tag{3.7}$$

As $\boldsymbol{X}_t$ at each time point is generated by its past $\boldsymbol{X}_{p,t}$, as indicated by our ":=" notation in Eq. 3.7, difference equations describing VAR model can be seen as a form of structural assignment of variables at time t based on variables at past times (implying acyclicity of the corresponding graph). In addition, the exogenous variables of each structural equation correspond to the components of the innovation vectors η_t. Since these components for all time points are jointly independent, all conditions of Definition 1 are satisfied and the time series can be considered as an SCM. Interestingly, the graph describing assignments at all times, called *full-time graph*, is potentially infinite.

The VAR modelling of our observational data has been defined in Eq. 3.3 for $\tilde{\boldsymbol{X}}_t$. An example graph of such a VAR model (of order 2) is represented in Figure 3.6A, where $\tilde{\boldsymbol{X}}_t$ is structurally assigned by the past two states. The corresponding SCM graph that incorporates the hidden state Z_t is illustrated in Figure 3.6B for the Markov switching VAR model defined in Eq. 3.5, where the hidden state Z_t is only dependent on its immediate past Z_{t-1}.

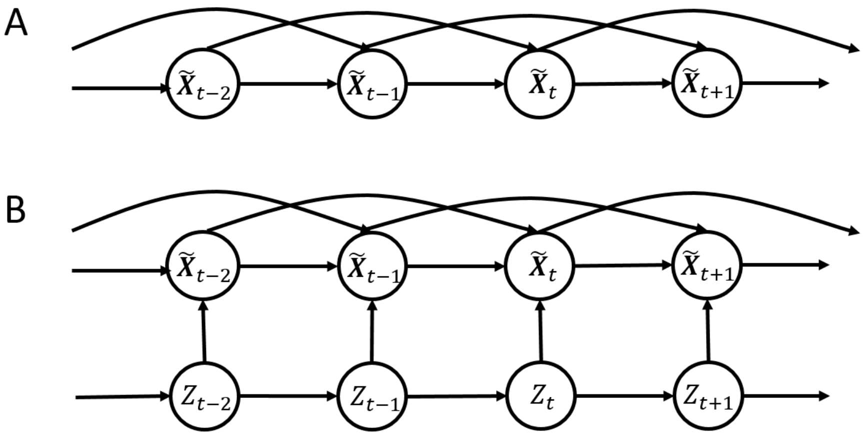

Figure 3.6: SCMs for inhomogeneous VAR models for observation signals. (A) SCM for an example inhomogeneous VAR(2) model of the observation data $\tilde{\boldsymbol{X}}_t$. The SCM is consistent with Eq. 3.3. (B) State-dependent SCM for the inhomogeneous VAR(2) model in (A) with the incorporation of hidden states Z_t. The SCM corresponds to Eq. 3.5.

For state-dependent peri-event data, the unbiased peri-event snapshots $\tilde{\boldsymbol{X}}_{t'+t}$ for the state $Z_t = 0$ can be obtained by gathering observation signals $\tilde{\boldsymbol{X}}_t$ for peri-event time $t' + t$ where $Z_t = 0$ and $t' = [-T/2, T/2]$ with the peri-event window T (see Section 3.2.2.3). As seen in Figure 3.7A, the SCM formalism for VAR model of the unbiased peri-event snapshots $\tilde{\boldsymbol{X}}_{t+t'}|Z_t = 0$ can be seen as conditioned on the yellow hidden state node for $t' = 0$ such that $Z_t = 0$.

To estimate the inhomogeneous VAR model, we are interested in obtaining an unbiased estimate of the conditional distribution $P(\tilde{\boldsymbol{X}}_{t'+t}|\tilde{\boldsymbol{X}}_{p,t'+t}, Z_t = 0)$, which is critical for the Maximum Likelihood (ML) estimation formulas (see Section 3.2.4.2 and Appendix A.3.1). In Figure 3.7A, at peri-event time $t' = 0$, as the node for the hidden state $Z_{t'=0}$ is observed due to Assumption 1 (and conditioned on), due to the Markov properties of SCMs, the past hidden states $Z_{t'}$ where $t' < 0$ are irrelevant to the estimation of the conditional $P(\tilde{\boldsymbol{X}}_{t'=1}|\tilde{\boldsymbol{X}}_{p,t'=0}, Z_{t'=0} = 0)$. However, at peri-event time $t' \neq 0$, the hidden states $Z_{t'+t}$ are not observed by conditioning on $Z_t = 0$ and therefore generate unblocked paths, e.g. from $\tilde{\boldsymbol{X}}_{t'=-4}$ to $\tilde{\boldsymbol{X}}_{t'=-1}$ through $Z_{t'=-4}$, $Z_{t'=-3}$, $Z_{t'=-2}$ and $Z_{t'=-1}$.

$\tilde{\boldsymbol{X}}_{t'=-4}$ is chosen as an example for the observations not included in $\tilde{\boldsymbol{X}}_{p,t'=-1}$ in the VAR(2) model

This illustrates the fact that the Markov properties cannot be, strictly speaking, satisfied for the graph where the hidden states are marginalized, represented in Figure 3.7B. Indeed, $\tilde{\boldsymbol{X}}_{t'+t}|Z_t = 0$ cannot be assumed only dependent on $\tilde{\boldsymbol{X}}_{p,t'+t}|Z_t = 0$ for arbitrary perievent times t'. However, we make the assumption that state varies only with low probability in the peri-event time window, such that the peri-event hidden states satisfy approximately $Z_{t'+t} = 0$ where $t' \neq 0$ (see also Section 3.2.2.2).

This approximation entails that all peri-event hidden states nodes $Z_{t'+t}$ where $t' \neq 0$ are observed, blocking the paths through hidden states. Therefore, we can approximate the SCM for peri-event snapshots into the form presented in Figure 3.7B such that each node represents the conditional peri-event variable conditioned on $Z_t = 0$ and the Markovianity is approximately satisfied.

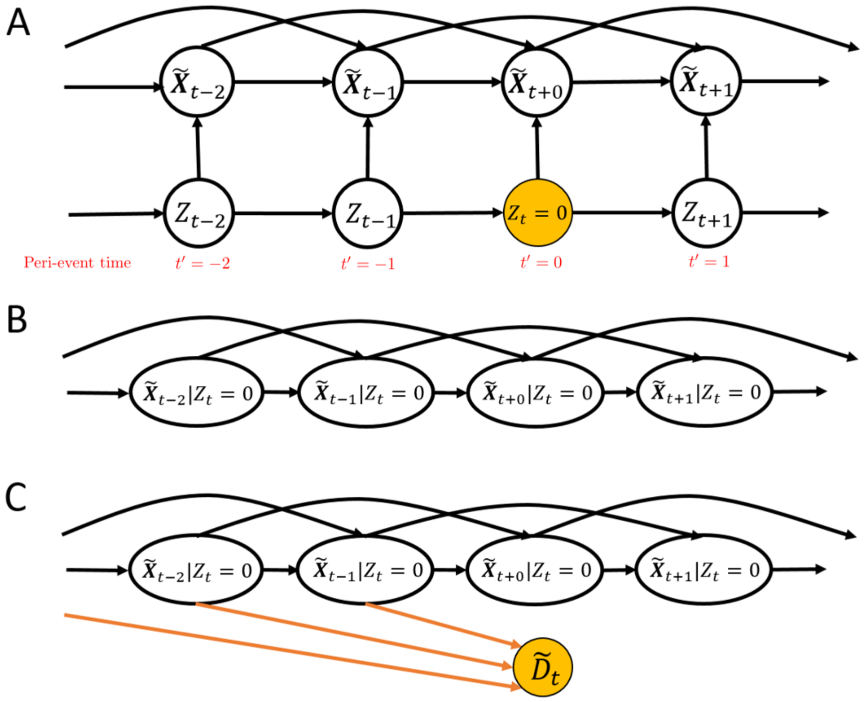
A
$\tilde{X}_{t-2}$
$\tilde{X}_{t-1}$
$\tilde{X}_{t+0}$
$\tilde{X}_{t+1}$
Z_{t-2}
Z_{t-1}
$Z_t = 0$
Z_{t+1}
Peri-event time
$t' = -2$
$t' = -1$
$t' = 0$
$t' = 1$
B
$\tilde{X}_{t-2}|Z_t = 0$
$\tilde{X}_{t-1}|Z_t = 0$
$\tilde{X}_{t+0}|Z_t = 0$
$\tilde{X}_{t+1}|Z_t = 0$
C
$\tilde{X}_{t-2}|Z_t = 0$
$\tilde{X}_{t-1}|Z_t = 0$
$\tilde{X}_{t+0}|Z_t = 0$
$\tilde{X}_{t+1}|Z_t = 0$
$\tilde{D}_t$

Figure 3.7: SCMs for VAR model of peri-event snapshots. (A) An SCM for VAR model of peri-event snapshots gathered from the full-time SCM in Figure 3.6B by setting the *reference points* as $\{t|Z_t = 0\}$. (B) The SCM in (A) can be approximated into a conditional graph when the peri-event states $Z_{t+t'}$ are of high probability to be $Z_{t+t'} = 0$. (C) An SCM illustrating the detection of peri-event snapshots in (B) based on past N_D states. The yellow node marks the detection node.

3.2.3.4 *Event detection generates selection bias in time series models*

This result presented in Section 3.2.3.2 can be extended to time series models, explaining why the event detection procedure modelled in Section 3.2.2.3 induces selection bias. Here we directly apply the recoverability theory of the two-node SCM model in Section 3.2.3.2 to the simplified peri-event SCM model shown in Figure 3.7B.

The detection of peri-event snapshots based on the *detection signal* $\tilde{D}_t$ is equivalent to adding a node for detection based on the past N_D nodes (Figure 3.7D). When peri-event snapshots are obtained by applying the detection procedure to a state-dependent VAR model such as the Markov switching model of Eq. 3.5, we are interested in using the resulting distribution of snapshot panel data conditioned on $\tilde{D}_t$, i.e., $P(\tilde{\mathbf{X}}_{t'+t}|\tilde{\mathbf{X}}_{p,t'+t}, Z_t = 0, \tilde{D}_t \geqslant d_0)$ to recover the conditional probability characterizing the markovian dynamics $P(\tilde{\mathbf{X}}_{t'+t}|\tilde{\mathbf{X}}_{p,t'+t}, Z_t = 0)$ for t' in a peri-event time window. For a better understanding, we point out here that the time points t in the conditional $P(\tilde{\mathbf{X}}_{t'+t}|\tilde{\mathbf{X}}_{p,t'+t}, Z_t = 0, \tilde{D}_t \geqslant d_0)$ is the same as t_n, indicating a *reference point* where the event is detected (Section 3.2.2.3). Here we omit the index n to see clearly the comparison between the two conditionals.

In this context, comparing the graphical relationship between $\tilde{\mathbf{X}}_{t'+t}$ and $\tilde{\mathbf{X}}_{p,t'+t}$ to the relationship between Y and X in Section 3.2.3.2, we can conclude easily that $P(\tilde{\mathbf{X}}_{t'+t}|\tilde{\mathbf{X}}_{p,t'+t}, Z_t = 0)$ is identifiable from the snapshot data $P(\tilde{\mathbf{X}}_{t'+t}| \tilde{\mathbf{X}}_{p,t'+t}, Z_t = 0, \tilde{D}_t \geqslant d_0)$ for points after the detection time point $t' \geqslant 0$ due to d separation. Similarly, based on the case presented in Figure 3.5B,C, $P(\tilde{\mathbf{X}}_{t'+t}|\tilde{\mathbf{X}}_{p,t'+t}, Z_t = 0)$ are not identifiable for time points before the detection time point, i.e., $t' < 0$.

This theoretical result provides insights about challenges for identifying the markovian dynamics of the system due to the selection process of peri-event snapshots for generic time series models. However, this does not preclude, a priori, that enforcing more assumptions on the model would lead to identifiability for a broader range of time points. In the following, we use a parametric (linear Gaussian) setting to estimate model parameters based on such detection in the inhomogeneous time setting.

3.2.4 *VAR model estimation for peri-event snapshots*

Having clarified the theory of selection bias in peri-event snapshot detection and investigating the nature of such bias with SCMs, we will next explore how to correct it. As a prerequisite, we first introduce a general statistical estimation framework for inhomogeneous VAR models based on peri-event snapshots.

Here the bold tilted $\boldsymbol{X}_t$ without the tilde refers to multi-variate random variables undelying a stochastic process

Repeating the VAR model in Eq. 3.7, we see

$$\boldsymbol{X}_t \coloneqq A_t \boldsymbol{X}_{p,t} + \boldsymbol{\eta}_t \, , \boldsymbol{\eta}_t \sim \mathcal{N}(\boldsymbol{k}_t, \Sigma_t) \tag{3.8}$$

The model is determined by the model order p and 3 model parameters: the coefficient matrix A_t, the innovations mean $\boldsymbol{k}_t$ and the innovations covariance Σ_t.

3.2.4.1 *VAR model order selection*

The determination of model order p falls into the category of classical model selection. If the model order is too low, the corresponding VAR(p) model will not be able to reconstruct enough past dynamics in the signal. On the contrary, a VAR model with higher order will over-fit the signals. In both cases, the estimation of residuals, thus the estimation of innovations' mean and covariance, will be inaccurate, such that spurious results might be obtained in subsequent analyses (e.g. causality analysis, see Section 4.2.2.1).

Two common ways to optimize the model order are the Akaike information criterion (AIC) [1] and the Bayesian information criterion (BIC) [168]. They both introduce a penalty term in the log-likelihood function to compensate for the effect caused by over-fitting with over-complex models:

$$\mathrm{IC}(p) = -\log(\mathcal{L}(p)) + \mathcal{P}(p)\,. \tag{3.9}$$

The model order is selected as the order that minimizes the information criterion. The penalty term $\mathcal{P}(p)$ involves the effect of model complexity by punishing on the number of parameters, which scales as pd^2 for a d-dimensional VAR(p) models. For AIC the penalty term is just the proportional to the number of parameters, scaling again as $\mathcal{P}(p) = pd^2$. BIC takes into account the effect of sample size $T - p$ as the log-likelihood function also increases with the number of samples, and the corresponding penalty term is $\mathcal{P}(p) = \frac{1}{2}pd^2\log(T-p)$ (for derivation see Appendix A.3.4).

A critical issue that is often overlooked is the model order selection for the multi-trial case. The difficulty is to decide what is the equivalent number of parameters and equivalent sample size in the multi-trial and non-stationary case. Despite previous attempts (for a review see Appendix A.3.2.1), we propose here an extended version of BIC that is appropriate for non-stationary signals with multi-trial structures: for the multi-trial inhomogeneous case as in Eq. 3.8, the penalty term should be $\mathcal{P}(p) = \frac{1}{2}Tpd^2\log(N)$ (for an elaborated proof see Appendix A.3.4).

3.2.4.2 *Estimation of VAR parameters*

Besides the optimization of VAR model order, the model parameters should be estimated consistently from the empirical peri-event snapshots.

Mathematically, with N *i.i.d.* samples $\{\boldsymbol{X}_t^{(n)}\}(n = 1...N)$, the coefficient matrix A_t can be estimated as a function of two covariance matrices (for a derivation of Ordinary Least Square (OLS) estimation or an ML estimation see Appendix A.3.1):

Here the $\{\boldsymbol{X}_t^{(n)}\}$ are not confined to long time series or peri-events snapshots

$$\widehat{A}_t = \widehat{\Sigma}_{\boldsymbol{X}_t\boldsymbol{X}_p}(\widehat{\Sigma}_{\boldsymbol{X}_p})^{-1} \tag{3.10}$$

where the covariance matrices are estimated from the N-sampled data as:

$$\widehat{\Sigma}_{\boldsymbol{X}_t\boldsymbol{X}_p} = \frac{1}{N}\sum_{n=1}^{N}(\boldsymbol{X}_t^{(n)} - \widehat{\mathbb{E}}[\boldsymbol{X}_t])(\boldsymbol{X}_{p,t}^{(n)} - \widehat{\mathbb{E}}[\boldsymbol{X}_{p,t}])^\top\,, \tag{3.11}$$

and

$$\widehat{\Sigma}_{\boldsymbol{X}_p} = \frac{1}{N}\sum_{n=1}^{N}(\boldsymbol{X}_{p,t}^{(n)} - \widehat{\mathbb{E}}[\boldsymbol{X}_{p,t}])(\boldsymbol{X}_{p,t}^{(n)} - \widehat{\mathbb{E}}[\boldsymbol{X}_{p,t}])^\top\,, \tag{3.12}$$

where $\widehat{\mathbb{E}}$ indicates the empirical means estimated as $\widehat{\mathbb{E}}[\boldsymbol{X}_t] = \frac{1}{N}\sum_{n=1}^{N}\boldsymbol{X}_t^{(n)}$ and $\widehat{\mathbb{E}}[\boldsymbol{X}_{p,t}] = \frac{1}{N}\sum_{n=1}^{N}\boldsymbol{X}_{p,t}^{(n)}$.

The innovations' mean and covariance, estimated as the residual mean and residual covariance matrix, take the following form:

$$\widehat{\boldsymbol{k}}_t = \widehat{\mathbb{E}}[\boldsymbol{X}_t] - \widehat{A}_t\widehat{\mathbb{E}}[\boldsymbol{X}_{p,t}] \tag{3.13}$$

$$\widehat{\Sigma}_t = \frac{1}{N}\sum_{n=1}^{N}(\boldsymbol{X}_t^{(n)} - \widehat{A}\boldsymbol{X}_{p,t}^{(n)} - \widehat{\boldsymbol{k}}_t)(\boldsymbol{X}_t^{(n)} - \widehat{A}\boldsymbol{X}_{p,t}^{(n)} - \widehat{\boldsymbol{k}}_t)^\top \tag{3.14}$$

In practice, when dealing with multi-trial data, the trials are assumed to be repeated observations of the same process or processes deemed sufficiently similar to correspond to the same underlying dynamical model. Considering the VAR model we exploit in Eq. 3.3 is inhomogeneous in time, the *i.i.d.* samples we use to estimate the parameters are gathered for all trials at each time point.

3.2.4.3 *Bias in estimating VAR model parameters with peri-event snapshots*

Based on the equations for estimating VAR model parameters, we can show here specifically how the detection procedure (represented by the yellow detection node in Figure. 3.7C) causes the selection bias in recovering the system dynamics.

Briefly, we start by fitting the panel data assuming it corresponds to samples from an inhomogeneous VAR model described in Eq. 3.7 where $n = 1 \ldots N$:

This is the VAR model for peri-event snapshots

$$\tilde{\boldsymbol{X}}_{t'+t} \coloneqq A_{t'+t}\tilde{\boldsymbol{X}}_{p,t'+t} + \boldsymbol{\eta}_{t'+t}\,, \boldsymbol{\eta}_{t'+t} \sim \mathcal{N}(\boldsymbol{k}_{t'+t}, \Sigma_{t'+t}) \tag{3.15}$$

Notably, these quantities are all conditioned on $Z_t = 0$ (consistent with Section 3.2.3.3) while we omit the conditions to ease notation. As demonstrated in Section 3.2.4.2, the parameters of such model can be inferred from the time-resolved second-order statistics such that

These equations are just peri-event version of the ones in Section 3.2.4.2

$$\widehat{A}_t = \widehat{\Sigma}_{\tilde{\boldsymbol{X}}_{t'+t}\tilde{\boldsymbol{X}}_{p,t'+t}}(\widehat{\Sigma}_{\tilde{\boldsymbol{X}}_{p,t'+t}})^{-1} \tag{3.16}$$

where the covariance matrices are estimated from the N-trials of data as:

$$\widehat{\Sigma}_{\tilde{\boldsymbol{X}}_{t'+t}\tilde{\boldsymbol{X}}_{p,t'+t}} = \frac{1}{N}\sum_{n=1}^{N}(\tilde{\boldsymbol{X}}_{t'+t} - \widehat{\mathbb{E}}[\tilde{\boldsymbol{X}}_{t'+t}])(\tilde{\boldsymbol{X}}_{p,t'+t} - \widehat{\mathbb{E}}[\tilde{\boldsymbol{X}}_{p,t'+t}])^\top, \tag{3.17}$$

and

$$\widehat{\Sigma}_{\tilde{\boldsymbol{X}}_{p,t'+t}} = \frac{1}{N}\sum_{n=1}^{N}(\tilde{\boldsymbol{X}}_{p,t'+t} - \widehat{\mathbb{E}}[\tilde{\boldsymbol{X}}_{p,t'+t}])(\tilde{\boldsymbol{X}}_{p,t'+t} - \widehat{\mathbb{E}}[\tilde{\boldsymbol{X}}_{p,t'+t}])^\top, \tag{3.18}$$

where the empirical means are estimated as $\widehat{\mathbb{E}}[\tilde{\boldsymbol{X}}_{t'+t}] = \frac{1}{N}\sum_{n=1}^{N} \tilde{\boldsymbol{X}}_{t'+t}$ and $\widehat{\mathbb{E}}[\tilde{\boldsymbol{X}}_{p,t'+t}] = \frac{1}{N}\sum_{n=1}^{N} \tilde{\boldsymbol{X}}_{p,t'+t}$.

Clearly, the accurate recovery of the coefficient matrix is guaranteed by an accurate estimation of the two covariance matrices in Eq. 3.17 and Eq. 3.18 from data. However, these statistics are conditioned on the snapshot selection criterion based on $\tilde{D}_t$, while we want to assess the unconditional quantities.

Specifically, as snapshots are detected in observed time series $\tilde{\boldsymbol{X}}_t$ using condition $\tilde{D}_{t_n} > d_0$, the covariance matrices we obtain directly from the detected peri-events snapshots are estimates of the conditional covariance matrices $\widehat{\Sigma}_{\tilde{\boldsymbol{X}}_{t'+t_n}\tilde{\boldsymbol{X}}_{p,t'+t_n}|Z_{t_n}=0,\tilde{D}_{t_n}>d_0}$ and $\Sigma_{\tilde{\boldsymbol{X}}_{p,t'+t_n}|Z_{t_n}=0,\tilde{D}_{t_n}>d_0}$, may differ from the real (unconditional) ones, i.e. $\widehat{\Sigma}_{\tilde{\boldsymbol{X}}_{t'+t}\tilde{\boldsymbol{X}}_{p,t'+t}|Z_t=0}$ and $\Sigma_{\tilde{\boldsymbol{X}}_{p,t'+t}|Z_t=0}$.

For notations check Section 3.2.3.4

3.2.5 *Debiasing based on threshold variations: the DeSnap algorithm*

The event selection bias problem in uncovering state-dependent network dynamics has been fully described in the above sections. In this section, we will show that it is possible to correct the bias by sampling from multiple conditions, i.e., the conditions where the threshold d_0 takes multiple values instead of a single one, resulting in multiple datasets of the peri-event snapshots. Specifically, we derive a Debiased Snapshot estimation (*DeSnap*) algorithm, which relies on a Gaussian approximation and exploits the variation in the second-order statistics of the data induced by variations of detection thresholds.

3.2.5.1 *Relationship between the unconditional and conditional statistics*

Apparently, as the bias in practice originates from estimating the VAR model with conditioned covariance matrices, the key of correction is to establish a relationship between 1) the conditional covariance matrices obtained from observed peri-event snapshots and 2) the unconditional covariance reflecting the *state-dependent* network dynamics. The differences of the two conditions are addressed in the last paragraph in Section 3.2.4.3.

To ease notations, in this section we will denote current state of peri-event snapshot $\tilde{\boldsymbol{X}}_{t'+t}|Z_t = 0$ as $\mathbf{X}_t$ and the past states as $\tilde{\boldsymbol{X}}_{p,t'+t}|Z_t = 0$ as $\mathbf{X}_{p,t}$. We also simplify the detection signal in the event-hosting state $\tilde{D}_t|Z_t = 0$ as D. Consequently, the detected peri-event snapshots $\tilde{\boldsymbol{X}}_{t'+t_n}|D_{t_n} \geqslant d_0, Z_{t_n} = 0$ is denoted as $\mathbf{X}_t|D \geqslant d_0$. Similarly, the detected past state $\tilde{\boldsymbol{X}}_{p,t'+t_n}|D_{t_n} \geqslant d_0, Z_{t_n} = 0$ are denoted as $\mathbf{X}_{p,t}|D \geqslant d_0$.

The new simplified notations are bold, untilted $\mathbf{X}_t$ without tilde

We start the derivation by representing the snapshot values at peri-event time point t as an extended state variable $\mathbf{Y}_t$, by concatenating $\mathbf{X}_t$ and $\mathbf{X}_{p,t}$, where $t \in [-T/2, T/2]$:

$$\mathbf{Y}_t = \begin{bmatrix} \mathbf{X}_t \\ \mathbf{X}_{p,t} \end{bmatrix}$$

such that the first and second-order statistics, written in the following forms, can be estimated directly from data (μ and Σ denotes the mean and covariance of a random variable):

$$\mu_{\mathbf{Y}_t|D \geqslant d_0} = \begin{bmatrix} \mu_{\mathbf{X}_t|D \geqslant d_0} \\ \mu_{\mathbf{X}p|D \geqslant d_0} \end{bmatrix}, \Sigma_{\mathbf{Y}_t|D \geqslant d_0} = \begin{bmatrix} \sum_{\mathbf{X}_t|D \geqslant d_0} & \Sigma_{\mathbf{X}_t\mathbf{X}_p|D \geqslant d_0} \\ \sum_{\mathbf{X}_p\mathbf{X}_t|D \geqslant d_0} & \Sigma_{\mathbf{X}_p|D \geqslant d_0} \end{bmatrix} \tag{3.19}$$

Critically, here we make a joint Gaussian assumption of the lagged observations $\mathbf{Y}_t$ and the *detection signal* D, so that we can take advantage of the following property of multi-variate Gaussian distributions: the conditional probability for one subset of components conditioned on the remaining ones is still a Gaussian distribution (see e.g. [168]).

Under this Gaussian assumption, to derive the probabilities for condition *over threshold* (i.e., $D \geqslant d_0$), we start from for the probabilities of each values of $d \in D$ where $d \geqslant d_0$. The conditional distribution of $\mathbf{Y}_t|D = d$ is also Gaussian with mean $\mu_{\mathbf{Y}_t|D=d}$ and variance $\Sigma_{\mathbf{Y}_t|D=d}$, such that:

$$\mu_{\mathbf{Y}_t|D=d} = \mu_{\mathbf{Y}_t} + \Sigma_{\mathbf{Y}_t D}\Sigma_D^{-1}(d - \mu_D) \tag{3.20}$$

$$\Sigma_{\mathbf{Y}_t|D=d} = \Sigma_{\mathbf{Y}_t} - \Sigma_{\mathbf{Y}_t D}\Sigma_D^{-1}\Sigma_{\mathbf{Y}_t D}^T \tag{3.21}$$

where μ_D is the mean of D_t over time: $\mu_D = \mathbb{E}[D_t]$.

"at" means that only points in D *that equals* d *are detected as reference points*

Practically, $\mathbf{Y}_t|D = d$ refers to the lagged snapshots detected *at* the threshold d. What is of more interest are the condition where the snapshots are detected *over* the threshold d. As the conditions $D = d$ are not overlapping for different values of d, the joint probability of $\mathbf{Y}_t$ and $D|D \geqslant d_0$ is the sum of the infinite joint probability of $\mathbf{Y}_t$ and $D|D = d$ (where $d \geqslant d_0$)

$$P\left(\mathbf{Y}_t, D \geqslant d_0\right) = \int_{d_0}^{+\infty} P\left(\mathbf{Y}_t, D = d\right) dd \tag{3.22}$$

Factorizing the joint probability into conditionals, the conditional probability of $\mathbf{Y}_t|D \geqslant d_0$ can then be derived as:

$$P\left(\mathbf{Y}_t \mid D \geqslant d_0\right) = \int_{d_0}^{+\infty} \frac{P(D = d)}{P\left(D \geqslant d_0\right)} P\left(\mathbf{Y}_t \mid D = d\right) dd \tag{3.23}$$

$P(D = d)$ *and* $P\left(D \geqslant d_0\right)$ *are not necessarily computable directly from the data, for details see Section 3.2.5.2*

Eq. 3.23 indicates that the lagged snaphsots Y_t detected *over* threshold d follows a Gaussian mixture distribution comprised of infinite Gaussian distributions $P\left(\mathbf{Y}_t \mid D = d\right)$, where $P(D = d)$ and $P\left(D \geqslant d_0\right)$ are both constants given any d and d_0. The mean and covariance of this Gaussian mixture is a function of the mean and covariance of each element [168]. By plugging in Eq. 3.20 and Eq. 3.21, we can derive the two statistics in the following expressions (see Appendix A.4 for detailed derivation):

$$\mu_{\mathbf{Y}_t|D\geqslant d_0} = \mu_{\mathbf{Y}_t} + \Sigma_{\mathbf{Y}_t D}\Sigma_D^{-1}\left(\bar{d} - \mu_D\right), \tag{3.24}$$

$$\Sigma_{\mathbf{Y}_t|D\geqslant d_0} = \Sigma_{\mathbf{Y}_t} + \Sigma_{\mathbf{Y}_t D}\Sigma_D^{-1} c(d_0)\Sigma_D^{-1}\Sigma_{\mathbf{Y}_t D}^{T}. \tag{3.25}$$

where $\bar{d}$ is the average of D_t over the threshold d_0:

$$\bar{d} = \mathbb{E}[D \mid D \geqslant d_0] = \int_{d_0}^{+\infty} dP(D = d)dd, \tag{3.26}$$

$c(d_0)$ is a scalar statistic of D_t:

$$c(d_0) = \int_{d_0}^{+\infty} \frac{P(D = d)}{P\left(D \geqslant d_0\right)} \left(d - \mu_D\right)^2 dd - \left(\bar{d} - \mu_D\right)^2 - \Sigma_D. \tag{3.27}$$

To understand the role of each variable in Eq. 3.24 and Eq. 3.25, we catagorize them into 3 groups and make the following statements:

- What is knwon empirically from peri-event snapshots are the conditional statistics $\mu_{\mathbf{Y}_t|D\geqslant d_0}$, $\Sigma_{\mathbf{Y}_t|D\geqslant d_0}$ (which can be estimated according to Section 3.2.4.2), and the binned conditions d (which we can specify on our need for detection).

- What we are interested in recovering, are the unconditional mean $\mu_{\mathbf{Y}_t}$ and covariance matrix $\Sigma_{\mathbf{Y}_t}$.

- $\Sigma_{\mathbf{Y}_t D}\Sigma_D^{-1}$, μ_D and $c(d_0)$ are intermediate unknown variables that help us estimated the unconditional statistics.

Therefore, Eq. 3.24 and Eq. 3.25 shows the desired link between the unconditional $\mu_{\mathbf{Y}_t}$, $\Sigma_{\mathbf{Y}_t}$ and the conditional statsitics $\mu_{\mathbf{Y}_t|D\geqslant d_0}$ and $\Sigma_{\mathbf{Y}_t|D\geqslant d_0}$. The expressions suggest that at each time point, the differences between the conditional statistics (conditioned on thresholding over d_0) and the real (unconditional) statistics, are linear functions of other statistics that only depends on the detection threshold d_0 :

$$\mu_{\mathbf{Y}_t|D\geqslant d_0} - \mu_{\mathbf{Y}_t} = f_\mu(\bar{d}(d_0)) \tag{3.28}$$

$$\Sigma_{\mathbf{Y}_t|D \geqslant d_0} - \Sigma_{\mathbf{Y}_t} = f_{\Sigma}(c(d_0)) \tag{3.29}$$

where f_{μ} and f_{Σ} represent the linear relationships. This idea paves the way for recovering the former from the latter, as we will propose in Section. 3.2.5.2.

3.2.5.2 *Details of DeSnap algorithm*

As mentioned at the beginning of Section 3.2.5.1, all the equations derived in Section 3.2.5.1 apply to the signals in a single state $Z_t = 0$. For a uni-state signals, the full-time time series of detection signal D are from the same state, where statistics of D, like μ_D and $c(d_0)$, can be easily obtained by exploiting the distribution of D.

Correction is more challenging in the case where the signal is a mixture of multiple states, where the hidden states Z_t are largely unobserved. With Assumption 1 in Section 3.2.2.3, the *reference points* $\{t_n|D \geqslant d_0\}$ always fall in the desired state $Z_t = 0$, suggesting that statistics of the observations $\mathbf{Y}_t$ calculated in the condition of $D \geqslant d_0$ (e.g. $\mu_{\mathbf{Y}_t|D \geqslant d_0}$, $\Sigma_{\mathbf{Y}_t|D \geqslant d}$) are also in the state $Z_t = 0$ (for an example see Section 3.3.2). However, it is hard to recover D for every hidden state $Z_t = 0$, making it difficult to estimate the statistics related to D due to unobserved probabilities of $P(D \geqslant d_0)$ and $P(D = d|d \geqslant d_0)$ in Eq. 3.26 and Eq. 3.27.

reminder: D *refers to* $\tilde{D}_t|Z_t = 0$

Actually, taking advantage of the linear relationships achieved under Gaussian assumption in Eq. 3.28 and Eq. 3.29, these intermediate variables and the unconditional statistics can all be retrieved by performing three linear regressions.

- First, with the snapshots and a given set of binned thresholds d (which must satisfy $d \geqslant d_0$ but should not be too large to limit the sample size of $P(\mathbf{Y}_t|D = d)$), we can regress d over $\mu_{\mathbf{Y}_t|D=d}$ in Eq. 3.24 to get the coefficient p_t and the intercept q_t corresponding to:

$$p_t = \Sigma_{\mathbf{Y}_t D} \Sigma_D^{-1}, \tag{3.30}$$

$$q_t = \mu_{\mathbf{Y}_t} - \Sigma_{\mathbf{Y}_t D} \Sigma_D^{-1} \mu_D . \tag{3.31}$$

- Secondly, q_t is a linear function of p_t as $q_t = \mu_{\mathbf{Y}_t} - p_t \mu_D$. Thus we can regress p_t over q_t to estimate the mean of D (μ_D) as the coefficient and $\mu_{\mathbf{Y}_t}$ as the intercept.

- Finally, Eq. 3.25 can be reorganized as:

$$\Sigma_{\mathbf{Y}_t|D \geqslant d_0} = \Sigma_{\mathbf{Y}_t} + c(d_0) p_t p_t^T, \tag{3.32}$$

  For a given threshold d_0, $c(d_0)$ is a constant for all elements of the covariance matrix at all time points of the snapshots. Regressing $p_t p_t^T$ over $\Sigma_{\mathbf{Y}_t|D \geqslant d}$ for any single element across time, we can estimate $c(d_0)$, by which we are able to retrieve $\Sigma_{\mathbf{Y}_t}$ from Eq. 3.32.

Notably, considering the linear form of covariance matrices in Eq. 3.32, it is theoretically possible to regress $c(d_0)$ over $\Sigma_{\mathbf{Y}_t|D \geqslant d}$ at different threshold d_0. However, practically it is better to avoid this regression because sample size can be limited with higher values of d_0.

3.2.6 *Summary of Methods*

This section is designed to summarize all the information in Method section to form an overview of the selection bias problem in analysing peri-event snapshots and the ideas behind correction, while establishing a standard snapshot analysis procedure to prepare for the applications in the Results Section.

3.2.6.1 *Analysis and correction of the bias in peri-event snapshots*

For a thorough investigation of the nature of bias in peri-event snapshot analysis, we re-state the key interpretations here to get the scattered ideas into shape.

With the snapshot analysis framework, we assume that events prone to emerge when the state space trajectory of the system crosses a specific region $\mathcal{E}_0$ such that we define the $\mathcal{E}_0$-passing hidden states as $Z_t = 0$. According to *Takens theorem*, the state space trajectory can be reconstructed by delayed embedding of the past N_D states of the system's observation $\tilde{\boldsymbol{X}}_t$, where $\mathcal{E}_0$ in the hidden state space are preserved as a region $\mathcal{E}_D$ in the embedded space.

By design, the detection signal $\tilde{D}_t$ projects the state-space embedding into a 1-dimensional space. By thresholding over the detection signal $\tilde{D}_t$ to find the *reference points* $\{t_n | \tilde{D}_t \geqslant d_0\}$, we are detecting the states in $\mathcal{E}_D$ that are located on one side of the hyperplane in state space associated to the threshold value d_0, as illustrated in Figure 3.4. This is naturally a biased selection of states in $\mathcal{E}_D$ corresponding to the fact that $P(\tilde{D}_t \geqslant d_0 | Z_t = 0) \neq 1$.

Peri-event snapshots can be modelled as inhomogeneous VAR models, where the estimation of time-varying model parameters depends on the conditional probability $P(\tilde{\boldsymbol{X}}_{t'+t} | \tilde{\boldsymbol{X}}_{p,t'+t}, Z_t = 0)$. Further, we combined VAR model with the sample selection bias theory in the framework of SCM, explaining in the non-parametric settings why conditioning on the detection $\tilde{D}_t \geqslant d_0$ leads to non-recoverability of the conditional probability.

However, with an extra Gaussian assumption, it is possible to obtain an unbiased estimate of the model parameters. To dig into the problem, we found that two covariance matrices estimated from biased samples $\tilde{\boldsymbol{X}}_{t'+t} | \tilde{D}_t \geqslant d_0$ and $\tilde{\boldsymbol{X}}_{p,t'+t} | \tilde{D}_t \geqslant d_0$ lead to a biased estimation of the VAR coefficient matrix A_t (Section 3.2.3.4). This implies that if the covariance matrices can be corrected into the unbiased version, we are able to recover the real dynamics of the system with unbiased A_t. This further motivated us to establish a link between the covariance matrices conditioned and unconditioned on the detection criterion $\tilde{D}_t \geqslant d_0$. We then propose that by setting multiple detection thresholds and applying the *DeSnap* algorithm with three linear regressions, we are able to recover the unconditioned covariance matrices, thus the unbiased system dynamics associated to the target state $Z_t = 0$ where the events occur.

3.2.6.2 *Standard procedure in treating peri-event snapshots*

Here, we present a standard procedure for treating peri-event snapshots from detection to estimation and bias correction. For any original signal under study, we first perform the detection procedure described in Section 3.2.1: 1) depending on field knowledge, obtain the detection signal $\tilde{D}_t$ by applying a causal filter; 2) apply a threshold d_0 on the detection signal to detect the *reference points* $\{t_n | \tilde{D}_t \geqslant d_0\}$; 3) extract peri-event snapshots within a window around the reference points.

The inhomogeneous VAR model is then applied to model detected peri-event snapshots, providing the time-varying estimation of coefficient matrices, innovations mean and innovations variance, as described in Section 3.2.4.2. The VAR model described by these model parameters is referred to as the "conditional" or "uncorrected" model. With these model parameters, the peri-event snapshots can be approximated by a Gaussian process, enabling us to simulate multiple Monte-Carlo realizations of the process and calculate the power spectrograms using a Morlet wavelet transform. Notably, it is possible to calculate the spectrograms directly from the observed snapshots instead of using the Monte-Carlo re-simulations; however, we do the latter to keep consistency with the following "corrected" model.

As will seen in Section 3.3.3.1, sometimes the snapshots are re-aligned by the peri-event peak to compare with event-triggered activities

Afterward, multiple thresholds $d_n \geqslant d_0$ higher than the original threshold d_0 are applied to the detection signal $d_n \geqslant d_0$ to obtain multiple datasets of peri-event snapshots, with which we perform the *DeSnap* algorithm as three linear regressions to uncover the unconditional (real) statistics of the snapshots (i.e. μ_{Y_t}, Σ_{Y_t}). The "unconditional" or "corrected" model can be reconstructed with a Gaussian process using the model parameters calculated according to Section 3.2.4.2. Similarly, power spectrograms can be estimated by multiple Monte-Carlo simulations of the Gaussian process.

This standard procedure will be performed for different datasets in the Results section, illustrating the outcome of each analysis step in different conditions.

3.3 RESULTS

In Section 3.3.1 and Section 3.3.2, we first validate our *DeSnap* algorithm in simulations with known system dynamics. Specifically, we simulate bivariate oscillatory VAR(2) processes, with or without Markov-switching hidden state, to illustrate how the detection of peri-event snapshots biases the estimation of model coefficients as a representation of state-dependent system dynamics, and how *DeSnap* corrects for this bias. In Section 3.3.3, we will apply the *DeSnap* algorithm to a type of transient neural events, the SPW-Rs, to test the algorithm and to explore the underlying network dynamics.

3.3.1 *Validation on single-state VAR(2) process*

We first test our method for the simple case of a bivariate stationary VAR(2) model to see how well the bias correction could perform in the situation where the state is homogeneous across time.

Despite the temporal homogeneity, similarly to the example of Figure 3.1, we show that event detection introduces a time-inhomogeneity in the peri-event snapshots exclusively caused by the selection bias, as explained in Section 3.2.2.1. We will assess the ability of *DeSnap* to recover the time-homogeneous parameters.

3.3.1.1 *Simulation procedure*

The dynamics of the system is controlled by a constant coefficient matrix $A_t = A$, a constant non-zero innovations mean $\boldsymbol{k}_t = \boldsymbol{k}$, and a constant innovation covariance $\Sigma_t = \Sigma$, as defined in 3.3. Entries of this coefficient matrix were randomly generated and then selected such that the VAR(2) model is stationary while allowing the occurrence of intrinsic local oscillations, which we detect as events.

$$A = \begin{bmatrix} -0.5751 & 1 & -0.9408 & 1 \\ 0 & 1.7263 & 0 & -0.9737 \end{bmatrix}, \boldsymbol{k} = \begin{bmatrix} 0 \\ 0.65 \end{bmatrix}, \Sigma = \begin{bmatrix} 0.5 & 0 \\ 0 & 0.5 \end{bmatrix}$$

We simulated the process for 1×10^5 s with a sampling rate of 1000Hz. By calculating its power spectrum, we know that the simulated signals have maximum spectral power at 79.6Hz.

The snapshots are detected by the following procedure (in line with what has been described in Section 3.2.6.2): 1) the original signals (Figure 3.8A, top and middle) are filtered in a narrow band around its power peak (74.6-84.6Hz) with an order 49 Finite Impluse Response filter; 2) then the filtered pair of signals are summed up (Figure 3.8A, bottom); 3) afterward, d_0 = mean + 3*standard deviation of $\tilde{D}_t$ is selected as a threshold for event detection (marked by a black line in Figure 3.8A, bottom), where all points over the threshold $\{t_n | \tilde{D}_t \geqslant d_0\}$ can be seen as *reference points* marking the location of events; 4) peri-event snapshots are extracted by an 801-ms-long window around the *reference points* (Figure 3.8A, black traces). As a result, we detected snapshots comprising 5946 oscillatory events.

3.3.1.2 *Results*

Model order selection is first performed for the peri-event snapshots with the BIC method proposed in Section 3.2.4.1, which results in an optimized model order of 2, matching the true model structure.

With this model order, we first calculated the time-varying mean, covariance matrix, and the resulting autoregressive coefficients for the "uncorrected" model, as described in Section 3.2.6.2. They are visualized in blue traces in Figure 3.8B-C. Furthermore, with these time-varying statistics, the samples of the snapshot panel can be approximated by Monte-Carlo simulation of the fitted time-inhomogeneous VAR model, with which we estimated the power spectrograms using reconstructed simulations of the snapshots (see Section 3.2.6.2). The spectrograms are shown in Figure 3.8D.

We next obtain the "corrected" or "unconditional" estimation of the model by applying our bias correction method, i.e., the *DeSnap* algorithm. Similarly, we can obtain the corresponding statistics, i.e., time-varying mean, covariance, model coefficients, which are shown in orange traces in Figure 3.8B-C and the power spectrograms in Figure 3.8D are estimated through Monte-Carlo simulation of the corrected model.

As shown in Figure 3.8B, compared to uncorrected averaged event waveforms, corrected waveforms match well the ground truth (black) as time-invariant for both variables in the VAR(2) process. As shown in Figure 3.8C, time-varying bias in covariance matrices (left subfigure showing the results of the 1^{st} element), together with the autoregressive coefficient matrices (right subfigure for the 1^{st} element), are both well-recovered after applying the correction algorithm. Power spectrograms of the detected event snapshots are closer to ground truth (stationarity) after correction for both variables.

These results overall support that our bias correction method is able to deal well with selection bias for the snapshots detected in one stationary state.

3.3.2 *Validation on two-state VAR(2) process*

As the objective of our approach is to estimate state-dependent system dynamics for a specific value of the state (associated to the emergence of events), we test the performance of our *DeSnap* algorithm for a two-state Markov switching model, implementing alternations between a non-oscillatory regime and the oscillatory regime modelled in the Section 3.3.1.

3.3.2.1 *Simulation procedure*

The simulation of the oscillatory process (denoted as "state 0") still follows the parameter settings described in Section 3.3.1. The non-oscillatory regime (denoted as "state 1") corresponds to a constant VAR coefficient matrix $A_t = A'$, zero-valued innovation mean $\boldsymbol{k}_t = \boldsymbol{k}'$ and the same innovations covariance $\Sigma_t = \Sigma$, whose values are designed such that the process shows weaker variance and less oscillations in the 74.6-84.6-Hz band compared to the oscillatory process:

the coefficient matrix ensures uni-directional causal effect, which is useful for Section 4.3.4.1

$$A' = \begin{bmatrix} 0.5 & 1 & 0.3 & 1 \\ 0 & -1.5 & 0 & -0.7 \end{bmatrix}, \boldsymbol{k}' = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \Sigma = \begin{bmatrix} 0.5 & 0 \\ 0 & 0.5 \end{bmatrix}$$

The Markov switching model's hidden state dynamics is determined by the transition probabilities: $P(\text{state } 0|\text{state } 1) = P(\text{state } 1|\text{state } 0) = 0.0001$.

Similarly to Section 3.3.1.1, we simulated the process for 2×10^5 s with a sampling rate of 1000Hz, and obtained a snapshot panel of 23674 trials after applying the same detection procedure. Example traces of the original signals and the detection signal are presented in Figure 3.9A (with a similar design as Figure 3.8A).

Notably, in this example, although thresholding ensures that *reference points* are all in state 0, the finite and random temporal duration of state 0 implies that the extracted snapshot also includes points from state 0 with an increasing probability as time deviates from the *reference points* (i.e. as the absolute values for peri-event time t' increases). This leads to a state mixing which tends to bias the estimation (especially close to edges of the peri-event time window). Nevertheless, the following results will show that it does not affect much the performance of the correction at time points close to the center of the detected event.

3.3.2.2 *Results*

Following the procedure explained in 3.3.1.2, we calculated the uncorrected and corrected time-varying statistics and model parameters of the two-state peri-event panel data, leading to the results shown in Figure 3.9B-E.

We can see from the figure that *DeSnap* performs well in the two-state case, especially in the middle of the snapshots, close to the peri-event time $t' = 0$. Figure 3.9B shows that similar to Figure 3.8B, the bias in averaged event waveforms can be well-corrected despite the state mixing at the border of the window. In Figure 3.9C-D, large bias in the time-varying covariance matrix and the coefficient matrix can be corrected by our method (exemplified by the traces of two elements each), similar to Figure 3.8C.

The mixing of states does cause a small bias, the size of which increases with the temporal distance from the *reference point* (better illustrated in Figure 3.9D, right). However, this mixing-induced bias is negligible compared to the selection bias.

Similarly, for power spectrograms, regardless of the mixing of states, the selection bias correction method performs well on the two-state case. The mixing bias is an intrinsic, inevitable problem of the threshold-based detection approach but does not harm the performance of our selection bias correction method.

3.3.3 *Application to in-vivo hippocampal recordings*

3.3.3.1 *Data structure and results*

We use the *DeSnap* approach to analyze the transient dynamics of the hippocampal SPW-Rs [29, 192], whose key role in reactivation of memory engrams has been introduced in Section 1.1.3.3.

We perform the analysis on 16 pairs of 2.88h-long local field potential signals recorded in the CA1 hippocampal subfield in one anesthetized macaque with a sampling rate of 667Hz. The two signals in each pair are recorded simultaneously in the stratum radiatum ("sr") and pyramidal layer ("pl"), as illustrated by the diagrams in Figure 3.10A (right insets). An example signal trace of this "sr"-"pl" pair is presented in Figure 3.10A (top and middle).

The snapshots detection procedure is similar to what was performed in the above simulations. The only differences are: 1) for ripples, we filter the original broadband signal in the ripple band (90-190Hz, where ripples have been reported to be significant [140, 192]) with a causal filter; 2) as the detection signal is forward-shifted by the causal filter, we use a non-centered detection window to counterbalance the time lag induced by this filter such that the SPW-R events appear in the middle of the window.

As seen in the following, aligning by "pl" signals, are more sensitive to the ripple-band oscillations, ensures better visualization in Figure 3.11

We obtain a snapshot consisting of 1928 SPW-R events. Notably, for the uncorrected condition, we re-aligned the events by the peri-event peaks in "pl" channels (and marked as "peri-event" in the figures). This is a more straightforward way to compare the bias-corrected results with the widely adopted event-triggered snapshots in experimental studies.

For autoregressive modeling, the optimal model order is set to 2 based on the BIC method we proposed for inhomogeneous multi-trial datasets (see Section 3.2.4.1). Then the *DeSnap* debiasing procedure was applied, resulting in the corrected time-varying statistics presented in Figure 3.10B, C.

Comparing the averaged temporal profiles of SPW-R-based *state* in "sr" and "pl" channels calculated with the "pl"-aligned peri-event datasets and corrected model (Figure 3.10B), we can conclude that both channel's profiles are attenuated after correction. Thus it is likely that in the uncorrected version (which is classically used by neuroscientists), the temporal profiles of transient states reflected by events are erroneously amplified due to the event selection procedure.

Unlike in Figures 3.8, 3.9, here we don't plot the comparison of covariance matrices and coefficient matrices as there is no ground truth

In Figure 3.10C, uncorrected spectrogram reflects a classical difference between activities recorded in different layers of the CA1 subfields: "pl" signals are more sensitive to the high-frequency component of the event, the ripple oscillation (>90Hz), while "sr" is more sensitive to the lower-frequency gamma activity (50-90Hz). Interestingly, the bias correction method removed a peak in a high-frequency band in the power spectrogram of the "pl" channel (Figure 3.10C), shifting location and time-width of the peak reflecting the ripple oscillation to a lower frequency range and a longer duration (≈20 ms), more in line with the properties estimated form single-trial analysis (see Figure S5 in Ramirez-Villegas et al. [193]).

This suggests common event selection procedures of SPW-Rs in neuroscience might result in a biased power spectrogram that misrepresents single-trial activities. As a consequence, it is critical to run a correction for event selection such as *DeSnap* when exploiting peri-event data in order to avoid misestimating key parameters tightly related to the underlying neural mechanisms, such as the frequency and duration of high-frequency oscillations.

3.3.3.2 *Filter invariant spectrograms matches state-dependent network dynamics*

We conducted an additional analysis on the peri-event snapshots of SPW-Rs to illustrate the power of our bias correction method in recovering underlying state-dependent dynamics based on detected events.

In contrast to the detection procedure described in Section 3.3.3.1 where the original signal is filtered within a single frequency band (90-190 Hz), here we compare the bias correction performance on hippocampal signals after they are filtered in theta, high gamma, and ripple bands. These bands have been reported to be associated with three different hippocampal transient events, as reviewed in Section 1.1.4.3.

Figure 3.11A, B reflect the correction of spectrograms after the detection is performed in 10 shifted filter bands. The filter bands are presented as two groups: in Figure 3.11A the 5 filter bands are shifted within theta and low-gamma bands; in Figure 3.11B the other 5 filter bands range within high-gamma and ripple bands.

Interestingly, we found that as the filter's frequency band is shifted, the *corrected* spectrograms of the high gamma and ripple signals remain unchanged, as seen in Figure 3.11B for an example channel pair. Comparing Figure 3.11B to Figure 3.11A, we can see that the corrected spectrograms of theta events are different from the corrected spectrograms in A but also remain relatively constant when the filter is shifted within the theta band (i.e., first three subplots). In short, the corrected spectrograms categorizes the events obtained with different filter bands into two clusters: one cluster consisting of theta events, the other cluster comprised of high-gamma and ripple events.

This separation of events based on filter band invariance is further confirmed by a similarity analysis of these 10 spectrograms for either the uncorrected and corrected models. Comparing Figure 3.11C (top left) and Figure 3.11C (top right), the corrected spectrograms are better clustered into two categories that are consistent with A and B. We also mapped these spectrograms in the high dimensional space by multidimensional scaling, as shown in Figure 3.11C (bottom). In the corrected model (left), the 10 pairs of spectrograms also appear as two clusters, while the clustering is not clear for the uncorrected model (Figure 3.11C (bottom right)).

Notably, the two clusters are different from the two groups: the low-gamma bands are outliers that belong to neither cluster

To illustrate the consistent contrast of clustering between the uncorrected model and the corrected model, we calculate the cluster quality for all channel pairs, as shown in Figure 3.11D. Based on the two groups presented in Figure 3.11A and Figure 3.11B, the cluster quality is defined as the ratio between inter-group distance and intra-group distance of each point in Figure 3.11D. Intuitively, large cluster quality values indicate that both groups are more clustered. As Figure 3.11D shows, all cluster quality values for 16 channel pairs for the corrected model are significantly higher than the values for the uncorrected model ($p < 0.001$), confirming that the spectrograms of two groups form two clusters after correction.

Based on the previous idea that the corrected model captures the network dynamics related to the state, this result suggests that the ripple and high gamma events occur in one state, and the theta events occur in another state. Actually, this conclusion is consistent with the previous experimental findings mentioned in Section 1.1.4.3 that the theta events occur in one state similar to the REM stage coupled to REM PGO waves, while the high-gamma and ripple events occur in another state resembling the NREM state marked by pre-REM PGO waves.

3.4 DISCUSSION

In summary, this chapter provides evidence for selection bias issues when exploiting peri-event snapshots to infer the underlying network dynamics. By analyzing and formulating the event detection procedure mathematically, we provided a theory with dynamical systems and SCMs to account for the nature of such selection bias.

In particular, we proposed a correction approach for the bias with inhomogeneous VAR models. The correction algorithm is based on a Gaussian assumption that ensures a linear relationship between the correction term and statistics of the threshold, enabling the underlying system dynamics to be recovered via three linear regressions. Further, we validated the correction algorithm with simulated VAR(2) systems, whose results suggest that the algorithm has the potential of recovering system dynamics underlying observed events.

Reversely, such capability of the correction algorithm is able to deepen the understanding of the mechanism underlying certain events. This has been confirmed by our results in applying the correction methods to transient hippocampal events defined in three different frequency bands, where the clustering of events matching the differentiation of transient states in the literature. Thus this correction method can be potentially applied to multiple unknown events and check whether they are generated by the same underlying mechanism, which will help to reduce the complexity of research when investigating certain events.

One needs to be careful when applying our correction algorithm to data, as the Gaussian assumption must be satisfied. Besides, as the regression depends on setting multiple large-valued thresholds to detect relatively rare events, the number of detected peri-event snapshots decreases with higher thresholds. Thus it is important to guarantee a relatively large sample size. Nevertheless, the third regression has already been designed to avoid estimation bias due to sample size, as addressed in Section 3.2.5.2.

In general, as this correction method provides a reconstruction of the system dynamics via VAR models, it naturally forms the basis of more advanced model-based analysis and overcome the spontaneous event problem we described in Section 3.1. We will show in Section 4.3.4 how this bias correction method can be combined with causality measures to reflect state-dependent causal interactions between two brain regions.

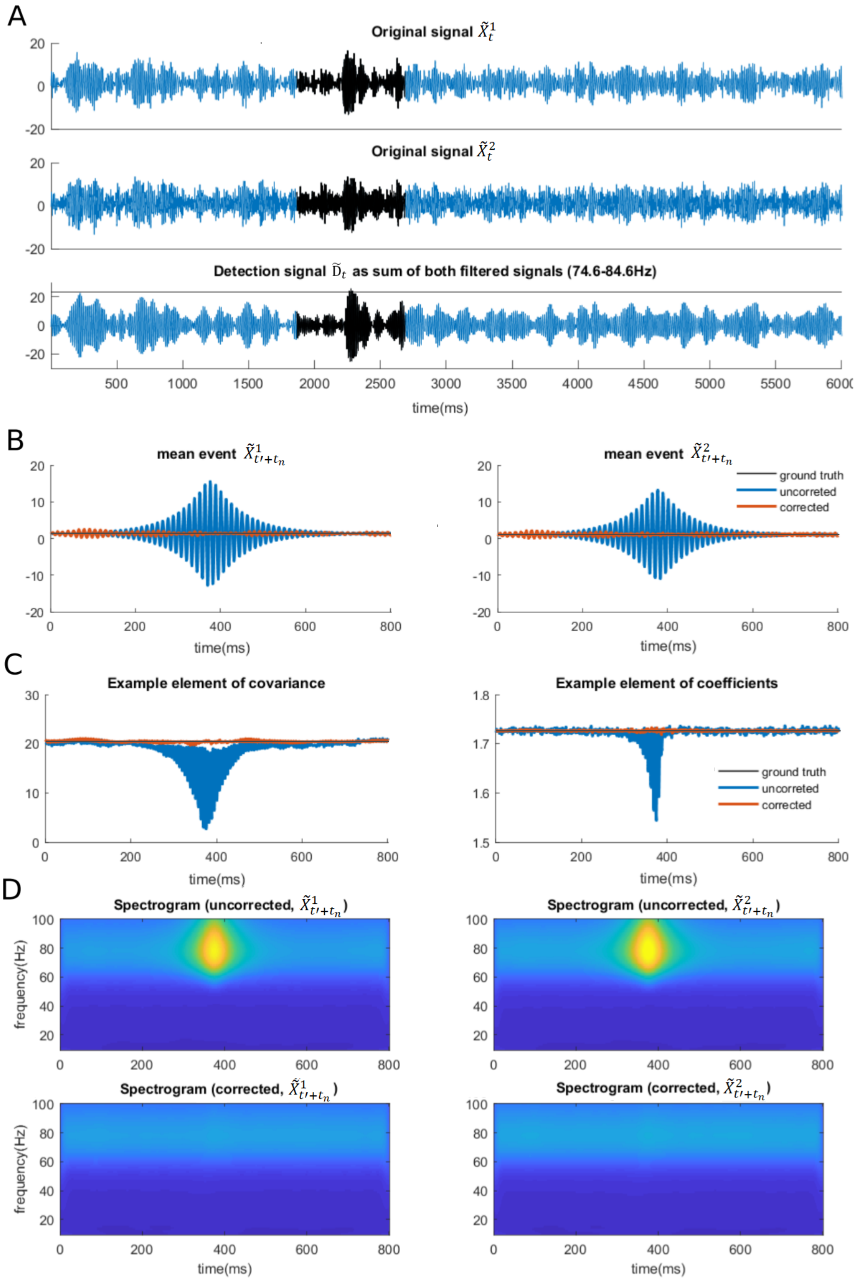

Figure 3.8: Experiments with a stationary two dimensional (one-state) oscillatory VAR(2) process. (A) Time course of both variables for one realization, together with the detection signal (bottom trace). Horizontal line indicates the detection threshold d_0. (B) Time-resolved peri-event mean estimate based on the panel data, without and with correction for selection bias. The peri-event time $t' = 0$ is at the center of the window. (C) Time-resolved peri-event estimate of an example covariance coefficient (left graph) and autoregressive coefficient (right) based on the panel data. The peri-event time $t' = 0$ is the same as (B). (D) Peri-event spectrogram estimate of each variable of the model uncorrected (top) and corrected (bottom).The peri-event time $t' = 0$ is the same as (B).

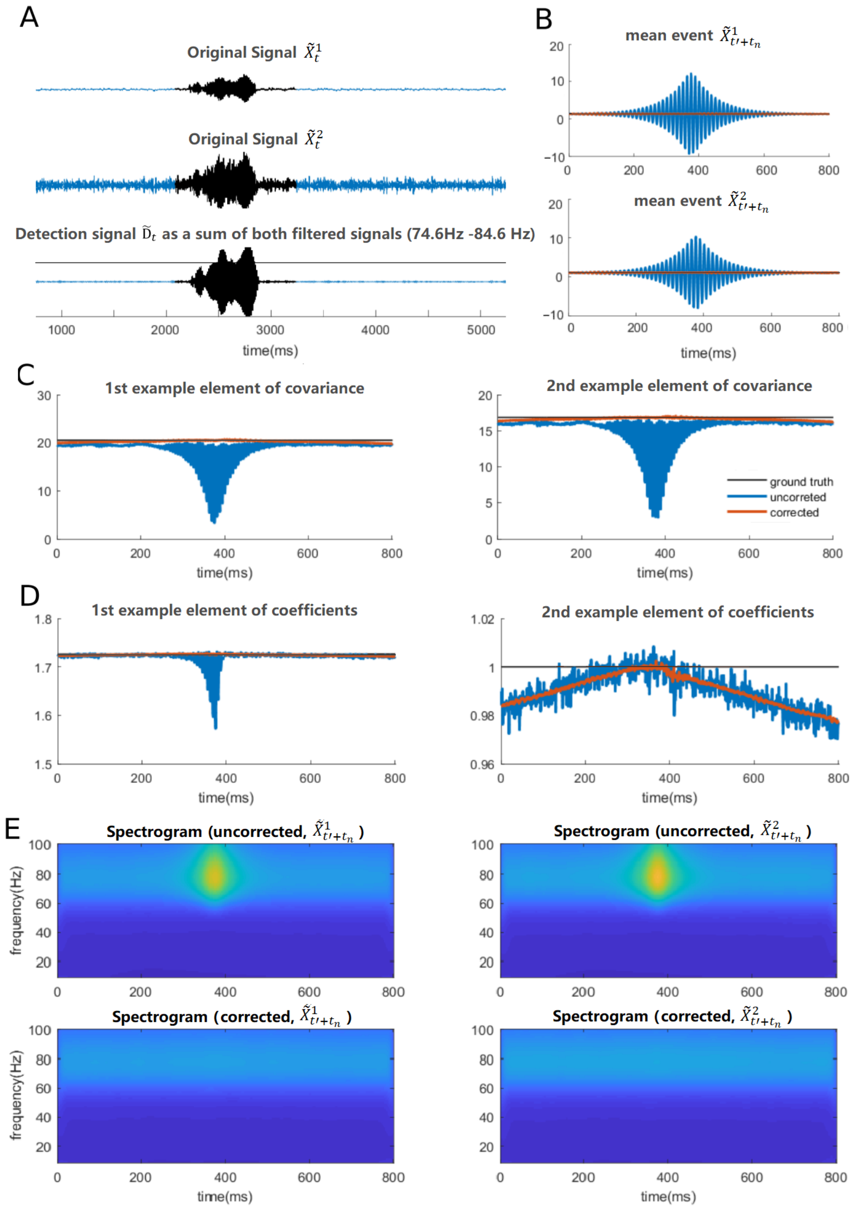

Figure 3.9: Experiments with a two-state oscillatory Markov switching model. (A) Time course of both variables for one realization, together with the detection signal (bottom trace). Black regions indicate a selected snapshot. Horizontal line indicates the detection threshold d_0. (B) Time-resolved peri-event mean estimate based on the panel data, without and with correction for selection bias. The peri-event time $t' = 0$ is at the center of the window. (C) Time-resolved peri-event estimate of two example elements of covariance matrices. The peri-event time $t' = 0$ is the same as (B). (D) Same as (C) for autoregressive coefficient. The peri-event time $t' = 0$ is the same as (B). (E) Peri-event spectrogram estimate of each variable of the model uncorrected (top) and corrected (bottom). The peri-event time $t' = 0$ is the same as (B).

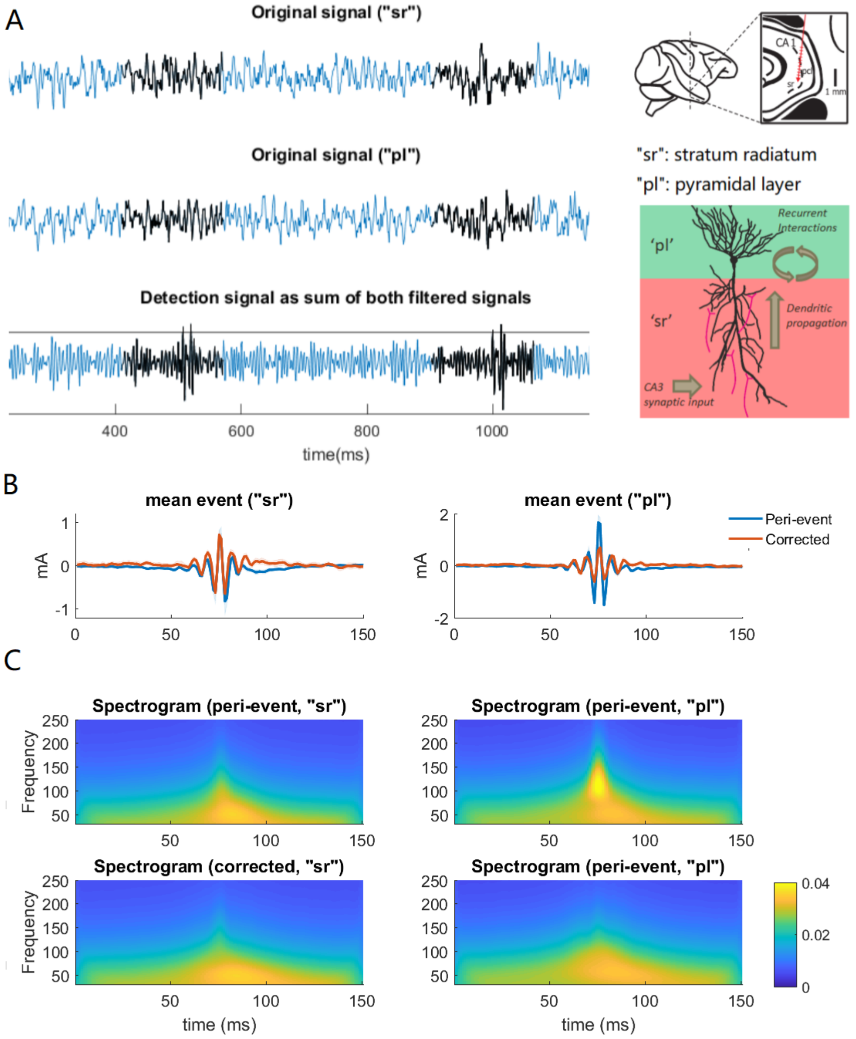

Figure 3.10: Experiments with hippocampal recordings of SPW-Rs in anesthetized macaque. (A) Exemplary traces, together with the detection signal (bottom trace). Black regions indicate a selected snapshot. Horizontal line indicates the detection threshold d_0. Right inset indicates the positioning of the recording electrode and the putative hippocampal subfields associated to each channel ("pl": pyramidal layer, "sr": stratum radiatum). (B) Time-resolved peri-event mean estimate based on the panel data, without and with correction for selection bias. The "uncorrected" model in this case are peri-event snapshots aligned by the peaks of "pl". The peri-event time $t' = 0$ is at t=100 ms due to an un-balanced window. (C) Peri-event spectrogram estimate of each variable of the model uncorrected (top) and corrected (bottom). The peri-event time $t' = 0$ is the same as (B).

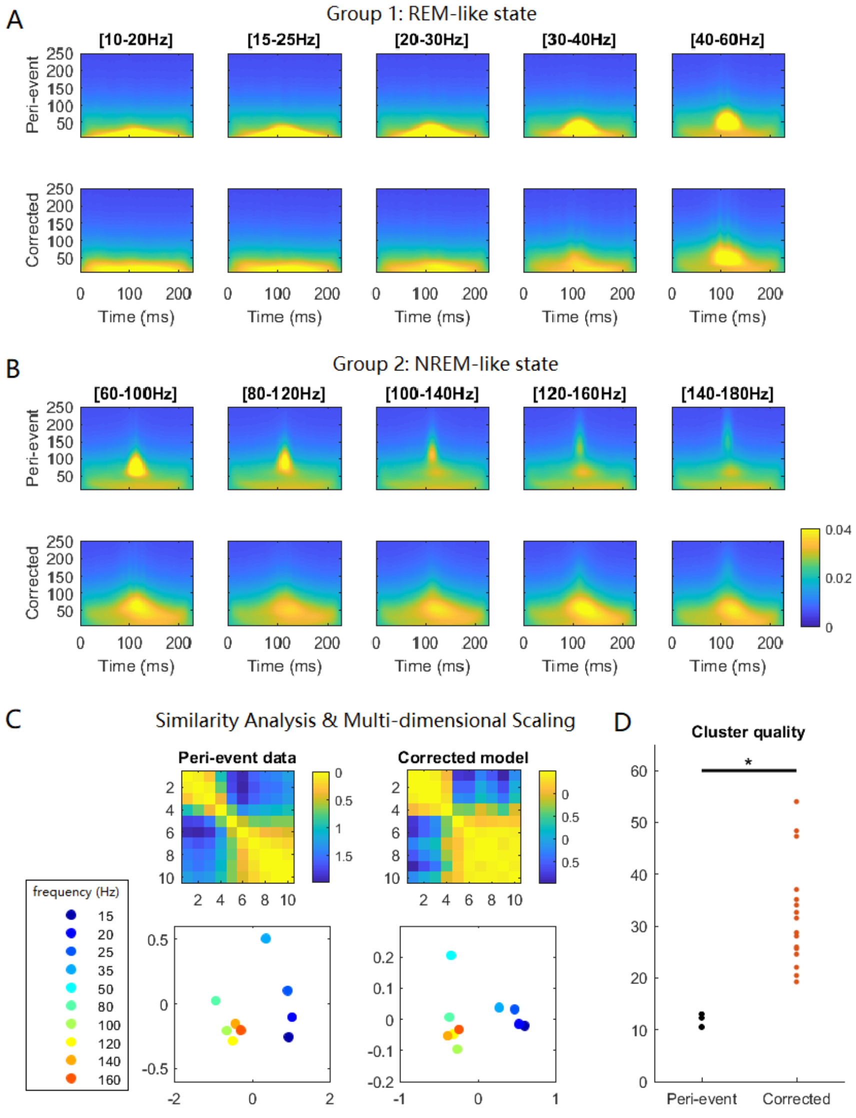

Figure 3.11: State-dependent filter-band invariance of corrected power spectrograms. (A) (top row) Power spectrograms of SPW-Rs aligned by "pl" in an example channel pair. The detection for events in different subfigures are based on a group of different filtering frequency bands ranging from the theta band to the low-gamma band (in the REM-like state). (bottom row) Power spectrograms of SPW-Rs calculated by the corrected model. Different subplots represent the corrected power spectrogram based on the same frequency band as the top row. (B) The same as (A) but the filtering bands are in another group ranging from high-gamma to ripple bands (in the NREM-like state). (C) Similarity analysis and multi-dimensional scaling of the power spectrograms in the 10 frequency bands shown in (A) and (B). (top) Similarity matrix within 10 frequency bands of the two groups in (A) and (B) for 'pl'-aligned peri-event snapshots (left) and the corrected model (right). (bottom) Multi-dimensional scaling of 10 frequency bands of the two groups in (A) and (B) for 'pl'-aligned peri-event snapshots (left) and the corrected model (right). Color code marks the center frequency of the corresponding filtering band. (D) Cluster quality calculated as the ratio between inter-and intra-group distances for the multi-dimensional scaling of REM and NREM bands, repeated 16 times by using different pairs of electrodes to fit the VAR models (N=16). The black star indicates a significant difference between the two groups ($p = 3.05 \times 10^{-5}$).

4 CAUSAL INVESTIGATION OF PERI-EVENT DATA

4.1 INTRODUCTION

motivation for causal analysis for transient events

As addressed in Section 1.2, one way of uncovering the mechanisms of brain-wide interactions is to investigate the causal interactions between neural events. In this chapter, we focus on the causal interactions between PGO waves and other transient events, especially SPW-Rs. This is a critical problem because researchers are interested in characterizing how one event drives another in order to understand how they are coordinated together.

In order to understand how these transient phenomena operate, causality measures based on observed neural time series can be very helpful to quantify the underlying transient influences between brain structures. Candidate measures range from Granger Causality (GC), originating from the econometrics literature[89], to its non-linear extensions (e.g.[149, 148, 64]) and, more widely-acknowledged, its information-theoretic generalization: Transfer Entropy (TE) [209, 244]. These measures quantify causal influences based on the ability of putative causes to improve the predictability of future observed effects. However, GC and TE are devised for stationary signals, which limits there applicability to uncover causal interactions that vary rapidly in time, as in the context of characterizing the interplay between transient events (elaborated in Section 4.2.4.1).

Moreover, the characterization of causality should arguably be based on the notions of counterfactuals and manipulation (elaborated in Section 4.2.1 and Section 4.2.4.2) instead of predictability of (unmanipulated) observations. In this sense, one can argue that TE and GC only measure statistical dependencies in the time series instead of actual causality. Therefore, although they have been widely used for assessing the significance of causal links, whether they are appropriate quantities to measure the strength of these links is still debated [227].

A more promising direction for causality analysis is through SCMs, as introduced in Section 3.2.3.1. SCMs and their associated graphical representation allow causality measures to be evaluated by implementing putative interventions on the system under study. Causality measures in the context SCMs have been investigated in Ay and Krakauer [3], which discussed how to account for knockout experiments and introduces a measure of *information flow*. Furthermore, the work of Janzing et al. [111] provides interesting theoretical justifications for this kind of measure and extends it to assess *causal strength* (CS) of an arbitrary set of arrows in a graphical model. Compared to GC or TE, information flow and CS have the benefit of being local, in the sense that it depends only on the direct causes of the observed effects. Besides, taking advantage of the interventional treatment, the SCM-based CS has the potential to be extended to intuitively accounted for the counterfactual condition where the cause is unchanged regardless of the causal link, which can be inspired by the *causal impact* methodology [198, 23].

In this chapter, we will review the key notions of counterfactuals and manipulations in causal analysis and clarify the differentiation between the cause and the causal mechanism (Section 4.2.1). After reviewing the theoretical insights of GC, TE, CS and causal impact (Section 4.2.2), we will extend them as time-varying information-theoretic quantities in the context of inhomogeneous VAR models, enabling the estimation of time-resolved causality from peri-event data (Section 4.2.3). Some theoretical and practical issues of these causality measures will be discussed in Section 4.2.4 within the SCM framework, suggesting the benefits of time-varying CS over TE while both suffer from the insensitivity to deterministic time-varying perturbations (Section 4.2.4.4). To address this issue, we propose an extension of DCS, relative DCS (rDCS), which captures causal influences even when they are mediated by deterministic changes in the innovations of the peri-event time series, and matches the intuition behind causal impact (Section 4.2.5).

We will validate these causality measures on simulated toy models (Section 4.3.1 and Section 4.3.2) and simulated thalamocortical spindles (Section 4.3.3), which points out an inconsistency issue with different peri-event alignment methods. We propose to address the alignment issue by combining the *DeSnap* algorithm introduced in Chapter 3 with causality measures to eliminate the effect of selection bias caused by different alignments when characterizing transient causal interactions underlying neural events. The effectiveness of such combination is further validated with toy models (Section 4.3.4.1), thalamocortical spindles (Section 4.3.4.2) and hippocampal SPW-Rs (Section 3.3.3.1).

4.2 METHOD

To flesh out the above critiques regarding the appropriateness of GC and TE for causality analysis, we first review some concepts in SCM-based causality to set the basis for interpreting and evaluating causality measures in the following sections.

causality measures include GC, TE, causal strength, causal impact

This will be followed by a review of several causality measures that are proposed as candidates for addressing transient causal interactions between neural systems. With a general mathematical introduction of each measure, we will show how to extend it to time-varying time series models (e.g., the inhomogenous VAR model discussed in Chapter 3).

Later, we will discuss the time-varying causality measures in the context of causality principles. We will demonstrate the shortcomings of each measure in characterizing transient causal interactions, and by addressing them pertinently, propose a novel measure that overcomes these shortcomings.

4.2.1 *Principles of causality studies*

Directly following what has been addressed in Section 4.1, investigations of causal interactions in neural signals frequently overlooked principles of causal manipulations and instead focused on distilling statistical dependencies for predicting the signals' future values. Here we discuss several basic principles and concepts of the field of causality research and demonstrate how they can be applied to investigating the causal interactions between transient events occurring in different brain regions.

4.2.1.1 *Counterfactuals*

One important direction of causality analysis is based on the counterfactual theories of causation, which dates back to David Hume [106] and has been extensively developed by David Lewis [137].

In this theory, causality boils down to comparing the *actual* outcomes and the *counterfactual* outcomes. The actual outcome means what has indeed happened (e.g., an event A), while counterfactual outcomes refer to a hypothetical condition where the event A does not occur. This forms the basis of causal investigations between two variables in empirical science, where control experiments are designed with mutually exclusive treatments to implement *manipulations* of the system, such that the results obtained with the two treatments form a pair of actual and counterfactual outcomes. However, the feasibility of obtaining the counterfactual conditions by manipulation is only confined to the systems which can be carefully controlled, while in reality, many physical and physiological processes cannot be easily located and repeated. For example, as described in Section 1.1.2 and Section 2.1, the neurons generating some transient events (e.g., the PGO waves) are too sparsely distributed spatially to be recorded with current electrophysiological techniques, creating obstacles for inhibiting these events to observe the resulting transient effects in another brain region at a fine temporal scale.

In fact, practically, in many cases like this, the counterfactual analysis suffers the problem that the counterfactual condition is difficult to define, e.g., the counterfactual event for one actual event can be expressed in different forms depending on the context. Moreover, even if the definition is clear, the counterfactual condition is usually hardly accessible unless in some specific cases. Then the critical question is whether we can reconstruct the counterfactual condition given the observational data only reflecting the actually observed condition.

see the comparison in Section 4.2.1.3

4.2.1.2 *Interventions in SCMs*

Fortunately, in the framework of SCMs (as introduced in Section 3.2.3.1), an operation named *intervention* has been incorporated to perform pseudo-manipulation of the system described by the SCM to obtain the counterfactual condition. More specifically, it refers to artificially change one arrow or one node in the SCM to observe the outcome of the change.

Classical interventions in an SCM are *perfect interventions* and amount to imposing a fixed value on a random variable, but the framework allows a much broader class of modifications of the SCM. For instance, instead of a constant, one can do *soft interventions* where we impose the value of a variable to be drawn from a given distribution, independently from other variables in the SCM.

This allows us to perform an intervention of removing the effect of an arrow, which is equivalent to feeding the effect node with an independent copy of the cause node with the same distribution. Mathematically, taking the SCM defined in Section 3.2.3.1, where a directed acyclic graph ($\mathcal{G}$) is described by the following structural equations

$$V_j := f_j(\mathbf{PA}_j, N_j), j = 1, \dots, d.$$

$\mathbf{PA}_j$ are the variables indexed by the set of parents of vertex j in $\mathcal{G}$. Intervening on V_k consists in replacing its structural assignment by a new one:

$$V_k := \widetilde{f_k}(\widetilde{\mathbf{PA}_k}, \widetilde{N_k}) .$$

The resulting modified distribution $\widetilde{P}_V = P_V^{do(V_k := \widetilde{f_k}(\widetilde{\mathbf{PA}_k}, \widetilde{N_k}))}$ is called *intervention distribution* (see e.g. [186, chapter 6]). Meanwhile, other structural equations and the marginal distribution of the parents $\mathbf{PA}_k$ are kept unchanged.

We will discuss in the Section 4.2.3 how to obtain candidate counterfactual conditions for time series signals with appropriate intervention distributions in SCMs.

4.2.1.3 *Independence of cause and mechanism*

Another essential principle to mention to construct a reasonable counterfactual condition is the Independence of cause and mechanism [186].

The gist of this principle is that the cause and the mechanism that drives the effect are essentially independent of each other and thus should be decomposed. For example, for an abstract causal relationship where an event C causes another event E. E is driven by C through a physical process that modulates the system of E and exists regardless of the occurence of C. As a specific example in neuroscience, in the propagation of spikes between two neurons, cause corresponds to the spiking activities of the pre-synaptic neurons. The mechanism refers to the synaptic machinery that leads to the excitatory post-synaptic potentials, contributing to the spiking of post-synaptic neurons. Without the pre-synaptic spikes, no post-synaptic firing would occur, which will not happen either without synaptic connectivity. Thus, in characterizing causality, we would need to consider both the cause and the mechanism.

A direct benefit of this principle is to help clarify the counterfactual condition. Still, in the example of C causing E, a natural counterfactual scenario is "if the event C does not happen, the event E does not occur". However, taking into account of the mechanism, the original statement can be detailed as "When C happens, and if there is a mechanism linking C to E, then E happens". Therefore, a better counterfactual condition should be "if C does not happen or if there is no mechanism linking C to E, then E does not happen".

We will elaborate on the application of this principle in Section 4.2.3 and Section 4.2.5.

4.2.2 *Review of candidate causality measures*

Having introduced the basic principles of causality analysis, in this section we will review several widely used causality measures which can be potentially applied to understand the transient interplay between neural events.

Here, the measures will be introduced as their classical definitions in a stationary bivariate system consisting two variables X^1 and X^2. As introduced in Section 3.2.4.2, in a homogeneous bivariate VAR model, at each time t, the current state $\boldsymbol{X}_t = [X^1_t, X^2_t]^\top$ is a linear function of the past p states gathered in the vector

The notations are designed to be consistent with the VAR model for peri-event snapshots, as introduced in Section 3.2.4.2

$$\boldsymbol{X}_{p,t} = [{\boldsymbol{X}^1_{p,t}}^\top, {\boldsymbol{X}^2_{p,t}}^\top]^\top = [X^1_{t-1}, X^1_{t-2}, ..., X^1_{t-p}, X^2_{t-1}, X^2_{t-2}, ..., X^2_{t-p}]^\top \tag{4.1}$$

and the exogeneous inputs as the innovation term

$$\boldsymbol{\eta} = [\eta^1, \eta^2]^\top$$

in the following form

$$\boldsymbol{X}_t := A\boldsymbol{X}_{p,t} + \boldsymbol{\eta}, \boldsymbol{\eta} \sim \mathcal{N}(\mathbf{k}, \Sigma), \tag{4.2}$$

where $A = \begin{bmatrix} \mathbf{a}^\top & \mathbf{b}^\top \\ \mathbf{c}^\top & \mathbf{d}^\top \end{bmatrix}$, $k = \begin{bmatrix} k^1 \\ k^2 \end{bmatrix}$, $\Sigma = \begin{bmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{bmatrix}$ are the model parameters. Elements of the coefficient matrix **a**,**b**,**c** and **d** are all p-dimensional vectors.

Expanding the vector version of Eq. 4.2 using Eq. 4.1, the equation can be rewritten as:

$$X_t^1 \;=\; \mathbf{a}^\top \boldsymbol{X}_{p,t}^1 + \mathbf{b}^\top \boldsymbol{X}_{p,t}^2 + \eta^1, \quad \eta_t^1 \sim \mathcal{N}(k^1, \sigma_1), \tag{4.3}$$

$$X_t^2 \;=\; \mathbf{c}^\top \boldsymbol{X}_{p,t}^1 + \mathbf{d}^\top \boldsymbol{X}_{p,t}^2 + \eta^2, \quad \eta^2 \sim \mathcal{N}(k^2, \sigma_2). \tag{4.4}$$

Applying this statistical model to data requires estimation of the model VAR model parameters, resorting to estimating of the statistics of X_t^1, X_t^2, $\boldsymbol{X}_{p,t}^1$ and $\boldsymbol{X}_{p,t}^2$ as random variables. Notably, in the homogeneous VAR model, the subscripts (t and p,t) just mark the relative temporal relationships between the current and the past state but do not confine the variables to each time point. Therefore, data at all time points can be gathered as *i.i.d.* samples of X_t^1, X_t^2, $\boldsymbol{X}_{p,t}^1$ and $\boldsymbol{X}_{p,t}^2$ as required in the VAR estimation procedure presented in Section 3.2.4.2. This will result in a time-invariant estimate of their mean, (co)variance and the model parameters.

If causal interactions are bi-directional in such a bi-variate system, the reviewed causality measures are supposed to reveal significant causal effects in both directions. Here we clarify that in the later sections (e.g., Results), we focus on the dominant direction of causations to check whether the causality measures reflect a qualitative view of the system.

4.2.2.1 *Granger causality*

The Granger causality (GC), as well as it extention, the Transfer Entropy (TE) (Section 4.2.2.2), is based on Wiener's principle of causality. According to the principle, Granger [89] postulates the existence of (Granger-)causality from X^2 to X^1 if knowledge of $\boldsymbol{X}_{p,t}^2$, in addition to $\boldsymbol{X}_{p,t}^1$, will allow better prediction of X_t^1. This actually resorts to the comparation of two conditions, where the first condition can be understood as the actual condition and the second as a counterfactual condition:

this is not necessary the only way to define the counterfactual condition

- predict X_t^1 with both $\boldsymbol{X}_{p,t}^1$ and $\boldsymbol{X}_{p,t}^2$
- predict X_t^1 with only $\boldsymbol{X}_{p,t}^1$

Importantly, in the general definition of this notion, $\boldsymbol{X}_{p,t}^1$ and $\boldsymbol{X}_{p,t}^2$ refer to the past of these time series, without further specification of a particular model order, such that in our notation p should be understood as potentially infinite. Choosing p actually raises issues (debates that are discussed in Section 4.2.4.1), thus we ask the reader to bear with us that p remains unspecified.

The classical Granger causality measure has been defined for the stationary VAR models as given in Eq. 4.2, where the coefficient matrix A and the innovations $\boldsymbol{\eta}$ are assumed homogenous across time. This model describing the actual condition, as Eqs. 4.3-4.4- is referred to as the *full model* [80], where the modelling of the first variable X_t^1 is dependent on both variables X_t^1 and X_t^2. To test whether X^2 causes X^1, the estimated innovation variance of X_t^1 (σ_1 in Eq. 4.3) also reflects the asymptotic mean squared residual error ($\hat{\sigma}_1$) of the forecast of X_t^2 under the assumption that both $\boldsymbol{X}_{p,t}^1$ and $\boldsymbol{X}_{p,t}^2$ contribute to X_t^1.

Under the proposed counterfactual condition where X_t^1 is predicted only by $\boldsymbol{X}_{p,t}^1$, we have a *reduced model*

$$X_t^1 = \mathbf{a}'^\top \boldsymbol{X}_{p',t}^1 + \eta^{1\prime}, \quad \eta^{1\prime} \sim \mathcal{N}(k^1, \sigma_1'). \tag{4.5}$$

where the model order p', the coefficient $\mathbf{a}'$, the innovations mean k^1 and innovations variance σ_1' are different from the corresponding terms in Eq. 4.3 and should be re-estimated.

for estimation methods see Section 3.2.4.2

If X^2 causes X^1, then the full model is a better model of the data compared to the reduced model. This means that the mean squared error of the reduced model, $\hat{\sigma}_1'$ should be larger than the full model, $\hat{\sigma}_1$. Then the Granger causality can be defined as the log ratio of the residual variance between the reduced model and the full model, which leads to estimate the magnitude of Granger causality as

$$GC(X^2 \to X^1) = \frac{1}{2} \log \left(\frac{\hat{\sigma}_1'}{\hat{\sigma}_1} \right) . \tag{4.6}$$

The factor 1/2 chosen in this case is chosen for consistency with TE (see Section 4.2.2.2).

While the above linear VAR model is the most widely used, Granger causality has been extended to non-linear models following the same predictive approach [149, 148, 64].

4.2.2.2 *Transfer Entropy*

TE is an information-theoretic implementation of Wiener's principle, where the performance of prediction between the actual and counterfactual conditions is quantified with conditional entropy. In information theory, the conditional entropy measures the amount of information needed to describe the outcome of a random variable given that the value of another random variable. Thus conditional entropy is directly formulation of the Wiener's principle.

Quantifying to which amount X^2 is Granger causes X^1 then results in the following quantity named transfer entropy,

$$\begin{aligned} TE(X^2 \to X^1) &= H(X_t^1 | \boldsymbol{X}_{p,t}^1) - H(X_t^1 | \boldsymbol{X}_{p,t}^1, \boldsymbol{X}_{p,t}^2) \\ &= \int p(x_t^1, \boldsymbol{X}_{p,t}^1, \boldsymbol{X}_{p,t}^2) \log \left(\frac{p(x_t^1 | \boldsymbol{X}_{p,t}^1, \boldsymbol{X}_{p,t}^2)}{p(x_t^1 | \boldsymbol{X}_{p,t}^2)} \right) dx_t^1 \, d\boldsymbol{X}_{p,t}^1 \, d\boldsymbol{X}_{p,t}^2 . \end{aligned} \tag{4.7}$$

Interestingly, introducing the widely used Kullback-Leibler (KL) divergence D_{KL} between to probability densities

$$D_{KL}(p \| q) = \int p(x) \log \frac{p(x)}{q(x)} dx \tag{4.8}$$

to quantify the discrepancy between two distributions, TE can be rewritten as an expected KL-divergence between the corresponding conditional probabilities,

$$TE(X^2 \to X^1) = \mathbb{E}_{(\boldsymbol{X}_{p,t}^1, \boldsymbol{X}_{p,t}^2)} \left[D_{KL} \left(p(X_t^1 | \boldsymbol{X}_{p,t}^1, \boldsymbol{X}_{p,t}^2) \| p(X_t^1 | \boldsymbol{X}_{p,t}^1) \right) \right] . \tag{4.9}$$

Section 4.2.4.2 will clarify why TE is not applying counterfactual analysis, but here we still refer the second conditional as the quotation-marked "counterfactual" to keep symmetry with CS/DCS and rDCS

We will abusively denote (making the expectation with respect to conditioning variables implicit)

$$TE(X^2 \to X^1) = D_{KL} \left(p(X_t^1 | \boldsymbol{X}_{p,t}^1, \boldsymbol{X}_{p,t}^2) \| p(X_t^1 | \boldsymbol{X}_{p,t}^1) \right) . \tag{4.10}$$

This formulization can be understood as a probabilistic comparison between the actual and "counterfactual" conditions via KL divergence.

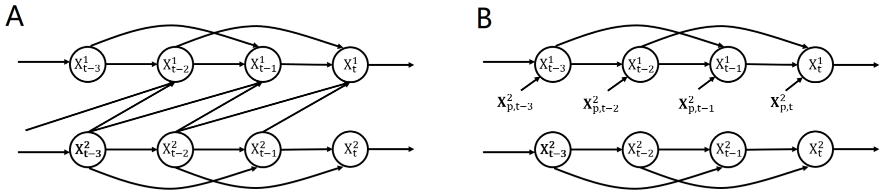

Figure 4.1: Illustration of the intervention implemented in CS. (A) Structural causal model of a bi-variate VAR(2) model defined in Eq. 4.13 and Eq. 4.14 with uni-directional coupling from X^2 to X^1. (B) The intervention implemented in devising CS is to break the causal arrows and send an independent copy $\boldsymbol{X}^2_{p,t}$ to X^1_t at each time point. This diagram applies to both CS and DCS (Section 4.2.3.2). The difference is: CS assumes a homogeneous VAR model while DCS depends on a inhomogeneous VAR model.

As noticed in [8], under stationary VAR assumptions the analytic expression of KL divergence between Gaussian distributions applied to Eq. 4.10 leads directly to

$$\mathrm{GC}(X^2 \to X^1) = \mathrm{TE}(X^2 \to X^1)\,, \tag{4.11}$$

such that TE appears as a generalization of GC. Therefore, in the following we will only focus on TE in the time series models under Gaussian assumptions.

4.2.2.3 *Causal strength*

To overcome the limitation of TE and GC (see Section 4.2.4 for a discussion), [3] has proposed a measure of *information flow* to quantify the influence of some variables on others in a system, which has been further studied and generalized in [111] as a measure of the *causal strength* (CS) of an arbitrary set of arrows in a graphical model.

This measure respects an intuitive notion of locality: the strength of the influence mediated by a given set of arrows pointing to an effect variable depends only on the input to it (the marginal of the input variable) and on the other arrows pointing to the same variables. A particular instantiation of this measure for time series will be introduced in Section 4.2.3.2.

In the context of our bivariate time series, CS also addresses the difference between the actual and counterfactual conditions with KL divergence. The actual condition is the same as devised in TE/GC where X^1_t is assumed to be dependent on both $\boldsymbol{X}^1_{p,t}$ and $\boldsymbol{X}^2_{p,t}$.

Following [3, 111] we can construct the counterfactual condition by implementing the intervention that replaces the arrow $\boldsymbol{X}^2_{p,t} \to X^1_t$ by an arrow injecting instead ${\boldsymbol{X}^2_{p,t}}'$ as an independent copy of $\boldsymbol{X}^2_{p,t}$ with the same marginal, as illustrated in Figure. 4.1A. Importantly, compared to [111] but in line with [3], we propose to replace the multivariate vector $\boldsymbol{X}^2_{p,t}$ by a copy without enforcing independence between the components of this vector, in order to preserve the structure of the SCM. The intervention distribution results in the entailed conditional probability while the marginals $p(\boldsymbol{X}^2_{p,t}) = p({\boldsymbol{X}^2_{p,t}}')$:

notations here for intervention is consistent with the general form in Section 4.2.1.2

$$\begin{aligned} p^{do(X_t^1 := f(\boldsymbol{X}_{p,t}^1, \boldsymbol{X}_{p,t}^{2}{}', \eta_t^1))}(X_t^1|\boldsymbol{X}_{p,t}^1) &= \sum_{\boldsymbol{X}_{p,t}^{2}{}'} p(X_t^1|\boldsymbol{X}_{p,t}^1, \boldsymbol{X}_{p,t}^2) p(\boldsymbol{X}_{p,t}^{2}{}') \\ &= \sum_{\boldsymbol{X}_{p,t}^2} p(X_t^1|\boldsymbol{X}_{p,t}^1, \boldsymbol{X}_{p,t}^2) p(\boldsymbol{X}_{p,t}^2) \end{aligned}$$

Then KL divergence quantifies the distance between $p(X_t^1|\boldsymbol{X}_{p,t}^1, \boldsymbol{X}_{p,t}^2)$ and $p^{do(X_t^1 := f(\boldsymbol{X}_{p,t}^1, \boldsymbol{X}_{p,t}^{2}{}', \eta_t^1))}(X_t^1|\boldsymbol{X}_{p,t}^1, \boldsymbol{X}_{p,t}^2)$, leading to the (implicit) form of CS

$$\begin{aligned} &CS(X^2 \to X^1) \\ &= \mathbb{E}_{(\boldsymbol{X}_{p,t}^1, \boldsymbol{X}_{p,t}^2)} \left[D_{KL}(p(X_t^1|\boldsymbol{X}_{p,t}^1, \boldsymbol{X}_{p,t}^2) \| p^{do(X_t^1 := f(\boldsymbol{X}_{p,t}^1, \boldsymbol{X}_{p,t}^{2}{}', \eta_t^1))}(X_t^1|\boldsymbol{X}_{p,t}^1, \boldsymbol{X}_{p,t}^2)) \right] . \end{aligned} \tag{4.12}$$

4.2.2.4 *Causal impact and regression discontinuity.*

In line with the potential outcome framework [198], Brodersen et al. [23] introduced a Bayesian framework to quantify the causal impact of an intervention at a given time point n on an observed time series $\{y_k\}$. As illustrated in Figure 4.2, it relies on observed pre-intervention data $\{y_1, ..., y_n\}$ covariates and priors on time series parameters to extrapolate a distribution of potential outcome sample paths $\{y_{\tilde{n+1}}, ..., y_{\tilde{m}}\}$ under the counterfactual condition that no intervention had been performed. Comparing the posterior predictive density of these unobserved counterfactual responses to the observed time course $y_{n+1}, ..., y_m$ (under intervention) thus allows to quantify the effect of the intervention. Such contrasting strategy is also present in a variant of regression discontinuity designs in economics and social sciences. In particular, *regression discontinuity in time* assesses causal effects by comparing outcomes distributions on time intervals before and after the onset of a policy change [93]. One key difference of these approaches with respect to TE and CS is the contrasting of properties of the models over different time intervals, one before the intervention and one after. One specificity of such framework is that the interventional distribution is often observed (e.g. a patient is treated at a given time), while the counterfactual (unobserved) scenario is one where no intervention is performed (what would have happend if the patient had not been treated). This strategy makes intuitive sense in the context of non-stationary data, while TE and CS have not been designed for capturing such effects, as they have been applied chiefly in static (for CS) or stationary settings (for TE and the natural extension of CS to time series, DCS, described in the Section 4.2.3.1 and Section 4.2.3.2).

4.2.3 *Time-varying extension of causality measures*

The neural events are by nature non-stationary, hence for modeling these signals with stochastic difference equations, while keeping the benefits of linearity, we will extend the causality measures (TE and CS) in Section 4.2.2 into the inhomogeneous VAR models, where the coefficient of matrices in Eq. 4.2 are assumed time dependent. Consistent with the compact form in Section 3.2.4.2, the bi-variate non-stationary VAR model takes the form

$$X_t^1 = \mathbf{a}_t^\top \boldsymbol{X}_{p,t}^1 + \mathbf{b}_t^\top \boldsymbol{X}_{p,t}^2 + \eta_t^1 , \quad \eta_t^1 \sim \mathcal{N}(k_t^1, \sigma_{1,t}) . \tag{4.13}$$

$$X_t^2 = \mathbf{c}_t^\top \boldsymbol{X}_{p,t}^1 + \mathbf{d}_t^\top \boldsymbol{X}_{p,t}^2 + \eta_t^2 , \quad \eta_t^2 \sim \mathcal{N}(k_t^2, \sigma_{2,t}) , \tag{4.14}$$

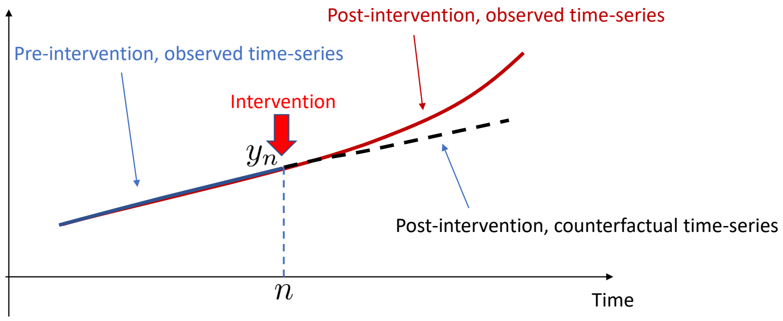
Post-intervention, observed time-series
Pre-intervention, observed time-series
Intervention
y_n
Post-intervention, counterfactual time-series
n
Time

Figure 4.2: Illustration of the principle of the regression discontinuity in time, and causal impact methodologies. The observed time series where an intervention occurs at a specified time point n is contrasted with the counterfactual scenario that no intervention was performed.

The model can be represented in the framework of SCM illustrated in Figure 4.1A.

The estimation of inhomogeneous VAR model parameters and statistics should follow the procedures introduced in Section 3.2.4.2. Compared to stationary VAR model, the time-varying version requires estimating the probabilities $p(X_t^1|\boldsymbol{X}_{p,t}^1, \boldsymbol{X}_{p,t}^2)$ and $p(X_t^1|\boldsymbol{X}_{p,t}^1)$ with the samples obtained at each time point, instead of gathering the samples along the whole time series (see Section 3.2.4.2). For models based on arbitary distributions, the conditional distributions need to be estimated empirically. Fortunately, with our Gaussian assumption of the VAR model, the expressions of time-varying TE and CS can be derived in explicit forms.

As both TE and CS are defined as the KL divergence of two conditionl probabilities representing the actual and counterfactual conditions (which are Gaussian according to the joint Gaussian assumption), we first present here the general form of the KL divergence between two univariate Gaussians $\mathcal{N}(\mu_a, \sigma_a^2)$ and $\mathcal{N}(\mu_c, \sigma_c^2)$:

$$D_{KL}(\mathcal{N}(\mu_a, \sigma_a^2)||\mathcal{N}(\mu_c, \sigma_c^2)) = \log\frac{\sigma_c^2}{\sigma_a^2} - \frac{1}{2} + \frac{\sigma_a^2 + (\mu_a - \mu_c)^2}{2\sigma_c^2} \tag{4.15}$$

This equation, as derived in Appendix A.5.1, enables us to calculate TE and CS after establishing the explicit expression of the two conditional distributions accordingly.

For both measures, the actual condition is described by Eq. 4.13, where

$$\mathcal{N}(\mu_a, \sigma_a^2) = p(X_t^1|\boldsymbol{X}_{p,t}^1, \boldsymbol{X}_{p,t}^2)\,.$$

The corresponding mean and variance of the conditional probability, as derived in Appendix A.5.3, take the form:

$$\mu_a = \mathbf{a}_t^\top \boldsymbol{X}_{p,t}^1 + \mathbf{b}_t^\top \boldsymbol{X}_{p,t}^2 + k_t^1 \tag{4.16}$$

$$\sigma_a^2 = \sigma_{1,t} \tag{4.17}$$

The conditional probability of the counterfactual condition will be shown individually for TE and CS in the following.

4.2.3.1 *TE based on non-stationary VAR models*

For TE, as the conditional probability representing the "counterfactual" condition takes the form

$$\mathcal{N}(\mu_c, \sigma_c^2) = p(X_t^1 | \boldsymbol{X}_{p,t}^1)\,.$$

Resulting from the same model in Eq. 4.13, the mean and variance can be derived as (for details ses Appendix A.5.3)

$$\mu_c = \mathbf{a}_t^\top \boldsymbol{X}_{p,t}^1 + \mathbf{b}_t^\top \mathbb{E}[\boldsymbol{X}_{p,t}^2 | \boldsymbol{X}_{p,t}^1] + k_t^1 \tag{4.18}$$

$$\sigma_c^2 = \mathbf{b}_t^\top \mathrm{Cov}[\boldsymbol{X}_{p,t}^2 | \boldsymbol{X}_{p,t}^1] \mathbf{b}_t + \sigma_{1,t} \tag{4.19}$$

Plugging the expressions of μ_a, μ_c, σ_a^2 and σ_c^2 into Eq. 4.15, the KL divergence can be derived as

$$\mathrm{TE} = \mathrm{D}_{\mathrm{KL}}(\mathcal{N}(\mu_a,\, \sigma_a^2) || \mathcal{N}(\mu_c,\, \sigma_c^2)) = \log \frac{\mathbf{b}_t^\top \mathrm{Cov}[\boldsymbol{X}_{p,t}^2 | \boldsymbol{X}_{p,t}^1] \mathbf{b}_t + \sigma_{1,t}}{\sigma_{1,t}}$$

Notably, "Cov" here refer to the covariance operations, while "Σ" refers to defined variable representing the covariance matrix

As $\boldsymbol{X}_{p,t}^1$ and $\boldsymbol{X}_{p,t}^2$ are jointly Gaussian, the conditional variance takes the form

$$\mathrm{Cov}[\boldsymbol{X}_{p,t}^2 | \boldsymbol{X}_{p,t}^1] = \Sigma_{\boldsymbol{X}_p^2} - \Sigma_{\boldsymbol{X}_p^1 \boldsymbol{X}_p^2} \Sigma_{\boldsymbol{X}_p^1}^{-1} \Sigma_{\boldsymbol{X}_p^2 \boldsymbol{X}_p^1}\,.$$

Covariance matrices $\Sigma_{\boldsymbol{X}_p^1} = \mathrm{Cov}[\boldsymbol{X}_{p,t}^1]$, $\Sigma_{\boldsymbol{X}_p^1 \boldsymbol{X}_p^2} = \mathrm{Cov}[\boldsymbol{X}_{p,t}^1, \boldsymbol{X}_{p,t}^2]$, $\Sigma_{\boldsymbol{X}_p^2} = \mathrm{Cov}[\boldsymbol{X}_{p,t}^2]$ and $\Sigma_{\boldsymbol{X}_p^2 \boldsymbol{X}_p^1} = \mathrm{Cov}[\boldsymbol{X}_{p,t}^2, \boldsymbol{X}_{p,t}^1]$ are parts of the lagged covariance matrix $\Sigma_{\boldsymbol{X}_{p,t}}$ defined in Eq. 3.12.

Therefore, the expression of time-varying TE should be

$$\mathrm{TE} = \frac{1}{2} \log \frac{\sigma_{1,t} + \mathbf{b}_t^\top \Sigma_{\boldsymbol{X}_p^2} \mathbf{b}_t - \mathbf{b}_t^\top \Sigma_{\boldsymbol{X}_p^1 \boldsymbol{X}_p^2} \Sigma_{\boldsymbol{X}_p^1}^{-1} \Sigma_{\boldsymbol{X}_p^2 \boldsymbol{X}_p^1} \mathbf{b}_t}{\sigma_{1,t}} \tag{4.20}$$

4.2.3.2 *Dynamic causal strength (DCS)*

The counterfactual probability in CS, as explained in Section 4.2.2.3, is the interventional distribution after removing the causal arrow from $\boldsymbol{X}_{p,t}^2$ to X_t^1 and substituting it with an independent copy $\boldsymbol{X}_{p,t}^2{}'$such that $p(\boldsymbol{X}_{p,t}^2{}') = p(\boldsymbol{X}_{p,t}^2)$:

$$\mathcal{N}(\mu_c, \sigma_c^2) = p^{\mathrm{do}(X_t^1 := f(\boldsymbol{X}_{p,t}^1, \boldsymbol{X}_{p,t}^2{}', \eta_t^1))}(X_t^1 | \boldsymbol{X}_{p,t}^1, \boldsymbol{X}_{p,t}^2)\,.$$

The corresponding model is:

$$X_t^1 = \mathbf{a}_t^\top \boldsymbol{X}_{p,t}^1 + \mathbf{b}_t^\top \boldsymbol{X}_{p,t}^2{}' + \eta_t^1\,, \quad \eta_t^1 \sim \mathcal{N}(k_t^1,\, \sigma_{1,t})\,.$$

Mean and variance of this counterfactual probability are derived as (for derivation see Appendix A.5.5):

$$\mu_c = \mathbf{a}_t^\top \boldsymbol{X}_{p,t}^1 + \mathbf{b}_t^\top \mathbb{E}[\boldsymbol{X}_{p,t}^2] + k_t^1 \tag{4.21}$$

$$\sigma_c^2 = \mathbf{b}_t^\top \mathrm{Cov}[\boldsymbol{X}_{p,t}^2] \mathbf{b}_t + \sigma_{1,t} \tag{4.22}$$

After plugging in the terms and reorganizing the terms, we can reach the expression of DCS

$$\mathrm{DCS} = \mathrm{D}_{\mathrm{KL}}(\mathcal{N}(\mu_a,\, \sigma_a^2) || \mathcal{N}(\mu_c,\, \sigma_c^2)) = \frac{1}{2} \log \frac{\mathbf{b}_t^\top \mathrm{Cov}[\boldsymbol{X}_{p,t}^2] \mathbf{b}_t + \sigma_{1,t}}{\sigma_{1,t}} \tag{4.23}$$

In this paper, we call this time-varying measure (absolute) Dynamic Causal Strength (DCS). We chose a different name than the ones proposed in these papers to avoid confusion between this measure and both the ones proposed in the literature and the novel measure we propose in Section 4.2.5.

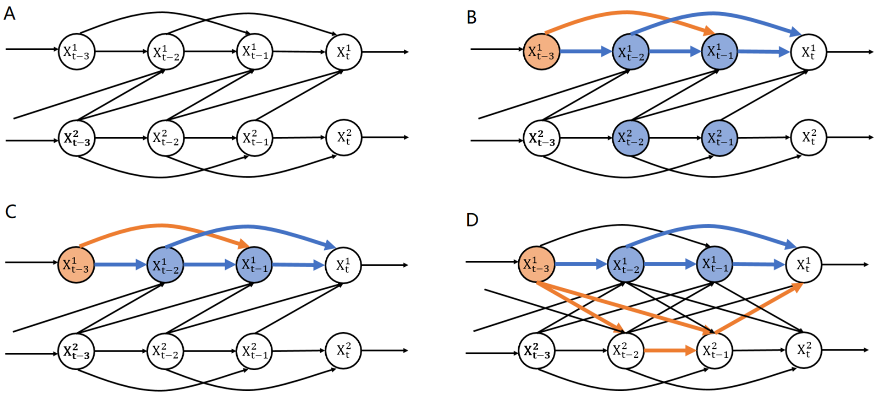

Figure 4.3: D-separation of bi-variate VAR(2) model. (A) Structural causal model of a bi-variate VAR(2) model defined in Eq. 4.13 and Eq. 4.14 with uni-directional coupling from X^2 to X^1. (B) Conditioning on both past states of X^1 and X^2 blocks all paths from X^1_{t-3} to X^1_t. Blue nodes represents conditioned nodes while blue arrows marks blocked paths. Orange arrows marks the unblocked paths. (C) Conditioning on past states of X^1 alone blocks all paths from X^1_{t-3} to X^1_t in the uni-directional case. Color codes are the same as (B). (D) Conditioning on past states of X^1 alone does not block all paths from X^1_{t-3} to X^1_t in the bi-directional case. Color codes are the same as (B).

4.2.4 *Discussion of the candidate causality measures*

In general, TE and GC statistics are two commonly used measures of causal strength for investigating interactions between brain regions (e.g., [244]). Based on the observational conditional distribution of the neural signals being analyzed, these two measures estimate a quantity that is easily interpretable from a forecasting perspective. However, potential problems would arise when they are interpreted based on SCMs and when they are applied to non-stationary signals like transient events.

These problems are often neglected during applications, while we discuss several of them in this section. We will show that DCS is better in some aspects but still not suitable for dealing with a typical case in analyzing transient events.

4.2.4.1 *TE estimation is non-local*

The first critical problem to be addressed is that the extension of TE or GC to time-varying versions is largely limited by its non-locality.

In the original definitions of GC (Section 4.2.2.1) and TE (Section 4.2.2.2), we remain imprecise regarding the number of past states p used for the forecast in the model. While one can exploit classical model order selection techniques (e.g. AIC and BIC discussed in Section 3.2.4.1) to select the best order for the full model, in case of bi-directional coupling, the reduced model of Eq. 4.5 is misspecified (in a generic case) for any finite order.

This can be easily seen by exploiting the d-separation criterion presented in Section 3.2.3.1, as illustrated in Figure 4.3. Figure 4.3A shows the SCM of a second-order stationary VAR model with uni-directional coupling from X^2 to X^1, which is represented by causal arrows from $\boldsymbol{X}^2_{p,t} = [X^2_{t-2}, X^2_{t-1}]^\top$ to X^1_t. Figure 4.3B shows the estimation in the full model, where conditioning on both $\boldsymbol{X}^1_{p,t}$ and $\boldsymbol{X}^2_{p,t}$ blocks all the path from X^1_{t-3} to X^1_t. According to the "chain" rule of d-separation, X^1_{t-3} and X^1_t are conditional independent (conditioned on $\boldsymbol{X}^1_{p,t}$ and $\boldsymbol{X}^2_{p,t}$).

For such an uni-directionally-coupled system, a finite order for the reduced model also guarantees such conditional independence, as seen in Figure 4.3C where all paths are blocked by conditioning. However, in the same system with bi-directional coupling, for any $k > p$ (i.e. $k > 2$), there is always a path from X^1_{t-k} to X^1_t going through nodes of X^2 that is unblocked by $(X^1_{t-p}, \cdots, X^1_{t-1})$. As Figure 4.3D shows, 2 paths from X^1_{t-3} to X^1_t are not blocked by conditioning on $\boldsymbol{X}^1_{p,t}$. Under faithfulness assumptions, this implies that there is conditional dependence between X^1_t and its remote past samples, no matter on how many finite past states we are conditioning on. This further implies that to minimize the forecast error of X^1_t in the reduced model one should ideally exploit the past information of this time series up to $p = +\infty$.

This issue has been both raised and addressed in the literature, in particular by resorting to Autoregressive Moving Average models and state space models for defining an appropriate reduced model (e.g. [9, 216]). However, this remains a important limitation when extending TE to time-varying versions, where the model is assumed to be stationary only locally in time. For example, in the non-stationary VAR model (Eq. 4.13 and Eq. 4.14), we assume a constant linear model in a 1-point time window. The non-locality of TE is particularly problematic for such time varying model assumption because with the local stationarity, only a limited number of past states can be observed. Thus it is impossible to retrieve the infinite past state to estimate an appropriate reduced model.

4.2.4.2 *TE does not implement an intervention*

Another interesting way of understanding the problem of TE is to look at the "counterfactual" condition it defines.

As introduced in Section 4.2.2.2, for a uni-directionally-coupled system where X^2 causes X^1. In the actual condition, X^1_t is dependent on both $\boldsymbol{X}^1_{p,t}$ and $\boldsymbol{X}^2_{p,t}$ (as in the full model) while the "counterfactual" condition assumes X^1_t is dependent only on $\boldsymbol{X}^1_{p,t}$ (as in the reduced model).

Actually, the "counterfactual" condition ignores the influence of past states of X^2 on X^1, which corresponds to marginalizing an SCM (i.e., removing observed variables from the graph). In other words, the "counterfactual" condition is modelling the observational data with another SCM (i.e., another model), violating the basic idea of counterfactual analysis. Intuitively, this is problematic because the counterfactual analysis in TE does not compare the actual and counterfactual conditions based on their different outcomes in the same model (e.g., different observed distributions) but by interpreting the same data with different models. In this case, the "counterfactual" outcomes given that X^1_t is dependent only on $\boldsymbol{X}^1_{p,t}$ is neither observed or reconstructed based on the observed data, which is why TE only reflects observational predictive dependency instead of causality.

By comparison, as addressed in Section 4.2.1.2 and Section 4.2.3.2, this shortcoming of TE can be overcome by DCS, where the intervention in the SCM framework enables the reconstruction of the counterfactual probabilities where the causal link does not exist.

4.2.4.3 *TE underperforms for strongly synchronized signals*

Besides, it has also been pointed out that the definition of TE in Eq. 4.10 has some other non-intuitive implications [3, 111]. In particular, there are situations in which $\mathrm{TE}(X^2 \to X^1)$ almost vanishes, although the influence is intuitively clear. How frequent are the practical situations in which we have these detrimental effects is unclear; however, theoretical analysis suggests that this can happen when the time series are strongly correlated.

To see this, we can derive with Eq. 4.9 in the case where x^2 is a deterministic function of x^1 such that TE vanishes. In an extreme case where X_t^2 is propotional to X_t^1 such that $X_t^2 = kX_t^1$, representing a time-wise synchronization of the two signals, the conditional variance will be

$$\Sigma_{\boldsymbol{X}_p^2|\boldsymbol{X}_p^1} = \Sigma_{\boldsymbol{X}_p^2} - \Sigma_{\boldsymbol{X}_p^2\boldsymbol{X}_p^1}\Sigma^{-1}_{\boldsymbol{X}_p^1}\Sigma_{\boldsymbol{X}_p^1\boldsymbol{X}_p^2} = \Sigma_{\boldsymbol{X}_p^2} - k\Sigma_{\boldsymbol{X}_p^2}\cdot(\frac{1}{k^2}\Sigma^{-1}_{\boldsymbol{X}_p^2})\cdot k\Sigma_{\boldsymbol{X}_p^2} = 0$$

Plugging into Eq. A.5.3 yields,

$$\mathrm{TE}(X^2 \to X^1) = \log\frac{\mathbf{b}_t^\top \mathrm{Cov}[\boldsymbol{X}_{p,t}^2|\boldsymbol{X}_{p,t}^1]\mathbf{b}_t + \sigma_{1,t}}{\sigma_{1,t}} = \log\frac{\sigma_{1,t}}{\sigma_{1,t}} = \log 1 = 0$$

However, strong correlation between two observed time series does not usually imply that causal interactions between them are week, from an SCM perspective. We will illustrate this theoretical prediction in Section 4.3.1 and compare with the results of DCS to show that DCS does not suffer from this non-intuitive vanishing problem.

4.2.4.4 *Insensitivity of TE and DCS to deterministic perturbations*

While several intuitive properties make DCS a good candidate to quantify causal influences, we exhibit a counterintuitive property common to TE and DCS in the context of peri-event time series. It is common in neuroscience to observe evoked potentials, a waveform that appears consistently in response to a stimulus. More generally, neural activities in relation to events are likely to have a deterministic component appearing in the peri-event snapshot distribution (potentially even after the Desnap bias correction implemented in Chapter 3), and such component may reflect mechanisms involved in causal influences.

Consider an example bi-variate VAR(1) model in the following form

$$X_t^1 \coloneqq \eta_t^1, \tag{4.24a}$$

$$X_t^2 \coloneqq cX_{t-1}^1 + dX_{t-1}^2 + \eta_t^2, \tag{4.24b}$$

with $c, d \neq 0$ and a stationary innovation for X^2, $\eta_t^2 \sim \mathcal{N}(0,1)$, but a non-stationary innovation for X^1, $\eta_t^1 \sim \mathcal{N}(\alpha\delta_{t\,t_0}, 1)$, which models a (soft) intervention on the innovation expectation through the unit impulse at time t_0 represented by the Kronecker delta

$$\delta_{tt_0} = \begin{cases} 1, & \text{for } t = t_0\,, \\ 0, & \text{otherwise}\,. \end{cases}$$

Then it can be easily shown that the expected time course of X^2 is

$$\mathbb{E}X_t^2 = \begin{cases} \alpha c d^{t-t_0+1}, & t \geqslant t_0 + 1 \\ 0, & \text{otherwise.} \end{cases}$$

This witnesses the causal influence of $X^1_{t_0}$ on subsequent values of X^2_t at subsequent time, which for large α results in large deviations from the baseline expectation of X^2_t for t prior to t_0.

Paradoxically, for a broad class of models that includes the example above, such a deterministic causal influence cannot be detected by TE or DCS.

Proposition 3. *For a linear VAR model, TE and DCS values are unaffected by interventions on the innovations' expectation at any time point.*

Proof. Consider the order p VAR model of Eq. (3.3). Consider the intervention at time t_0 that transforms η_{t_0} in $\eta_{t_0} + \alpha$. To compute the intervention distribution of the new variables denoted $(\tilde{X}^1, \tilde{X}^2)$ changes with respect to the distribution of the original variables, we can examine the difference with respect to (X^1, X^2) that has the same innovations, except for $\eta^1_{t_0}$ for which we remove a constant α. (X^1, X^2) is then distributed according to the original distribution (before intervention), and the difference $(U, V) = (\tilde{X}^1 - X^1, \tilde{X}^2 - X^2)$ follows the equations

$$\begin{aligned} U_t &= \mathbf{a}^\top \mathbf{U}_{p,t} + \mathbf{b}^\top \mathbf{V}_{p,t} + \delta_{t\,t_0} \\ V_t &= \mathbf{c}^\top \mathbf{U}_{p,t} + \mathbf{d}^\top \mathbf{V}_{p,t} \end{aligned}$$

which is a deterministic difference equation with a unique solution making $\boldsymbol{X}$ and $\tilde{\boldsymbol{X}}$ coincide before the intervention[1] (u_t, v_t). As a consequence, the intervention distribution $\widetilde{P}$ is a shifted version of the original distribution:

$$\widetilde{P}(x^2_t, \boldsymbol{X}^2_{p,t}, \boldsymbol{X}^1_{p,t}) = P(x^2_t - u_t, \boldsymbol{X}^2_{p,t} - \mathbf{u}_{p,t}, \boldsymbol{X}^1_{p,t} - \mathbf{v}_{p,t})$$

which implies the same for conditional marginal distributions, e.g.

$$\widetilde{P}(x^2_t | \boldsymbol{X}^2_{p,t}) = P(x^2_t - u_t | \boldsymbol{X}^2_{p,t} - \mathbf{u}_{p,t}, \boldsymbol{X}^1_{p,t} - \mathbf{v}_{p,t})$$

As a consequence TE on the intervention distribution writes

$$\begin{aligned} \mathrm{TE}(\widetilde{X}^1 \to \widetilde{X}^2) &= \int \widetilde{p}(x^2_t, \boldsymbol{X}^2_{p,t}, \boldsymbol{X}^1_{p,t}) \log \frac{\widetilde{p}(x^2_t | \boldsymbol{X}^2_{p,t}, \boldsymbol{X}^1_{p,t})}{\widetilde{p}(x^2_t | \boldsymbol{X}^2_{p,t})} \, dx^2_t d\boldsymbol{X}^2_{p,t} d\boldsymbol{X}^1_{p,t} \\ &= \int p(x^2_t - u_t, \boldsymbol{X}^2_{p,t} - \mathbf{u}_{p,t}, \boldsymbol{X}^1_{p,t} - \mathbf{v}_{p,t}) \log \frac{p(x^2_t - u_t | \boldsymbol{X}^2_{p,t} - \mathbf{u}_{p,t}, \boldsymbol{X}^1_{p,t} - \mathbf{v}_{p,t})}{p(x^2_t - u_t | \boldsymbol{X}^2_{p,t} - \mathbf{u}_{p,t})} dx^2_t d\boldsymbol{X}^2_{p,t} d\boldsymbol{X}^1_{p,t} \\ &= \int p(x^2_t, \boldsymbol{X}^2_{p,t}, \boldsymbol{X}^1_{p,t}) \log \frac{p(x^2_t | \boldsymbol{X}^2_{p,t}, \boldsymbol{X}^1_{p,t})}{p(x^2_t | \boldsymbol{X}^2_{p,t})} \, dx^2_t d\boldsymbol{X}^2_{p,t} d\boldsymbol{X}^1_{p,t} = \mathrm{TE}(X^1 \to X^2) \,. \end{aligned}$$

The same reasoning can be applied to DCS leading to invariance as well. □

This result is not what we would expect from a measure of influence, because in the above example of Eq. 4.24, setting a large α intuitively leads to a large influence of X^1 on X^2 provided $c \neq 0$. Provided that TE and DCS can be made arbitrary small by reducing Σ_t , TE and DCS would detect no influence despite this strong effect on the mean of X^2_t.

4.2.5 *A novel measure: relative Dynamic Causal Strength*

4.2.5.1 *Motivation*

To deal with the problem that neither TE nor DCS is applicable in the case where the transient events are driven by a deterministic exogenous input,

[1] Because initial conditions of this deterministic linear system are set to zero before the intervention at t_0

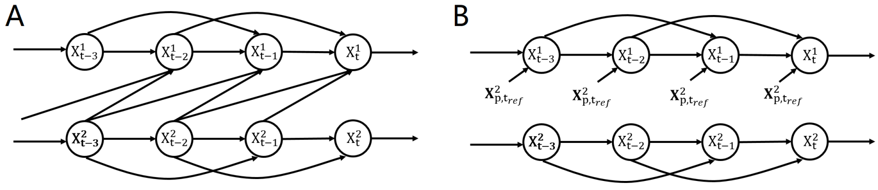

Figure 4.4: Illustration of the intervention implemented in rDCS. (A) Structural causal model of a bi-variate VAR(2) model defined in Eq. 4.13 and Eq. 4.14 with uni-directional coupling from X^2 to X^1. (B) The intervention implemented in devising rDCS is to break the causal arrows and send an independent copy of the stationary state $\boldsymbol{X}^2_{p,t_{ref}}$ to X^1_t at each time point.

we propose a novel measure, the relative Dynamic Causal Strength (rDCS), as an improved version of DCS with modifications specifically designed for this problem.

Recalling the causality principles discussed in Section 4.2.1.3, the cause and the mechanism triggered by the cause to generate the effect should be considered separately. In the specific problem we are investigating, the cause is the past states of X^2 as $\boldsymbol{X}^2_{p,t}$, while the mechanism can be represented by the model in Eq. 4.13 and symbolized by the corresponding causal arrow in the SCMs. In the measures we have introduced so far, DCS only exploits the case where the causal arrow is deleted as a counterfactual condition but does not address the change in the cause itself.

In the case where X^2 experiences a deterministic exogenous input in a transient window, the cause increases significantly; thus, intuitively, the causal effect should also be enhanced even if the causal arrow remains the same (i.e., the coefficient $\mathbf{b}$ stays unchanged). According to the principle of independence between cause and mechanism, apart from intervening on the causal arrow, further intervention can be implemented on the cause node to construct a counterfactual condition where the cause receives no time-varying innovations.

Therefore, inspired by causal impact (Section 4.2.2.4) which characterizes the difference between the current state and a baseline state, we propose (additionally to DCS) to replace the marginal of $\boldsymbol{X}^2_{p,t}$ by the marginal $\boldsymbol{X}^2_{p,t_{ref}}$ for a reference period t_{ref}. The reference period t_{ref} is typically chosen to be a stationary period before the occurence of the transient deterministic pertabations and statistics of $\boldsymbol{X}^2_{p,t_{ref}}$ can be averaged by statistics of $\boldsymbol{X}^2_{p,t}$ within this period. This leads to the *relative Dynamical Causal Strength* (rDCS)

$$\begin{aligned}&\mathrm{rDCS}^{t}_{t_{ref}}(X^2 \to X^1) = \\ &\mathbb{E}_{(\boldsymbol{X}^1_{p,t}, \boldsymbol{X}^2_{p,t})}\left[D_{KL}(p(X^1_t|\boldsymbol{X}^2_{p,t}, \boldsymbol{X}^2_{p,t})\|p^{do(X^1_t := f(\boldsymbol{X}^1_{p,t}, \boldsymbol{X}^2_{p,t_{ref}}, \eta^1_t))}(X^1_t|\boldsymbol{X}^1_{p,t}, \boldsymbol{X}^2_{p,t}))\right]\end{aligned} \tag{4.25}$$

Intuitively, the *relativeness* originates from the comparison between the current past states $\boldsymbol{X}^2_{p,t}$ and the reference past states $\boldsymbol{X}^2_{p,t_{ref}}$. It is then natural to predict that in the uni-directional case, $\mathrm{rDCS}(X^2 \to X^1) = \mathrm{DCS}(X^2 \to X^1)$ for any reference time t_{ref} if X^2 is stationary because stationarity implies that the marginal distributions of $\boldsymbol{X}^2_{p,t_{ref}}$ and $\boldsymbol{X}^2_{p,t}$ are identical. As a particular case, this result implies that a transient loss of causal link from X^2 to X^1 will lead to $\mathrm{rDCS} = 0$, while for a stationary bivariate system, $\mathrm{DCS} = \mathrm{rDCS}$ is constant.

4.2.5.2 *Time-varying implementation*

Similar to the time-varying TE and DCS (Section 4.2.3.1 and Section 4.2.3.2), rDCS can be estimated in the instantaneous manner with an explicit form under the Gaussian assumption.

The actual condition where $\mathcal{N}(\mu_a, \sigma_a^2) = p(X_t^1 | \boldsymbol{X}_{p,t}^1, \boldsymbol{X}_{p,t}^2)$ still holds for the calculation of rDCS (Section 4.2.3). For the counterfactual condition with the probability

$$\mathcal{N}(\mu_c, \sigma_c^2) = p^{do(X_t^1 := f(\boldsymbol{X}_{p,t}^1, \boldsymbol{X}_{p,t_{ref}}^2, \eta_t^1))}(X_t^1 | \boldsymbol{X}_{p,t}^1, \boldsymbol{X}_{p,t}^2)$$

where the model describing the intervened SCM (Figure 4.4B) is

$$X_t^1 = \mathbf{a}_t^\top \boldsymbol{X}_{p,t}^1 + \mathbf{b}_t^\top \boldsymbol{X}_{p,t_{ref}}^2{}' + k_t^1 + \eta_t^1.$$

Similarly, as elaborated in Appendix A.5.5, the mean and variance can be derived as

$$\mu_c = \mathbf{a}_t^\top \boldsymbol{X}_{p,t}^1 + \mathbf{b}_t^\top \mathbb{E}[\boldsymbol{X}_{p,t_{ref}}^2] + k_t^1 \tag{4.26}$$

$$\sigma_c^2 = \mathbf{b}_t^\top \mathrm{Cov}[\boldsymbol{X}_{p,t_{ref}}^2] \mathbf{b}_t + \sigma_{1,t} \tag{4.27}$$

where the covariance variance of the reference state takes the form

$$\mathrm{Cov}[\boldsymbol{X}_{p,t_{ref}}^2] = \mathbb{E}[(\boldsymbol{X}_{p,t_{ref}}^2 - \mathbb{E}[\boldsymbol{X}_{p,t_{ref}}^2])(\boldsymbol{X}_{p,t_{ref}}^2 - \mathbb{E}[\boldsymbol{X}_{p,t_{ref}}^2])^\top].$$

By putting these statistic into the expression of KL-divergence in Eq. 4.15

$$D_{KL}(\mathcal{N}(\mu_a, \sigma_a^2) || \mathcal{N}(\mu_c, \sigma_c^2))$$

we can obtain the explicit form of rDCS

$$\begin{aligned} \mathrm{rDCS}(X^2 \to X^1) = \frac{1}{2} \log \frac{\sigma_{1,t} + \mathbf{b}_t^\top \mathrm{Cov}[\boldsymbol{X}_{p,t_{ref}}^2] \mathbf{b}_t}{\sigma_{1,t}} - \frac{1}{2} \\ + \frac{1}{2} \cdot \frac{\sigma_{1,t} + \mathbf{b}_t^\top \mathbb{E}[(\boldsymbol{X}_{p,t}^2 - \mathbb{E}[\boldsymbol{X}_{p,t_{ref}}^2])(\boldsymbol{X}_{p,t}^2 - \mathbb{E}[\boldsymbol{X}_{p,t_{ref}}^2])^\top] \mathbf{b}_t}{\sigma_{1,t} + \mathbf{b}_t^\top \mathrm{Cov}[\boldsymbol{X}_{p,t_{ref}}^2] \mathbf{b}_t} \end{aligned} \tag{4.28}$$

4.3 RESULTS

In the Results section, we first focus on illustrating the properties of TE, DCS and rDCS with toy models in simulations. The problem of vanishing TE occurring with synchronized signals and the benefits of DCS in the same situation will be validated in Section 4.3.1. Next, we will simulate a simple uni-directionally coupled VAR system with rhythmic perturbations of the cause variable to generate transient events, where we will show that rDCS is able to reflect the change of causal effects due to this perturbation while TE and DCS fail. From this example, we will raise an interesting phenomenon: the choice of reference point to align the peri-event panel data influences the estimation of causality measures. This relates to the selection bias issue investigated in Chapter 3.

To address this problem, we will combine the selection bias correction algorithm (i.e., the *DeSnap* algorithm) introduced in Chapter 3 with the calculation of causality measures based on the corrected model parameters. We will show the causality estimates based on such a combination better reflect the underlying state-dependent dynamics. This is validated with the two-state stationary process we investigated in Section 3.3.2, the biophysically realistic simulation of thalamocortical spindles and recordings of hippocampal SPW-Rs.

4.3.1 Validation on strongly-correlated signals

As mentioned in Section 4.2.2.3, TE underperforms when the cause and effect signals are strongly correlated with each other, while DCS are unaffected by the synchrony between the pair of signals under analysis.

Here, to illustrate such contrast, we simulated a bivariate dynamical system in the form of two synchronized continuous harmonic oscillators $x(t)$ and $y(t)$, with uni-directional coupling (i.e., $x(t)$ driving $y(t)$):

Here $x(t)$ and $y(t)$ are irrelevant to the rest of the thesis

$$\begin{cases} \frac{d^2x}{dt^2} &= -2\zeta_x \omega_x \frac{dx}{dt} - \omega_x^2 x + n_x\,, \\ \frac{d^2y}{dt^2} &= -2\zeta_y \omega_y \frac{dy}{dt} - \omega_y^2 y + cx + n_y\,. \end{cases} \tag{4.29}$$

In this system, $x(t)$ is designed as an under-damped oscillator ($\zeta_x = 0.015722$) which approximately oscillates at a limit cycle with a period $T_x = 200$ and angular frequency $\omega_x = 2\pi/T_x = 0.0314$. To achieve synchrony, $y(t)$ is also designed as an under-damped oscillator ($\zeta_y = 0.2$) whose intrinsic oscillation gradually vanishes and finally follows the oscillation of $x(t)$ with a coupling strength of $c = 0.098$. For $y(t)$, $T_y = 20$, $\omega_y = 2\pi/T_y = 0.314$. We also add some small Gaussian white innovations to both oscillators: $n_x \sim \mathcal{N}(0, 0.02)$, $n_y \sim \mathcal{N}(0, 0.005)$. Adding this noise allows fitting a VAR model to the the signals to assess the causal interactions with TE and DCS. VAR parameter estimation would fail with deterministic signals by causing the covariance matrix estimates to be singular.

Using the Euler method with a time step of 1 and random initial points ($\mathcal{N}(0,1)$), we simulated 2000 trials of this uni-directionally coupled system with 1000-point length. We discarded the first 500 points to ensure that the time series reaches a sufficient level of synchronization. We can see this system as a stationary VAR(2) process because numerical simulation with the Euler method generates data with its past two states.

Figure 4.5 (left panel) shows the results of time-varying TE and DCS for assessing the causal effects between $x(t)$ and $y(t)$. Calculation is performed in both the ground truth direction ($x(t) \to y(t)$) and the opposite direction.

We first look at the control experiment. Consistent with the system's stationarity, TE is constant in both directions while being higher in the ground-truth direction. DCS in the ground-truth direction stays at a relatively high level, despite some small oscillation under a frequency similar to the intrinsic oscillation frequency of $x(t)$.

With respect to the detection of causal direction, both measures are able to detect the correct direction (i.e., causation for $x(t) \to y(t)$ is much larger than in the opposite direction). It is also reasonable that DCS in both directions is higher than TE, according to its definition in section 4.2.3.2. However, from the control experiment, we cannot conclude that the smaller TE values are due to its definition or due to the strong synchrony in the signals.

Therefore, we introduced a transient decrease of the noise variance in the cause signals ($x(t)$). The logic of designing this transient change is the following: the level of synchronization will increase with weaker noise, but the system and input remain the same because the noise change is negligible to the signal amplitude; thus if TE is suitable for synchronized signals, its values are expected to stay constant. However, as the results show in Figure 4.5 (right panel), there is a transient decrease of TE during the interval where noise variance is decreased, suggesting that TE performs poorly in the cases where the cause and effect signals are synchronized.

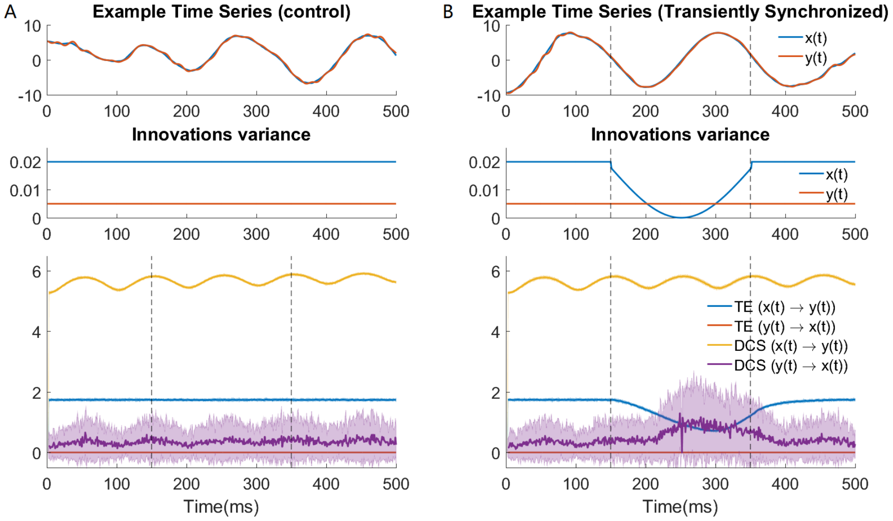

Figure 4.5: TE fails when the signals are strongly synchronized. (A) Control experiments where synchrony is not changed. (top) Example of the bivariate signal in the control experiment. (middle) Time-varying design of innovation's variance for both variables in the control experiment. (bottom) Time-varying TE and DCS results in the control experiment. (B) TE underperforms during transient increased synchrony induced by a tiny change in noise variance. The transient change can be seen as an event. Subfigure designs are the same as (A).

4.3.2 *Validation on simulated perturbation events with nonzero innovations*

In this section, we directly address the benefits of rDCS over the other proposed causality measures when applied to signals driven by deterministic perturbations. To illustrate this specific property, we designed some simple transient events perturbing the innovation parameters of a stationary process with uni-directional coupling. The events are generated by feeding the cause signal with non-zero time-varying innovations such that both signals will exhibit temporal oscillations. We refer to these events as *perturbation events* in the following sections.

Interestingly, exploiting this toy model draws attention to a problem of event alignment, which will be extensively explained in this section.

4.3.2.1 *Simulation procedure*

We simulated a non-stationary uni-directionally-coupled autoregressive system defined in Eq. 4.13 and Eq. 4.14. The causal direction is $X^2 \to X^1$. The system is designed as a bivariate VAR(4) process with a time-invariant coefficient matrix: $\mathbf{a}^\top = [-0.55, -0.45, -0.55, -0.85]$, $\mathbf{b}^\top = [1.4, -0.3, 1.5, 1.7]$, $\mathbf{c}^\top = [0, 0, 0, 0]$ and $\mathbf{d}^\top = [0.9, -0.25, 0, 0.25]$. Uni-directional interactions is ensured by setting the coupling strength in the opposite direction (i.e. $\mathbf{c}$) as zero for all lags. All entries of this coefficient matrix is randomly generated and fixed to guarantee the stability of the VAR(4) system.

We implement the non-stationarity in η_t^2, the innovations of 'cause' process $\{X_t^2\}$. Both innovations η_t^1 and η_t^2 are drawn from a Gaussian distribution with a variance of 0.1 (with no correlation in between, i.e. $\mathrm{Cov}[\eta_t^1, {\eta_t^2}^\top] = 0$); the difference is that $\mathbb{E}[\eta_t^1] = k_t^1 = 0$ while $\mathbb{E}[\eta_t^2] = k_t^2$ is non-zero and time-varying. We designed the time-varying profile of k_t^2 as a Morlet-shaped waveform to mimic the oscillatory properties of neural event signals: $k_t^2 = H\exp(-(\alpha x)^2/2)\cos(5\alpha x)$, where $\alpha = 2/25$ is a constant controlling the event duration, and $H = 10$ is the amplitude of the highest peak in the center of the event. The total duration of the Morlet-shaped waveform is 101 ms. The innovation's mean designed for both variables are shown in Figure 4.6A (bottom).

We generated 1000 trials of the VAR(4) process with a length of 700 ms (and a sampling frequency of 1kHz). The initial 200 points are discarded to ensure that the time series is stationary, resulting in 500 points for each trial.

From the simulations, we can extract a dataset of multi-trial event snapshots in the following way: after detecting local maxima of amplitude exceeding 5SD a chosen signal (see below), we define this maximum as a reference point of the event in each trial and put it at the center the peri-event time-window of 400 ms (for better readability, we only visualize a 120-ms window around the event, with 60 ms at each side of the alignment position). The averaged event waveform is illustrated in Figure 4.6A (top). We will show in the next section that based on different choices of the reference point (i.e., different alignment methods), the model estimation and the calculation of causality measures will perform dramatically differently.

4.3.2.2 *Effect of alignment on model estimation and causality measures*

The criterion for choosing the reference points determines how events are aligned across multiple trials. While alignment might seem a trivial step at first blush, its influence on VAR model estimation turns out to be critical. Intuitively, as the model estimation depends on the assumption that all observations at each time point are *i.i.d.* samples, jittered alignments violate the assumption and lead to biased estimation of the time-varying model parameters.

In the following parts, we explore the effect of alignments by comparing the model estimation and the performance of causality measures in the three scenarios: 1) *ground-truth alignment*; 2) *aligning by the cause*; 3) *aligning by the effect*.

Ground-truth alignment refers to the case where the detection point is the peak of innovations mean of the cause signal. As the innovation is designed as deterministic (i.e., identical) across trials, with this alignment, the data points gathered across trials at each time point can be assumed sampled *i.i.d.* from the original innovation distribution. The averaged waveform of the snapshots with this alignment is presented in Figure 4.6A (top).

However, empirically, even if there is a deterministic input to all the recorded trials, the innovation is a hidden variable in the system that cannot be observed directly. Practically, detection points have to be chosen from the observed signals, i.e., either the cause signal or the effect signal. Then it is almost inevitable that different trials in a snapshot will be extracted with small time shifts (i.e., jitters), as in various commonly-adopted criteria of alignment (e.g., alignment on the power peak or the onset time). In our simulation, *aligning by the cause* means aligning by the peak of $\{X_t^2\}$, and *aligning by the effect* means aligning by the peak of $\{X_t^1\}$.

We are then interested in how these different alignment methods and the resulted time shifts would affect the performance of causality measures, especially in the time-varying case that we focus on.

The multi-trial BIC criteria we proposed in Section 3.2.4.1 is able to reconstruct the true model order: 4. We fitted the extracted events with VAR(4) model with non-zero innovations mean (Eq. 4.13 and Eq. 4.14), so that they

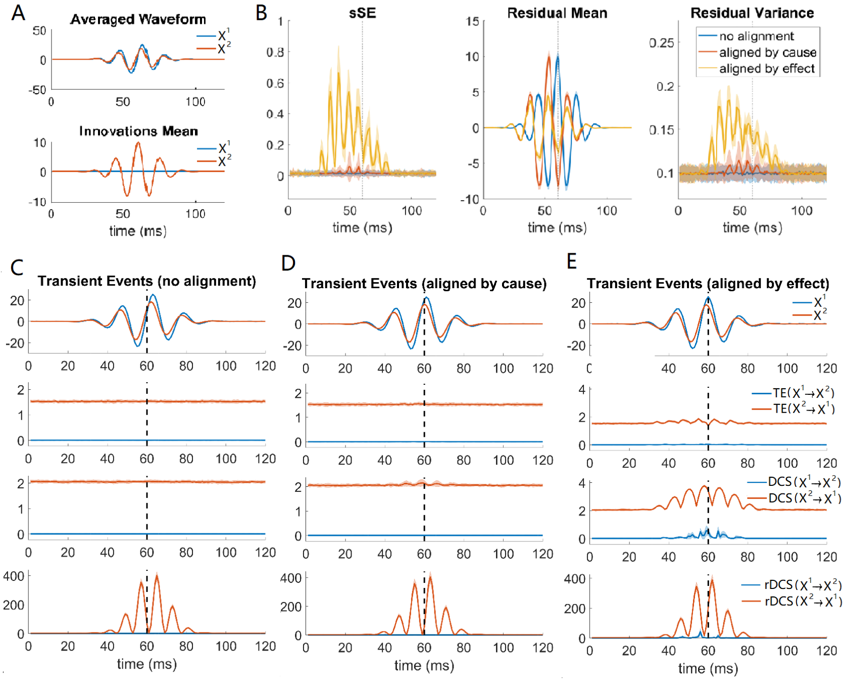

Figure 4.6: Causal analysis for simulated perturbation events with non-zero innovations. (A) (top) Trial-averaged signals with *ground-truth alignment* (aligned on innovations mean). (bottom) Deterministic innovation added into cause and effect signals. (B) (left) Sum of squared errors (sSE) of coefficient matrix estimation in three cases: *ground-truth alignments*, *aligning by the cause* and *aligning by the effect*. For the case *ground-truth alignments* the gray dashed line marks the peak of the innovation; for the other two cases it marks the peak of the signals. (middle) Residual mean as an estimation of the innovations mean in the same three cases as in (left). The red line is lagged compared to the blue one is because a delay before the innovations peak causes the signals' peak. (right) Estimated innovations (residual) variance in the same three cases. The ground-truth innovations variance is 0.1. (C) Averaged event waveform and causality measures (TE, DCS and rDCS) results of the perturbation events with *ground-truth alignments*. The gray dashed line is the same as in (B). (D) The same as (d) with *aligning by the cause*. (E) The same as (d) with *aligning by the effect*.

can be directly applied to calculate the causality measures TE, DCS and rDCS.

Figure 4.6B shows the accuracy of inhomogeneous VAR coefficient estimation. The accuracy is represented by the summed squared error across all the matrix coefficients at each time point (denoted as 'sSE' for 'summed squared error'). In the scenario *ground-truth alignment*, the sSE remains relatively constant over time. This is consistent with the model design, i.e., the coefficient matrix is set to be time-invariant, also suggesting that coefficient estimation is rather accurate in this case. With the scenario of *aligning by the cause* ($\{X_t^2\}$), the accuracy is similar to the scenario *ground-truth alignment*, despite some small fluctuations around the center of the event. This is reasonable because aligning on the peak of the cause signals is almost equivalent to aligning on the innovation when the innovation's mean is strong enough relative to the innovation's variance (i.e., the deterministic part of the signal is stronger than its stochastic part). In comparison, due to intrinsic dynamics, the peaks of cause signals that appear around the same time point may arrive at the effect signals with some shifts in time. Thus the scenario *aligning by the effect* induces strong errors in coefficient estimation.

origin of the alignment problem

The idea that *aligning by the effect* induces time shifts is confirmed by the estimation of innovation as residual mean and variance, as shown in Figure 4.6B (middle, right). In the scenario of *ground-truth alignment*, the amplitude and shape of the non-stationary innovation are fully reconstructed, together with its variance estimated around the true value 0.1. It is similar to the scenario of *aligning by the cause*, despite some small deviation of the variance. In the scenario of *aligning by the effect*, the dis-alignment of innovations peak lowers the amplitude of the estimated residual mean and induces large errors in the residual variance.

Figure 4.6C, Figure 4.6D, Figure 4.6E show the corresponding results of how causality measures perform in the three alignment scenarios. During the periods where no transient events occur, all three measures are able to detect a time-invariant stronger causal effect in the ground-truth direction ($X^2 \rightarrow X^1$) compared to the opposite direction. Besides, in line with the theoretical predictions, DCS is higher than TE.

The different effects of alignment emerge during the time intervals where we add non-stationary non-zero innovations. In the first two scenarios, TE and DCS are almost constant across time, while rDCS shows stronger causal effects in the ground-truth direction. In line with the results for model estimation, the time-invariance properties of TE and DCS are due to the fact that innovations are well-aligned and regressed out by the model with a non-zero mean. Therefore TE and DCS, which do not depend on the mean of signals, would not show any time-varying difference related to the innovations mean. The rDCS is, on the other hand, dependent on the signals mean, and thus reveals a time-varying trend in a shape resembling the rectified profile of the cause signal., suggesting that rDCS is able to reflect both the existence of connectivity between two variables but also the change of the cause variable.

By comparison, the time shifts caused by *aligning on the effect* lead to an increased residual variance resulting from the poor alignment of the innovations' mean profile across trials. Thus there is a time-varying pattern of causal effects detected by all the causality measures.

As a summary so far, in the presence of deterministic innovations, rDCS is able to reveal the causal effect caused by both the cause and the mechanism, while TE and DCS are only able to account for the effect caused by the mechanism. Interestingly, different alignment methods lead to different results of TE and DCS due to temporal jittering. While in practice, such jittering can be considered as temporal smoothing, which does not affect much causal analysis based on time-varying models, in the next section, we show that alignment can be problematic when the events exhibit deterministic waveforms.

4.3.3 *Application to simulated thalamocortical spindle oscillations*

Type-I and Type-II spindles

The Costa model is able to generate two types of spindles (for example, see Figure 4.7B). During the N2 stage of NREM sleep, spindles can be triggered by cortical K-complexes (in the cortical Pyr neurons) through thalamocortical interactions (referred to as *Type-I*) or generated spontaneously in the thalamus (*Type-II*) by the hyperpolarization of its pacemaker, the GABAergic neurons in the reticular thalamic nucleus. After initial generation, both types of spindles are transmitted to the cortex, and the Type-I cortical spindles typically co-occur with the positive peak of the K-complexes. These two spindles are both within the spindle oscillation range (7-12 Hz) but are visually distinguishable by their amplitude, duration and waveform.

detection procedure of two subtypes

We simulated the thalamocortical model for 2×10^8 ms (56 minutes) with a sampling frequency of 1000 Hz. To detect spindle waveforms, we set a threshold in the membrane potentials of the thalamocortical neurons (denoted as V_t for TC population): peaks of V_t above -52 mV indicate the existence of spindles; a detected spindle is classified as Type-I if the highest peak of V_t is above -45 mV; otherwise, they are sorted as Type-II spindles. A typical Type-I spindle can be 1.5 sec in length, while the duration of a typical Type-II spindle can be 2-3 sec.

After the detection procedure, we obtained 1282 Type-I and 1221 Type-II spindle events that are simultaneously simulated in TC and Pyr neurons. All the spindles, each as a trial, are aligned by their highest peak either in the membrane potential of TC neurons or Pyr neurons.

4.3.3.1 *Hypotheses on the ground truth for dominant directions of causation.*

As a critical point to validate the causality measures, we discuss here the "ground-truth" directions of these two subtypes of events. As TC and Pyr neurons are anatomically reciprocally connected in the model, here we are more interested in which neuronal population *drives* the other due to their internal dynamics, which can be superficially understood as the summed effects of both the connectivity and the activity of one of the neuronal population as the cause.

Interestingly, we cannot give a categorical ground truth answer to this question without making further assumptions. This is largely due to the complexity of the considered system, where defining interventions to assess causality would need to be carefully designed in order to preserve some properties of the subsystem while changing others. Such an approach would require a deep understanding of the system's dynamics in different interventional regimes, which is beyond the scope of this thesis. Instead, we rely on qualitative observations of the system's dynamics to make conjectures on the ground truth causality that we expect to infer with our approaches.

Based on visual inspection (e.g., Figure 4.7B) and theoretical exploration of the dynamics of the thalamocortical system ([242, 39]), we conclude that the causal interactions between Pyr and TC neurons for Type-I spindles

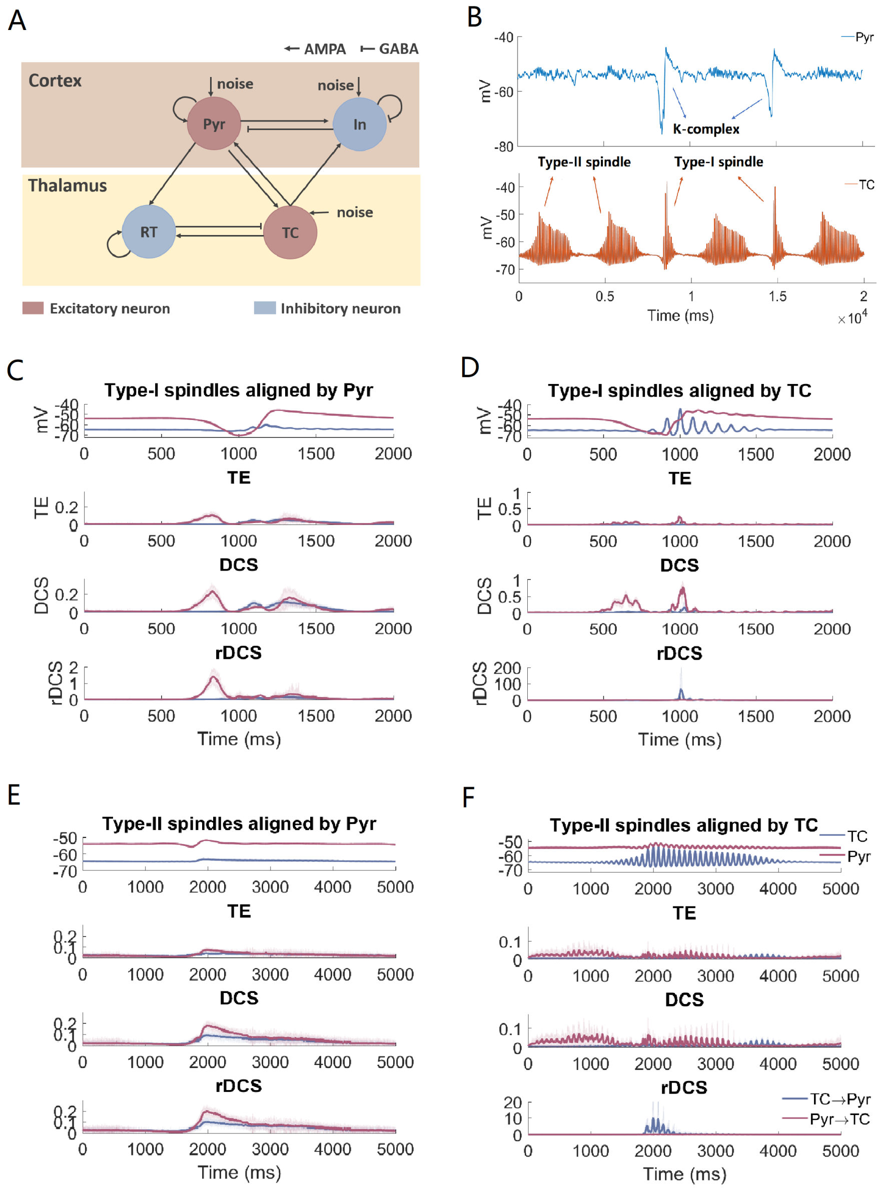

Figure 4.7: Causal analysis with simulated spindles in a thalamocortical neural mass model. (A) Model Structure as part of Figure 2.1B. The thalamocortical spindles are generated within a neural circuit that consists of four neuron types whose interactions are shown in the diagram. K-complexes and spindles are observed typically in the cortical pyramidal neurons and thalamocortical neurons. (B) Example waveforms of Type-I (and the associated K-complexes) and Type-II spindles in the Pyr and TC neurons during a short period of simulation. (C) Averaged event waveform and causality measures (TE, DCS and rDCS) for Type-I spindles aligned by Py neurons. Shades reflect time-varying standard deviations across 20 repeated simulation samples, as explained in the main text. (first row) Averaged waveforms for thalamo-cortical Type-I spindles aligned by Pyr neurons. (second - fourth row) time-varying TE, DCS and rDCS for Type-I spindles aligned by Pyr neurons. (D) Averaged event waveform and causality measures (TE, DCS and rDCS) for Type-I spindles aligned by TC neurons. Designs are the same as in (C). (E) Averaged event waveform and causality measures (TE, DCS and rDCS) Type-II spindles aligned by Pyr neurons. Designs are the same as in (C). (F) Averaged event waveform and causality measures (TE, DCS and rDCS) Type-II spindles aligned by TC neurons. Designs are the same as in (C).

undergoes three transient phases (qualitatively). As stated in Section 4.2.2, here we list the *dominant* causal direction in each phase:

1. 1^{st} phase: Pyr→TC

   In the first phase, *strong* negative noise perturbs the Pyr neurons and triggers the K-complex. The negative peak hyperpolarizes TC neurons and initiates Type-I spindles in the DOWN→UP transition. Thus the causal interactions in Pyr→TC dominate in this phase.

2. 2^{nd} phase: TC→Pyr

   In the second phase, the T-current in TC neurons is de-inactivated and activated, generating spindle-range oscillations in the TC neurons, which drives the spindle rhythms in Pyr neurons riding on the positive peak of the K-complex. Thus the dominant causal direction of this phase is TC→Pyr.

3. 3^{rd} phase: Pyr→TC

   The third phase starts shortly after the second phase and is partially overlapping with the latter. During the third phase, K-complex evolution in the Pyr neurons attenuates the spindle envelop in TC neurons, leading to its termination. Thus the causal interaction in the direction Pyr→TC is dominant again.

Similarly, circuit dynamics underlying Type-II spindles also consist of two phases:

for the different effect in Pyr triggered by strong/weak perturbations, one can refer to Figure 4 in [242]

1. 1^{st} phase: Pyr→TC

   In the first phase, *weak* negative perturbations entering the Pyr neurons does not trigger K-complex but instead, induce the slight de-inactivation of T-current in TC neurons that further develops into spindle-like rhythms. Thus the initial trigger of Type-II spindles in TC neurons is driven by Pyr neurons.

2. 2^{nd} phase: TC→Pyr

   Similar to the second phase of Type-I spindles, oscillations of Type-II spindles in TC neurons propagate to Pyr neurons after it has been initially triggered. Thus the dominant causal direction is TC→Pyr. Unlike Type-I spindles, Type-II spindles are terminated via the internal dynamics of TC neurons instead of by cortical K-complexes.

As such, the dominant drivers of both subtypes of spindles are transiently switching. Therefore this dataset does not provide a fixed “cause”->“effect” relation between the signals such that we can *align by the cause* or *align by the effect* the peri-event data, as was allowed in the case of the perturbation events (see Section 4.3.2). However, in the following section, we can still make interesting observations regarding the inferred causation depending on the alignment procedure.

4.3.3.2 *Model estimation and results of causality measures*

For both alignments of both types of spindles, the BIC selects reasonably small order for the VAR model (i.e., around 4). The results are shown in Figure 4.7C-F. The time-varying results are smoothed in a 20-ms window for better visualization.

Similar to the experiments of the perturbation events (Section 4.3.2), we also calculated the standard deviations of the causality measures for spindles. The variability originates from generating repeated simulations of the thalamocortical system for the same duration, performing the same detection of spindles, and calculating causality measures. We generated 20 trials of the same process, with the standard deviation plotted as shades in Figure 4.7C-F.

Consistent with results in the experiment of the perturbation events, TE for spindles for all four alignment cases are close to but smaller than DCS, and both are weaker than rDCS. rDCS reveals distinct profiles in all four cases, likely due to the highly deterministic nature of the simulated spindles.

When Type-I spindles are aligned by Pyr neurons (Figure 4.7C), all three causality measures show three peaks that match the three phases that we hypothesized for ground-truth causal interactions (Section 4.3.3.1): 1) a peak appearing at the initial phase of the K-complex in the direction Pyr → TC is in line with the first phase; 2) slightly higher causal effects from TC to Pyr than the other way round around the beginning of the thalamic spindle oscillations is consistent with the second phase; 3) causal effects from Pyr to TC are present from the beginning of spindles but become stronger after the positive peak of K-complex. This is in line with what we conjectured for the third phase. However, rDCS results amplify the first peak, which is likely due to alignment by the negative peak of the K-complex in Pyr neurons.

In the case where Type-I spindles are aligned by TC neurons (Figure 4.7D), the three peaks are still present in TE and DCS but the temporal relationship is slightly different, which is likely due to different alignment methods. More importantly, results in rDCS show an enormous peak in the direction TC→Pyr which overrides all other peaks. This is because by aligning on the peak of TC spindles, the events are highly synchronized in TC neurons, leading to a selection bias of the snapshots as introduced in Chapter 3. As rDCS calculation is dependent on the mean of events relative to the stationary state (Eq. 4.28), such large peaks in TC signals lead to very large values in rDCS in the direction where the TC signal serves as the cause. Such selection bias will also be seen in the next results.

For Type-II spindles, alignment has a stronger effect on the estimation of causality measures. When Type-II spindles are aligned by Pyr neurons (Figure 4.7E), spindle oscillations in TC neurons are greatly blurred, leading to small amplitudes. In this case, results for TE, DCS and rDCS all reveal bi-directional interactions during the spindle period (in the second phase of the conjectured ground truth interactions). However, in contrast to what we hypothesized as ground truth in Section 4.3.3.1 where causal effects from TC to Pyr dominate, these results show a reversed rank of the two directions. This is also likely caused by selection bias on Pyr activity which leads to larger amplitudes in the waveforms of Type-II spindles in Pyr neurons compared to TC spindles.

When Type-II spindles are aligned by TC neurons (Figure 4.7F), as in the same case in Type-I spindles, rhythmic peaks in thalamic spindles (TC activities) are well-aligned, leading to the transiently amplified rDCS results in the direction TC→Pyr due to selection bias. This peak in rDCS is consistent with the conjecture on the second phase of Type-II spindles, but its huge amplitude makes other temporal interactions invisible from the figure (i.e., the first phase).

To summarize, in the case of simulated thalamocortical spindles, aligning on either signal leads to opposite causal inference results. Specifically, while

TE and DCS are sometimes able to recover the ground-truth causal directions, rDCS is highly sensitive to the alignment methods: aligning by either signal amplifies the peak in the direction where this signal acts as the cause.

However, such sensitivity is not desired for event-based causal interactions. As we addressed in Section 3.1, the events are detected to reflect the transient network mechanisms defined as states. In the specific problem of spindle events, the state refers to the period when the T-current in the TC and RT neurons are de-inactivated, a pre-requisite to generate spindle-like rhythmic patterns (Section 1.1.3.2). Therefore, the de-inactivating states and spindle events are of similar length. Nevertheless, not all de-inactivation periods share exactly the same duration and temporal dynamics due to random noise, allowing some space for jittering when we align them. Therefore selection bias is possible to occur when we align the spindles by either the peak of Pyr or TC activities.

4.3.4 *Combination of DeSnap and causality measures*

In the results of the perturbation events and both subtypes of spindles, we observed that alignment significantly affects the characterization of causal interactions with all measures, especially rDCS. This is because the calculation of rDCS involves replacing the cause with a reference state, making it sensitive to the time-varying statistics estimated by the model. As briefly mentioned in Section 3.4, the selection bias correction method (*DeSnap* algorithm) we proposed in Chapter 3 can be combined with the rDCS measure to overcome this problem.

Intuitively, the causality measure calculated after applying the selection bias correction method reflects the state-dependent causal interaction between two systems in the brain. In the following sections, we will calculate the *DeSnapped* causality measure to validate their performance in a series of examples that we have already investigated in previous sections.

4.3.4.1 *DeSnapped causality analysis of oscillatory events in the two-state stationary system*

First, as a simple example, we will illustrate the correction of event-based causality into state-based causality with the oscillatory events we already explored in the two-state stationary system in Section 3.3.2. This bivariate system, consisting of variables X^1 and X^2, has been designed to be unidirectionally coupled ($X^2 \rightarrow X^1$, see Section 3.3.2.1). Figure 3.9 has illustrated comprehensively the correction process that proves effective in recovering the state-dependent parameters of the dynamics. Notably, we will skip the example of the oscillatory events detected in the uni-state stationary process, as explained in Section 3.3.1, because if the *DeSnapped* causality measures perform well enough for the two-state case, it naturally generalizes to the simpler uni-state model.

Here, based on the estimated inhomogeneous VAR model parameters, especially the coefficient matrix and the residual covariance, we are able to calculate the three causality measures (TE, DCS, rDCS) directly based on Eq. A.5.3, Eq. 4.23 and Eq. 4.28. . The comparison of revealed causal interactions between the uncorrected model and corrected model can be visualized in Figure 4.8.

For the detailed calculation of rDCS in the corrected model, we refer to Eq. A.29

Before the correction, all three measures are able to detect causal directions in line with the ground truth, i.e., X^2 causing X^1 but not in the oppo-

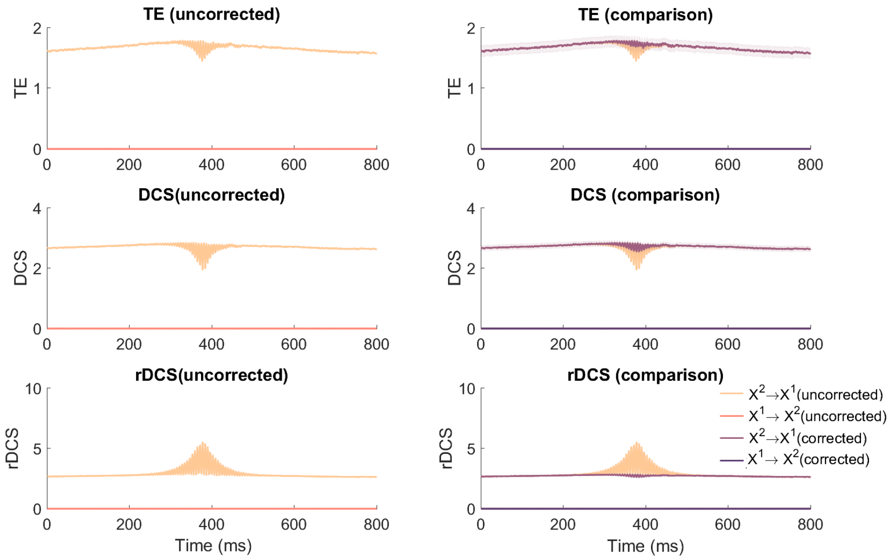

Figure 4.8: *DeSnapped* causality measures of events in two-state stationary VAR process. (left) Time-varying TE, DCS and rDCS calculated based on the uncorrected model. (right) Time-varying TE, DCS and rDCS calculated based on the corrected model after applying the *DeSnap* algorithm. For details see Section 3.3.2 and Section 3.2.6.2.

site direction. This is consistent with the ground truth direction. However, the time-varying causality measures are not constant across time (consistent with the state-dependent stationarity). Instead, we observe a transient oscillatory pattern in the calculated TE, DCS and rDCS in the ground truth directions, which can be explained by the following logic. Due to the threshold-based detection of the oscillatory events, in the uncorrected model, we are focused specifically on the period accompanied by strong oscillations in the signal (see Section 3.3.2). This is a biased characterization of the whole stationary state that affects the model parameter estimations, eventually resulting in the transient oscillatory pattern.

Consistent with the correction performance in the model dynamics in Figure 3.9 that recovers the state-dependent stationarity, we can see that the three causality measures calculated with the corrected model are rather constant across time. This matches the generation mechanism of the state that X^2 causing X^1 with constant connectivity (**b** in Eq. 4.13), but the cause remains unchanged across time.

This result suggests that combining the selection bias correction method with causality measures provides knowledge of the state-dependent causal interaction underlying events but not the event itself. In the following section, we will also show with the spindle example, (Section 4.3.3) the combination also provides a proper characterization of causal effects that avoids alignment-induced biases.

4.3.4.2 *DeSnapped causality analysis of simulated spindles*

Recalling the results obtained by applying the three causality measures on simulated spindles with different alignment methods (Section 4.3.3), we know that alignment greatly affects the characterization of time-varying thalamocortical causal interactions due to the deterministic waveform of the events. Here we show how applying the *DeSnap* algorithm before calculating the causality measures is helpful for the recovery of state-dependent causal interactions. The detection signal (Section 3.2.2.3) is obtained by calculating cosine similarities between windows of the original signals with the spindle templates in each neuronal population. The template is obtained by averaging the detected events in Section 4.3.3. The results of TE, DCS and rDCS are shown in Figure 4.9.

For Type-I spindles (Figure 4.9A), all three causality measures reveal the time-varying causal interactions that are consistent with the hypothesized ground truth we proposed in Section 4.3.3.1. In the first phase, where the initiation of K-complex in Pyr is supposed to drive TC activities, we observe a small peak in the direction Pyr→TC. In the second phase, where thalamic spindles in TC neurons drive cortical spindles in Pyr neurons, the results show stronger causal effects from TC to Pyr neurons. In the third phase, where spindle propagation still persists, the positive phase of the K-complex in Pyr neurons drives the termination of thalamic spindles, which is in line with the results of TE, DCS and rDCS.

Similarly, causality results for Type-II spindles, as illustrated in Figure 4.9B, match the two phases described as the hypothesized ground truth. In the first phase, weak perturbation to the Pyr neurons drives the de-inactivation of TC neurons, resulting in the peak (Pyr→TC) at the beginning of the spindles. In the second phase, thalamic spindles are transmitted to Pyr neurons while Pyr activities modulate the envelope of spindle oscillations. Therefore bi-directional interactions are detected while TC→Pyr is the dominant direction. The difference between TC, DCS and rDCS is that rDCS amplifies the contrast between the dominant and non-dominant directions due to its definition.

As a summary of this section, in comparison to the uncorrected causal analysis on the same datasets of thalamocortical spindles (Section 4.3.3), we found that after correction for the selection bias, all three causality measures are able to reveal causal interactions in line with the ground truth directions in a time-varying manner. Compared to TE and DCS, rDCS recovers the causal effect by the cause, which is consistent with its proposed benefits as discussed in Section 4.2.5. Notably, although the connectivities between the two neuronal populations are time-invariant in the model, TE and DCS do not show temporally constant causal interactions. This is likely due to the intrinsic non-linear interactions of neural mass models that cannot be addressed with causality measures based on linear inhomogeneous VAR models.

4.3.4.3 *Application to hippocampal SPW-Rs*

The validation so far for the combination of *DeSnap* algorithm and causality measures are all conducted with simulation signals. In this section, we apply the *DeSnapped* TE, DCS and rDCS on the hippocampal SPW-Rs to test the performance of the measures in real data.

"sr": stratum radiatum; "pl":pyramidal layer of CA1

As explored in Section 3.3.3.1, the *DeSnap* algorithm is effective to correct the selection bias for the internal dynamics underlying the SPW-Rs recorded in the "sr" and "pl" layer of the CA1 hippocampal subfield. Based on the cellular mechanism introduced in Section 1.1.3.3, the cell bodies of CA1 pyramidal neurons are located in the pyramidal ("pl") layer while their dendritic trees spread into the *stratum radiatum* ("sr") layer (illustrated in Figure 3.10)A (right). Therefore, the ground truth dominant causal direction underlying this system is "sr"→"pl".

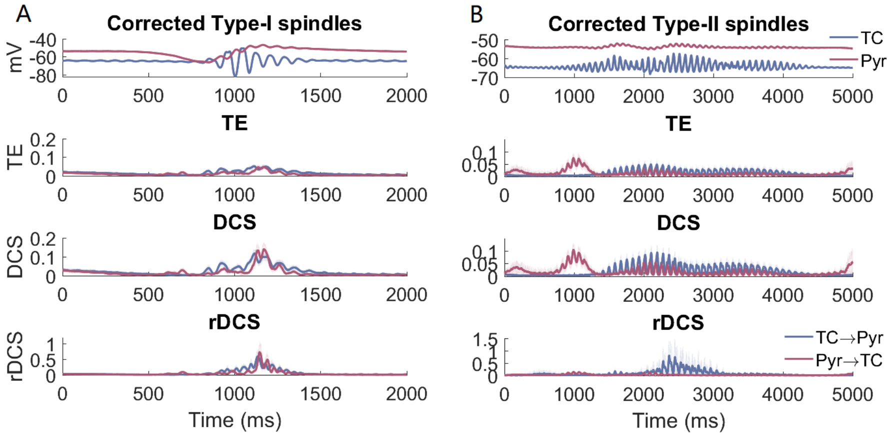

Figure 4.9: *DeSnapped* causal analysis of two subtypes of simulated spindles. (A) *DeSnapped* causality measures (TE, DCS and rDCS) for Type-I spindles. Shades reflect time-varying standard deviations across 20 repeated signal samples, as explained in the main text. (first row) Corredcted averaged waveforms for thalamo-cortical Type-I spindles in TC and Pyr neurons. (second - fourth row) Time-varying TE, DCS and rDCS for Type-I spindles calculated after applying the selection bias correction method. (B) *DeSnapped* causality measures (TE, DCS and rDCS) for Type-II spindles. Designs are the same as in (A).

Consistent with Section 4.3.4.1 and Section 4.3.4.2, we also calculate the peri-SPW-R causality measures with the peri-event snapshots detected and aligned on the peaks of either "sr" or "pl". This corresponds to the uncorrected model as defined in Section 3.2.6.2. The results are visualized as a comparison between the uncorrected and corrected model in Figure 4.10. The standard deviation plotted in the figure originates from the variability over 16 channel pairs.

The results show that regardless of alignment methods, all three causality measures are stronger in the ground truth direction ("sr"→"pl") at the stationary stages where events do not occur (i.e., at the edge of the peri-event window). This confirms their effectiveness in the stationary case, as explained in Section 4.2.2. TE and DCS appear relatively unaffected by different alignment methods, as shown in Figure 4.10A, B (second-third row). When the event occurs (i.e., around the center of the peri-event window), the dominant causal direction becomes unclear as TE/DCS in both directions exhibit transient peaks, although their values in the ground truth direction are generally higher. However, bias correction by performing the *DeSnap* algorithm has an effect on TE/DCS calculation. The peak in the ground truth direction is weakened after correction, suggesting that the peak might reflect spurious strong transient causal effects in the ground-truth direction.

The effect of bias correction is more significant for the calculation of rDCS. When the events are aligned by "sr" or "pl", the uncorrected rDCS both show a peak in the direction where the aligned variable acts as the cause. This is consistent with similar problems showed in Section 4.3.4.1 and Section 4.3.4.2, and supports our conjecture that the peaks in opposite directions are caused by selection bias. Indeed, after correcting, the peaks in opposite directions disappear while the corrected rDCS remains stronger in the ground truth direction.

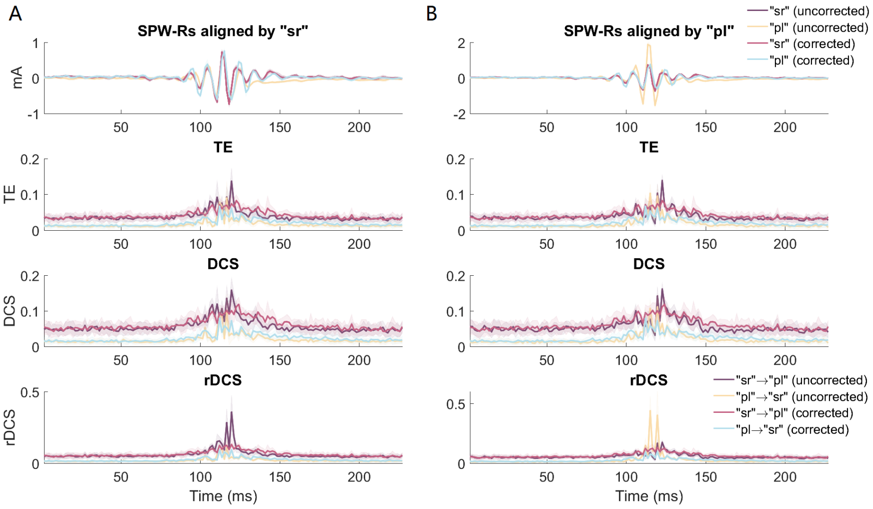

Figure 4.10: *DeSnapped* causal analysis of SPW-Rs aligned by "sr" and "pl". (A) *DeSnapped* causality measures (TE, DCS and rDCS) for SPW-Rs aligned by "sr". (first row) Corredcted and Uncorrected averaged waveforms for SPW-Rs in the regions "sr" and "pl". Shades reflect time-varying standard deviations across 16 channel pairs, as explained in the main text. (second - fourth row) Time-varying TE, DCS and rDCS for SPW-Rs calculated after applying the selection bias correction method. (B) *DeSnapped* causality measures (TE, DCS and rDCS) for SPW-Rs aligned by "pl". Designs are the same as in (A).

Although SPW-Rs are triggered by CA3 inputs into CA1, they are not necessarily deterministic

These results confirm in real data that the *DeSnap* algorithm is helpful to remove biased causal effects detected due to different alignment methods. Interestingly, the results on SPW-Rs do not show that rDCS performs better than TE/DCS. As we argued in Section 4.2.5, rDCS is a better causality measure when transient events result from deterministic perturbations, while the recorded SPW-Rs are not necessary triggered by such deterministic components. The weak causal effect in the non-ground-truth direction ("pl"→"sr") is likely due to the back-propagation within CA1 pyramidal neurons caused by the high-frequency recurrent activities within the local circuit in the "pl" region.

4.4 discussion

In summary, in this chapter, we have discussed the benefits and shortcomings of two time-varying causality measures (TE and DCS) in characterizing causal interactions based on peri-event data. To address their insensitivity to deterministic perturbations, we proposed a novel measure, the rDCS, that implements an intervention on both the cause and the mechanism in the SCM framework. We compared the performance of these causality measures on perturbation events with non-zero-meaned time-variant innovations, oscillatory events detected in stationary signals, simulated thalamocortical spindles and hippocampal SPW-Rs. The superiority of rDCS is supported by the perturbation events presented in Section 4.3.2. As causality analysis of transient events aims at uncovering the network mechanisms underlying these phenomena (e.g., addressing whether one event *drives* the other), we propose to use rDCS as, in theory, it captures both the effects due to changes in the cause and the propagation of the activity of the cause through anatomical connections.

However, we show that the data preparation procedure (i.e., the detection procedure and the alignment of the events) potentially affect the detection of causal effects and the quantification of their strength, especially rDCS. In principle, the directionality of peri-event causal interactions is a property of the state-dependent underlying transient mechanism and is not supposed to vary with the detection procedure. However, we showed non-intuitive results that the detected causal direction is dependent on the alignment methods (i.e., the triggering region). We hypothesize that this is due to the selection bias elaborated in Chapter 3, which is supported by the consistent alignment-dependent rDCS results in all the tested datasets. Actually, aligning neural events based on the activity in a single region is a common practice in event-related brain research. These results obtained with such peri-event data collection methods are then questionable because they may be affected by selection bias.

The solution to this problem is thus directly linked to the framework we proposed in Chapter 3. Events triggered by the peri-event peaks only reflect network dynamics when the observations exhibit certain patterns but not necessarily mapped to specific regions of the space of hidden states (see Section 3.2.2.1). Therefore the peak-triggered events reflect the time period that the observation is high-values, naturally resulting to a strong peak in causality measures in the direction where the triggering variable acts as the cause.

recalling that rDCS account for the change in the cause in addition to the mechanism, stronger observations corresponds to larger causes

Therefore, in order to reveal the state-dependent causal interactions based on peri-event data, we combined the calculation of causality measures with the selection bias correction algorithm (i.e., *DeSnap*) proposed in Chapter 3. The combination is straightforward because the time-varying causality measures are defined with the VAR model parameters, which are directly corrected with the *DeSnap* algorithm. We showed with different types of data that *DeSnapped* causality measure corrected for the alignment-dependence and recovers the real causal directions (Section 1.1.3.2 and Section 4.3.4.3).

Notably, as designed in Section 3.2.2.3, with *DeSnap* the peri-event snapshots are obtained with *reference points* $\{t_n | \tilde{D}_{t_n} \geqslant d_0\}$. Thus with neighboring reference points, we can extract temporally overlapping trials, leading to a smoothing of events. It is still not clear whether the weakened peak in the wrong direction is a result of the smoothing or due to the correction. Further investigations should be done to clarify the role of smoothing in bias correction

Part III

OUTLOOK

5 OUTLOOK

5.1 EXPLORING EVENT-TRIGGERED PLASTICITY WITH DATA-DRIVEN MODELLING

Theoretical predictions are expected to guide experimental explorations. Our theoretical results presented in Chapter 2 predicts that two subtypes of PGO waves trigger opposite plastic changes in the cortical networks. This modelling work exemplifies how to characterize the mesoscopic changes of synaptic strength based on the circuit activities. Specifically, such investigation can be based on the population STDP rule, which considers the cross-correlations of pre- and post-synaptic activities.

In models where all variables are accessible, it is natural to implement the STDP rule. However, if one wants to check the physiological existence of the opposite PGO-triggered plasticity, the two key variables, i.e. the presynaptic current and post-synaptic firing rate (especially the former) are technically hard to track. Indeed, measuring synaptic strength directly has been a challenge due to experimental limitations and sparsity of neurons. This is especially true when one aims to address population activities underlying PGO waves: with electrophysiological approaches it is technically challenging to visualize the vast number of the synapses or spines spanning the whole primary visual cortex with imaging techniques.

As a consequence, we seek a more indirect way to infer the changes procuded by the STDP rule from data. Considering that we already have one model of PGO waves that proves effective in the reproduction of many PGO-triggered features, we can assume that the mechanisms capture the major circuit interactions. To better reproduce the specific event-triggered network dynamics, we resort to finding more biologically plausible parameter sets for experimental PGO waves. Theoretically, this points to the problem of inferring parameters of differential equations describing a system whose underlying mechanisms are already known.

Data-driven modelling of neural systems (more broadly, data-driven dynamical systems) have long been a great challenge for the community. Dynamic causal models attempt to model the brain-wide systems with simplification of neuronal populations, but the shared model in each brain structure reduces its biological plausibility [75]. More generically, many methods in the field of nonlinear filtering have been applied to such problems [85, 125]. Still, these analytical methods are better suited for the characterization of simpler systems. More recently, simulation-based Bayesian inference approaches have also been developed to infer Hodgkin-Huxley mechanistic models of the single neuron activities [84]. While promising, the computational cost of Bayesian inference combined with repeated runs of an external (typically continuous-time) simulator remains an issue as the number of parameters increases. Our model consists of 16 differential equations describing the evolution of hidden variables, together with several linear output functions modelling the LFPs. Practically, estimation of this system with the above methods remains challenging. In contrast, we chose and approach in which both data-based optimization and mechanistic modelling are based

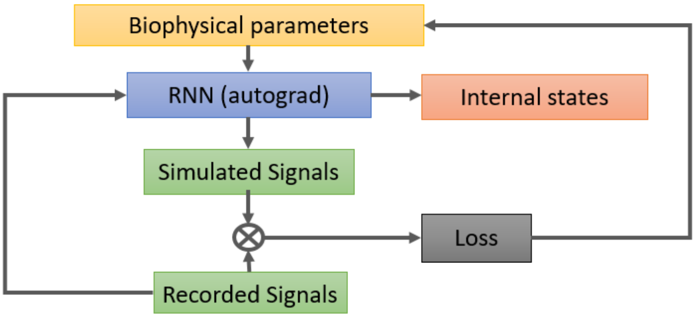

Figure 5.1: General framework of the hybrid model to learn internal dynamics from empirical data.

on a discrete-time dynamical model that nicely aligns with the VAR models exploited in Chapters 3-4.

We propose to adopt the current deep learning frameworks to train such systems with complicated dynamics and a large number of parameters. Specifically, we will present the design of an ongoing project. In this project, we seek to train the experimental peri-event data with a flexible type of Recurrent Neural Network (RNN), where the main idea is illustrated in Figure 5.1.

In brief, the internal structure of the RNN is adapted to incorporate the dynamics of a mechanistic model with unknown model parameters that entail biological meaning. The model is a simplification of the cortex module within the PGO model introduced in Section 2.2.3.1. By training the network with recorded peri-event data, we aim to retrieve the model dynamics constituting the internal states in the RNN. The training procedure is implemented by comparing the model output (as simulated signals) and the empirical (recorded) signals and minimizing a loss function defined as the distance between the simulated and empirical signals.

The detailed methodology will be elaborated in Section 5.2. We will explain the effectiveness and limits of such a model-fitting paradigm in recovering model parameters and hidden states based on preliminary results. Finally, we will discuss possible reasons why the training of RNN may fail and suggest potential solutions.

5.2 METHODOLOGY UNDERLYING DATA-DRIVEN MODELLING WITH RNN

5.2.1 *Simplified Neural Mass model of the cortex*

As mentioned in Section 5.1, models describing network dynamics in a neural system usually appear in the form of differential equations. The dimension of differential equations and the number of parameters will limit the performance of learning dynamical systems from data. Therefore, the model complexity should be carefully controlled.

To reduce model complexity, we simplified our PGO model into an isolated cortex model receiving thalamic inputs from TC neurons. The cortex structure, i.e., the Pyr and In neurons together with their connectivities, is preserved as we want to address intra-cortical plasticity. As seen in Figure 2.1B, both Pyr and In neurons receive inputs from TC neurons but not from RT neurons. Conversely, Pyr neurons send cortico-thalamic feedbacks

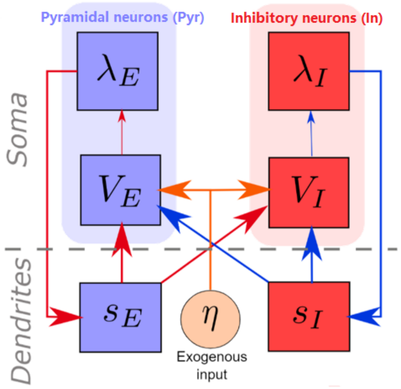

Figure 5.2: Sturctural illustration of the simplified cortical neural mass model. See the main text for explanation of the working mechanism. This figure is adapted from [200] with permission.

to both thalamic neurons. During bottom-up information transmission processes like the propagation of PGO waves, the thalamocortical connections are the major contributors to the circuit dynamics, while cortico-thalamic projections have minor effects. Therefore, we decide to replace the recurrent structure with a uni-directional thalamocortical projection from TC neurons to both Pyr and In neurons.

Another significant difference with respect to the original PGO model is the removal of the cortical sodium-modulated potassium channel, as introduced in Section 2.2.2.2. This is because it is designed as a phenomenological approximation of the cortical SO generation while the actual mechanism is far more complicated. Meanwhile, this current is the mechanism controlling the triggering of high-amplitude K-complexes in response to a strong perturbation to the Pyr neurons [39], suggesting a lack of variability in its dynamic behavior. Thus keeping this intrinsic current is limiting the model flexibility without gaining biophysical plausibility.

As a result, the simplified neural mass model of the cortex is illustrated as Figure 5.2. The Pyr neurons can be described by the following set of differential equations:

$$\tau_E \frac{dV_E}{dt} = -V_E + \nu_{E\to E} s_E - \nu_{I\to E} s_I + \eta \tag{5.1}$$

$$\tau_{Es} \frac{ds_E}{dt} = -s_E + \lambda_E \tag{5.2}$$

$$\lambda_E = \frac{Q_E}{1 + \exp\left(-\chi_E \left(V_E - V_{th,E}\right)\right)} \tag{5.3}$$

$$\mathrm{LFP}_E = \mu'_{E\to E} s_E - \mu'_{I\to E} s_I + \mu_\eta \eta \tag{5.4}$$

Similarly, the In neurons can be modelled with the following equations.

$$\tau_I \frac{dV_I}{dt} = -V_I + \nu_{I\to E} s_E - \nu_{I\to I} s_I + \alpha\eta \tag{5.5}$$

$$\tau_{Is} \frac{ds_I}{dt} = -s_I + \lambda_I \tag{5.6}$$

$$\lambda_I = \frac{Q_I}{1 + \exp\left(-\chi_I \left(V_I - V_{th,I}\right)\right)} \tag{5.7}$$

$$\mathrm{LFP}_I = \mu'_{E\to I} s_E - \mu'_{I\to I} s_I + \alpha\mu_\eta \eta \tag{5.8}$$

Eq. 5.1 and Eq. 5.5 describe the dynamics of membrane potentials of Pyr and In neurons (as V_E and V_I) processed in the soma. These two equations are consistent with the modelling of membrane potential adaptation (Eq. 2.18) as explained in Section 2.2.2.4. Intra-cortical synaptic currents s_E and s_I contributes to the change of both neuron populations with different synaptic strengths denoted as ν. The thalamic input η represents the thalamic LFPs generated mainly by thalamocortical synaptic currents from TC to both thalamic neurons. As the thalamocortical inputs are mainly influenced by the distance between neuronal populations, we model the differences between TC→Pyr and TC→In inputs with a scaling factor α. α takes the values from 0 to 1, reflecting the assumption that the TC→Pyr input is stronger than the TC→In input.

The membrane potentials are mapped to the population firing rates with a sigmoid function, yielding Eq. 5.3 and Eq. 5.7. In line with Section 2.2.2.3, the dynamics of the synaptic current depends on the convolution between the pre-synaptic firing rate and an alpha function representing the synaptic kinetics. This results in Eq. 5.2 and Eq. 5.6 describing the excitatory (AMPA) current s_E from Pyr neurons and the inhibitory (GABA) current s_I from In neurons.

Finally, consistently with Section 2.2.2.4, the cortical LFPs are modelled as a linear sum of the synaptic currents each neuronal population receives [154], as defined in Eq. 5.4 and Eq. 5.8.

5.2.2 *A bio-informed Recurrent Neural Networks*

Having established the aforementioned dynamical system as a simplified model of the cortex, in this section, we explain how to assimilate the dynamic structure into an RNN.

A typical RNN is comprised of 3 layers: an input layer, a hidden layer and an output layer, as illustrated in Figure. 5.3. The word "recurrent" refers to the loop that connects the hidden layer to itself. In fact, we can unfold such a network into a directed acyclic graph (also shown in Figure. 5.3) such that the dynamics of the output is only dependent on the hidden states defined within the hidden layer. The hidden layer dynamics only depends on the past hidden state and the current input at each time point. Compared to convolutional neural networks, such a recurrent structure incorporates a temporal relationship between the network inputs and outputs, which proves very effective for the data with sequence characteristics, e.g., time series data [213].

The specific RNNs commonly referred to in the community (as defined in PyTorch) use different units at fixed, successive time intervals. However, the idea of recurrence is encoded in the sharing of the parameters of the functional relationship linking units at successive time steps: the detailed relationships between the current and next hidden state may range from a linear mapping to various types of non-linearities. In the latter case, generalized RNNs with complex architectures include Gated Recurrent Units (GRUs) [34] and Long Short-Term Memory(LSTM) networks [99], etc.. However, these neural networks all have fixed structures that do not fit our simplified neural mass model of the cortex. In fact, as the structures essentially encode a time-discretized set of differential equations, the non-linearity in the hidden states can be implemented in a much more flexible way to incorporate any biologically meaningful non-linear dynamics.

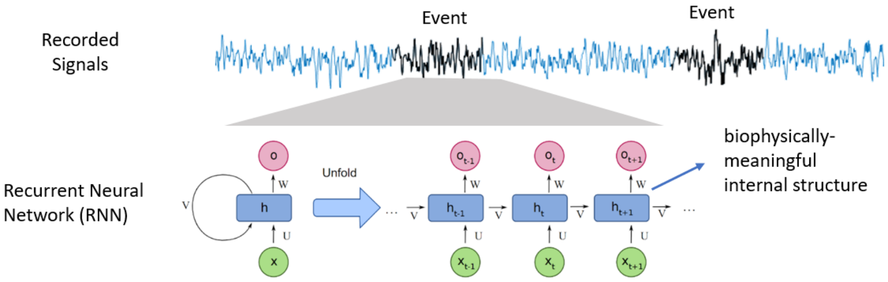

Figure 5.3: RNN designed for peri-event data. A typical RNN can be unfolded into different layers corresponding to different time points. The internal structure of an RNN can be designed to be biologically meaningful. The unfolded RNN receives a peri-event signal as a time series as the model input.

To implement the bio-informed RNN, the first step is to discretize the set of differential equations of the simplified cortex model to match the stepwise nature of the RNN. The equations presented in Section 5.2.1 can be discretized to be the following set of difference equations:

$$V_E(t) = (1 - \frac{\Delta t}{\tau_E})V_E(t-1) + \frac{\Delta t}{\tau_E}\nu_{E\to E}s_E(t-1) - \frac{\Delta t}{\tau_E}\nu_{I\to E}s_I(t-1) + \eta(t-1)$$

$$V_I(t) = (1 - \frac{\Delta t}{\tau_I})V_I(t-1) + \frac{\Delta t}{\tau_I}\nu_{E\to I}s_E(t-1) - \frac{\Delta t}{\tau_E}s_I(t-1) + \alpha\eta(t-1)$$

$$s_E(t) = \frac{\Delta t}{\tau_{E_s}}s_E(t-1) + (1 - \frac{\Delta t}{\tau_{E_s}})\lambda_E(t-1)$$

$$s_I(t) = \frac{\Delta t}{\tau_{I_s}}s_I(t-1) + 1 - \frac{\Delta t}{\tau_{I_s}}\lambda_I(t-1)$$

$$\lambda_E(t) = \text{Sigmoid}(V_E(t))$$

$$\lambda_I(t) = \text{Sigmoid}(V_I(t))$$

$$\text{LFP}_E(t) = \mu'_{E\to E}s_E(t) - \mu'_{I\to E}s_I(t) + \mu_\eta\eta(t)$$

$$\text{LFP}_E(t) = \mu'_{E\to I}s_E(t) - \mu'_{I\to I}s_I(t) + \alpha\mu_\eta(t)$$

From these equations, we know that η is independent of the past states of any other variable. Thus, consistent with the modelling idea introduced in Section 5.2.1, it can be understood as the input to the model. The variable LFP is seen as the output variable as it is only dependent on the current values of the other variables. The other six variables, i.e., the membrane potentials V_E and V_I, firing rates λ_E and λ_I, and intra-cortical synaptic currents s_E and s_I, are designed to be the hidden states. The temporal dependence of different variables in the model is visualized in Figure 5.4.

5.2.3 *Implementation and training*

As elaborated in Section 3.1, we are interested in the transient state underlying the observed events. Thus, we only train the RNN with peri-event data and assume that the network parameters, as a state-dependent property, remain unchanged. The compact form of the RNN (Figure 5.3) can be interpreted as a three-layer structure; the "depth" of the network lies in the unfolded version where each time step is one layer. In this sense, our RNN

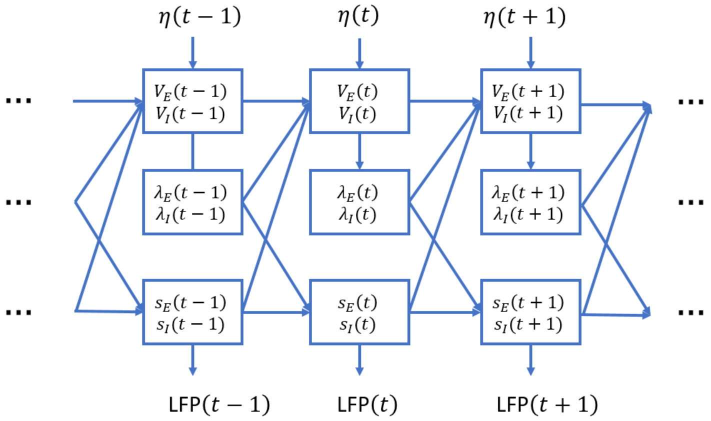

Figure 5.4: Detailed implementation of the biologically-informed RNN. Blue blocks denote the hidden states at each time point. Blue arrows represent the dependencies between the current state and the past states. The structure is based on the difference equations listed in Section 5.2.2.

has the layer numbers that equal to the time duration of the peri-event snapshots. Receiving a time-wise input, the hidden states are updated at each time point. At the same time, the network parameters are assumed to be constant over time.

For the training, we feed the RNN with peri-event thalamic LFPs as inputs. The targets for training, corresponding to the network outputs, are the peri-event cortical LFP snapshots. We optimize the network parameters by minimizing a loss function as the mean squared error (MSE) between the target and the network outputs.

The RNN is implemented with PyTorch, which enables automatic calculation of gradient descent optimization. The network is defined as an instance of the *nn* PyTorch module. Attempted optimizers include Adam[121], Adamax[199] and RMSprop[107]. Notably, we use the trick of batch training, i.e., we train the network with a huge batch consisting of all the trials for many epochs repeated. In this way, we are able to stabilize the learning curve when applying some fast-training optimizers like the RMSprop or Adam.

5.3 PRELIMINARY TRAINING RESULTS

We first test whether such a design would achieve our goal of recovering the hidden states. The validation is most effective if we can successfully recover the hidden states with data generated from systems with ground-truth known mechanisms (e.g., known structures and parameters). Thus, we simulated the linear version of dynamical systems defined in Section 5.2.1. The linearity is implemented by defining a linear relationship instead of the sigmoid functions in Eq. 5.3 and Eq. 5.7. After simulation, we obtained a bi-variate 2000-point-long snapshot as training data (i.e., corresponding to the model input η and target LFP_E).

We first train the RNN in the simplest scenario where only one parameter, the Pyr→Pyr connectivity $\nu_{E\to E}$, needs to be optimized. With a learning rate of 0.01 of the optimizer Adam, the loss converges to stabilization, with the parameter converging to its true value. The results of optimizing any one of the other parameters are consistently effective. This suggests that training the RNN in this way is able to recover the parameters and thus the hidden states.

However, when the number of unknown parameters increases to over 4, with the same training procedure, the loss still converges, but the parameters converge to values biased from their true values. The bias is dependent on the initial values of the parameters before the training. Yet, the recovered hidden states do not deviate much from the real time series, suggesting that the training procedure with RNN has the potential of estimating the hidden states with a few unknown parameters.

With this knowledge at hand, we applied the same procedure to peri-event data reflecting thalamocortical PGO events (as LFP signals). Surprisingly, although the loss function converges after around 60 trials, the parameters do not converge, indicating that the minimum is not achieved. The reason behind the failed training is still unclear and would require further investigation.

5.4 POSSIBLE PROBLEMS AND POTENTIAL SOLUTIONS

In this section, we will briefly discuss the potential reasons why the same training procedure did not work in real data.

Noise is the biggest problem in real data. For the simulated data, each input is mapped with a target deterministically. However, the noise would be huge in the real data and vary from trial to trial, making it hard to minimize it with an MSE loss. Alternatively, a loss function based on KL divergence might be a better option to characterize the noise distribution of the large batch of data we feed into the network.

Besides, as a classical neural network training problem, the initial values of parameter sets before training should also greatly influence convergence. Suppose the initial values deviate too much from the real values (given that the model is correct), with the automatic gradient descent algorithm. In that case, the parameters might change in any direction, thus becoming farther away from their true values. Unfortunately, although the parameters are designed as biologically meaningful, we have no knowledge about their true values. Thus, a feasible way is to train the RNN with random initial values repeatedly and check the consistency of training results obtained with different ranges of intial values.

A more inherent problem is the non-linearity entailed in the sigmoid function. Testing on this model shows that with the sigmoid non-linearity, the system is easily saturated. This suggests that parameter sets ensuring a non-saturated working region might be limited. Thus the standard parameter optimization depending on back-propagation might drive the parameters away from the region. As the preliminary results show that parameter optimization is effective on a linearized modification of the model, an alternative solution could be to train the RNN not with the original LFP data but with the linear Gaussian process approximation of the event snapshots, which we obtained in Chapter 3. In this way, the training data is linear in nature, allowing a more complicated linear structure to model it.

More fundamentally, training directly with an RNN might be too primitive as an application of deep learning to the problem of data-driven dynamic modelling. Instead, a new framework has recently been proposed to model differential equations with real data, a method named Neural Ordinary Differential Equations (Neural ODE) [33]. With this network, any sophisticated ODE solver can be applied to both the forward computation and backward propagation of errors. Potentially, this is promising to improve the precision compared with a simple RNN that helps minimize the loss function.

6 GENERAL CONCLUSION

In summary, in this thesis, we use both biological and statistical models to investigate what network effects are triggered by spontaneous transient events occurring during sleep. The network effects can be categorized into two aspects: the transient causal interactions one event-hosting region exerts on another, and the plastic effects one event triggers on the hosting network. This coincides with the two key questions we wanted to address in Section 1.1.2: 1) how transient events are coordinated and 2) how transient events trigger synaptic rescaling. Both aspects are critical to understanding the mechanisms underlying the memory consolidation and homeostasis functions of sleep.

To address the first question, in Chapter 3, we pointed out that the transient events are markers of emergent dynamic properties of the underlying neural system. Therefore it is critical to recover the state-dependent network dynamics by exploiting peri-event data properly. As the spontaneity of the event occurrence challenges the collection of peri-event data, frequently leading to selection bias of the underlying states, we established a theoretical framework that accounts for the origin of such selection bias during thresholding detection. We formulated the bias problem for peri-event snapshots in the frameworks of dynamical systems and SCMs. More importantly, we proposed a *DeSnap* algorithm to correct the selection bias in peri-event snapshots based on VAR models with Gaussian assumptions. The performance of correction is validated with simulated signals with a single stationary state and with the Markovian alternation between two states, suggesting that our bias correction algorithm is able to recover the system dynamics when the event-hosting states are unobserved and transiently switching across time. The *DeSnap*-recovered hippocampal states underlying SPW-R and theta events match well with experimental findings, also suggesting the effectiveness of bias correction.

Further and along the same line, in Chapter 4, we investigated how to characterize the state-dependent transient causal interactions with peri-event snapshots. Understanding such causal interplay requires knowledge of the transient system dynamics, which we have already established in Chapter 3. By modelling these system dynamics with time-varying VAR models, we formulated the time-varying extension of TE and CS (i.e., DCS) in the framework of SCM. We also proposed a novel measure, the rDCS, that is based on interventional causality principles and is able to deal with the situations where TE and DCS fail. By applying these causality measures to a series of simulated and real data, we demonstrated that without *DeSnap* correction, calculating causality measures on peri-event snapshots aligned by the peak of one event-hosting region significantly bias the detection of peri-event causal interactions. This problem is critical as peri-event data exploited in the community are often triggered by a single hosting region. We propose that network properties based on event-triggered dynamics should be estimated after correcting for the selection bias.

To answer the second question, in Chapter 2 we built a mechanistic model of PGO waves that covers three key regions generating three important transient events. In particular, we reproduced at the mesoscopic level the transient electrophysiological properties of two subtypes of PGO waves during different sleep stages, as well as SOs and spindles during NREM sleep. We extrapolated on the event-triggered plasticity effects with these replicated transient characteristics in cortical population activities. As results, we found that pre-REM and REM PGO waves trigger opposite plastic changes in cortical circuits, which can be an interesting mechanism that coordinates synaptic rescaling through the switches between NREM and REM sleep. Our results on spindle-triggered LTP also support the differentiated functional roles between isolated spindles and SO-spindle coupling. These theoretical predictions on event-triggered plastic effects can be validated with experimental results using a data-driven biophysical model proposed in Chapter 5. This might be a promising framework to recover hidden variables from experimental data for mechanistic investigations.

Apart from plasticity effects, the PGO model, based on its biologically plausible model assumptions, is able to validate the mechanistic hypothesis proposed by experimentalists (i.e., in this sense, addressing the first question). The successful reproduction of two subtypes of PGO waves supports the inter-regional cellular mechanism under PGO wave generation that is previously under debate. The simulated pre-REM PGO waves trigger a DOWN state followed by a DOWN→UP transition, which has been proposed to implement the rescaling of cortical neurons' firing patterns while integrating new memory traces when co-occurring with SPW-Rs. This might match the experimental findings for PGO-ripple coupling and suggest a coordinated mechanism between SOs, SPW-Rs, and PGO waves.

Together, this thesis improves the knowledge of event-triggered network effects by making predictions with mechanistic models and proposing generic frameworks for event-based data analysis. Further experiments can be carried out to validate the theoretical predictions while the data analysis approaches we developed pave the way for achieving more fruitful outcomes based on experimental recordings.

BIBLIOGRAPHY

[1] Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723.

[2] Axmacher, N., Elger, C. E., and Fell, J. (2008). Ripples in the medial temporal lobe are relevant for human memory consolidation. *Brain: A Journal of Neurology*, 131(Pt 7):1806–1817.

[3] Ay, N. and Krakauer, D. C. (2007). Geometric robustness theory and biological networks. *Theory in Biosciences = Theorie in Den Biowissenschaften*, 125(2):93–121.

[4] Aydore, S., Pantazis, D., and Leahy, R. M. (2013). A note on the phase locking value and its properties. *Neuroimage*, 74:231–244.

[5] Bareinboim, E., Tian, J., and Pearl, J. (2014). Recovering from selection bias in causal and statistical inference. In *AAAI*, pages 2410–2416.

[6] Barkai, E. and Hasselmo, M. E. (1994). Modulation of the input/output function of rat piriform cortex pyramidal cells. *Journal of Neurophysiology*, 72(2):644–658.

[7] Barnes, C. A., Jung, M. W., McNaughton, B. L., Korol, D. L., Andreasson, K., and Worley, P. F. (1994). LTP saturation and spatial learning disruption: effects of task variables and saturation levels. *The Journal of Neuroscience*, 14(10):5793–5806.

[8] Barnett, L., Barrett, A. B., and Seth, A. K. (2009). Granger causality and transfer entropy are equivalent for gaussian variables. *Physical review letters*, 103(23):238701.

[9] Barnett, L. and Seth, A. K. (2015). Granger causality for state-space models. *Physical Review. E, Statistical, Nonlinear, and Soft Matter Physics*, 91(4):040101.

[10] Bartram, J., Kahn, M. C., Tuohy, S., Paulsen, O., Wilson, T., and Mann, E. O. (2017). Cortical up states induce the selective weakening of subthreshold synaptic inputs. *Nature Communications*, 8(1):665.

[11] Beierlein, M. (2014). Synaptic mechanisms underlying cholinergic control of thalamic reticular nucleus neurons. *The Journal of Physiology*, 592(19):4137–4145.

[12] Benington, J. H. and Heller, H. C. (1995). Restoration of brain energy metabolism as the function of sleep. *Progress in Neurobiology*, 45(4):347–360.

[13] Bi, G. Q. and Poo, M. M. (1998). Synaptic modifications in cultured hippocampal neurons: dependence on spike timing, synaptic strength, and postsynaptic cell type. *The Journal of Neuroscience*, 18(24):10464–10472.

[14] Biørn, E. (2016). *Econometrics of panel data: Methods and applications*. Oxford University Press.

[15] Bishop, C. M. (2006). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg.

[16] Bista, P., Cerina, M., Ehling, P., Leist, M., Pape, H.-C., Meuth, S. G., and Budde, T. (2015). The role of two-pore-domain background k^+ (k_2p) channels in the thalamus. *Pflugers Archiv: European Journal of Physiology*, 467(5):895–905.

[17] Bista, P., Meuth, S. G., Kanyshkova, T., Cerina, M., Pawlowski, M., Ehling, P., Landgraf, P., Borsotto, M., Heurteaux, C., Pape, H.-C., Baukrowitz, T., and Budde, T. (2012). Identification of the muscarinic pathway underlying cessation of sleep-related burst activity in rat thalamocortical relay neurons. *Pflugers Archiv: European Journal of Physiology*, 463(1):89–102.

[18] Bizzi, E. (1966). Discharge patterns of single geniculate neurons during the rapid eye movements of sleep. *Journal of Neurophysiology*, 29(6):1087–1095.

[19] Bliss, T. V. and Lomo, T. (1973). Long-lasting potentiation of synaptic transmission in the dentate area of the anaesthetized rabbit following stimulation of the perforant path. *The Journal of Physiology*, 232(2):331–356.

[20] Bowker, R. M. (1985). Variability in the characteristics of pontogeniculooccipital spikes during paradoxical sleep. *Experimental Neurology*, 87(2):212–224.

[21] Boyce, R., Glasgow, S. D., Williams, S., and Adamantidis, A. (2016). Causal evidence for the role of REM sleep theta rhythm in contextual memory consolidation. *Science*, 352(6287):812–816.

[22] Boyce, R., Williams, S., and Adamantidis, A. (2017). REM sleep and memory. *Current Opinion in Neurobiology*, 44:167–177.

[23] Brodersen, K. H., Gallusser, F., Koehler, J., Remy, N., Scott, S. L., et al. (2015). Inferring causal impact using bayesian structural time-series models. *The Annals of Applied Statistics*, 9(1):247–274.

[24] Brooks, D. C. (1968). Waves associated with eye movement in the awake and sleeping cat. *Electroencephalography and Clinical Neurophysiology*, 24(6):532–541.

[25] Buzsaki, G. (1984). Long-term changes of hippocampal sharp-waves following high frequency afferent activation. *Brain Research*, 300(1):179–182.

[26] Buzsaki, G. (1998). Memory consolidation during sleep: a neurophysiological perspective. *Journal of sleep research*, 7(S1):17–23.

[27] Buzsaki, G. (2015). Hippocampal sharp wave-ripple: A cognitive biomarker for episodic memory and planning. *Hippocampus*, 25(10):1073–1188.

[28] Buzsaki, G., Haas, H. L., and Anderson, E. G. (1987). Long-term potentiation induced by physiologically relevant stimulus patterns. *Brain Research*, 435(1-2):331–333.

[29] Buzsaki, G., Horvath, Z., Urioste, R., Hetke, J., and Wise, K. (1992). High-frequency network oscillation in the hippocampus. *Science*, 256(5059):1025–7.

[30] Callaway, C. W., Lydic, R., Baghdoyan, H. A., and Hobson, J. A. (1987). Pontogeniculooccipital waves: spontaneous visual system activity during rapid eye movement sleep. *Cellular and Molecular Neurobiology*, 7(2):105–149.

[31] Cekic, S., Grandjean, D., and Renaud, O. (2018). Multiscale bayesian state-space model for granger causality analysis of brain signal. *Journal of applied statistics*, 46(1):1–19.

[32] Chauvette, S., Seigneur, J., and Timofeev, I. (2012). Sleep oscillations in the thalamocortical system induce long-term neuronal plasticity. *Neuron*, 75(6):1105–1113.

[33] Chen, R. T., Rubanova, Y., Bettencourt, J., and Duvenaud, D. (2018). Neural ordinary differential equations. *arXiv preprint arXiv:1806.07366*.

[34] Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

[35] Clawson, B. C., Durkin, J., and Aton, S. J. (2016). Form and function of sleep spindles across the lifespan. *Neural plasticity*, 2016:6936381.

[36] Clemens, Z., Molle, M., Eross, L., Jakus, R., Rasonyi, G., Halasz, P., and Born, J. (2011). Fine-tuned coupling between human parahippocampal ripples and sleep spindles. *The European Journal of Neuroscience*, 33(3):511–520.

[37] Cohen, B. and Feldman, M. (1968). Relationship of electrical activity in pontine reticular formation and lateral geniculate body to rapid eye movements. *Journal of Neurophysiology*, 31(6):806–817.

[38] Costa, M. S., Born, J., Claussen, J. C., and Martinetz, T. (2016a). Modeling the effect of sleep regulation on a neural mass model. *Journal of Computational Neuroscience*, 41(1):15–28.

[39] Costa, M. S., Weigenand, A., Ngo, H.-V. V., Marshall, L., Born, J., Martinetz, T., and Claussen, J. C. (2016b). A thalamocortical neural mass model of the eeg during nrem sleep and its response to auditory stimulation. *PLoS computational biology*, 12(9):e1005022.

[40] Cox, R., Hofman, W. F., and Talamini, L. M. (2012). Involvement of spindles in memory consolidation is slow wave sleep-specific. *Learning & Memory*, 19(7):264–267.

[41] Csicsvari, J., Hirase, H., Mamiya, A., and Buzsaki, G. (2000). Ensemble patterns of hippocampal CA3-CA1 neurons during sharp wave-associated population events. *Neuron*, 28(2):585–594.

[42] Datta, S. (1995). Neuronal activity in the peribrachial area: relationship to behavioral state control. *Neuroscience and Biobehavioral Reviews*, 19(1):67–84.

[43] Datta, S. (1997). Cellular basis of pontine ponto-geniculo-occipital wave generation and modulation. *Cellular and Molecular Neurobiology*, 17(3):341–365.

[44] Datta, S. (2000). Avoidance task training potentiates phasic pontine-wave density in the rat: A mechanism for sleep-dependent plasticity. *The Journal of Neuroscience*, 20(22):8607–8613.

[45] Datta, S. (2006). Activation of phasic pontine wave generator a mechanism for sleep dependent memory processing. *Sleep and Biological Rhythms*.

[46] Datta, S. (2012). Phasic pontine-wave (p-wave) generation. In *Sleep and brain activity*, pages 147–164. Elsevier.

[47] Datta, S., Calvo, J., Quattrochi, J., and Hobson, J. (1992). Cholinergic microstimulation of the peribrachial nucleus in the cat. i. immediate and prolonged increases in ponto-geniculo-occipital waves. *Archives italiennes de biologie*, 130(4):263.

[48] Datta, S. and Hobson, J. A. (1994). Neuronal activity in the caudolateral peribrachial pons: relationship to PGO waves and rapid eye movements. *Journal of Neurophysiology*, 71(1):95–109.

[49] Datta, S. and Hobson, J. A. (1995). Suppression of ponto-geniculo-occipital waves by neurotoxic lesions of pontine caudo-lateral peribrachial cells. *Neuroscience*, 67(3):703–712.

[50] Datta, S., Li, G., and Auerbach, S. (2008). Activation of phasic pontine-wave generator in the rat: a mechanism for expression of plasticity-related genes and proteins in the dorsal hippocampus and amygdala. *The European Journal of Neuroscience*, 27(7):1876–1892.

[51] Datta, S., Mavanji, V., Patterson, E. H., and Ulloor, J. (2003). Regulation of rapid eye movement sleep in the freely moving rat: local microinjection of serotonin, norepinephrine, and adenosine into the brainstem. *Sleep*, 26(5):513–520.

[52] Datta, S., Mavanji, V., Ulloor, J., and Patterson, E. H. (2004). Activation of phasic pontine wave generator prevents rapid eye movement sleep deprivation induced learning impairment in the rat: a mechanism for sleep-dependent plasticity. *The Journal of Neuroscience*, 24(6):1416–1427.

[53] Datta, S. and O'Malley, M. W. (2013). Fear extinction memory consolidation requires potentiation of pontine-wave activity during REM sleep. *The Journal of Neuroscience*, 33(10):4561–4569.

[54] Datta, S., Siwek, D. F., and Huang, M. P. (2009). Improvement of two-way active avoidance memory requires protein kinase a activation and brain-derived neurotrophic factor expression in the dorsal hippocampus. *Journal of Molecular Neuroscience*, 38(3):257–264.

[55] Datta, S., Siwek, D. F., Patterson, E. H., and Cipolloni, P. B. (1998). Localization of pontine PGO wave generation sites and their anatomical projections in the rat. *Synapse*, 30(4):409–423.

[56] Deboer, T. (2018). Sleep homeostasis and the circadian clock: Do the circadian pacemaker and the sleep homeostat influence each other's functioning? *Neurobiology of sleep and circadian rhythms*, 5:68–77.

[57] Deboer, T., Ross, R. J., Morrison, A. R., and Sanford, L. D. (1999). Electrical stimulation of the amygdala increases the amplitude of elicited ponto-geniculo-occipital waves. *Physiology & Behavior*, 66(1):119–124.

[58] Deboer, T., Sanford, L. D., Ross, R. J., and Morrison, A. R. (1998). Effects of electrical stimulation in the amygdala on ponto-geniculo-occipital waves in rats. *Brain Research*, 793(1-2):305–310.

[59] Deschenes, M. and Hu, B. (1990). Membrane resistance increase induced in thalamic neurons by stimulation of brainstem cholinergic afferents. *Brain Research*, 513(2):339–342.

[60] Destexhe, A., Bal, T., McCormick, D. A., and Sejnowski, T. J. (1996). Ionic mechanisms underlying synchronized oscillations and propagating waves in a model of ferret thalamic slices. *Journal of Neurophysiology*, 76(3):2049–2070.

[61] Destexhe, A., Contreras, D., Sejnowski, T. J., and Steriade, M. (1994). A model of spindle rhythmicity in the isolated thalamic reticular nucleus. *Journal of Neurophysiology*, 72(2):803–818.

[62] Destexhe, A., Neubig, M., Ulrich, D., and Huguenard, J. (1998). Dendritic low-threshold calcium currents in thalamic relay cells. *The Journal of Neuroscience*, 18(10):3574–3588.

[63] Diekelmann, S. and Born, J. (2010). The memory function of sleep. *Nature Reviews. Neuroscience*, 11(2):114–126.

[64] Diks, C. and Wolski, M. (2016). Nonlinear granger causality: guidelines for multivariate analysis. *Journal of Applied Econometrics*, 31(7):1333–1351.

[65] Ding, M., Bressler, S. L., Yang, W., and Liang, H. (2000). Short-window spectral analysis of cortical event-related potentials by adaptive multivariate autoregressive modeling: data preprocessing, model validation, and variability assessment. *Biological Cybernetics*, 83(1):35–45.

[66] Disney, A. A., Aoki, C., and Hawken, M. J. (2007). Gain modulation by nicotine in macaque v1. *Neuron*, 56(4):701–713.

[67] Ego-Stengel, V. and Wilson, M. A. (2010). Disruption of ripple-associated hippocampal activity during rest impairs spatial learning in the rat. *Hippocampus*, 20(1):1–10.

[68] Eichler, M. and Didelez, V. (2010). On granger causality and the effect of interventions in time series. *Lifetime data analysis*, 16(1):3–32.

[69] Eom, K. (1999). Time-varying autoregressive modeling of HRR radar signatures. *IEEE transactions on aerospace and electronic systems*, 35(3):974–988.

[70] Eschenko, O., Moelle, M., Born, J., and Sara, S. J. (2006). Elevated sleep spindle density after learning or after retrieval in rats. *The Journal of Neuroscience*, 26(50):12914–12920.

[71] Eschenko, O., Ramadan, W., Moelle, M., Born, J., and Sara, S. J. (2008). Sustained increase in hippocampal sharp-wave ripple activity during slow-wave sleep after learning. *Learning & Memory*, 15(4):222–228.

[72] Fernandez-Mendoza, J., Lozano, B., Seijo, F., Santamarta-Liebana, E., Ramos-Platon, M. J., Vela-Bueno, A., and Fernandez-Gonzalez, F. (2009). Evidence of subthalamic PGO-like waves during REM sleep in humans: a deep brain polysomnographic study. *Sleep*, 32(9):1117–1126.

[73] Fogel, S. M. and Smith, C. T. (2006). Learning-dependent changes in sleep spindles and stage 2 sleep. *Journal of Sleep Research*, 15(3):250–255.

[74] Fries, P. (2015). Rhythms for cognition: communication through coherence. *Neuron*, 88(1):220–235.

[75] Friston, K. J., Harrison, L., and Penny, W. (2003). Dynamic causal modelling. *Neuroimage*, 19(4):1273–1302.

[76] Fung, P. K. and Robinson, P. A. (2013). Neural field theory of calcium dependent plasticity with applications to transcranial magnetic stimulation. *Journal of Theoretical Biology*, 324:72–83.

[77] Gais, S. and Born, J. (2004). Low acetylcholine during slow-wave sleep is critical for declarative memory consolidation. *Proceedings of the National Academy of Sciences of the United States of America*, 101(7):2140–2144.

[78] Gais, S., Moelle, M., Helms, K., and Born, J. (2002). Learning-dependent increases in sleep spindle density. *The Journal of Neuroscience*, 22(15):6830–6834.

[79] Garside, P., Arizpe, J., Lau, C.-I., Goh, C., and Walsh, V. (2015). Cross-hemispheric alternating current stimulation during a nap disrupts slow wave activity and associated memory consolidation. *Brain Stimulation*, 8(3):520–527.

[80] Geweke, J. F. (1984). Measures of conditional linear dependence and feedback between time series. *Journal of the American Statistical Association*, 79(388):907–915.

[81] Girardeau, G., Benchenane, K., Wiener, S. I., Buzsaki, G., and Zugaro, M. B. (2009). Selective suppression of hippocampal ripples impairs spatial memory. *Nature Neuroscience*, 12(10):1222–1223.

[82] Girardeau, G., Cei, A., and Zugaro, M. (2014). Learning-induced plasticity regulates hippocampal sharp wave-ripple drive. *The Journal of Neuroscience*, 34(15):5176–5183.

[83] Girardeau, G., Inema, I., and Buzsaki, G. (2017). Reactivations of emotional memory in the hippocampus-amygdala system during sleep. *Nature Neuroscience*, 20(11):1634–1642.

[84] Goncalves, P. J., Lueckmann, J.-M., Deistler, M., Nonnenmacher, M., Öcal, K., Bassetto, G., Chintaluri, C., Podlaski, W. F., Haddad, S. A., Vogels, T. P., et al. (2020). Training deep neural density estimators to identify mechanistic models of neural dynamics. *Elife*, 9:e56261.

[85] Gontier, C. and Pfister, J.-P. (2020). Identifiability of a binomial synapse. *Frontiers in Computational Neuroscience*, 14:558477.

[86] Gonzalo, J. and Pitarakis, J.-Y. (2002). Lag length estimation in large dimensional systems. *Journal of Time Series Analysis*, 23(4):401–423.

[87] GonzÃ¡lez-Rueda, A., Pedrosa, V., Feord, R. C., Clopath, C., and Paulsen, O. (2018). Activity-dependent downscaling of subthreshold synaptic inputs during slow-wave-sleep-like activity inÂ vivo. *Neuron*, 97(6):1244–1252.e5.

[88] Gott, J. A., Liley, D. T. J., and Hobson, J. A. (2017). Towards a functional understanding of PGO waves. *Frontiers in Human Neuroscience*, 11:89.

[89] Granger, C. W. J. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica : journal of the Econometric Society*, 37(3):424.

[90] Gridchyn, I., Schoenenberger, P., O'Neill, J., and Csicsvari, J. (2020). Assembly-specific disruption of hippocampal replay leads to selective memory deficit. *Neuron*.

[91] Grosmark, A. D., Mizuseki, K., Pastalkova, E., Diba, K., and Buzsaki, G. (2012). REM sleep reorganizes hippocampal excitability. *Neuron*, 75(6):1001–1007.

[92] Hamilton, J. D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica: Journal of the econometric society*, pages 357–384.

[93] Hausman, C. and Rapson, D. S. (2018). Regression discontinuity in time: Considerations for empirical applications. *Annual Review of Resource Economics*, 10(1):533–552.

[94] Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica: Journal of the econometric society*, pages 153–161.

[95] Hernán, M. A., Hernández-Díaz, S., and Robins, J. M. (2004). A structural approach to selection bias. *Epidemiology*, 15(5):615–625.

[96] Hesse, W., Moeller, E., Arnold, M., and Schack, B. (2003). The use of time-variant EEG granger causality for inspecting directed interdependencies of neural assemblies. *Journal of Neuroscience Methods*, 124(1):27–44.

[97] Hobson, J. A. (1965). The effects of chronic brain-stem lesions on cortical and muscular activity during sleep and waking in the cat. *Electroencephalography and Clinical Neurophysiology*, 19:41–62.

[98] Hobson, J. A. (2009). REM sleep and dreaming: towards a theory of protoconsciousness. *Nature Reviews. Neuroscience*, 10(11):803–813.

[99] Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

[100] Horwitz, R. I. and Feinstein, A. R. (1978). Alternative analytic methods for case-control studies of estrogens and endometrial cancer. *New England Journal of Medicine*, 299(20):1089–1094.

[101] Hsiao, C. (2014). *Analysis of panel data*. Number 54. Cambridge university press.

[102] Hu, B., Bouhassira, D., Steriade, M., and Deschenes, M. (1988). The blockage of ponto-geniculo-occipital waves in the cat lateral geniculate nucleus by nicotinic antagonists. *Brain Research*, 473(2):394–397.

[103] Hu, B., Steriade, M., and Deschenes, M. (1989a). The cellular mechanism of thalamic ponto-geniculo-occipital waves. *Neuroscience*, 31(1):25–35.

[104] Hu, B., Steriade, M., and Deschenes, M. (1989b). The effects of brainstem peribrachial stimulation on neurons of the lateral geniculate nucleus. *Neuroscience*, 31(1):13–24.

[105] Hu, B., Steriade, M., and Deschenes, M. (1989c). The effects of brainstem peribrachial stimulation on perigeniculate neurons: the blockage of spindle waves. *Neuroscience*, 31(1):1–12.

[106] Hume, D. (1748). *An enquiry concerning human understanding*. Oxford ; New York : Oxford University Press.

[107] Igel, C. and Hüsken, M. (2000). Improving the rprop learning algorithm. In *Proceedings of the second international ICSC symposium on neural computation (NC 2000)*, volume 2000, pages 115–121. Citeseer.

[108] Inglis, W. L. and Semba, K. (1996). Colocalization of ionotropic glutamate receptor subunits with NADPH-diaphorase-containing neurons in the rat mesopontine tegmentum. *The Journal of Comparative Neurology*, 368(1):17–32.

[109] Jadhav, S. P., Kemere, C., German, P. W., and Frank, L. M. (2012). Awake hippocampal sharp-wave ripples support spatial memory. *Science*, 336(6087):1454–1458.

[110] Jansen, B. H., Zouridakis, G., and Brandt, M. E. (1993). A neurophysiologically-based mathematical model of flash visual evoked potentials. *Biological Cybernetics*, 68(3):275–283.

[111] Janzing, D., Balduzzi, D., Grosse-Wentrup, M., and SchÃ¶lkopf, B. (2013). Quantifying causal influences. *The Annals of Statistics*, 41(5):2324–2358.

[112] Ji, D. and Wilson, M. A. (2007). Coordinated memory replay in the visual cortex and hippocampus during sleep. *Nature Neuroscience*, 10(1):100–107.

[113] Jouvet, M. (1959). Correlations electromyographiques du sommeil chez le chat decortique et mesencephalique chronique. *C. R. Soc. Biol.*

[114] Kamondi, A., Williams, J. A., Hutcheon, B., and Reiner, P. B. (1992). Membrane properties of mesopontine cholinergic neurons studied with the whole-cell patch-clamp technique: implications for behavioral state control. *Journal of Neurophysiology*, 68(4):1359–1372.

[115] Kang, Y. and Kitai, S. T. (1990). Electrophysiological properties of pedunculopontine neurons and their postsynaptic responses following stimulation of substantia nigra reticulata. *Brain Research*, 535(1):79–95.

[116] Karashima, A., Katayama, N., and Nakao, M. (2010). Enhancement of synchronization between hippocampal and amygdala theta waves associated with pontine wave density. *Journal of Neurophysiology*, 103(5):2318–2325.

[117] Karashima, A., Nakamura, K., Horiuchi, M., Nakao, M., Katayama, N., and Yamamoto, M. (2002). Elicited ponto-geniculo-occipital waves by auditory stimuli are synchronized with hippocampal theta-waves. *Psychiatry and Clinical Neurosciences*, 56(3):343–344.

[118] Karashima, A., Nakamura, K., Watanabe, M., Sato, N., Nakao, M., Katayama, N., and Yamamoto, M. (2001). Synchronization between hippocampal theta waves and PGO waves during REM sleep. *Psychiatry and Clinical Neurosciences*, 55(3):189–190.

[119] Karashima, A., Nakao, M., Honda, K., Iwasaki, N., Katayama, N., and Yamamoto, M. (2004). Theta wave amplitude and frequency are differentially correlated with pontine waves and rapid eye movements during REM sleep in rats. *Neuroscience Research*, 50(3):283–289.

[120] Kaufman, L. S. and Morrison, A. R. (1981). Spontaneous and elicited PGO spikes in rats. *Brain Research*, 214(1):61–72.

[121] Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

[122] Klinzing, J. G., Niethard, N., and Born, J. (2019). Mechanisms of systems memory consolidation during sleep. *Nature Neuroscience*.

[123] Koehler, A. B. and Murphree, E. S. (1988). A comparison of the akaike and schwarz criteria for selecting model order. *Applied statistics*, 37(2):187.

[124] Krishnan, G. P., Chauvette, S., Shamie, I., Soltani, S., Timofeev, I., Cash, S. S., Halgren, E., and Bazhenov, M. (2016). Cellular and neurochemical basis of sleep stages in the thalamocortical network. *eLife*, 5.

[125] Kutschireiter, A., Surace, S. C., and Pfister, J.-P. (2020). The hitchhiker's guide to nonlinear filtering. *Journal of mathematical psychology*, 94:102307.

[126] Ladenbauer, J., Ladenbauer, J., Kulzow, N., de Boor, R., Avramova, E., Grittner, U., and Floel, A. (2017). Promoting sleep oscillations and their functional coupling by transcranial stimulation enhances memory consolidation in mild cognitive impairment. *The Journal of Neuroscience*, 37(30):7111–7124.

[127] Lasota, A. and Mackey, M. C. (2013). *Chaos, fractals, and noise: stochastic aspects of dynamics*, volume 97. Springer Science & Business Media.

[128] Latchoumane, C.-F. V., Ngo, H.-V. V., Born, J., and Shin, H.-S. (2017). Thalamic spindles promote memory formation during sleep through triple phase-locking of cortical, thalamic, and hippocampal rhythms. *Neuron*, 95(2):424–435.e6.

[129] Lee, A. K. and Wilson, M. A. (2002). Memory of sequential experience in the hippocampus during slow wave sleep. *Neuron*, 36(6):1183–1194.

[130] Lee, K. H. and McCormick, D. A. (1995). Acetylcholine excites GABAergic neurons of the ferret perigeniculate nucleus through nicotinic receptors. *Journal of Neurophysiology*, 73(5):2123–2128.

[131] Leminen, M. M., Virkkala, J., Saure, E., Paajanen, T., Zee, P. C., Santostasi, G., Hublin, C., Mueller, K., Porkka-Heiskanen, T., Huotilainen, M., and Paunio, T. (2017). Enhanced memory consolidation via automatic sound stimulation during non-REM sleep. *Sleep*, 40(3).

[132] Leonard, C. S. and Llinas, R. (1994). Serotonergic and cholinergic inhibition of mesopontine cholinergic neurons controlling REM sleep: an in vitro electrophysiological study. *Neuroscience*, 59(2):309–330.

[133] Leonard, S. (1990). Electrophysiology of mammalian pedunculopontine and laterodorsal tegmental neurons in vitro: Implications for the control of REM sleep. *Brain Cholinergic System*.

[134] Lerma, J. and GarcÃa-Austt, E. (1985). Hippocampal theta rhythm during paradoxical sleep. effects of afferent stimuli and phase relationships with phasic events. *Electroencephalography and Clinical Neurophysiology*, 60(1):46–54.

[135] Levenstein, D., Buzsaki, G., and Rinzel, J. (2019). NREM sleep in the rodent neocortex and hippocampus reflects excitable dynamics. *Nature Communications*, 10(1):2478.

[136] Levenstein, D., Watson, B. O., Rinzel, J., and Buzsaki, G. (2017). Sleep regulation of the distribution of cortical firing rates. *Current Opinion in Neurobiology*, 44:34–42.

[137] Lewis, D. (1973). *Counterfactuals*. Oxford Blackwell.

[138] Li, W., Ma, L., Yang, G., and Gan, W.-B. (2017). REM sleep selectively prunes and maintains new synapses in development and learning. *Nature Neuroscience*, 20(3):427–437.

[139] Lim, A. S., Lozano, A. M., Moro, E., Hamani, C., Hutchison, W. D., Dostrovsky, J. O., Lang, A. E., Wennberg, R. A., and Murray, B. J. (2007). Characterization of REM-sleep associated ponto-geniculo-occipital waves in the human pons. *Sleep*, 30(7):823–827.

[140] Logothetis, N. K., Eschenko, O., Murayama, Y., Augath, M., Steudel, T., Evrard, H. C., Besserve, M., and Oeltermann, A. (2012). Hippocampal-cortical interaction during periods of subcortical silence. *Nature*, 491(7425):547–553.

[141] Lopes da Silva, F. H., Hoeks, A., Smits, H., and Zetterberg, L. H. (1974). Model of brain rhythmic activity. *Kybernetik*, 15(1):27–37.

[142] Louie, K. and Wilson, M. A. (2001). Temporally structured replay of awake hippocampal ensemble activity during rapid eye movement sleep. *Neuron*, 29(1):145–156.

[143] Luebke, J. I., Greene, R. W., Semba, K., Kamondi, A., McCarley, R. W., and Reiner, P. B. (1992). Serotonin hyperpolarizes cholinergic low-threshold burst neurons in the rat laterodorsal tegmental nucleus in vitro. *Proceedings of the National Academy of Sciences of the United States of America*, 89(2):743–747.

[144] Lundqvist, M., Herman, P., Warden, M. R., Brincat, S. L., and Miller, E. K. (2018). Gamma and beta bursts during working memory readout suggest roles in its volitional control. *Nature communications*, 9(1):1–12.

[145] Luo, X. Trace: Tennessee research and creative exchange.

[146] Maaten, L. v. d. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605.

[147] Maingret, N., Girardeau, G., Todorova, R., Goutierre, M., and Zugaro, M. (2016). Hippocampo-cortical coupling mediates memory consolidation during sleep. *Nature Neuroscience*, 19(7):959–964.

[148] Marinazzo, D., Liao, W., Chen, H., and Stramaglia, S. (2011). Nonlinear connectivity by granger causality. *Neuroimage*, 58(2):330–338.

[149] Marinazzo, D., Pellicoro, M., and Stramaglia, S. (2008). Kernel method for nonlinear granger causality. *Physical review letters*, 100(14):144103.

[150] Marshall, L., Helgadottir, H., Moelle, M., and Born, J. (2006). Boosting slow oscillations during sleep potentiates memory. *Nature*, 444(7119):610–613.

[151] Marshall, L., Moelle, M., Hallschmid, M., and Born, J. (2004). Transcranial direct current stimulation during sleep improves declarative memory. *The Journal of Neuroscience*, 24(44):9985–9992.

[152] Massimini, M., Huber, R., Ferrarelli, F., Hill, S., and Tononi, G. (2004). The sleep slow oscillation as a traveling wave. *The Journal of Neuroscience*, 24(31):6862–6870.

[153] Mavanji, V. and Datta, S. (2003). Activation of the phasic pontine-wave generator enhances improvement of learning performance: a mechanism for sleep-dependent plasticity. *The European Journal of Neuroscience*, 17(2):359–370.

[154] Mazzoni, A., Linden, H., Cuntz, H., Lansner, A., Panzeri, S., and Einevoll, G. T. (2015). Computing the local field potential (LFP) from integrate-and-fire network models. *PLoS Computational Biology*, 11(12):e1004584.

[155] McCarley, R. W., Nelson, J. P., and Hobson, J. A. (1978). Ponto-geniculo-occipital (PGO) burst neurons: correlative evidence for neuronal generators of PGO waves. *Science*, 201(4352):269–272.

[156] McCormick, D. A. (1992). Cellular mechanisms underlying cholinergic and noradrenergic modulation of neuronal firing mode in the cat and guinea pig dorsal lateral geniculate nucleus. *The Journal of Neuroscience*, 12(1):278–289.

[157] McCormick, D. A. and Prince, D. A. (1987a). Acetylcholine causes rapid nicotinic excitation in the medial habenular nucleus of guinea pig, in vitro. *The Journal of Neuroscience*, 7(3):742–752.

[158] McCormick, D. A. and Prince, D. A. (1987b). Actions of acetylcholine in the guinea-pig and cat medial and lateral geniculate nuclei, in vitro. *The Journal of Physiology*, 392:147–165.

[159] McDevitt, E. A., Duggan, K. A., and Mednick, S. C. (2015). REM sleep rescues learning from interference. *Neurobiology of Learning and Memory*, 122:51–62.

[160] Mednick, S. C., McDevitt, E. A., Walsh, J. K., Wamsley, E., Paulus, M., Kanady, J. C., and Drummond, S. P. A. (2013). The critical role of sleep spindles in hippocampal-dependent memory: a pharmacology study. *The Journal of Neuroscience*, 33(10):4494–4504.

[161] Moelle, M., Eschenko, O., Gais, S., Sara, S. J., and Born, J. (2009). The influence of learning on sleep slow oscillations and associated spindles and ripples in humans and rats. *The European Journal of Neuroscience*, 29(5):1071–1081.

[162] Moelle, M., Yeshenko, O., Marshall, L., Sara, S. J., and Born, J. (2006). Hippocampal sharp wave-ripples linked to slow oscillations in rat slow-wave sleep. *Journal of Neurophysiology*, 96(1):62–70.

[163] Moeller, E., Schack, B., Arnold, M., and Witte, H. (2001). Instantaneous multivariate EEG coherence analysis by means of adaptive high-dimensional autoregressive models. *Journal of Neuroscience Methods*, 105(2):143–158.

[164] Morin, A., Doyon, J., Dostie, V., Barakat, M., Hadj Tahar, A., Korman, M., Benali, H., Karni, A., Ungerleider, L. G., and Carrier, J. (2008). Motor sequence learning increases sleep spindles and fast frequencies in post-training sleep. *Sleep*, 31(8):1149–1156.

[165] Moroni, F., Nobili, L., Iaria, G., Sartori, I., Marzano, C., Tempesta, D., Proserpio, P., Lo Russo, G., Gozzo, F., Cipolli, C., De Gennaro, L., and Ferrara, M. (2014). Hippocampal slow EEG frequencies during NREM sleep are involved in spatial memory consolidation in humans. *Hippocampus*, 24(10):1157–1168.

[166] Morrison, A. R. and Pompeiano, O. (1966). Vestibular influences during sleep. II. effects of vestibular lesions on the pyramidal discharge during desynchronized sleep. *Archives Italiennes de Biologie*, 104(2):214–230.

[167] Muller, L. and Destexhe, A. (2012). Propagating waves in thalamus, cortex and the thalamocortical system: experiments and models. *Journal of Physiology-Paris*, 106(5-6):222–238.

[168] Murphy, K. P. (2012). *Machine learning: a probabilistic perspective*. MIT press.

[169] Mushiake, H., Kodama, T., Shima, K., Yamamoto, M., and Nakahama, H. (1988). Fluctuations in spontaneous discharge of hippocampal theta cells during sleep-waking states and PCPA-induced insomnia. *Journal of Neurophysiology*, 60(3):925–939.

[170] Nelson, J. P., McCarley, R. W., and Hobson, J. A. (1983). REM sleep burst neurons, PGO waves, and eye movement information. *Journal of Neurophysiology*, 50(4):784–797.

[171] Ngo, H.-V. V., Martinetz, T., Born, J., and MÃ¶lle, M. (2013). Auditory closed-loop stimulation of the sleep slow oscillation enhances memory. *Neuron*, 78(3):545–553.

[172] Ngo, H.-V. V., Miedema, A., Faude, I., Martinetz, T., MÃ¶lle, M., and Born, J. (2015). Driving sleep slow oscillations by auditory closed-loop stimulation-a self-limiting process. *The Journal of Neuroscience*, 35(17):6630–6638.

[173] Niethard, N., Ngo, H.-V. V., Ehrlich, I., and Born, J. (2018). Cortical circuit activity underlying sleep slow oscillations and spindles. *Proceedings of the National Academy of Sciences of the United States of America*, 115(39):E9220–E9229.

[174] Nir, Y., Staba, R. J., Andrillon, T., Vyazovskiy, V. V., Cirelli, C., Fried, I., and Tononi, G. (2011). Regional slow waves and spindles in human sleep. *Neuron*, 70(1):153–169.

[175] Nishida, M., Pearsall, J., Buckner, R. L., and Walker, M. P. (2009). REM sleep, prefrontal theta, and the consolidation of human emotional memory. *Cerebral Cortex*, 19(5):1158–1166.

[176] Oda, S., Sato, F., Okada, A., Akahane, S., Igarashi, H., Yokofujita, J., Yang, J., and Kuroda, M. (2007). Immunolocalization of muscarinic receptor subtypes in the reticular thalamic nucleus of rats. *Brain Research Bulletin*, 74(5):376–384.

[177] Ognjanovski, N., Broussard, C., Zochowski, M., and Aton, S. J. (2018). Hippocampal network oscillations rescue memory consolidation deficits caused by sleep loss. *Cerebral Cortex*, 28(10):3711–3723.

[178] Ognjanovski, N., Maruyama, D., Lashner, N., Zochowski, M., and Aton, S. J. (2014). CA1 hippocampal network activity changes during sleep-dependent memory consolidation. *Frontiers in Systems Neuroscience*, 8:61.

[179] Ognjanovski, N., Schaeffer, S., Wu, J., Mofakham, S., Maruyama, D., Zochowski, M., and Aton, S. J. (2017). Parvalbumin-expressing interneurons coordinate hippocampal network dynamics required for memory consolidation. *Nature Communications*, 8:15039.

[180] Ong, J. L., Lo, J. C., Chee, N. I. Y. N., Santostasi, G., Paller, K. A., Zee, P. C., and Chee, M. W. L. (2016). Effects of phase-locked acoustic stimulation during a nap on EEG spectra and declarative memory consolidation. *Sleep Medicine*, 20:88–97.

[181] Oyanedel, C. N., DurÃ¡n, E., Niethard, N., Inostroza, M., and Born, J. (2020). Temporal associations between sleep slow oscillations, spindles and ripples. *The European Journal of Neuroscience*.

[182] Pare, D., Curro Dossi, R., Datta, S., and Steriade, M. (1990). Brainstem genesis of reserpine-induced ponto-geniculo-occipital waves: an electrophysiological and morphological investigation. *Experimental Brain Research*, 81(3):533–544.

[183] Partlo, L. A. and Sainsbury, R. S. (1996). Influence of medial septal and entorhinal cortex lesions on theta activity recorded from the hippocampus and median raphe nucleus. *Physiology & Behavior*, 59(4-5):887–895.

[184] Pearl, J. (2000). *Causality: models, reasoning and inference*, volume 29. Cambridge Univ Press.

[185] Peters, J., Janzing, D., and Schölkopf, B. (2013). Causal inference on time series using restricted structural equation models. In *Advances in Neural Information Processing Systems*, pages 154–162.

[186] Peters, J., Janzing, D., and Schölkopf, B. (2017). *Elements of Causal Inference – Foundations and Learning Algorithms*. MIT Press.

[187] Peyrache, A., Khamassi, M., Benchenane, K., Wiener, S. I., and Battaglia, F. P. (2009). Replay of rule-learning related neural patterns in the prefrontal cortex during sleep. *Nature Neuroscience*, 12(7):919–926.

[188] Popa, D., Duvarci, S., Popescu, A. T., Lena, C., and Pare, D. (2010). Coherent amygdalocortical theta promotes fear memory consolidation during paradoxical sleep. *Proceedings of the National Academy of Sciences of the United States of America*, 107(14):6516–6519.

[189] Puentes-Mestril, C. and Aton, S. J. (2017). Linking network activity to synaptic plasticity during sleep: Hypotheses and recent data. *Frontiers in Neural Circuits*, 11:61.

[190] Puentes-Mestril, C., Roach, J., Niethard, N., Zochowski, M., and Aton, S. J. (2019). How rhythms of the sleeping brain tune memory and synaptic plasticity. *Sleep*, 42(7).

[191] Rajan, J. J. and Rayner, P. J. (1996). Generalized feature extraction for time-varying autoregressive models. *IEEE transactions on signal processing*, 44(10):2498–2507.

[192] Ramirez-Villegas, J. F., Besserve, M., Murayama, Y., Evrard, H. C., Oeltermann, A., and Logothetis, N. K. (2020). Coupling of hippocampal theta and ripples with pontogeniculooccipital waves. *Nature*.

[193] Ramirez-Villegas, J. F., Logothetis, N. K., and Besserve, M. (2015). Diversity of sharp-wave–ripple lfp signatures reveals differentiated brain-wide dynamical events. *Proceedings of the National Academy of Sciences*, 112(46):E6379–E6387.

[194] Rasch, B., Pommer, J., Diekelmann, S., and Born, J. (2009). Pharmacological REM sleep suppression paradoxically improves rather than impairs skill memory. *Nature Neuroscience*, 12(4):396–397.

[195] Roach, J. P., Pidde, A., Katz, E., Wu, J., Ognjanovski, N., Aton, S. J., and Zochowski, M. R. (2018). Resonance with subthreshold oscillatory drive organizes activity and optimizes learning in neural networks. *Proceedings of the National Academy of Sciences of the United States of America*, 115(13):E3017–E3025.

[196] Robinson, P. A. (2011). Neural field theory of synaptic plasticity. *Journal of Theoretical Biology*, 285(1):156–163.

[197] Rosanova, M. and Ulrich, D. (2005). Pattern-specific associative long-term potentiation induced by a sleep spindle-related spike train. *The Journal of Neuroscience*, 25(41):9398–9405.

[198] Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688.

[199] Ruder, S. (2016). An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*.

[200] Safavi, S., Panagiotaropoulos, T., Kapoor, V., Ramirez-Villegas, J. F., Logothetis, N. K., and Besserve, M. (2020). Uncovering the organization of neural circuits with generalized phase locking analysis. *bioRxiv*.

[201] Sakai, K. and Jouvet, M. (1980). Brain stem PGO-on cells projecting directly to the cat dorsal lateral geniculate nucleus. *Brain Research*, 194(2):500–505.

[202] Sanchez, R. and Leonard, C. S. (1994). NMDA receptor-mediated synaptic input to nitric oxide synthase-containing neurons of the guinea pig mesopontine tegmentum in vitro. *Neuroscience Letters*, 179(1-2):141–144.

[203] Sanchez, R. and Leonard, C. S. (1996). NMDA-receptor-mediated synaptic currents in guinea pig laterodorsal tegmental neurons in vitro. *Journal of Neurophysiology*, 76(2):1101–1111.

[204] Sano, K., Iwahara, S., Senba, K., Sano, A., and Yamazaki, S. (1973). Eye movements and hippocampal theta activity in rats. *Electroencephalography and Clinical Neurophysiology*, 35(6):621–625.

[205] Schabus, M., Gruber, G., Parapatics, S., Sauter, C., KlÃ¶sch, G., Anderer, P., Klimesch, W., Saletu, B., and Zeitlhofer, J. (2004). Sleep spindles and their significance for declarative memory consolidation. *Sleep*, 27(8):1479–1485.

[206] Schellenberger Costa, M., Weigenand, A., Ngo, H.-V. V., Marshall, L., Born, J., Martinetz, T., and Claussen, J. C. (2016). A thalamocortical neural mass model of the EEG during NREM sleep and its response to auditory stimulation. *PLoS Computational Biology*, 12(9):e1005022.

[207] Schlogl, A., Roberts, S., and Pfurtscheller, G. (2000). A criterion for adaptive autoregressive models. In *Proceedings of the 22nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (Cat. No.00CH37143)*, pages 1581–1582. IEEE.

[208] Schoenwald, S. V., de Santa-Helena, E. L., Rossatto, R., Chaves, M. L., and Gerhardt, G. J. (2006). Benchmarking matching pursuit to find sleep spindles. *Journal of Neuroscience Methods*, 156(1):314–321.

[209] Schreiber, T. (2000). Measuring information transfer. *Physical review letters*, 85(2):461.

[210] Seibt, J., Richard, C. J., Sigl-Glockner, J., Takahashi, N., Kaplan, D. I., Doron, G., de Limoges, D., Bocklisch, C., and Larkum, M. E. (2017). Cortical dendritic activity correlates with spindle-rich oscillations during sleep in rodents. *Nature Communications*, 8(1):684.

[211] Sejnowski, T. J. and Destexhe, A. (2000). Why do we sleep? *Brain research*, 886(1-2):208–223.

[212] Seth, A. K. (2010). A MATLAB toolbox for granger causal connectivity analysis. *Journal of Neuroscience Methods*, 186(2):262–273.

[213] Sherstinsky, A. (2020). Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network. *Physica D: Nonlinear Phenomena*, 404:132306.

[214] Silvestri, A. J. and Kapp, B. S. (1998). Amygdaloid modulation of mesopontine peribrachial neuronal activity: implications for arousal. *Behavioral Neuroscience*, 112(3):571–588.

[215] Sirota, A., Csicsvari, J., Buhl, D., and Buzsaki, G. (2003). Communication between neocortex and hippocampus during sleep in rodents. *Proceedings of the National Academy of Sciences of the United States of America*, 100(4):2065–2069.

[216] Solo, V. (2016). State-space analysis of granger-geweke causality measures with application to fMRI. *Neural Computation*, 28(5):914–949.

[217] Soma, S., Shimegi, S., Osaki, H., and Sato, H. (2012). Cholinergic modulation of response gain in the primary visual cortex of the macaque. *Journal of Neurophysiology*, 107(1):283–291.

[218] Staresina, B. P., Bergmann, T. O., Bonnefond, M., van der Meij, R., Jensen, O., Deuker, L., Elger, C. E., Axmacher, N., and Fell, J. (2015). Hierarchical nesting of slow oscillations, spindles and ripples in the human hippocampus during sleep. *Nature Neuroscience*, 18(11):1679–1686.

[219] Steriade, M., Deschenes, M., Domich, L., and Mulle, C. (1985). Abolition of spindle oscillations in thalamic neurons disconnected from nucleus reticularis thalami. *Journal of Neurophysiology*, 54(6):1473–1497.

[220] Steriade, M., Domich, L., Oakson, G., and Deschenes, M. (1987). The deafferented reticular thalamic nucleus generates spindle rhythmicity. *Journal of Neurophysiology*, 57(1):260–273.

[221] Steriade, M. and LlinÃ¡s, R. R. (1988). The functional states of the thalamus and the associated neuronal interplay. *Physiological Reviews*, 68(3):649–742.

[222] Steriade, M., McCormick, D. A., and Sejnowski, T. J. (1993a). Thalamocortical oscillations in the sleeping and aroused brain. *Science*, 262(5134):679–685.

[223] Steriade, M., Nunez, A., and Amzica, F. (1993b). Intracellular analysis of relations between the slow (<1 hz) neocortical oscillation and other sleep rhythms of the electroencephalogram. *The Journal of Neuroscience*, 13(8):3266–3283.

[224] Steriade, M., Pare, D., Bouhassira, D., Deschenes, M., and Oakson, G. (1989). Phasic activation of lateral geniculate and perigeniculate thalamic neurons during sleep with ponto-geniculo-occipital waves. *The Journal of Neuroscience*, 9(7):2215–2229.

[225] Steriade, M., Pare, D., Datta, S., Oakson, G., and Curro Dossi, R. (1990). Different cellular types in mesopontine cholinergic nuclei related to ponto-geniculo-occipital waves. *The Journal of Neuroscience*, 10(8):2560–2579.

[226] Steriade, M., Timofeev, I., and Grenier, F. (2001). Natural waking and sleep states: a view from inside neocortical neurons. *Journal of Neurophysiology*, 85(5):1969–1985.

[227] Stokes, P. A. and Purdon, P. L. (2017). A study of problems encountered in granger causality analysis from a neuroscience perspective. *Proceedings of the National Academy of Sciences of the United States of America*, 114(34):E7063–E7072.

[228] Stowell, R. D., Sipe, G. O., Dawes, R. P., Batchelor, H. N., Lordy, K. A., Whitelaw, B. S., Stoessel, M. B., Bidlack, J. M., Brown, E., Sur, M., and Majewska, A. K. (2019). Noradrenergic signaling in the wakeful state inhibits microglial surveillance and synaptic plasticity in the mouse visual cortex. *Nature Neuroscience*, 22(11):1782–1792.

[229] Sullivan, D., Csicsvari, J., Mizuseki, K., Montgomery, S., Diba, K., and Buzsaki, G. (2011). Relationships between hippocampal sharp waves, ripples, and fast gamma oscillation: influence of dentate and entorhinal cortical activity. *Journal of Neuroscience*, 31(23):8605–8616.

[230] Sun, Y.-G., Pita-Almenar, J. D., Wu, C.-S., Renger, J. J., Uebele, V. N., Lu, H.-C., and Beierlein, M. (2013). Biphasic cholinergic synaptic transmission controls action potential activity in thalamic reticular nucleus neurons. *The Journal of Neuroscience*, 33(5):2048–2059.

[231] Takens, F. (1981). Detecting strange attractors in turbulence. In *Dynamical systems and turbulence, Warwick 1980*, pages 366–381. Springer.

[232] Tallon-Baudry, C. and Bertrand, O. (1999). Oscillatory gamma activity in humans and its role in object representation. *Trends in cognitive sciences*, 3(4):151–162.

[233] Tamaki, M., Huang, T.-R., Yotsumoto, Y., Hamalainen, M., Lin, F.-H., Nanez, J. E., Watanabe, T., and Sasaki, Y. (2013). Enhanced spontaneous oscillations in the supplementary motor area are associated with sleep-dependent offline learning of finger-tapping motor-sequence task. *The Journal of Neuroscience*, 33(34):13894–13902.

[234] Tamaki, M., Matsuoka, T., Nittono, H., and Hori, T. (2008). Fast sleep spindle (13-15 hz) activity correlates with sleep-dependent improvement in visuomotor performance. *Sleep*, 31(2):204–211.

[235] Tononi, G. and Cirelli, C. (2003). Sleep and synaptic homeostasis: a hypothesis. *Brain research bulletin*, 62(2):143–150.

[236] Tononi, G. and Cirelli, C. (2014). Sleep and the price of plasticity: from synaptic and cellular homeostasis to memory consolidation and integration. *Neuron*, 81(1):12–34.

[237] Tononi, G. and Cirelli, C. (2020). Sleep and synaptic down-selection. *The European Journal of Neuroscience*, 51(1):413–421.

[238] Vandekerckhove, M. and Wang, Y.-L. (2018). Emotion, emotion regulation and sleep: An intimate relationship. *AIMS neuroscience*, 5(1):1–17.

[239] Varela, C. (2014). Thalamic neuromodulation and its implications for executive networks. *Frontiers in Neural Circuits*, 8:69.

[240] Varela, C. and Sherman, S. M. (2007). Differences in response to muscarinic activation between first and higher order thalamic relays. *Journal of Neurophysiology*, 98(6):3538–3547.

[241] Watson, B. O., Levenstein, D., Greene, J. P., Gelinas, J. N., and Buzsaki, G. (2016). Network homeostasis and state dynamics of neocortical sleep. *Neuron*, 90(4):839–852.

[242] Weigenand, A., Schellenberger Costa, M., Ngo, H.-V. V., Claussen, J. C., and Martinetz, T. (2014). Characterization of k-complexes and slow wave activity in a neural mass model. *PLoS Computational Biology*, 10(11):e1003923.

[243] Wetzel, W., Ott, T., and Matthies, H. (1977). Post-training hippocampal rhythmic slow activity elicited by septal stimulation improves memory consolidation in rats. *Behavioral Biology*, 21(1):32–40.

[244] Wibral, M., Pampu, N., Priesemann, V., Siebenhuehner, F., Seiwert, H., Lindner, M., Lizier, J. T., and Vicente, R. (2013). Measuring information-transfer delays. *Plos One*, 8(2):e55809.

[245] Wierzynski, C. M., Lubenov, E. V., Gu, M., and Siapas, A. G. (2009). State-dependent spike-timing relationships between hippocampal and prefrontal circuits during sleep. *Neuron*, 61(4):587–596.

[246] Williams, J. A. and Reiner, P. B. (1993). Noradrenaline hyperpolarizes identified rat mesopontine cholinergic neurons in vitro. *The Journal of Neuroscience*, 13(9):3878–3883.

[247] Wilson, H. R. and Cowan, J. D. (1972). Excitatory and inhibitory interactions in localized populations of model neurons. *Biophysical Journal*, 12(1):1–24.

[248] Wilson, H. R. and Cowan, J. D. (1973). A mathematical theory of the functional dynamics of cortical and thalamic nervous tissue. *Kybernetik*, 13(2):55–80.

[249] Wilson, M. T., Fulcher, B. D., Fung, P. K., Robinson, P. A., Fornito, A., and Rogasch, N. C. (2018). Biophysical modeling of neural plasticity induced by transcranial magnetic stimulation. *Clinical Neurophysiology*, 129(6):1230–1241.

[250] Wilson, M. T., Goodwin, D. P., Brownjohn, P. W., Shemmell, J., and Reynolds, J. N. J. (2014). Numerical modelling of plasticity induced by transcranial magnetic stimulation. *Journal of Computational Neuroscience*, 36(3):499–514.

[251] Yang, G., Lai, C. S. W., Cichon, J., Ma, L., Li, W., and Gan, W.-B. (2014). Sleep promotes branch-specific formation of dendritic spines after learning. *Science*, 344(6188):1173–1178.

[252] Ylinen, A., Bragin, A., Nadasdy, Z., Jando, G., Szabo, I., Sik, A., and Buzsaki, G. (1995). Sharp wave-associated high-frequency oscillation (200 hz) in the intact hippocampus: network and intracellular mechanisms. *The Journal of Neuroscience*, 15(1 Pt 1):30–46.

[253] Zhou, Y., Lai, C. S. W., Bai, Y., Li, W., Zhao, R., Yang, G., Frank, M. G., and Gan, W.-B. (2020). REM sleep promotes experience-dependent dendritic spine elimination in the mouse cortex. *Nature Communications*, 11(1):4819.

[254] Zhu, J. J., Lytton, W. W., Xue, J. T., and Uhlrich, D. J. (1999). An intrinsic oscillation in interneurons of the rat lateral geniculate nucleus. *Journal of Neurophysiology*, 81(2):702–711.

[255] Zhu, J. J. and Uhlrich, D. J. (1998). Cellular mechanisms underlying two muscarinic receptor-mediated depolarizing responses in relay cells of the rat lateral geniculate nucleus. *Neuroscience*, 87(4):767–781.

[256] Zutshi, I., Brandon, M. P., Fu, M. L., Donegan, M. L., Leutgeb, J. K., and Leutgeb, S. (2018). Hippocampal neural circuits respond to optogenetic pacing of theta frequencies by generating accelerated oscillation frequencies. *Current Biology*, 28(8):1179–1188.e3.

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

AIC	Akaike Information Criterion
AMPA	α-amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid
BIC	Bayesian Information Criterion
CA1	Cornu ammonis field 1
CA3	Cornu ammonis field 3
C-PBL	Caudolateral Peribrachial
CS	Causal Strength
DCM	Dynamic Causal Modelling
DCS	Dynamic Causal Strength
DeSnap	Debiased Snapshot
EEG	electroencephalogram
GABA	Gamma-aminobutyric acid
GC	Granger Causality
GRU	Gated Recurrent Unit
In	Inhibitory (neurons)
KL	Kullback-Leibler
LFP	Local Field Potential
LGin	intra-Geniculate inter(neurons)
LGN	Lateral Geniculate Nucleues
LSTM	Long Short-Term Memory
LTD	Long-term depression
LTP	Long-term potentiation
mAChR	muscarinic Acetylcholine Receptor
ML	Maximum Likelihood
MSE	Mean Squared Errors
MUA	Multi Unit Activity
nAChR	nicotinic Acetylcholine Receptor
NET-fMRI	Neural-Event-Triggered functional Magnetic Resonance Imaging
NREM	Non-Rapid-Eye-Movement
ODE	Ordinary Differential Equation

OLS	Ordinary Least Square
PBL	Peribrachial (area)
PGN	Peri-Geniculate Nucleues
PGO	Ponto-Geniculo-occipital
PSC	Postsynaptic Current
PV+	parvalbumin-expressing
Pyr	Pyramidal (neurons)
rDCS	relative Dynamic Causal Strength
REM	Rapid-Eye-Movement
RNN	Recurrent Neural Network
R-PBL	Rostal Peribrachial
RT	Reticular thalamic (neurons)
SCM	Structural Causal Model
SD	Standard Deviation
SEM	Structural Equation Models
SHY	Synaptic Homeostasis Hypothesis
SO	Slow Oscillations
SPW-R	Sharp Wave-Ripples
sSE	Sum of squared errors
STDP	Spike-Time-Dependent-Plasticity
TC	Thalamocortical (neurons)
TE	Transfer Entropy
TMS	Transcranial Magnetic Stimulation
TRN	Thalamic Reticular Nucleus
VAR	Vector Autoregressive

STATEMENT OF CONTRIBUTIONS

The work presented in this thesis are mainly done by Kaidi Shao (K.S.) under the supervison of Dr. Michel Besserve (M.B.). Other collaborators that contributed include Prof. Dr. Nikos K. Logothetis (N.K.L.) and Dr. Juan F. Ramirez Villegas (J.F.R.V.).

Currently, contents in Chapter 2 has been submitted, while contents in Chapter 3 and Chapter 4 are in preparation as manuscripts to be submitted. In the following sections the contribution of each chapter will be elaborated.

CONTRIBUTION OF CHAPTER 2

- Conceptualization: M.B., J.F.R.V., K.S.
- Methodology: K.S., M.B., J.F.R.V.
- Simulation: K.S.
- Validation: K.S.
- Formal analysis: K.S.
- Data Curation: K.S.
- Visualization: K.S.
- Writing - original draft preparation: K.S., M.B.
- Writing - review and editing: M.B., J.F.R.V., K.S.
- Supervision: M.B.
- Funding acquisition: N.K.L.

M.B. and J.F.R.V. devised the project inspired by previous research, where K.S. further completed the picture. K.S. constructed the full model after literature review, during which M.B. and J.F.R.V. gave advice. K.S. simulated the model and tuned model parameters. K.S. did further analysis on the simulation results. K.S. designed and plotted the figures. K.S. and M.B. wrote the manuscript. J.F.R.V. reviewed the manuscript and gave suggestions. M.B. supervised the whole process. N.K.L. offered advice and provided funding.

CONTRIBUTION OF CHAPTER 3

- Conceptualization: M.B., K.S.
- Methodology: M.B., K.S.
- Simulation: K.S.
- Formal analysis: K.S.
- Resources: N.K.L.
- Visualization: K.S.
- Data Curation: K.S.

- Writing - original draft preparation: K.S., M.B.
- Writing - review and editing: M.B., K.S.
- Supervision: M.B.
- Funding acquisition: N.K.L.

M.B. and K.S. conceptualized the main idea. M.B. designed a toy model to illustrate the problem and K.S. implemented the whole rest parts. M.B. and K.S. derived the equations of selection bias correction. K.S. implemented the equations and applied the method to causality measures. N.K.L collected the empirical data. K.S. applied the methods to simulated and empirical data and plotted the figures. K.S. and M.B. wrote the manuscripts. M.B. supervised the whole process. N.K.L. offered advice and provided funding.

CONTRIBUTION OF CHAPTER 4

- Conceptualization: M.B.
- Methodology: M.B., K.S.
- Simulation: K.S.
- Formal analysis: K.S.
- Resources: N.K.L.
- Visualization: K.S.
- Data Curation: K.S.
- Writing - original draft preparation: K.S., M.B.
- Writing - review and editing: M.B., K.S.
- Supervision: M.B.
- Funding acquisition: N.K.L.

M.B. conceptualized the main idea while K.S. complimented the interpretations. M.B. and K.S. derived the TE, DCS and rDCS based on VAR models. K.S. implemented the measures. N.K.L collected the empirical data. K.S. applied the methods to simulated and empirical data and plotted the figures. K.S. and M.B. wrote the manuscripts. M.B. supervised the whole process. N.K.L. offered advice and provided funding.

ACKNOWLEDGMENT

It feels like the time when I was writing my Master's thesis in the same lab is as vivid as yesterday's memory when I am wrapping up for my Ph.D.'s work. The accomplishment of my Ph.D. is indebted to many people who offered me invaluable help in the last three years. I am extending my gratefulness to them here.

First and foremost, I would like to express my sincerest gratitude to my supervisor Dr. Michel Besserve. Michel is such a brilliant scientist that he is the scientific hero that I would always want to become myself. Throughout the four years of working with him, constantly amazed by his ideas with sparkling intelligence, I have gradually understood what kind of mindset and knowledge base one would need to think fundamentally of scientific problems and form elegant theories. Besides, I cannot imagine before how such an intelligent person can be at the same time super gentle and considerate. I am very grateful that Michel has always been able to think on the student's side and never spared any effort to support me during difficult times. Michel has always been providing much more than I expected or even deserved. I am very sure that my Ph.D. would never be accomplished like this without him.

It's an honor that Prof. Nikos Logothetis could provide me with the opportunity to do my Ph.D. in his lab. Nikos and the fascinating research conducted in his lab are what I have been admiring for a long time. Being inside this group myself, I have been even impressed by his intelligence, insights, and active devotion to front-line research, rare for a distinguished scientist like him. His strong determination against the infuriating injustice against animal experiments is constantly encouraging me during the hard times. Besides, his lifelong passion for classical music and piano playing also motivates me to keep practicing myself, which turns out very helpful for both science and life.

I want to thank my advisory board members and my second reader for providing constructive advice and support that guided my projects in the right direction. Thanks, Prof. Steffen Gais, for stressing the importance of validation with real data, which is critical to bear in mind for a computational neuroscientist. Thanks, Prof. Anna Levina, for helping me a lot scientifically and setting a unique role model for female scientists: from her, I can have a glimpse of how a well-balanced mentality can manage to be friendly without losing oneself at the same time. Thanks, Prof. Martin Giese, whom I have always been impressed by the solid theoretical background, for being willing to help with my thesis evaluation when I was cornered with administrative problems. Thanks, Prof. Masataka Watanabe, for being willing to evaluate my defense and provided insightful guidance in all aspects of science and life during his stay in the lab.

I would also like to thank the IT technician Joachim Werner for fixing my endlessly crashing Linux virtual machine and other technical problems. Besides, I am fortunate to have him as a neighbor. His kindness and patience helped me deal with many personal issues in the lab and at home.

I would also like to thank my current and former group members and collaborators: Shervin Safavi, Roxana Zaraati and Hongbiao Chen and Juan Ramirez-Villegas, for having many inspiring discussions and mutual supports. Especially, I am very grateful to Shervin has been guiding me since my Master's time in the same lab; he is always so nice and supportive and has helped me in all aspects of both research and life. Roxana always impressed me with her elegant intelligence. Thanks for taking care of me during the conference we participated in together. Thanks, Hongbiao, for bringing so many delicious snacks and constantly reminded me of the importance of health while I was pushing myself too hard.

I would also like to thank my lab colleagues, Mingyu Yang, Hao Mei, Kun Wang, Leonid Federov, Alireza Saedi, Weiyi Xiao, Rui Kimura, Ryo Iwai, Shuchen Wu, Dongmei Shi, etc., for making the lab feel like a family. Without their company and all the juicy discussions, delicious food, outings and movie nights we enjoyed together, I could not imagine how hard it might be to endure the corona time.

I would also like to thanks my close friends - Yue Zhang, Jialin Zhou, Haven Feng, Jiatong Liu, Jiexiu Zhai, Kun Wang, for sharing delicious food, inspiring discussions and mental supports during spare times.

Finally, I cannot tell how much I am grateful to my parents, grandfather, and boyfriend Zhijian Jiang: without their unconditional love, patience, and healing comforts during insomniac nights, I guess I could not have survived the painful spiritual crisis that made me who I am now. They are like my Patronus against the dementors that always give me the courage to confront real life.

Part IV

APPENDIX

APPENDIX A

A.1 DERIVATION OF SYNAPSE REPRESENTATION

A.1.1 *Alpha function*

For a synapse current (type m) from presynaptic population k' to postsynaptic population k with connectivity strength $N_{k'k}$, assume its impulse response is an alpha function:

$$h_m(t) = \gamma_m^2 \cdot t \cdot \exp(-\gamma_m t)$$

If we perform a Laplace transform to the impulse response, we would obtain:

$$\mathcal{L}[h_m(t)](s) = \mathcal{L}[\gamma_m^2 t \exp(-\gamma_m t)](s) = \gamma_m^2 \cdot \frac{\Gamma(2)}{(s+\gamma_m)^2} = \frac{\gamma^2}{(s+\gamma_m)^2}$$

Here the impulse response equals:

$$h_m(t) = \frac{s_{mk}(s)}{N_{k'k} \cdot Q_{k'}(s)} = \frac{\gamma_m^2}{s^2 + 2\gamma_m s + \gamma_m^2}$$

If we rearrange the terms, we can get:

$$s^2 \cdot s_{mk}(s) + 2s \cdot s_{mk}(s) + \gamma_m^2 \cdot s_{mk}(s) = \gamma_m^2 \cdot N_{k'k} \cdot Q_{k'}(s)$$

After an inverse Laplace transform and a rearrangement of the terms, we can reach the differential equation that we want:

$$\ddot{s}_{mk} = \gamma_m^2 (N_{k'k} \cdot Q_{k'}(V_{k'}(t)) - s_{mk}) - 2\gamma_m \dot{s}_{mk}$$

A.1.2 *'Two-exponential' inpulse response function*

Similarly, for the 'two-exponential' type of synaptic impulse response:

$$h_m(t) = B(\exp(-t/\tau_1) - \exp(-t/\tau_2))$$

The Laplace transformed $h_m(t)$ would take the following form:

$$\mathcal{L}[h_m(t)](s) = \mathcal{L}[B(\exp(-t/\tau_1) - \exp(-t/\tau_2))](s) = \frac{B}{s+\tau_1^{-1}} - \frac{B}{s+\tau_2^{-1}}$$

Then we include the input and output of the system in the Laplace domain:

$$\frac{s_{mk}(s)}{N_{k'k} \cdot Q_{k'}(s)} = \frac{B(\tau_2^{-1} - \tau_1^{-1})}{s^2 + (\tau_2^{-1} + \tau_1^{-1})s + \tau_2^{-1}\tau_1^{-1}}$$

If we rearrange the terms, we can get:

$$s^2 \cdot s_{mk}(s) + (\tau_2^{-1} + \tau_1^{-1})s \cdot s_{mk}(s) + \tau_2^{-1}\tau_1^{-1} \cdot s_{mk}(s) = B(\tau_2^{-1} - \tau_1^{-1}) \cdot N_{k'k} \cdot Q_{k'}(s)$$

After an inverse Laplace transform and a rearrangement of the terms, The differential version is obtained after a reverse Laplace transform

$$\ddot{s}_{mk} = B(\tau_2^{-1} - \tau_1^{-1}) N_{k'k} \cdot Q_{k'}(V_{k'}(t)) - \tau_2^{-1}\tau_1^{-1} s_{mk} - (\tau_2^{-1} + \tau_1^{-1})\dot{s}_{mk}$$

A.2 FULL LIST OF EQUATIONS FOR PGO MODEL

A.2.1 *Neuron populations*

Firing rate function of population k, where $k \in \{t, r, g, l, c\}$, representing the TC neurons, the RT neuron, the intra-LG interneurons, the PGO transferring neurons and the PGO triggering neurons.

$$Q_k = \frac{Q_k^{max}}{1 + \exp(-(V_k - \theta_k)/\sigma_k)}$$

Membrane potential adaptations for population t, r, g, l

$$\begin{aligned} \tau_t \dot{V}_t = \; & -J_L^t - J_{AMPA}(s_{at}) - J_{nAChR}(s_{nt}) - J_{GABA}(s_{gt}) - \\ & C_m^{-1} \tau_t (I_{LK}^t + I_T^t + I_h) \end{aligned}$$

$$\tau_r \dot{V}_r = -J_L^r - J_{AMPA}(s_{ar}) - J_{GABA}(s_{gr}) - J_{nAChR}(s_{nr}) \\ - J_{mAChR}(s_{mr}) - C_m^{-1} \tau_r (I_{LK}^r + I_T^r)$$

$$\tau_g \dot{V}_g = -J_L^g - J_{nAChR}(s_{ng}) - J_{mAChR}(s_{ng})$$

$$\tau_l \dot{V}_l = -J_L^l - J_{AMPA}(s_{al}) - J_{NMDA}(s_{dl}) - C_m^{-1} \tau_l (I_T^l + I_{ACh})$$

A.2.2 *Synaptic currents*

Synaptic currents are denoted in the form of $J(s_{ij})$. $i \in \{a, g, n, m, d\}$ states the synaptic type AMPA, GABA, nAChR, mAChR and NMDA, whereas $j \in \{t, r, g, l, c\}$ indicates the postsynaptic population in Section A.2.1.

General leaky currents of population k, where $k \in \{t, r, g, l, c\}$

$$J_L^k = (V_k - E_L^k)$$

AMPA synapse in postsynaptic population t

$$J_{AMPA}(s_{at}) = s_{at} \cdot (V_r - E_{at})$$

$$\ddot{s}_{at} = \gamma_{at}^2 (\phi_n - s_{at}) - 2\gamma_{at} \dot{s}_{at}$$

GABA synapse in postsynaptic population t

$$J_{GABA}(s_{gt}) = s_{gt} \cdot (V_t - E_{gt})$$

$$\ddot{s}_{gt} = \gamma_{gt}^2 \cdot (N_{gt} \cdot Q_r(V_r) + N_{gt} \cdot Q_g(V_g) - s_{gt}) - 2\gamma_{gt} \dot{s}_{gt}$$

nAChR synapse in postsynaptic population t

$$J_{nAChR}(s_{nt}) = s_{nt} \cdot (V_t - E_{nt})$$

$$\ddot{s}_{nt} = B_{nt}(\tau_{nt,2}^{-1} - \tau_{nt,1}^{-1}) N_{nt} Q_l(V_l) - \tau_{nt,2}^{-1} \tau_{nt,1}^{-1} s_{nt} - (\tau_{nt,2}^{-1} + \tau_{nt,1}^{-1}) \dot{s}_{nt}$$

mAChR synapse in postsynaptic population t

$$J_{mAChR}(s_{mt}) = s_{mt} \cdot (V_t - E_{mt})$$

$$\ddot{s}_{mt} = \gamma_{mt}^2 (N_{mt} \cdot Q_l(V_l) - s_{mt}) - 2\gamma_{mt} \dot{s}_{mt}$$

AMPA synapse in postsynaptic population r

$$J_{AMPA}(s_{ar}) = s_{ar} \cdot (V_r - E_{ar})$$

$$\ddot{s}_{ar} = \gamma_{ar}^2\left((N_{ar}) \cdot Q_t(V_t) - s_{ar}\right) - 2\gamma_{ar}\dot{s}_{ar}$$

GABA synapse for self-feedback in population r

$$J_{GABA}(s_{gr}) = s_{gr} \cdot (V_r - E_{gr})$$

$$\ddot{s}_{gr} = \gamma_{gr}^2\left((N_{gr}) \cdot Q_r(V_r) - s_{gr}\right) - 2\gamma_{gr}\dot{s}_{gr}$$

nAChR synapse in postsynaptic population r

$$J_{nAChR}(s_{nr}) = s_{nr} \cdot (V_r - E_{nr})$$

$$\ddot{s}_{nr} = B_{nr}(\tau_{nr,2}^{-1} - \tau_{nr,1}^{-1})N_{nr}Q_l(V_l) - \tau_{nr,2}^{-1}\tau_{nr,1}^{-1}s_{nr} - (\tau_{nr,2}^{-1} + \tau_{nr,1}^{-1})\dot{s}_{nr}$$

mAChR synapse in postsynaptic population r

$$J_{mAChR}(s_{mr}) = s_{mr} \cdot g_{mAChR}(V_r) \cdot (V_r - E_{mr})$$

$$g_{mAChR}(V_r) = \frac{1}{1 + \exp((V_r + 66.3)/29.1)}$$

$$\ddot{s}_{mr} = B_{mr}(\tau_{mr,2}^{-1} - \tau_{mr,1}^{-1})N_{mr}Q_l(V_l) - \tau_{mr,2}^{-1}\tau_{mr,1}^{-1}s_{mr} - (\tau_{mr,2}^{-1} + \tau_{mr,1}^{-1})\dot{s}_{mr}$$

nAChR synapse in postsynaptic population g

$$J_{nAChR}(s_{ng}) = s_{ng} \cdot (V_g - E_{ng})$$

$$\ddot{s}_{ng} = B_{ng}(\tau_{ng,2}^{-1} - \tau_{ng,1}^{-1})N_{ng}Q_l(V_l) - \tau_{ng,2}^{-1}\tau_{ng,1}^{-1}s_{ng} - (\tau_{ng,2}^{-1} + \tau_{ng,1}^{-1})\dot{s}_{ng}$$

mAChR synapse in postsynaptic population g

$$J_{mAChR}(s_{mg}) = s_{mg} \cdot (V_g - E_{mg})$$

$$\ddot{s}_{mg} = B_{mg}(\tau_{mg,2}^{-1} - \tau_{mg,1}^{-1})N_{mg}Q_l(V_l) - \tau_{mg,2}^{-1}\tau_{mg,1}^{-1}s_{mg} - (\tau_{mg,2}^{-1} + \tau_{mg,1}^{-1})\dot{s}_{mg}$$

AMPA synapse in postsynaptic population l

$$J_{AMPA}(s_{al}) = s_{al} \cdot (V_l - E_{al})$$

$$\ddot{s}_{al} = B_{al}(\tau_{al,2}^{-1} - \tau_{al,1}^{-1})N_{al}Q_c(V_c) - \tau_{al,2}^{-1}\tau_{al,1}^{-1}s_{al} - (\tau_{al,2}^{-1} + \tau_{al,1}^{-1})\dot{s}_{al}$$

NMDA synapse in postsynaptic population l

$$J_{NMDA}(s_{dl}) = s_{dl} \cdot g_{NMDA}(V_l)(V_r - E_{dl})$$

$$g_{NMDA}(V_l) = \frac{1}{1 + \exp(-0.0062V_l)}[Mg^{2+}]_o/3.57$$

$$\ddot{s}_{dl} = B_{dl}(\tau_{dl,2}^{-1} - \tau_{dl,1}^{-1})N_{dl}Q_c(V_c) - \tau_{dl,2}^{-1}\tau_{dl,1}^{-1}s_{dl} - (\tau_{dl,2}^{-1} + \tau_{dl,1}^{-1})\dot{s}_{dl}$$

A.2.3 *Intrinsic currents*

Intrinsic currents are presented as I_i^k, with the subscription $i \in \{LK, T, h, IR\}$ describing the current types (K^+ leaky currents, T-currents, h-currents and k^+ inward rectification currents). The superscription indicates the neuron population where the intrinsic currents are located.

Potassium leaky currents of population t

$$I_{LK}^t = \bar{g}_{LK}^t \cdot (V_t - E_K)$$

Low threshold Calcium T-current for population t

$$I_T^t = \bar{g}_T^t \cdot (m_\infty^t)^2 \cdot h_T^t \cdot (V_t - E_{Ca})$$

$$m^t_\infty = \frac{1}{1+\exp(-(V_t+59)/6.2)}$$

$$\dot{h}^t_T = (h^t_\infty - h^t_T)/\tau^t_h$$

$$h^t_\infty = \frac{1}{1+\exp(-(V_t+81)/4)}$$

$$\tau^t_h = (30.8+(211.4+\exp((V_t+115.2)/5))/(1+\exp((V_t+86)/3.2)))/3^{1.2}$$

Anomalous inward rectifier h-current for population t

$$I^t_h = \overline{g}_h \cdot (m_{h1} + g_{inc} m_{h2}) \cdot (V_t - E_h))$$

$$\dot{m}_{h1} = (m^h_\infty(1-m_{h2}) - m_{h1})/\tau^h_m - k_3 * P_h m_{h1} + k_4 m_{h2}$$

$$\dot{m}_{h2} = k_3 P_h m_{h1} - k_4 m_{h2}$$

$$P_h = k_1[Ca]^4/(k_1[Ca]^4 + k_2)$$

$$[\dot{Ca}] = -\alpha_{Ca} I^t_T - ([Ca] - Ca_0)/\tau_{Ca}$$

Potassium leaky currents of population r

$$I^r_{LK} = \overline{g}^r_{LK} \cdot (V_r - E_K)$$

Low threshold Calcium T-current for population r

$$I^r_T = \overline{g}^r_T \cdot (m^r_\infty)^2 \cdot h^r_T \cdot (V_r - E_{Ca})$$

$$m^r_\infty = \frac{1}{1+\exp(-(V_r+52)/7.4)}$$

$$\dot{h}^r_T = (h^r_\infty - h^r_T)/\tau^r_h$$

$$h^r_\infty = \frac{1}{1+\exp(-(V_r+80)/5)}$$

$$\tau^t_h = (85+1/(\exp((V_r+48)/4))/(1+\exp(-(V_r+407)/50)))/3^{1.2}$$

Low threshold Calcium T-current for population l

$$I^l_T = \overline{g}^l_T \cdot \left(m^l_\infty\right)^2 \cdot h^l_T \cdot (V_l - E_{Ca})$$

$$m^l_\infty = \frac{1}{1+\exp(-(V_l+50.6)/0.44)+\exp((V_l+50.6)/17.4)}$$

$$\dot{h}^l_T = (h^l_\infty - h^l_T)/\tau^l_h$$

$$h^r_\infty = \frac{1}{1+\exp((V_l+65)/2.7)}$$

$$\tau^t_h = (52+(211.4+\exp((V_l+115.2)/5))/(1+\exp((V_l+86)/3.2)))/\exp(1.2\log(3))$$

K^+ Inward rectifier of cholinergic input in population l

$$I^l_{IR} = \overline{g}^l_{IR} \cdot g_{IR}(V_l) \cdot (V_l - E_K)$$

$$g^l_{IR} = \frac{1}{1+\exp((V_l+66.3)/29.1)}$$

A.2.4 *State modulation*

$$\Delta[\mathrm{ACh}](t) = -\cos(\frac{2\pi}{T}t)$$

$$\bar{g}^{t}_{LK}(t) = -0.5(\bar{g}^{t}_{LK,NREM} - \bar{g}^{t}_{LK,REM}) \cdot (\Delta[\mathrm{ACh}](t) - 1) + \bar{g}^{t}_{LK,REM})$$

$$\bar{g}^{r}_{LK}(t) = 0.5(\bar{g}^{r}_{LK,REM} - \bar{g}^{r}_{LK,NREM}) \cdot (\Delta[\mathrm{ACh}](t) + 1) + \bar{g}^{r}_{LK,NREM})$$

A.3 VAR MODEL ESTIMATION

A.3.1 *Estimation of VAR model parameters*

We consider stochastic process $\mathbf{X}_t$ generated by the linear Gaussian vector autoregression of Eq. 3.3:

$$\boldsymbol{X}_t = A_t \boldsymbol{X}_{p,t} + \boldsymbol{\eta}_t, \boldsymbol{\eta}_t \sim \mathcal{N}(\mathbf{k}_t, \Sigma_t) \quad \mathrm{Cov}\left[\boldsymbol{\eta}_t, \boldsymbol{X}_{p,t}\right] = 0 \tag{A.1}$$

As stated in Section 3.2.4.2, the VAR model parameters can be estimated from data. Here we derive the OLS estimation and ML estimation of VAR model parameters.

A.3.1.1 *OLS estimation of VAR model parameters*

For the case where $\boldsymbol{k}_t = 0$, multiplying both sides of Eq. A.1 with the transpose of past states $\boldsymbol{X}_{p,t}$ yields

$$\boldsymbol{X}_t \boldsymbol{X}_{p,t}^{\top} = A_t \boldsymbol{X}_{p,t} \boldsymbol{X}_{p,t}^{\top} + \boldsymbol{\eta}_t \cdot \boldsymbol{X}_{p,t}^{\top}$$

Taking the expectation at both sides eliminates the term related to the innovation

$$\mathbb{E}\left[\boldsymbol{X}_t \boldsymbol{X}_{p,t}^{\top}\right] = A_t \mathbb{E}\left[\boldsymbol{X}_{p,t} \boldsymbol{X}_{p,t}^{\top}\right] + 0$$

Therefore

$$\widehat{A_t} = \mathbb{E}\left[\boldsymbol{X}_t \boldsymbol{X}_{p,t}^{\top}\right] \mathbb{E}\left[\boldsymbol{X}_{p,t} \boldsymbol{X}_{pt}^{\top}\right]^{-1}.$$

For the case where $\boldsymbol{k}_t \neq 0$, first we demean Eq. A.1 by substracting the expectation on both sides

$$\boldsymbol{X}_t - \mathbb{E}\left[\boldsymbol{X}_t\right] = A_t \left(\boldsymbol{X}_{p,t} - \mathbb{E}\left[\boldsymbol{X}_p, t\right]\right) + \left(\boldsymbol{\eta}_t - \mathbf{k}_t\right),$$

This equation times $\left(\boldsymbol{X}_t - \mathbb{E}\left[\boldsymbol{X}_t\right]\right)^{\top}$ at both sides yields

$$\begin{aligned}
&\left(\boldsymbol{X}_t - \mathbb{E}\left[\boldsymbol{X}_t\right)\right] \left(\boldsymbol{X}_{p,t} - \mathbb{E}\left[\boldsymbol{X}_{p,t}\right]\right)^{\top} \\
&= A_t \left(\boldsymbol{X}_{p,t} - \mathbb{E}\left[\boldsymbol{X}_{p,t}\right]\right) \left(\boldsymbol{X}_{p,t} - \mathbb{E}\left[\boldsymbol{X}_{p,t}\right)\right] + \left(\boldsymbol{\eta}_t - \mathbf{k}_t\right) \left(\boldsymbol{X}_{p,t} - \mathbb{E}\left[\boldsymbol{X}_{p,t}\right]\right)^{\top}
\end{aligned}$$

Similarly, apply expectation operation at both sides, we get

$$\mathbb{E}\left[\left(\boldsymbol{X}_t - \mathbb{E}\left[\boldsymbol{X}_t\right]\right) \left(\boldsymbol{X}_{p,t} - \mathbb{E}\left[\boldsymbol{X}_{p,t}\right]\right)^{\top}\right] = A_t \mathbb{E}\left[\left(\boldsymbol{X}_{p,t} - \mathbb{E}\left[\boldsymbol{X}_{p,t}\right]\right) \left(\boldsymbol{X}_{p,t} - \mathbb{E}\left[\boldsymbol{X}_{p,t}\right]\right)^{\mathsf{T}}\right]$$

because

$$\mathbb{E}\left[\left(\boldsymbol{\eta}_t - \mathbf{k}_t\right) \left(\boldsymbol{X}_{p,t} - \mathbb{E}\left[\boldsymbol{X}_{p,t}\right]\right)^{\top}\right] = \mathrm{Cov}\left(\boldsymbol{\eta}_t, \boldsymbol{X}_{p,t}\right) = 0.$$

Therefore

$$\mathrm{Cov}\left(\boldsymbol{X}_t, \boldsymbol{X}_{p,t}\right) = A_t \,\mathrm{Cov}\left(\boldsymbol{X}_{p,t}, \boldsymbol{X}_{p,t}\right)$$

$$\widehat{A_t} = \mathrm{Cov}\left(\boldsymbol{X}_t, \boldsymbol{X}_{p,t}\right) \mathrm{Cov}\left(\boldsymbol{X}_{p,t}, \boldsymbol{X}_{p,t}\right)^{-1} = \boldsymbol{\Sigma}_{\boldsymbol{X}_t \boldsymbol{X}_{p,t}} \left(\boldsymbol{\Sigma}_{\boldsymbol{X}_{p,t}}\right)^{-1} \tag{A.2}$$

For an *i.i.d.* dataset with sample size N, we can estimate two covariance matrices as

$$\Sigma_{\boldsymbol{X}_t \boldsymbol{X}_p} = \frac{1}{N} \sum_{n=1}^{N} (\boldsymbol{X}_t^{(n)} - \mathbb{E}[\boldsymbol{X}_t])(\boldsymbol{X}_{p,t}^{(n)} - \mathbb{E}[\boldsymbol{X}_{p,t}])^\top$$

$$\Sigma_{\boldsymbol{X}_p} = \frac{1}{N} \sum_{n=1}^{N} (\boldsymbol{X}_{p,t}^{(n)} - \mathbb{E}[\boldsymbol{X}_{p,t}])(\boldsymbol{X}_{p,t}^{(n)} - \mathbb{E}[\boldsymbol{X}_{p,t}])^\top$$

where mean can be estimated as $\mathbb{E}[\boldsymbol{X}_t] = \frac{1}{N} \sum_{n=1}^{N} \boldsymbol{X}_t^{(n)}$, $\mathbb{E}[\boldsymbol{X}_{p,t}] = \frac{1}{N} \sum_{n=1}^{N} \boldsymbol{X}_{p,t}^{(n)}$. Therefore the VAR model coefficient A_t can be estimated as

$$\widehat{A_t} = \boldsymbol{\Sigma}_{\boldsymbol{X}_t \boldsymbol{X}_{p,t}} \left(\boldsymbol{\Sigma}_{\boldsymbol{X}_{p,t}}\right)^{-1}$$

From the equation

$$\mathbb{E}\left[\boldsymbol{X}_t\right] = \widehat{\mathbf{A}}_t \mathbb{E}\left[\boldsymbol{X}_{p,t}\right] + \mathbf{k}_t$$

we can estimate the innovation's mean

$$\widehat{\mathbf{k}}_t = \mathbb{E}\left[\boldsymbol{X}_t\right] - \widehat{A_t} \mathbb{E}\left[\boldsymbol{X}_{p,t}\right]$$

The innovation's variance is estimated by the covariance of the residuals $\boldsymbol{\eta}_t^{(n)} = \boldsymbol{X}_t^{(n)} - \widehat{\mathbf{A}} \boldsymbol{X}_{p,t}^{(n)}$ yielding the form

$$\widehat{\Sigma_t} = \frac{1}{N} \sum_{n=1}^{N} \left(\boldsymbol{X}_t^{(n)} - \widehat{\mathbf{A}} \boldsymbol{X}_{p,t}^{(n)}\right) \left(\boldsymbol{X}_t^{(n)} - \widehat{\mathbf{A}} \boldsymbol{X}_{p,t}^{(n)}\right)^\top$$

A.3.1.2 *ML estimation of VAR model parameters*

For the dataset with N-sized *i.i.d* samples, the likelihood function is the product of the likelihood of each sample

$$\mathcal{L}\left(A_t, \Sigma_t; \boldsymbol{X}_t\right) = \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi |\Sigma_t|}} \exp\left(-\frac{1}{2}\left(\boldsymbol{X}_t^{(n)} - A_t \boldsymbol{X}_{p,t}^{(n)}\right) \Sigma_t^{-1} \left(\boldsymbol{X}_t^{(n)} - A_t \boldsymbol{X}_{p,t}^{(n)}\right)^\top\right)$$

The log likelihood thus takes the form

$$\begin{aligned} l\left(A_t, \Sigma_t; \boldsymbol{X}_t\right) =& \log \mathcal{L}\left(A_t; \Sigma_t; \boldsymbol{X}_t\right) \\ =& -\frac{N}{2}\log(2\pi) - \frac{N}{2}\log|\Sigma_t| - \frac{1}{2}\sum_{n=1}^{N}\left(\boldsymbol{X}_t^{(n)} - A_t \boldsymbol{X}_{p,t}^{(n)}\right) \\ & \boldsymbol{\Sigma}_t{}^{-1}\left(\boldsymbol{X}_t^{(n)} - A_t \boldsymbol{X}_{p,t}^{(n)}\right)^\top \end{aligned}$$

Thus the ML estimation is equivalent to finding the parameter values that makes its derivative to zero

$$\frac{dl\left(A_t, \Sigma_t; \boldsymbol{X}_t\right)}{dA_t} = -2 \sum_{n=1}^{N} \Sigma_t^{-1} \left(\boldsymbol{X}_t^{(n)} - A_t \boldsymbol{X}_{p,t}^{(n)}\right) \boldsymbol{X}_{p,t}^{(n)\top} = 0 \tag{A.3}$$

Reorganizing the terms yields

$$\sum_{n=1}^{N} \left(\boldsymbol{X}_t^{(n)} - \widehat{\mathbf{A}}_t \boldsymbol{X}_{p,t}^{(n)}\right) \boldsymbol{X}_{p,t}^{(n)\top} = 0.$$

Thus the coefficient matrix can be estimated as

$$\widehat{\mathbf{A}}_t = \left(\sum_{n=1}^{N} \boldsymbol{X}_t^{(n)} \boldsymbol{X}_{p,t}^{(n)\top} \right) \left(\sum_{n=1}^{N} \boldsymbol{X}_{p,t}^{(n)} \boldsymbol{X}_{p,t}^{(n)\top} \right)^{-1}$$

This is consistent with the OLS-estimated coefficient matrix in Eq. A.2.

A.3.2 *VAR model order selection*

Two common ways to optimize model order are the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). They both introduce a penalty term in the log-likelihood function to compensate for the effect caused by over-fitting with over-complex models:

$$\mathrm{IC}(p) = -\log(\mathcal{L}(p)) + \mathcal{P}(p)\,. \tag{A.4}$$

The model order is selected as the order that minimize the information criterion. The penalty term $\mathcal{P}(p)$ involves the effect of model complexity by punishing on the number of parameters, which equals pd^2 for a d-dimensional VAR(p) models. For AIC the penalty term is just the number of parameters $\mathcal{P}(p) = pd^2$. BIC takes into account of the effect of sample size $T - p$ as the log-likelihood function also increase with number of sample, and the corresponding penalty term is $\mathcal{P}(p) = \frac{1}{2} pd^2 \log(T - p)$ (for derivation see Appendix A.3.4).

In the rest of this paper, we focus on BIC and let alone AIC as AIC is more likely to result in over-fitting while BIC tends to reconstruct the true model order [123]. Although it was argued that the over-fitting feature of AIC vanishes for VAR(p) processes with the dimensional higher than 3 [86], in the bivariate case mainly dealt with in this paper we prefer to stick to BIC.

A.3.2.1 *Multi-trial selection of VAR model orders*

A critical issue that are often neglected is the model order selection for the multi-trial case. The difficulty is to decide what is the equivalent number of parameters and equivalent sample size in the multi-trial and non-stationary case. Naive ideas that there are $(T - p) \cdot pd^2$ parameters and $N(T - p)$ samples might be misleading.

Previous studies proposed various methods to estimate time-varying VAR model (for an overall review see [31]), together with some model order selection criteria for each specific type of methods. However, they didn't provide a specific penalty term in Eq. A.4 which is applicable for non-stationary and multi-trial cases.

For example, Ding et al. proposed the short windowing method which assumes local stationarity in the process [65], which is also a ML estimator of the coefficients in a time window. Model order in this study was optimized by the classical AIC by assuming $\mathcal{L}(p)$ as the product of the likelihood of each sample in all the trials. However, the penalty term was not revised accordingly.

In the studies using adaptive methods, similar treatment was applied in [96]. Another related paper suggested to perform model order selection individually for each short window, and select the maximum order for the whole process [163]. While this idea might sound reasonable, one is likely to end up with a very large model order, especially with the time-varying model in Eq. (3.3), where coefficients are changing at each time point (i.e. window length is 1 time point). Even when we pick the averaged model order instead of the maximum, it is not a convincing way to evaluate the overall effect of a fixed model order.

Other criteria may also lack generality. Schloegl proposed to use a revised version of mean square error for model order selection combined with a state-space estimation for time-varying models [207], but this may not apply to multi-trial case. [191] and [69] derived in a Bayesian approach the criteria for time-varying VAR models, which is similar to our proposal below. However, they assume that the coefficient matrix is composed of some basis functions, and the criteria depend on the number of basis functions, which differs from the general BIC form in Eq. A.4.

Therefore we propose here an extended version of BIC that is appropriate for non-stationary signals with multi-trial structures: for multi-trial homogeneous VAR(p) models, the penalty term should be $\mathcal{P}(p) = \frac{1}{2}pd^2\log(N(T-p))$; for the multi-trial inhomogeneous case the penalty term should be $\mathcal{P}(p) = \frac{1}{2}Tpd^2\log(N)$ (for proof see Appendix A.3.4).

Note that for single-trial homogeneous case where $N = 1$, the penalty term reduces to $pd^2\log(T-P)$, which is consistent with the classical form of BIC.

A.3.3 *Derivation of Hessian of the likelihood function of the coefficient matrix* A_t

The Hessian matrix can be obtained by calculating the derivative of the first-order derivative of the log-likelihood given in Eq. A.3, which can be rewritten as:

$$\frac{dl}{dA_t} = \sum_{n=1}^{N} \Sigma_t^{-1}\boldsymbol{X}_t^{(n)}\boldsymbol{X}_{p,t}^{(n)\top} - A_t\boldsymbol{X}_{p,t}^{(n)}\boldsymbol{X}_{p,t}^{(n)\top} \tag{A.5}$$

We define a matrix C as the sum of the covariance matrix of the lagged state $\boldsymbol{X}_{p,t}$:

$$C = \sum_{n=1}^{N} \boldsymbol{X}_{p,t}^{(n)}\boldsymbol{X}_{p,t}^{(n)\top} \tag{A.6}$$

Each entry of the matrices A_t, Σ_t^{-1} and C at the i^{th} row and j^{th} column can be denoted as a_{ij}, σ_{ij}, c_ij. Then the second-order derivative of the log-likelihood w.r.t. element a_{ij} and a_{km} is:

$$\frac{d^2 l}{da_{ij}da_{km}} = -c_{mi}\cdot\sigma_{kj} \tag{A.7}$$

The $d \times pd$ dimensional coefficient matrix A_t can be vectorized as a pd^2-dimensional vector by concatenating each column. If the Hessian matrix is organized as $H = \frac{d^2 l}{d\mathit{vec}(A_t)^\top d\mathit{vec}(A_t)}$, then the Hessian matrix can be rewritten as a compact form:

$$H = -C \otimes \Sigma_t^{-1} \tag{A.8}$$

A.3.4 *Derivation of BIC for multi-trial VAR models*

The key to model order selection is to find the optimal model order in the class of VAR models which maximizes the compensated likelihood function after adding a penalty term. In the Bayesian perspective this likelihood probability is $p(\{\boldsymbol{X}_t\}|p)$. This likelihood conditioned on model order is not equal to the one conditioned on the estimated coefficient (i.e. $p(\{\boldsymbol{X}_t\}|\widehat{A}, p)$ as $\widehat{A_t}$ are just one choice of parameters given the p-ordered model with the conditional probability $p(\widehat{A_t}|p)$. In fact, with Laplace approximation the likelihood can be elaborated as:

$$p(\{\boldsymbol{X}_t\}|p) = \int p(\{\boldsymbol{X}_t\}, A_t|p) dA_t \approx p(\{\boldsymbol{X}_t\}|\widehat{A_t}, p) p(\widehat{A_t}|p) (2\pi)^{\frac{pd^2}{2}} |H|^{\frac{-1}{2}} \quad \text{(A.9)}$$

with H as the Hessian matrix of coefficient A_t defined in Eq. A.8.

Taking logarithm of both sides, and we can get the log-likelihood as:

$$\log p(\{\boldsymbol{X}_t\}|p) \approx \log p(\{\boldsymbol{X}_t\}|\widehat{A_t}, p) + \log p(\widehat{A_t}|p) + \frac{pd^2}{2} \log(2\pi) - \frac{1}{2} \log |H| \quad \text{(A.10)}$$

The second and third term are independent of sample size. With sufficiently large sample size, the first term dominates as it grows with sample size. Therefore, comparing $\log p(\{\boldsymbol{X}_t\}|p)$ given different order p would require the knowledge how H changes with sample size.

Traditional BIC based on single-trial stationary VAR(p) processes assumes that H increase linearly with sample size $T - p$, i.e. $H \approx (T - p) H_0$ for a constant H_0. Then the logarithm is:

$$-\frac{1}{2} \log |H| \approx -\frac{1}{2} \log |(T - p) H_0| = -\frac{pd^2}{2} \log(T - p) - \frac{pd^2}{2} \log |H_0| \quad \text{(A.11)}$$

Keeping the first term that depends on sample size, we can obtain the likelihood probability as

$$\log p(\{\boldsymbol{X}_t\}|p) \approx \log p(\{\boldsymbol{X}_t\}|\widehat{A_t}, p) - \frac{1}{2} \log(T - p) \quad \text{(A.12)}$$

The information criteria in Eq. A.4 $IC(p) = -\log p(\{\boldsymbol{X}_t\}|p)$, and $\mathcal{L}(p) = p(\{\boldsymbol{X}_t\}|\widehat{A_t}, p)$. Thus the penalty term of traditional BIC is $\frac{pd^2}{2} \log(T - p)$.

For derivation of BIC in multi-trial cases, let us start again with the simpler homogeneous case. Given the T i.i.d samples from a stationary d-dimension VAR(p) model, the Hessian matrix of A_t takes the form in Eq. A.8 (for derivation see Appendix A.3.3). The determinant of the Hessian matrix is

$$|H| = |C|^d \cdot |\Sigma_t^{-1}|^{pd} \quad \text{(A.13)}$$

Then the logarithm of the determinant of the Hessian matrix is

$$\log |H| = d \log |C| - pd \log |\Sigma_t^{-1}| \quad \text{(A.14)}$$

If N is very large and asymptotically $C = N(T - p) \cdot C_0$ (where $C_0 = E[\boldsymbol{X}_{p,t} \boldsymbol{X}_{p,t}^\top]$, the covariance of p lagged-states for each time point t, is supposed to be a constant matrix with the dimension of pd for a stationary process), then

$$\log |H| = d \log |NtC_0| - pd \log |\Sigma_t^{-1}| = pd^2 \log(N(T - p)) + d \log |C_0| - pd \log |\Sigma| \quad \text{(A.15)}$$

$$\log |H| = pd^2 \log(N(T - p)) + d \log |C_0| - pd \log |\Sigma| \quad \text{(A.16)}$$

The second and the third term can be ignored with large N, So the penalty term should be: $pd^2 \log(N(T - p))$.

Then for a non-stationary case, we can assume that there are $T-p$ inhomogenous VAR models, and each VAR(p) model has N samples. The Hessian matrix of each VAR model is denoted as H_t, and the covariance matrix of each model is C_{t0}. If we assume that each VAR model has the same order, then we can sum up $\log|H_t|$:

$$\sum_{n=1}^{N}\log|H_t| = \sum_{n=1}^{N}(pd^2\log(N) + d\log|C_{t0}| - pd\log|\Sigma_t|) \tag{A.17}$$

$$\sum_{n=1}^{N}\log|H_t| = (T-p)pd^2\log(N) + d\sum_{n=1}^{N}\log|C_{t0}| - pd(\sum_{n=1}^{N}\log|\Sigma_t|) \tag{A.18}$$

Thus similarly we can ignore the second and the third term, and the penalty term is $(T-p)pd^2\log(N)$.

A.4 CORRECTION OF COEFFICIENT ESTIMATION BIAS CAUSED BY SELECTION

The calculation of many time-varying causality measures depends on the accurate estimation of autoregressive coefficient matrices. However, estimation with snapshots detected via thresholding tend to introduce selection bias into the estimated statistics, thus leading to erroneous estimation of causality measures.

If we assume that the peri-event snapshots under study can be modelled as a multi-variate autoregressive process $\mathbf{X}_t$, where the current state is a linear combination of the previous states:

$$\mathbf{X}_t = A_t\mathbf{X}_{p,t} + \boldsymbol{\eta}_t\,, \boldsymbol{\eta}_t \sim \mathcal{N}(\boldsymbol{k}_t, \Sigma_t)\,.$$

Notably, the notations for peri-event snapshots $\mathbf{X}_t$ is different from notations for the general time series $\boldsymbol{X}_t$.

This is a k^{th}-order m-variate vector autoregressive process, with the current state defined as

$$\mathbf{X}_t = \left[X_t^1, X_t^2, \cdots, X_t^m\right]^T$$

and the past state as $\mathbf{X}_{p,t} = [\mathbf{X}_{t-1}, \mathbf{X}_{t-2}, \cdots, \mathbf{X}_{t-p}]^T$. Innovations $\boldsymbol{\eta}_t$ are time-inhomogeneous Gaussian random variables, where $\mathbb{E}[\eta] = \boldsymbol{k}_t, \mathrm{Cov}[\eta] = \Sigma_t$.

In line with the results in Section A.3.1, the estimation of two covariance matrices $\Sigma_{\mathbf{X}_t\mathbf{X}_p}$ and $\Sigma_{\mathbf{X}_p}$ determines the estimation of VAR coefficient matrix as $\widehat{A_t} = \widehat{\Sigma}_{\mathbf{X}_t\mathbf{X}_p}\left(\widehat{\Sigma}_{\mathbf{X}_p}\right)^{-1}$. The innovations mean and variances depends on the estimation of coefficient (see Section A.3.1).

As snapshots are detected in time series $\tilde{\boldsymbol{X}}_t$ using condition $D_{t_0} > d_0$, the covariances we obtain directly from the panel data estimation procedure are estimates of the conditional covariance matrices $\Sigma_{\mathbf{X}_t\mathbf{X}_p|D_{t_0}>d_0}$ and $\Sigma_{\mathbf{X}_p|D_{t_0}>d_0}$, may differ from the real (unconditional) ones.

Therefore our *DeSnap* procedure introduces a new approach to reduce the selection bias covariance matrices as follows. If we represent the snapshot values at peri-event time point t as a lagged state $\mathbf{Y}_t$, by concatenating $\mathbf{X}_t$ and $\mathbf{X}_p$, where $t \in [-T/2, T/2]$:

$$\mathbf{Y}_t = \begin{bmatrix} \mathbf{X}_t \\ \mathbf{X}_{p,t} \end{bmatrix}$$

then second-order statistics of panel data approximate the conditional mean of the snapshots and can be written as :

$$\mu_{\mathbf{Y}_t|D_{t_0}\geqslant d_0} = \begin{bmatrix} \mu_{\mathbf{X}_t|D_{t_0}\geqslant d_0} \\ \mu_{\mathbf{X}p|D_{t_0}\geqslant d_0} \end{bmatrix}, \Sigma_{\mathbf{Y}_t|D_{t_0}\geqslant d_0} = \begin{bmatrix} \sum_{X_t|D_{t_0}\geqslant d_0} & \Sigma_{\mathbf{X}_t\mathbf{X}_p|D_{t_0}\geqslant d_0} \\ \sum_{\mathbf{X}_p\mathbf{X}_t|D_{t_0}\geqslant d_0} & \Sigma_{\mathbf{X}_p|D_{t_0}\geqslant d_0} \end{bmatrix} \tag{A.19}$$

For simplicity, we omit the time indices of D_{t_0} in the notations and refer to the *detection signal*, denoted by D. We now show how to exploit information in the snapshots to estimate the unconditional covariance under a joint Gaussian assumption of $\mathbf{Y}_t$ and D. For each values of $d \in D$ where $d \geqslant d_0$, the conditional distribution of $\mathbf{Y}_t|D = d$ is also Gaussian with mean $\mu_{\mathbf{Y}_t|D=d}$ and variance $\Sigma_{\mathbf{Y}_t|D=d}$, such that:

$$\mu_{\mathbf{Y}_t|D=d} = \mu_{\mathbf{Y}_t} + \Sigma_{\mathbf{Y}_t D}\Sigma_D^{-1}(d-\mu_D) \tag{A.20}$$

$$\Sigma_{\mathbf{Y}_t|D=d} = \Sigma_{\mathbf{Y}_t} - \Sigma_{\mathbf{Y}_t D}\Sigma_D^{-1}\Sigma_{\mathbf{Y}_t D}^T \tag{A.21}$$

The conditional distribution of $\mathbf{Y}_t|D \geqslant d_0$ can then be computed as:

$$P(\mathbf{Y}_t \mid D \geqslant d_0) = \int_{d_0}^{+\infty} \frac{P(D=d)}{P(D \geqslant d_0)} P(\mathbf{Y}_t \mid D = d)\, dd$$

The mean and covariance of this Gaussian mixture is a function of the mean and covariance of each element. For the mean we get

$$\begin{aligned}
&\mu_{\mathbf{Y}_t|D\geqslant d_0} \\
&= \textstyle\int_{d_0}^{+\infty} \frac{P(D=d)}{P(D\geqslant d_0)} \mu_{\mathbf{Y}|D=d}\, dd, \\
&= \textstyle\int_{d_0}^{+\infty} \frac{P(D=d)}{P(D\geqslant d_0)} \left(\mu_{\mathbf{Y}_t} + \Sigma_{\mathbf{Y}_t D}\Sigma_D^{-1}(d-\mu_D)\right) dd, \\
&= \mu_{\mathbf{Y}_t} + \Sigma_{\mathbf{Y}_t D}\Sigma_D^{-1} \textstyle\int_{d_0}^{+\infty} \frac{P(D=d)}{P(D\geqslant d_0)} (d-\mu_D)\, dd, \\
&= \mu_{\mathbf{Y}_t} + \Sigma_{\mathbf{Y}_t D}\Sigma_D^{-1}\left(\overline{d}-\mu_D\right),
\end{aligned}$$

where $\overline{d}$ is the average of $D = d \geqslant d_0$

For the covariance, we use the law of total covariance (for two random variables X and Y)

$$\mathrm{Cov}(X,Y) = \mathbb{E}\left[\mathrm{Cov}(X,Y,D)\right] + \mathrm{Cov}\left(\mathbb{E}\left[X|D\right], \mathbb{E}\left[Y|D\right]\right)$$

to obtain

$$\begin{aligned}
&\Sigma_{\mathbf{Y}_t|D\geqslant d}, \\
&= \textstyle\int_{d_0}^{+\infty} \frac{P(D=d)}{P(D\geqslant d_0)} \left(\Sigma_{\mathbf{Y}_t} - \Sigma_{\mathbf{Y}_t D}\Sigma_D^{-1}\Sigma_{\mathbf{Y}_t D}^T\right) dd, \\
&+ \textstyle\int_{d_0}^{+\infty} \frac{P(D=d)}{P(D\geqslant d_0)} \left(\mu_{\mathbf{Y}|D=d} - \mu_{\mathbf{Y}|D\geqslant d_0}\right)\left(\mu_{\mathbf{Y}|D=d} - \mu_{\mathbf{Y}|D\geqslant d_0}\right)^T dd, \\
&= \Sigma_{\mathbf{Y}_t} + \Sigma_{\mathbf{Y}_t D}\Sigma_D^{-1} c \Sigma_D^{-1}\Sigma_{\mathbf{Y}_t D}^T,
\end{aligned}$$

where $c = \int_{d_0}^{+\infty} \frac{P(D=d)}{P(D \geqslant d_0)} (d - \mu_D)^2 \, dd - \left(\bar{d} - \mu_D\right)^2 - \Sigma_D$.

As a result, we have

$$\mu_{\mathbf{Y}_t|D \geqslant d_0} = \mu_{\mathbf{Y}_t} + \Sigma_{\mathbf{Y}_t D} \Sigma_D^{-1} \left(\bar{d} - \mu_D\right) , \tag{A.22}$$

$$\Sigma_{\mathbf{Y}_t|D \geqslant d} = \Sigma_{\mathbf{Y}_t} + \Sigma_{\mathbf{Y}_t D} \Sigma_D^{-1} c \Sigma_D^{-1} \Sigma_{\mathbf{Y}_t D}^T . \tag{A.23}$$

What can be estimated from peri-event panels in Eq. A.20, A.22 and A.23 are the conditional statistics $\mu_{\mathbf{Y}_t|D \geqslant d_0}$, $\Sigma_{\mathbf{Y}_t|D \geqslant d}$ (which we can estimate from Eq. A.19, and the binned conditions d (which we can specify on our need). What we are interested in recovering, are the unconditional mean $\mu_{\mathbf{Y}_t}$ and covariance matrix $\Sigma_{\mathbf{Y}_t}$. Some intermediate unknown variables that help us estimated the unconditional statistics are $\Sigma_{\mathbf{Y}_t D} \Sigma_D^{-1}$, μ_D and c. For a uni-state signals, μ_D and c can be easily obtained by exploiting the distribution of D; however, if the signal is a mixture of multiple states, these statistics are largely unobserved. Actually, these intermediate variables and the unconditional statistics can all be retrieved by performing three linear regressions. First, with the snapshot and a given set of binned d (which must satisfy $d \geqslant d_0$ but should not be too large to limit the sample size of $P(\mathbf{Y}_t|D = d)$), we can regress d over $\mu_{\mathbf{Y}_t|D=d}$ in Eq. A.22 to get the coefficient a_t and the intercept b_t corresponding to:

$$p_t = \Sigma_{\mathbf{Y}_t D} \Sigma_D^{-1} , \tag{A.24}$$

$$q_t = \mu_{\mathbf{Y}_t} - \Sigma_{\mathbf{Y}_t D} \Sigma_D^{-1} \mu_D . \tag{A.25}$$

Secondly, b_t is a linear function of a_t as $q_t = \mu_{\mathbf{Y}_t} - p_t \mu_D$. Thus we can regress p_t over q_t to estimate the mean of D (μ_D) as the coefficient and $\mu_{\mathbf{Y}_t}$ as the intercept.

Finally, Eq. A.23 can be reorganized as:

$$\Sigma_{\mathbf{Y}_t|D \geqslant d} = \Sigma_{\mathbf{Y}_t} + c p_t p_t^T , \tag{A.26}$$

For a given threshold d_0, $c(d_0)$ is a constant for all elements of the covariance matrix at all time points of the snapshots. Regressing $p_t p_t^T$ over $\Sigma_{\mathbf{Y}_t|D \geqslant d}$ for any single element across time, we can estimate $c(d_0)$, by which we are able to retrieve $\Sigma_{\mathbf{Y}_t}$ from Eq. 3.32. Sometimes, as event extraction induces temporal correlations, we can also apply a first order difference in the panel such that $\Delta_t(\Sigma_{\mathbf{Y}_t|D \geqslant d}) = \Delta_t(\Sigma_{\mathbf{Y}_t}) + c\Delta_t(p_t a_t^T) = c\Delta_t(p_t p_t^T)$. Then, regressing $\Delta_t(p_t p_t^T)$ over $\Delta_t(\Sigma_{\mathbf{Y}_t|D \geqslant d})$, we can similarly calculate c and retrieve $\Sigma_{\mathbf{Y}_t}$ from Eq. A.26.

A.5 DERIVATIONS FOR TIME-VARYING CAUSALITY MEASURES

A.5.1 *Derivation of KL-divergence between two uni-variate Gaussians*

Time-varying TE, DCS and rDCS are formulated as the KL divergence between the corresponding actual and counterfactual conditions. Thus we first present the KL divergence between two uni-variate Gaussian variables.

Consistent with Section 4.2.3, we denote the gaussian for the actual condition as $p(x) = \mathcal{N}(\mu_a, \sigma_a^2)$, and the counterfactual gaussian as $q(x) = \mathcal{N}(\mu_c, \sigma_c^2)$. Then the KL divergence between $p(x)$ and $q(x)$ is

$$D_{KL}(p\|q) = -\int p(x)\log q(x)dx + \int p(x)\log p(x)dx$$

$$= \int [\log(p(x)) - \log(q(x))]\, p(x)dx$$

$$= \int \left[-\frac{1}{2}\log(2\pi) - \log(\sigma_a) - \frac{1}{2}\left(\frac{x-\mu_a}{\sigma_a}\right)^2 + \frac{1}{2}\log(2\pi) + \log(\sigma_c) + \frac{1}{2}\left(\frac{x-\mu_c}{\sigma_c}\right)^2\right] \times$$

$$\frac{1}{\sqrt{2\pi}\sigma_a}\exp\left[-\frac{1}{2}\left(\frac{x-\mu_a}{\sigma_a}\right)^2\right]dx$$

$$= \int \left\{\log\frac{\sigma_c}{\sigma_a} + \frac{1}{2}\left[\left(\frac{x-\mu_c}{\sigma_c}\right)^2 - \left(\frac{x-\mu_a}{\sigma_a}\right)^2\right]\right\} \times \frac{1}{\sqrt{2\pi}\sigma_a}\exp\left[-\frac{1}{2}\left(\frac{x-\mu_a}{\sigma_a}\right)^2\right]dx$$

$$= \mathbb{E}_p\left[\log\frac{\sigma_c}{\sigma_a} + \frac{1}{2}\left[\left(\frac{x-\mu_c}{\sigma_c}\right)^2 - \left(\frac{x-\mu_a}{\sigma_a}\right)^2\right]\right]$$

$$= \log\frac{\sigma_c}{\sigma_a} + \frac{1}{2\sigma_c^2}\mathbb{E}_p\left[(X-\mu_c)^2\right] - \frac{1}{2\sigma_a^2}\mathbb{E}_p\left[(X-\mu_a)^2\right]$$

$$= \log\frac{\sigma_c}{\sigma_a} + \frac{1}{2\sigma_c^2}\mathbb{E}_p\left[(X-\mu_c)^2\right] - \frac{1}{2}$$

Note that

$$(X-\mu_c)^2 = (X-\mu_a+\mu_a-\mu_c)^2$$

$$= (X-\mu_a)^2 + 2(X-\mu_a)(\mu_a-\mu_c) + (\mu_a-\mu_c)^2$$

Therefore,

$$\begin{aligned} D_{KL}(\mathcal{N}(\mu_a,\, \sigma_a^2)\|\mathcal{N}(\mu_c,\, \sigma_c^2)) &= \log\frac{\sigma_c}{\sigma_a} + \frac{1}{2\sigma_c^2}\mathbb{E}_p\left[(X-\mu_a)^2\right] \\ &\quad + 2(\mu_a-\mu_c)\mathbb{E}_p\left[X-\mu_a\right] + (\mu_a-\mu_c)^2 - \frac{1}{2} \\ &= \frac{1}{2}\log\frac{\sigma_c^2}{\sigma_a^2} + \frac{\sigma_a^2 + (\mu_a-\mu_c)^2}{2\sigma_c^2} - \frac{1}{2} \end{aligned} \tag{A.27}$$

A.5.2 *Conditional mean and variance for the actual condition*

The actual condition defined for TE, DCS and rDCS is that the current state of X_t^1 is dependent on the past of both X^1 and X^2, denoted as

$$\mathcal{N}(\mu_a, \sigma_a^2) = p(X_t^1 | \boldsymbol{X}_{p,t}^1, \boldsymbol{X}_{p,t}^2)\,.$$

The dynamics of X_t^1 is described by the structural equation

$$X_t^1 = \mathbf{a}_t^\top \boldsymbol{X}_{p,t}^1 + \mathbf{b}_t^\top \boldsymbol{X}_{p,t}^2 + \eta_t^1\,, \quad \eta_t^1 \sim \mathcal{N}(k_t^1,\, \sigma_{1,t})\,. \tag{A.28}$$

Therefore, the time-varying conditional mean and variance can be derived as

$$\mu_a = \mathbb{E}[\mathbf{a}_t^\top \boldsymbol{X}_{p,t}^1 + \mathbf{b}_t^\top \boldsymbol{X}_{p,t}^2 + \eta_t^1 | \boldsymbol{X}_{p,t}^1, \boldsymbol{X}_{p,t}^2] = \mathbf{a}_t^\top \boldsymbol{X}_{p,t}^1 + \mathbf{b}_t^\top \boldsymbol{X}_{p,t}^2 + k_t^1$$

$$\sigma_a^2 = \mathrm{Var}[\mathbf{a}^\top \boldsymbol{X}_{p,t}^1 + \mathbf{b}^\top \boldsymbol{X}_{p,t}^2 + \eta_t^1 | \boldsymbol{X}_{p,t}^1, \boldsymbol{X}_{p,t}^2] = \mathrm{Var}[\eta_t^1 | \boldsymbol{X}_{p,t}^1, \boldsymbol{X}_{p,t}^2] = \sigma_{1,t}$$

A.5.3 *Transfer Entropy*

For TE, as the conditional probability representing the "counterfactual" condition takes the form

$$\mathcal{N}(\mu_c, \sigma_c^2) = p(X_t^1 | \boldsymbol{X}_{p,t}^1) \,.$$

Recalling Section 4.2.2.2, TE does not rely on counterfactuals but here the "counterfactual" is just used for symmetry reasons

Resulting from the same model in Eq. A.28, the mean and variance can be derived as

$$\mu_c = \mathbb{E}[\mathbf{a}_t^\top \boldsymbol{X}_{p,t}^1 + \mathbf{b}_t^\top \boldsymbol{X}_{p,t}^2 + \eta_t^1 | \boldsymbol{X}_{p,t}^1] = \mathbf{a}_t^\top \boldsymbol{X}_{p,t}^1 | \boldsymbol{X}_{p,t}^1 + \mathbf{b}_t^\top \mathbb{E}[\boldsymbol{X}_{p,t}^2 | \boldsymbol{X}_{p,t}^1] + k_t^1$$

$$\sigma_c^2 = \mathrm{Var}[\mathbf{a}^\top \boldsymbol{X}_{p,t}^1 + \mathbf{b}^\top \boldsymbol{X}_{p,t}^2 + \eta_t^1 | \boldsymbol{X}_{p,t}^1] = \mathbf{b}_t^\top \mathrm{Cov}[\boldsymbol{X}_{p,t}^2 | \boldsymbol{X}_{p,t}^1] \mathbf{b}_t + \sigma_{1,t}$$

while

$$\mathbb{E}_{(\boldsymbol{X}_{p,t}^1, \boldsymbol{X}_{p,t}^2)}[(\mu_a - \mu_c)^2] = \mathbb{E}_{(\boldsymbol{X}_{p,t}^1, \boldsymbol{X}_{p,t}^2)}[(\mathbf{b}_t^\top (\boldsymbol{X}_{p,t}^2 - \mathbb{E}[\boldsymbol{X}_{p,t}^2 | \boldsymbol{X}_{p,t}^1]))^2]$$
$$= \mathbf{b}_t^\top \mathbb{E}[(\boldsymbol{X}_{p,t}^2 - \mathbb{E}|\boldsymbol{X}_{p,t}^2])(\boldsymbol{X}_{p,t}^2 - \mathbb{E}|\boldsymbol{X}_{p,t}^2 | \boldsymbol{X}_{p,t}^1])^\top] \mathbf{b}_t = \mathbf{b}_t^\top \mathrm{Cov}[\boldsymbol{X}_{p,t}^2 | \boldsymbol{X}_{p,t}^1]] \mathbf{b}_t$$

Plugging the expressions of μ_a, μ_c, σ_a^2 and σ_c^2 into Eq. A.27, the KL divergence can be derived as

$$\mathrm{TE} = D_{KL}(\mathcal{N}(\mu_a,\ \sigma_a^2) || \mathcal{N}(\mu_c,\ \sigma_c^2)) = \frac{1}{2} \log \frac{\mathbf{b}_t^\top \mathrm{Cov}[\boldsymbol{X}_{p,t}^2 | \boldsymbol{X}_{p,t}^1] \mathbf{b}_t + \sigma_{1,t}}{\sigma_{1,t}}$$

As $\boldsymbol{X}_{p,t}^1$ and $\boldsymbol{X}_{p,t}^2$ are jointly Gaussian, the conditional variance takes the form

$$\mathrm{Cov}[\boldsymbol{X}_{p,t}^2 | \boldsymbol{X}_{p,t}^1] = \Sigma_{\boldsymbol{X}_p^2} - \Sigma_{\boldsymbol{X}_p^1 \boldsymbol{X}_p^2} \Sigma_{\boldsymbol{X}_p^1}^{-1} \Sigma_{\boldsymbol{X}_p^2 \boldsymbol{X}_p^1} \,.$$

Therefore, the expression of time-varying TE should be

$$\mathrm{TE} = \frac{1}{2} \log \frac{\sigma_{1,t} + \mathbf{b}_t^\top \Sigma_{\boldsymbol{X}_p^2} \mathbf{b}_t - \mathbf{b}_t^\top \Sigma_{\boldsymbol{X}_p^1 \boldsymbol{X}_p^2} \Sigma_{\boldsymbol{X}_p^1}^{-1} \Sigma_{\boldsymbol{X}_p^2 \boldsymbol{X}_p^1} \mathbf{b}_t}{\sigma_{1,t}}$$

A.5.4 *Dynamic Causal Strength*

For DCS, the counterfactual condition is

$$\mathcal{N}(\mu_c, \sigma_c^2) = p^{do(X_t^1 := f(\boldsymbol{X}_{p,t}^1, \boldsymbol{X}_{p,t}^{2}{}', \eta_t^1))}(X_t^1 | \boldsymbol{X}_{p,t}^1, \boldsymbol{X}_{p,t}^2)$$

with $X_t^1 = \mathbf{a}_t^\top \boldsymbol{X}_{p,t}^1 + \mathbf{b}_t^\top \boldsymbol{X}_{p,t}^{2}{}' + \eta_t^1$, and $\boldsymbol{X}_{p,t}^{2}{}'$ is a random sample drawn from the distribution $p(\boldsymbol{X}_{p,t}^2)$.

Then the conditional mean is:

$$\mu_c = \mathbb{E}[\mathbf{a}_t^\top \boldsymbol{X}_{p,t}^1 + \mathbf{b}_t^\top \boldsymbol{X}_{p,t}^{2}{}' + \eta_t^1 | \boldsymbol{X}_{p,t}^1, \boldsymbol{X}_{p,t}^2] = \mathbf{a}_t^\top \boldsymbol{X}_{p,t}^1 + \mathbf{b}_t^\top \mathbb{E}[\boldsymbol{X}_{p,t}^2] + \eta_t^1$$

The conditional variance is:

$$\sigma_a^2 = \mathrm{Var}[\mathbf{a}^\top \boldsymbol{X}_{p,t}^1 + \mathbf{b}^\top \boldsymbol{X}_{p,t}^2 + \eta_t^1 | \boldsymbol{X}_{p,t}^1, \boldsymbol{X}_{p,t}^2]$$
$$= \mathrm{Var}[\mathbf{b}^\top \boldsymbol{X}_{p,t}^{2}{}' | \boldsymbol{X}_{p,t}^2] + \mathrm{Var}[\eta_t^1] = \mathbf{b}^\top \mathrm{Cov}[\boldsymbol{X}_{p,t}^2] \mathbf{b} + \sigma_{1,t}$$

with $\mathrm{Cov}[\boldsymbol{X}^2_{p,t}] = \mathbb{E}[(\boldsymbol{X}^2_{p,t} - \mathbb{E}|\boldsymbol{X}^2_{p,t}])(\boldsymbol{X}^2_{p,t} - \mathbb{E}|\boldsymbol{X}^2_{p,t}])^\top]$.

DCS defined as the KL divergence between the actual and counterfactual conditions:

$$\mathrm{DCS}(X^2 \to X^1) = \mathbb{E}_{(\boldsymbol{X}^1_{p,t}, \boldsymbol{X}^2_{p,t})}\left[\frac{1}{2}\log\frac{\sigma_c^2}{\sigma_a^2} - \frac{1}{2} + \frac{1}{2}\cdot\frac{\sigma_a^2 + (\mu_a - \mu_c)^2}{\sigma_c^2}\right]$$

$$\mathrm{DCS}(X^2 \to X^1) = \frac{1}{2}\log\frac{\sigma_c^2}{\sigma_a^2} - \frac{1}{2} + \frac{1}{2}\cdot\frac{\sigma_a^2 + \mathbb{E}_{(\boldsymbol{X}^1_{p,t}, \boldsymbol{X}^2_{p,t})}[(\mu_a - \mu_c)^2]}{\sigma_c^2}$$

while

$$\begin{aligned}\mathbb{E}_{(\boldsymbol{X}^1_{p,t}, \boldsymbol{X}^2_{p,t})}[(\mu_a - \mu_c)^2] &= \mathbb{E}_{(\boldsymbol{X}^1_{p,t}, \boldsymbol{X}^2_{p,t})}[(\mathbf{b}_t^\top(\boldsymbol{X}^2_{p,t} - \mathbb{E}[\boldsymbol{X}^2_{p,t}]))^2] \\ &= \mathbf{b}_t^\top \mathbb{E}[(\boldsymbol{X}^2_{p,t} - \mathbb{E}|\boldsymbol{X}^2_{p,t}])(\boldsymbol{X}^2_{p,t} - \mathbb{E}|\boldsymbol{X}^2_{p,t}])^\top]\mathbf{b}_t = \mathbf{b}_t^\top \mathrm{Cov}[\boldsymbol{X}^2_{p,t}]\mathbf{b}_t\end{aligned}$$

Therefore the expression of DCS can be obtained by plugging-in the means and variances of the two Gaussian distributions,

$$\begin{aligned}\mathrm{DCS}(X^2 \to X^1) &= \frac{1}{2}\log\frac{\mathbf{b}_t^\top \mathrm{Cov}[\boldsymbol{X}^2_{p,t}]\mathbf{b}_t + \sigma_{1,t}}{\sigma_{1,t}} - \frac{1}{2} \\ &\quad + \frac{1}{2}\cdot\frac{\sigma_{1,t} + \mathbf{b}_t^\top \mathrm{Cov}[\boldsymbol{X}^2_{p,t}]\mathbf{b}_t}{\mathbf{b}_t^\top \mathrm{Cov}[\boldsymbol{X}^2_{p,t}]\mathbf{b}_t + \sigma_{1,t}} = \frac{1}{2}\log\frac{\mathbf{b}_t^\top \mathrm{Cov}[\boldsymbol{X}^2_{p,t}]\mathbf{b}_t + \sigma_{1,t}}{\sigma_{1,t}}\end{aligned}$$

A.5.5 *Relative Dynamic Causal Strength*

The relative Causal Strength is 'relative' in the sense that the intervention involves an independent copy of a baseline state $\boldsymbol{X}^2_{p,t_{ref}}$, instead of the lagged state $\boldsymbol{X}^2_{p,t}{}'$, such that $X^{1\prime}_t = \mathbf{a}_t^\top \boldsymbol{X}^1_{p,t} + \mathbf{b}_t^\top \boldsymbol{X}^2_{p,t_{ref}} + \eta^1_t$ for the counterfactual condition.

$$\mathcal{N}(\mu_c, \sigma_c^2) = p^{do(X^1_t := f(\boldsymbol{X}^1_{p,t}, \boldsymbol{X}^2_{p,t_{ref}}, \eta^1_t))}(X^1_t | \boldsymbol{X}^1_{p,t}, \boldsymbol{X}^2_{p,t})$$

The conditional mean and variance for the counterfactual condition should be revised as:

$$\mu_c = \mathbb{E}[\mathbf{a}_t^\top \boldsymbol{X}^1_{p,t} + \mathbf{b}_t^\top \boldsymbol{X}^2_{p,t_{ref}} + \eta^1_t | \boldsymbol{X}^1_{p,t}, \boldsymbol{X}^2_{p,t}] = \mathbf{a}_t^\top \boldsymbol{X}^1_{p,t} + \mathbf{b}_t^\top \mathbb{E}[\boldsymbol{X}^2_{p,t_{ref}}] + \eta^1_t$$

$$\begin{aligned}\sigma_c^2 &= \mathrm{Var}[\mathbf{a}_t^\top \boldsymbol{X}^1_{p,t} + \mathbf{b}_t^\top \boldsymbol{X}^2_{p,t_{ref}} + \eta^1_t | \boldsymbol{X}^1_{p,t}, \boldsymbol{X}^2_{p,t}] \\ &= \mathrm{Var}[\mathbf{a}_t^\top \boldsymbol{X}^2_{p,t_{ref}} | \boldsymbol{X}^2_{p,t}] + \mathrm{Var}[\eta^1_t] = \mathbf{b}_t^\top \mathrm{Cov}[\boldsymbol{X}^2_{p,t_{ref}}]\mathbf{b}_t + \sigma_{1,t}\end{aligned}$$

with $\mathrm{Cov}[\boldsymbol{X}^2_{p,t_{ref}}] = \mathbb{E}[(\boldsymbol{X}^2_{p,t_{ref}} - \mathbb{E}|\boldsymbol{X}^2_{p,t_{ref}}])(\boldsymbol{X}^2_{p,t_{ref}} - \mathbb{E}|\boldsymbol{X}^2_{p,t_{ref}}])^\top]$.

$$\begin{aligned}\mathbb{E}_{(\boldsymbol{X}^1_{p,t}, \boldsymbol{X}^2_{p,t})}[(\mu_a - \mu_c)^2] &= \mathbb{E}_{(\boldsymbol{X}^1_{p,t}, \boldsymbol{X}^2_{p,t})}[(\mathbf{b}_t^\top(\boldsymbol{X}^2_{p,t} - \mathbb{E}[\boldsymbol{X}^2_{p,t_{ref}}]))^2] \\ &= \mathbf{b}_t^\top \mathbb{E}[(\boldsymbol{X}^2_{p,t} - \mathbb{E}|\boldsymbol{X}^2_{p,t_{ref}}])(\boldsymbol{X}^2_{p,t} - \mathbb{E}|\boldsymbol{X}^2_{p,t_{ref}}])^\top]\mathbf{b}_t\end{aligned}$$

Therefore, the non-zero-mean Causal Strength with reference states is

$$\begin{aligned}\mathrm{rDCS}(X^2 \to X^1) &= \frac{1}{2}\log\frac{\sigma_{1,t} + \mathbf{b}_t^\top \mathrm{Cov}[\boldsymbol{X}^2_{p,t_{ref}}]\mathbf{b}_t}{\sigma_{1,t}} - \frac{1}{2} \\ &\quad + \frac{1}{2}\cdot\frac{\sigma_{1,t} + \mathbf{b}_t^\top \mathbb{E}[(\boldsymbol{X}^2_{p,t} - \mathbb{E}[\boldsymbol{X}^2_{p,t_{ref}}])(\boldsymbol{X}^2_{p,t} - \mathbb{E}[\boldsymbol{X}^2_{p,t_{ref}}])^\top]\mathbf{b}_t}{\sigma_{1,t} + \mathbf{b}_t^\top \mathrm{Cov}[\boldsymbol{X}^2_{p,t_{ref}}]\mathbf{b}_t}\end{aligned}$$

Notably, when rDCS is applied to the *DeSnap*-corrected model (Section 3.2.6.2) where we don't have explicit access of the *i.i.d.* samples for $\boldsymbol{X}^2_{\mathrm{p,t}}$, the term $\mathbb{E}[(\boldsymbol{X}^2_{\mathrm{p,t}} - \mathbb{E}[\boldsymbol{X}^2_{\mathrm{p,t}_{ref}}])(\boldsymbol{X}^2_{\mathrm{p,t}} - \mathbb{E}[\boldsymbol{X}^2_{\mathrm{p,t}_{ref}}])^\top]$ should be expanded as:

$$\begin{aligned}
&\mathbb{E}[(\boldsymbol{X}^2_{\mathrm{p,t}} - \mathbb{E}[\boldsymbol{X}^2_{\mathrm{p,t}_{ref}}])(\boldsymbol{X}^2_{\mathrm{p,t}} - \mathbb{E}[\boldsymbol{X}^2_{\mathrm{p,t}_{ref}}])^\top] \\
&= \mathbb{E}[\boldsymbol{X}^2_{\mathrm{p,t}}{\boldsymbol{X}^2_{\mathrm{p,t}}}^\top] - \mathbb{E}[\boldsymbol{X}^2_{\mathrm{p,t}}\mathbb{E}[\boldsymbol{X}^2_{\mathrm{p,t}_{ref}}]^\top] - \mathbb{E}[\mathbb{E}[\boldsymbol{X}^2_{\mathrm{p,t}_{ref}}]^\top {\boldsymbol{X}^2_{\mathrm{p,t}}}^\top] + \mathbb{E}[\mathbb{E}[\boldsymbol{X}^2_{\mathrm{p,t}_{ref}}]\mathbb{E}[\boldsymbol{X}^2_{\mathrm{p,t}_{ref}}]^\top] \\
&\qquad = \mathrm{Cov}(\boldsymbol{X}^2_{\mathrm{p,t}}) + \mathbb{E}[\boldsymbol{X}^2_{\mathrm{p,t}}]\mathbb{E}[\boldsymbol{X}^2_{\mathrm{p,t}}]^\top - \mathbb{E}[\boldsymbol{X}^2_{\mathrm{p,t}}]\mathbb{E}[\boldsymbol{X}^2_{\mathrm{p,t}_{ref}}]^\top \\
&\qquad\qquad - \mathbb{E}[\boldsymbol{X}^2_{\mathrm{p,t}_{ref}}]\mathbb{E}[\boldsymbol{X}^2_{\mathrm{p,t}}]^\top + \mathbb{E}[\boldsymbol{X}^2_{\mathrm{p,t}_{ref}}]\mathbb{E}[\boldsymbol{X}^2_{\mathrm{p,t}_{ref}}]^\top \qquad \text{(A.29)}
\end{aligned}$$

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/`