ABSTRACT

Title of dissertation: EXTRACTING NEURONAL DYNAMICS AT HIGH SPATIOTEMPORAL RESOLUTIONS: THEORY, ALGORITHMS, AND APPLICATION

Alireza Sheikhattar, Doctor of Philosophy, 2018

Dissertation directed by: Professor Behtash Babadi
Department of Electrical & Computer Engineering

Analyses of neuronal activity have revealed that various types of neurons, both at the single-unit and population level, undergo rapid dynamic changes in their response characteristics and their connectivity patterns in order to adapt to variations in the behavioral context or stimulus condition. In addition, these dynamics often admit parsimonious representations. Despite growing advances in neural modeling and data acquisition technology, a unified signal processing framework capable of capturing the adaptivity, sparsity and statistical characteristics of neural dynamics is lacking. The objective of this dissertation is to develop such a signal processing methodology in order to gain a deeper insight into the dynamics of neuronal ensembles underlying behavior, and consequently a better understanding of how brain functions.

The first part of this dissertation concerns the dynamics of stimulus-driven neuronal activity at the single-unit level. We develop a sparse adaptive filtering framework for the identification of neuronal response characteristics from spiking

activity. We present a rigorous theoretical analysis of our proposed sparse adaptive filtering algorithms and characterize their performance guarantees. Application of our algorithms to experimental data provides new insights into the dynamics of attention-driven neuronal receptive field plasticity, with a substantial increase in temporal resolution.

In the second part, we focus on the network-level properties of neuronal dynamics, with the goal of identifying the causal interactions within neuronal ensembles that underlie behavior. Building up on the results of the first part, we introduce a new measure of causality, namely the Adaptive Granger Causality (AGC), which allows capturing the sparsity and dynamics of the causal influences in a neuronal network in a statistically robust and computationally efficient fashion. We develop a precise statistical inference framework for the estimation of AGC from simultaneous recordings of the activity of neurons in an ensemble.

Finally, in the third part we demonstrate the utility of our proposed methodologies through application to synthetic and real data. We first validate our theoretical results using comprehensive simulations, and assess the performance of the proposed methods in terms of estimation accuracy and tracking capability. These results confirm that our algorithms provide significant gains in comparison to existing techniques. Furthermore, we apply our methodology to various experimentally recorded data from electrophysiology and optical imaging: 1) Application of our methods to simultaneous spike recordings from the ferret auditory and prefrontal cortical areas reveals the dynamics of top-down and bottom-up functional interactions underlying attentive behavior at unprecedented spatiotemporal resolutions; 2)

Our analyses of two-photon imaging data from the mouse auditory cortex shed light on the sparse dynamics of functional networks under both spontaneous activity and auditory tone detection tasks; and 3) Application of our methods to whole-brain light-sheet imaging data from larval zebrafish reveals unique insights into the organization of functional networks involved in visuo-motor processing.

EXTRACTING NEURONAL DYNAMICS AT HIGH SPATIOTEMPORAL RESOLUTIONS: THEORY, ALGORITHMS, AND APPLICATION

by

Alireza Sheikhattar

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2018

Advisory Committee:
Professor Behtash Babadi, Chair/Advisor
Professor Shihab Shamma
Professor Jonathan Z. Simon
Dr. Jonathan Fritz
Dr. Misha Ahrens
Professor Patrick Kanold, Dean's Representative

Acknowledgments

First and foremost, I would like to take this opportunity to express my profound gratitude to my advisor, Professor Behtash Babadi, for his valuable guidance and friendly support throughout my PhD period. This dissertation would not have been possible without his constant encouragement and untiring supervision. Thanks for enlightening my mindset toward research, through which I learned that the sky is the limit to the knowledge one can gain, and the contribution one can make to science. Thanks for the countless hours of brainstorming and scientific discussions we had through these years, through which I have broadened my horizons in science, and developed a researcher mindset. It has been, and will be an honor to have the opportunity to work under his supervision, as one of his first PhD students. As a supervisor, he has always inspired me to acquire a never-give-up mindset, and encouraged me to push the boundaries.

I would like to thank Prof. Shihab Shamma, Prof. Jonathan Simon, Prof. Patrick Kanold, Dr. Jonathan Fritz, and Dr. Misha Ahrens, who kindly agreed to serve on my dissertation committee, and generously shared their invaluable time and expertise to enrich this dissertation.

I would also like to thank my colleagues, Majid Mirbagheri, Sahar Akram, Sina Miran, Abbas Kazemipour, and Proloy Das for their support, friendship and the countless hours of brainstorming, exchanging ideas and insightful discussions. Moreover, I would like to express my gratitude to all my collaborators, Nikolas Francis, Diego Elgueda, Ji Liu, Yu Mu, Mikail Rubinov, Maarten Zwart, and Jing

Lim, for sharing their broad knowledge, extensive expertise and rich data recordings with me.

I would like to thank all my wonderful friends who made this journey a memorable and pleasant one, Mehregan, Alborz, Sahar, Ladan, Mahshid, Ali, Sara, Arian, Niloofar, Farhad, Pooya, Sina, and all other dear friends.

Lastly and most importantly, this dissertation is for sure dedicated to my family - my parents, Maryam and Ali Asghar, and my sisters, Leila, Zahra and Mahsa, for their lifetime love and endless support. They have always believed in me, and encouraged me to follow my dreams since I was a child. Although we have been far apart all these years, they always stood by me in the worst of times, and supported me in all manners. They have lost a lot due to my studies abroad.

August 2018
Alireza Sheikhattar

Table of Contents

List of Figures

List of Abbreviations

2P	Two-Photon
A1	Primary Auditory Cortex
ACF	Autocovariance Function
AGC	Adaptive Granger Causality
AR	Autoregressive
ARMA	Autoregressive Moving Average
AUC	Area Under Curve
BY	Benjamini-Yekutieli
CDF	Cumulative Distribution Function
CIF	Conditional Intensity Function
CoSaMP	Compressed Sampling Matching Pursuit
CS	Compressed Sensing
EEG	Electroencephalography
EM	Expectation Maximization
FAR	False Alarm Rate
FDR	False Discovery Rate
FOV	Field of View
FWER	Family-wise Error Rate
GABA	Gamma-Aminobutyric Acid
GC	Granger Causality
GLM	Generalized Linear Model
KS	Kolmogorov-Smirnov
LFP	Local Field Potential
LMS	Least Mean Squares
ML	Maximum Likelihood
MSE	Mean Squared Error
MVAR	Multi-variate Autoregressive
NRC	Normalized Reverse Correlation
PFC	Prefrontal Cortex
PSD	Power Spectral Density
RLPPF	Regularized Likelihood Point Process Filter
RLS	Recursive Least Squares
ROC	Receiver Operating Characteristic
SCAD	Smoothly Clipped Absolute Deviation
SDPPF	Steepest Descent Point Process Filter
SGPPF	Sparse Greedy Point Process Filter
SNR	Signal-to-Noise Ratio
SPARLS	Sparse RLS
SPM	Sparsity Metric
SSPPF	Stochastic State Point Process Filter
STRF	Spectrotemporal Receptive Field
TDR	True Detection Rate
TORC	Temporally Orthogonal Ripple Combinations
ℓ_1-PPF	ℓ_1-regularized Point Process Filter

Chapter 1: Introduction

The brain is arguably the most complex dynamical system in nature, consisting of billions of interconnected neurons. It continuously processes internal and external inputs from various sources in real time, and integrates neural information from multiple streams through its many circuits in order to generate and control behavior. Characterizing the spatiotemporal dynamics of the neurons, as the core computational units of this sophisticated organic system, is crucial to deciphering the many mysteries of brain function.

Analyses of neuronal activity recorded from various types of neurons have revealed three main features: first, neuronal activity is remarkably stochastic in nature and exhibits significant variability over time and across trials; second, many neurons often undergo rapid changes in their response characteristics referred to as neuronal plasticity, in order to adapt to changing stimulus salience or behavioral context; and third, the neuronal dynamics often admit parsimonious descriptions. Examples include place cells in the hippocampus [1] and spectrotemporally tuned neurons in the primary auditory cortex [2] with sparse tuning characteristics.

This dissertation aims at developing a statistically robust and computationally efficient signal processing methodology in order to gain a deeper insight into the

dynamics of neuronal ensembles underlying behavior, and consequently a better understanding of the brain function. At a high level, this dissertation comprises two major parts. The first part concerns the dynamics of stimulus-driven neuronal activity at the single-unit level (Chapter 3), whereas in the second part, we inspect the network-level aspects of neuronal dynamics by probing the neuronal functional network dynamics underlying behavior (Chapters 4-6).

Part 1: Adaptive Sparse Identification of Neuronal Dynamics

The theory of point processes [3] has been recently adopted as a mathematical framework to model the stochasticity of neuronal data. Traditionally, these models have been used to predict the likelihood of self-exciting processes such as earthquake occurrences [4, 5], but have recently found significant applications in the analysis of neuronal data [6–12]. On the other hand, classic results in signal processing such as the Least Mean Squares (LMS) and Recursive Least Squares (RLS) algorithms [13] have created a framework to efficiently capture the dynamics of the parameters in linear observation models. Existing solutions in computational neuroscience have adopted this framework to estimate the dynamics of neuronal activity. For instance, in [7] an LMS-type point process filter was introduced to study plasticity in hippocampal neurons. In [14], more general adaptive filtering solutions based on approximations to the Chapman-Kolmogorov equation were introduced. Although quite powerful in analyzing neuronal data, these solutions do not account for the sparsity of the underlying parameters.

Finally, the theory of compressed sensing (CS) has provided a novel methodology for measuring and estimating statistical models governed by sparse underlying parameters [15–18]. In particular, for *static* linear and generalized linear models (GLM) with random covariates and sparsity of the parameters, the CS theory characterizes sharp trade-offs between the number of measurement, sparsity, and estimation accuracy [19]. The sparse solutions of CS are typically achieved using batch-mode convex programs and greedy techniques (See [20] for a review of these techniques). In online settings, sparse adaptive filters have only been introduced in the context of linear systems governed by sparse parameters such as communication channels [21–23]. Despite significant progress in all these research fronts, a cohesive analytical framework to simultaneously capture the dynamicity, sparsity and stochastic nature of neuronal dynamics had been lacking, and served as the motivation for the first part of this dissertation.

In Chapter 3, we indeed close this gap by integrating techniques from point process theory, adaptive filtering, and compressed sensing, and develop novel online methods for sparse neuronal system identification. To this end, we consider the problem of estimating time-varying stimulus modulation coefficients (e.g., receptive fields) from a sequence of binary observations in an online fashion. We model the spiking activity by a conditional Bernoulli point process, where the conditional intensity is a logistic function of the stimulus and its time lags. We then design a novel objective function by incorporating the forgetting factor mechanism of RLS-type algorithms into the ℓ_1-regularized maximum likelihood estimation of the point process parameters. We present theoretical guarantees that extend those of CS the-

ory and characterize fundamental trade-offs between the number of measurements, forgetting factor, model compressibility, and estimation error of the underlying point processes in the non-asymptotic regime. We then develop two adaptive filters for recursive estimation of the ℓ_1-regularized objective function based on proximal gradient techniques, as well as a filter for recursive computation of statistical confidence regions.

Next, we extend the proposed sparse adaptive filtering framework, and develop new adaptive greedy techniques and regularization-based filters beyond the ℓ_1-norm. Greedy algorithms iteratively identify and update the model parameters until a halting criterion is met. Regularization-based methods, on the other hand, perform sparse identification through convex optimization algorithms involving sparsity-inducing regularization schemes. Most conventional methods from both approaches [24–28] operate in batch mode, and therefore do not meet real-time requirements. Several solutions have been introduced for sparse adaptive estimation problem in the literature taking greedy [23, 29] or regularization-based [21, 22] approaches. However, all these algorithms are tailored for adaptive estimation of linear Gaussian models, which makes them inapplicable for neural data analysis with highly non-Gaussian statistics. In this dissertation, we develop a sparse greedy adaptive filter for point process data based on a novel choice of the proxy metric and low-complexity recursive computational update rules. Next, we extend the filtering algorithms by considering regularization schemes beyond the ℓ_1-norm. These approaches show excellent performance in terms of sparse estimation and tracking capabilities by taking advantage of greedy iterations and optimal properties of the

employed sparsity-inducing penalty functions, respectively.

In order to validate our algorithms, we provide simulation studies which reveal that the proposed adaptive filtering algorithms significantly outperform existing point process filters in terms of goodness-of-fit, mean square error and trackability. We finally apply our proposed filters to multi-unit spike recordings from ferret primary auditory cortex (A1) during passive stimulus presentation and during performance of a click rate discrimination task [30] in order to characterize the spectrotemporal receptive field (STRF) plasticity of A1 neurons. Application of our algorithm to these data provides new insights into the time course of attention-driven STRF plasticity, with orders of magnitude increase in temporal resolution from minutes to centiseconds, while capturing the underlying sparsity in a robust fashion. Aside from their theoretical significance, our results are particularly important in light of the recent technological advances in neural prostheses, which require real-time robust neuronal system identification from limited data. The results of this chapter were selected in part for nanosymposium presentation at the annual meeting of the *Society for Neuroscience (SfN 2015)*, were presented at the *IEEE Asilomar Conference on Signals, Systems and Computers* [31] and the international *Conference of Engineering in Medicine and Biology Society (EMBC 2016)* [32], and have been published in the *IEEE Transactions on Signal Processing* [33].

Part 2: Probing Neuronal Functional Network Dynamics at High Resolutions

In the second part of dissertation, we study neuronal dynamics from a network-level perspective. Converging lines of evidence in neuroscience, from neuronal network models and neurophysiology [34–41] to resting-state imaging [42–44], suggest that sophisticated brain function results from the emergence of distributed, dynamic, and sparse functional networks underlying the brain activity. These networks are highly dynamic and task-dependent, which allows the brain to rapidly adapt to abrupt changes in the environment resulting in robust function.

Recent technological advances in neural data acquisition have resulted in abundant pools of neural data across different modalities and time-scales. In particular, simultaneous recordings from a large number of neurons have provided valuable insights into the mechanisms of complex dynamic interactions among neurons, within neuronal populations and across brain regions. In order to exploit these modern-day neural data, computationally efficient time series analysis techniques capable of simultaneously capturing the dynamicity, sparsity and statistical characteristics of the underlying functional networks are required.

Historically, various techniques such as cross-correlogram [45] and joint peristimulus time histogram [46] analyses have been utilized for inferring the statistical relationship between pairs of spike trains [45–47]. Despite being widely used, these methods are unable to provide reliable estimates of the underlying directional pat-

terns of causal interactions among an ensemble of interacting neurons due to the intrinsic deficiencies in identification of directionality, low sensitivity to inhibitory interactions [48], and susceptibility to the indirect interactions and latent common inputs.

Methods based on Granger causality (GC) analysis have shown promise in addressing these shortcomings and have thus been employed for inferring functional interactions from neural data of different modalities [49–52]. The notion of GC was originally introduced by Wiener [53] based on the concept of temporal predictability, and later adapted into more pragmatic form by Granger [54] in the context of econometrics. The rationale behind GC analysis is based on two principles: the temporal precedence of cause over effect, and the unique information of cause about the effect. Given two time series $\{X_t, Y_t\}_{t=1}^{T}$, if including the history of Y_t can improve the prediction of X_{t+1}, it is implied that the history of Y_t contains unique information about X_t, not captured by other covariates. In this case, we say that Y_t has a G-causal link to X_t.

Numerous efforts have been dedicated to extending the bivariate GC measure to more general settings, such as the conditional form of GC in [55] for multivariate setting, and several frequency-domain variants of GC [56–58]. Despite significant advances in time series analysis using GC and its variants, when applied to neuronal data, the existing methods exhibit several drawbacks.

First, most existing methods for causality inference provide static estimates of the causal influences associated with the entire data duration. Although suitable for the analysis of stationary neural data, they are not able to capture the

rapid task-dependent changes in the underlying neural dynamics. To address this challenge, several time-varying measures of causality have been proposed in the literature based on Bayesian filtering and wavelets [59–65]. Second, there are very few causal inference approaches to take into account the sparsity of the functional networks [66–68]. As an example, authors in [66] introduced a method for sparse identification of functional connectivity patterns from large-scale functional imaging data. Despite their success in inferring sparse connectivity patterns, these techniques assume static connectivity structures. Third, most existing approaches are tailored for continuous-time data, such as electroencephalography (EEG) and Local Field Potential (LFP) recordings, which limits their utility when applied to binary neuronal spike recordings. These methods are generally based on MVAR modeling, with a few non-parametric exceptions [65, 69]. Some efforts have been made to adapt the MVAR modeling to neuronal spike trains [50, 70, 71]. For instance, the binary spikes were pre-processed in [50, 70] via a smoothing kernel, which significantly distorts the temporal details of the neuronal dynamics. In addition, the frequency-domain GC analysis techniques implicitly assume that the data have rich oscillatory dynamics. Although this assumption is valid for steady-state EEG responses or resting-state recordings, spike trains recorded from cortical neuronal ensembles often do not exhibit any oscillatory behavior.

In order to address the third challenge, point process modeling and estimation have been successfully employed in capturing the stochastic dynamics of binary neuronal spiking data, as mentioned earlier [12, 72]. This framework has been particularly utilized for inferring functional interactions in neuronal ensembles from

spike recordings [67, 72–76]. A maximum likelihood (ML)-based approach was introduced in [72] based on a network likelihood formulation of the point process model; a model-based Bayesian approach based on point process likelihood models with sparse priors on the connectivity pattern was introduced in [67]. Among the more recent results, an information-theoretic measure of causality, referred to as the directed information, is proposed in [75]; a static GC measure based on point process likelihoods is proposed in [74]. However, a modeling and estimation framework to *simultaneously* take into account the dynamicity and sparsity of the G-causal influences as well as the statistical properties of binary neuronal spiking data had been lacking, which served as motivation for the second part of this dissertation.

We indeed fill this gap in Chapter 4 by developing a novel dynamic measure of GC by integrating the forgetting-factor mechanism of recursive least squares (RLS), point process modeling and sparse estimation. To this end, we first exploit the prevalent parsimony of neurophysiological time-constants manifested in neuronal spiking dynamics, such as those in sensory neurons with sharp tunings, as well as the potential low-dimensional structure of the underlying functional networks. These features can be captured by point process models in which the cross-history dependence of the neurons are described by sparse vectors. The significance of sparsity in our approach is two-fold. First, while the functional networks may not be truly sparse, they can often be parsimoniously described by a sparse set of significant functional links. Our models can indeed capture these significant links through sparse cross-history dependence. Second, sparsity enables stable estimation in the face of limited data. This is particularly important for adaptive estimation, where

the goal is to reliably estimate a large number of cross-history parameters using short effective observation windows.

Building up on the results of Chapter 3, we then employ the exponentially-weighted log-likelihood framework to recursively estimate the model parameters via sparse adaptive filtering, thereby defining a dynamic measure of GC, which we call the Adaptive Granger Causality (AGC) measure. Next, we develop a statistical inference framework for the proposed AGC measure by extending classical results on the analysis of deviance to our sparse dynamic point process setting. Moreover, we derive a non-central chi-square filtering and smoothing algorithm to track the dynamics of the underlying distributions involved in characterizing the statistical significance of the detected GC interactions.

In Chapters 5 and 6, we demonstrate the utility of our proposed methodologies through application to synthetic and real data. First, we examine the validity of our theoretical results through multiple simulation studies and numerical examples in Chapter 5, and carry out a comprehensive evaluation of the performance of the proposed methods for both discrete spiking data and continuous-valued optical imaging data. We provide numerical examples to assess the identification and tracking capabilities of the AGC inference method for synthetically generated spike trains, which reveal remarkable performance gains compared to existing techniques, in both detecting the causal links and avoiding false detections, while capturing the dynamics of the causal interactions in a neuronal ensemble. We also test the robustness of our methods to the choice of parameters, and evaluate how they affect the performance metrics. We further test the robustness of our methods against latent

confounding causal effects using comprehensive numerical studies, involving both deterministic and stochastic latent common inputs, and confounding effects due to network subsampling, which reveal that our proposed techniques provide a degree of immunity to these latent confounding effects. We finally present a simulation study to verify the utility of a static variant of our GC inference method tailored for continuous-valued data.

In Chapter 6, we present the application of our methodology to various experimental datasets from both electrophysiology and optical imaging. We first present our results on the AGC analysis of spiking data from two experimental recordings: 1) spike trains inferred from two-photon (2P) calcium imaging of the mouse auditory cortex under spontaneous activity; and 2) simultaneous spike recordings from the ferret auditory and prefrontal cortices under a tone-detection task. Our analyses of the 2P imaging data from the mouse auditory cortex reveals unique sparse dynamic features of the functional networks under spontaneous activity. Application of our methods to simultaneous spike recordings from the ferret auditory and prefrontal cortices extracts the dynamics of inter-cortical (top-down and bottom-up) functional interactions underlying attentive behavior at unprecedented spatiotemporal resolutions.

We then demonstrate the applicability of our GC inference methodology to two different optical imaging datasets: 1) 2P imaging data from the mouse primary auditory cortex during auditory tasks, and 2) whole-brain light-sheet imaging data from the larval zebrafish during fictive motor behavior. Our analysis of the 2P imaging data sheds light on the transient emergence of localized functional networks with

sparse configuration and preferred orientations under auditory task performance. Finally, our analysis of the light-sheet imaging data from the entire brain of the larval zebrafish brings new insights into the functional organization of the large-scale networks involved in visuo-motor processing. In particular, the latter analysis led to a comprehensive spectral analysis of the whole-brain imaging data and simultaneous electrophysiological recordings from motor neurons in the tail, which resulted in the discovery of a synchronized network of neurons with predominant oscillatory dynamics and forming new hypotheses on their functional role in motor behavior. The comprehensive study in the first sections of Chapters 4, 5 and 6, including the theory, simulation and applications to the neural spiking data, was presented first at the *Conference on Information Science and Systems (CISS 2016)* [77], and later at the *IEEE Asilomar Conference on Signals, Systems and Computers* [78] and the annual meeting of the *Society for Neuroscience (SfN 2017)*, and finally published in the *Proceedings of the National Academy of Sciences* [79]. The study in the second sections of the aforementioned chapters, including the theory, simulation study and applications to the continuous-valued 2P imaging data from the mouse A1, was published in *Neuron* [80].

In addition to their utility in analyzing neuronal data, our techniques have potential application in extracting functional network dynamics in other domains beyond neuroscience, such as social networks or gene regulatory networks, thanks to the plug-and-play nature of the algorithms used in our inference framework. We close this dissertation by discussing the limitations of our approach and outlining future directions of research to follow.

Chapter 2: Preliminaries and Notations

In this chapter, we give a brief introduction to point processes and discuss how it can be utilized to capture neuronal spiking statistics (see [3] for a detailed treatment). We will use the following notation throughout the dissertation: parameter vectors are denoted by bold-face greek letters, and the scalar parameters are shown by regular-type letters. For example, $\boldsymbol{\omega} = [\omega_1, \omega_2, \cdots, \omega_M]'$ denotes an M-dimensional parameter vector consisting of M scalar parameters ω_m for $m = 1, \ldots, M$, with $[\cdot]'$ denoting the transpose operator.

2.1 Overview on Point Process Theory

Consider a stochastic process defined by a sequence of discrete events occurring at random points in time, noted by $\tau_1^J = [\tau_1, \tau_2, \cdots, \tau_J]'$, and a counting measure given by

$$dN(\tau) = \sum_{k=1}^{J} \delta(\tau - \tau_k), \quad \text{and} \quad N(\tau) = \int_0^\tau dN(u), \tag{2.1}$$

where $\delta(.)$ is the Dirac's measure. The Conditional Intensity Function (CIF) for this process, denoted by $\lambda(\tau|H_\tau)$, is defined as

$$\lambda(\tau|H_\tau) := \lim_{\varepsilon \to 0} \frac{\mathbb{P}\left(N(\tau+\varepsilon) - N(\tau) = 1|H_\tau\right)}{\varepsilon}, \tag{2.2}$$

where H_τ denotes the history of the process as well as the covariates up to time τ. The CIF can be interpreted as the *instantaneous rate* given the history of the process and the covariates. For a point process with a CIF $\lambda(\tau|H_\tau)$, the likelihood of a sample path $N(\tau)$ with J events at $\tau_1 < \tau_2 < \cdots < \tau_J$ in the interval $[0, \mathcal{T}]$ is given by:

$$p(N(\tau)) = \exp\left(\int_0^{\mathcal{T}} \log \lambda(\tau|H_\tau) dN(\tau) - \int_0^{\mathcal{T}} \lambda(\tau|H_\tau) d\tau\right). \tag{2.3}$$

A point process model is fully characterized by its CIF. For instance, $\lambda(\tau|H_\tau) = \lambda$ corresponds to the homogenous Poission process with rate λ. A discretized version of this process can be obtained by binning $N(\tau)$ within an observation interval of $[0, \mathcal{T}]$ by bins of length Δ, that is

$$n_t := N(t\Delta) - N((t-1)\Delta), \quad t = 1, 2, \cdots, T, \tag{2.4}$$

where $T := \lceil \mathcal{T}/\Delta \rceil$ and $N(0) := 0$. Throughout this dissertation, $\{n_t\}_{t=1}^T$ will be considered as the observed spiking sequence, which will be used for estimation purposes. Also, by approximating Eq. 2.2 for small $\Delta \ll 1$, and denoting $\lambda_t := \lambda(t\Delta|H_{t\Delta})$, we have:

$$\begin{aligned} \mathbb{P}(n_t = 0) &= 1 - \lambda_t \Delta + o(\Delta), \\ \mathbb{P}(n_t = 1) &= \lambda_t \Delta + o(\Delta), \\ \mathbb{P}(n_t \geq 2) &= o(\Delta). \end{aligned} \tag{2.5}$$

In discrete time, the orderliness of the process is equivalent to the requirement that with high probability not more than one event fall into any given bin. In practice, this can always be achieved by choosing Δ small enough. An immediate consequence

of Eq. 2.5 is that $\{n_t\}_{t=1}^T$ can be approximated by a sequence of Bernoulli random variables with success probabilities $\{\lambda_t\Delta\}_{t=1}^T$.

A popular class of models for the CIF is given by Generalized Linear Models (GLM). In its general form, a GLM consists of two main components: an observation model, which is given by Eq. 2.5 in our case, and an equation expressing some (possibly nonlinear) function of the observation mean as a *linear* combination of the covariates. In modeling neuronal dynamics, the effective neural covariates consist of extrinsic covariates (e.g., the neural stimuli) as well as intrinsic covariates (e.g., the self- or cross-history of the neuronal activity).

Let s_t denote the stimulus at time bin t, $[\theta_0, \theta_1, \cdots, \theta_{M-2}]'$ denote the vector of stimulus modulation parameters, and μ denote the baseline firing rate. We adopt a logistic regression model for the CIF as follows:

$$\text{logit}(\lambda_t\Delta) := \log\left(\frac{\lambda_t\Delta}{1-\lambda_t\Delta}\right) = \mu + \sum_{i=0}^{M-2}\theta_i s_{t-i}. \tag{2.6}$$

By defining $\boldsymbol{\omega} := [\mu, \theta_0, \theta_1, \cdots, \theta_{M-2}]'$ and $\mathbf{x}_t := [1, s_t, \cdots, s_{t-M+2}]'$, we can equivalently write:

$$\lambda_t\Delta = \text{logit}^{-1}(\boldsymbol{\omega}'\mathbf{x}_t) := \frac{\exp(\boldsymbol{\omega}'\mathbf{x}_t)}{1+\exp(\boldsymbol{\omega}'\mathbf{x}_t)}. \tag{2.7}$$

The model above is also known as the logistic-link CIF model. Another popular model in the computational neuroscience literature is the log-link model where $\lambda_t\Delta = \exp(\boldsymbol{\omega}'\mathbf{x}_t)$. The significance of the logistic-link model is that unlike the log-link, $\text{logit}^{-1}(.)$ maps the real line $(-\infty, +\infty)$ to the unit probability interval $(0, 1)$, making it a feasible model for describing statistics of binary events independent of the scaling of the covariates and modulation parameters.

Despite capturing the stimulus dependence in quite a general form, the GLM model in Eq. 2.7 represents a static model. We therefore generalize this model to the dynamic setting by allowing temporal variability of the modulation parameters:

$$\lambda_t \Delta = \text{logit}^{-1}(\boldsymbol{\omega}_t' \mathbf{x}_t) = \frac{\exp(\boldsymbol{\omega}_t' \mathbf{x}_t)}{1 + \exp(\boldsymbol{\omega}_t' \mathbf{x}_t)}, \tag{2.8}$$

where $\boldsymbol{\omega}_t := [\mu_t, \theta_{t,0}, \theta_{t,1}, \ldots, \theta_{t,M-2}]'$ represents the time-varying parameter vector at time t. Throughout the rest of the dissertation, we refer to $\mathbf{x}_t$ and $\boldsymbol{\omega}_t$ as the covariate vector and the modulation parameter vector at time t, respectively.

In order to have a framework allowing multi-timescale dynamics, we consider piece-wise constant dynamics for the modulation parameter vector. That is, we assume that $\boldsymbol{\omega}_t$ remains constant over time windows of arbitrary length $W \geq 1$ samples, for some integer W. By segmenting the corresponding spiking data $\{n_t\}_{t=1}^{T}$ into $K := \frac{T}{W}$ windows of length W samples each, the latter assumption implies that the CIF for each time point $(k-1)W + 1 \leq t \leq kW$ is governed by $\boldsymbol{\omega}_t = \boldsymbol{\omega}_k$, for $k = 1, 2, \cdots, K$. Note that number of spiking samples T is assumed to be an integer multiple of window size W, without loss of generality.

In our applications of interest, the modulation parameter vector exhibits a degree of sparsity [81,82]. That is, only certain components in the stimulus modulation have significant contribution in determining the statistics of the process. These components can be thought of as the preferred or intrinsic tuning features of the underlying neuron. To be more precise, for a sparsity level $S < M$, we denote by $\mathcal{S} \subset \{1, 2, \cdots, M\}$ the support of the S highest elements of $\boldsymbol{\omega}$ in absolute value,

and by $\boldsymbol{\omega}_S$ the best S-term approximation to $\boldsymbol{\omega}$. We also define

$$\sigma_S(\boldsymbol{\omega}) := \|\boldsymbol{\omega} - \boldsymbol{\omega}_S\|_1 \tag{2.9}$$

to capture the compressibility of the parameter vector $\boldsymbol{\omega}$. Recall that for $\mathbf{x} \in \mathbb{R}^M$, the ℓ_1-norm is defined as $\|\mathbf{x}\|_1 := \sum_{i=1}^{M} |x_i|$. When $\sigma_S(\boldsymbol{\omega}) = 0$, the parameter $\boldsymbol{\omega}$ is called S-sparse. If $\sigma_S(\boldsymbol{\omega}) = \mathcal{O}(S^{1-\frac{1}{\xi}})$ for some $\xi \in (0, 1)$, the parameter is called (ξ, S)-compressible [26].

Chapter 3: Adaptive Identification of Sparse Neuronal Models

In this chapter, we develop a novel sparse adaptive filtering framework for estimation of the dynamics of a neuronal model under plasticity from its spiking activity. We provide an algorithmic framework for sparse adaptive filtering of point process data based on greedy techniques and regularized optimization, and present theoretical guarantees on their performance.

We further assess the performance of the proposed filtering algorithms in terms of goodness-of-fit, estimation accuracy and trackability using simulation studies. Finally, we present the application of our filtering algorithms to experimentally recorded spiking data from the ferret primary auditory cortex during attentive behavior. Note that we only consider purely extrinsic covariates (e.g., acoustic stimuli) for GLM models in this chapter, although most of our results can be generalized to incorporate intrinsic covariates as well, as discussed in the forthcoming chapters.

3.1 ℓ_1-regularized Point Process Adaptive Filtering

In this section, we first formulate the sparse adaptive estimation problem, and later provide algorithmic solutions and theoretical guarantees. We propose two efficient filtering algorithms for adaptive estimation of the sparse time-varying

modulation coefficients from point process observations, through recursive solution of a sequence of ℓ_1-regularized ML problems via proximal algorithms. Moreover, we describe how statistical confidence intervals can also be constructed in a recursive fashion for our estimates.

We provide a rigorous theoretical analysis on the consistency of the estimated sparse parameter vectors, and thereby extend the classical CS guarantees to the more complex setting of dynamic point process models.

3.1.1 Problem Formulation

The main estimation problem of this chapter can be stated as follows: *given binary observations* $\{n_t\}_{t=1}^T$ *and covariates* $\{\mathbf{x}_t\}_{t=-M+1}^T$ *from a point process with a CIF given by Eq. 2.8, the goal is to estimate the* M*-dimensional parameter vectors* $\{\boldsymbol{\omega}_t\}_{t=1}^T$ *in an online and stable fashion.*

To establish a sparse adaptive filtering framework, we first construct a new objective function tailored specifically for our sparse dynamic point process setting. Recall that for small choices of bin size $\Delta \ll 1$, the point process statistics can be simplified using the Bernoulli approximation, as was shown in Eq. 2.5. Based on this approximation, the log-likelihood of the observation n_t at time t can be expressed as:

$$\begin{aligned} \log p(n_t) &\approx n_t \log(\lambda_t \Delta) + (1 - n_t) \log(1 - \lambda_t \Delta) \\ &= n_t(\mathbf{x}_t' \boldsymbol{\omega}_t) - \log\left(1 + \exp\left(\mathbf{x}_t' \boldsymbol{\omega}_t\right)\right). \end{aligned} \tag{3.1}$$

Assuming conditional independence of the spiking events, the joint log-likelihood

of the observations within window i evaluated at $\boldsymbol{\omega}$ is given by:

$$\ell_i(\boldsymbol{\omega}) := \sum_{j=1}^{W} \left\{ n_{(i-1)W+j} \mathbf{x}'_{(i-1)W+j} \boldsymbol{\omega} - \log\left(1 + \exp(\mathbf{x}'_{(i-1)W+j} \boldsymbol{\omega})\right) \right\}. \tag{3.2}$$

In order to explicitly enforce adaptivity in the log-likelihood function, we adopt the forgetting factor mechanism of the RLS algorithm, where the log-likelihood of each window is exponentially weighted regressively in time, with a forgetting factor $0 < \beta \leq 1$. That is, the effective data log-likelihood up to and including window k is taken to be:

$$\ell^{\beta}(\boldsymbol{\omega}_k) := \sum_{i=1}^{k} \beta^{k-i} \ell_i(\boldsymbol{\omega}_k) \tag{3.3}$$

for some $0 < \beta \leq 1$. Note that for $\beta = 1$, $\ell^1(\boldsymbol{\omega}_k)$ coincides with the natural data log-likelihood. Moreover, if we replace the Bernoulli log-likelihood with the Gaussian log-likelihood, then $\ell^{\beta}(\boldsymbol{\omega}_k)$ coincides with the conventional RLS objective function.

Next, in order to explicitly enforce sparsity, we adopt the ℓ_1-regularization mechanism. That is, at every window k, we seek to solve an ℓ_1-regularized ML problem of the form:

$$\widehat{\boldsymbol{\omega}}_k = \underset{\boldsymbol{\omega}_k}{\operatorname{argmax}} \quad \left\{ \ell^{\beta}(\boldsymbol{\omega}_k) - \gamma \|\boldsymbol{\omega}_k\|_1 \right\}, \tag{3.4}$$

where γ is a regularization parameter controlling the trade-off between the log-likelihood fit and the sparsity of estimated parameters. In the next subsection, we proceed with the development of recursive filters to track the solutions of the ℓ_1-regularized ML problem sequence in the more general time-varying setting.

3.1.2 Algorithm Development

Several standard optimization techniques, such as interior point methods, can be used to find the maximizer of Eq. 3.4. However, most of these techniques operate in batch mode and do not meet the real-time requirements of the adaptive filtering setting where the observations arrive in a streaming fashion. In order to avoid the increasing runtime complexity and memory requirements of the batch-mode computation, we seek a recursive approach which can perform low-complexity updates in an online fashion upon the arrival of new data in order to form the estimates. To this end, we adopt the proximal gradient approach. Each iteration of the proximal algorithm moves the previous iterate along the gradient of the log-likelihood function, which will then pass through a shrinkage operator. For more details on the proximal gradient algorithm, see Appendix A.2.

Before proceeding further with our development, we introduce a more compact notation for convenience. Let $\mathbf{n}_k := [n_{(k-1)W+1},\ n_{(k-1)W+2}, \ldots, n_{kW}]'$ denote the vector of observed spikes within window k, for $k = 1, 2, \ldots, K$. Similarly, let $\boldsymbol{\lambda}_k := \left[\lambda_{(k-1)W+1},\ \lambda_{(k-1)W+2}, \ldots, \lambda_{kW}\right]'$ denote the vector of CIFs within window k. By extending the domain of the $\text{logit}^{-1}(\cdot)$ to vectors in a component-wise fashion, we define $\boldsymbol{\lambda}_k(\boldsymbol{\omega})$ for any window k and any parameter $\boldsymbol{\omega}$ to be:

$$\boldsymbol{\lambda}_k(\boldsymbol{\omega}) := \frac{1}{\Delta} \text{logit}^{-1}\left(\mathbf{X}_k\boldsymbol{\omega}\right), \tag{3.5}$$

where $\mathbf{X}_k := \left[\mathbf{x}_{(k-1)W+1}, \mathbf{x}_{(k-1)W+2}, \ldots, \mathbf{x}_{kW}\right]'$ is the data matrix of size $W \times M$ with rows corresponding to the covariate vectors in window k. Suppose that at window

k, we have an iterate denoted by $\widehat{\boldsymbol{\omega}}_k^{(\ell)}$, for $\ell = 0, 1, \cdots, L$, with L being an integer denoting the total number of iterations. The gradient of $\ell^\beta(\cdot)$ evaluated at $\widehat{\boldsymbol{\omega}}_k^{(\ell)}$ can be written as:

$$\nabla_{\boldsymbol{\omega}} \ell^\beta \left(\widehat{\boldsymbol{\omega}}_k^{(\ell)} \right) = \sum_{i=1}^{k} \beta^{k-i} \mathbf{X}_i' \boldsymbol{\varepsilon}_i \left(\widehat{\boldsymbol{\omega}}_k^{(\ell)} \right) =: \mathbf{g}_k \left(\widehat{\boldsymbol{\omega}}_k^{(\ell)} \right), \tag{3.6}$$

where $\boldsymbol{\varepsilon}_i(\cdot) := \mathbf{n}_i - \boldsymbol{\lambda}_i(\cdot)\Delta$ represents the innovation vector of the point process at window i. The innovation vector $\boldsymbol{\varepsilon}_i$ can be thought of as the counterpart of the conventional innovation vector in adaptive filtering of linear models. The proximal gradient iteration for the ℓ_1-regularized ML problem can be written in the compact form as:

$$\widehat{\boldsymbol{\omega}}_k^{(\ell+1)} = \mathcal{S}_{\gamma\alpha} \Big(\widehat{\boldsymbol{\omega}}_k^{(\ell)} + \alpha \mathbf{g}_k \left(\widehat{\boldsymbol{\omega}}_k^{(\ell)} \right) \Big) \tag{3.7}$$

where $\mathcal{S}_\tau(\cdot)$ is the element-wise soft thresholding operator at a level of τ given in Appendix A.2. The final estimate at window k is obtained following the L^{th} iteration, and is denoted by $\widehat{\boldsymbol{\omega}}_k := \widehat{\boldsymbol{\omega}}_k^{(L)}$. In order to achieve a recursive updating rule for $\mathbf{g}_k$, we can rewrite Eq. 3.6 as:

$$\mathbf{g}_k \left(\widehat{\boldsymbol{\omega}}_k^{(\ell)} \right) = \beta \, \mathbf{g}_{k-1} \left(\widehat{\boldsymbol{\omega}}_k^{(\ell)} \right) + \mathbf{X}_k' \boldsymbol{\varepsilon}_k \left(\widehat{\boldsymbol{\omega}}_k^{(\ell)} \right). \tag{3.8}$$

However, in an adaptive setting, we only have access to values of $\mathbf{g}_{k-1}$ evaluated at $\widehat{\boldsymbol{\omega}}_{k-1}^{(1:L)}$! In order to turn Eq. 3.8 into a fully recursive updating rule, all the previous CIF vectors $\{\boldsymbol{\lambda}_i(\cdot)\}_{i=1}^{k-1}$ should be recalculated at the most recent set of iterates $\widehat{\boldsymbol{\omega}}_k^{(1:L)}$. In order to overcome this computational burden, we exploit the smoothness of the logistic function and employ the Taylor series expansion of the CIF to approximate the required recursive update. In what follows, we consider

the zeroth order and first order expansions, which result in two distinct, yet fully recursive, updating rules for Eq. 3.8.

Zeroth Order Expansion: By retaining only the first term in the Taylor series expansion of the CIF $\boldsymbol{\lambda}_i\left(\widehat{\boldsymbol{\omega}}_k^{(\ell)}\right)$ around $\widehat{\boldsymbol{\omega}}_i$, we get:

$$\boldsymbol{\lambda}_i\left(\widehat{\boldsymbol{\omega}}_k^{(\ell)}\right)\Delta \approx \boldsymbol{\lambda}_i\left(\widehat{\boldsymbol{\omega}}_i\right)\Delta, \tag{3.9}$$

where $\boldsymbol{\lambda_i}(\widehat{\boldsymbol{\omega}}_i)\Delta = \text{logit}^{-1}(\mathbf{X}_i\widehat{\boldsymbol{\omega}}_i)$. Substituting this approximation in Eq. 3.6, we can express the zeroth order approximation to the gradient at window k, denoted by $\mathbf{g}_k^0(\cdot)$, as:

$$\mathbf{g}_k^0\left(\widehat{\boldsymbol{\omega}}_k^{(\ell)}\right) = \sum_{i=1}^{k}\beta^{k-i}\,\mathbf{X}_i'\boldsymbol{\varepsilon}_i(\widehat{\boldsymbol{\omega}}_i). \tag{3.10}$$

It is then straightforward to obtain a recursive form as:

$$\mathbf{g}_k^0\left(\widehat{\boldsymbol{\omega}}_k^{(\ell)}\right) = \beta\;\mathbf{g}_{k-1}^0\left(\widehat{\boldsymbol{\omega}}_k^{(\ell)}\right) + \mathbf{X}_k'\boldsymbol{\varepsilon}_k\left(\widehat{\boldsymbol{\omega}}_k^{(\ell)}\right).$$

The shrinkage step will be then given by:

$$\widehat{\boldsymbol{\omega}}_k^{(\ell+1)} = \mathcal{S}_{\gamma\alpha}\Big(\widehat{\boldsymbol{\omega}}_k^{(\ell)} + \alpha\mathbf{g}_k^0\left(\widehat{\boldsymbol{\omega}}_k^{(\ell)}\right)\Big). \tag{3.11}$$

We refer to the resulting filter as the ℓ_1-regularized Point Process Filter of the Zeroth Order (ℓ_1-PPF$_0$). A pseudo-code is given in Algorithm 1.

First Order Expansion: If instead, we retain the first two terms in the Taylor expansion, Eq. 3.9 will be replaced by:

$$\boldsymbol{\lambda}_i\left(\widehat{\boldsymbol{\omega}}_k^{(\ell)}\right)\Delta \approx \boldsymbol{\lambda}_i\left(\widehat{\boldsymbol{\omega}}_i\right)\Delta + \boldsymbol{\Lambda}_i\left(\widehat{\boldsymbol{\omega}}_i\right)\mathbf{X}_i\left(\widehat{\boldsymbol{\omega}}_k^{(\ell)} - \widehat{\boldsymbol{\omega}}_i\right), \tag{3.12}$$

Algorithm 1 ℓ_1-regularized Point Process Filter of the Zeroth Order (ℓ_1-PPF$_0$)

Inputs: $\mathbf{n}_k$, $\mathbf{X}_k$, $\mathbf{g}_{k-1}$, $\widehat{\boldsymbol{\omega}}_k^{(0)}$, and L .

1: **for** $\ell = 0, \ldots, L-1$ **do**

2: $\quad \boldsymbol{\lambda}_k \Delta = \text{logit}^{-1}\left(\mathbf{X}_k \widehat{\boldsymbol{\omega}}_k^{(\ell)}\right)$

3: $\quad \boldsymbol{\varepsilon}_k = \mathbf{n}_k - \boldsymbol{\lambda}_k \Delta$

4: $\quad \mathbf{g}_k = \beta\, \mathbf{g}_{k-1} + \mathbf{X}_k' \boldsymbol{\varepsilon}_k$

5: $\quad \widehat{\boldsymbol{\omega}}_k^{(\ell+1)} = \mathcal{S}_{\gamma\alpha}\Big[\widehat{\boldsymbol{\omega}}_k^{(\ell)} + \alpha \mathbf{g}_k\Big]$

6: **end for**

Output: $\widehat{\boldsymbol{\omega}}_k := \widehat{\boldsymbol{\omega}}_k^{(L)}$.

where $\boldsymbol{\Lambda}_i(\widehat{\boldsymbol{\omega}}_i)$ is a diagonal $W \times W$ matrix with the (m,m)-th diagonal element given by $\lambda_{(i-1)W+m}\Delta(1 - \lambda_{(i-1)W+m}\Delta)$. Using the first order approximation above, we can improve the resulting approximation to the gradient, denoted by $\mathbf{g}_k^1$, as:

$$\mathbf{g}_k^1\left(\widehat{\boldsymbol{\omega}}_k^{(\ell)}\right) = \sum_{i=1}^{k} \beta^{k-i}\, \mathbf{X}_i'\Big(\boldsymbol{\varepsilon}_i(\widehat{\boldsymbol{\omega}}_i) - \boldsymbol{\Lambda}_i(\widehat{\boldsymbol{\omega}}_i)\mathbf{X}_i\big(\widehat{\boldsymbol{\omega}}_k^{(\ell)} - \widehat{\boldsymbol{\omega}}_i\big)\Big). \tag{3.13}$$

By defining:

$$\begin{aligned} \mathbf{u}_k &:= \sum_{i=1}^{k} \beta^{k-i}\, \mathbf{X}_i'\Big(\boldsymbol{\varepsilon}_i(\widehat{\boldsymbol{\omega}}_i) + \boldsymbol{\Lambda}_i(\widehat{\boldsymbol{\omega}}_i)\mathbf{X}_i\widehat{\boldsymbol{\omega}}_i\Big) \\ \mathbf{B}_k &:= \sum_{i=1}^{k} \beta^{k-i}\, \mathbf{X}_i'\boldsymbol{\Lambda}_i(\widehat{\boldsymbol{\omega}}_i)\mathbf{X}_i, \end{aligned} \tag{3.14}$$

we can express $\mathbf{g}_k^1\left(\widehat{\boldsymbol{\omega}}_k^{(\ell)}\right)$ as:

$$\begin{aligned} \mathbf{g}_k^1\left(\widehat{\boldsymbol{\omega}}_k^{(\ell)}\right) &= \mathbf{u}_k - \mathbf{B}_k\widehat{\boldsymbol{\omega}}_k^{(\ell)} \\ &= \beta\, \mathbf{g}_{k-1}^1\left(\widehat{\boldsymbol{\omega}}_k^{(\ell)}\right) + \mathbf{X}_k'\boldsymbol{\varepsilon}_k(\widehat{\boldsymbol{\omega}}_k). \end{aligned} \tag{3.15}$$

The shrinkage step is then given by:

$$\widehat{\boldsymbol{\omega}}_k^{(\ell+1)} = \mathcal{S}_{\gamma\alpha}\Big(\widehat{\boldsymbol{\omega}}_k^{(\ell)} + \alpha \mathbf{g}_k^1\left(\widehat{\boldsymbol{\omega}}_k^{(\ell)}\right)\Big). \tag{3.16}$$

It is then straightforward to check that both $\mathbf{u}_k$ and $\mathbf{B}_k$ can be updated recursively as:

$$\mathbf{u}_k = \beta\,\mathbf{u}_{k-1} + \mathbf{X}_k'\Big(\boldsymbol{\varepsilon}_k\left(\widehat{\boldsymbol{\omega}}_k^{(L)}\right) + \boldsymbol{\Lambda}_k\left(\widehat{\boldsymbol{\omega}}_k^{(L)}\right)\mathbf{X}_k\widehat{\boldsymbol{\omega}}_k^{(L)}\Big),$$

$$\mathbf{B}_k = \beta\,\mathbf{B}_{k-1} + \mathbf{X}_k'\boldsymbol{\Lambda}_k\left(\widehat{\boldsymbol{\omega}}_k^{(L)}\right)\mathbf{X}_k.$$

Note that the update rules for both $\mathbf{B}_k$ and $\mathbf{u}_k$ involve simple rank-W operations. We refer to the resulting filter as the ℓ_1-regularized Point Process Filter of the First Order (ℓ_1-PPF$_1$). A pseudo-code is given in Algorithm 2.

Algorithm 2 ℓ_1-regularized Point Process Filter of the First Order (ℓ_1-PPF$_1$)

Inputs: $\mathbf{n}_k$, $\mathbf{X}_k$, $\mathbf{u}_{k-1}$, $\mathbf{B}_{k-1}$, $\widehat{\boldsymbol{\omega}}_k^{(0)}$, and L .

for $\ell = 0, \ldots, L-1$ **do**
  $\boldsymbol{\lambda}_k^{(\ell)}\Delta = \text{logit}^{-1}\left(\mathbf{X}_k\widehat{\boldsymbol{\omega}}_k^{(\ell)}\right)$
  $\boldsymbol{\varepsilon}_k^{(\ell)} = \mathbf{n}_k - \boldsymbol{\lambda}_k^{(\ell)}\Delta$
  $\mathbf{g}_k^{(\ell)} = \beta\big(\mathbf{u}_{k-1} - \mathbf{B}_{k-1}\widehat{\boldsymbol{\omega}}_k^{(\ell)}\big) + \mathbf{X}_k'\boldsymbol{\varepsilon}_k^{(\ell)}$
  $\widehat{\boldsymbol{\omega}}_k^{(\ell+1)} = \mathcal{S}_{\gamma\alpha}\big[\widehat{\boldsymbol{\omega}}_k^{(\ell)} + \alpha\mathbf{g}_k^{(\ell)}\big]$
end for
$(\boldsymbol{\Lambda}_k)_{m,m} = (\boldsymbol{\lambda}_k^{(L)})_m\Delta\big(1 - (\boldsymbol{\lambda}_k^{(L)})_m\Delta\big)$, $m = 1, \cdots, W$
$\mathbf{u}_k = \beta\,\mathbf{u}_{k-1} + \mathbf{X}_k'\big(\boldsymbol{\varepsilon}_k^{(L)} + \boldsymbol{\Lambda}_k\mathbf{X}_k\widehat{\boldsymbol{\omega}}_k^{(L)}\big)$
$\mathbf{B}_k = \beta\,\mathbf{B}_{k-1} + \mathbf{X}_k'\boldsymbol{\Lambda}_k\mathbf{X}_k$

Output: $\widehat{\boldsymbol{\omega}}_k := \widehat{\boldsymbol{\omega}}_k^{(L)}$.

Remark. The computational complexity of ℓ_1-PPF$_0$ and ℓ_1-PPF$_1$ algorithms can be shown to be linear and quadratic in M per iteration, respectively. Our results in Section 3.3 will reveal that both filters outperform existing filters of the same complexity, respectively. Furthermore, ℓ_1-PPF$_1$ exhibits superior performance over ℓ_1-PPF$_0$ as expected, although with an additional cost of $\mathcal{O}(M)$ in computational complexity per iteration. Our theoretical analysis in the next subsection reveals appropriate choices for γ, β and the trade-offs therein.

3.1.3 Theoretical Guarantees and Trade-offs

In order to quantify the trade-offs involving our choice of the objective function in Eq. 3.4, we proceed in the tradition of performance analysis result of the RLS algorithm [13] by characterizing the geometric properties of the estimates $\boldsymbol{\omega}_k$ in a stationary environment where $\boldsymbol{\omega_k} = \boldsymbol{\omega}$ for all k. Our analysis, however, is quite general and avoids ad hoc assumptions such as direct averaging or covariate independence which are usually invoked in the analysis of least squares problems.

We have the following theoretical result regarding the consistency of the sparse parameter vector estimates:

Theorem 3.1 *Suppose that binary observations from a point process with a CIF given by Eq. 2.8 are given over K windows of length W each. Consider the setting where $\boldsymbol{\omega}_k = \boldsymbol{\omega}$ for all k. Then, under mild technical assumptions, for an arbitrarily chosen positive constant $d > 0$, there exist constants C, C', and C'' such that for $M > 10S$, $1 - \frac{C'}{S^2 \log M} \leq \beta < 1$, $K \geq \frac{\log 2}{\log(\frac{1}{\beta})}$, and a choice of $\gamma = C''\sqrt{\frac{\log M}{1-\beta}}$, any solution $\widehat{\boldsymbol{\omega}}$ to Eq. 3.4 satisfies the bound*

$$\|\widehat{\boldsymbol{\omega}} - \boldsymbol{\omega}\|_2 \leq C\sqrt{(1-\beta)S\log M} + \sqrt{C\sigma_S(\boldsymbol{\omega})}\sqrt[4]{(1-\beta)S\log M},$$

with probability at least $1 - \frac{5}{M^d}$. The constants C, C', and C'' are only functions of d, $p_{\min}$, $p_{\max}$, σ^2, B, and W, and are explicitly given in Appendix A.1.

Proof 3.1 *The proof of Theorem 3.1 is given in Appendix A.1.*

Remarks. The result of Theorem 3.1 has four major implications. First, assuming that $\sigma_S(\boldsymbol{\omega}) = 0$, the error bound scales with $\sqrt{(1-\beta)S\log M}$, the sparsity level, as opposed to $\sqrt{(1-\beta)M}$ for the ML estimate. This is consistent with results from conventional CS, where given T observations, the error bound of the ℓ_1-regularized estimate scales as $\sqrt{\frac{S\log M}{T}}$ as opposed to $\sqrt{\frac{M}{T}}$ obtained by the least squares estimate [19, 83, 84]. Note that the latter implies a putative performance gain of order $\mathcal{O}\left(\frac{M}{S\log M}\right)$ in terms of estimation error, and thereby results in the robustness of the estimate when the underlying parameter is sparse. Nevertheless, the bound holds for general non-sparse $\boldsymbol{\omega}$, but is sharpest when $\sigma_S(\boldsymbol{\omega})$ is negligible, i.e., the parameter vector is nearly S-sparse. If the $\boldsymbol{\omega}$ is (ξ, S)-compressible, the second term scales as $S^{\frac{3\xi-2}{4\xi}}$. In particular, as $\xi \to 0$ resulting in an S-sparse parameter vector, the second term vanishes.

Second, the theorem prescribes a lower bound on the forgetting factor akin to the bounds obtained in CS theory for the total number of observations. For instance, the result of [85] for CS under Toeplitz sensing measurements for the linear model requires $T = \mathcal{O}(S^2\log M)$ number of measurements to achieve a similar scaling of the error bound. In our case, the role of the number of measurements is transferred to forgetting factor by taking $\frac{1}{1-\beta}$ as the *effective* length of the measurements. In the absence of the forgetting factor ($\beta = 1$), by a careful limiting process, our results require $T = \mathcal{O}(S^2\log M)$ measurements. The latter case can be compared to the result of [19] for point process models with independent and identically distributed covariate vectors, which requires $\mathcal{O}(S\log M)$ for stability. The loss of $\mathcal{O}(S)$ is incurred due to the shift structure and hence high dependence of the covariate vectors

in our case.

Third, the theorem reveals the scaling of the regularization parameter in terms of M and β. In particular, this scaling is significant as it reveals another role for the forgetting factor mechanism: not only the forgetting factor mechanism allows for adaptivity of the estimates, it also influences the scaling of the ℓ_1-regularization term with respect to the log-likelihood term. Fourth, unlike conventional results in the analysis of adaptive filters which concern the expectation of the error in the asymptotic regime, our result holds for a single realization with probability polynomially approaching 1, in the non-asymptotic regime.

Note that the objective function in Eq. 3.4 is clearly concave, and assuming that the matrix of the covariate vectors is full-rank, will be strictly concave with a unique maximizer. However, the result of Theorem 3.1 does not require the uniqueness of the maximizer and holds for any maximizer of the objective function.

3.1.4 Constructing Confidence Intervals

Characterizing the statistical confidence bounds associated with the estimates is of utmost importance in neural data analysis, as it allows to test the validity of various hypotheses. Although construction of confidence bounds for linear models in the absence of regularization is well understood and widely applied, regularized ML estimates are usually deemed as point estimates for which the construction of statistical confidence regions is not straightforward. A series of remarkable results in high-dimensional statistics [86–88] have recently addressed this issue by providing

techniques to construct confidence intervals for ℓ_1-regularized ML estimates of linear models as well as GLMs. These approaches are based on a careful inspection of the Karush-Kuhn-Tucker (KKT) conditions for the regularized estimates, which admits a procedure to decompose the estimates into a bias term plus an asymptotically Gaussian term (referred to as 'de-biasing' in [87]), which can be computed using a nodewise regression [89] of the covariates.

In what follows, we give a brief description of how the methods of [87] apply to our setting, and leave the details to Appendix A.3. Using the result of [87], the estimate $\widehat{\boldsymbol{\omega}}_k$ as the maximizer of 3.4 can be decomposed as:

$$\widehat{\boldsymbol{\omega}}_k = \widehat{\boldsymbol{\Theta}}_k \mathbf{g}_k(\widehat{\boldsymbol{\omega}}_k) + \widehat{\mathbf{w}}_k, \tag{3.17}$$

where $\widehat{\boldsymbol{\Theta}}_k$ is an approximate inverse to the Hessian of $\ell^\beta(\boldsymbol{\omega})$ evaluated at $\widehat{\boldsymbol{\omega}}_k$, $\mathbf{g}_k$ is the gradient of $\ell^\beta(\boldsymbol{\omega})$ previously defined in Eq. 3.6, and $\widehat{\mathbf{w}}_k$ is an unbiased and asymptotically Gaussian random vector with a covariance matrix of $\mathsf{cov}(\widehat{\mathbf{w}}_k) = \widehat{\boldsymbol{\Theta}}_k \mathbf{G}_k(\widehat{\boldsymbol{\omega}}_k) \widehat{\boldsymbol{\Theta}}_k'$, with

$$\mathbf{G}_k(\widehat{\boldsymbol{\omega}}_k) := \sum_{i=1}^{k} \beta^{2(k-i)} \mathbf{X}_i' \boldsymbol{\varepsilon}_i(\widehat{\boldsymbol{\omega}}_k) \boldsymbol{\varepsilon}_i(\widehat{\boldsymbol{\omega}}_k)' \mathbf{X}_i. \tag{3.18}$$

The first term in Eq. 3.17 is a bias term which can be directly computed given $\widehat{\boldsymbol{\Theta}}_k$. Given $\mathsf{cov}(\widehat{\mathbf{w}}_k)$, statistical confidence bounds for the second term can be constructed at desired levels in a standard way. The main technical issue in the aforementioned procedure in our setting is the computation of $\widehat{\boldsymbol{\Theta}}_k$ in a recursive fashion. Since the rows of $\widehat{\boldsymbol{\Theta}}_k$ are computed using ℓ_1-regularized least squares, we use the SPARLS algorithm [21] as an efficient method to carry out the computation in a recursive

fashion. Algorithm 3 summarized the recursive computation of confidence intervals for the m-th component of $\widehat{\mathbf{w}}_k$.

Algorithm 3 Recursive Construction of the Confidence Regions for the m-th Component of $\widehat{\mathbf{w}}_k$.[a]

Inputs: $\mathbf{n}_k$, $\mathbf{X}_k$, $\mathbf{u}_k$, $\mathbf{B}_k$, and $\widehat{\boldsymbol{\omega}}_k$, $\mathbf{G}_{k-1}$, m, γ_m, and $\widehat{\boldsymbol{\psi}}_m^{(0)}$.

$\mathbf{g}_k = \mathbf{u}_k - \mathbf{B}_k\widehat{\boldsymbol{\omega}}_k$

$\mathbf{G}_k = \beta^2\mathbf{G}_{k-1} + \mathbf{X}_k'\boldsymbol{\varepsilon}_k\boldsymbol{\varepsilon}_k'\mathbf{X}_k$

for $\ell = 0, \ldots, L-1$ **do**

$\quad \widehat{\boldsymbol{\psi}}_m^{(\ell+1)} = \mathcal{S}_{\gamma_m\alpha}\Big[\widehat{\boldsymbol{\psi}}_m^{(\ell)} - \alpha\Big((\mathbf{B}_k)_{m,\setminus m} - (\mathbf{B}_k)_{\setminus m,\setminus m}\widehat{\boldsymbol{\psi}}_m^{(\ell)}\Big)\Big]$

end for

$\tau_m^2 = (\mathbf{B}_k)_{m,m} - \widehat{\boldsymbol{\psi}}_m^{(L)}(\mathbf{B}_k)_{m,\setminus m}'$

$(\mathbf{c})_m = 1$, and $(\mathbf{c})_{\setminus m} = -\widehat{\boldsymbol{\psi}}_m^{(L)}$

$(\widehat{\boldsymbol{\Theta}}_k)_m = \frac{1}{\tau_m^2}\mathbf{c}$

$\widehat{\sigma}_{k,m}^2 := (\widehat{\boldsymbol{\Theta}}_k)_m\mathbf{G}_k(\widehat{\boldsymbol{\Theta}}_k)_m'$

$(\widehat{\mathbf{w}}_k)_m = (\widehat{\boldsymbol{\omega}}_k)_m - (\widehat{\boldsymbol{\Theta}}_k)_m\mathbf{g}_k$

Output: $\mathcal{CR}_{k,m} := [(\widehat{\mathbf{w}}_k)_m \pm \Phi^{-1}(1-\alpha/2)\,\widehat{\sigma}_{k,m}]$

[a]For a matrix $\mathbf{A} \in \mathbb{R}^{M\times M}$, we denote by $(\mathbf{A})_{m,\setminus m}$ the m-th row with the m-th element removed, and by $(\mathbf{A})_{\setminus m,\setminus m}$ the submatrix of $\mathbf{A}$ with both the m-th row and column removed.

3.2 A Family of Sparse Adaptive Algorithms for Spiking Data

In this section, we introduce two new classes of sparse adaptive filtering algorithms for point process data: First, we develop a sparse greedy point process filter based on a novel choice of the proxy metric and low-complexity recursive update rules. Second, we extend our proposed ℓ_1-PPF adaptive filtering framework by incurring sparsifying regularization schemes beyond the ℓ_1-norm. These approaches can improve the estimation performance of the adaptive filters by taking advantage of greedy iterations and optimal properties of the employed sparsity-inducing penalty functions, respectively.

3.2.1 Greedy Approach

We develop a framework for greedy adaptive filtering which can be integrated with a variety of existing greedy techniques [24–26]. We choose the Compressed Sampling Matching Pursuit algorithm (CoSaMP) [26] for presentation of our greedy framework here. Inspired by [29], we modify the update procedure of CoSaMP to the adaptive setting by integrating the forgetting factor mechanism of RLS algorithm into the proxy identification and estimation steps. In our adaptive scheme, the gradient of the effective log-likelihood function defined earlier in Eq. 3.6, namely the *score function*, is chosen as the proxy signal for identification of support set. Intuitively speaking, the score function is an indication of the sensitivity of the data log-likelihood function with respect to each component of the parameter vector, which makes it a suitable candidate for the proxy metric. The forgetting factor-based scheme of the proposed proxy metric enables us to capture the variations of the support set over time. Note that the proposed proxy function for the point process log-likelihood is analogous to the correlation-based proxy signal used in greedy algorithms for linear Gaussian models.

As for the estimation step, we compute the ML estimate based on the adaptive log-likelihood objective function $\ell_k^{\beta}(\boldsymbol{\omega}_k)$ at each time step k, by performing a simple gradient ascent update as follows:

$$\widehat{\boldsymbol{\omega}}_{k\ |\Omega_k}^{(\ell+1)} = \widehat{\boldsymbol{\omega}}_{k\ |\Omega_k}^{(\ell)} + \alpha \mathbf{g}_{k\ |\Omega_k}^{(\ell)} \tag{3.19}$$

where $\alpha > 0$ is the step size, $\ell = 0, 1, \ldots, L-1$ is the iteration index, $\mathbf{g}_{k\ |\Omega_k}^{(\ell)} :=$

$\mathbf{g}_k(\widehat{\boldsymbol{\omega}}_{k\,|\Omega_k}^{(\ell)})$, and the subscripted notation $\widehat{\boldsymbol{\omega}}_{k\,|\Omega_k}^{(\ell)}$ represents the restriction of $\widehat{\boldsymbol{\omega}}_k^{(\ell)}$ to the updated merged support set Ω_k.

The gradient function $\mathbf{g}_k$ plays a central role in the proposed greedy procedure, both in the proxy identification and the estimation steps. Hence, a recursive update rule for the gradient function is required in order to attain a low-complexity real-time greedy algorithm. We consider the fully recursive update rule derived in Eq. 3.15 based on the first-order Taylor series expansion. The same update procedure can be applied for the gradient ascent update in the estimation stage, by replacing $\widehat{\boldsymbol{\omega}}_k$ with the index-restricted version $\widehat{\boldsymbol{\omega}}_{k\,|\Omega_k}$.

Algorithm 4 Sparse Greedy Point Process Filter (SGPPF)

Inputs: n_k, $\mathbf{x}_k$, $\widehat{\boldsymbol{\omega}}_k^{(0)}$, $\mathbf{u}_{k-1}$, and $\mathbf{B}_{k-1}$.

for $\ell = 1, 2, \ldots, L-1$ **do**
- $\lambda_k^{(\ell)}\Delta = \text{logit}^{-1}\left(\mathbf{x}_k'\widehat{\boldsymbol{\omega}}_k^{(\ell)}\right)$
- $\kappa_k^{(\ell)} = \lambda_k^{(\ell)}\Delta(1 - \lambda_k^{(\ell)}\Delta)$
- $\varepsilon_k^{(\ell)} = n_k - \lambda_k^{(\ell)}\Delta$
- $\mathbf{p}_k^{(\ell)} := \mathbf{g}_k(\widehat{\boldsymbol{\omega}}_k^{(\ell)}) = \beta(\mathbf{u}_{k-1} - \mathbf{B}_{k-1}\widehat{\boldsymbol{\omega}}_k^{(\ell)}) + \varepsilon_k^{(\ell)}\mathbf{x}_k$
- $\Omega_k = \text{Supp}(\mathbf{p}_k^{(\ell)}, 2S) \cup \mathcal{S}_k^{(\ell)}$
- $\mathbf{g}_{k\,|\Omega_k}^{(\ell)} = \beta(\mathbf{u}_{k-1\,|\Omega_k} - \mathbf{B}_{k-1\,|\Omega_k}\widehat{\boldsymbol{\omega}}_{k\,|\Omega_k}^{(\ell)}) + \varepsilon_k^{(\ell)}\mathbf{x}_{k\,|\Omega_k}$
- $\widehat{\boldsymbol{\omega}}_{k\,|\Omega_k}^{(\ell+1)} = \widehat{\boldsymbol{\omega}}_{k\,|\Omega_k}^{(\ell)} + \alpha\mathbf{g}_{k\,|\Omega_k}^{(\ell)}$
- $\mathcal{S}_k^{(\ell+1)} = \text{Supp}(\widehat{\boldsymbol{\omega}}_{k\,|\Omega_k}^{(\ell+1)}, S)$
- $\widehat{\boldsymbol{\omega}}_{k\,|\mathcal{S}_k^C}^{(\ell+1)} = \mathbf{0}$

end for

$\mathbf{u}_k = \beta\,\mathbf{u}_{k-1} + \left(\varepsilon_k^{(L)} + \kappa_k\mathbf{x}_k'\widehat{\boldsymbol{\omega}}_k^{(L)}\right)\mathbf{x}_k$

$\mathbf{B}_k = \beta\,\mathbf{B}_{k-1} + \kappa_k^{(L)}\mathbf{x}_k\mathbf{x}_k'$

Output: $\widehat{\boldsymbol{\omega}}_k := \widehat{\boldsymbol{\omega}}_k^{(L)}$

Algorithm 4 gives the summary of our proposed adaptive greedy algorithm procedure at time step k, which we refer to as *Sparse Greedy Point Process Filter (SGPPF)*. The algorithm recursively updates the score function as the proxy metric

(line 5), and updates the support set by selecting the $2S$ highest correlated components from the proxy denoted by function $\text{Supp}(\mathbf{p}_{\mathbf{k}}^{(\ell)}, 2S)$. It then merges the new components with the previous support set $\mathcal{S}_k^{(\ell)} := \text{Supp}(\widehat{\boldsymbol{\omega}}_k^{(\ell)})$, updates the restricted gradient (line 7), and performs a gradient ascent (line 8), followed by pruning to S largest set of components denoted by $\mathcal{S}_k^{(\ell+1)}$ (line 9 and 10). Finally, it updates $\mathbf{u}_k$ and $\mathbf{B}_k$ for next time step (line 12 and 13).

3.2.2 Regularization-based approach

In this subsection, we adopt an alternative approach by casting the adaptive sparse identification problem as an ML problem regularized by a sparsity-inducing penalty function. We develop a unified regularized likelihood-based framework for sparse identification of time-varying tuning features of the underlying neuronal model. We regularize the adaptive point process log-likelihood objective function in Eq. 3.3 by a general sparsity-inducing penalty function, and solve a regularized ML problem at each time step k as follows:

$$\widehat{\boldsymbol{\omega}}_k = \underset{\boldsymbol{\omega}_k}{\text{argmax}}\ \ell_k^{\beta}(\boldsymbol{\omega}_k) - \gamma \sum_{m=1}^{M} \mathcal{J}_R(|\omega_{k,m}|) \tag{3.20}$$

where $\gamma > 0$ is the regularization parameter and $\mathcal{J}_R(\cdot)$ denotes a separable sparsity-inducing regularization function typically in form of a non-smooth norm.

The commonly-used sparsity-inducing penalty is the ℓ_1-norm [27, 28]. The ℓ_1-norm penalty $\gamma\|\cdot\|_1$ penalizes all the parameters uniformly with a regularization level γ, leading to an overall shrinkage of the estimated parameters, and therefore biased estimates of the true parameters [90]. In [89], the authors prove an inherent

shortcoming of the ℓ_1-norm penalty which implies that consistent variable selection and optimal estimation cannot be attained simultaneously.

To resolve this issue, we extend the regularized likelihood framework of the ℓ_1-PPF adaptive filters developed earlier in Section 3.1 to a family of sparse adaptive filters which we refer to as *Regularized Likelihood Point Process Filters (RLPPF)*, by replacing the ℓ_1-norm penalty with specific sparsity-inducing penalty functions with optimal variable selection and prediction properties. In particular, we select a non-concave penalty function called *smoothly clipped absolute deviation (SCAD)* [90]. It is proven that this method enjoys the so-called oracle properties, namely it performs nearly as accurate as the genie-aided model where true sparse support of the parameters is known in advance. The SCAD penalty has a re-weighted-ℓ_1 form $\mathcal{J}_R(|\omega_{k,m}|) = \widehat{c}_{k,m}|\omega_{k,m}|$, which cleverly assigns non-uniform data-dependent weights $\widehat{c}_{k,m}$ to different components $\omega_{k,m}$, in a way that the out-of-support near-zero coefficients are penalized further (with larger weights) as compared to the in-support coefficients with larger absolute values. Unlike the uniform shrinkage effect of the ℓ_1-norm penalty, the shrinkage rule obtained from the SCAD penalty systematically sets the small parameter components to zero and returns nearly-unbiased values of the significant non-zero components. Following the techniques in [33], we use proximal algorithms to recursively solve the regularized ML problems in RLPPF with the SCAD penalty. Due to the separability of the SCAD penalty, it turns out that the resulting algorithm has the same computational complexity as that using the ℓ_1-norm penalty, while nearly achieving the optimal performance of the genie-aided estimator.

3.3 Applications

In this section, we apply the proposed algorithms to the simulated data as well as experimentally recorded spiking data from the ferret primary auditory cortex. In our simulation studies, we compare the performance of our proposed filters with two of the state-of-the-art point process filters, namely the steepest descent point process filter (SDPPF) [7] and the stochastic state point process filter (SSPPF) [14]. These adaptive filters are based on approximate solutions to the Chapman-Kolmogorov forward equation obtained by a steepest descent and a Gaussian approximation procedure, respectively.

3.3.1 Simulation Study 1: MSE and Sparse Recovery Learning Curves

First, we consider a stationary environment where $\boldsymbol{\omega}$ is constant over time. We use a bin size of $\Delta = 1\ ms$ and window size of $W = 1$ sample, for a total observation window of $\mathcal{T} = 30\ sec$ ($K = 30000$). The length of the parameter vector $\boldsymbol{\omega} = [\mu, \boldsymbol{\theta}]$ is chosen as $M = 101$. For each realization, we draw a sparse parameter vector $\boldsymbol{\theta}$ of fixed length $M - 1 = 100$ and sparsity $S = 3$. The support $\mathcal{S}$ and values of the nonzero components of $\boldsymbol{\theta}$ are chosen randomly and the values are normalized so that $\|\boldsymbol{\theta}\|_2 = 10$. The binary spike train $\{n_k\}_{k=1}^{K}$ is generated as a single realization of conditionally independent Bernoulli trials with success rate $\lambda_k\Delta$. The stimulus input sequence $\{s_k\}_{k=-M+1}^{K}$ is drawn from an i.i.d. Gaussian distribution $\mathcal{N}(0, \sigma^2)$. The stimulus variance is chosen as $\sigma^2 = 0.01$ small enough so that the average spiking rate $\bar{\lambda}\Delta = 0.13 \ll 1$ to ensure that the Bernoulli approximation is valid. All

the simulations are done with $L = 1$ iteration per time step. The step size is chosen as $\simeq 9 \times 10^{-4}$ (See Appendix A.2 for details).

For a given forgetting factor and step size, we select an optimal value for the regularization parameter γ by performing a two-fold even-odd cross validation procedure: first, the data are split into two sets of even and odd samples in an interleaved manner. Then, one set is used as the training set for estimation of the parameter vectors $\boldsymbol{\omega}_k$ and the other is used to assess the goodness-of-fit of the estimates $\widehat{\boldsymbol{\omega}}_k$ with respect to the log-likelihood of the observations. We repeat the process switching the role of the two sets and take the average as the overall measure of fit.

Let $\widehat{\mathbb{E}}$ denote the averaging operator with respect to realizations. We consider two performance metrics: the normalized mean squared error (MSE) defined as $\mathrm{MSE}_k := 10 \log_{10} \left(\widehat{\mathbb{E}} \|\widehat{\boldsymbol{\omega}}_k - \boldsymbol{\omega}_k\|^2 / \widehat{\mathbb{E}} \|\boldsymbol{\omega}_k\|^2 \right)$ to evaluate MSE performance in dB at time step k; and the out-of-support energy defined as $\mathrm{SPM}_k := \widehat{\mathbb{E}} \|\widehat{\boldsymbol{\theta}}_k - (\widehat{\boldsymbol{\theta}}_k)_S\|^2 / \mathbb{E} \|\widehat{\boldsymbol{\theta}}_k\|^2$ to represent a sparsity metric (SPM), where $(\widehat{\boldsymbol{\theta}}_k)_S$ denotes the restriction of $\widehat{\boldsymbol{\theta}}_k$ to the support S. Ideally, SPM_k must be equal to zero at all times. The averaging is carried out over a total of 1000 realizations, ensuring that the standard deviation of the ensemble following convergence is below 0.1 dB for all algorithms.

Figure 3.1 shows the corresponding learning curves for the four algorithms. According to Figure 3.1–A, the ℓ_1-PPF_1 achieves the lowest stationary MSE measure of -11.8 dB, followed ℓ_1-PPF_0 which achieves an MSE of -9.5 dB. The SSPPF and SDPPF algorithms respectively achieve an MSE of -2.7 dB and -1.9 dB, which

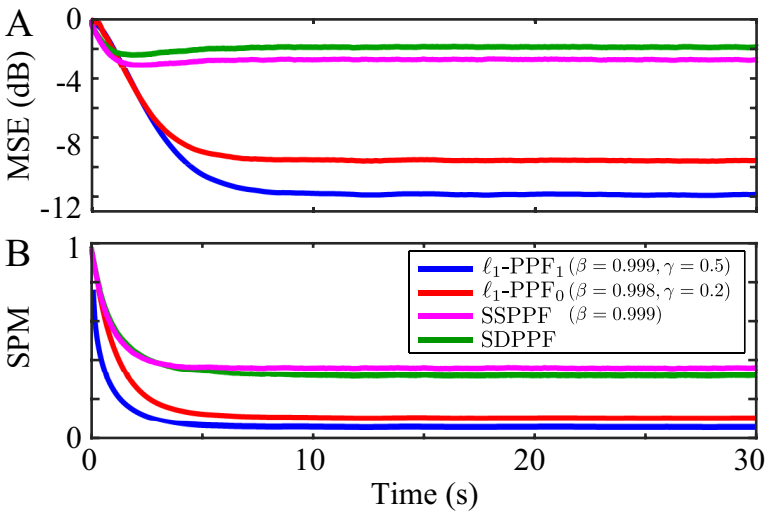

Figure 3.1: Learning curves of the adaptive filtering algorithms in a stationary environment. A) MSE vs. time, B) SPM vs. time.

reveals a gap of at least ≈ 8 dB with respect to our proposed filters.

3.3.2 Simulation Study 2: Tracking and Goodness-of-fit Performance

In the second simulation scenario, we consider a more realistic setting where $\boldsymbol{\omega}_k$ evolves in time. Furthermore, as in the case of real data applications, we assume that the support of $\boldsymbol{\omega}_k$ is not available as a performance benchmark and resort to statistical goodness-of-fit test. These tests for point process models have been developed as an application of the time-rescaling theorem [91, 92] and consist of the Kolmogorov-Smirnov (KS) test for assessing the conditional intensity estimation accuracy, and the Autocovariance Function (ACF) test to assess the conditional independence assumption. We skip the details, and refer the readers to the aforementioned references for a detailed treatment.

As in the previous case, we consider a bin size of $\Delta = 1ms$, window size of $W = 1$, and a total observation window of $\mathcal{T} = 60sec$ ($K = 60000$ bins). The stimulus is

generated as in the previous case. For the parameter vector $\boldsymbol{\omega}_k$, we choose a fixed baseline rate of $\mu_k = -2.51$ to set the baseline spiking rate to $\bar{\lambda}\Delta \approx 0.1$, and select a sparse modulation vector $\boldsymbol{\theta}_k$ of length $M = 100$ with a support $\mathcal{S} = \{1, 10, 20\}$ of size $S = 3$, and respective values of $(\boldsymbol{\theta}_k)_{\{1,10,20\}} = \{10, -5, 5\}$ for $k \leq K/2$. Halfway through the test, at $k = K/2 + 1$, the largest component $(\boldsymbol{\theta}_k)_1$, drops rapidly and linearly to 0, within a window of length 1 *sec* and remains zero for the rest of the run.

Figure 3.2 shows the performance of all four algorithms in the aforementioned setting. Each row (A through D) shows the true time-varying parameter vector (dashed traces) as well as the filtered estimates (solid traces) in the left panel. In particular, the gray solid traces show the out-of-support components which must ideally be equal to zero. The colored hulls around $(\widehat{\boldsymbol{\theta}}_k)_1$ show the 95% confidence intervals (note that confidence intervals for SDPPF cannot be directly obtained and require averaging over multiple realizations). The middle and right panels show the KS and ACF test results at a 95% confidence, respectively. For the quadratic algorithms ℓ_1-PPF$_1$ and SSPPF, a forgetting factor of $\beta = 0.9995$ is chosen. The regularization parameter for ℓ_1-PPF$_1$ is chosen as $\gamma = 1$, obtained by the aforementioned two-fold even-odd cross validation. For the zeroth order algorithm ℓ_1-PPF$_0$, a smaller forgetting factor of $\beta = 0.995$ is chosen to ensure stability, and a value of $\gamma = 0.1$ is used based on cross validation. The step size of $\simeq 5 \times 10^{-3}$ is chosen to be the same for both algorithms. These settings ensure that all the algorithms are tuned in their optimal operating point for fairness of comparison.

Figures 3.2–A and 3.2–B reveal three striking performance gaps between the

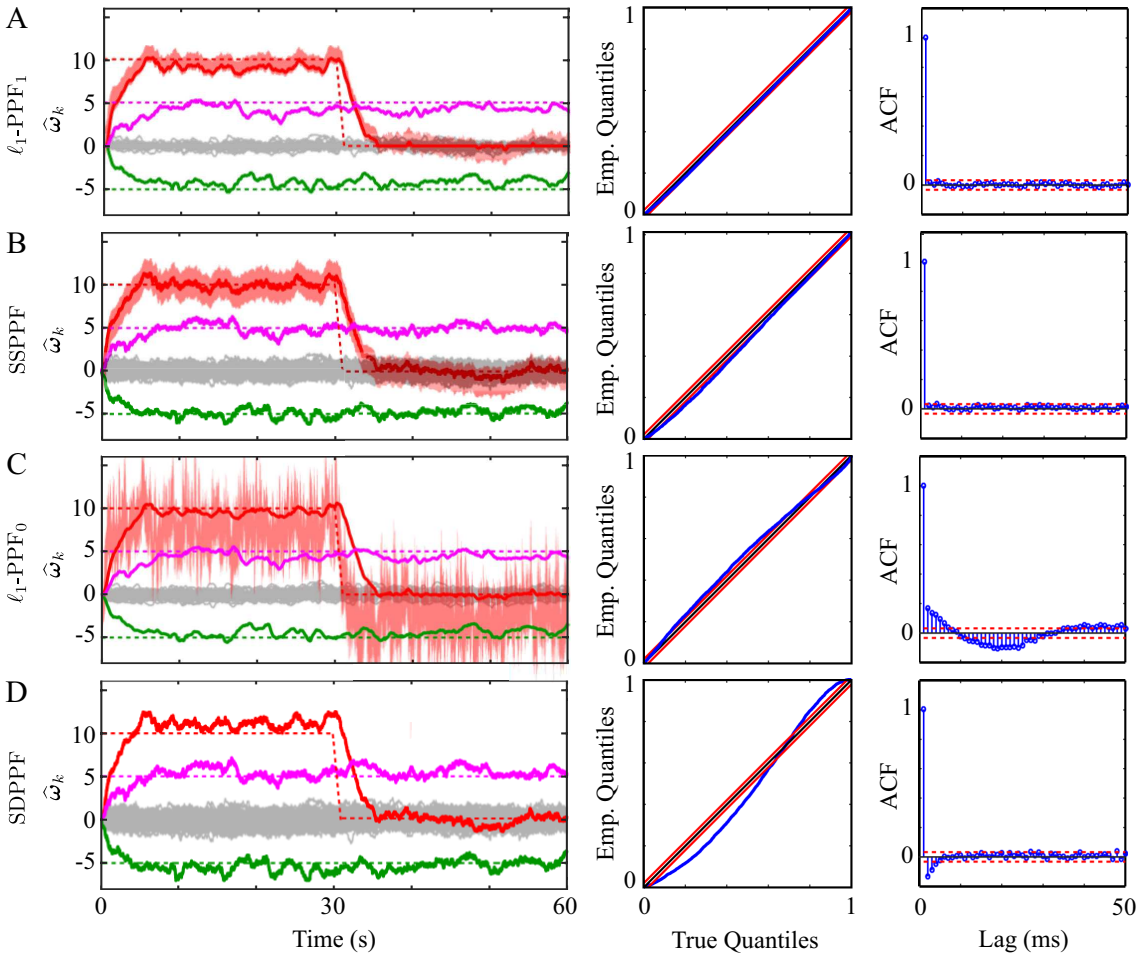

Figure 3.2: Performance comparison of the adaptive filtering algorithms: A) ℓ_1-PPF$_1$, B) SSPPF, C) ℓ_1-PPF$_0$, and D) SDPPF. In each row, the left panel shows the true parameter vector with dashed traces and the estimates with solid traces. Colored hulls show the 95% confidence intervals for one of the components. The middle and right panels show the corresponding KS and ACF test plots, respectively. Red traces show confidence regions at a level of 95% for both tests.

two second-order algorithms (with the same computational complexity, quadratic in M): first, the out-of-support components (gray traces) of ℓ_1-PPF$_1$ are significantly smaller than those of SSPPF; second, the confidence regions of ℓ_1-PPF$_1$ are narrower than those of SSPPF; and third, ℓ_1-PPF$_1$ fully passes the KS test, while SSPPF marginally does so. Similarly, comparing the two first-order algorithms (with the same computational complexity, linear in M) Figure 3.2–C and 3.2–D reveal that the ℓ_1-PPF$_0$ significantly suppresses the out-of-support components as compared to SDPPF. Moreover, ℓ_1-PPF$_0$ provides confidence bounds, which cannot be directly obtained for SDPPF. Finally, ℓ_1-PPF$_0$ marginally fails the KS test, whereas SDPPF does so significantly. Both algorithms fail the ACF test, which shows that the second-order corrections embedded in ℓ_1-PPF$_1$ and SSPPF are necessary to achieve a better goodness-of-fit, with a price of higher computational complexity.

We also inspect the estimated firing probability $\lambda_k(\widehat{\boldsymbol{\omega}}_k)\Delta$ for the four algorithms in Figure 3.3. In addition, we include the probability estimated by the normalized reverse correlation (NRC) method, which is commonly used in neural data analysis, and fits the modulation parameters using a linear model. Figure 3.3 shows the true spiking probability (blue solid trace) and the resulting spikes (black vertical lines). In the subsequent rows (B through F), the true and estimated probabilities are shown by dashed blue and solid red traces, respectively. A comparison of all the rows reveals that ℓ_1-PPF$_1$ and ℓ_1-PPF$_0$ outperform SSPPF and SDPPF, respectively, in terms of estimating the true probability. The NRC method is inferior to the preceding four algorithms, and results in negative estimates of the probability due to its use of a linear model (as opposed to logistic).

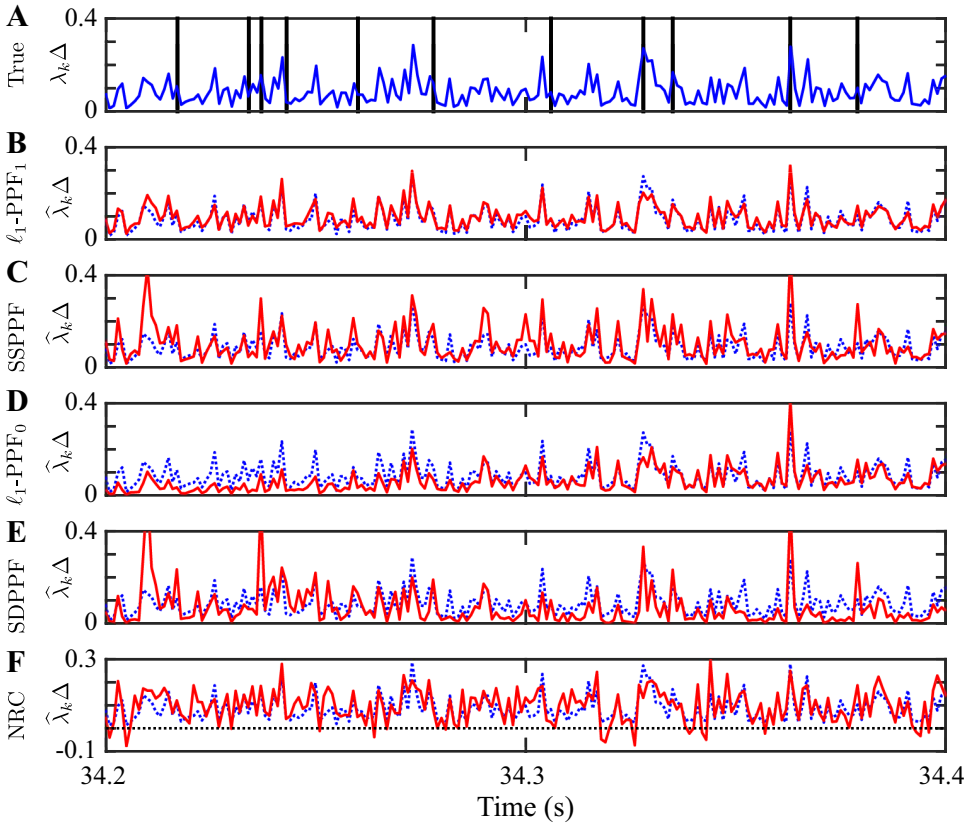

Figure 3.3: Firing rate estimates for adaptive filtering algorithms within an interval of 34.2 $s \leq t \leq$ 34.4 s: A) true rate (blue solid trace) and spikes (black vertical lines), B) ℓ_1-PPF$_1$, C) SSPPF, D) ℓ_1-PPF$_0$, E) SDPPF, and F) normalized reverse correlation (NRC). In rows B through F, the dashed blue traces and solid red traces show the true rate and the estimated rate, respectively.

3.3.3 Application to Real Data: Dynamic Analysis of Spectrotemporal Receptive Field Plasticity

The responses of neurons in the primary auditory cortex (A1) can be characterized by their spectrotemporal receptive fields (STRFs), where each neuron is tuned to a specific region in the time-frequency plane, and only significantly spikes when the acoustic stimulus contains spectrotemporal contents matching its tuning region [2] (See, for example, Figure 3.4, top row, leftmost panel). Several exper-

imental studies have revealed that receptive fields undergo rapid changes in their characteristics during attentive behavior in order to capture salient stimulus modulations [30, 93, 94]. In [30], it is suggested that this rapid plasticity has a significant role in the functional processes underlying active listening. However, most of the widely-used estimation techniques (e.g., normalized reverse correlation) provide static estimates of the receptive field with a temporal resolution of the order of minutes. Moreover, they do not systematically capture the inherent sparsity manifested in the receptive field characteristics.

In the context of our model, the STRF can be modeled as an $(I \times J)$-dimensional matrix, where I and J denote the number of time lags and frequency bands, respectively. By vectorizing this matrix, we obtain an $(M-1)$-dimensional vector $\boldsymbol{\theta}_k$ at window k, where $M = I \times J + 1$. Augmenting the baseline rate parameter μ_k, we can model the activity of the A1 neurons using the logistic CIF with a parameter $\boldsymbol{\omega}_k := [\mu_k, \boldsymbol{\theta}_k]'$. The stimulus vector at time t, $\mathbf{s}_t$ is given by the vectorized version of the spectrogram of the acoustic stimulus with J frequency bands and I lags. In order to capture the sparsity of the STRF in the time-frequency plane, we further represent $\boldsymbol{\theta}_k$ over a Gaussian time-frequency dictionary consisting of Gaussian windows centered around a regular subset of the $I \times J$ time-frequency plane. That is, for $\boldsymbol{\theta}_k = \mathbf{F}\boldsymbol{\xi}_k$, where $\mathbf{F}$ is the dictionary matrix and $\boldsymbol{\xi}_k$ is the sparse representation of the STRF. The estimation procedures of this chapter can be applied to $\boldsymbol{\xi}_k$, by absorbing the dictionary matrix into the data matrix $\mathbf{X}_k$ at window k.

We apply our proposed adaptive filter ℓ_1-PPF$_1$ to multi-unit spike recordings from the ferrets A1 during a series of passive listening conditions and active auditory

task conditions (data from the Neural Systems Laboratory, Institute for Systems Research, University of Maryland, College Park). During each active task, ferrets attended to the temporal dynamics of the sounds, and discriminated the rate of acoustic clicks [93]. The STRFs were estimated from the passive condition, where the quiescent animal listened to a series of broadband noise-like acoustic stimuli known as Temporally Orthogonal Ripple Combinations (TORC). The experiment consisted of 2 active and 11 passive blocks. Within each passive block, 30 TORCs were randomly repeated a total of 4-5 times each. In our analysis, we pool the spiking data corresponding to the same repeated TORC within each block. Therefore, the time axis corresponds to the experiment time modulo repetitions within each block. We discretize the resulting duration of $\mathcal{T} = 990s$ to time bins of size $\Delta = 1\ ms$, and segment data to windows of size $W = 10$ samples ($10\ ms$). The STRF dimensions are 50×50, regularly spanning lags of 1 to 50 ms and frequency bands of 0.5 kHz to 16 kHz (in logarithmic scale). The dictionary $\mathbf{F}$ consists of 13×13 Gaussian atoms, evenly spaced within the STRF domain. Each atom is a two-dimensional Gaussian kernel with a variance of $D^2/4$ per dimension, where D denotes the spacing between the atoms. We selected a forgetting factor of $\beta = 0.9998$, a step size of $\alpha = \frac{4(1-\beta)}{MW\bar{\sigma}^2}$, where $\bar{\sigma}^2$ is the average variance of the spectrogram components, $L = 1$ iteration per sample, and a regularization parameter of $\gamma = 40$ via two-fold even-odd cross validation.

Figure 3.4, top row, depicts five snapshots taken at $\{180, 360, 540, 630, 990\}$ sec corresponding to the end-points of the $\{2, 4, 6, 7, 11\}$th passive tasks. The bottom row shows the time-course of five selected points (marked as A through D in the

leftmost panel of the top row) of the STRF during the experiment. The STRF snapshots at times 180 and 540 *sec* correspond to 90 *secs* after the two active tasks, respectively, and verify the sharpening effect of the excitatory region ($\sim$ 30 $msec, 8\ kHz$) due to the animal's attentive behavior following the active task reported in [30]. Moreover, the STRF snapshots at times 360 and 630 *sec* reveal the weakening of the excitatory region long after the active task and returning to the pre-active state, highlighting the plasticity of A1 neurons. Previous studies have revealed the STRF dynamics with a resolution of the order of minutes [94]. Our result in Figure 3.4 provides a temporal resolution of the order of centiseconds (3 orders of magnitude increase), while capturing the STRF sparsity in a robust fashion.

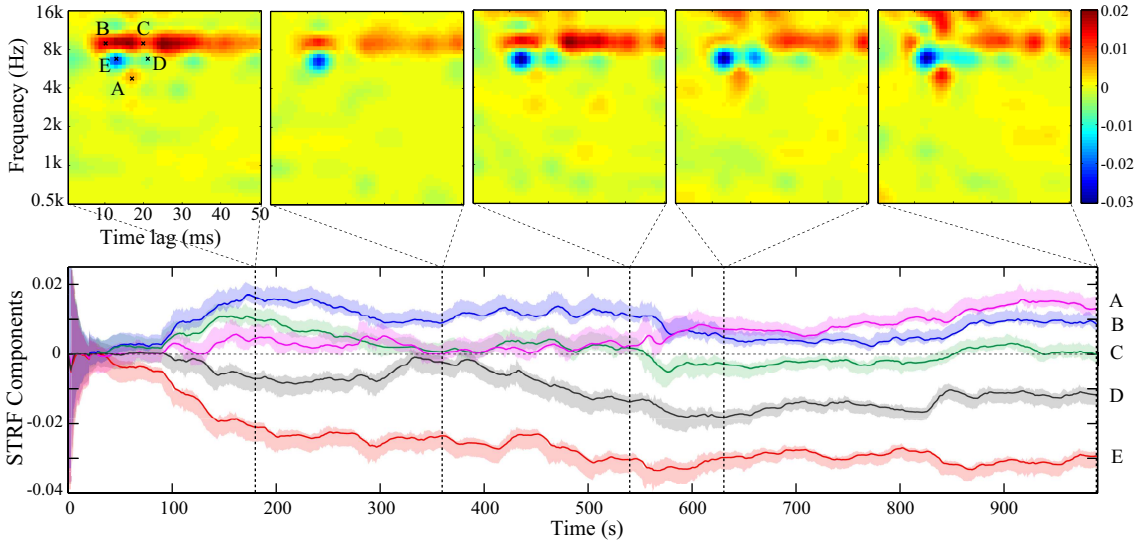

Figure 3.4: The time-course of task-dependent STRF plasticity of a ferret A1 neuron. The top row shows snapshots of the STRF at five selected points in time, marked by the dashed vertical lines in the bottom graph. The bottom graph shows the time-course of five selected points (A through E) in the STRF marked on the leftmost panel of the top row.

3.3.4 Application to Real Data: Sparse Adaptive Point Process Filters

We apply the proposed sparse adaptive filters in Section 3.2 to the same multi-unit spiking data set from the ferret A1 as the previous subsection. We binned the experiment duration $\mathcal{T} = 1017s$ to bins of size $\Delta = 1ms$, and segmented bins by windows of length $W = 5$. For the adaptive filtering setup, we selected the parameters $\beta = 0.9999$ and $\alpha = (1-\beta)/(200\bar{\kappa})$ to be equal for both greedy and regularized ML-based filters and set the $\gamma = 15$ for the RLPPF tuned by a two-fold even-odd cross validation.

Figure 3.5 depicts five snapshots of the estimated STRFs corresponding to

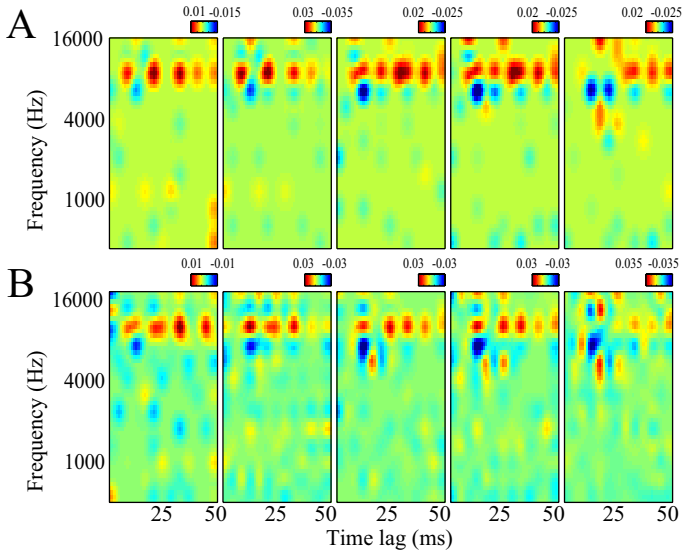

Figure 3.5: Adaptive sparse estimation of spectrotemporal receptive fields from the multi-unit spiking recordings. Each row depicts snapshots of the estimated STRFs for five selected time points during the experiment obtained by A) SGPPF, B) RLPPF with SCAD penalty.

five selected time points, obtained by SGPPF (Fig. 3.5-A) and RLPPF with SCAD penalty (Fig. 3.5-B). The first and second pair of STRFs correspond to respective pre-active and post-active conditions, and the last one is showing the long-term passive condition. Inspection of these figures reveal that both sparse adaptive filters effectively capture the sparsity manifested in the STRF domain while the adaptive nature of the estimates reveals the time-course of the receptive field plasticity. These results are consistent with the sharpening effect of low-latency features in temporal tasks [93], and provide a significant improvement in the temporal resolution of the STRF estimates.

Chapter 4: Inference of Neuronal Functional Network Dynamics via Adaptive Granger Causality Analysis

Studies of complex network systems, such as social networks, financial markets or the human brain, generally aim at understanding how the myriad components of a system interact together and collectively generate a macroscopic behavior. The problem of extracting and quantifying these functional interactions among network entities based on local observations is a key challenge in the study of complex networked systems. In particular, addressing this challenge is crucial in systems neuroscience, as it provides the opportunity to gain insights into how the brain, as one of the most mysterious and sophisticated systems in the universe, functions and coordinates behavior.

In this chapter, we present our theoretical results and algorithmic framework for modeling, estimation, and statistical inference for extracting functional neuronal network dynamics in the sense of Granger. First, we develop a dynamic measure of GC tailored for binary-natured neuronal spiking data recordings from an ensemble of sparsely interacting neurons. Later on, we describe a static variant of our GC inference framework with application to continuous-valued modalities of imaging data with linear Gaussian statistics.

4.1 Adaptive Granger Causality Inference from Ensemble Neuronal Spiking Activity

In this section, we present a new dynamic GC inference framework for neural spiking data, for which we integrate several techniques from adaptive filtering, compressed sensing, point process theory, and high-dimensional statistics. We use an exponential weighting scheme inspired by the RLS to design a recursive algorithm for computation of the dynamic GC measure in an online fashion. To assess the statistical significance of the new measure, we formulate a novel test statistic specifically tailored for our sparse dynamic setting, present theoretical results on its distribution, and further characterize the test strengths corresponding to the detected GC interactions.

4.1.1 The Adaptive Granger Causality (AGC) Measure

Consider simultaneous spike recordings from an ensemble of C neurons indexed by $c = 1, 2, \cdots, C$, denoted by $\{\{n_t^{(c)}\}_{t=1}^{T}\}_{c=1}^{C}$ over the time bins $t = 1, \cdots, T$. At time t, the spiking statistics of each neuron (c) are modeled via the CIF formulation of Eq. 2.8 using a sparse modulation parameter vector $\boldsymbol{\omega}_t^{(c)} = [\mu_t^{(c)}, \boldsymbol{\omega}_t^{(c,1)'}, \boldsymbol{\omega}_t^{(c,2)'}, \cdots, \boldsymbol{\omega}_t^{(c,C)'}, \boldsymbol{\theta}_t^{(c)}]'$ consisting of a scalar baseline firing parameter $\mu_t^{(c)}$, a collection of sparse history dependence parameter vectors $\{\boldsymbol{\omega}_t^{(c,\tilde{c})}\}_{\tilde{c}=1}^{C}$ of size M_H, in which $\boldsymbol{\omega}_t^{(c,\tilde{c})}$ represents the contribution of the spiking history of neuron $(\tilde{c})$ to the CIF of neuron (c), and $\boldsymbol{\theta}_t^{(c)}$ accounts for the stimulus modulation vector (e.g., receptive field), as we

had in the previous chapter. Let $h_{t,i}^{(c)} := \sum_{j=t-1-b_i}^{t-1-b_{i-1}} n_j^{(c)}$ be the spike count of neuron (c) within the i-th spike counting window of length $W_{H,i}$, where $b_i := \sum_{j=1}^{i} W_{H,j}$ for $i = 1, 2, \cdots, M_H$ and $b_0 = 0$. The covariates associated with the ensemble activity are given by $\mathbf{x}_t := [1, \mathbf{h}_t^{(1)\prime}, \mathbf{h}_t^{(2)\prime}, \cdots, \mathbf{h}_t^{(C)\prime}, \mathbf{s}_t']'$, where $\mathbf{h}_t^{(c)} := [h_{t,1}^{(c)}, h_{t,2}^{(c)}, \cdots, h_{t,M_H}^{(c)}]'$ denotes the history of spike counts of neuron (c) within non-overlapping windows of $W_H = [W_{H,1}, \ldots, W_{H,M_H}]$ up to a lag of $L_H := \sum_{i=1}^{M_H} W_{H,i}$, and $\mathbf{s}_t \in \mathbb{R}^{M_s}$ is the vector of neural stimuli in effect at bin t. We refer to this model, where the history of *all* the neurons in the ensemble are taken into account, as the *full model*. Fig. 4.1 shows an example of the neuronal ensemble and the corresponding covariates for $C = 3$.

In order to assess the G-causal influences, a likelihood-based GC measure has

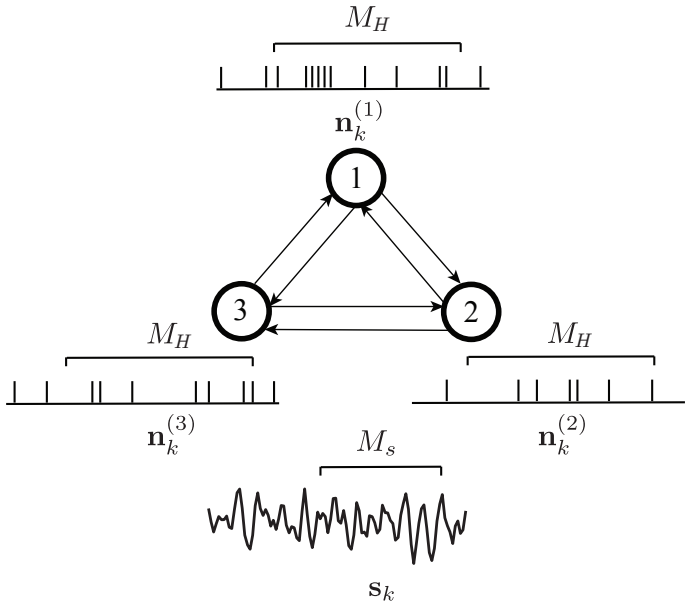

Figure 4.1: An example of the neuronal ensemble model for $C = 3$ neurons. The CIF of neuron (2) can be expressed as $\lambda_k^{(2)} = \text{logit}^{-1}\left(\mu_k^{(2)} + \boldsymbol{\omega}_k^{(2,1)\prime}\mathbf{n}_k^{(1)} + \boldsymbol{\omega}_k^{(2,2)\prime}\mathbf{n}_k^{(2)} + \boldsymbol{\omega}_k^{(2,3)\prime}\mathbf{n}_k^{(3)} + \boldsymbol{\theta}_k^{(2)\prime}\mathbf{s}_k\right)$.

been proposed in [74] for point process models. Consider neuron (c) as the target neuron with an observation vector $\mathbf{n}^{(c)} := [n_1^{(c)}, n_2^{(c)}, \cdots, n_K^{(c)}]'$. Let $\mathcal{H}^{(c)}$ denote the history of the covariates of neuron (c). The parameter vector and covariate history of neuron (c) after excluding the effect of neuron $(\tilde{c})$ are denoted by $\boldsymbol{\omega}_k^{(c\backslash\tilde{c})}$ and $\mathcal{H}^{(c\backslash\tilde{c})}$, respectively, and compose the so-called *reduced model*. The log-likelihood ratio statistic associated with the G-causal influence of neuron $(\tilde{c})$ on neuron (c) can be defined as:

$$\mathcal{F}_{(\tilde{c}\mapsto c)} := s\big(\widehat{\boldsymbol{\omega}}^{(c,\tilde{c})}\big) \log \frac{\mathcal{L}(\widehat{\boldsymbol{\omega}}^{(c)} |\, \mathbf{n}^{(c)}, \mathcal{H}^{(c)})}{\mathcal{L}(\widehat{\boldsymbol{\omega}}^{(c\backslash\tilde{c})} |\, \mathbf{n}^{(c)}, \mathcal{H}^{(c\backslash\tilde{c})})}, \tag{4.1}$$

where $\mathcal{L}(\widehat{\boldsymbol{\omega}}|\mathbf{n}, \mathcal{H})$ denotes the likelihood of estimated parameter vector $\widehat{\boldsymbol{\omega}}$ given the observation sequence $\mathbf{n}$ and the history of the covariates included in the model $\mathcal{H}$, and $s(\boldsymbol{\omega}) := \text{sign}(\sum_l \widehat{\omega}_l)$. Based on this formulation, the GC effect from neuron $(\tilde{c})$ to neuron (c) can be measured as the reduction in the point process log-likelihood of neuron (c) in the reduced model as compared with the full model. Note that the signum function determines the effective aggregate excitatory or inhibitory nature of this influence. This form of GC, conditioned on the mutual set of covariates (the spiking history of all other neurons in the ensemble) is referred to as *conditional Granger causality*, which allows to effectively distinguish between the direct and indirect causal interactions among an ensemble of simultaneously-acquired time series.

Most existing formulations of GC leverage the MVAR modeling framework [55–64, 66, 70], which pertains to data with linear Gaussian statistics. The GC measure in Eq. 4.1, however, benefits from the likelihood-based inference methodology and

covers a wide range of complex statistical models. Both the MVAR-based GC measure and its log-likelihood-based point process variant of [74] assume that the underlying time series are stationary, i.e., the modulation parameters are all static. In many scenarios of interest, however, the underlying dynamics exhibit a degree of non-stationarity, in which the underlying parameters change in time. An example of such a scenraio is the task-dependent receptive field plasicity phenomenon [7,30,33]. In addition, ML estimation used by these techniques does not capture the underlying sparsity of the parameters and often exhibits poor performance, when the data length is short or the number of neurons C is large.

In order to account for possible time-variability of the ensemble parameters and their underlying sparsity, we introduce the AGC measure, which is capable of capturing the dynamics of G-causal influences in the ensemble. To this end, we make two major modifications to the classical GC measure. First, we leverage the exponentially-weighted log-likelihood formulation of Eq. 3.3 to induce adaptivity into the GC measure. Second, we exploit the possible sparsity of the ensemble parameters using the sparse parameter estimates obtained through the ℓ_1-regularized ML procedure of Eq. 3.4 from Chapter 3. Replacing the standard data log-likelihoods in Eq. 4.1 by their sparse adaptive counterparts given in Eqs. 3.3 and 3.4, we define the AGC measure from neuron $(\tilde{c})$ to neuron (c) at time window k as:

$$\mathcal{F}_{k,\beta}^{(\tilde{c}\mapsto c)} := s_k(\widehat{\boldsymbol{\omega}}_k^{(c,\tilde{c})})\big(\ell_k^{\beta}(\widehat{\boldsymbol{\omega}}_k^{(c)}) - \ell_k^{\beta}(\widehat{\boldsymbol{\omega}}_k^{(c\setminus\tilde{c})})\big). \tag{4.2}$$

Although these modifications bring about crucial advantages in capturing the functional network dynamics in a robust fashion, they require construction of a statis-

tical inference framework in order for the proposed AGC measure to be useful. We address these issues in the forthcoming sections.

4.1.2 The AGC Inference Framework

Due to the stochastic and often biased nature of GC estimates, nonzero values of GC do not necessarily imply existence of G-causal influences. Hence, a statistical inference framework is required to assess the significance of the potential G-causal interactions extracted from the neural data.

Consider two nested GLM models, referred to as full and reduced models, with respective parameters $\boldsymbol{\omega}^{(F)} := \boldsymbol{\omega}^{(c)}$ and $\boldsymbol{\omega}^{(R)} = \boldsymbol{\omega}^{(c\setminus\tilde{c})}$, in which the latter is a special case of the former. In order to assess the statistical significance of a GC link, one can test for the null hypothesis $H_0 : \boldsymbol{\omega} = \boldsymbol{\omega}^{(R)}$ against the alternative $H_1 : \boldsymbol{\omega} = \boldsymbol{\omega}^{(F)}$. The test statistic often used for statistical inference of two nested models is referred to as the *deviance difference* of the two models and is defined as,

$$D(\widehat{\boldsymbol{\omega}}^{(F)}; \widehat{\boldsymbol{\omega}}^{(R)}) := 2\big(\ell(\widehat{\boldsymbol{\omega}}^{(F)}) - \ell(\widehat{\boldsymbol{\omega}}^{(R)})\big) \tag{4.3}$$

where $\ell(\cdot)$ is the log-likelihood and $\widehat{\boldsymbol{\omega}}^{(F)}$ and $\widehat{\boldsymbol{\omega}}^{(R)}$ denote the parameter estimates with the respective dimensions of $M^{(F)}$ and $M^{(R)}$ under the full and reduced models, respectively. The deviance difference for the likelihood-based GC is twice the right-hand-side of Eq. 4.1, modulo the signum function.

To perform the foregoing hypothesis test, the distributions of the deviance difference under both null and alternative hypotheses need to be characterized. The asymptotic distribution of the deviance difference statistic under both hypotheses

has been studied in the literature in the context of classical likelihood-ratio tests. It has been proven in [95, 96] that under certain regularity conditions, the deviance difference statistics asymptotically follow a chi-squared distribution with $M^{(d)} := M^{(F)} - M^{(R)}$ degrees of freedom, as the data length goes to infinity, when null hypothesis is true. Furthermore, under the same regularity conditions, the deviance difference has been proven to asymptotically converge in distribution to a non-central chi-squared with $M^{(d)}$ degrees of freedom and with a non-centrality parameter ν, under a sequence of local alternative hypotheses [97, 98].

The aforementioned classical distributional inference results cannot be readily extended to our AGC measure for two main reasons: first, the log-likelihoods are replaced by their exponentially-weighted counterparts, which suppresses their dependence on the data length N due to the forgetting factor mechanism. Second, unlike ML estimates which are asymptotically unbiased, the ℓ_1-regularized ML estimates are biased, and hence violate the common asymptotic normality assumptions.

In order to address these challenges, inspired by the recent results in high-dimensional regression [86, 87], we define the *adaptive de-biased deviance* as:

$$D_{k,\beta}(\widehat{\boldsymbol{\omega}}_k;\boldsymbol{\omega}_k) := \frac{1+\beta}{1-\beta}\Big(2\big(\ell_k^\beta(\widehat{\boldsymbol{\omega}}_k)-\ell_k^\beta(\boldsymbol{\omega}_k)\big)-\dot{\boldsymbol{\ell}}_k^\beta(\widehat{\boldsymbol{\omega}}_k)'\ddot{\boldsymbol{\ell}}_k^\beta(\widehat{\boldsymbol{\omega}}_k)^{-1}\dot{\boldsymbol{\ell}}_k^\beta(\widehat{\boldsymbol{\omega}}_k)\Big), \tag{4.4}$$

where $\dot{\boldsymbol{\ell}}_k^\beta(\cdot)$ and $\ddot{\boldsymbol{\ell}}_k^\beta(\cdot)$ are the gradient vector and Hessian matrix of the exponentially-weighted log-likelihood function $\ell_k^\beta(\cdot)$, and $\boldsymbol{\omega}_k$ and $\widehat{\boldsymbol{\omega}}_k$ denotes the true and estimated parameter vector at time window k, respectively. The adaptive de-biased deviance is composed of two main terms: the first term is twice the exponentially-weighted log-likelihood ratio statistic, which is analogous to the standard deviance difference,

whereas the second is a bias correction term. The bias correction term compensates for the effect of the ℓ_1-regularization bias imposed in favor of enforcing sparsity in the estimate $\widehat{\boldsymbol{\omega}}_k$. The effect of forgetting factor mechanism appears in the form of the scaling $(1+\beta)/(1-\beta)$. Note that the bias term has a quadratic form and would vanish, if the log-likelihoods were evaluated at the ML estimates, since $\dot{\boldsymbol{\ell}}_k^{\beta}(\widehat{\boldsymbol{\omega}}_k^{\mathsf{ML}}) = 0$. Finally, we define a test statistic for AGC inference referred to as *adaptive de-biased deviance difference* as follows:

$$D_{k,\beta}^{(\tilde{c} \,\mapsto\, c)} := D_{k,\beta}(\widehat{\boldsymbol{\omega}}_k^{(c)}; \boldsymbol{\omega}_k^{(c)}) - D_{k,\beta}(\widehat{\boldsymbol{\omega}}_k^{(c\setminus\tilde{c})}; \boldsymbol{\omega}_k^{(c\setminus\tilde{c})}). \tag{4.5}$$

In the following, we will mainly work with $D_{k,\beta}^{(\tilde{c} \,\mapsto\, c)}$, as opposed to its biased version given by $\mathcal{F}_{k,\beta}^{(\tilde{c}\mapsto c)}$ in Eq. 4.2. Note that $\mathcal{F}_{k,\beta}^{(\tilde{c}\mapsto c)} = \frac{1}{2} s_k(\widehat{\boldsymbol{\omega}}_k^{(c,\tilde{c})}) \Big(\frac{D_{k,\beta}^{(\tilde{c}\mapsto c)}}{1+\beta} + B_{k,\beta}^{(\tilde{c}\mapsto c)} \Big)$, where $B_{k,\beta}^{(\tilde{c}\mapsto c)}$ is the difference of the bias terms of the full and reduced models.

In what follows, we develop the AGC inference procedure in four major stages: (1) efficient computation of $D_{k,\beta}^{(\tilde{c} \,\mapsto\, c)}$ from the data in recursive form, (2) distributional inference of $D_{k,\beta}^{(\tilde{c} \,\mapsto\, c)}$ under both the absence and presence of a GC link, (3) the false discovery rate control procedure, and (4) the statistical significance assessment of the detected GC links. Fig. 4.2 shows a schematic depiction of the overall inference procedure, which we will discuss next.

(1) Recursive Computation of the AGC Measure: The computation of the adaptive de-biased deviance differences $D_{k,\beta}^{(\tilde{c} \,\mapsto\, c)}$ for all the possible $|\mathcal{C}|$ links and at all times k is required for our statistical analysis. Therefore, in order for the analysis to scale favorably with the network size C and the data length K, it is crucial to develop an efficient framework for the computation of the AGC measure.

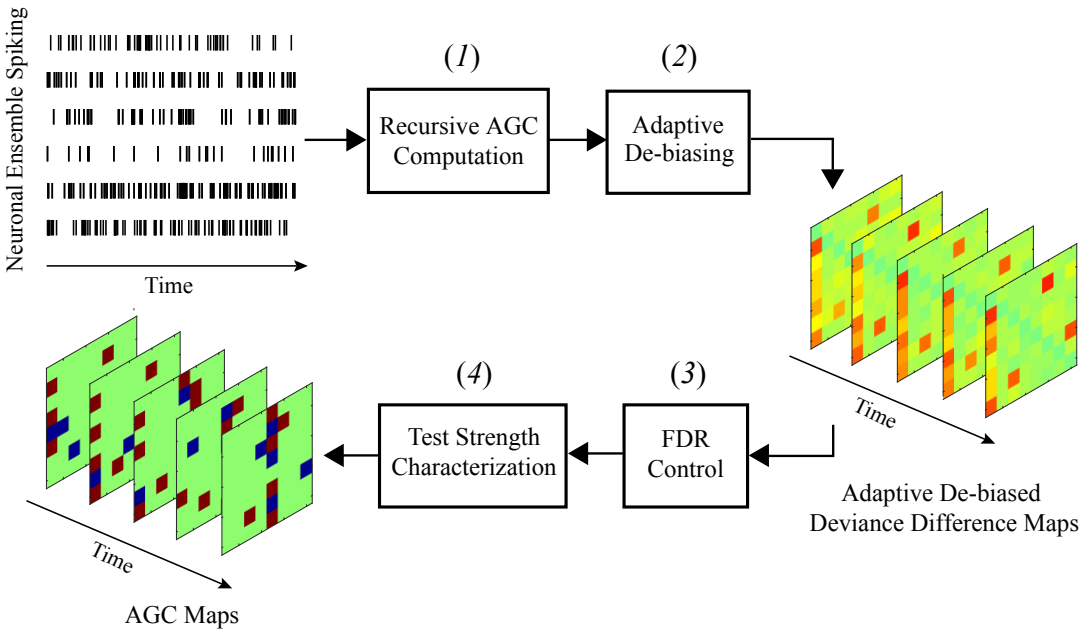

Figure 4.2: Schematic depiction of the inference procedure for the AGC measure.

The RLS-inspired exponential weighting of the log-likelihoods in Eq. 3.3 indeed paves the way for the recursive computation of the AGC measure. We design low-complexity recursive update rules for computation of $\ell_k^{\beta}(\widehat{\boldsymbol{\omega}}_k)$ for a generic estimate $\widehat{\boldsymbol{\omega}}_k$, from which the AGC measure of Eq. 4.2 can be computed. This step comprises the *Recursive AGC Computation* block in Fig. 4.2.

In order to achieve recursive computation, we exploit the smoothness of the point process log-likelihood function, and approximate each scalar-valued log-likelihood function $\ell_i(\widehat{\boldsymbol{\omega}}_k)$ using a second order Taylor's series expansion around $\widehat{\boldsymbol{\omega}}_i$ for $i \leq k$. Retaining the first three terms of the expansion yields:

$$\ell_i(\widehat{\boldsymbol{\omega}}_k) \approx \ell_i(\widehat{\boldsymbol{\omega}}_i) + (\widehat{\boldsymbol{\omega}}_k - \widehat{\boldsymbol{\omega}}_i)'\dot{\boldsymbol{\ell}}_i(\widehat{\boldsymbol{\omega}}_i) + \frac{1}{2}(\widehat{\boldsymbol{\omega}}_k - \widehat{\boldsymbol{\omega}}_i)'\ddot{\boldsymbol{\ell}}_i(\widehat{\boldsymbol{\omega}}_i)(\widehat{\boldsymbol{\omega}}_k - \widehat{\boldsymbol{\omega}}_i), \tag{4.6}$$

where $\dot{\boldsymbol{\ell}}_i(\cdot)$ and $\ddot{\boldsymbol{\ell}}_i(\cdot)$ denote the gradient vector and Hessian matrix with respect to

$\boldsymbol{\omega}$, which can be computed from Eq. 3.2 for the logit-linked GLM model as follows:

$$\dot{\boldsymbol{\ell}}_i(\widehat{\boldsymbol{\omega}}_i) = \mathbf{X}_i'\boldsymbol{\varepsilon}_i, \tag{4.7}$$

$$\ddot{\boldsymbol{\ell}}_i(\widehat{\boldsymbol{\omega}}_i) = -\mathbf{X}_i'\boldsymbol{\Lambda}_i\mathbf{X}_i, \tag{4.8}$$

where $\boldsymbol{\varepsilon}_i := \mathbf{n}_i - \boldsymbol{\lambda}_i(\widehat{\boldsymbol{\omega}}_i)\Delta$ denotes the point process innovation vector at time window i, and $\boldsymbol{\Lambda}_i := \text{diag}\left(\boldsymbol{\lambda}_i\Delta \odot (1 - \boldsymbol{\lambda}_i\Delta)\right)$ is a $W \times W$ diagonal matrix with $(\boldsymbol{\Lambda}_i)_{m,m} := \lambda_{(i-1)W+m}(\widehat{\boldsymbol{\omega}}_i)\Delta(1 - \lambda_{(i-1)W+m}(\widehat{\boldsymbol{\omega}}_i)\Delta)$ as the m-th diagonal element obtained from the second-order derivative of the logistic log-likelihood function. Substituting the quadratic Taylor's approximation of Eq. 4.6 into Eq. 3.2 and rearranging terms will lead to the following recursive update rule for the adaptive log-likelihoods at time step k:

$$\ell_k^{\beta}(\widehat{\boldsymbol{\omega}}_k) = a_k + \widehat{\boldsymbol{\omega}}_k'\mathbf{u}_k + \frac{1}{2}\,\widehat{\boldsymbol{\omega}}_k'\mathbf{B}_k\widehat{\boldsymbol{\omega}}_k, \tag{4.9}$$

where

$$\begin{aligned}
a_k &= \sum_{i=1}^{k} \beta^{k-i}\big(\mathbf{1}_W{}'\boldsymbol{\ell}_i(\widehat{\boldsymbol{\omega}}_i) - \widehat{\boldsymbol{\omega}}_i'\mathbf{X}_i'\boldsymbol{\varepsilon}_i - \frac{1}{2}\widehat{\boldsymbol{\omega}}_i'\mathbf{X}_i'\boldsymbol{\Lambda}_i\mathbf{X}_i\widehat{\boldsymbol{\omega}}_i\big), \\
\mathbf{u}_k &= \sum_{i=1}^{k} \beta^{k-i}\mathbf{X}_i'\big(\boldsymbol{\varepsilon}_i + \boldsymbol{\Lambda}_i\mathbf{X}_i\widehat{\boldsymbol{\omega}}_i\big), \\
\mathbf{B}_k &= -\sum_{i=1}^{k} \beta^{k-i}\mathbf{X}_i'\boldsymbol{\Lambda}_i\mathbf{X}_i,
\end{aligned} \tag{4.10}$$

in which $\boldsymbol{\ell}_i(\widehat{\boldsymbol{\omega}}_i) := [\ell_{(i-1)W+1}(\widehat{\boldsymbol{\omega}}_i), \ldots, \ell_{iW}(\widehat{\boldsymbol{\omega}}_i)]'$ denotes the vector of log-likelihoods corresponding to the i-th time window, and $\mathbf{1}_W := [1, \ldots, 1]'$ is the vector of all ones of length W. It is easy to see that a_k, $\mathbf{u}_k$ and $\mathbf{B}_k$ also admit recursive update rules

at time step k:

$$
\begin{aligned}
a_k &= \beta\, a_{k-1} + \mathbf{1}_W{}'\boldsymbol{\ell}_k(\widehat{\boldsymbol{\omega}}_k) - \widehat{\boldsymbol{\omega}}_k'\mathbf{X}_k'\boldsymbol{\varepsilon}_k - \frac{1}{2}\widehat{\boldsymbol{\omega}}_k'\mathbf{X}_k'\boldsymbol{\Lambda}_k\mathbf{X}_k\widehat{\boldsymbol{\omega}}_k, \\
\mathbf{u}_k &= \beta\, \mathbf{u}_{k-1} + \mathbf{X}_k'\left(\boldsymbol{\varepsilon}_k + \boldsymbol{\Lambda}_k\mathbf{X}_k\widehat{\boldsymbol{\omega}}_k\right), \\
\mathbf{B}_k &= \beta\, \mathbf{B}_{k-1} - \mathbf{X}_k'\boldsymbol{\Lambda}_k\mathbf{X}_k.
\end{aligned} \tag{4.11}
$$

By performing the recursive computation of Eq. 4.9 for both the full model and the reduced model, a fully recursive update procedure for the AGC measure of Eq. 4.2 is obtained, which enables us to track the G-causal interactions among the neurons in an online fashion. This fully recursive procedure can be further extended to our proposed statistical inference framework based on the de-biased deviance statistics. To this end, we obtain a recursive update rule for the quadratic bias terms in Eq. 4.4. The recursion for the score statistic evaluated at the current estimate, $\dot{\boldsymbol{\ell}}_k^{\beta}(\widehat{\boldsymbol{\omega}}_k)$, is readily available through a similar treatment using the Taylor's series expansion and is employed in the ℓ_1–PPF$_1$ filtering procedure for estimating the maximizers of ℓ_1-regularized ML problems recursively in Eq. 3.15 of Chapter 3. This update rule simplifies to:

$$
\dot{\boldsymbol{\ell}}_k^{\beta}(\widehat{\boldsymbol{\omega}}_k) = \mathbf{u}_k + \mathbf{B}_k\widehat{\boldsymbol{\omega}}_k. \tag{4.12}
$$

The inverse Hessians $\ddot{\boldsymbol{\ell}}_k^{\beta}(\widehat{\boldsymbol{\omega}}_k)^{-1}$ can also be efficiently computed via the Woodbury matrix identity applied to the update rule of the quadratic bias term from Eq. 4.4. When the Hessians are not invertible, a recursive implementation of the node-wise regression procedure of [87] can be used, which is developed in [31] using the SPARLS iteration [21] for RLS-type exponentially weighted log-likelihoods, as

Algorithm 5 Recursive update rule for $\ell_k^{\beta}(\widehat{\boldsymbol{\omega}}_k)$

Inputs: $\mathbf{n}_k$, $\mathbf{X}_k$, $\widehat{\boldsymbol{\omega}}_k$, a_{k-1}, $\mathbf{u}_{k-1}$, and $\mathbf{B}_{k-1}$.

$\mathbf{y}_k = \mathbf{X}_k \widehat{\boldsymbol{\omega}}_k$

$\boldsymbol{\lambda}_k \Delta = \text{logit}^{-1}(\mathbf{y}_k)$

$\boldsymbol{\varepsilon}_k = \mathbf{n}_k - \boldsymbol{\lambda}_k \Delta$

$\boldsymbol{\Lambda}_k = \text{diag}\left(\boldsymbol{\lambda}_k \Delta \odot (1 - \boldsymbol{\lambda}_k \Delta)\right)$

$a_k = \beta\, a_{k-1} + \mathbf{1}_W{}' \boldsymbol{\ell}_k(\widehat{\boldsymbol{\omega}}_k) - \mathbf{y}_k' \boldsymbol{\varepsilon}_k - \frac{1}{2}\mathbf{y}_k' \boldsymbol{\Lambda}_k \mathbf{y}_k$

$\mathbf{u}_k = \beta\, \mathbf{u}_{k-1} + \mathbf{X}_k' \left(\boldsymbol{\varepsilon}_k + \boldsymbol{\Lambda}_k \mathbf{y}_k\right)$

$\mathbf{B}_k = \beta\, \mathbf{B}_{k-1} - \mathbf{X}_k' \boldsymbol{\Lambda}_k \mathbf{X}_k$

Output: $\ell_k^{\beta}(\widehat{\boldsymbol{\omega}}_k) = a_k + \widehat{\boldsymbol{\omega}}_k' \mathbf{u}_k + \frac{1}{2}\widehat{\boldsymbol{\omega}}_k' \mathbf{B}_k \widehat{\boldsymbol{\omega}}_k$

presented in Appendix A.3. Algorithm 5 summarizes the recursive computation of the exponentially-weighted log-likelihoods at window k.

(2) Asymptotic Distributional Analysis of the AGC measure: We next present our main theoretical result, which extends the asymptotic inference results of the classical deviance difference statistic to our adaptive de-biased variant:

Theorem 4.1 *Consider simultaneous spike train observations* $\{\{n_t^{(c)}\}_{t=1}^{T}\}_{c=1}^{C}$ *from an ensemble of* C *neurons. Let* $\widehat{\boldsymbol{\omega}}_k^{(c)}$ *and* $\widehat{\boldsymbol{\omega}}_k^{(c\setminus\tilde{c})}$ *denote the estimated sparse parameter vectors of neuron* (c) *at time window* k *in two nested logit-linked point process GLM models, where the contribution of neuron* $(\tilde{c})$ *is suppressed in the latter. Suppose that the adaptive estimation is carried out through solving the* ℓ_1*-regularized ML problem of Eq. 3.4 at time window* k*. Then,*

i) in the absence of a GC link from $(\tilde{c})$ *to* (c)*, we have* $D_{k,\beta}^{(\tilde{c} \,\mapsto\, c)} \to \chi^2\big(M^{(d)}\big)$*, and*

ii) in the presence of a GC link from $(\tilde{c})$ *to* (c)*, and assuming that the cross-history coefficients from* $(\tilde{c})$ *to* (c) *scale at least as* $\mathcal{O}\left(\sqrt{\frac{1-\beta}{1+\beta}}\right)$*, then* $D_{k,\beta}^{(\tilde{c} \,\mapsto\, c)} \to$

$$\chi^2\big(M^{(d)}, \nu_k^{(\tilde{c}\mapsto c)}\big),$$

as $\beta \to 1$, *where* $M^{(d)} := M^{(F)} - M^{(R)}$ *is the dimensionality difference of the two nested models, and* $\nu_k^{(\tilde{c}\ \mapsto\ c)} > 0$ *is the corresponding non-centrality parameter and is only a function of the true model parameter of neuron* (c) *at time* k.

Proof 4.1 *The proof is given in Appendix B.*

Remarks. Theorem 4.1 has two major implications. First, it establishes that our proposed adaptive de-biased deviance difference statistic admits simple asymptotic distributional characterization. Given that these asymptotic distributions form the main ingredients of the forthcoming inference procedure, the second block in Fig. 4.2 serves to highlight the significance of adaptive de-biasing. The output of the second block is the de-biased deviance differences corresponding to all pairs of neurons (shown in 2D as deviance difference maps).

Second, given that for $\nu_k^{(\tilde{c}\mapsto c)} = 0$, the non-central chi-squared distribution coincides with the chi-squared distribution, the non-centrality parameter plays a key role in separating the distributions under the null and alternative hypotheses: when the deviance difference is close to zero, the null hypothesis H_0 is likely to be true, i.e., no GC link. When the deviance difference is large, the alternative H_1 is likely to be true, i.e., a GC link exists (See Remark 2 in the Appendix B for further discussion). The non-centrality parameter $\nu_k^{(\tilde{c}\mapsto c)}$, however, is a complicated function of the true values of the parameters, and can not be directly observed. In what follows, we initially assume that an estimate $\widehat{\nu}_k^{(\tilde{c}\mapsto c)}$ is at hand, and later on

derive an algorithm for its estimation. Moreover, we will show how to translate the deviance differences to statistically interpretable AGC links.

(3) False Discovery Rate Control: We next describe a statistical inference procedure for simultaneous assessment of the statistical significance of all possible GC interactions among the neurons in the ensemble. In the multiple hypothesis testing problem, a group of interconnected null hypotheses are tested simultaneously, where the probability that at least one true null would be rejected (joint false positive) can increase considerably.

Several solutions have been proposed to handle this problem such as the well-known Bonferroni correction [99, 100], where the probability of incorrectly rejecting at least one null among all the hypotheses (also referred to as *family-wise error rate (FWER)*) is controlled. Here, we take an alternative approach given by the Benjamini-Yekutieli (BY) procedure [101], which is proved to be among the most effective solutions to the multiple testing problem. The BY procedure aims at controlling the false discovery rate (FDR), which is the expected ratio of incorrectly rejected null hypotheses or namely "false discoveries", at a desired significance level α.

We use part (i) of the result of Theorem 4.1 in order to control the FDR in a multiple hypothesis testing framework. In order to identify significant GC interactions while avoiding spurious false positives, we conduct multiple hypothesis tests on the set of $|\mathcal{C}| := C \times (C-1)$ pairwise possible GC interactions $\mathcal{C} := \{(\tilde{c} \mapsto c) \,|\, \tilde{c}, c = 1, \ldots, C,\ c \neq \tilde{c}\}$ among the ensemble of C neurons at each time step k. The

null hypothesis $H_{0,k}^{(\tilde{c}\mapsto c)}$ corresponds to lack of a GC link from neuron $(\tilde{c})$ to (c) at time step k. Thus, rejection of the null hypothesis amounts to discovering a GC link $(\tilde{c} \mapsto c)$ at time step k. We first compute $D_{k,\beta}^{(\tilde{c} \mapsto c)}$ for all possible links in $\mathcal{C}$. Based on Theorem 4.1, under null hypothesis $H_{0,k}^{(\tilde{c} \mapsto c)}$, we have $D_{k,\beta}^{(\tilde{c} \mapsto c)} \to \chi^2(M^{(d)})$ as $\beta \to 1$. Hence, by virtue of convergence in distribution, for β close to 1, thresholding the test statistic results in a consistent approximation to limiting the false positive rate, namely type I error: the null hypothesis $H_{0,k}^{(\tilde{c} \mapsto c)}$ is rejected at a confidence level of $1-\alpha$, if $D_{k,\beta}^{(\tilde{c} \mapsto c)} > \mathrm{F}^{-1}_{\chi^2(M^{(d)})}(1-\alpha)$, where $\mathrm{F}^{-1}_{\chi^2(M^{(d)})}(\cdot)$ is the inverse CDF of a χ^2 distribution with $M^{(d)}$ degrees of freedom. Using the BY procedure, we can thus control the mean FDR at a rate of $\bar{\alpha} := \dfrac{(|\mathcal{C}|+1)\alpha}{2|\mathcal{C}|\log|\mathcal{C}|}$ for all tests. This stage forms the *FDR Control* block in Fig. 4.2.

(4) Test Strength Characterization via J-statistic: Next, we use part (ii) of the result of Theorem 4.1 to assess the significance of the tests for the detected GC links. Under the alternative hypothesis, Theorem 4.1 implies that $H_{1,k}^{(\tilde{c}\mapsto c)} : D_{k,\beta}^{(\tilde{c} \mapsto c)} \to \chi^2\big(M^{(d)}, \nu_k^{(\tilde{c}\mapsto c)}\big)$ as $\beta \to 1$. Hence, by virtue of convergence in distribution, the false negative rate, namely the type II error can be estimated by $\eta_k^{(\tilde{c} \mapsto c)} := \mathrm{F}_{\chi^2(M^{(d)},\widehat{\nu}_k)}\big(\mathrm{F}^{-1}_{\chi^2(M^{(d)})}(1-\alpha)\big)$, at a confidence level of $1-\alpha$, where $\mathrm{F}_{\chi^2(M^{(d)},\widehat{\nu}_k)}(\cdot)$ represents the CDF of a non-central χ^2 distribution with $M^{(d)}$ degrees of freedom and the estimate $\widehat{\nu}_k$ of the corresponding non-centrality parameter $\nu_k^{(\tilde{c} \mapsto c)}$. It can be seen that the true positive rate defined as $1-\eta_k^{(\tilde{c} \mapsto c)}$ is increasing in $\widehat{\nu}_k$, due to the monotonically decreasing property of the non-central χ^2 CDF function with respect to the non-centrality parameter. In other words, the

larger $\widehat{\nu}_k$ takes values, the closer to one the true positive rate will be. Hence, the non-centrality parameter $\widehat{\nu}_k$ can be interpreted as an implicit measure of test power for underlying GC link.

In order to quantify the significance of an estimated GC link, we utilize the Youden's *J-statistic*, which is an effective measure often used for summarizing the overall performance of a diagnostic test. The J-statistic in our setting is given by:

$$J_k^{(\tilde{c} \mapsto c)} := 1 - \alpha - \mathrm{F}_{\chi^2(M^{(d)},\widehat{\nu}_k)}\left(\mathrm{F}^{-1}_{\chi^2(M^{(d)})}(1-\alpha)\right), \tag{4.13}$$

for a fixed significance level α. Note that the J-statistic can take values in $[0, 1]$. The case of $J_k^{(\tilde{c} \mapsto c)}$ being close to one represents high sensitivity and specificity of the test statistic, which coincides with large values of non-centrality. One advantage of the J-statistic over the conventional p-value is that it accounts for both type I and type II errors. In the context of GC analysis, the J-statistic for each possible link can serve as a normalized indicator of how reliable the detected link is. For consistency, we assign a value of $J_k^{(\tilde{c} \mapsto c)} = 0$, when the null hypothesis is not rejected. The foregoing statistical test strength characterization forms the last block of Fig. 4.2.

Fig. 4.3 demonstrates the hypothesis testing framework and the FDR control procedure, and summarizes the quantities involved. Fig. 4.3-A illustrates the hypothesis testing by showing the distributions under null H_0 and alternative H_1, and the areas corresponding to type I and type II errors, given a confidence level $1 - \alpha$. Fig. 4.3-B exhibits the receiver operating characteristic (ROC) curves for different values of $(M^{(d)}, \nu)$, as well as how the J-statistic is calculated for $\alpha = 0.05$.

Algorithm 6 summarizes the FDR control procedure based on BY rule along

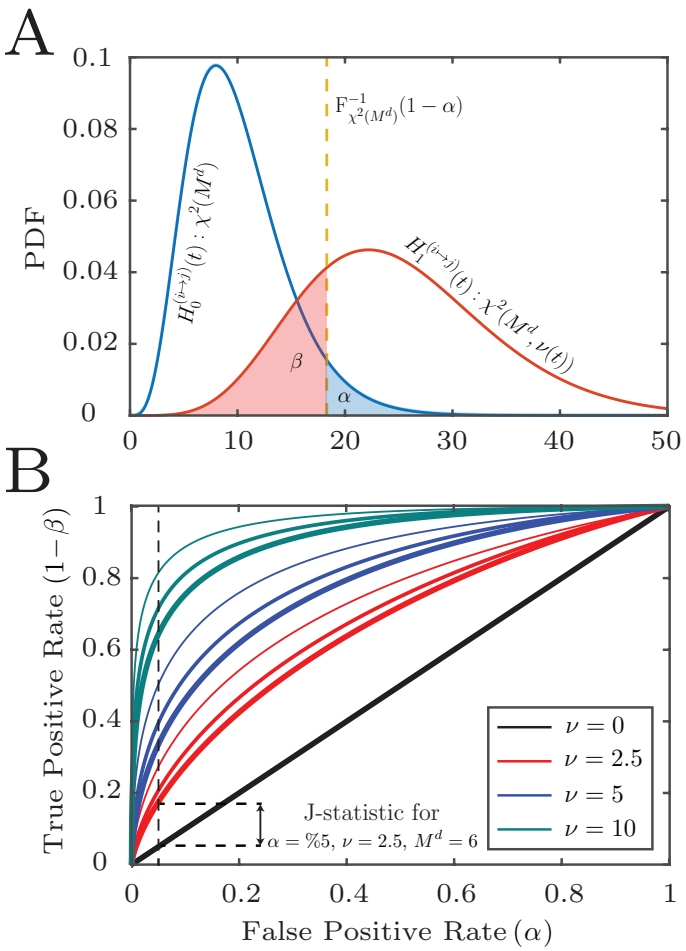

Figure 4.3: Illustration of the hypothesis testing framework and the FDR control procedure. A) PDFs of H_0 and H_1 for $M^{(d)} = 10$, $\nu = 15$, and $\alpha = 0.05$, B) ROC curves for different values of $M = \{2\text{ (narrow)}, 4\text{ (medium)}, 6\text{ (thick)}\}$ and $\nu = \{0\text{ (black)}, 2.5\text{ (red)}, 5\text{ (blue)}, 10\text{ (green)}\}$.

with the test strength characterization via J-statistics. Given the estimates of deviance differences $D_{k,\beta}^{(\tilde{c}\mapsto c)}$ and the non-centrality parameters $\nu_k^{(\tilde{c}\mapsto c)}$, the G-causal links can be detected at a fixed FDR α, and their corresponding test strengths can be assessed via the J-statistics computed at the mean FDR $\bar{\alpha}$.

It remains to estimate the unknown non-centrality parameters $\nu_k^{(\tilde{c}\mapsto c)}$ given the observed deviance differences $D_{k,\beta}^{(\tilde{c}\mapsto c)}$. Under the assumption that $\nu_k^{(\tilde{c}\mapsto c)}$ changes

smoothly in time, this can be carried out efficiently using a non-central χ^2 filtering and smoothing algorithm, which is discussed next.

Algorithm 6 BY FDR Control and Characterizing the J-statistics

Input: $\Big\{\{D_{k,\beta}^{(\tilde{c}\mapsto c)}, \widehat{\nu}_k^{(\tilde{c}\mapsto c)}\}_{=1}^{K} \,|\, (\tilde{c} \mapsto c) \in \mathcal{C}\Big\}$, $M^{(d)}$, and α.

1: **for** $k = 1, 2, \ldots, K$ **do**
2: **for** $(\tilde{c} \mapsto c) \in \mathcal{C}$ **do**
3: Define p-values $p_k^{(\tilde{c}\mapsto c)} := 1 - \mathrm{F}_{\chi^2(M^{(d)})}\left(D_k^{(\tilde{c}\mapsto c)}\right)$
4: **end for**
5: Sort the calculated p-values as $p_k^{(m_1)} \leq p_k^{(m_2)} \leq \cdots \leq p_k^{(m_{|\mathcal{C}|})}$ where $\{m_1, \cdots, m_{|\mathcal{C}|}\}$ is a permutation of $\{1, \cdots, |\mathcal{C}|\}$
6: Find largest $i_{\max}$ for which $p_k^{(m_i)} \leq \alpha_i := \frac{i\alpha}{|\mathcal{C}| \log\left(|\mathcal{C}|\right)}$
7: Reject all null hypotheses $\{H_0^{(m_i)} | i \leq i_{\max}\}$ associated with the GC links $m = m_1, m_2, \ldots, m_{i_{\max}}$
8: $J_k^{(m_i)} = 0$ for $i = i_{\max} + 1, \ldots, |\mathcal{C}|$
9: $J_k^{(m_i)} = 1 - \bar{\alpha} - \mathrm{F}_{\chi^2(M^{(d)}, \widehat{\nu}_k^{(m_i)})}\left(\mathrm{F}^{-1}_{\chi^2(M^{(d)})}(1-\bar{\alpha})\right)$ for $i = 1, \ldots, i_{\max}$
10: **end for**

Output: $\left\{\{J_k^{(\tilde{c}\mapsto c)}\}_{k=1}^{K} \,|\, (\tilde{c} \mapsto c) \in \mathcal{C}\right\}$

Non-central χ^2 Filtering and Smoothing Algorithm: In order to estimate the unknown non-centrality parameters $\nu_k^{(\tilde{c} \,\mapsto\, c)}$ given in Theorem 4.1, we make two additional assumptions. First, although the result of Theorem 4.1 establishes convergence in distribution as $\beta \to 1$, we make the assumption that $D_{k,\beta}^{(\tilde{c}\mapsto c)}$ is a sample drawn from a $\chi^2\big(M^{(d)}, \nu_k^{(\tilde{c} \,\mapsto\, c)}\big)$ density, when β is close to 1. This assumption is akin to the common adoption of a Gaussian density to parametrically describe uncertainties which are known to converge in distribution to a Gaussian random variable, thanks to the law of large numbers. In what follows, the dependence of $D_{k,\beta}^{(\tilde{c} \,\mapsto\, c)}$ and $\nu_k^{(\tilde{c} \,\mapsto\, c)}$ on c, $\tilde{c}$, and β will be suppressed for notational convenience.

Second, we assume that ν_k changes smoothly in time. Based on this assump-

tion, we construct a state-space model and develop recursive filtering and smoothing algorithms to compute smoothed estimates of ν_k from the observed deviance data $\{D_k\}_{k=1}^K$. To this end, given that $\nu_k \geq 0$, we define the exponential link $\nu_k = \exp(z_k)$, for some random variable z_k in the range of $(-\infty, \infty)$ and impose first-order autoregressive dynamics of the form:

$$z_k = \rho z_{k-1} + e_k, \tag{4.14}$$

where $0 < \rho \leq 1$ is a scaling factor, and $e_k \sim \mathcal{N}(0, \sigma_e^2)$ is a zero-mean i.i.d. Gaussian random variable with a variance of σ_e^2. Together with the assumption of $D_k \sim \chi^2\big(M^{(d)}, \nu_k\big)$, Eq. 4.14 forms a state-space model describing the dynamics of ν_k.

The parameters ρ and σ_e^2 are unknown, and need to be estimated. Assuming that the values of ρ and σ_e^2 are known, we can estimate $\{z_k\}_{k=1}^K$ given the sequence of deviance differences $\{D_k\}_{k=1}^K$ using approximate state-space smoothing [8]. The resulting estimator consists of two steps: a forward filter, and a backward fixed interval smoother.

For the filtering algorithm, we exploit the unimodal property of non-central chi-squared distribution, and make a recursive Gaussian approximation to the posterior probability density function $p(z_k|D_{1:k})$, where the posterior modes and variances are computed recursively [8]. Let $z_{k|l}$ and $\sigma_{k|l}^2$ denote the respective mode and variance of the state variable z_k, given the deviance samples up to and including time l, $\{D_i\}_{i=1}^l$. Using the Bayes' rule and substituting the non-central chi-squared density

function into the log-posterior, we get:

$$z_{k|k} := \underset{z_k}{\operatorname{argmax}} \left\{ -\frac{(D_k + \exp(z_k))}{2} + \frac{\xi}{2}(\log D_k - z_k) + \log I_\xi(\zeta_k) - \frac{(z_k - z_{k|k-1})^2}{2\sigma^2_{k|k-1}} \right\}, \tag{4.15}$$

where $\zeta_k := \sqrt{D_k \exp(z_k)}$, and $I_\xi(.)$ denotes the modified Bessel function of the first kind of order $\xi := M^{(d)}/2 - 1$. Note that in Eq. 4.15 a Gaussian approximation is applied to the density $p(z_k|D_{1:k-1}) \sim \mathcal{N}(z_{k|k-1}, \sigma^2_{k|k-1})$, where the mode and variance are easily derived from Eq. 4.14 as $z_{k|k-1} = \rho z_{k-1|k-1}$ and $\sigma^2_{k|k-1} = \rho^2 \sigma^2_{k-1|k-1} + \sigma^2_e$. From Eq. 4.15, the posterior mode $z_{k|k}$ can be computed as the solution to the following nonlinear equation:

$$z_k = z_{k|k-1} + \frac{\sigma^2_{k|k-1}}{2}\big(\zeta_k \, r_\xi(\zeta_k) - \exp(z_k)\big), \tag{4.16}$$

where the function $r_\xi(\zeta) := I_{\xi+1}(\zeta)/I_\xi(\zeta)$ is the ratio of modified Bessel functions of the first kind with order difference of one. This nonlinear equation can be solved numerically using iterative techniques such as Newton's method.

Given $z_{k|k}$, the posterior variance $\sigma^2_{k|k}$ can be computed as the negative inverse of the second order derivative of the log-posterior at $z_{k|k}$:

$$\sigma^2_{k|k} = \left(\left(\sigma^2_{k|k-1}\right)^{-1} + \frac{\exp(z_{k|k})}{2} - \frac{\zeta^2_{k|k}}{4}\left(1 - \frac{I_{\xi-1}(\zeta_{k|k}) I_{\xi+1}(\zeta_{k|k})}{I_\xi(\zeta_{k|k})^2}\right)\right)^{-1}, \tag{4.17}$$

where $\zeta_{k|k} := \sqrt{D_k \exp(z_{k|k})}$, and we used the recurrence relation $I_{\xi-1}(\zeta) = I_{\xi+1}(\zeta) + (2\xi/\zeta) I_\xi(\zeta)$ to simplify the update rule. Unlike the ordinary Bessel functions, the modified Bessel functions of the first kind $I_\xi(\cdot)$ are exponentially growing. This could cause numerical stability issues for the recursive update rules of Eqs. 4.16 and 4.17, as the input ζ_k may take large values through recursion leading to extremely large

values of the modified Bessel functions. To resolve potential numerical instability, we use the following sharp bounds on the ratio of Bessel function [102]:

$$\sqrt{\zeta^2+(\xi+1)^2}-(\xi+1)\leq \zeta\, r_\xi(\zeta) \leq \sqrt{\zeta^2+(\xi+1/2)^2}-(\xi+1/2).$$

We select the upper-bound as the more accurate approximate of the ratio $\zeta\, r_\xi(\zeta)$ in Eq. 4.16 for large values of ζ. Moreover, the second order Bessel ratio in Eq. 4.17 can be replaced using a sharp upper bound on the Turánian of the modified Bessel functions of the first kind, $I_\xi(\zeta)^2 - I_{\xi-1}(\zeta)I_{\xi+1}(\zeta)$ [103]:

$$\frac{I_\xi(\zeta)^2 - I_{\xi-1}(\zeta)I_{\xi+1}(\zeta)}{I_\xi(\zeta)^2} \leq \frac{1}{\sqrt{\zeta^2+\xi^2-1/4}}. \tag{4.18}$$

Given filtered outputs $z_{k|k}$ and $\sigma^2_{k|k}$ obtained from the forward filtering algorithm, we next perform backward smoothing using the fixed interval smoothing algorithm [8], yielding the smoothed posterior modes $z_{k|K}$ and variances $\sigma^2_{k|K}$ for $k = K, K-1, \ldots, 1$ as follows:

$$\begin{aligned} z_{k-1|K} &= z_{k-1|k-1} + s_k(z_{k|K} - z_{k|k-1}) \\ \sigma^2_{k-1|K} &= \sigma^2_{k-1|k-1} + s_k^2(\sigma^2_{k|K} - \sigma^2_{k|k-1}), \end{aligned} \tag{4.19}$$

where $s_k := \rho\sigma^2_{k-1|k-1}/\sigma^2_{k|k-1}$ is the backward smoothing gain. It should be noted that unlike the forward filtering, the backward smoothing step results in an overall batch-mode algorithm, as it refines the preceding filtered estimates $z_{k|k}$ using the deviance data D_i for $i > k$. Nevertheless, for real-time implementations one can always resort to the filtered estimates of the non-centrality parameters. Statistical confidence regions for both the filtered estimates $\widehat{z}_k^{\mathsf{filtered}} \sim \mathcal{N}(z_{k|k}, \sigma^2_{k|k})$ and smoothed estimates $\widehat{z}_k^{\mathsf{smoothed}} \sim \mathcal{N}(z_{k|K}, \sigma^2_{k|K})$ can be computed at each time step k

Algorithm 7 Non-central χ^2 Filtering and Smoothing Algorithm

Input: D_k, $M^{(d)}$, ρ, σ_e^2, $z_{0|0}$ and $\sigma_{0|0}^2$.

for $k = 1, 2, \ldots, K$ **do**

  Define $\xi := M^{(d)}/2 - 1$ and $\zeta_{k|k} := \sqrt{D_k \exp(z_{k|k})}$

  $z_{k|k-1} = \rho\, z_{k-1|k-1}$

  $\sigma_{k|k-1}^2 = \rho^2 \sigma_{k-1|k-1}^2 + \sigma_e^2$

  $z_{k|k} = z_{k|k-1} + \frac{\sigma_{k|k-1}^2}{2}\left(\sqrt{\zeta_{k|k}^2 + (\xi+1/2)^2} - (\xi+1/2) - \exp(z_{k|k})\right)$

  $\sigma_{k|k}^2 = \left((\sigma_{k|k-1}^2)^{-1} + \frac{\exp(z_{k|k})}{2} - \frac{\zeta_{k|k}^2}{4\sqrt{\zeta_{k|k}^2 + \xi^2 - 1/4}}\right)^{-1}$

  $\widehat{\nu}_k^{\mathsf{filtered}} = \exp(z_{k|k})$

  $\mathcal{CR}_k^{\mathsf{filtered}} = \left[\exp\left(z_{k|k} \pm \Phi^{-1}(1-\epsilon/2)\sigma_{k|k}\right)\right]$

end for

Given $\{z_{k|k}\}_{k=1}^K$ and $\{\sigma_{k|k}^2\}_{k=1}^K$

for $k = K, K-1, \ldots, 1$ **do**

  $z_{k-1|K} = z_{k-1|k-1} + s_k(z_{k|K} - z_{k|k-1})$

  $\sigma_{k-1|K}^2 = \sigma_{k-1|k-1}^2 + s_k^2(\sigma_{k|K}^2 - \sigma_{k|k-1}^2)$

  $\widehat{\nu}_{k-1}^{\mathsf{smoothed}} = \exp(z_{k-1|K})$

  $\mathcal{CR}_{k-1}^{\mathsf{smoothed}} = \left[\exp\left(z_{k-1|K} \pm \Phi^{-1}(1-\epsilon/2)\sigma_{k-1|K}\right)\right]$

end for

Output: Filtered estimates $(\widehat{\nu}_{1:K}^{\mathsf{filtered}}, \mathcal{CR}_{1:K}^{\mathsf{filtered}})$, and smoothed estimates $(\widehat{\nu}_{1:K}^{\mathsf{smoothed}}, \mathcal{CR}_{1:K}^{\mathsf{smoothed}})$

and mapped to those of $\widehat{\nu}_k^{\mathsf{filtered}} = \exp\left(\widehat{z}_k^{\mathsf{filtered}}\right)$ and $\widehat{\nu}_k^{\mathsf{smoothed}} = \exp\left(\widehat{z}_k^{\mathsf{smoothed}}\right)$ in a straightforward fashion. Algorithm 7 summarizes the non-central χ^2 filtering and smoothing procedure.

In order to simultaneously smooth z_k's and estimate the unknown parameters ρ and σ_e^2, an Expectation-Maximization (EM) approach can be used [104]. The details of this EM-based approach are given in subsection 4.1.4.

4.1.3 Summary of Advantages of AGC Inference over Existing Work

Algorithm 8 summarizes the overall AGC inference procedure. Choices of the parameters Θ involved in Algorithm 8 and its computational complexity are dis-

Algorithm 8 AGC Inference from Ensemble Neuronal Spiking

Input: Spike trains $\{\{n_t^{(c)}\}_{t=1}^T\}_{c=1}^C$ and parameters Θ.

1: **for** $c, \tilde{c} = 1, \ldots, C$, $\tilde{c} \neq c$ **do**

2: Recursively estimate the sparse time-varying modulation parameter vectors $\{\widehat{\boldsymbol{\omega}}_k^{(c)}\}_{k=1}^K$ and $\{\widehat{\boldsymbol{\omega}}_k^{(c\backslash\tilde{c})}\}_{k=1}^K$ corresponding to full and reduced GLMs using ℓ_1–PPF$_1$ (Algorithm 2),

3: Recursively compute the adaptive de-biased deviance differences $\{D_{k,\beta}^{(\tilde{c}\mapsto c)}\}_{k=1}^K$ (Algorithm 5),

4: Perform non-central χ^2-squared filtering and smoothing to estimate the non-centrality parameters $\{\widehat{\nu}_k^{(\tilde{c}\mapsto c)}\}_{k=1}^K$ from $\{D_{k,\beta}^{(\tilde{c}\mapsto c)}\}_{k=1}^K$ (Algorithm 7),

5: **end for**

6: **for** $k = 1, \ldots, K$ **do**

7: Apply BY rejection rule to the ensemble set of GC tests to control FDR at rate α (Algorithm 6),

8: Compute AGC maps $\widehat{\boldsymbol{\Phi}}_k \in [-1, 1]^{C\times C}$ based on the J-statistics as $(\widehat{\boldsymbol{\Phi}}_k)_{c,\tilde{c}} := s_k\big(\widehat{\boldsymbol{\omega}}_k^{(c,\tilde{c})}\big) J_k^{(\tilde{c}\mapsto c)}$ (Algorithm 6).

9: **end for**

Output: AGC maps $\{\widehat{\boldsymbol{\Phi}}_k\}_{k=1}^K$.

cussed in the next subsection. Here, we summarize the advantages of our methodology over existing work:

1) Sparse dynamic GLM modeling provides more accurate estimates of the parameters [33], and hence more reliable detection of the GC links, as compared to existing static methods based on ML. We further examine this aspect of our methodology in the next chapter, using an illustrative simulation study;

2) Relating the non-centrality parameters to the test strengths of the detected GC links is novel, and is not employed by existing techniques. In light of Theorem 4.1 and the need for estimating the non-centrality parameters, we devised a non-central χ^2 filtering and smoothing algorithm to exploit the entire observed data for obtaining reliable estimates;

3) Exponential weighting of the log-likelihoods admits construction of adaptive

filters for estimating the network parameters in a recursive fashion, which significantly reduces the computational complexity of our inference procedure; and

4) Characterization of AGC via the J-statistic as a normalized measure of hypothesis test strength for each detected GC link is novel, and can be further utilized for graph-theoretic analysis of the inferred functional networks. By viewing the J-statistic as a surrogate for link strength, the AGC networks can be refined by thresholding the J-statistics, and access to the distribution of the J-statistics in a network allows to perform further hypothesis tests regarding the network function.

In Chapter 5, we illustrate these advantages by comparing our methodology with two representative techniques for inferring functional network dynamics using a comprehensive range of simulation studies.

4.1.4 Parameter Selection and Computational Complexity

In this subsection, we describe how the various parameters involved in our proposed AGC inference procedure are selected, and discuss the underlying trade-offs.

Forgetting Factor: As discussed earlier in Chapter 3, the effective block length of the filter is determined by $N_{\mathsf{eff}} = \frac{W}{1-\beta}$ in the adaptive filtering setting with a forgetting factor mechanism β and window size W. It was shown in the remarks of Theorem 3.1 that the estimation error scales as $\mathcal{O}(\sqrt{S \log M / N_{\mathsf{eff}}})$ in the ℓ_2 sense, where S denotes the sparsity level. Thus, the forgetting factor β controls the trade-off between the estimation and tracking performance of the filter. That

is, a choice of β close to 1 corresponds to a large effective block length N_{eff}, which in turn results in a more accurate estimation of the modulation parameters $\widehat{\boldsymbol{\omega}}_k$, and consequently the AGC, at the cost of losing the trackability of the underlying dynamics. On the other extreme, a choice of β far from 1 reduces the effective block length, and thereby results in capturing the fast dynamics of the underlying time-varying process, although the estimation accuracy degrades. As discussed in the remarks following the proof of Theorem 4.1, the proposed statistical testing procedure enables us to detect G-causal links associated with true cross-history components of the order of $\boldsymbol{\omega}_k^{\beta} = \mathcal{O}(\sqrt{1-\beta})$. Hence, a choice of β close to 1 will increase the test strengths. In the applications of interest in this dissertation, the underlying dynamics are slower than the sampling rate, which allows us to choose forgetting factor values sufficiently close to 1. While it may be beneficial to tune β via cross-validation, our numerical experiments show that the resulting values of β turn to be close to 1 (i.e., $1-\beta \in [10^{-4}, 10^{-2}]$, depending on the choice of W). Therefore, in order to simplify the cross-validation procedure, we fixed the value of β close to 1 in our analysis. It is noteworthy that the usage of the forgetting factor mechanism mitigates the problem of choosing a window size faced by GC inference methods based on sliding-window processing.

Model Order Selection: Our model selection procedure is grounded in the compressed sensing theory. In contrast to classical model order selection procedures (e.g., AIC), compressed sensing suggests choosing large model orders followed by sparse regularization to avoid overfitting. Indeed, our recent results on extending the theoretical guarantees of compressed sensing to processes with non-i.i.d. and

history dependent covariates [33, 105], show that recovery of sparse history kernels with large ambient dimensions M is possible from a limited number of observations N, in which N may be comparable or smaller than M, as long as the sparsity level S is small enough. In more precise terms, long kernels of self-history can be robustly estimated given an effective number of observations N_{eff} scaling sub-linearly with M and S.

The benefit of employing such models with long self-history kernels is two-fold: first, long self-history kernels M_H^{self} enable us to maximally capture the intrinsic spiking statistics of a unit. Second, due to the autoregressive nature of these models, long self-history kernels allow for estimation, and thereby correcting for the effects of latent confounding variables, which cannot be explained by the cross-history influences from other units. Thus, we choose $M_H^{\mathsf{self}} > M_H^{\mathsf{cross}}$ to maximally capture the aforementioned intrinsic and latent confounding effects. At the same time, smaller values of M_H^{cross} are beneficial in increasing the statistical test strengths, as they directly set the statistical thresholds for multiple hypothesis testing. We present two illustrative numerical experiments in subsection 5.1.4 that corroborate our choices for these parameters.

Adaptive Filtering Parameters: In order to achieve an estimation performance with high accuracy, we select the effective block length $N_{\mathsf{eff}} \gg M^{(F)}$ to be larger than the kernel length. We use non-overlapping spike counting windows of length W_H for parameterizing the self- and cross-history kernels, where W_H is often chosen to be comparable to the filtering window length W.

For the adaptive filtering setting, we first standardize the matrix of covariates

(i.e., zero-mean columns with unit norm), and then apply the ℓ_1–PPF_1 adaptive filter, with a step size of $\varsigma := \frac{1-\beta}{cW}$, where c is a constant often chosen in the range $c \in [1, 10]$, to achieve different levels of smoothing.

The regularization parameter γ for the ℓ_1–PPF_1 is chosen as $\gamma = \mathcal{O}\big(\sqrt{\log M/N_{\mathsf{eff}}}\big)$, based on the results of Theorem 3.1 discussed in Chapter 3, and the asymptotic scaling requirement in [87], to obtain consistent ℓ_1-regularized ML estimates. In order to adapt this parameter to different neurons, we choose $\gamma^{(c)} = \bar{\gamma}^{(c)}\sqrt{\bar{\kappa}^{(c)} \log M/N_{\mathsf{eff}}}$ for neuron (c), where $\bar{\kappa}^{(c)} := \text{var}(\mathbf{n}^{(c)}) = \bar{\lambda}\Delta^{(c)}(1 - \bar{\lambda}\Delta^{(c)})$, followed by tuning the normalized regularization parameter $\bar{\gamma}^{(c)}$ for each neuron in a data-driven fashion via the even-odd two-fold cross validation procedure.

Note that when the underlying functional network is fully connected, the cross-validation procedure for tuning the regularization parameter γ is expected to choose values near zero (i.e., no sparsity in the parameter vectors), and hence our methodology can adapt to non-sparse network connectivity as well. For the applications of interest in this work, the cross-validation procedure consistently resulted in sparse functional networks.

Finally, it is possible to generalize the ℓ_1-regularization scheme to have a different regularization parameter $\gamma^{(c,\tilde{c})}$ for the cross-history parameters of units (c) and $(\tilde{c})$. Theoretical analysis, however, suggest that there is little benefit in terms of estimation accuracy in doing so, which comes at the cost of higher computational complexity in the cross-validation and bias correction stages. More precisely, separate regularization of each of the cross-history parameters may result in better constants in the error rate, but the asymptotic scaling of the rate remains unchanged. For

instance, as mentioned earlier, the result of Theorem 3.1 implies that the estimation error scales as $\mathcal{O}(\sqrt{S \log M / N_{\mathsf{eff}}})$, which is optimal modulo the logarithmic factor. By viewing the concatenation of several sparse vectors as another sparse vector, we use a single regularization parameter that is tuned appropriately via cross-validation in order to select a sparse model at a near optimal error rate. Nevertheless, the ℓ_1-PPF$_1$ procedure can be generalized in a straightforward fashion to accommodate multiple regularization parameters, thanks to the separable nature of the ℓ_1 norm and the underlying proximal algorithms (See Appendix A.2 for further details).

Parameters of the Non-central χ^2 Filtering and Smoothing: For the non-central χ^2 filtering and smoothing algorithm, we select the scaling factor $\rho \in [\beta, 1]$ close (or equal) to one to promote temporal continuity. The state variance σ_e^2 plays the role of a smoothing factor for non-centrality parameter estimates $\widehat{\nu}_k$, and can be determined in two ways. First, we can choose a small value, e.g. in the range $[10^{-7}, 10^{-4}]$ suggested by our numerical experiments, which results in smooth estimates of $\widehat{\nu}_k$ for a wide range of settings.

Second, σ_e^2 can be systematically estimated via the expectation maximization (EM) algorithm [106] in a data-driven fashion using the observed deviance data $D_{1:K} := \{D_k\}_{k=1}^K$. We take $z_{1:K} := \{z_k\}_{k=1}^K$ as the set of latent variables for the EM algorithm. Given an estimate $\widehat{\sigma}_e^{2,(\ell)}$ at the ℓ-th iteration, the E-step at the $(\ell+1)$-st iteration computes:

$$\mathbb{E}_{\mathbf{z}}\Big[\log p\big(D_{1:K}, z_{1:K}|\sigma_e^2\big)\big|D_{1:K}, \widehat{\sigma}_e^{2,(\ell)}\Big] = -\frac{K}{2}\log(\sigma_e^2)$$
$$- \frac{1}{2\sigma_e^2}\sum_{k=1}^{K}\Big\{\big(\sigma_{k|K}^2+z_{k|K}^2\big)+\rho^2\left(\sigma_{k-1|K}^2+z_{k-1|K}^2\right)$$
$$- 2\rho\left(\sigma_{k-1,k|K}^2 + z_{k-1|K}z_{k|K}\right)\Big\} + \mathsf{cnst.}, \tag{4.20}$$

where $\mathbb{E}_{\mathbf{z}}\big[.|D_{1:K}, \widehat{\sigma}_e^{2,(\ell)}\big]$ denotes the expectation operator with respect to the latent variables given the complete set of deviance data $D_{1:K}$ and the current estimate of the parameter $\widehat{\sigma}_e^{2,(\ell)}$, and $\mathsf{cnst.}$ denotes all terms not dependent on σ_e^2. It is noteworthy that calculation of the E-step involves computation of the smoothed means and variances $\mathbb{E}_{\mathbf{z}}\big[z_k^2 \,|\, D_{1:K}, \widehat{\sigma}_e^{2,(\ell)}\big] = \sigma_{k|K}^2 + z_{k|K}^2$, which are readily available from the non-central chi-squared smoothing given by Eq. 4.19, and the covariance terms $\mathbb{E}_{\mathbf{z}}\big[z_{k-1}z_k|D_{1:K}, \widehat{\sigma}_e^{2,(\ell)}\big] = \sigma_{k-1,k|K}^2 + z_{k-1|K}z_{k|K}$, which can be computed using a state-space covariance smoothing algorithm [104] as $\sigma_{k-1,k|K}^2 = s_k\sigma_{k|K}^2$. The M-step gives the update for $\widehat{\sigma}_e^{2,(\ell+1)}$ by maximizing Eq. 4.20 as follows:

$$\widehat{\sigma}_e^{2,(\ell+1)} = \frac{1}{K}\sum_{k=1}^{K}\Big\{\big(\sigma_{k|K}^2 + z_{k|K}^2\big)+\rho^2\big(\sigma_{k-1|K}^2 + z_{k-1|K}^2\big) - 2\rho\left(\sigma_{k-1,k|K}^2 + z_{k-1|K}z_{k|K}\right)\Big\}. \tag{4.21}$$

Computational Complexity Considerations: The computational complexity of Algorithm 8 (per cross-validation iteration) is linear in the total data length T and quadratic in the network size C and parameter orders M, due to the RLS-type adaptive filtering procedure used [33]. However, the high number of cross-validation iterations required to tune the regularization parameters increases the overall runtime of the algorithm. Substantial reduction of the run-

time can be achieved by parallel implementation: the cross-validation steps for each unit can be done independently of the others, and therefore using a natural parallel implementation, the runtime would reduce by $1/C$. Note that we have not used this parallel scheme in our current implementation deposited on GitHub (`https://github.com/Arsha89/AGC_Analysis`). In order to efficiently analyze data from high-density neuronal recordings, we suggest the use of a parallel implementation and view it as a future work.

4.2 Granger Causality Analysis of Optical Imaging Data

We finally present a static variant of our GC methodology to be applied to optical imaging data. In particular, we apply this method to data from two different imaging experiments: two-photon imaging data from mouse A1 during different auditory tasks and behavioral conditions, and light-sheet imaging data from the entire brain of larval zebrafish during locomotive behavior. The results will be discussed in chapter 6.

In order to capture the functional dependencies within an ensemble of nodes with continuous-valued activities, and the sparsity of interactions thereof, we employ sparse multivariate autoregressive models. Similar to our methodology for spiking data, we introduce a variant of GC which accounts for sparse interactions, estimate the model parameters using fast optimization methods, and perform statistical tests to assess the significance of possible GC interactions, while controlling the FDR to avoid spurious detection of the GC links.

Modeling: Consider a sequence of calcium indicator fluorescence measurements from a set of C neurons indexed by $c = 1, 2, \ldots, C$ within a slice, denoted by $\{y_{r,n}^{(c)}\}_{r=1:R,n=1:N}^{c=1:C}$ over time bins $n = 1, \ldots, N$, and across R trial repetitions indexed by $r = 1, \ldots, R$. We adopt a sparse vector autoregressive (VAR) framework [66] for modeling the slow-decaying and transient dynamics of the calcium fluorescence signals as well as the cross-dependencies among the neurons.

Suppose that the fluorescence observation vector of neuron (c) at the r-th repetition is represented by $\mathbf{y}_r^{(c)} := [y_{r,1}^{(c)}, \ldots, y_{r,N}^{(c)}]'$, and let $\bar{\mathbf{y}}^{(c)} := \left[\mathbf{y}_1^{(c)\prime}, \mathbf{y}_2^{(c)\prime}, \ldots, \mathbf{y}_R^{(c)\prime}\right]'$ denote the zero-mean total observation vector, containing the set of all observation vectors $\mathbf{y}_r^{(c)}$ from all trials $r = 1, \ldots, R$. The effective neural covariates taken into account in our models are each neuron's self-history of activity and the history of activities of other neurons in the ensemble. We consider a lag of L_H samples within which the possible neuronal interactions may occur. Then, we segment L_H into M windows of lengths $W_{H,1}, W_{H,2}, \ldots, W_{H,M}$ such that $\sum_{i=1}^{M} W_{H,i} = L_H$. Let $b_m := \sum_{l=1}^{m} W_{H,l}$ for $m = 1, \ldots, M$, and $b_0 = 0$. Let

$$h_{r,n,m}^{(c)} := \frac{1}{W_{H,m}} \sum_{k=n-1-b_m}^{n-1-b_{m-1}} y_{r,k}^{(c)} \tag{4.22}$$

represent the average activity of neuron (c) within the m-th window lag of length $W_{H,m}$ with respect to time n and at trial r. We can then define the vector of history covariates from neuron (c), effective at time n and trial r as $\mathbf{h}_{r,n}^{(c)} := [h_{r,n,1}^{(c)}, h_{r,n,2}^{(c)}, \ldots, h_{r,n,M}^{(c)}]'$. Next, let $\mathbf{x}_{r,n} := \left[\mathbf{h}_{r,n}^{(1)\prime}, \mathbf{h}_{r,n}^{(2)\prime}, \ldots, \mathbf{h}_{r,n}^{(C)\prime}\right]'$ denote the vector of covariates from all neurons at time n and trial r.

In order to represent the covariates in a more compact form, we consider the

$N \times MC$ matrix $\mathbf{X}_r := \left[\mathbf{x}_{r,1}, \mathbf{x}_{r,2}, \ldots, \mathbf{x}_{r,N}\right]'$ which contains in its rows the covariate vectors at all times $n = 1, \ldots, N$ within trial r. Finally, let $\bar{\mathbf{X}} := \left[\mathbf{X}_1', \mathbf{X}_2', \ldots, \mathbf{X}_R'\right]'$ represent the matrix of all covariates with standardized columns (i.e., zero-mean columns with unit norm), capturing the covariates $\mathbf{X}_r$ for all the trials $r = 1, \ldots, R$.

The VAR model can then be expressed as:

$$\bar{\mathbf{y}}^{(c)} = \bar{\mathbf{X}}\boldsymbol{\omega}^{(c)} + \bar{\boldsymbol{\varepsilon}}^{(c)} \tag{4.23}$$

where $\bar{\boldsymbol{\varepsilon}}^{(c)} := [\boldsymbol{\varepsilon}_1^{(c)'}, \ldots, \boldsymbol{\varepsilon}_R^{(c)'}]' \sim \mathcal{N}(\mathbf{0}, {\sigma^{(c)}}^2\mathbf{I})$ is a zero-mean Gaussian noise vector of size $R \times N$ with variance ${\sigma^{(c)}}^2$, and $\boldsymbol{\omega}^{(c)}$ is a parameter vector accounting for the interactions in the network, for $c = 1, 2, \ldots, C$.

In agreement with the parsing of the covariates in the matrix $\bar{\mathbf{X}}$, the parameter vector $\boldsymbol{\omega}^{(c)} := \left[\boldsymbol{\omega}^{(c,1)'}, \boldsymbol{\omega}^{(c,2)'}, \ldots, \boldsymbol{\omega}^{(c,C)'}\right]$ in Eq. 4.23 is composed of a collection of cross-history dependence vectors $\{\boldsymbol{\omega}^{(c,\tilde{c})}\}^{\tilde{c}=1:C}$, where $\boldsymbol{\omega}^{(c,\tilde{c})}$ represents the contribution of the history of neuron $(\tilde{c})$ to the activity of neuron (c) via the corresponding covariate vector $\mathbf{h}_{r,n}^{(c)}$ encoded in matrix $\bar{\mathbf{X}}$. In particular, the component $\boldsymbol{\omega}^{(c,c)}$ is important in capturing the slow calcium florescence decay in an autoregressive fashion, and thereby excluding the transient effects of fluorescence decay from the GC analysis.

Next, we invoke the hypothesis of sparsity in the interactions among the neurons in the network, as previous section. This hypothesis is grounded in a body of well-accepted evidence from theoretical and experimental studies [34, 36–39, 41]. In our model, the sparsity of the interactions can be captured through the sparsity of the parameter vector $\boldsymbol{\omega}^{(c)}$: when only very few components of $\boldsymbol{\omega}^{(c)}$ are non-zero,

neuron $\boldsymbol{\omega}^{(c)}$ is only affected by the activity history of a few neurons in the network. In addition, as the dimension of the parameter vector given by MC scales with the network size C, the hypothesis of sparsity enables the detection of salient interactions within a large network, and thereby mitigates overfitting, especially when the observations are noisy and trials are limited in number.

Estimation: In order to define a framework for inferring a possible GC link ($\tilde{c} \mapsto c$), two nested models are taken into account: 1) the VAR model in Eq. 4.23, where the contributing covariates from all the neurons are taken into account, referred to as the full model, and 2) the same model in which the covariates and parameters of a single neuron ($\tilde{c}$) on neuron (c), $\tilde{c} \neq c$ are excluded, to which we refer as the reduced model. The parameters and covariates associated with the reduced model are denoted by $\boldsymbol{\omega}^{(c\backslash\tilde{c})}$ and $\mathbf{X}^{\backslash\tilde{c}}$, respectively.

The sparse parameter vector associated with either of the two models can be estimated by solving an ℓ_1-regularized maximum likelihood (ML) problem for each neuron as follows:

$$\widehat{\boldsymbol{\omega}} = \underset{\boldsymbol{\omega}}{\operatorname{argmin}} \Big(\frac{1}{2} \|\bar{\mathbf{y}} - \mathbf{X}\boldsymbol{\omega}\|_2^2 + \gamma \|\boldsymbol{\omega}\|_1 \Big) \tag{4.24}$$

where $\mathbf{X}$ takes the two values of $\bar{\mathbf{X}}$ and $\bar{\mathbf{X}}^{\backslash\tilde{c}}$ for the full and reduced models, respectively, the ℓ_1-norm is defined as $\|\boldsymbol{\omega}\|_1 := \sum_{m=1}^{M} |\omega_m|$, and $\gamma \geq 0$ is a regularization parameter tuning the sparsity level, which can be selected based on analytical results on ℓ_1-regularized ML problems [33, 87] or via cross-validation. This ℓ_1-regularized ML problem can be solved efficiently using proximal algorithms [20,107], as discussed in Appendix A.2. Given the parameter estimate $\widehat{\boldsymbol{\omega}}$, the corresponding variance asso-

ciated with the two full or reduced model can be computed as $\widehat{\sigma}^2 = \frac{1}{NR}\|\bar{\mathbf{y}} - \mathbf{X}\widehat{\boldsymbol{\omega}}\|_2^2$.

Inference: The conventional measures of GC are based on ML estimates of the VAR parameters, and not the regularized ML as in our case. Hence, as before, we need to modify the GC measure and the corresponding deviance statistics, to account for the estimation bias incurred due to ℓ_1-regularization.

To this end, we modify the deviance difference statistic corresponding to the full and reduced models to compensate for the bias incurred due to sparse regularization. Building on the theoretical results from section 4.1 in the context of point processes, and recent results from high dimensional statistics [87], the bias can be computed for the full model as $B^{(c)} := \mathbf{g}^{(c)\prime}\mathbf{H}^{(c)^{-1}}\mathbf{g}^{(c)}$, where $\mathbf{g}^{(c)} := \bar{\mathbf{X}}'(\bar{\mathbf{y}}^{(c)} - \bar{\mathbf{X}}\widehat{\boldsymbol{\omega}}^{(c)})/\widehat{\sigma}^{(c)2}$ and $\mathbf{H}^{(c)} := -\bar{\mathbf{X}}'\bar{\mathbf{X}}/\widehat{\sigma}^{(c)2}$ are the gradient and Hessian of the log-likelihood function for the Gaussian VAR model of Eq. 4.23, respectively. Similarly, the bias $B^{(c\backslash\tilde{c})}$ for the reduced model can be computed by replacing the matrix of covariates and parameter estimate by $\bar{\mathbf{X}}^{\backslash\tilde{c}}$ and $\widehat{\boldsymbol{\omega}}^{(c\backslash\tilde{c})}$, respectively.

The deviance difference statistic associated with the two nested full and reduced models can be expressed as:

$$D^{(\tilde{c}\mapsto c)} := NR \log \frac{\widehat{\sigma}^{(c\setminus\tilde{c})2}}{\widehat{\sigma}^{(c)2}} - B^{(\tilde{c}\mapsto c)} \tag{4.25}$$

where $B^{(\tilde{c}\mapsto c)} := B^{(c)} - B^{(c\setminus\tilde{c})}$ denotes the difference of bias terms corresponding to the full and reduced models. Note that the first term coincides with the log-likelihood ratio statistic for Gaussian data [108], and captures the prediction improvement of the full model over the reduced model.

We finally employ the inference framework presented in the previous section to

simultaneously test the statistical significance of all possible GC interactions and to control the FDR at a given significance level α. This inference framework integrates an extension of classical results on analysis of deviance, and a multiple hypothesis testing procedure based on the Benjamini-Yekutieli FDR control [101]. The weights of the detected links are similarly characterized using the Youdens J-statistic, and the excitatory or suppressive nature of GC links are determined by the effective sign of estimated cross-history parameters associated with shorter latencies.

Chapter 5: Validation of the Theoretical Framework Using Comprehensive Simulation Studies

This chapter contains the simulation results and the numerical examples used for validating the proposed algorithms and theoretical results discussed in Chapter 4. In the first section, we assess the performance of AGC inference for neural spiking data in terms of estimation accuracy and tracking capability through several simulated examples and comparisons. In the second section, we provide a simulation study for static GC inference from continuous-valued observations in the context of optical imaging data.

5.1 Simulation Studies for Neuronal Spiking Data

In this section, we carry out a comprehensive evaluation of the performance of the AGC inference method in terms of both identification and tracking of G-causal influences from neural spike trains through simulation studies and comparisons with two representative techniques for functional network inference.

5.1.1 A Simulated Example: AGC Inference for Neuronal Spike Trains

We consider a network of $C = 8$ functionally inter-connected neurons indexed by $c = 1, 2, \ldots, 8$, where each neuron is causally linked to a group of other neurons through a set of inhibitory or excitatory links.

As illustrated in Fig. 5.1, the network connectivity pattern undergo three main evolution states in time, each covering one-third ($40s$) of the simulation period: 1) the first static state, where neuron (1) plays a dominant role, causally influencing all other neurons, 2) the intermediate dynamic state, where neuron (1) loses the dominant role to neuron (5), as its causal influences smoothly decay, while a new set of causal interactions from neuron (5) to all the other neurons emerge, 3) the final static state, where all the causal links from neuron (1) are completely vanished and the links from neuron (5) are stabilized. The network also comprises three static causal links, e.g. $(3 \mapsto 7)$, which remain constant throughout.

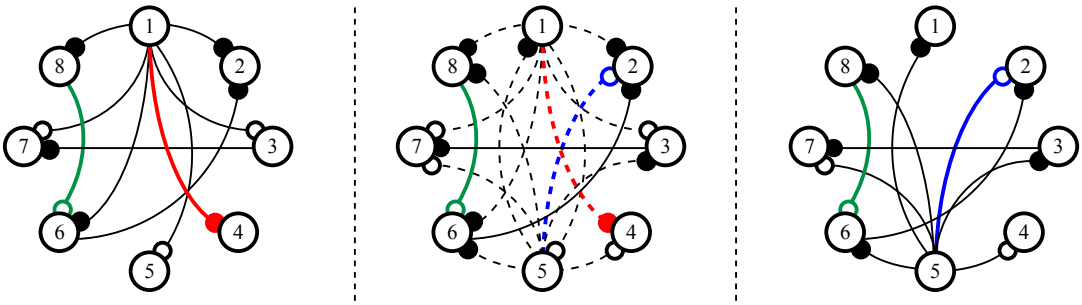

Figure 5.1: Three states of the functional network evolution, where a network of 8 neurons (vertices) are interacting through static (solid edges) or dynamic (dashed edges) causal links of inhibitory (open circles) or excitatory (filled circles) nature. The selected G-causal links under study are color-coded in blue, red and green.

An observation period of $\mathcal{T} = 120$ s is discretized to $T = 120$ k bins of length $\Delta = 1$ ms. We use a point process model with Bernoulli spiking statistics to generate the binary spike trains for all neurons, where the CIF is modeled given the dynamic GLM of Eq. 2.8. Note that the G-causal pattern of Fig. 5.1 is unknown to the estimator, and is to be inferred from the simulated spike trains. Further details on the parameter selection and estimation procedure are given later in this section.

Fig. 5.2–A shows a realization of the simulated spike trains indicated by black vertical lines for all 8 neurons within three sample windows of length $1s$, with end-points at $\{40, 60, 120\}$ s, selected from the three segments of the simulation. Fig. 5.2–B shows the time-course of the estimated non-centrality parameters $\widehat{\nu}_k$ and their 95% confidence intervals obtained by the non-central χ^2 filtering and smoothing algorithm associated with four selected GC links: 1) $(1 \mapsto 4)$ a dynamic weakening GC link (red), 2) $(5 \mapsto 2)$ a dynamic strengthening GC link (blue), 3) $(8 \mapsto 6)$ a static link (green), and 4) $(8 \mapsto 2)$ a non-existing GC link (magenta). Black traces show the shifted observed deviances $D_k - M^{(d)}$. Fig. 5.2–C represents the time-course of the estimated J-statistics associated with four selected GC links plotted in four separate panels, where the FDR is controlled at a rate $\alpha = 0.1$.

In Fig. 5.2–B, the estimates of $\widehat{\nu}_k$ corresponding to the three existing GC links take significant values, correctly identifying the G-causal interactions, while $\widehat{\nu}_k$ takes values close to zero for the non-existing link, implying no significant G-causal interaction. The time-course of changes for both dynamic links and the static link is closely tracked by the non-centrality parameters, albeit with an apparent delay. This delay is due to the choice of the effective window length, and highlights the

trade-off between estimation accuracy and delay. While it is possible to reduce this delay by choosing smaller effective windows, for the sake of accuracy of parameter estimation and thereby robust detection of the AGC links, we have chosen the effective window length to be 10 s (a fraction of the 40 s transition period) to incur a tolerable delay. The aforementioned performance is echoed in the test strengths quantified by the J-statistics shown in Fig. 5.2–C. Even though the non-centrality parameters in Fig. 5.2–C track the changes of the network parameters much faster, the J-statistics may lag behind due to the conservative statistical thresholds set by the FDR control procedure. By choosing a higher FDR level, the J-statistics will capture the changes much faster, but at the expense of possibly more false discoveries. It is noteworthy that our proposed method distinguishes the direct GC links from the indirect ones, as it correctly detects the direct GC links ($8 \mapsto 6$) and ($6 \mapsto 2$), but rejects the existence of the corresponding indirect link ($8 \mapsto 2$).

The top row in Fig. 5.2–D shows the ground truth G-causal maps plotted at 9 time instances (three per segment). Each map Φ_k represents an 8×8 color-coded array showing the excitatory, inhibitory and no-GC links in red, blue and green colors, respectively. The AGC maps estimated by our method are shown in the second row, where each entry $(\widehat{\mathbf{\Phi}}_k)_{c,\tilde{c}}$ represents the J-statistic $J_k^{(\tilde{c} \mapsto c)}$ of the estimated GC link $(\tilde{c} \mapsto c)$, where the excitatory or inhibitory nature of the links is determined by the sign of the AGC measure, accounting for the aggregate cross-history contribution. Note that such excitatory or inhibitory nature is not indicative of the morphological identity of the connections.

We compare the AGC maps with two other methods: the static GC method of

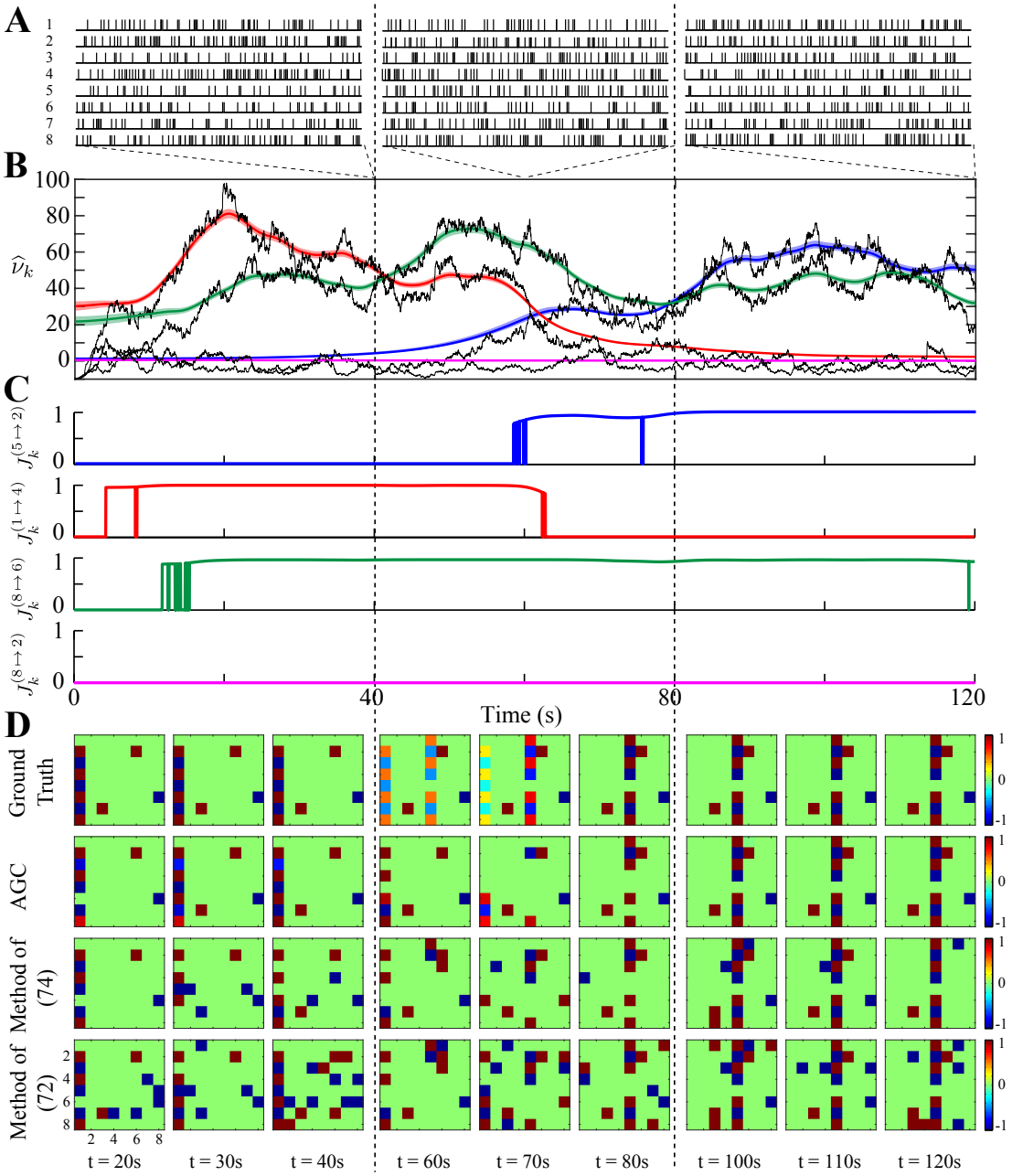

Figure 5.2: Functional network dynamics inference from simulated spikes. A) one realization of simulated spikes within $1s$ windows selected at $\{40, 60, 120\}s$, B) estimated non-centrality $\widehat{\nu}_k$ across time corresponding to 4 selected GC links (color-coded in Fig. 5.1), along with the shifted deviance differences $D_k - M^{(d)}$ (black traces) and the 95% confidence regions for each estimated trace $\widehat{\nu}_k$ (colored traces), C) four panels of estimated J-statistics J_k corresponding to the selected GC links, D) performance comparison of the causal inference methods: 1) the proposed AGC method (second row), 2) static GC method in [74] (third row), and 3) functional connectivity method in [72] (last), along with the true causality maps (first row). Each panel represents the estimated 8×8 causality map at a specific time.

[74] (third row), and the functional connectivity analysis of [72] (final row). In order to adapt these methods to the time-varying setting, we used non-overlapping window segments whose length is chosen to match the effective window length $N_{\mathsf{eff}} := \frac{W}{1-\beta}$ of our method. Within each window, the signed binary functional connectivity is estimated using the methods outlined in [72] and [74]. The true model order and the same significance level are used for all methods. Fig. 5.2–D (last two rows) shows the connectivity maps obtained by [72] and [74], at the same 9 time instances as in the previous rows. On a qualitative level, Fig. 5.2–D reveals the favorable performance of our proposed framework in terms of both identification and tracking of the GC influences. The method of [74] results in both high false positive and false negative errors, and fails to track the GC dynamics due to highly variable parameter estimates. Similarly, the method of [72] shows poor false positive rejection and tracking performance. In the spirit of easing reproducibility, we have archived a MATLAB implementation that fully generates Fig. 5.2 on GitHub (`https://github.com/Arsha89/AGC_Analysis`).

Quantitative Performance Comparison In order to quantify the foregoing performance comparison between the AGC inference and the methods of [72] and [74], we repeated the previous simulation for $R = 500$ realizations of spike trains randomly generated based on the network dynamics in Fig. 5.1, and computed two performance metrics, true detection rate (TDR) and false alarm rate (FAR), for each repetition. In what follows, we describe the computation of the TDR and FAR performance metrics for the AGC inference method, and discuss the details of

statistical tests performed on these metrics.

Given the continuous nature of the AGC links, as opposed to the binary connectivity measures of the other two methods, we binarize the resulting J-statistics for fairness of comparison. To this end, let $A_R^{(\tilde{c}\mapsto c)}$ be the fraction of times within a time window where the AGC link $(\tilde{c} \mapsto c)$ is identified with high statistical significance $J_k^{(\tilde{c}\mapsto c)} > J_{\mathsf{th}}$. We call an AGC link active within a given time window if $A_R^{(\tilde{c}\mapsto c)} > A_{\mathsf{th}}$, and inactive otherwise. We selected the thresholds to be $A_{\mathsf{th}} = \frac{1}{3}$ and $J_{\mathsf{th}} = \frac{1-\bar{\alpha}}{3}$.

The TDR at each time window is computed as the ratio of the correctly identified links to the total number of existing GC links. The FAR at each time k is computed as the ratio of spuriously detected links to the total number of non-existent links. Given the ground truth GC map shown in Fig. 5.2–D, these performance metrics can be computed for the static (first and last) segments of the experiment in a straightforward fashion. For the middle segment, where the GC influences undergo dynamic changes, we define the ground truth as follows: a threshold of $G_{\mathsf{th}} = \frac{1}{4}$ is used to binarize the ground truth GC links, which linearly ascend from 0 to 1 (emerging link) or descend from 1 to 0 (vanishing link) in the middle segment. For each repetition of the simulation, the FAR and TDR metrics are computed for each of the three segments by averaging over the time windows within, resulting in two summary statistics.

The area under curve (AUC) performance metric further summarizes the two TDR and FAR metrics into a single summary statistic, by computing the area under the ROC curve. The ROC performance curves are obtained by varying the values

of mean FDR $\bar{\alpha} \in [0, 1]$ for AGC and the statistical thresholds of [74] and [72] and plotting the corresponding (TDR,FAR) pairs averaged across repetitions.

Due to the highly non-Gaussian nature of the empirical distributions of the paired difference metrics, we have used the non-parametric Wilcoxon signed-rank test for comparison. The corresponding effect sizes are computed in the form of rank correlation $r := \mathscr{W}/\mathscr{S}$, where $\mathscr{W}$ is the Wilcoxon signed-rank statistic and $\mathscr{S}$ is the total sum of ranks [109].

Fig. 5.3–A represents the performance results in terms of TDR and FAR, which are shown in green and red, respectively. Boxes indicate the mean values as well as the 90% confidence intervals pooled across all repetitions. Based on the Wilcoxon signed-rank test with $p < 0.001$, our method has a significantly lower FAR compared with both [72] (effect sizes of $r = 1$ for all segments) and [74] ($r = 0.8, 0.86$ and 0.98 for the three segments, respectively). Our achieved TDRs are also significantly higher than those of [74] ($r = 0.73, 0.996$ and 0.94 for the three segments, respectively), and are only outperformed by [72] in the middle segment ($r = 0.19, 0.86$ and 0.27 for the three segments, respectively). It is noteworthy that our method is the only one with consistently low FAR ($< 1\%$), while maintaining high TDR. Finally, both methods in [72] and [74] output *binary* connectivity maps, as opposed to AGC which provides normalized *continuous-valued* test strengths of the detected GC links. Figure 5.3–B exhibits the ROC performance curves for the three segments of simulation, obtained by varying the significance levels for all three methods. The corresponding AUC values for the three segments are indicated on top of Fig 5.3–B. While the methods of [74] and [72] exhibit similar ROC performances,

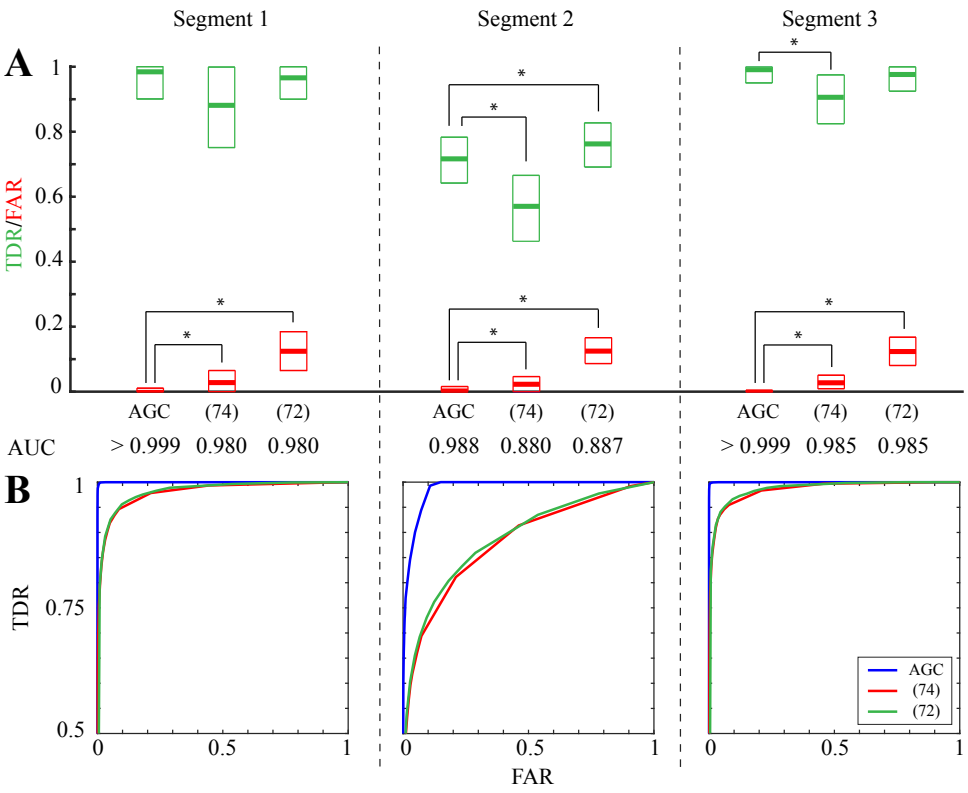

Figure 5.3: A) Performance comparison of AGC inference with the methods of [74] and [72] in terms of TDR (green) and FAR (red) for the three segments of the simulation period. Boxes represent the mean and 90% confidence intervals. Stars indicate significant difference with effect size of $r \geq 0.8$ (Wilcoxon signed-rank test, $p < 0.001$), B) ROC performance curves of the AGC inference (blue) and the methods of [74] (red) and [72] (green) for the three segments. The corresponding AUC values for the three methods are reported at the top of ROC curves.

the AGC achieves higher AUC values, particularly in the middle segment. We expect the performance gap between the AGC inference and the other two methods to increase for larger networks with higher sparsity.

Empirical Validation of the Results of Theorem 4.1: In order to validate the results of Theorem 4.1, we inspect the empirical distributions of the observed deviance differences $D_k^{(\tilde{c} \mapsto c)}$ for some selected links at arbitrary time points, and

compare them with the associated theoretical fits. To this end, we use the same ensemble of $R = 500$ simulated spike train realizations from the previous simulation study.

Fig. 5.4 exhibits the resulting histograms and theoretical density fits (solid curves), as predicted by Theorem 4.1, for two representative GC links $(1 \mapsto 7)$ and $(5 \mapsto 2)$ from Fig. 5.1 at two selected time points of 40 s (endpoint of the first segment) and 120 s (endpoint of the third segment). Note that the GC link $(1 \mapsto 7)$ was present in the first segment of the experiment and vanished in the last

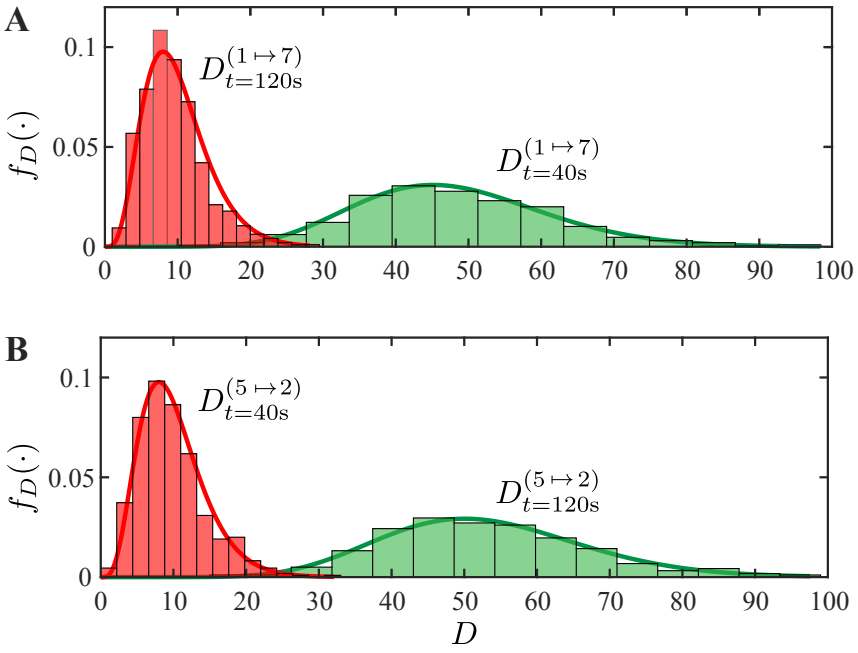

Figure 5.4: Empirical and theoretical fits to the distributions of the adaptive de-biased deviance difference $D_k^{(\tilde{c}\mapsto c)}$ for two selected links from Fig. 5.1. The empirical densities are shown as histograms using 15 bins (colored bars) and the theoretical fits are plotted as solid curves. (A) Empirical and theoretical densities of $D_k^{(1\mapsto 7)}$ at $t = 40$ s (existing GC link) and $t = 120$ s (non-existing GC link), (B) Empirical and theoretical densities of $D_k^{(5\mapsto 2)}$ at $t = 40$ s (non-existing GC link) and $t = 120$ s (existing GC link).

segment, and the GC link ($5 \mapsto 2$) did not exist in the first segment, but emerged in the last segment. The theoretical density $\chi^2(M^{(d)})$ from part (i) of the theorem is plotted for $M^{(d)} = 10$. The theoretical density from part (ii), i.e., $\chi^2(M^{(d)}, \nu_k^{(\tilde{c} \mapsto c)})$, is plotted for $M^{(d)} = 10$ and the non-centrality parameter estimates $\widehat{\nu}_k^{(\tilde{c} \mapsto c)}$ obtained by subtracting $M^{(d)}$ from the average deviance differences across the 500 realizations.

As it can be observed from Fig. 5.4, the theoretical predictions closely match the empirical estimates of the densities, even at a practical value of $\beta = 0.999$ close enough to unity and $W = 10$ (i.e., $N_{\mathsf{eff}} = 10000$). We confirmed that similar results hold for the rest of the links in the network, but have only plotted those corresponding to the aforementioned representative links for the sake of brevity.

Numerical Choices of Parameters: We now elaborate on the details of the parameter selection and estimation procedures used for the foregoing simulated example. We selected the modulation parameter vectors to be the same for all the G-causal interactions, and set to $\boldsymbol{\omega}_{\mathsf{exc.}} = [1, 0, 0, 2, 0, 0, 0, 0, 0, 1]$ for excitatory links and $\boldsymbol{\omega}_{\mathsf{inh.}} = -\boldsymbol{\omega}_{\mathsf{exc.}}$ for the inhibitory links, where each component corresponds to a uniform non-overlapping spike counting window of length 10 bins (or 10 ms). The modulation parameter vector associated with the non-existing G-causal links (such as $(8 \mapsto 2)$) is set to all zeros. The self-history dependence for all neurons is chosen to be of inhibitory and static nature to maintain stable behavior for simulation purposes. The norm of all non-zero parameter vectors is normalized to 1. The average spiking probability is set to $\bar{\lambda}\Delta \approx 0.07 \ll 1$ by choosing the baseline firing parameter $\mu_k = -2.597$ to be the same for all neurons.

To model the dynamics of the G-causal links in the second segment of the simulation, we enforce a linear time evolution for all the coefficients of underlying parameter vector, with a respective decay and growth for the links associated with neurons (1) and (5). For estimation of G-causal interactions, we select the sparse parameter vector associated with the full GLM model of neuron (c) to be in form of $\boldsymbol{\omega}_k^{(c)} = [\mu_k^{(c)}, \boldsymbol{\omega}_k^{(c,1)'}, \boldsymbol{\omega}_k^{(c,2)'}, \ldots, \boldsymbol{\omega}_k^{(c,C)'}]'$ of length $M^{(F)}$, composed of the scalar baseline parameter $\mu_k^{(c)}$, and sub-vectors $\boldsymbol{\omega}_k^{(c,c)}$ of length M_H^{self}, and $\boldsymbol{\omega}_k^{(c,\tilde{c})}$ of length M_H^{cross} for $\tilde{c} \neq c$, denoting the respective parameter vectors tuning the self-history dependence and the cross-history effects from neuron $(\tilde{c})$. We select $M_H^{\mathsf{cross}} = M_H^{\mathsf{self}} = 10$ history components associated with the respective kernel lengths of $L_H^{\mathsf{cross}} = L_H^{\mathsf{self}} = 100$ ms, obtained by non-overlapping windows of length $W_H = 10$ bins. Note that $M^{(F)} = 81$, and $M^{(R)} = M^{(F)} - M_H^{\mathsf{cross}} = 71$.

We employ the sparse adaptive filter ℓ_1–PPF_1 to estimate the sparse parameter vectors $\widehat{\boldsymbol{\omega}}_k$ at every time step k for both the full and reduced models. For the ℓ_1–PPF_1 filtering algorithm, an effective block length of $N_{\mathsf{eff}} = 10k$ is selected with a window size of $W = 20$, forgetting factor of $\beta = 0.998$ chosen sufficiently close to one, step size of $\varsigma := \frac{1-\beta}{W}$, and $L = 1$ number of iterations. The regularization parameter is tuned for each cell separately $\bar{\gamma}^{(c)} \in [0.3, 0.5]$, via the two-fold even-odd cross validation [33]. For the χ^2 filtering and smoothing algorithm, the smoothing and scaling factors are selected as $\sigma_e^2 = 5 \times 10^{-6}$ and $\rho = 1$, respectively, using an initialization of $z_{0|0} = 0$ and $\sigma_{0|0}^2 = 1$ in the EM algorithm.

For the performance comparison in Fig. 5.2–D, we have adapted the methods in [74] and [72], which are designed for static connectivity inference, to the time-

varying setting in full fairness. First, due to the batch-mode computation of these static methods, we divided the total $T = 120k$ observed bins to non-overlapping window segments of length $W^{\mathsf{ML}} = 10k$, matching the effective block length N_{eff} of our dynamic method. Both methods compute the ML estimates of the network parameters for each segment. We have therefore selected the true model orders $M^{\mathsf{ML}} = 10$ for both methods, matching the selected model order for our AGC inference method, so as to have a fair statistical comparison and to ensure that they operate at their optimal performance. (note that the dimensionality difference $M^{(d)}$ has a particularly pivotal role in the inference procedure).

The method in [74] computes a static GC connectivity map obtained from nested full and reduced ML estimates, followed by an FDR control procedure for correction of multiple comparison errors. The method in [72] performs a likelihood-ratio test to assess the significance of each pair-wise interaction. The same significance levels are chosen for the statistical tests in both methods, to match our FDR rate of $\alpha = 0.1$. Finally, both methods have been modified to the logit-linked GLM setting, in order to ensure their consistency with the generative model used for simulating the spike trains.

5.1.2 Robustness of AGC Inference to the Choice of Parameters

We inspect the robustness of the proposed AGC inference with respect to the choice of three major parameters: the dimensionality difference $M^{(d)}$, the regularization parameter γ, and the effective block length of the adaptive filter $T_{\mathsf{eff}} :=$

$N_{\text{eff}}\Delta = \frac{W\Delta}{1-\beta}$. We consider three different choices for each parameter, and for each choice, we run the simulated example in Fig. 5.2 for $R = 100$ repetitions, where a random sequence of spike trains are generated at each repetition based on the network dynamics of Fig. 5.1. In each setting, the rest of the parameters are chosen as described earlier in the previous subsection.

Robustness to the choice of $M^{(d)}$: For the dimensionality difference $M^{(d)}$, we consider three settings of $M^{(d)} \in \{10, 15, 20\}$. Fig. 5.5 shows the TDR and FAR performance results for different choices of $M^{(d)}$. While the FAR values remain consistently low (i.e., < 0.01, on average), as expected the larger choices of $M^{(d)}$ would impose stricter statistical thresholds on the hypothesis tests (See Fig. 4.3–B), leading to slight degradation of the TDR performance.

Robustness to the choice of γ: For the choice of regularization parameter,

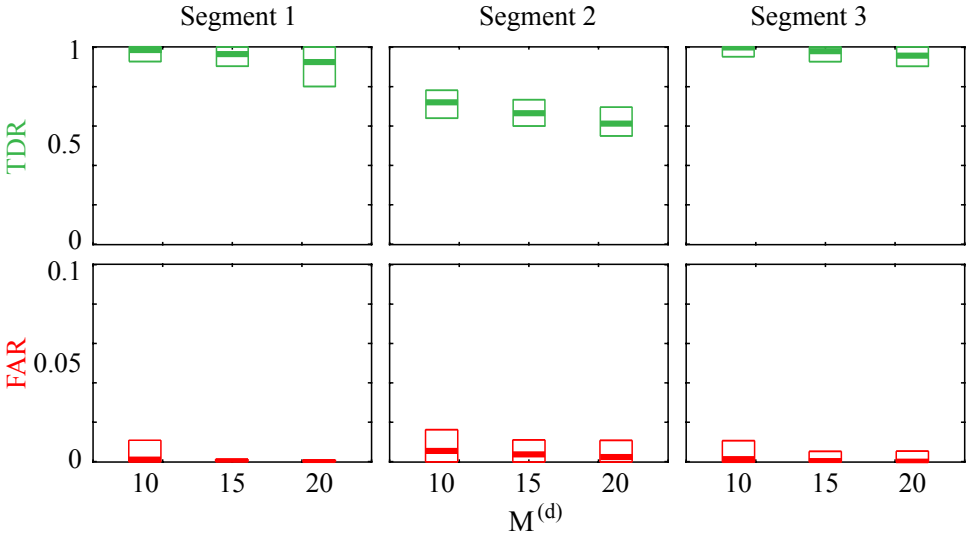

Figure 5.5: Performance of the AGC inference for three different values of $M^{(d)} \in \{10, 15, 20\}$, in terms of TDR (top row) and FAR (bottom row).

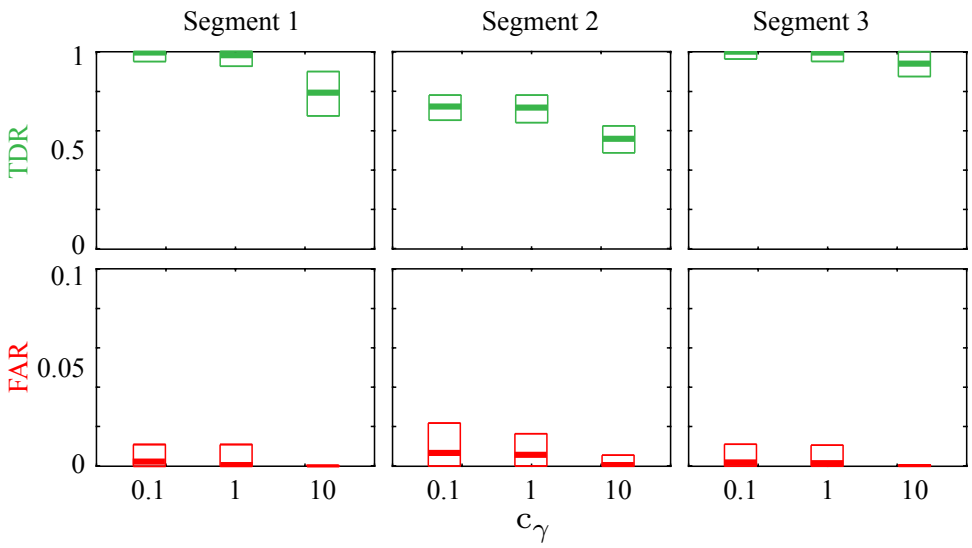

Figure 5.6: Performance of the AGC inference for three different scalings γ for $c_\gamma \in \{0.1, 1, 10\}$, in terms of TDR (top row) and FAR (bottom row).

we consider three different settings for $\boldsymbol{\gamma} = c_\gamma \boldsymbol{\gamma}^*, c_\gamma \in \{0.1, 1, 10\}$, where c_γ denotes a scaling factor and $\boldsymbol{\gamma}^*$ represents the optimally tuned regularization parameter vector obtained from cell-by-cell two-fold even-odd cross-validation. Fig. 5.6 reveals the robustness of the AGC inference with respect to the choice of the regularization parameter. It can be observed that the resulting performance metrics show resilience to under-regularization ($c_\gamma = 0.1$), while the TDR performance notably degrades due to over-regularization ($c_\gamma = 10$). This is due to the fact that larger choices of γ would shrink the inferred cross-history coefficients and thereby remove weaker GC effects, which would lead to reduced TDR (and FAR) performance for all segments. The optimally-tuned choice of the regularization parameter $\boldsymbol{\gamma} = \boldsymbol{\gamma}^*$ obtained via cross-validation achieves a favorable TDR-FAR performance trade-off.

Robustness to the choice of T_{eff}: For the effective filtering length, we select three different settings of $T_{\text{eff}} \in \{5, 10, 20\}$ s. Fig. 5.7 exhibits the significant

influence of effective filtering length T_{eff} on the performance of AGC inference, where as expected larger choices of T_{eff} would increase both the TDR and FAR metrics. In other words, larger effective number of samples for GC inference at each time step would increase both the capability of correct identification (due to increased estimation accuracy of the existing links) and the risk of false detection (due to increased effective observation noise).

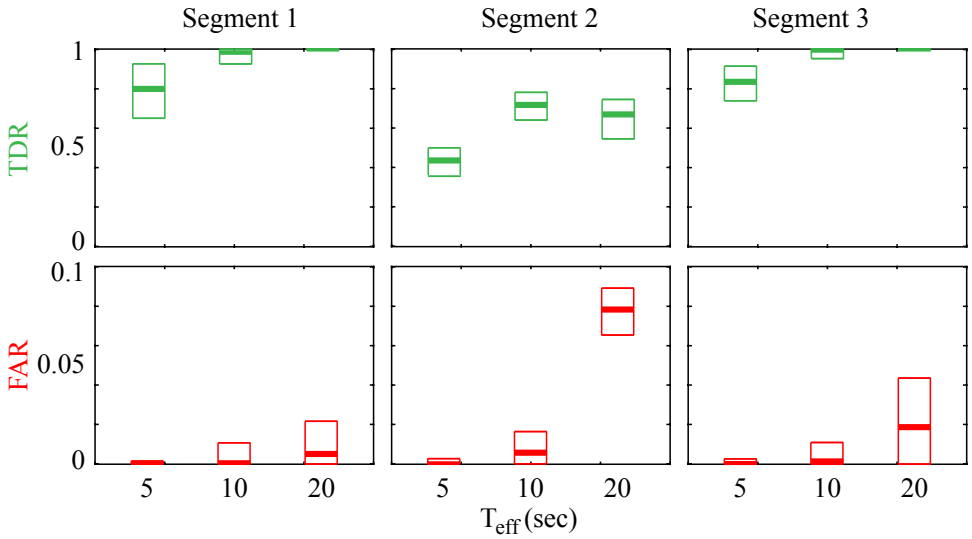

Figure 5.7: Performance of the AGC inference for three different values of $T_{\text{eff}} \in \{5, 10, 20\}$ s, in terms of TDR (top row) and FAR (bottom row).

5.1.3 The Roles of Adaptive Sparse Estimation and Bias Correction in AGC Inference

In this subsection, we inspect the roles of the bias correction procedure as well as sparse estimation in our proposed AGC inference method using an illustrative simulation study. We examine how these features affect the performance in terms of correct identification of the GC links and avoiding false positives. To this end,

we compare the performance of our AGC inference method with a biased variant in which the bias correction stage is removed, as well as the static ML-based GC inference [74], in which the dynamics and sparsity are not taken into account.

We consider $R = 100$ realizations of a random network configuration comprising $C = 10$ neurons, causally interacting through $N_{\mathsf{Links}} = 10$ randomly selected directional links. For each repetition and given the network configuration, a sequence of spike trains with a duration of $\mathcal{T} = 30$ s is generated with a bin size of $\Delta = 1$ ms. The average baseline spiking probability is set to $\bar{\lambda}\Delta = 0.05$. For spike generation, we use a logit-linked GLM model with a static block-sparse parameter vector $\boldsymbol{\omega}^{(\tilde{c},c)}$ with a support set of $\mathcal{S} = \{1, 5, 10\}$, and respective values of $(\boldsymbol{\omega}^{(\tilde{c},c)})_{\mathcal{S}} = \{2, -1, 1\}$ to model the self- and cross-history dependencies among neurons. Each history component is associated with a non-overlapping history window of $W_H = 5$ time bins. The sign of the history kernel determines the aggregate excitatory or inhibitory effect of a causal link. We assume self-excitatory behavior for all neurons, and the excitatory/inhibitory nature of the cross-history interactions are randomly selected for each link. For GLM estimation, we select $M_H^{\mathsf{cross}} = M_H^{\mathsf{self}} = 10$ history components, associated with spike counting windows of length 5 time bins. For the ℓ_1–PPF_1 algorithm, we select the forgetting factor $\beta = 0.999$, and a filtering window size of $W = 5$ bins (corresponding to an effective window length of 5 s), $c_w = 1$ and $L = 1$ number of iterations. The regularization parameter γ is tuned for each neuron via two-fold even-odd cross-validation. The χ^2 filtering and smoothing algorithm parameters are chosen as $\sigma_e^2 = 5 \times 10^{-6}$ and $\rho = 1$, and the FDR is controlled at a significance level of $\alpha = 0.1$.

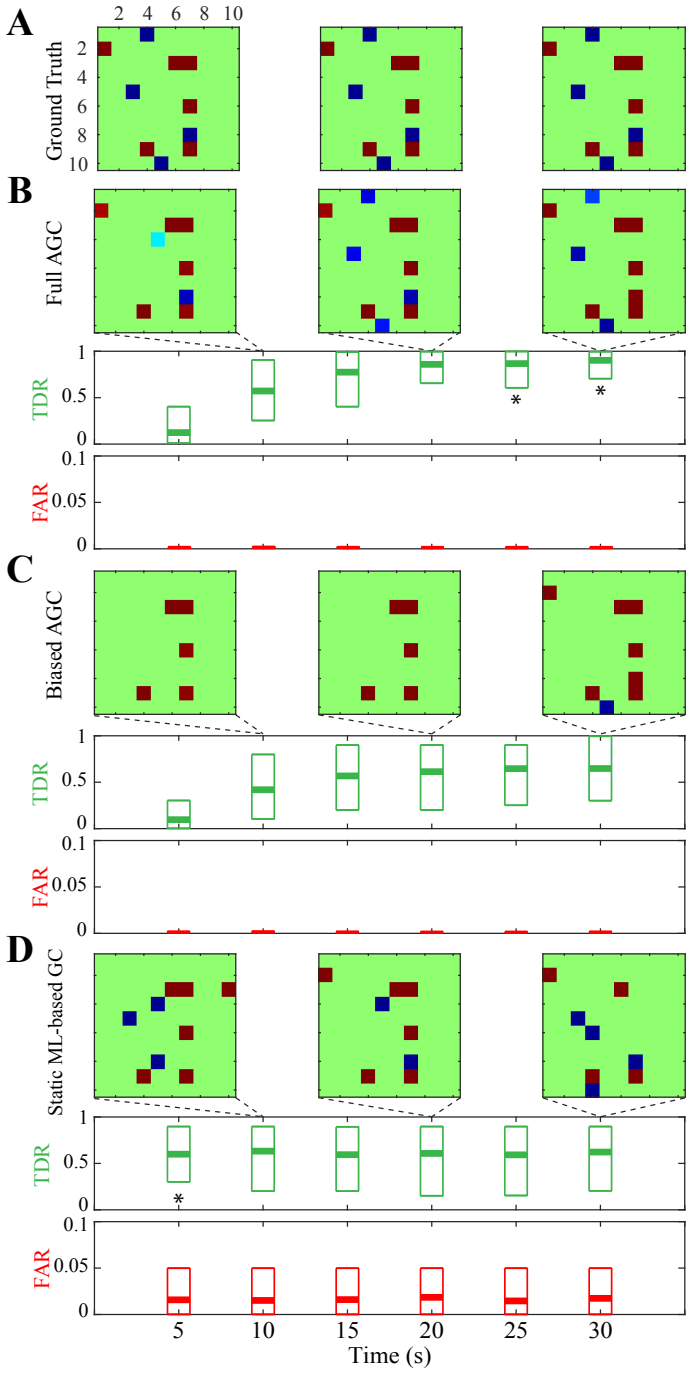

Figure 5.8: Performance comparison of AGC inference (B) to its biased variant (C) and static ML-based GC inference (D). A) the ground truth GC maps in a network of 10 neurons. The top rows of B, C and D, correspond to three snapshots of the network inference result for a given realization, and the bottom rows show the TDR and FAR metrics computed for the six non-overlapping segments, pooled across 100 realizations. Stars indicate significant differences between the AGC and static ML-based GC, with effect sizes of $r \geq 0.8$ (Wilcoxon signed-rank test, $p < 0.001$).

Fig. 5.8–A shows the static ground truth GC maps for a selected realization, plotted at three time instances in the form of 10×10 matrices. Fig. 5.8–B, C and D show the results for the full AGC inference method, its biased variant with no bias correction, and the ML-based static GC method, respectively. Each panel consists of three rows: snapshots of the detected GC maps for one realization (top row), and the TDR (second row) and FAR (third row) performance metrics computed for consecutive non-overlapping 5 s windows, and pooled across the $R = 100$ repetitions. Boxes show the mean and 90% confidence regions. The static GC maps in Fig. 5.8–D are estimated for non-overlapping windows of length $T_{\mathsf{eff}} = 5$ s, equal to the effective filtering block length of the AGC method.

Figs. 5.8–B and C reveal the favorable FAR performance of the AGC method as compared to the static ML, even in the absence of bias correction. In particular, based on the Wilcoxon signed-rank test with a p-value of $p < 0.001$, the FAR performance of AGC inference is significantly lower than that of the static ML for all segments (effect sizes of $r = 0.44, 0.38, 0.44, 0.52, 0.37$, and 0.37 in the six segments, respectively). However, the lack of bias correction (Fig. 5.8–C) results in lower TDR performance compared to the AGC method. The TDR performance of the AGC method is also significantly higher than that of the static ML for the last 4 segments, but is outperformed by the static ML in the first segment (effect sizes of $r = 1, 0.17, 0.73, 0.74, 0.88$ and 0.90, in the six segments, respectively), which is expected due to the initialization period of ~ 5 s for the AGC method.

This illustrative example shows that the static ML-based approach that does not account for sparsity overfits the parameters when applied to limited data, and

hence results in low true positive performance and a high number of spurious link detections. In comparison, the AGC method provides favorable TDR and FAR performance, but only after the initialization period, which is of the order of the effective window length. In addition, this example highlights the crucial role of bias correction for the deviance difference statistics in our proposed statistical inference procedure.

5.1.4 Robustness Against Latent Confounding Causal Effects: Three Simulation Studies

In this subsection, we investigate the robustness of our proposed AGC inference method with respect to *latent confounding causal effects*, which is one of the major challenges in causality inference. When the two time series X_t and Y_t subject to GC inference are driven by a third latent common process Z_t, with possibly different latencies, i.e., $(X_t \leftarrow Z_t \rightarrow Y_t)$, the GC inference may lead to spurious detection of causal effects between X_t and Y_t. This is due to the fact that the common information from Z_t introduced in both X_t and Y_t cannot be captured by the conditional covariates due to the latent nature of Z_t, and may result in false positive errors, thereby limiting the reliability of GC inference.

Although the original form of the GC measure does not take into account the latent confounding causal effects, several solutions have been proposed in the literature to resolve this issue. As an example, a variant of GC called "partial G-causality" is introduced in [110], which shows superior performance in terms of

removing the effects of hidden confounding influences compared to the conditional GC.

Our proposed method for AGC inference mitigates this issue through several mechanisms. First, the hypothesis of sparsity allows for stable estimation of high order GLM models, with large number of self-history components in both the reduced and full models used in the conditional GC measure. Hence, we expect that the latent effects are partially captured via the high order self-history parameters due to the autoregressive nature of the GLM models, which promotes the detection of the actual GC links between the units using the cross-history components. This feature is akin to estimating latent Moving Average (MA) components using autoregressive models in the ARMA modeling paradigm.

Second, by explicitly modeling the dynamics of the non-centrality parameters describing the deviance statistics, and thereby using the χ^2 filtering and smoothing algorithm to reliably estimate them, we expect that only the temporally-salient G-causal effects are captured, and the transient G-causal links possibly due to confounding influences manifested in the deviance statistics are suppressed.

Third, the non-centrality parameter estimates allow us to characterize the test strengths for the rejected nulls (i.e., detected GC interactions) obtained by the FDR-controlled multiple hypothesis testing framework, in a model-based fashion. The resulting J-statistics can be further used to reject the detected GC links with low test strength, which may be due to transient latent effects.

In order to demonstrate these features, we test the performance of the proposed AGC inference method in the presence of confounding causal effects under the fol-

lowing three scenarios: 1) confounding deterministic common input, 2) confounding stochastic common input, and 3) confounding effects due to network subsampling.

Scenario 1: Confounding Effects Due to Latent Deterministic Common Input. We first consider an illustrative two-neuron example. We consider a setting with a GC-link from neuron (2) to (1), and no GC link in the opposite direction. We also consider a hidden (confounding) source (H) affecting both neurons. We assess the robustness of AGC inference method in terms of two performance metrics: the detected false alarm rate (FAR) corresponding to the link ($1 \mapsto 2$), and the true detection rate (TDR) for correctly identifying the link ($2 \mapsto 1$), all in the presence of the confounding source ($H \mapsto 1, 2$) (Fig. 5.9–A).

We assume a stationary environment with static GC links, and we use the same spiking statistics based on logit-linked GLM model. We consider the case with no self-history dependence in order to more specifically inspect the trade-off between the cross-history and the latent confounding influences on our GC inference procedure. To model the cross-history dependence associated with the GC link ($2 \mapsto 1$), we select a uniform modulation vector $\boldsymbol{\omega}_k^{(1,2)} = \frac{1}{\sqrt{W^{(1,2)}}}\mathbf{1}_{W^{(1,2)}}$ covering a window of $W^{(1,2)}$ time bins, where $\mathbf{1}_{W^{(1,2)}}$ denotes the vector of all ones of length $W^{(1,2)}$. The effect of the latent hidden source is later added to the contributing effects in the GLM models for both neurons. For the estimation of the GLM models, a larger number of $M_H^{\mathsf{self}} = 5 \times M_H^{\mathsf{cross}}$ self-history components are considered compared to the cross-history in order to better capture the effects of the latent confounding influences.

In this first scenario, we model a sinusoidal signal $x_k = A_H \sin(2\pi k/200)$ afferent to neurons (1) and (2) as the latent common input with different delays. We consider a phase difference of $\pi/2$ between the latent inputs to neurons (1) and (2) to account for the delay. For simulation of this scenario, we select a ground-truth cross-history window of $W^{(1,2)} = 100$ time bins, and a uniform non-overlapping spike counting window of length $W_H = 50$ for parameterizing the history components, and $M_H^{\mathsf{cross}} = 20$ number of cross-history components.

For fairness of comparison, we choose the mean power $\mathcal{E}_H$ of the latent confounding source to be equal to the mean power of the G-causal link $\mathcal{E}_{GC} := \mathrm{var}(\boldsymbol{\omega}_k^{(1,2)'} \mathbf{x}_k^{(1,2)})$ for the first two scenarios, and denote them by $\mathcal{E}$. We run the simulation for $R = 50$ repetitions, where a spike train of $T = 180k$ samples covering a duration of $\mathcal{T} = 180$ s is generated with $\Delta = 1$ ms time bins for each realization. For the ℓ_1–PPF$_1$ sparse filtering setup, we selected an effective block length of $N_{\mathsf{eff}} = \frac{W}{1-\beta}$ in the set $\{10k, 20k, 100k\}$, with respective average spiking probabilities of $\bar{\lambda}\Delta \in \{0.1, 0.05, 0.01\}$. The regularization parameter γ is tuned for both neurons via two-fold even-odd cross-validation. For the χ^2 filtering and smoothing setup, we selected a scaling factor $\rho = 1$, and a smoothing factor $\sigma_e^2 = 10^{-4}$. We infer the GC links for each repetition, and finally measure the mean FAR and TDR across all realizations.

Table 5.1 exhibits the (FAR, TDR) performance pairs for six different settings of $(\bar{\lambda}\Delta, \mathcal{E})$ for the sinusoidal latent source, pooled across all the repetitions. Each row and column correspond to specific choices of the average spiking probability $\bar{\lambda}\Delta$ and mean confounding power $\mathcal{E}$, respectively. The effective number of spikes per filtering

window $n_{\mathsf{eff}} := N_{\mathsf{eff}}\bar{\lambda}\Delta$ is chosen to be the same across all rows. We selected two different values of $\mathcal{E} \in \{0.01, 0.05\}$. The FDR is controlled at the respective rates of $\alpha = 0.1$ and 0.05 for $\mathcal{E} = 0.01$ and 0.05. The respective small and large values of FARs and TDRs in the entries of Table 5.1 reveal the utility of our proposed method in suppressing the effect of confounding latent causal influences, while identifying the true G-causal links between the two neurons with high sensitivity and specificity. In addition, they suggest that high-order self-history components are capable of capturing deterministic latent effects. It is worth mentioning that the six different settings in Table 5.1 are chosen to span the low-spiking ($\bar{\lambda}\Delta = 0.01$) and high-spiking ($\bar{\lambda}\Delta = 0.1$) regimes, both in presence of weak ($\mathcal{E} = 0.01$) and strong ($\mathcal{E} = 0.05$) confounding effects.

In order to illustrate the aforementioned features of our proposed method in detecting the salient effects, and characterizing the corresponding test powers, we have shown one realization from Table 5.1 in Fig. 5.9, corresponding to the setting $(\bar{\lambda}\Delta, \mathcal{E}) = (0.05, 0.01)$. The corresponding spike trains of neurons (1) and (2) within a short window of 4 s are shown in Fig. 5.9–B. Fig. 5.9–C shows the time-course

$\bar{\lambda}\Delta$ \ $\mathcal{E}$		0.01	0.05
0.01	FAR	0.01 ± 0.05	0.13 ± 0.20
	TDR	0.66 ± 0.32	1
0.05	FAR	0.11 ± 0.11	0.15 ± 0.10
	TDR	0.89 ± 0.11	1
0.1	FAR	0.15 ± 0.09	0.12 ± 0.06
	TDR	0.90 ± 0.08	1

Table 5.1: Performance metrics of AGC inference in presence of a latent deterministic sinusoidal common input. Entries show mean +/- standard deviation.

of the estimated non-centrality parameter $\widehat{\nu}_k^{(1\mapsto 2)}$ associated with the false positive error $(1 \mapsto 2)$ (blue trace), and $\widehat{\nu}_k^{(2\mapsto 1)}$ associated with the true positive excitatory G-causal link $(2 \mapsto 1)$ (red trace). The black traces show the shifted deviance differences $D_{k,\beta} - M^{(d)}$. Fig. 5.9–D shows the time-course of the estimated J-statistics corresponding to the existing GC link $(2 \mapsto 1)$ (red trace) and the non-existing $(1 \mapsto 2)$ link (blue). As expected, the existing GC link is detected in a temporally-salient fashion with high test strength, whereas the non-existing link is overwhelmingly rejected, with low test strength otherwise. In order to highlight

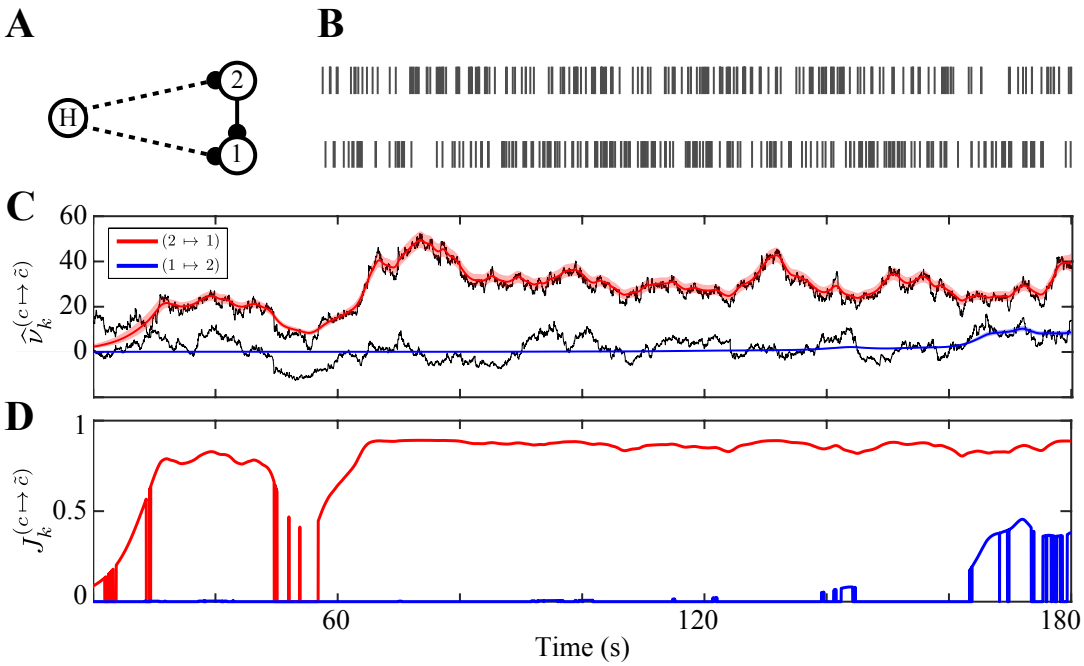

Figure 5.9: Performance of the AGC inference method in presence of the latent confounding causal effects corresponding to the realization from Table 5.1 with median performance metric pair (FAR, TDR) at the setting $(\bar{\lambda}\Delta, \mathcal{E}) = (0.05, 0.01)$. A) dual-neuron network model with hidden source, B) spike trains of both neurons within 4 s window, C) estimated non-centrality $\widehat{\nu}_k$ corresponding to GC links $(1 \mapsto 2)$ (blue) and $(2 \mapsto 1)$ (red) across time, along with the 95% confidence regions, and the shifted deviance differences $D_{k,\beta} - M^{(d)}$ (black traces), D) estimated J-statistics J_k for both GC links.

the effect of capturing the latent input using long self-history kernels, a sample of the estimated self-history coefficients of neuron (2) in the sinusoidal input scenario is shown in Fig. 5.10. The coefficients that are away from zero at a significance level of 90% are highlighted in black. The estimated sparse high-order self-history components are able to capture the sinusoidal latent input based on the temporal correlations in the spiking history of the neuron.

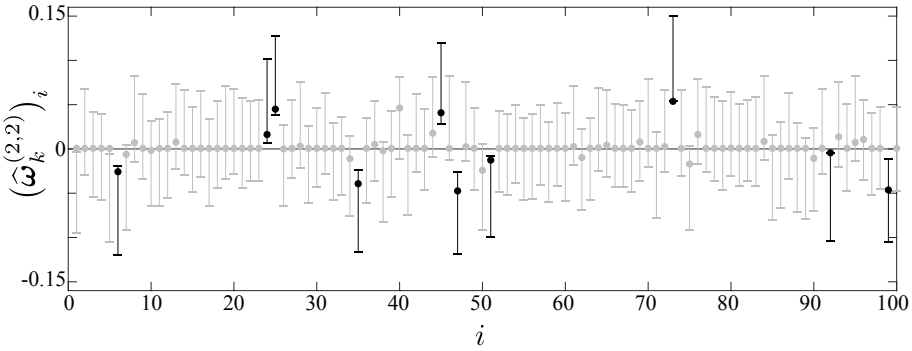

Figure 5.10: A sample of the self-history coefficient of neuron (2) in the latent sinusoidal common input scenario for a generic trial and time window. The error bars show 90% confidence intervals. Coefficients that are significantly away from zero are further highlighted in black.

Scenario 2: Confounding Effects Due to Latent Stochastic Common Input. For the second scenario, we consider a similar setting as the previous one, but instead of a deterministic input, we generate a high-order AR process to model a general stochastic latent confounding effect. We use a block-sparse structure for the AR kernel with parameters $\boldsymbol{\omega}_{\mathsf{AR}} = [0.7, 0, 0, 0, -0.05, 0, 0, 0, 0.02]'$, where each coefficient is associated with a non-overlapping spike counting window of length $W_H = 25$ bins. The AR coefficients are normalized to result in a stable process. We selected an arbitrary delay of 40 bins between the common input to the two neurons. A uniform ground-truth cross-history window of length $W^{(1,2)} = 50$, and

$M_H^{\mathsf{cross}} = 10$ number of cross-history components are selected for this setting. All the other parameters used for AGC inference are chosen similar to the previous scenario.

In the same vein as Table 5.1, Table 5.2 exhibits the (FAR, TDR) performance pairs for six different settings of $(\bar{\lambda}\Delta, \mathcal{E})$ for the AR latent source. Similarly, the respective small and large values of FARs and TDRs in the entries of Table 5.2 confirm the utility of our proposed method in suppressing the effect of stochastic latent common causal influences.

$\bar{\lambda}\Delta$ \ $\mathcal{E}$		0.01	0.05
0.01	FAR	0.07 ± 0.18	0.05 ± 0.16
	TDR	0.81 ± 0.29	1
0.05	FAR	0.10 ± 0.09	0.06 ± 0.07
	TDR	0.92 ± 0.11	1
0.1	FAR	0.09 ± 0.06	0.06 ± 0.05
	TDR	0.93 ± 0.05	1

Table 5.2: Performance metrics of AGC inference in presence of a latent stochastic AR common input. Entries show mean +/- standard deviation.

Scenario 3: Confounding Effects due to Network Subsampling. In the third scenario, we evaluate the performance of our proposed AGC inference method in the context of the more general confounding setting of *network subsampling*. This scenario occurs when the observable neurons are subsampled from a large neuronal network, and are prone to significant confounding effects from the unobserved portion of the network, as illustrated in Fig. 5.11. This scenario often happens in the analysis of experimentally recorded data, in which the observable neuronal ensemble consists of a small subset of a larger latent network of neurons, due to the physical limitations of data acquisition.

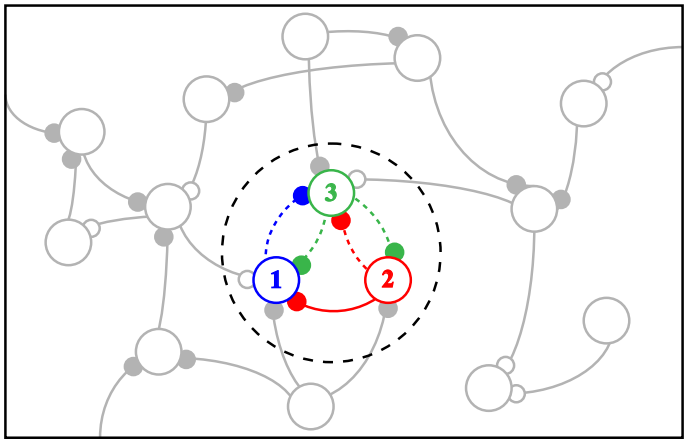

Figure 5.11: A schematic depiction of the network subsampling scenario. A small observable subnetwork of three neurons (within the dashed circle) are sampled from a large latent neuronal network. The observable subnetwork and the interactions within are represented by blue, red and green colors, while the latent neurons and interactions are shown in gray.

In order to test the robustness of our method to the problem of network subsampling, we consider a network of $C = 20$ neurons, where the AGC inference is performed on a small subnetwork of $C_{\mathsf{sub.}} = 3$ observable neurons. We repeat the network subsampling simulation for $R = 100$ realizations, where a random network configuration consisting of $N_{\mathsf{Links}} = 40$ links randomly selected out of 380 possible directional links is considered for each realization. To determine the observable ensemble for AGC inference, we randomly select a subset of $C_{\mathsf{sub.}}$ neurons, such that there would be at least one direct latent common input to a pair of causally-linked observable neurons (e.g. neurons 1 and 2 in Fig. 5.11). For simulation, we consider a static setting for the GC links, where the underlying parameters remain constant throughout the entire duration. We use a block-sparse kernel of $\boldsymbol{\omega}_H = [2, -1, 0, 0, -0.5, 0, 0, 0, 0, -0.5]'$ with non-overlapping history windows of length $W_H = 5$ bins, to model the self- and cross-history dependence among the

causally interacting neurons. The effective excitatory or inhibitory natures of the GC links are determined by positive ($\boldsymbol{\omega}^{(\tilde{c},c)} = +\boldsymbol{\omega}_H$) or negative polarity ($\boldsymbol{\omega}^{(\tilde{c},c)} = -\boldsymbol{\omega}_H$) of the kernel. We assume the self-history dependence to be of excitatory nature for all neurons, and the probability of excitatory or inhibitory cross-history dependence is set to 50% for all links.

For each realization, we generate spike trains with a total duration of $\mathcal{T} = 180$ s with $\Delta = 1$ ms time bins. For estimation of the GLM models, a total number of $M_H^{\mathsf{cross}} = 10$ cross-history and $M_H^{\mathsf{self}} = 20$ self-history components are considered with a spike counting window of length 5 for parameterizing the history components.

We repeat the network subsampling simulation and perform the AGC inference for six different settings of $(\bar{\lambda}\Delta, \mathcal{E})$ pairs, similar to the Tables 5.1 and 5.2, where two different values of the mean GC link power $\mathcal{E} \in \{0.01, 0.03\}$ and three different values of the average spiking probability $\bar{\lambda}\Delta \in \{0.01, 0.05, 0.1\}$ are selected. We use the same parameter settings for the ℓ_1–PPF_1 filter and the non-central χ^2 filtering and smoothing as in the previous two scenarios for the three different $\bar{\lambda}\Delta$ settings. The regularization parameter γ is tuned for each observable neuron separately via two-fold even-odd cross-validation. The FDR is controlled at a significance level of $\alpha = 0.1$.

As before, we evaluate the performance of AGC inference in terms of two performance metrics: FAR and TDR within the observable network across all realizations. Table 5.3 summarizes the performance results for the six different settings. The resulting metrics reveal the remarkable performance of our proposed AGC inference in suppressing the false positives due to the latent confounding causal effects

$\bar{\lambda}\Delta$ \ $\mathcal{E}$		0.01	0.03
0.01	FAR	0.01 ± 0.03	0.04 ± 0.07
	TDR	0.74 ± 0.27	0.99 ± 0.01
0.05	FAR	0.01 ± 0.01	0.01 ± 0.02
	TDR	0.71 ± 0.14	1
0.1	FAR	0.01 ± 0.01	0.01 ± 0.01
	TDR	0.68 ± 0.11	1

Table 5.3: Performance metrics of AGC inference under network subsampling. Entries show mean +/- standard deviation.

(low FAR rate of $\sim 1\%$, on average), while maintaining high true detection rates (high TDR rate of $\sim 70\%$, on average). Together with the results of the two foregoing scenarios, these results corroborate our earlier assessment of the AGC inference in maintaining a degree of immunity to latent confounding effects.

5.2 A Simulation Study on GC Inference from Imaging Data

To validate the GC inference method introduced in section 4.2 for imaging data modalities, and particularly highlight the roles of sparse estimation and bias correction, we provide an illustrative simulation study with known ground truth causal links. To this end, we compare the performance of our proposed method with the conventional ML-based GC inference technique, as well as GC inference from sparse estimates with no bias correction.

We consider a network of $C = 8$ neurons, causally inter-connected based on the pattern shown in Fig. 5.12–B (the leftmost panel). For each neuron, we simulate a sequence of fluorescence measurements of length $T = 150$ samples, across $R = 10$ trial repetitions, based on the VAR model in 4.23, driven by a zero-mean

Gaussian sequence of variance $\sigma^2 = 0.002$. Fig. 5.12–A shows a simulated trial for all 8 neurons. To simulate the fluorescence activity, we consider a random sequence of discrete events uniformly distributed over time with a low probability of $p = 0.33$ event/trial, to represent the VAR innovation sequences (i.e., spikes). We select a sparse parameter vector $\boldsymbol{\theta}$ of length $M = 20$ with a support set of $\mathcal{S} = \{1, 5, 15, 20\}$ of sparsity $S = 4$, and respective values of $\boldsymbol{\theta}_{\mathcal{S}} = \{0.3, 0.05, 0.05, 0.05\}$ to model the self and cross-history dependencies among neurons. The cross-history parameters are chosen to be the same for all the existing G-causal links, i.e., $\boldsymbol{\omega}^{(c,\tilde{c})} = \boldsymbol{\theta}$, and have a positive or negative sign for the excitatory or suppressive GC links, respectively. In addition to the cross-history effects, we assume self-history dependence of excitatory nature, $\boldsymbol{\omega}^{(c,c)} = \boldsymbol{\theta}$, and a scalar fluorescence baseline parameter, $\mu^{(c)} = 0.01$, consistent across all neurons. The underlying ground truth functional pattern is unknown to the estimator, and is to be inferred from the simulated fluorescence traces. For estimation of GC links, we consider $M_H^{\mathsf{cross}} = M_H^{\mathsf{self}} = 50$ history components associated with the lag of $L_H = 50$ samples (with $W_{H,m} = 1$). For sparse estimation, the regularization parameter γ is separately tuned for each neuron via cross-validation across trials. The FDR is controlled at a rate of $\alpha = 0.05$, for simultaneous testing of 56 possible GC links.

Figs. 5.12–B and C show the comparison of our proposed method (2nd panels) with: 1) sparse estimation without bias correction (3rd panels), and 2) the conventional GC inference based on ML estimates (4th panels), in graphical and matrix forms, respectively. The GC maps of Fig. 5.12–C represent 8×8 color-coded arrays, where red, blue and green show excitatory, suppressive and no link,

respectively. The first method (3rd panels) is the biased variant of our proposed GC inference framework, where we adopt a similar estimation procedure and inference, but disregard the bias correction term incurred by the sparse regularization in Eq. 4.25. In the second method, the sparsity of the parameters is not taken into account, and the parameter estimation and deviance computation are performed based on ML estimates.

Figs. 5.12–B and C reveal that our presented method outperforms the two other compared methods in terms of both identification of the true GC links and avoiding false discoveries: while our proposed method matches the ground truth, removing the bias correction step results in low hit rate, and not accounting for sparsity results in high false alarm rate. This illustrative example highlights the crucial role of bias correction for the deviance difference statistics in our proposed

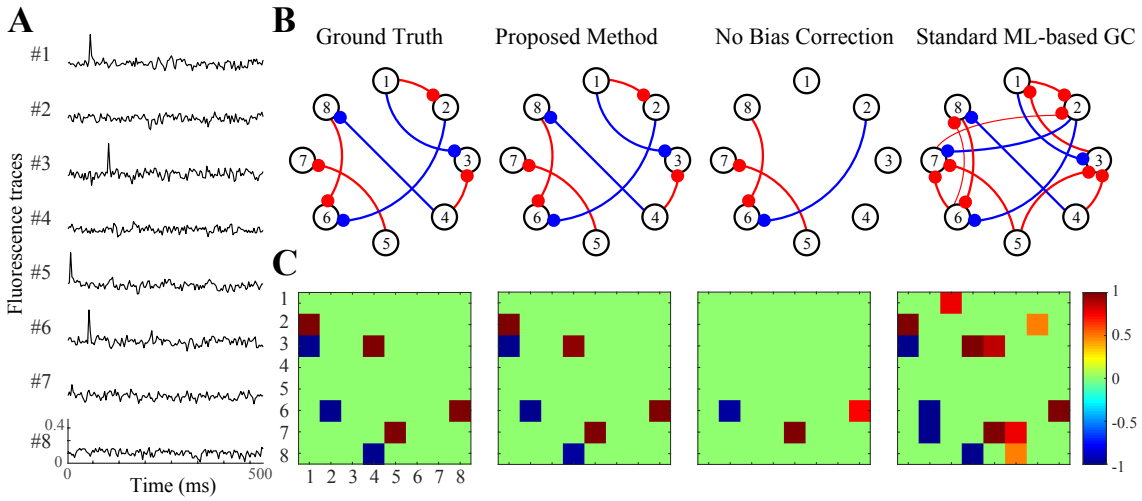

Figure 5.12: Illustrative simulated example of GC network inference with known ground truth. A) Simulated fluorescence traces from a single trial. B) Graphical functional network maps corresponding to the ground truth, our proposed inference method with and without bias correction, and the ML-based GC inference. C) The network maps of the abovementioned methods in matrix form.

statistical inference procedure. In addition, it shows that ML-based approaches that do not account for sparsity overfit the parameters when applied to limited data, and hence result in significant false discoveries.

Chapter 6: Application to Experimental Data

This chapter contains our results from the application of the proposed GC inference methods to experimentally recorded data. In the first section, we present the results obtained from AGC analysis of neural spiking data. In the second section, we provide the GC inference results from continuous-valued optical imaging data.

6.1 Application to Neural Spiking Data

In this section, we use data from two experimental settings: 1) inferred spike trains from two-photon imaging data recorded from the mouse auditory cortex under spontaneous activity, and 2) simultaneous spike recordings from the ferret primary auditory (A1) and prefrontal cortices (PFC) under a tone-detection task.

6.1.1 Application 1: Spontaneous Activity in the Mouse Auditory Cortex

In this subsection, we apply our proposed method to experimentally recorded neuronal population data from the mouse auditory cortex. We imaged the spontaneous activity in the auditory cortex of an awake mouse with *in vivo* two-photon calcium imaging. Within an imaged field of view, the activity of $N_{\mathsf{cells}} = 219$ neurons

is recorded at a sampling rate of $f_s \approx 30$ Hz for a total duration of $\mathcal{T} \approx 22$ mins. Spike trains are inferred from the fluorescence traces using the constrained-foopsi technique [111]. For GC inference, a subset of $C = 20$ neurons exhibiting high spiking activity were selected, as many of the neurons in the ensemble are relatively silent. The FDR is controlled at a rate of $\alpha = 0.005$ for testing the $|\mathcal{C}| = 380$ possible GC links.

Figs. 6.1–A and 6.1–B show the time-course of the non-centrality estimates and J-statistics for four selected candidate GC links, respectively. These representative GC links consist of a persistent link $(6 \mapsto 4)$ (blue), two transient links $(1 \mapsto 18)$ (red) and $(2 \mapsto 12)$ (green), and an insignificant link $(8 \mapsto 9)$ (magenta). Figs. 6.1–C and 6.1–D show four snapshots of the AGC map estimates, respectively in the matrix form, and as a network overlaid on the slice, at time-stamps $\{8.33, 11.66, 16.66, 22.22\}$ mins. Other than the three color-coded significant links, the rest of the detected G-causal links are indicated by black arrows.

The detected G-causal maps are considerably sparse (maximum ~ 16 out of 380 possible links), with a few persistent GC links and a multitude of transient links emerging and vanishing over time. The sparsity of the AGC maps is consistent with sparse activity in auditory cortex [112]. A careful inspection of the spatial pattern of the AGC links reveals that the detected links correspond to distances in the range of $[150, 200]$ μm. These distances are consistent with *in vitro* measurements of the spatial patterns of intra-laminar connectivity within the mouse auditory cortex [113], showing a significant peak in the connection probability within the mean radial range of $[120, 200]$ μm. These results indicate that the proposed AGC method is able to

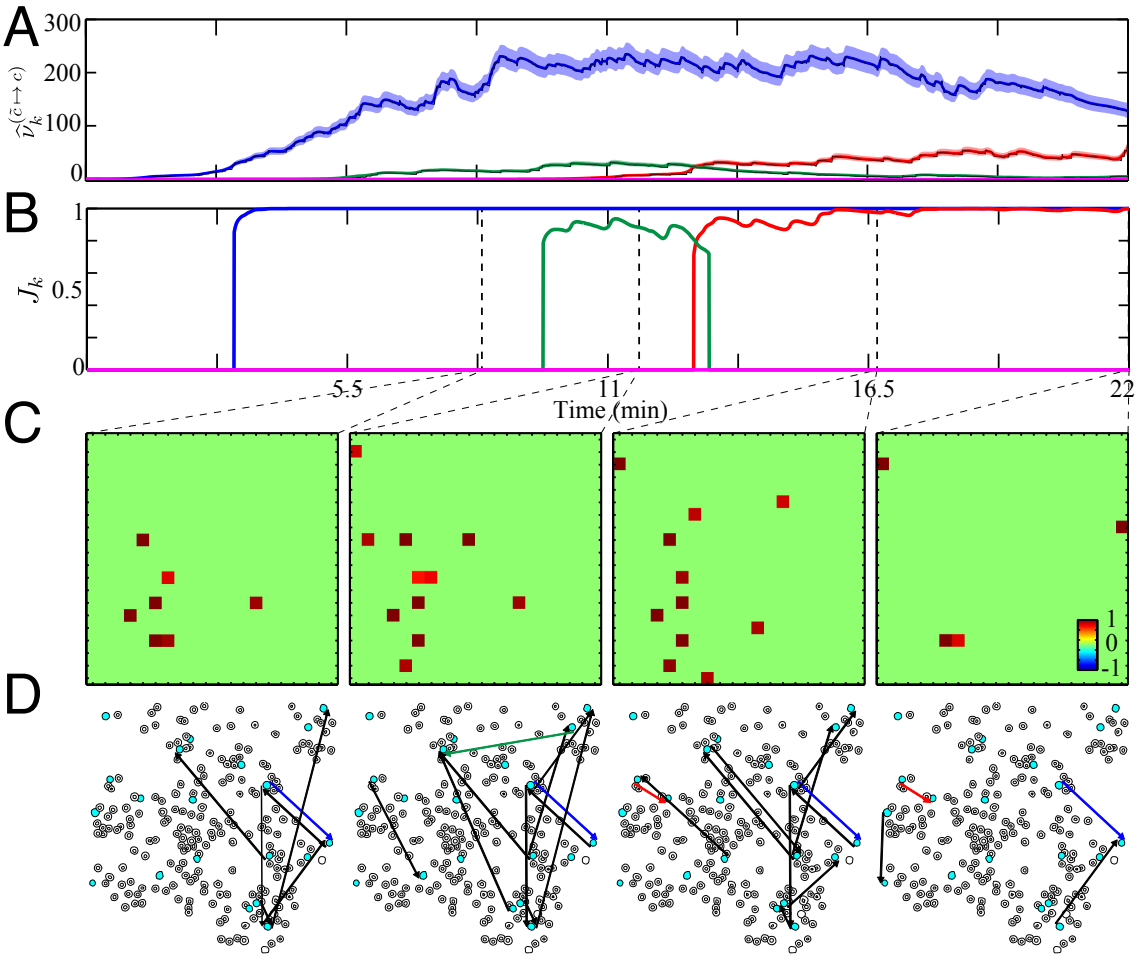

Figure 6.1: Adaptive G-causal interactions among ensemble of neurons in mouse auditory cortex under spontaneous activity. The time-course of estimated GC changes for four selected GC links obtained through A) non-centrality parameter $\widehat{\nu}_k$, and B) J-statistics J_k. C) AGC map estimates $\widehat{\Phi}_k$ at four selected points in time, marked by the dashed vertical lines in the top panel. D) network maps overlaid on the slice, showing cells with black circles and the selected cells highlighted in cyan. The detected GC links depicted in black directed arrows and colored for the selected links.

detect underlying connectivity patterns among neurons.

Numerical Choices of Parameters: We used time bins of length $\Delta =$ 33 ms, equal to the sampling interval. For the GLM models, we chose $M_H^{\text{cross}} = 3$ cross-history components associated with a block length of $L_H^{\text{cross}} = 10$ samples, ob-

tained by non-overlapping windows of length $[2, 4, 4]$ samples. To correct for possible latent confounding effects, we select a larger self-history kernel of $L_H^{\mathsf{self}} = 30$ samples segmented using windows of $[2, 4, 4, \ldots, 4]$ samples, giving a total of $M_H^{\mathsf{self}} = 8$ parameters. We corrected for the clustered spike detection effect of the constrained-foopsi method, using a masking window for rejecting multiple consecutive spikes. We selected an optimal data-driven masking window of size $W^{\mathsf{mask}} = 8$ samples, obtained by computing minimum rise-time of the calcium peaks inferred from the smoothed fluorescence traces of all cells. Then, the spikes detected within an interval of length W^{mask} are rejected.

We employ ℓ_1–PPF_1 algorithm for adaptive parameter estimation with a forgetting factor of $\beta = 0.999$, a window size of $W = 10$ bins, and $L = 1$ number of iterations. The regularization parameter was tuned for each cell via two-fold even-odd cross validation. The χ^2 filtering and smoothing algorithm parameters are chosen as $\rho = 0.999$, and $\sigma_e^2 = 10^{-3}$. The J-statistics are evaluated at the mean FDR for the detected GC links.

6.1.2 Application 2: Ferret Cortical Activity During Attentive Auditory Behavior

Studies of the prefrontal cortex (PFC) have revealed its association with high-level executive functions such as decision making and attention [114–116]. In particular, recent findings suggest that PFC is engaged in cognitive control of auditory behavior [116], through a top-down feedback to sensory cortical areas, resulting in

enhancement of goal-directed behavior. It is conjectured in [117] that the top-down feedback from PFC triggers adaptive changes in the receptive field properties of A1 neurons during active attentive behavior, in order to facilitate the processing of task-specific stimulus features. This conjecture has been examined in the context of visual processing, where top-down influences exerted on the visual cortical pathways have been shown to alter the functional properties of cortical neurons [118, 119].

In order to examine this conjecture at a single-unit level, we apply our proposed AGC inference method to single-unit spiking activities from an ensemble of neurons simultaneously recorded from two cortical regions of A1 and PFC in ferrets during a series of passive listening and active auditory task conditions. In this application, we sought to reveal the significant task-specific changes in the G-causal interactions within or between PFC and A1 regions at the single-unit level during active behavior. We used the spike data recordings from a large set of experiments (more than 35) conducted on three ferrets for GC inference analysis (data from the Neural Systems Laboratory, Institute for Systems Research, University of Maryland, College Park, MD). During each trial in an auditory discrimination task, the ferrets were presented with a random sequence of broadband noise-like acoustic stimuli known as temporally orthogonal ripple combinations (TORCs) along with randomized presentations of the target tone. Ferrets were trained to attend to the spectrotemporal features of the presented sounds, and discriminate the tonal target from the background reference stimuli (see [117] for details of the experimental procedures). Due to their broadband noise-like features, the TORCs and the corresponding neural responses admit efficient estimation of the spectrotemporal tuning

of the primary auditory neurons via sparse regression [33, 120].

Fig. 6.2 shows our results on a selected experiment in which a total number of $C = 9$ single-units were detected through spike sorting (5 units in A1 and 4 units in PFC detected from 4 electrodes per region). The selected experiment consists of three main blocks: passive listening pre-task, active task and passive listening post-task, composed of $R = 4, 4$, and 6 repetitions, respectively. Within each repetition, a complete set of 30 randomly permuted TORCs were presented along with a randomized presentation of the target tone at $f = 2.5$ kHz. Fig. 6.2–A shows the activity of all the units during the first repetition of each block, separated by vertical dashed lines. Fig. 6.2–B shows the time-courses of the inferred J-statistics, where each row represents the significant incoming GC links from all the other units. Each unit and its significant outgoing GC links are color-coded uniquely as labeled on the right side of each panel. For brevity, the significant GC links that show a degree of persistence during at least one block of the experiment are plotted. Fig. 6.2–C depicts the representative network maps of the detected GC links among the 9 units during the three main blocks, where each significant GC link from panel 6.2–B is indicated by a directional link. Finally, Fig. 6.2–D exhibits snapshots of the spectrotemporal receptive fields (STRFs) of all the five A1 units, taken at the endpoint of each block. The red arrow marks the tonal target.

Three major task-specific dynamic effects can be inferred from Fig. 6.2: 1) a significant bottom-up GC link from the target-tuned A1 unit during active behavior, 2) a persistent task-relevant top-down GC link, and 3) task-relevant plasticity and rapid tuning changes within A1. First, unit 4 in A1 shows strong frequency

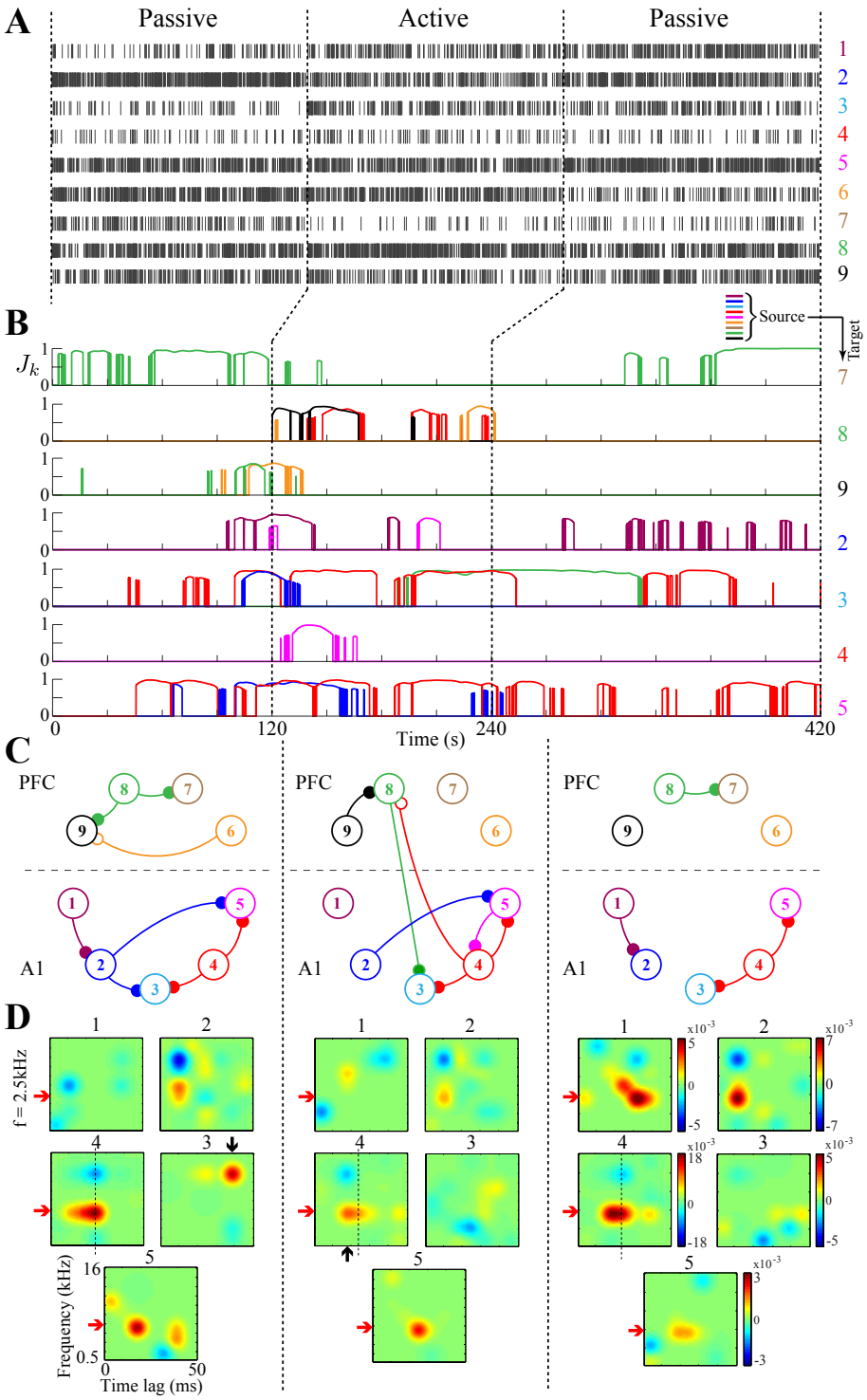

Figure 6.2: Dynamic inference of G-causal influences between single-units in ferret PFC and A1 during auditory task. A) Spike trains corresponding to the first repetition of each block, B) time-course of significant GC changes through J-statistics for selected single-units (FDR controlled at $\alpha = 0.1$), C) detected patterns of network AGC maps during three main blocks of experiment, D) STRF snapshots of A1 units at the endpoints of three blocks of experiment.

selectivity to the target around $f = 2.5$ kHz during the whole experiment (vertical dashed lines, Fig. 6.2–D). Moreover, its STRF dynamics reveal a plastic shift of the target-tuned facilitative regions to shorter latencies following the active attentive behavior (upward arrow, mid. panel). Strikingly, a bottom-up GC link from the very same strongly target-tuned unit to PFC (red link, $4 \mapsto 8$) emerges during the active task (second row, Fig. 6.2–B), temporally preceding any top-down significant GC link.

The second effect appears as a strong top-down GC link (green link, $8 \mapsto 3$) which builds up during the active auditory behavior, and even persists during a few repetitions of the post-active condition (fifth row, Fig. 6.2–B). The onset of this top-down GC link coincides with a dramatic and rapid change in the STRF of the A1 unit 3, which was initially tuned to ~ 8 kHz (downward arrow, left panel, Fig. 6.2–D) but eventually gets suppressed at this non-target frequency (mid. panel, Fig. 6.2–D) by getting G-causally influenced by the PFC unit 8. This effect reveals the relationship between the top-down network dynamics and the changes in the tuning of the A1 units. We examine the dynamics of the parameters of the foregoing bottom-up and top-down links in detail in section 6.1.3, for further clarification. The third effect concerns the emergence and strengthening of frequency selectivity in some of the A1 units (e.g, units 1 and 2, Fig. 6.2–D, right panel) to the target tone, which alludes to a salient synaptic reinforcement effect within A1 during and after the active task.

In addition to these inter-region GC links, multiple instances of GC links within A1 (e.g., $5 \mapsto 4$) or within PFC (e.g., $9 \mapsto 8$) emerge or vanish during

the active block, which accounts for the task-specific network-level changes within the cortical regions that are involved in active listening. A salient instance of this phenomenon can be observed in the dynamics of unit 8, whose GC links within PFC significantly change during the active behavior: as it gets G-causally linked to the lower-level A1 region, its GC links to the other PFC units fade away (rows 1, 3 and 5, Fig. 6.2–B). It is noteworthy that the fluctuating instances of the J-statistics (e.g., Fig. 6.2–B, fourth row, third segment) are due to the FDR control procedure, and there is no evidence to believe that they have a neurophysiological basis. In order to reduce these fluctuations, one can choose a higher FDR rate. If these effects persist at high FDR rates, careful inspection of the cross-history coefficients is needed to assess their possible neurophysiological basis. Further discussion on the dynamics of cross-history coefficients is provided in subsection 6.1.3. Finally, careful inspection of Fig. 6.2 reveals a remarkable property of our proposed AGC inference method. Although the pattern of spiking activities of the units in A1 and PFC did not vary significantly across active-passive blocks of the experiment, the inferred G-causal dynamics reflect significant task-specific network-level changes among the units in the two cortical regions.

In order to validate our results in the absence of ground truth, we assess their reliability using surrogate data obtained by *random shuffling* and *network subsampling* in subsection 6.1.4, and verify the robustness of the inferred task-dependent functional network dynamics against the aforementioned adversarial perturbations. In conclusion, our methodology enabled the extraction of the top-down and bottom-up network-level dynamics that were previously conjectured in [117] to be involved

in active attentive behavior, at the neuronal scale with high temporal resolution. In subsection 6.1.5, we present our analysis of another experiment, which further corroborates our findings.

Numerical Choices of Parameters: We discretized the total duration of $\mathcal{T} = 420$ s using bins of length $\Delta = 1$ ms. The GLM modulation parameter $\boldsymbol{\omega}_k := [\mu_k; \boldsymbol{\omega}_k^{\mathsf{Hist}}; \boldsymbol{\omega}_k^{\mathsf{STRF}}]$ at time k consists of the baseline firing parameter μ_k, the history dependence vector $\boldsymbol{\omega}_k^{\mathsf{Hist}}$, as well as the STRF vector denoted by $\boldsymbol{\omega}_k^{\mathsf{STRF}}$. For the history dependence parameters, we selected $M_H^{\mathsf{cross}} = 3$ cross-history and $M_H^{\mathsf{self}} = 21$ self-history components associated with respective history block lengths of $L_H^{\mathsf{cross}} = 100$ ms and $L_H^{\mathsf{self}} = 1$ s, using non-overlapping windows of $W_H^{\mathsf{cross}} = [20, 30, 50]$ and $W_H^{\mathsf{self}} = [20, 30, 50, \ldots, 50]$ bins, respectively.

For the STRF parameters, we used a vectorized array of size $I \times J$, with $I = 50$ time lag bins, and $J = 50$ frequency bins in logarithmic scale, uniformly spanning time lags in the range of $[0, 50]$ ms, and frequencies in the range of $f \in [500, 16k]$ Hz, respectively. To capture the inherent sparsity of the STRFs in the time-frequency domain, we used a representation $\boldsymbol{\theta}_k^{\mathsf{STRF}} = \mathbf{F}\boldsymbol{\omega}_k^{\mathsf{STRF}}$, where $\mathbf{F}$ is a Gaussian time-frequency dictionary of 49 Gaussian atoms [33], and $\boldsymbol{\omega}_k^{\mathsf{STRF}}$ and $\boldsymbol{\theta}_k^{\mathsf{STRF}}$ denote the sparse representation of the STRF (with 49 parameters) and the vectorized STRF at time k, respectively. We used two-dimensional symmetric Gaussian kernels with a variance of ${d_F}^2/4$ as Gaussian atoms in time-frequency plane, where atoms are distributed on a grid of size 7×7 with a spacing of $d_F = 7$ bins. The vectorized array of the TORC sequence spectrograms with J frequency bins and I time lags (the same

as those used for the STRFs) is considered as the common stimulus sequence $\mathbf{s}_k$ in the GLM model.

We used the ℓ_1–PPF$_1$ filter to estimate the sparse parameter vectors $\widehat{\boldsymbol{\omega}}_k$ associated with the reduced and full GLMs for each neuron in a dynamic fashion. We selected a forgetting factor of $\beta = 0.9998$, a window size of $W = 8$, a step size $\varsigma = \frac{1-\beta}{5W}$, $L = 20$ number of iterations per step, and regularization parameters $\gamma^{(c)}$ tuned for each unit separately via two-fold even-odd cross validation. We chose the scaling factor $\rho = \beta$, and the smoothing factor $\sigma_e^2 = 5 \times 10^{-6}$ for the χ^2 filtering and smoothing algorithm. The FDR is controlled at the rate $\alpha = 0.1$, and the J-statistics computed at mean FDR $\bar{\alpha} = 0.0119$, testing for $|\mathcal{C}| = 9 \times 8 = 72$ possible GC links among the units.

6.1.3 Cross-history Coefficient Dynamics of the Top-down and Bottom-up Links in the Ferret A1-PFC Analysis

We examine the dynamics of the cross-history coefficients associated with the extracted top-down and bottom-up GC links in the ferret A1-PFC interaction during active behavior (See Fig. 6.2). Recall that two of the major findings of this analysis were: 1) emergence of a bottom-up inhibitory link from unit 4 in A1 to 8 in PFC, followed by 2) a top-down excitatory link from unit 8 in PFC to 3 in A1. The latter effect resulted in the disappearance of the frequency selectivity of unit 3 which was originally sharply tuned to $f = 8$ kHz. In addition, unit 4 which affects unit 8 is sharply tuned to the target frequency of $f = 2.5$ kHz.

In order to gain insight into the nature of these influences, we examine the time-course of the estimated underlying cross-history coefficients, and their corresponding confidence intervals. Fig. 6.3–A shows the time-course of the cross-history coefficients $\widehat{\boldsymbol{\omega}}_k^{(8,4)}$ (red traces) and $\widehat{\boldsymbol{\omega}}_k^{(3,8)}$ (green traces) corresponding to the bottom-up and top-down links, respectively. As mentioned earlier in 6.1.2, the cross-history coefficients consist of three components: a low-latency component corresponding to a cross-history window of 20 ms, a mid-latency component corresponding to a cross-history window of 30 ms, and a high-latency component corresponding to a cross-history window of 50 ms, which cover an overall cross-history window of 100 ms. The low-, mid- and high-latency components are distinguished by their line width in Fig. 6.3–A, as indicated in the figure legend.

Consistent with our AGC inference results of Fig. 6.2, these cross-history coefficients undergo major changes shortly after the onset of the active segment, some of which persist throughout a considerable portion of the post-active passive segment. Note that the observed delay of order $\sim$ 40 s in adaptive parameter estimation is consistent with the choice of effective window length $\frac{W}{1-\beta}$ for $W = 8$ and $\beta = 0.9998$.

In order to dissect these dynamics more carefully, we have plotted three snapshots of these coefficients together with their 90% confidence intervals in Fig. 6.3–B. The confidence intervals are obtained based on the de-biasing procedure to account for the bias of the adaptive ℓ_1-regularized ML estimates [33, 87]. Note that, unlike the conventional unbiased Gaussian case, the confidence intervals are not evenly centered around the estimates, which highlights the effect of bias correction. The

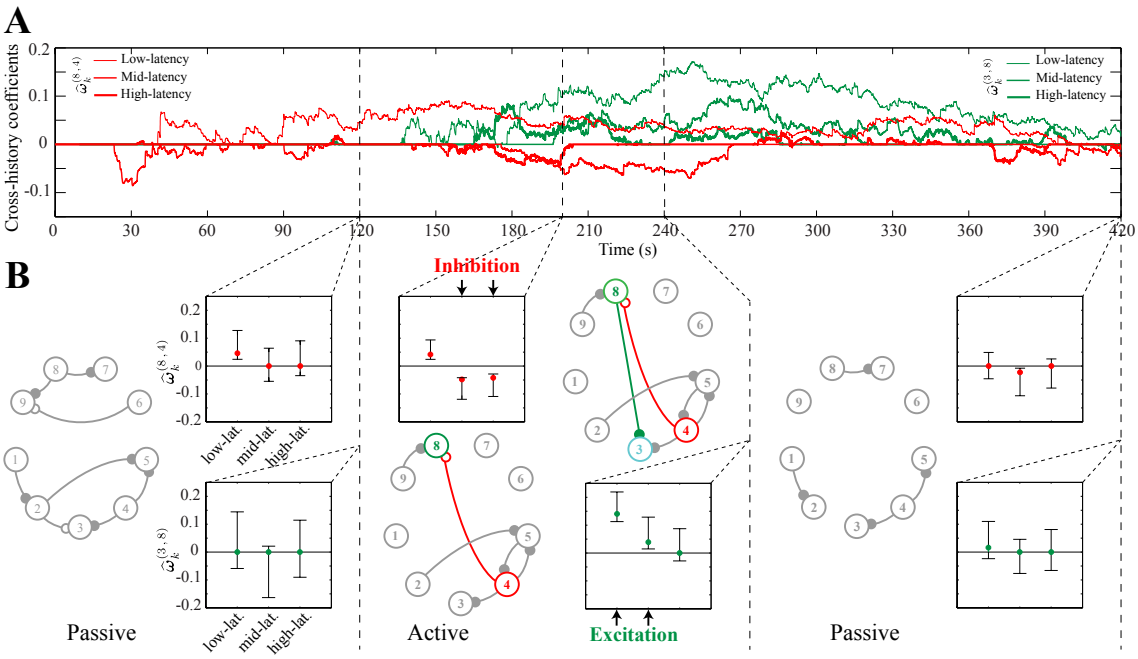

Figure 6.3: Dynamics of the cross-history coefficients for the bottom-up ($4 \mapsto 8$) and the top-down ($8 \mapsto 3$) links estimated from the simultaneous PFC and A1 recordings during the tone detection task (See Fig. 6.2). A) the low-, mid- and high-latency components are distinguished by their line widths, where red and green traces correspond to the bottom-up and top-down link, respectively. B) Bar plots of the cross-history coefficients and their 90% confidence intervals at times indicated by the dashed vertical lines. Each panel also depicts the inferred AGC network. Downward and upward vertical arrows in the middle panel highlight the significant changes in the coefficients.

confidence intervals are not shown in Fig. 6.3–A for graphical clarity. Each panel also shows the inferred AGC network from Fig. 6.2, in which the units not involved in the top-down and bottom-up GC links are grayed out for graphical simplicity.

The left panel shows that during the first passive task, most of the cross-history coefficients are insignificant, which is also reflected in the absence of any cross-region link in the inferred AGC network. The middle panel reveals the emergence of low-latency excitation together with strong mid- and high-latency inhibition from unit 4 to 8 (indicated by downward arrows), hence the overall inhibitory bottom-up

GC link. Similarly, the strong low- and mid-latency excitation from unit 8 to 3 (indicated by upward arrows) results in the top-down excitatory GC link. The latter excitation locks the activity of unit 3 to that of unit 8, and as a result the high frequency responsiveness of unit 3 is suppressed. Finally, the right panel shows that the cross-history coefficients return to the original setting of the pre-active condition.

As mentioned in the discussion following Fig. 6.2, the fluctuations of the J-statistics (e.g., red trace in Fig. 6.2–B, panel 8) are due to the FDR correction procedure, which results in rejecting the null hypotheses only corresponding to links with strong enough coefficients at a given time step. Therefore, the stochastic fluctuations of the cross-history coefficients (e.g., red traces in Fig. 6.3–A) lead to the fluctuations of the deviance statistics around the statistical thresholds set by the FDR control procedure in our multiple hypothesis testing framework.

6.1.4 Validation of the AGC Inference Results from the Ferret A1-PFC Experiment via Surrogate Data Analysis

Given the lack of access to ground truth in the analysis of real data, it is crucial to assess the reliability of our results using carefully devised surrogate data. To this end, we generate two sets of data using random shuffling and network subsampling procedures, and thereby evaluate the consistency of our results.

Analysis of Surrogate Data from Random Shuffling: We first assess the reliability of the inferred AGC networks in the analysis of the ferret A1-PFC

interaction through surrogate data obtained by random shuffling. To this end, we randomly shuffle the activity of single-units across different repetitions regardless of their active or passive nature (14 repetitions in total), such that each repetition of a single unit would be randomly aligned with different repetitions of other single units recorded at different experimental periods. We then infer the AGC network patterns for each shuffled composition of the repetitions. Our goal is to investigate whether our AGC inference procedure detects any significant GC pattern from the shuffled data.

We repeat the random shuffling procedure for $R = 100$ trials, and compute the J-statistics for different links across the whole experiment. We test the reliability of the detected significant links from the original unshuffled data by comparing their J-statistics to those pooled from the randomly shuffled surrogate data. For brevity, we focus on two of the most notable GC links: the top-down ($8 \mapsto 3$) link from PFC to A1 and the bottom-up ($4 \mapsto 8$) link from A1 to PFC.

Fig. 6.4 shows the time course of the J-statistics for these two representative GC links inferred from both the original (red and green traces) and surrogate (gray traces) data. In each panel, the black solid trace represents the mean J-statistics across the $R = 100$ randomly shuffled repetitions, and the colored hulls indicate the corresponding 95% confidence regions. It can be observed that the mean J-statistics from the surrogate data do not surpass the small value of 0.06, while the originally detected J-statistics take large values in the range of $\in (0.7, 1)$. For instance, the value of $J_k^{(8\mapsto 3)}$ at $t = 300$ s is significantly higher than those from the surrogate data (One-tailed Z-test, $p < 0.0001$).

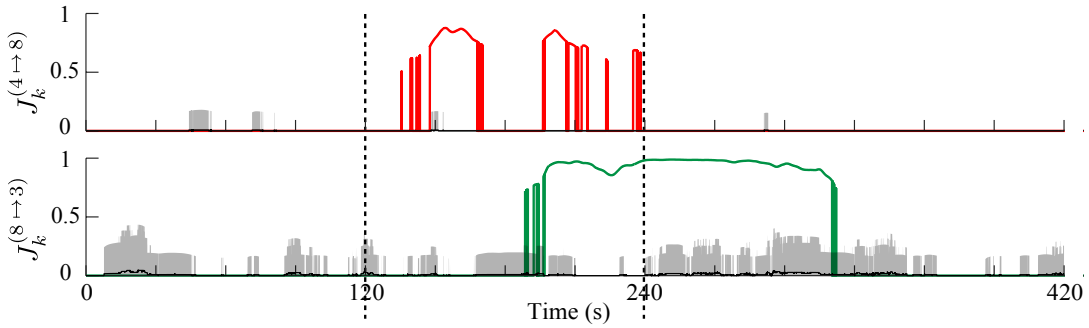

Figure 6.4: Analysis of surrogate data from random shuffling of the repetitions in the ferret A1-PFC experiment. The J-statistics of the $(4 \mapsto 8)$ and $(8 \mapsto 3)$ links inferred from the original data are shown in red and green traces, respectively. The average J-statistics obtained from the randomly shuffled ensemble are shown by black traces, with 95% confidence regions shown by the gray hulls. The J-statistics inferred from the original data show a significant statistical separation from those obtain from the surrogate data.

Moreover, the J-statistics of the surrogate data do not suggest any task-dependent behavior, as opposed to those from the original data. To illustrate this more precisely, suppose that the task-dependence behavior of the link $(8 \mapsto 3)$ were to be preserved in the surrogate data, i.e., this link would persist for blocks comparable in length to that of the original data. Given that this link is active with significant J-statistics for ~ 120s, then it would be expected that the average J-statistics of this link for the surrogate data would be close to $120/420 \approx 0.28$. However, the p-value of this observation with respect to the distribution of the J-statistics over the entire duration of the surrogate data is given by $p = 0.0007$ (One-tailed Z-test).

This analysis verifies that the highly significant AGC links inferred from the data vanish under random shuffling of the repetitions, and are therefore highly specific to the correct temporal ordering of the repetitions in the experiment.

Analysis of Surrogate Data from Network Subsampling: Next, we assess the reliability of the inferred AGC interactions in the analysis of the ferret A1-PFC interaction through surrogate data obtained by network subsampling. To this end, we investigate the robustness of the inferred AGC patterns and their time course against excluding a single or a group of neurons from the observed ensemble. For brevity, we focus on the two bottom-up and top-down AGC links and assess their reliability under three different network subsampling scenarios:

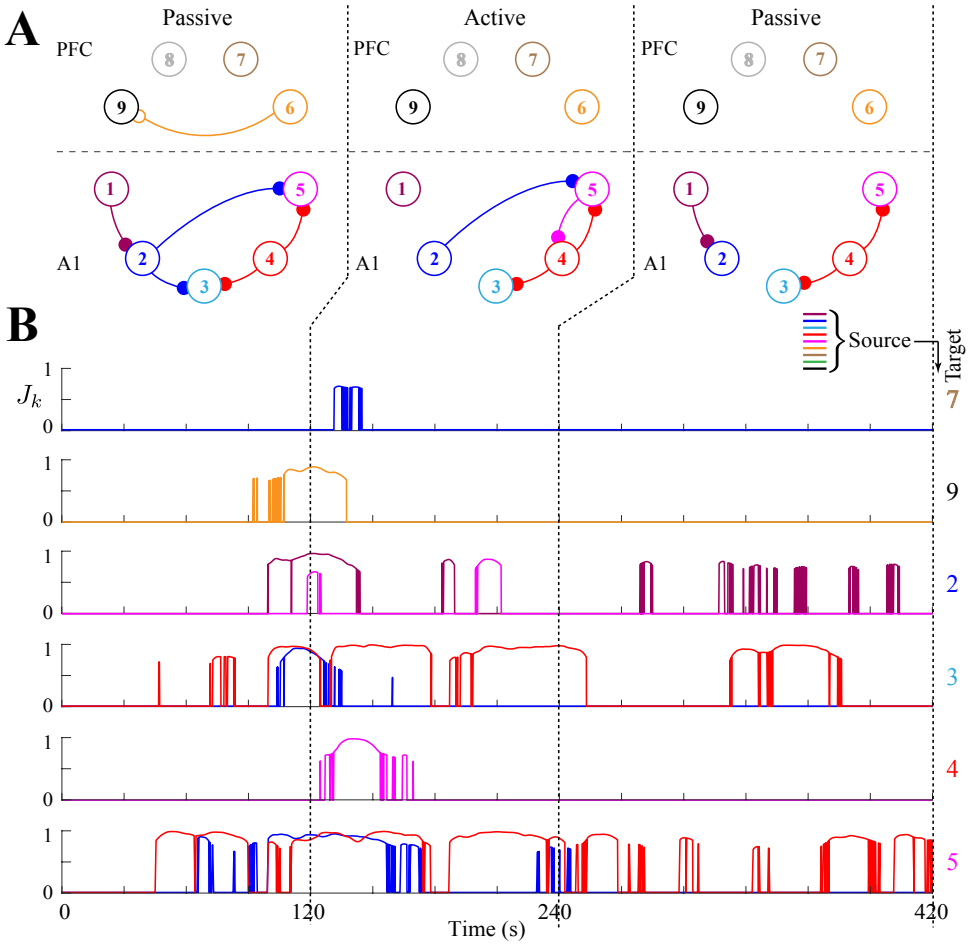

Figure 6.5: Analysis of surrogate data from network subsampling in the ferret A1-PFC experiment, where unit 8 (shown in gray) is excluded from the analysis. As expected, the significant bottom-up and top-down links between A1 and PFC vanish.

Scenario 1: We first exclude the single-unit 8, the only unit in PFC with a significant GC link to A1, from the analysis. We explore the presence of any possible new inter-region GC interactions, and expect that the top-down and bottom-up GC links between PFC and A1 would vanish due to the exclusion of unit 8. Fig. 6.5 shows the resulting AGC network maps and the time courses of the corresponding J-statistics. Indeed, the significant bottom-up and top-down interactions between A1 and PFC vanish, while the rest of the networks within A1 and PFC remain unchanged. The only notable exception is a small transient link from 2 to 7 for $t \in [135, 140]$s.

Scenario 2: Next, we exclude the single-unit 4 in A1, with a bottom-up link to PFC, and test the robustness of our method in terms of the new detected GC links, and expect that no new bottom-up links from A1 to PFC are discovered. Fig. 6.6 shows the resulting AGC network maps and the time courses of the corresponding J-statistics. As expected, the bottom-up link from unit 4 to 8 vanishes, while the rest of the AGC interactions, notably the top-down link from 8 to 3, remain unchanged.

Scenario 3: Finally, we consider a highly undersampled case where we restrict the observable set to the three units $\{3, 4, 8\}$ which are involved in the top-down and bottom-up interactions. We expect that the same bottom-up and top-down patterns between these units are discovered in the absence of all the other 6 neurons which did not exhibit any inter-region GC links. Fig. 6.7 shows the resulting AGC network maps and the time courses of the corresponding J-statistics. Indeed, the expected pattern of GC interaction between these three units is preserved, with the exception of a weak excitatory GC link from 3 to 4 with low statistical significance.

These results show that the inferred AGC maps and the time courses of the

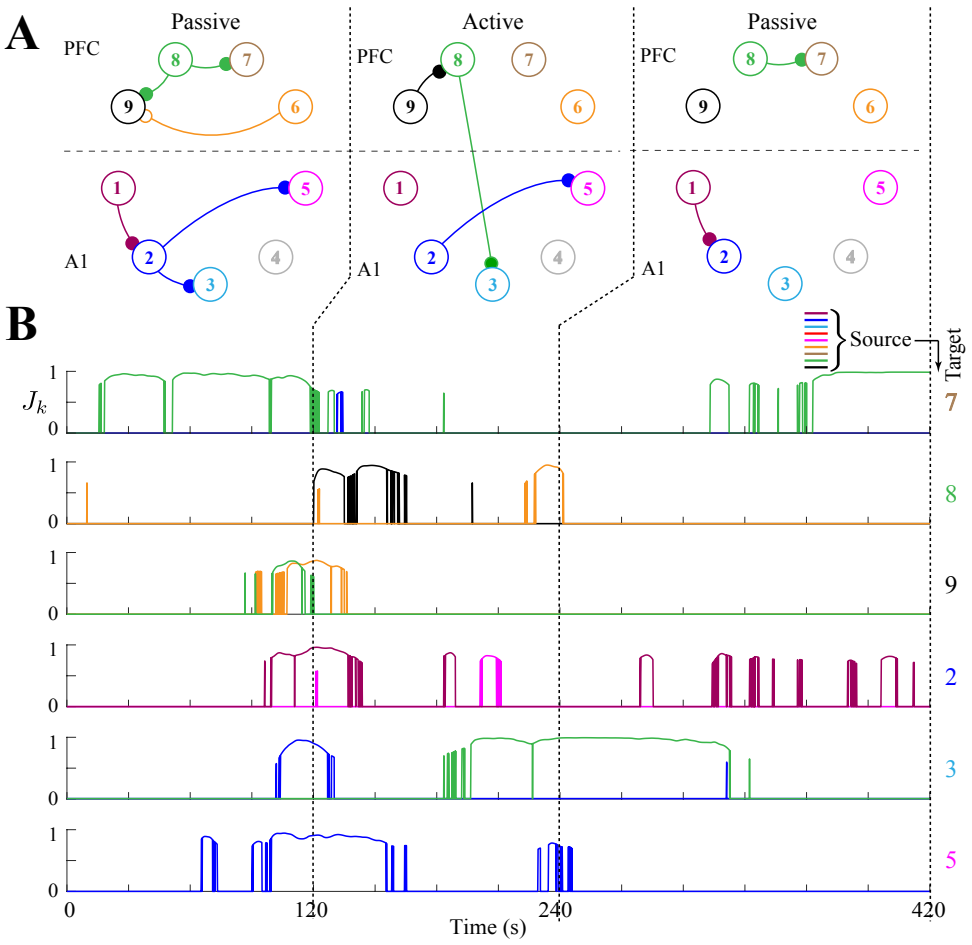

Figure 6.6: Analysis of surrogate data from network subsampling in the ferret A1-PFC experiment, where unit 4 (shown in gray) is excluded from the analysis. As expected, the bottom-up link from A1 to PFC vanishes.

corresponding J-statistics, and notably those pertaining to the bottom-up and top-down network structure, are robust to network subsampling. Hence, they are specific to the interactions between the single-units under study in this experiment, and there is no evidence to believe that they are the byproduct of this particular observable subsampled network of 9 neurons.

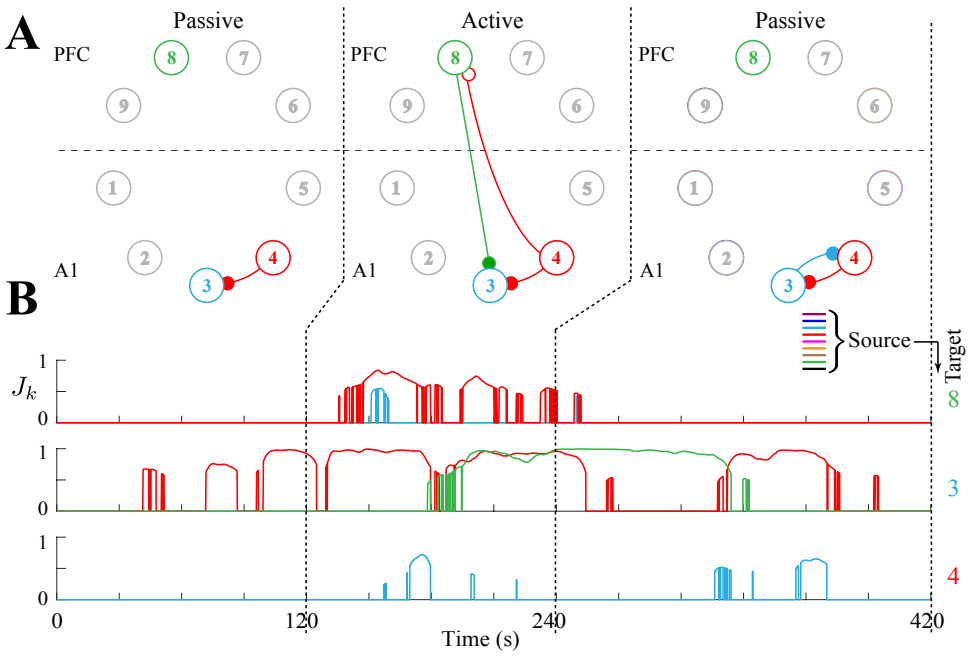

Figure 6.7: Analysis of surrogate data from network subsampling in the ferret A1-PFC experiment, where only units 3, 4 and 8 are included in the analysis (all other units shown in gray). As expected, the significant bottom-up and top-down links and their respective time courses are preserved.

6.1.5 Supporting Example: Ferret A1-PFC Interaction

We present the application of our proposed AGC inference on another instance of spike recordings from the same set of experiments on ferrets as described in 6.1.2, where the animal is performing a pure tone detection task [117].

Fig. 6.8 shows the results of our AGC inference for a selected experiment consisting of four main blocks: pre-active, active, and two post-active conditions, where each block is composed of $R = 5$ repetitions. Within each repetition, a complete set of 30 randomly permuted 1 sec-long TORCs was presented along with a randomized repetition of the target tone at $f = 8$ kHz. Fig. 6.8 shares the same

structural format as Fig. 6.2. A total number of $C = 8$ single-units are detected through spike sorting (4 units in each region), whose spike trains are shown in Fig. 6.8–A. For graphical convenience, we only plotted the spike trains within the last repetition of each block. Fig. 6.8–B shows the time-course of the changes in the J-statistics associated with detected GC links, where each row represents the corresponding significant GC influences from all units to a target unit, which passed the BY FDR control procedure. Each single-unit along with its significant outgoing GC link is color-coded uniquely as shown on the right. Fig. 6.8–C depicts these detected changes in the pattern of G-causal links among the 8 single-units during three main blocks of the experiment. Three STRF snapshots of all the four A1 units at the endpoints of the pre-active, active and post-active blocks are shown in Fig. 6.8–D, along with the target frequency $f = 8$ kHz indicated by a red arrow.

Fig. 6.8 reveals significant task-relevant changes in the pattern of G-causal interactions among the units within or across the PFC and A1 regions. The most striking observation is the identification of 4 bottom-up and 4 top-down GC links during active attentive behavior, which verifies the functional interaction (in the sense of Granger) between the higher-level PFC and the lower-level cortical region involved in active listening. The most significant and persistent bottom-up GC links, e.g. $(1 \mapsto 5)$, belong to the A1 unit 1, whose STRF characteristics show a frequency-selective suppression around the target frequency. As can be observed in Fig. 6.8–D, this A1 unit exhibits significant task-related plasticity [30], as its suppressive response to the target frequency vanishes entirely during the active attentive behavior (downward arrow, mid. panel, Fig. 6.8–D) while it G-causally

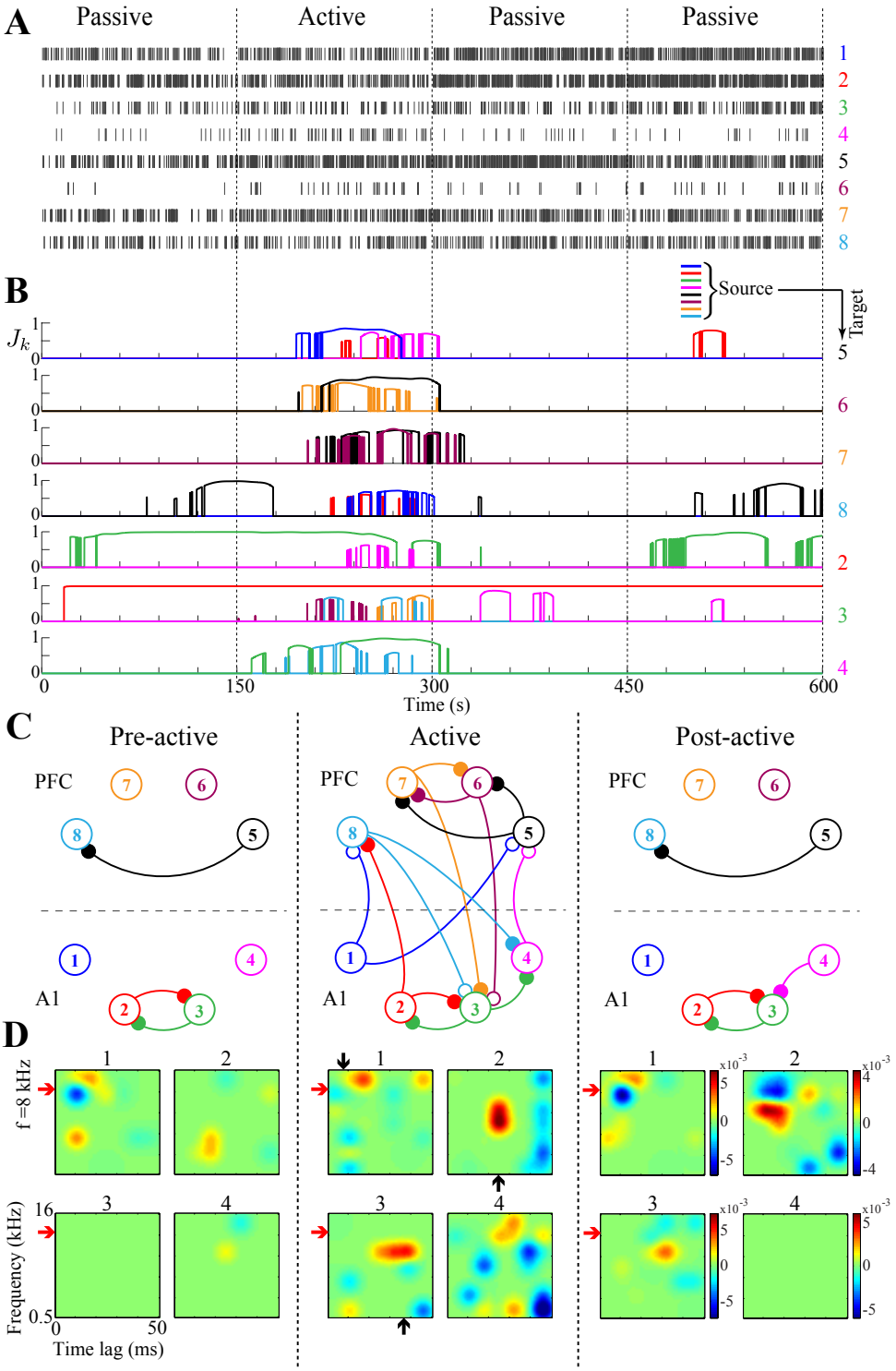

Figure 6.8: Adaptive GC inference from single-unit recordings in the ferret PFC and A1 during auditory tasks. (A) Spike trains corresponding to the last repetition of each block, (B) time-course of GC changes for the significant links through J-statistics, (C) inferred network AGC maps during pre-task, active task, and post-task passive conditions. D) STRF snapshots of A1 units at the endpoints of the three blocks of the experiment. Note the selective reduction in inhibition at 8 kHz (target tone) in A1 unit 1 during behavior (downward arrow, middle panel).

influences the higher level PFC units in an inhibitory fashion. Interestingly, unit 1 retrieves its original pre-active STRF after the active task is over. In addition to the detected inter-region GC links, several instances of task-relevant changes in GC links within A1 (e.g., $3 \mapsto 2$; see upward arrows, mid. panel) or within PFC (e.g., $5 \mapsto 6$) occur during active behavior. In addition, the pattern of GC links within PFC changes dramatically during active attentive behavior as compared to the passive conditions.

Choices of the Parameters: The total duration of $\mathcal{T} = 600$ s is binned by $\Delta = 1$ ms, and segmented by windows of length $W = 25$ bins. We applied the ℓ_1-PPF$_1$ adaptive filter to the spiking data of all single-units, where we selected a forgetting factor of $\beta = 0.9995$, a step size $\varsigma = \frac{1-\beta}{5W}$, $L = 20$ number of iterations per step, and regularization parameters tuned for each unit separately via two-fold even-odd cross validation. We consider the same dynamic GLM model to capture the spiking statistics as in the previous analysis, with the modulation coefficients accounting for both the ensemble spike history and stimuli. For the stimulus modulation, we consider a vectorized STRF array of size $I \times J$, with $I = 50$ time lags and $J = 50$ frequency bins in logarithmic scale represented by a Gaussian time-frequency dictionary [33], capturing the effect of the reference acoustic stimuli spectrogram. As for the ensemble history dependence, we select $M_H^{\mathsf{cross}} = 3$ cross-history and $M_H^{\mathsf{self}} = 21$ self-history components associated with respective non-overlapping spike counting windows of $W_H^{\mathsf{cross}} = [20, 30, 50]$ and $W_H^{\mathsf{self}} = [20, 30, 50, \ldots, 50]$ bins. The FDR is controlled at the rate $\alpha = 0.1$, testing for $|\mathcal{C}| = 56$ possible GC links among the 8 single-units.

6.2 Application to Optical Imaging Data

In this section, we present the results from application of our proposed GC inference methodology introduced earlier in section 4.2 to data from two different continuous-valued modalities of optical imaging: 1) two-photon (2P) calcium imaging data from mouse auditory cortex during auditory tasks, and 2) whole-brain light-sheet imaging data from the larval zebrafish during motor behavior.

6.2.1 Application 1: Probing the Functional Network Organization in the Mouse A1 During Auditory Task Performance

In this subsection, we report the findings obtained from application of our proposed static GC inference method to 2P imaging data in collaboration with the Kanold Laboratory, Systems and Developmental Neuroscience, Department of Biology, University of Maryland, College park. The following results are published in [80], and included in part here.

It is known that neuronal populations in A1 layer 2/3 (L2/3) exhibit heterogeneous frequency tuning in response to pure tones, and many neurons show occasional response or even no response to acoustic stimuli [121,122]. It is conjectured that such diversity in local tunings and response properties is likely due to the complex intra- and inter-laminar connections to L2/3 [123]. Such local heterogeneity and complex connectivity pattern raise the speculation that task-related information processing might differ across subpopulations.

We investigate the functional representation of task performance and behavioral choice at the network-level through GC inference analysis of 2P imaging data from A1 L2/3 of mice performing a tone-detection task. GC inference enables us to estimate and quantify the effective connectivity in the neuronal network by simultaneous analysis of tone-evoked responses in the ensemble of neurons. In addition, 2P imaging allows us to precisely identify the spatial location of each neuron, and thereby provides valuable insights into the spatial structure of the GC networks.

In the following, we perform statistical tests to compare the GC networks across three major task conditions: passive, hit, and miss, in terms of multiple network statistics such as the number, length, and direction of the GC links.

We use data from a large set of experiments ($N_{\mathsf{expt.}} = 80$, with total $N_{\mathsf{cells}} = 4316$ identified neurons) conducted on 10 awake behaving mice for GC analysis (data from the Kanold Lab, Department of Biology, University of Maryland, College park, MD). During task performance, mice were head-fixed and imaged *in vivo* in A1 using the 2P Ca^{2+} imaging technique. Sharing a similar experimental design as those in section 6.1.2, during each trial in a tone-detection task, the mice were trained to attend and follow a simple response-timing behavior by licking a waterspout only during a reward time interval shortly after the target tone onset. On hit trials, mice show licking behavior within the time interval and get rewarded with water flow, whereas they do not lick on miss trials. The number of trials was determined by the persistence of active behavior exhibited by the trained mouse (see [80] for details on the experimental procedures).

Figure 6.9–A shows a sample of inferred GC network maps within an imaged

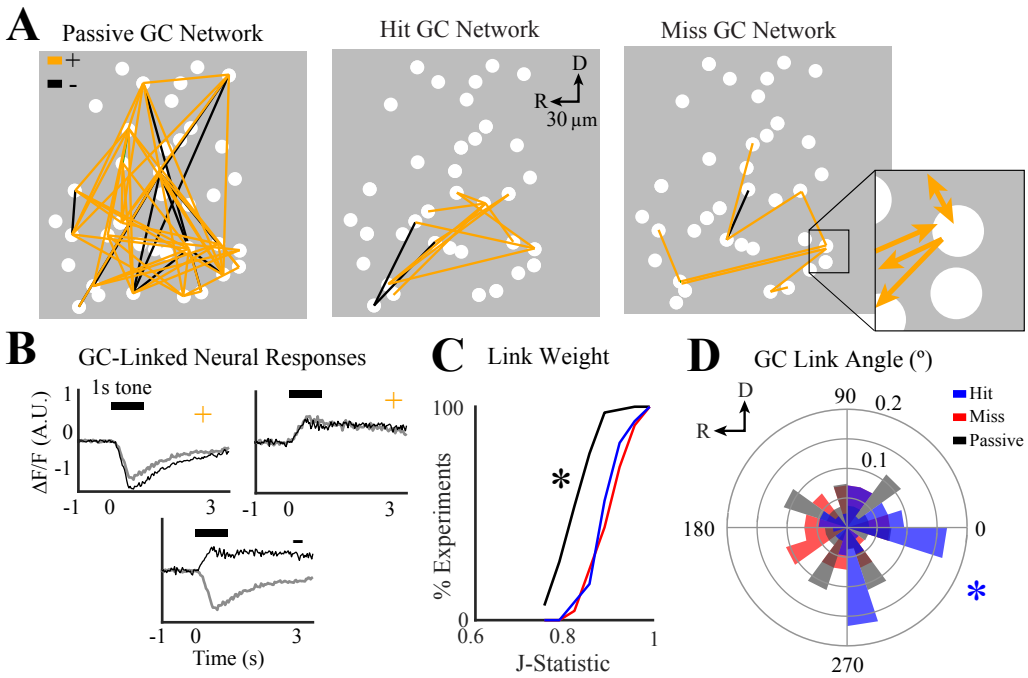

Figure 6.9: Inferred GC network structures in Mouse A1 L2/3 for different task conditions. A) Sample GC networks during passive (left), hit (middle), and miss (right) conditions. Excitatory (+) and suppressive (−) links are shown in orange and black colors, respectively. B) Sample $\Delta F/F$ traces associated with (+) and (−) GC-linked neurons. C) CDF of GC link weights, (*) indicates the difference between the passive and hit trials (bootstrap t-test, $p < 0.001$). D) Polar histogram of GC link average angles, (*) indicates the significant rostro-caudal directionality of links during hit condition (Rayleigh test, $p = 0.03$).

field of view across three different task conditions (passive, hit, and miss) from a selected experiment. The G-causal interactions with effective excitatory (positive) or suppressive (negative) nature are color-coded in orange and black, respectively. Fig. 6.9–B shows three selected $\Delta F/F$ traces associated with the G-causally linked neurons, where the detected effective excitatory or suppressive nature of links reflects the relative sign of the two associated traces.

Three major findings can be obtained from the GC inference analysis and the corresponding statistical tests: 1) Increase in test strengths of GC networks during

auditory task performance, 2) Orientation of GC interactions along the rostro-caudal axis during the hit condition, 3) Pruning of the GC links during auditory task performance.

First, we examine whether the GC link weights change significantly across different task conditions. We use the J-statistics measure obtained from the proposed GC inference as a natural representative reflecting the relative weights of the detected links. Fig. 6.9 represents the CDF of the J-statistics corresponding to the detected links for three major task conditions pooled across different experiments. Our analysis shows that the strengths of GC interactions were larger during the hit condition compared to the passive condition (Fig. 6.9–C; hit: 0.91 ± 0.005 versus passive: 0.84 ± 0.005, bootstrap t-test, $p < 0.001$), but not different for the hit versus miss conditions (hit-miss, 0.002 ± 0.006; bootstrap t-test, $p = 0.52$). This reveals that the auditory task performance boosts the strengths of functional interaction within the networked neurons.

It is known that A1 has a functionally anisotropic property, with the sensitivity to low and high frequencies arranged roughly in a rostro-caudal direction. In addition, this functional asymmetry is reflected in its underlying connectivity pattern [124]. Based on this evidence, we conjecture that the functional interactions within subnetworks involved in hit and miss conditions might differ in terms of spatial organization, in particular, in terms of their angular orientation with respect to the tonotopic axis of A1 (rostro-caudal). We computed the average GC link angle across all detected links for each experiment with the caudal direction taken as 0° reference. As shown in Fig. 6.9–D, the GC links with large test strengths ($J > 0.9$)

show significant directionality along the rostro-caudal axis during the hit condition ($-0.27^\circ \pm 0.18^\circ$, Rayleigh test, $p = 0.033$), and significant directionality was absent in the passive and miss conditions (Rayleigh test, $p > 0.05$). Hence, strong functional connections show preferential orientation along the tonotopic axis during tone detection.

We next performed statistical comparison of the size of GC networks and the organization of subnetworks within. Figs. 6.10–A and B show the CDF of the number of excitatory and suppressive links across different experiments. The average number of excitatory links was largest for passive trials (Fig. 6.10–A; hit, 6.8 ± 0.84; miss, 5.3 ± 0.78; passive, 24.7 ± 2.2; bootstrap t-test, $p < 0.001$ for both hit and miss), while the number of links between hit and miss trials was similar (bootstrap t-test, $p = 0.14$). Similarly, for suppressive links, more GC links were present in the passive condition (Fig. 6.10–B; hit, 1.3 ± 0.26; miss, 0.62 ± 0.15; passive, 5.3 ± 0.51; bootstrap t-test, $p < 0.001$ for both hit and miss), and the average number of links was similar between the hit and miss trials (bootstrap t-test, $p = 0.27$). This result reveals the predominant effect of task performance in decreasing the number of both excitatory and suppressive GC links. Further inspection of the GC networks unveils that groups of neurons form isolated subnetworks within the larger GC network. The CDFs of GC subnetwork size for three task conditions are shown in Fig. 6.10–C. The number of neurons within each isolated GC subnetwork was larger in the hit compared to miss condition (4.6 ± 0.4 versus 3.7 ± 0.41, bootstrap t-test, $p = 0.019$), and was largest in the passive condition (13.6 ± 0.8, bootstrap t-test, $p < 0.001$).

Finally, we performed statistical analysis of the length of the GC links, using

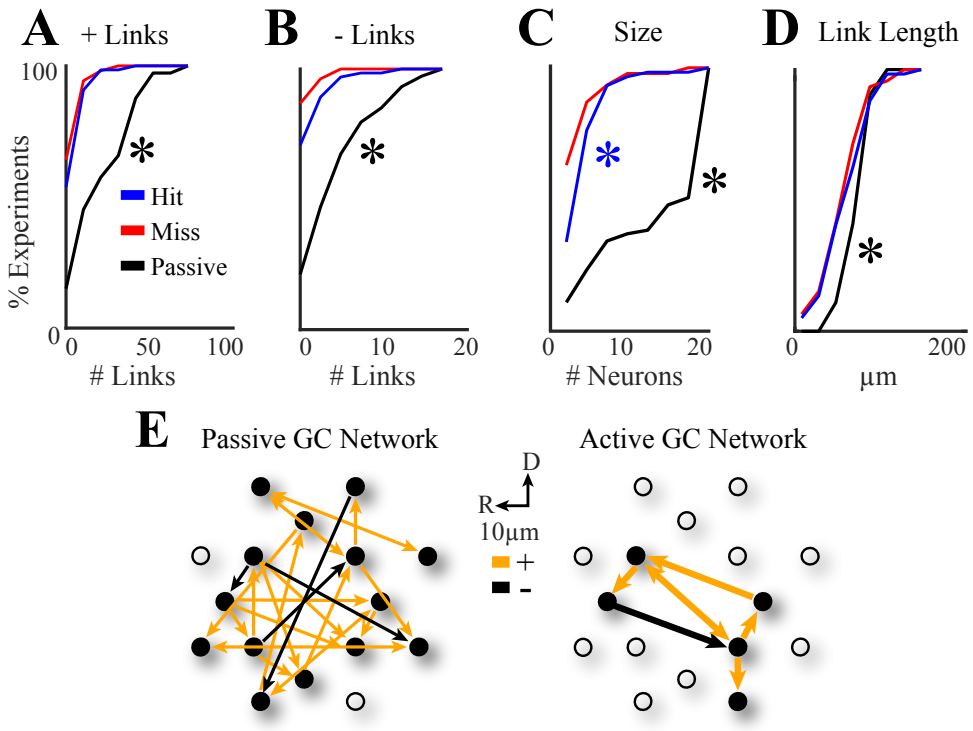

Figure 6.10: Formation of small and localized GC subnetworks during tone detection. A) and B) CDF for the number of (+) and (−) links, respectively. Black (*) indicates the significant difference in passive-hit comparison (bootstrap t-test, $p < 0.001$). C) CDF of GC subnetwork size, blue (*) shows significant difference in hit-miss comparison (bootstrap t-test, $p = 0.019$). D) CDF for GC link lengths. E) Illustration of passive and active (i.e., hit) GC networks.

the spatial information available thanks to the 2P imaging technique. Fig. 6.10–D shows the CDF of the GC link lengths for different task conditions. We found that not only did the number of links decrease during tone detection, but so did the length of the links (hit, $66.9 \pm 3.6mm$; miss, $62.8 \pm 4.03mm$; passive, $79.4 \pm 1.8mm$; bootstrap t-test: hit versus miss conditions, $p = 0.58$; hit or miss versus passive conditions, $p < 0.001$), indicating that during the hit condition, nearby cells were more likely to be GC linked. This finding reveals that tone detection reduces the area occupied by active neural networks in A1 L2/3.

In conclusion, our GC network inference analysis indicates that auditory tar-

get recognition correlates with the transient formation of small and localized sub-networks, having both effective excitatory and suppressive causal interactions that orient preferentially across the rostro-caudal tonotopic axis of A1 during tone detection.

6.2.2 Application 2: Extracting Large-scale Functional Network Maps of Larval Zebrafish from Whole-brain Imaging Data

It is known that the brain function arises from collective interactions among large-scale dynamic networks of neuronal populations spanning the entire brain. Analysis of these large-scale functional networks is only possible if the activity of large populations of neurons across the entire brain could be recorded simultaneously.

Although the number of simultaneously recorded neurons is growing for various modalities of electrophysiology [125] and optical imaging [126,127], the conventional neural imaging techniques and data acquisition methods could only acquire data from a small fraction of all the neurons in the brain. In addition, these methods mostly suffer from common physical constraints such as the trade-off between the imaging resolution and the effective area of the field of view. Hence, the functional networks among disparate population of neurons in different brain regions would be untraceable using these conventional recording techniques.

Recently, a novel imaging technique using light-sheet microscopy [128] is developed to record the activity of neurons in the larval zebrafish brain in vivo through

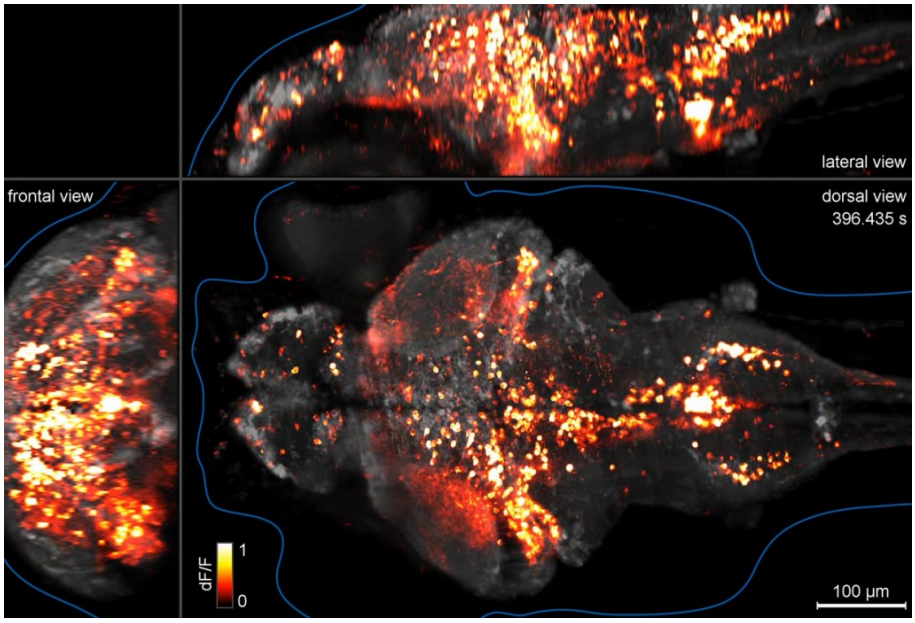

Figure 6.11: Whole-brain imaging of neuronal activity at cellular resolution. (credit: M. Ahrens, *Nature Methods*, 2013. [128])

genetically encoded calcium indicators. Unlike the conventional methods, it is reported that this new imaging technique is able to image the activity of almost all neurons (more than $> 80,000$) in the larval zebrafish brain at the single-cell resolution, covering the entire brain area. This imaging technique provides us with the opportunity, for the first time, to capture the functional organization of large-scale networks spanning the entire brain.

In what follows, we first briefly describe the experimental setup and two locomotion behavioral paradigms. Next, we present the results from the GC analysis of the light-sheet imaging data from larval zebrafish's entire brain during fictive motor behavior (data from the Ahrens Lab, Janelia Research Campus, Howard Hughes Medical Institute, Ashburn, VA).

Motivated by the findings from our GC inference analysis, we then perform

a comprehensive spectral analysis of the whole-brain imaging data with the goal of identifying neural oscillations and rhythmic activity across the brain.

Experimental Paradigm: Fictive Swimming Behavior Setup. The light-sheet imaging data is recorded from a larval zebrafish in fictive conditions. In the fictive behavior setup, the larval zebrafish is paralyzed and embedded in a drop of low melting point agarose, with a freed tail for electrophysiological recordings. Two electrodes record multi-unit extracellular signals from clusters of motor neuron axons, and provide a readout of intended locomotion. Fictive swim bouts are processed separately for the two left and right channels. For more details on the experimental procedure see [126, 129]. In the closed-loop experimental paradigm, a visual stimuli is presented to the zebrafish with fictive motion in the environment, time-locked to the fictive swim bouts, as a visual feedback to mimic the visual effect of swimming. In the open-loop paradigm, a visual stimuli with forward grating is shown to the zebrafish with no visual feedback from the fictive motion. The swimming statistics of the fictively behaving zebrafish change significantly in the open-loop condition, as the fish abruptly ceases to engage in behavior for long periods of time.

In what follows, we infer the functional networks underlying this motor behavior and investigate the network dynamics during the transition from the closed-loop to open-loop condition.

Whole-brain Functional Network Inference via GC Analysis: Authors in [128] performed correlational analysis across the entire brain to identify neuronal populations with correlated activity patterns, which led to the detection of two

functional networks spanning large areas of the hindbrain that may be involved in locomotion.

Despite its well-known merits, it is known that correlational analysis provides limited insights into the directional representation of functional network structure, due to the intrinsic deficiencies in identification of directionality, low sensitivity to inhibitory interactions, and susceptibility to the indirect interactions and confounding effects.

To address these shortcomings, we perform Granger causality analysis on the light-sheet imaging data from the entire brain of the larval zebrafish during fictive motion behavior to elucidate the functional circuitry of large-scale neuronal networks involved in locomotion. Before proceeding with the GC inference, we first need to handle the redundancy of highly correlated activity by reducing the dimensionality of the problem. To this end, we take an unsupervised learning approach to group the neural components within a dorsal projection of the brain based on their observed activity via the K-means clustering technique. For GC analysis, we select 8 neural clusters across different brain regions (forebrain, midbrain, and hindbrain) from the total number of 70 clusters, whose activity shows distinct statistics and significant variability. It can be observed from the clustering results (Fig. 6.12–B) that some neural clusters are localized within a brain region (e.g. cluster 5), while others appear as disparate populations scattered across brain (e.g. cluster 3). The symmetry of neural clusters and the coupling among distant populations are consistent with the known anatomy of the zebrafish brain.

Fig. 6.12 shows the GC inference results from light-sheet imaging of a selected

dorsal single-plane of the larval zebrafish brain during the fictive locomotion experiment. The selected experiment consists of a 100 s closed-loop segment followed by a $\approx$ 10 min open-loop fictive motion condition. Figs. 6.12–A and B represent the detected GC maps among the $N_{\mathsf{clust.}} = 8$ neural clusters corresponding to four 100 s segments of the experiment, respectively in the matrix form and as a network map overlaid on the selected dorsal plane. We selected the open-loop segments to be of equal duration as the closed-loop segment for consistency of the GC inference results. Each neural cluster is color-coded with a unique color (Fig. 6.12–A, left-most panel) and its location within the dorsal plane is indicated in the network map (Fig 6.12–B). The effective excitatory and inhibitory GC links are shown by red and blue colors, respectively. The imaged $\Delta F/F$ fluorescence traces corresponding to 3 selected neural clusters within a 40 s time window are shown in Fig. 6.12–C, along with the recorded electrophysiology signal from the tail (black trace) reflecting the fictive swim activity.

A careful inspection of Fig. 6.12 reveals three major effects: 1) significant inhibitory and excitatory GC interactions in the hindbrain; 2) GC link from the caudal hindbrain to optic tectum during the closed-loop setting; and 3) significant changes in the GC pattern following the onset of the open-loop condition. First, the neural clusters in the hindbrain (clusters $1, 2$ and 3) show persistent GC interactions during the locomotion behavior, representing the active neural regions engaged in motor processing. In particular, a strong inhibitory interaction is identified between the symmetric neural clusters 1 (purple) and 2 (dark green) located in the anterior hindbrain region, referred to as the "waist" region, and the caudal hindbrain, re-

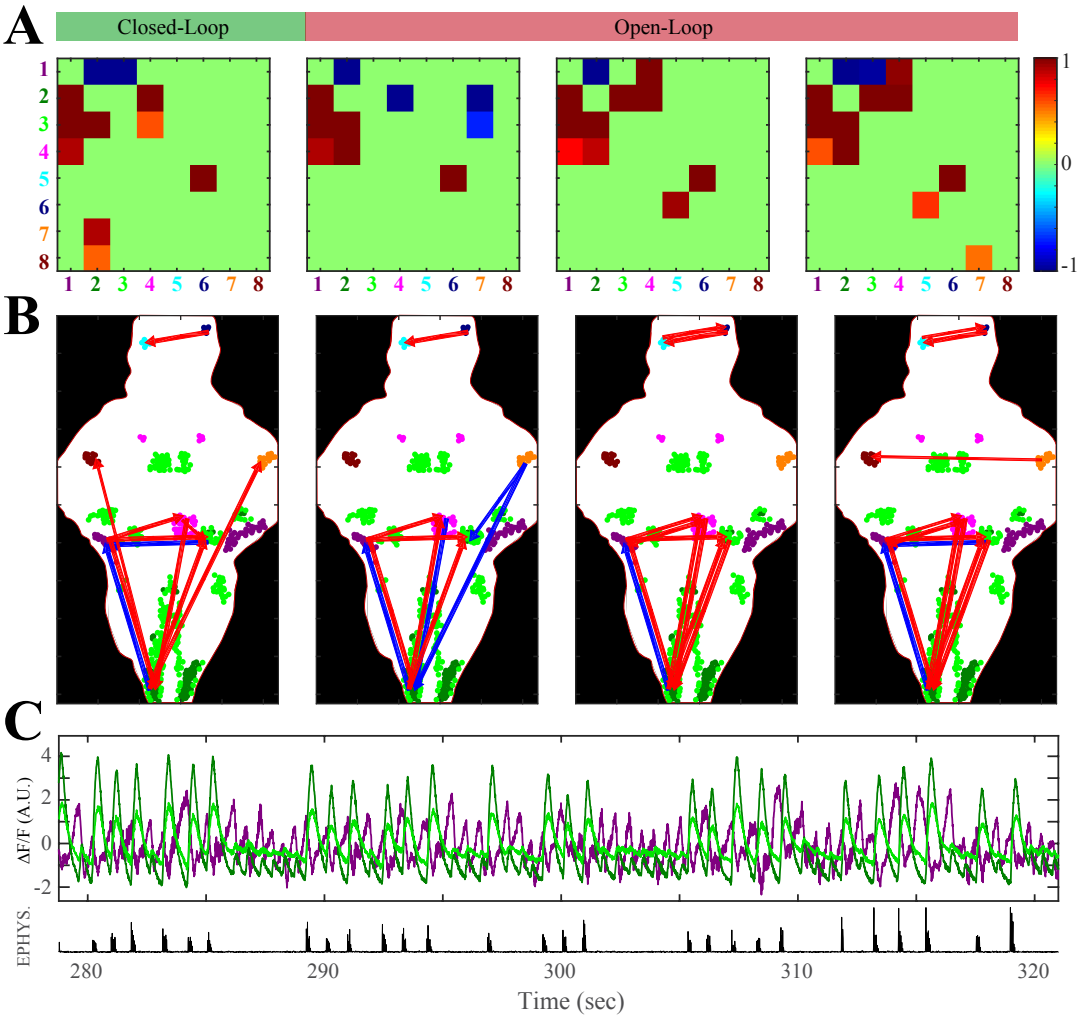

Figure 6.12: GC Inference of large-scale functional networks in the larval zebrafish brain during fictive locomotion behavior. A) Four detected GC maps among the 8 selected neural clusters in matrix form corresponding to a 100 s closed-loop segment, followed by three open-loop segments of equal duration. Each neural cluster is color-coded with a unique color. B) network maps overlaid on the selected dorsal plane. C) Imaged $\Delta F/F$ activity traces associated with three selected neural clusters within a 40 s window reflecting significant GC interactions.

spectively. This inhibitory interaction is reflected in Fig. 6.12–C in the form of an anti-correlated pair of activity traces.

The second effect emerges only during the closed-loop condition, in the form of a GC link from the caudal hindbrain (cluster 2) to the optic tectum (clusters 7

and 8). This effect reflects the visual feedback from the motor commands in caudal hindbrain to the optic tectum region during the closed-loop condition, as the neural activity in the caudal hindbrain is strongly time-locked to the swim bouts. The third effect appears shortly after the onset of the open-loop condition (Rows A and B, second panel), where the GC links among several neural clusters change significantly, such as the inhibitory link from the tectum (cluster 7) to hindbrain (clusters 2 and 3). This reflects the change in functional circuitry during the transition from the closed-loop to open-loop behavioral conditions. In the following, we focus on the significant inhibitory GC interaction identified in the hindbrain and the corresponding anti-correlated activity of the two neural clusters.

Spectral Analysis of Whole-brain Imaging Data in Pursuit of Neural Oscillations. It is generally known that the spatiotemporal patterns of neural activity reflect the functional structure of the underlying neuronal circuits. More specifically, the synchronized patterns of neural activity often arise from the functional interactions among neurons, and the resulting rhythmic activity may synchronize across multiple different brain regions to form large-scale brain networks. Hence, the identification of neural oscillations at the neuronal scale across the whole brain can provide us valuable insights into the functional circuitry underlying motor behavior.

Neural oscillations have been observed throughout the brain in a wide range of species and across different modalities [125,130], ranging from neuronal-scale rhythmic spike trains to population-level oscillatory LFP dynamics. Increasing evidence

suggests that synchronized neural activity is associated with a diverse set of potential functional roles such as neural binding [131], motor coordination [132], brain-wide communication [133] and selective attention [134]. For instance, based on the neural binding theory, oscillatory activity patterns are thought to enable the integration of neural information from a diverse set of sensory streams into a single cohesive form. Despite this growing evidence, the precise underlying mechanisms and the functional roles of different types of rhythmic neural activity are still subjects of controversy.

Despite being widely observed in a diverse set of vertebrates, to the best of our knowledge, there is very limited evidence on the presence of neural oscillations in the zebrafish brain [135]. We inspect the presence of oscillatory dynamics in the larval zebrafish brain using whole-brain light-sheet imaging during motor behavior, and investigate the possible role of these neural oscillations in visuo-motor processing. There is ample evidence that motor and sensory functions are modulated by rhythmic activity patterns, and these rhythmic patterns may be involved in the sensorimotor processing [134]. The emergence of rhythmic oscillatory neural activity is often regarded as a fundamental principle for movement generation and motor control across different species, reflecting a dynamical system structure of the motor cortex [125]. Examples include the rhythmic muscle contractions in medicinal leech [130] and the ~ 1 Hz rhythmic cortical activity in the walking monkey which are consistent with the movement rate [125].

Motivated by the identified GC-linked neural clusters with oscillatory dynamics in the hindbrain, we perform a thorough spectral analysis on the light-sheet

imaging data from the larval zebrafish brain during locomotion behavior in search of coordinated neural oscillations across the entire brain. We construct anatomical spectral maps comprised of multiple dorsal single-plane cross-sections spanning the whole-brain, inspecting the presence of neural oscillations of different frequencies as well as quantifying their extent at different regions through the entire brain.

To this end, we integrate techniques from sparse VAR-based regression and multi-taper spectral analysis [136, 137]. The dynamics of the imaged neural responses often rely on different factors: the transient decay of calcium indicators, the functional interactions among neurons, the sensory inputs, and the motor output. In order to capture and dissect the effects of different covariates, e.g. the transient dynamics of the calcium indicators, we employ a modified variant of the sparse VAR modeling framework which we discussed earlier in Section 4.2. In this new variant, rather than the regular AR-based windowing of regressors, we designed a symmetric Slepian windowing scheme for sparse regression. The advantage of this symmetric window structure is the zero-phase property of the resulting regression filters, which will prevent the phase distortion and the possible emergence of artifacts in spectral domain, when computing the response conditioned on the covariates. In other words, this symmetric structure preserves the spectral power of high-frequency components, conditioned on the effective regressors. In addition to symmetry, we replace the rectangular windows with the zero-th order discrete prolate spheroidal sequence (DPSS) or Slepian sequence [138]. This window function has the optimal property of maximizing the energy concentration in the main lobe of frequency response, which leads to minimal distortion in the spectral properties of conditioned responses. A

schematic depiction of this windowing procedure is shown in Fig. 6.13 along with the conventional AR-based rectangular windows. Note that this symmetric window structure restricts the number of windows M_h (the associated number of coefficients) and the number of samples per window W_h to odd numbers. After dissecting the covariate effects via sparse regression, we perform multitaper spectral analysis on the responses, conditioned on the transient calcium indicator and swim activity, to obtain a high-resolution characterization of the spectral profiles of all neural clusters across the brain [136,137]. Characterizing the power spectral densities (PSD) at high spectral resolutions helps us in precise detection and localization of the rhythmic activity patterns across the whole brain.

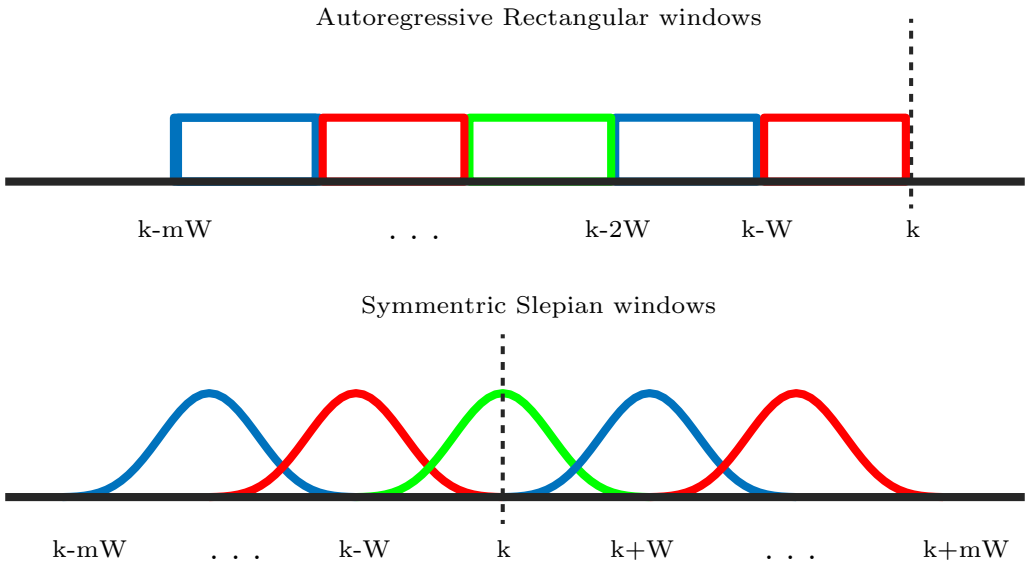

Figure 6.13: Two windowing schemes for sparse regressional analysis of neural responses conditioned on the effective covariates. The AR-based rectangular windows (top panel), and the overlapping symmetric Slepian windows (bottom panel).

Fig. 6.14 depicts the dorsal spectral heat maps of the larval zebrafish brain obtained from spectral analysis of the single-plane light sheet imaging data from the

entire brain. These heat maps represent the degree of rhythmicity associated with the neural response, conditioned on the swim bouts, at each point in the dorsal projection around specific peak frequencies. In order to quantify the measure of rhythmicity, we use the normalized ratio of PSD within a narrow band ($f_c \pm \delta f$) around a peak central frequency f_c. The frequency band width is selected as $\delta f \geq R_{\mathsf{MT}}$, where R_{MT} denotes the multi-taper spectral resolution.

It is noteworthy that due to the limitations imposed by the relatively slow temporal dynamics of calcium indicators, we are only able to detect the low-frequency brain rhythms with a frequency roughly up to $f_c \sim 7$ Hz. These spectral maps can be helpful in the identification of functionally coupled neuronal populations, and can provide insights into the large-scale functional networks, in particular, the brain circuits involved in sensorimotor processing.

Although more dominant in the hindbrain, the identified rhythmic neural clusters are scattered across both the dorsal and lateral projections spanning a large area of the brain. This widespread property along with temporal persistence in specific regions raise the speculation that the detected neural oscillations might be involved in or be a byproduct of the vital rhythmic signals such as cardiovascular or respiratory activity. The proximity of the frequency range of the heart beat of larval zebrafish of a certain age ($[2 \ldots 3]\, Hz$ at $[4 \ldots 6]$ days post fertilization (dpf)) to the one from the neural oscillations identified in specific brain regions further encourage this speculation.

There are several practical ways to examine this conjecture: 1) Inducing arrhythmia using a cardiovascular drug such as terfenadine; 2) Manipulating the heart

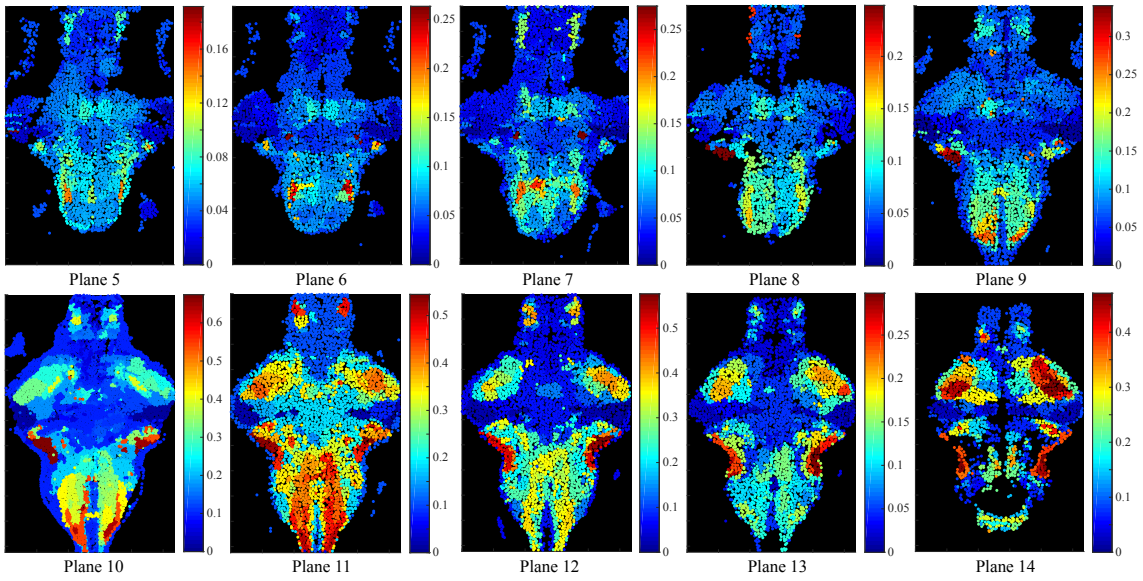

Figure 6.14: Anatomical spectral maps of the larval zebrafish brain during motion behavior in dorsal view, obtained by spectral analysis of the light-sheet imaging data. These maps consist of 10 single-plane cross-sections from the whole-brain stack, and represent the heat maps of the rhythmicity measure at the low frequencies of $f_c = \{2, 3.5, 4.5\}Hz$.

beat rate by changing the temperature or stress condition; and 3) Silencing the heart beat. In order to examine this conjecture, we conducted a simple experiment: In an attempt to manipulate the characteristics of the heart beat and the respiratory function, we added hot water to the petri-dish where the zebrafish was located during the fictive swim condition. We simultaneously recorded the heart beat through the first channel and the fictive swim bouts using the second channel, meanwhile imaging the neural activity of the entire brain.

If associated with the heart beats, it is expected that the neural oscillations would have similar spectral properties as the heart beat rhythms, and any change in the characteristics of the heart beat (including its rate) would be reflected in the detected brain rhythms. Based on our analysis of this experiment, the heart

beat rate increased as a result of the increase in the temperature, while there was no significant change in the characteristics of the neural oscillations. Hence, we rule out the possibility that the observed neural oscillations are mainly tied to vital rhythmic signals.

We next inspect the identified neural clusters showing significant rhythmic activity (Fig. 6.14) in more detail. Four major rhythmic neural clusters are identified across different brain regions: 1) An anatomically symmetric pair of neural clusters in the anterior hindbrain (planes 8 to 14); 2) A symmetric neural cluster in the caudal hindbrain close to the spinal cord (planes 5 to 11); 3) A localized symmetric pair of neuronal groups in the anterior hindbrain (planes 6 and 7); and 4) Symmetric clusters in the forebrain (planes 11 and 12).

The most predominant rhythmic activity appears in the anatomically symmetric pairs of neural clusters in the anterior hindbrain, referred to as the "waist" region. This cluster spans a wide range in the lateral projection, spanning many planes $[8 \dots 14]$ as shown in Fig. 6.14. The maximum rhythmicity occurs around plane 10 with a sharp peak frequency in the range $f_c \in [2 \dots 5]$ Hz. The second rhythmic neural cluster is located in the caudal hindbrain, spanning several ventral planes $[5 \dots 11]$, with the highest rhythmicity being around planes $[10, 11]$. The peak frequency of the rhythms within this cluster was in the frequency range $f_c \in [2 \dots 5]$ Hz for this experiment. The third rhythmic cluster is a relatively small localized symmetric pair in the anterior hindbrain. We speculate that this neuronal group pair might be the *locus coeruleus (LC)*, as its outline and location in the dorsal and lateral projections match those of the LC based on the registered brain atlases. The

LC is known to have widespread connections across the brain, and is involved in many functions including arousal, stress and cognitive control. The last dominant rhythmic cluster appears in the form of symmetric pairs in the forebrain. The degree of regularity and rhythmicity of oscillations across forebrain is often less than those in the hindbrain. The peak rhythmicity appears around plane 11, and is roughly located within the *habenula* region.

In the following, we focus on the aforementioned neural clusters with predominant oscillatory dynamics, and we will further inspect their temporal and spectral characteristics by precise estimation of their PSD and analysis of their phase-locking characteristics. In particular, we will inspect whether these rhythmic activities are synchronized with each other or to the swim behavior with the aim of detecting potential candidates involved in sensorimotor processing.

Fig. 6.15 shows our results from the inspection of the oscillatory dynamics of the two neural clusters with the most predominant rhythmic activity imaged during a fictive open-loop experiment. Fig. 6.15–A represents a selected dorsal projection (plane 10) where both neural clusters exhibit the highest rhythmicity. The waist and the caudal hindbrain clusters are shown in respectively red and blue colors within the dorsal projection of the brain in yellow (forehead in upward direction). Fig. 6.15–B shows the mean response traces, condition on the swim bouts, associated with the two neural regions along with the recorded sequence of swim activity (black traces) within a selected 20 s time window. The PSD estimates corresponding to the conditioned responses of both clusters are shown in Fig. 6.15–C. Fig. 6.15–D represents the phase histogram of the oscillatory activity of the two clusters pooled

across the swim onsets throughout the experiment.

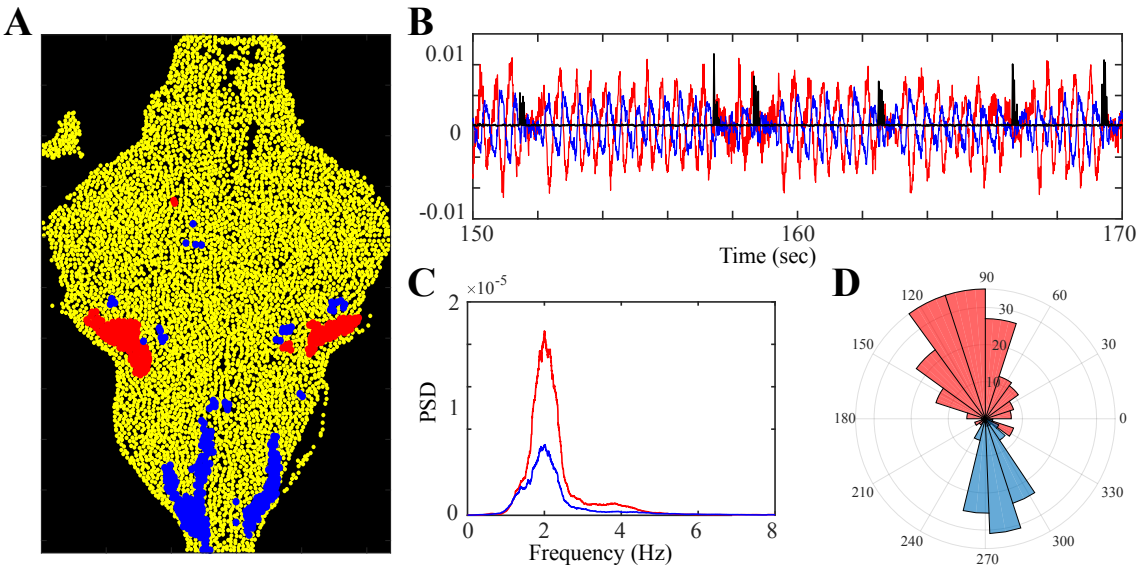

Figure 6.15: Neural clusters with predominant oscillatory dynamics in the hindbrain of larval zebrafish during fictive locomotion behavior. A) a dorsal projection of the waist (blue) and caudal hindbrain (red) clusters within the brain (yellow). B) the neural activity of the two clusters (conditioned on the swim bouts) along with the swim activity (black) within a 20 s time window. C) the estimated PSD for both neural clusters showing oscillatory power around 2 Hz. D) the phase histogram pooled across the swim onsets representing phase-locked antiphase activity of the two neural clusters.

Figs. 6.15–B, C and D reveal the highly oscillatory dynamics of the two neural regions in the hindbrain, which even persist during the absence of the swim bouts. Moreover, the oscillatory activity associated with the waist and caudal hindbrain clusters share the same narrow-band frequency characteristic around the peak $f_c \sim 2$ Hz, and oscillate in an antiphase fashion ($\Delta\phi \sim 180°$). This synchronized activity provides new evidence on the possible connection of these two neural clusters, and reflects a potential inhibitory functional interaction among them, as detected earlier in Fig. 6.12.

Moreover, Fig. 6.15–D further demonstrates a remarkable property of these two oscillations: both neural clusters are phase-locked to the swim activity. The neural activity in the waist often drops during swimming ($\phi \approx 90°$), while the activity in the caudal hindbrain rises at the swim bout onsets ($\phi \approx 270°$). This striking feature reveals the association of these neural oscillations to the motor behavior.

Next, we investigate the neural properties of the regions showing synchronized activity in the hindbrain, with the aim of detecting neuronal populations showing oscillatory activity. We specifically focus on the waist region which shows the highest degree of rhythmicity. This region is mostly composed of densely mixed neuropil structures with axons and dendrites from different brain regions. Apart from the neuropil, the waist region contains sparse neuronal populations according to high resolution whole-brain zebrafish atlases. To this end, we analyze the light-sheet imaging data, in an attempt to distinguish between rhythmic neurons and neuropil structures, and to identify the possible neuronal populations within or close to the waist region with predominant rhythmic activity.

Fig. 6.16 shows the results of our analysis of the light-sheet imaging data during the closed-loop condition, where the field of view is restricted to the waist region. Fig. 6.16–A depicts the raw imaging field (left panel) along with its high-pass filtered version (middle panel), where we are able to locate the neuronal populations (bulb-shaped structures) surrounding the smoother neuropil area. Careful inspection of the imaging field reveals the presence of a patch of neurons within the posterior waist region. The right panel represents the detected neural components (with sig-

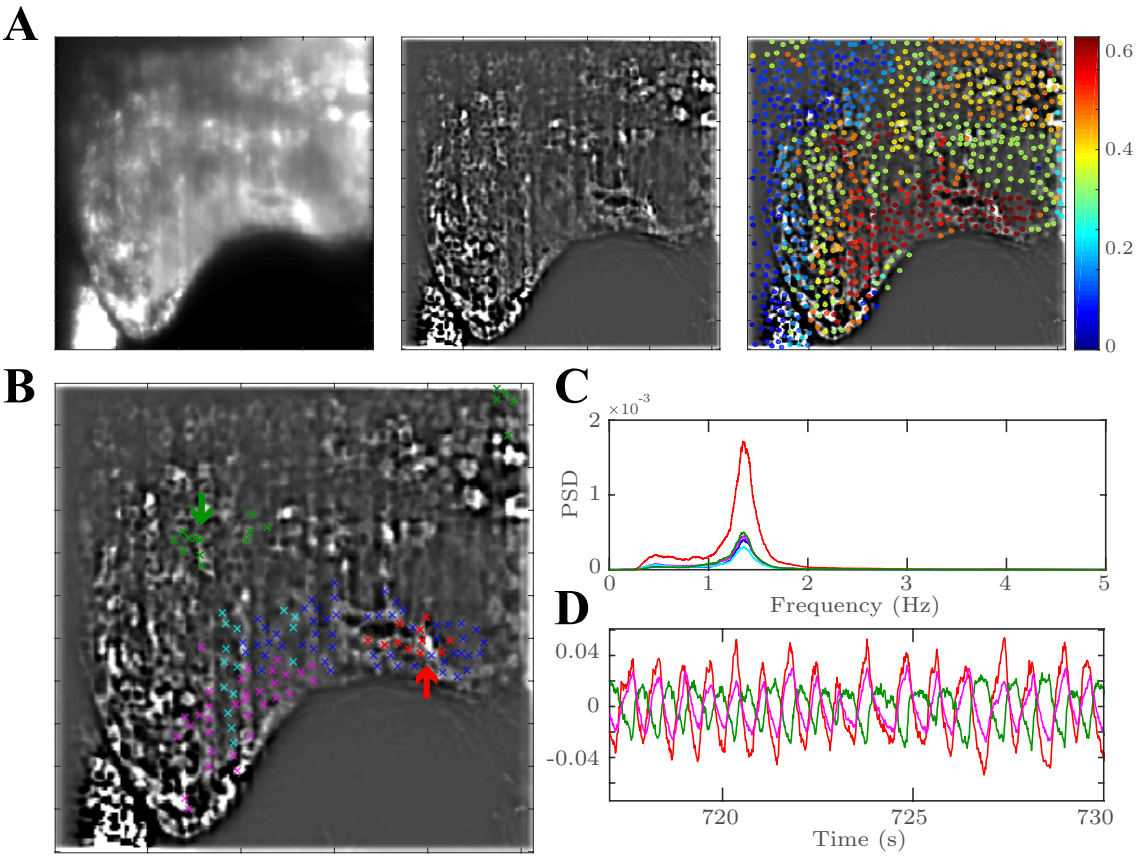

Figure 6.16: Identification of rhythmic neuronal populations within and around the waist neurpil using light-sheet imaging data during fictive locomotion behavior. A) the raw (left) and high-pass filtered imaging field (middle), and the corresponding heat map of rhythmicity overlaid on the FOV (right). B) the top five rhythmic neural clusters (colored crosses) overlaid on FOV. C) the estimated PSD with peak around $f_c \sim 1.35Hz$ and D) the rhythmic activity conditioned on the swim bouts within a 15 s time window associated with these neural clusters.

nificant variability) as filled circles overlaid on the imaging field, and color-coded based on the rhythmicity measure. It can be observed that the neural components within the waist region show the maximum rhythmicity.

Fig. 6.16–B exhibits the top five neural clusters showing the highest rhythmicity around the peak frequency $f_c \sim 1.35$ Hz using colored crosses overlaid on the imaging field. The corresponding PSD estimates are shown in Fig. 6.16–C, and

the activity traces (conditioned on the calcium indicator decay dynamics) associated with the top three rhythmic neural clusters are plotted in Fig. 6.16–D. The first major observation is that the identified neuronal patch in the posterior waist region (red arrow) shows the highest degree of rhythmicity with the most dominant spectral power around the peak frequency (red traces in 6.16–C and D). Such oscillatory activity appears throughout the rest of the waist region in an in-phase but weaker form. Another major observation is the presence of a small neuronal population close to the waist region (green arrow) showing oscillatory activity of the same frequency, but surprisingly in an antiphase fashion (red and green traces in 6.16–D). These close-to-waist neuronal populations appeared partially in Fig. 6.15–A showing synchronized activity in-phase with the caudal hindbrain region.

In summary, we detected isolated neuronal populations within the posterior waist region with dominant oscillatory activity, and small groups of neurons located close to the waist region which are precisely synchronized in an antiphase fashion. This pair of antiphase neural oscillations may have a potential role in the synchronization of motor control. This hypothesis remains to be tested in the future using optogenetic ablation and excitation studies.

Given the neuronal populations detected around the waist region, we next investigate their neurotransmitter identity and morphology using new imaging experiments. In general, neurons can be categorized as excitatory or inhibitory based on their neurotransmitters. Excitatory (inhibitory) neurons release neurotransmitters at their synapses upon arrival of an action potential, triggering positive (negative) changes in the membrane potential of the post-synaptic neurons. Two of

the most common excitatory and inhibitory neurotransmitters are glutamate and gamma aminobutyric acid (GABA), and the neurons generating these transmitters are called Glutamatergic and GABAergic, respectively. We inspect whether the identified rhythmic neurons are Glutamatergic or GABAergic using two different imaging experiments: 1) The two-colored light-sheet imaging from the double transgenic larval zebrafish, and 2) Voltage imaging from the GABA-labeled zebrafish.

Fig. 6.17–A shows the raw imaging field containing the waist and hindbrain regions of the larval zebrafish during locomotion behavior, where the green indicator is expressed in most neurons, and the GABAergic neurons are labeled by red indicators in the double transgenic fish. Fig. 6.17–B represents the high-pass filtered version of imaging field which facilitates the distinction of the neurons from the neuropil and identifying the precise location of neurons. We perform a similar spectral analysis of the activity of detected neural components (conditioned on the calcium indicator decay dynamics). The two neural clusters with high rhythmicity and antiphase synchronized activity are indicated in Fig. 6.17–C with yellow and magenta colored crosses overlaid on their associated regions. It can be observed that the neuronal cluster near the waist (magenta) roughly coincides with the red-labeled cells, and hence we speculate that they may be GABAergic.

To further validate this finding, we imaged the oscillatory neural activity observed around the waist region using voltage imaging. The significance of conducting such imaging experiment is twofold: First, it provides further evidence for the aforementioned conjecture about the GABAergic nature of the near-waist rhythmic neuronal cluster. Second, voltage imaging enables us to image the neural activity

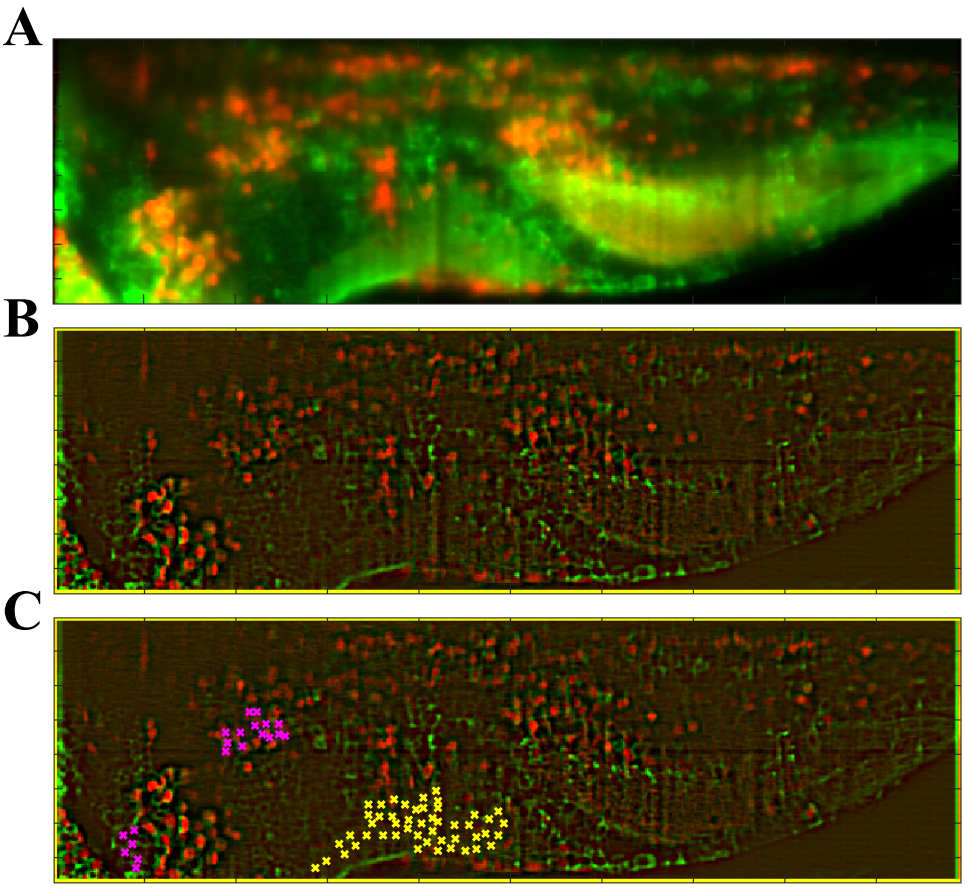

Figure 6.17: Inspection of the neural identity of the rhythmic neuronal cluster near the waist using two-color light-sheet imaging of the double transgenic zebrafish. A) the raw and B) high-pass filtered two-colored imaging field from the waist and hindbrain regions, where green and red indicators are expressed in all neurons and the GABAergic neurons, respectively. C) the two dominant oscillatory clusters: waist neuropil (yellow), and the near-waist neurons (magenta), where the second cluster roughly coincides with the red-labeled GABAergic neurons.

at a higher temporal resolution (up to $f < 500$ Hz) at the expense of lower SNR. At such a high temporal resolution, we are able to detect the neural spike events. This gives us the opportunity to discover the nature of the identified neural oscillation, and determine whether they are the result of smoothed rhythmic spike bursts or smoothed oscillatory LFP activity.

Fig. 6.18–A shows the dorsal projection of the larval zebrafish's hindbrain

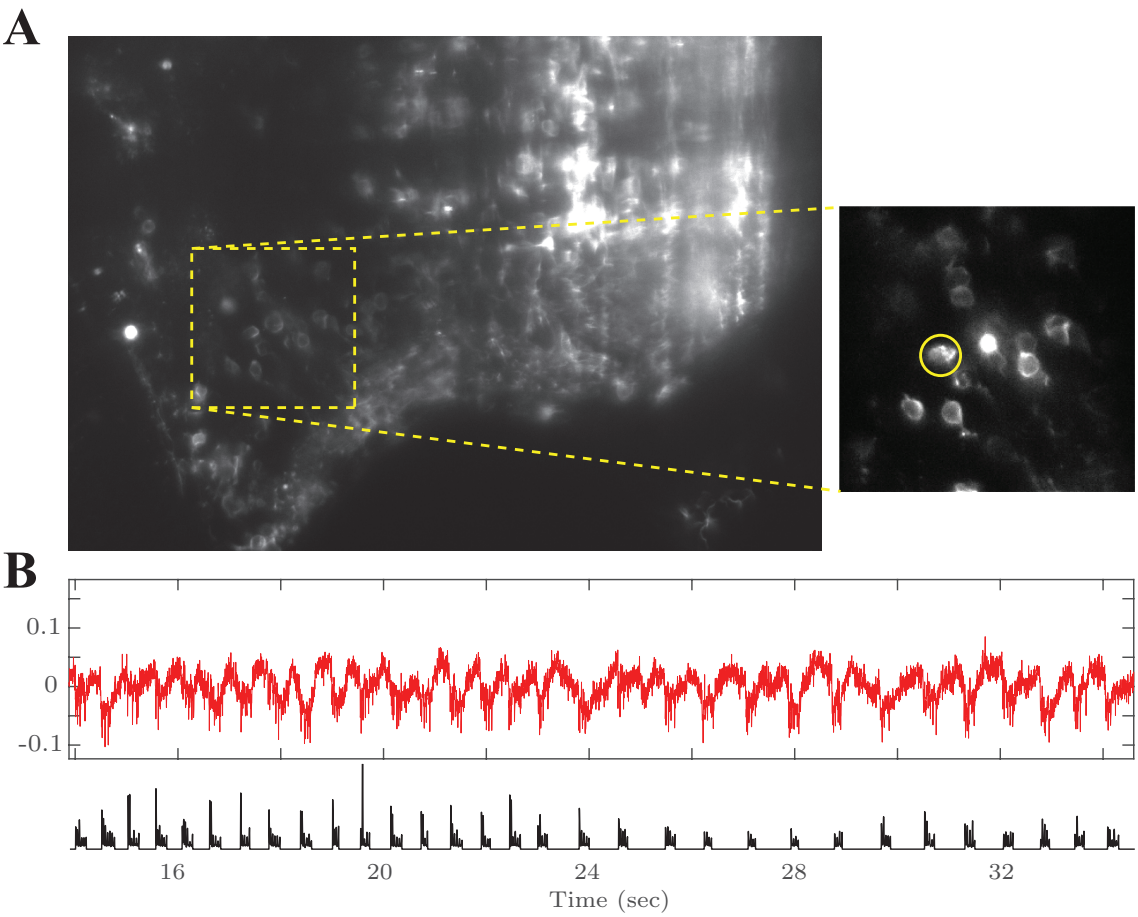

Figure 6.18: Inspection of the neural identity of the rhythmic neuronal cluster near the waist using voltage imaging. A) the FOV (yellow rectangle) within hindbrain whose location is set to image the activity of the target rhythmic neuronal cluster. B) the preprocessed voltage trace of the single neuron highlighted by a yellow circle, showing strong rhythmic spiking activity along with the swim activity within a 20 s time window.

during fictive motor behavior, where only a sparse subset of GABAergic neurons are labeled by voltage indicators. The voltage imaging field is restricted to a small window (yellow rectangle) targeting the rhythmic neuronal cluster identified near the waist. Fig. 6.18–B represents the voltage indicator response of a single neuron (Fig. 6.18–A, yellow circle) in the imaging field (red trace) preprocessed with sparse regression, along with the recorded sequence of swim activity (black trace) within a

20 s time window. The sharp drops in the imaged traces represent spike events.

Two remarkable observations can be made from Fig. 6.18: First, the oscillatory dynamics observed earlier in the target neuronal cluster are originally rhythmic bursts of spiking activity locked to the swim onset. Second, the rhythmic neuronal cluster near the waist is indeed GABAergic. This observation corroborates our speculation on the presence of local oscillators around the waist region in the form of a pair of excitatory and inhibitory clusters. The possible presence of excitatory-inhibitory neuronal clusters along with the strong anti-phase synchronized activity suggests a potential inhibitory connection in the waist region. In summary, our network-level GC and spectral analyses show the promise of these methods in forming hypotheses regarding the functional roles of various brain regions in the visu-motor behavior. These new hypotheses remain to be tested in the future using optogenetic ablation and excitation studies.

Chapter 7: Concluding Remarks and Future Directions

7.1 Summary and Extensions of our Contributions

In the first part of this dissertation, we proposed a sparse adaptive filtering framework for recursive estimation of the time-varying neuronal tuning characteristics from binary spiking data driven by continuous external stimuli. To this end, we integrated several techniques from point process theory, adaptive filtering, compressed sensing, optimization and statistics. We formulated the sparse adaptive estimation problem using an elegantly tailored objective function which enjoys from the trackability features of the RLS-type algorithms, sparsifying features of ℓ_1-minimization, and unlike the rate-based linear models commonly used to analyze spiking data, takes into account the binary statistics of the observations. We constructed a family of filtering algorithms called ℓ_1-regularized point process filters, namely ℓ_1-PPF, consisting of two adaptive filters, with respective linear and quadratic complexity requirements, for recursive solution of the sequence of sparse adaptive filtering problems in an online setting. We analyzed the consistency of the parametric solutions to these problems in a rigorous fashion, revealing novel trade-offs between various model parameters. Moreover, we characterized the statistical confidence regions for our estimates, and devised a recursive procedure to compute

them efficiently. We further extended this family of adaptive filters to accommodate greedy techniques and regularization-based approaches beyond the ℓ_1-norm.

We tested the performance of our algorithms on simulated as well as experimentally recorded spiking data. Our simulation studies revealed that the proposed filters outperform several existing point process filters. Application of our filters to real data from the ferret primary auditory cortex provided a high-resolution characterization of the time-course of spectrotemporal receptive field plasticity, with orders of magnitudes increase in temporal resolution. Although we focused on auditory neurons, we expect a similar favorable performance of our filters when applied to other sensory or motor neurons (e.g., neurons in primary or supplementary motor cortex [139]).

In the second part of this dissertation, we considered the problem of inferring functional network dynamics from neuronal data at high resolutions. Most widely adopted time series analysis techniques for quantifying functional causal relations among the nodes in a network assume static functional structures or otherwise enforce dynamics using sliding windows. While they have proven successful in analyzing stationary Gaussian time-series, when applied to spike recordings from neuronal ensembles undergoing rapid task-dependent dynamics, they hinder a precise statistical characterization of the sparse dynamic neuronal functional networks underlying adaptive behavior.

To address these shortcomings, we developed a dynamic inference paradigm for extracting functional neuronal network dynamics in the sense of Granger, by integrating techniques from adaptive filtering, compressed sensing, point process theory,

and high-dimensional statistics. We proposed a novel measure of time-varying GC, namely AGC, and demonstrated its utility through theoretical analysis, algorithm development, and application to synthetic and real data. Application of our techniques to simultaneous recordings from the ferret auditory and prefrontal cortical areas suggested evidence for the role of rapid top-down and bottom-up functional dynamics across these areas involved in robust attentive behavior. Our analysis of the mouse auditory cortical activity revealed unique features of the functional neuronal network structures underlying both spontaneous activity and auditory task performance at unprecedented spatiotemporal resolutions. Application of our methods to the whole-brain imaging data from larval zebrafish unveiled new insights into the functional organization of the large-scale neuronal networks involved in visuo-motor processing.

7.2 Limitations of our Approach

In closing, it is worth discussing two potential limitations of our proposed AGC inference methodology:

1. Confounding Effects due to Network Subsampling: A common criticism of statistical causality measures, such as the GC, directed information, or transfer entropy, is susceptibility to latent confounding causal effects arising from network subsampling. In practice, these methods are typically applied to a small subnetwork of the circuits involved in neuronal processing. Given that each neuron may receive thousands of synaptic inputs, lack of access to a large number of latent

confounding inputs can affect the validity of the causal inference results obtained by these methods.

We have evaluated the robustness of our method against such confounding effects using comprehensive numerical studies in subsection 5.1.4. These studies involve scenarios with deterministic and stochastic latent common inputs as well as confounding effects due to network subsampling, and suggest that our techniques indeed exhibit a degree of immunity to such confounding effects. We argue that this performance is due to explicit modeling of the dynamics of the Granger causal effects in the GLM framework, invoking the sparsity hypothesis, and employing sharp statistical inference procedures.

2. Biological Interpretation: The functional network characterization provided by our framework must not be readily interpreted as direct or synaptic connections that result in causal effects. Our analysis results in a sparse number of GC interactions between neurons that can appear and vanish over time in a task-specific fashion. While it is possible that these connections reflect synaptic contacts between neurons, as changes in synaptic strengths can be induced rapidly within minutes [140], the observed GC dynamics could also be due to other underlying mechanisms such as desynchronization of inputs, altered shunting or dendritic filtering. Thus, these plasticity effects remain to be tested with ground truth experiments. An alternative and inclusive view is that these links reflect a measure of information transferred from one neuron to another.

The relatively rapid switching of these links, however, must be interpreted with

caution: while some of the rapid fluctuations are due to the usage of the FDR control procedure (as discussed in sections 6.1.2 and 6.1.3 of Chapter 6), sudden emergence or disappearance of a link does not necessarily imply sudden changes in the causal structure or information transfer in the network. A sudden disappearance of a steady link most likely reflects the fact that given the amount of currently available data, there is not enough evidence to maintain the existence of the link at the group level with the desired statistical confidence; similarly, a sudden emergence of a link most likely implies that enough evidence has just been accumulated to justify its presence with statistical confidence. The gradual effects of these interactions are indeed reflected in the dynamics of the non-centrality parameters estimated by our methods.

7.3 Future Directions

The plug-and-play nature of the algorithms used in our framework enables them to be generalized for application to various other domains beyond neuroscience, such as the analysis of social networks or gene regulatory networks. As an example, the GLM models can be generalized to account for m-ary data or accommodate other link functions (such as log or probit-link) and other regularization schemes (e.g., re-weighted or group-sparse regularization), the forgetting factor mechanism for inducing adaptivity can be extended to state-space models governing the coefficient dynamics, and the FDR correction can be replaced by more recent techniques such as knockoff filters [141]. To ease reproducibility and

aid the adoption of our method, we have archived an implementation on GitHub (`https://github.com/Arsha89/AGC_Analysis`).

As demonstrated by the applications of our inference procedures, our framework provides a robust characterization of the dynamic statistical dependencies in the network in the sense of Granger at high temporal resolution. This characterization can be readily used at a phenomenological level to describe the dynamic network-level functional correlates of behavior, as demonstrated by our real data applications. More importantly, as demonstrated by our analysis of the whole-brain data from the larval zebrafish, this characterization can serve as a guideline in forming hypotheses for further testing of the direct causal effects using experimental procedures such as lesion studies, microstimulation, or optogenetics in animal models.

Appendix A: Supplementary Material on Chapter 3

A.1 Proof of Theorem 3.1

The proof is mainly based on the beautiful treatment of Negahban et al. [19]. The major difficulty in our case lies in the high inter-dependence of the covariates, which form a Toeplitz structure due to the setup of adaptive filtering. We address the latter issue by adopting techniques from another remarkable paper by Haupt et al. [85] to deal with the underlying inter-dependence. In the process, we also employ concentration inequalities for dependent random variables due to van de Geer [142].

Before proceeding with the proof, we need to make the following technical assumptions for our analysis:

(1) The stimulus sequence $\{s_t\}_{t=-M+1}^{T}$ consists of independent (but not necessarily identically distributed) random variables with a variance of σ^2 which are uniformly bounded by a constant $B > 0$ in absolute value. Note that with this assumption, two successive covariate vectors, say at times t and $t+1$, given respectively by $\mathbf{x}_t = [1, s_{t-M+2}, s_{t-M+3}, s_{t-M+4}, \cdots, s_t]$ and $\mathbf{x}_{t+1} = [1, s_{t-M+3}, s_{t-M+4}, \cdots, s_t, s_{t+1}]$ are highly *dependent*, as they have $M-3$ random variables in common. Hence, the independence assumption used in study-

ing least squares problem is violated.

(2) We further assume that for all times t, $0 < p_{\min} \leq \lambda_t \Delta \leq p_{\max} < 1$, for some constants $p_{\min}$ and $p_{\max}$, i.e., the probability of spiking does not reach its extremal values of 0 and 1, but can get arbitrarily close. This assumption can be realized due to the boundedness of the covariates and appropriate normalization of the stimulus modulation coefficients, and does not result in any practical loss of generality.

In order to proceed, we adopt the notion of Strong Restricted Convexity (RSC) introduced in [19]. For a twice differentiable log-likelihood with respect to $\boldsymbol{\omega}$, the RSC property of order S implies the existence of a lower quadratic bound on the negative log-likelihood:

$$\mathcal{D}_\ell(\boldsymbol{\Delta}, \boldsymbol{\omega}) := -\ell^\beta(\boldsymbol{\omega} + \boldsymbol{\Delta}) + \ell^\beta(\boldsymbol{\omega}) + \boldsymbol{\Delta}' \nabla \ell^\beta(\boldsymbol{\omega}) \geq \kappa \|\boldsymbol{\Delta}\|_2^2, \tag{A.1}$$

for a positive constant $\kappa > 0$ and all $\boldsymbol{\Delta} \in \mathbb{R}^M$ satisfying:

$$\|\boldsymbol{\Delta}_{\mathcal{S}^c}\|_1 \leq 3\|\boldsymbol{\Delta}_{\mathcal{S}}\|_1 + 4\sigma_S(\boldsymbol{\omega}). \tag{A.2}$$

for any index set $\mathcal{S} \subset \{1, 2, \cdots, M\}$ of cardinality S. The following key lemma establishes the RSC for $\ell^\beta(\boldsymbol{\omega})$:

Lemma A.1 *Let $\{\mathbf{x}_t\}_{t=1}^{KW}$ denote a sequence of covariates and let $\boldsymbol{\omega}$ denote the corresponding logistic parameters. Then, for an arbitrarily chosen positive constant $d > 0$, there exist constants C' and $\kappa > 0$ such that for $M > 10S$, $\beta \geq 1 - \frac{C'}{S^2 \log M}$ and $K \geq \frac{\log 2}{\log\left(\frac{1}{\beta}\right)}$ the negative log-likelihood $-\ell^\beta(\boldsymbol{\omega})$ satisfies the RSC of order S with*

constant $\frac{\kappa}{1-\beta}$ *with probability greater than* $1 - \frac{3}{M^d}$. *The constants* C' *and* κ *are only functions of* d, $p_{\min}$, $p_{\max}$, σ^2, B, W, *and are explicitly given in the proof.*

Proof A.1 *The proof is inspired by the elegant treatment of Negahban et al. [19]. The major difficulty in our setting is the high interdependence of successive covariates due to the shift structure induced by the adaptive setting, whereas in [19], the matrix of covariates is composed of i.i.d. rows. Using the Taylor's theorem,* $\mathcal{D}_\ell(\boldsymbol{\Delta}, \boldsymbol{\omega})$ *can be written as:*

$$\sum_{i=1}^{K}\sum_{j=1}^{W} \beta^{K-i} \frac{\exp\left(\mathbf{x}'_{(i-1)W+j}\boldsymbol{\omega}^\star\right)|\boldsymbol{\Delta}'\mathbf{x}_{(i-1)W+j}|^2}{\left(1+\exp\left(\mathbf{x}'_{(i-1)W+j}\boldsymbol{\omega}^\star\right)\right)^2},$$

with $\boldsymbol{\omega}^\star = \boldsymbol{\omega} + \tau\boldsymbol{\Delta}$ *for some* $\tau \in (0,1)$. *Since by hypothesis* $0 < p_{\min} \leq \lambda_i \Delta \leq p_{\max} < 1$, *we have:*

$$\frac{\exp\left(\mathbf{x}'_{(i-1)W+j}\boldsymbol{\omega}^\star\right)}{\left(1+\exp\left(\mathbf{x}'_{(i-1)W+j}\boldsymbol{\omega}^\star\right)\right)^2} \geq p_{\min}(1-p_{\max}).$$

We can therefore further lower bound $\mathcal{D}_\ell(\boldsymbol{\Delta}, \boldsymbol{\omega})$ *by:*

$$\mathcal{D}_\ell(\boldsymbol{\Delta}, \boldsymbol{\omega}) \geq p_{\min}(1-p_{\max})\sigma^2 N_\beta \left\{\boldsymbol{\Delta}'\mathbf{C}_\beta\boldsymbol{\Delta}\right\},$$

where $N_\beta := W\frac{1-\beta^K}{1-\beta}$, *and*

$$\mathbf{C}_\beta := \frac{1}{\sigma^2 N_\beta}\sum_{i=1}^{K}\sum_{j=1}^{W}\beta^{K-i}\mathbf{x}_{(i-1)W+j}\mathbf{x}'_{(i-1)W+j}. \tag{A.3}$$

Note that the matrix $\mathbf{C}_\beta$ *has highly inter-dependent elements due to the Toeplitz structure in the adaptive design. In order to establish the RSC condition, we show the stronger Restricted Eigenvalue (RE) property, which in turn implies RSC [143]. Let* $\delta \in (0,1)$ *be fixed so that* $\frac{1+\delta}{1-\delta} < \frac{M-S}{9S}$. *To do so, we need to bound the eigenvalues*

of $(\mathbf{C}_\beta)_\mathcal{S}$*, the restriction of* $\mathbf{C}_\beta$ *to a subset of columns and rows corresponding to indices in* $\mathcal{S} \subset \{1, 2, \cdots, M\}$ *with* $|\mathcal{S}| = rS$*, for some integer* $r > 1 + \frac{9(1+\delta)}{1-\delta}$ *such that* $rS \leq M$*. Note that the hypothesis of* $M > 10S$ *makes it possible to simultaneously choose* δ *and* r *satisfying the aforementioned inequalities.*

Without loss of generality, we replace the first entry of the covariate vectors $\mathbf{x}_t$ *by* σ *instead of* 1*, for presentational simplicity of the following treatment. For* $m, m' \neq 1$*, we have:*

$$(\mathbf{C}_\beta)_{m,m'} = \frac{1}{\sigma^2 N_\beta} \sum_{i=1}^{K} \sum_{j=0}^{W-1} \beta^{K-i} s_{(i-1)W+j+m-M} \times s_{(i-1)W+j+m'-M}.$$

For $m = m' = 1$*,*

$$(\mathbf{C}_\beta)_{1,1} = \frac{1}{\sigma^2 N_\beta} \sum_{i=1}^{K} W \beta^{K-i} \sigma^2 = \frac{1}{N_\beta} W \frac{1-\beta^K}{1-\beta} = 1,$$

and for $m \neq 1$*,*

$$(\mathbf{C}_\beta)_{m,1} = (\mathbf{C}_\beta)_{1,m} = \frac{1}{\sigma N_\beta} \sum_{i=1}^{K} \sum_{j=0}^{W-1} \beta^{K-i} s_{(i-1)W+j+m-M}.$$

We also have $\mathbb{E}\{(\mathbf{C}_\beta)_{m,m'}\} = \delta_{mm'}$*. Using Hoeffding's inequality [144] we get:*

$$\begin{aligned}
\mathbb{P}\left(|(\mathbf{C}_\beta)_{m,m} - 1| > t\right) &\leq 2\exp\left(-\frac{2N_\beta^2 t^2 \sigma^4}{B^4 \sum_{i=1}^{K} W \beta^{2(K-i)}}\right) \\
&= 2\exp\left(-\frac{2N_\beta^2 t^2 \sigma^4}{B^4 N_{\beta^2}}\right) \\
&\leq 2\exp\left(-\frac{2N_\beta t^2 \sigma^4}{B^4}\right), \qquad \text{(A.4)}
\end{aligned}$$

since $N_{\beta^2} = N_\beta \frac{1+\beta^K}{1+\beta} \leq N_\beta$*, for* $\beta \in [0, 1]$*. Similarly,*

$$\mathbb{P}\left(|(\mathbf{C}_\beta)_{1,m}| > t\right) \leq 2\exp\left(-\frac{2N_\beta t^2 \sigma^2}{B^2}\right), \qquad \text{(A.5)}$$

Next, we adopt the partitioning technique of Theorem 4 in [85]: for $m \neq m'$, each term in the summation defining $(\mathbf{C}_\beta)_{m,m'}$ is at most dependent on two other terms in the summation. Hence, it is possible to decompose $(\mathbf{C}_\beta)_{m,m'} = (\mathbf{C}_\beta)^{(1)}_{m,m'} + (\mathbf{C}_\beta)^{(2)}_{m,m'}$, where

$$(\mathbf{C}_\beta)^{(i)}_{m,m'} = \frac{1}{\sigma^2 N_\beta} \sum_{\ell=1}^{T_i} \beta^{K-\lfloor \frac{\pi_i(\ell)}{W} \rfloor} s_{\pi_i(\ell)+m-M} s_{\pi_i(\ell)+m'-M}, \quad i = 1, 2, \tag{A.6}$$

in which $\pi_1(\cdot)$ and $\pi_2(\cdot)$ are permutation operators over $\{1, 2, \cdots, KW\}$ and $T_1, T_2 \leq \frac{KW+1}{2}$, such that the summands in each of $(\mathbf{C}_\beta)^{(1)}_{m,m'}$ and $(\mathbf{C}_\beta)^{(2)}_{m,m'}$ are independent. Thus,

$$\begin{aligned} \mathbb{P}\left(|(\mathbf{C}_\beta)_{m,m'}| > t\right) &\leq \sum_{i=1}^{2} \mathbb{P}\left(\left|(\mathbf{C}_\beta)^{(i)}_{m,m'}\right| > \frac{t}{2}\right) \\ &\leq 4 \exp\left(-\frac{N_\beta t^2 \sigma^4}{8B^4}\right), \end{aligned} \tag{A.7}$$

where the first inequality follows from the union bound and the second inequality follows from the Hoeffding's inequality and the fact that $\sum_{\ell=1}^{T_i} \beta^{2K-2\lfloor \frac{\pi_i(\ell)}{W} \rfloor} \leq N_\beta$, for $i = 1, 2$.

Let $B_0 := \max\{B^2, \frac{B^4}{\sigma^2}\}$. Now, the inequalities of Eqs. A.5 and A.7 and the union bound yield:

$$\mathbb{P}\left(\bigcup_{\substack{m,m'=1 \\ m<m'}}^{M} \left\{ \left|(\mathbf{C}_\beta)_{m,m'}\right| > \frac{\delta}{2rS} \right\}\right) \leq 2M^2 \exp\left(-\frac{N_\beta \delta^2 \sigma^2}{32 B_0 r^2 S^2}\right),$$

where we have used $\binom{M}{2} < \frac{M^2}{2}$. Similarly, the inequality of Eq. A.4 and the union bounds yield:

$$\mathbb{P}\left(\bigcup_{m=1}^{M} \left\{ |(\mathbf{C}_\beta)_{m,m} - 1| > \frac{\delta}{2rS} \right\}\right) \leq 2M \exp\left(-\frac{N_\beta \delta^2 \sigma^2}{4 B_0 r^2 S^2}\right).$$

Now, by invoking the Gershgorin's disc theorem, the eigenvalues of any sub-matrix of $\mathbf{C}_\beta$ *restricted to an index set* S *with* $|\mathcal{S}| = rS$*, lie in the interval* $[1-\delta, 1+\delta]$ *with probability at least:*

$$\begin{aligned} &1 - 2M^2 \exp\left(-\frac{N_\beta \delta^2 \sigma^2}{32 B_0 r^2 S^2}\right) - 2M \exp\left(-\frac{N_\beta \delta^2 \sigma^2}{4 B_0 r^2 S^2}\right) \\ &\geq 1 - 3M^2 \exp\left(-\frac{N_\beta \delta^2 \sigma^2}{32 B_0 r^2 S^2}\right). \end{aligned}$$

Hence, by choosing $N_\beta \geq \frac{32 B_0 \sigma^2 r^2 (d+2)}{\delta^2} S^2 \log M$*, the probability above is greater than* $1 - \frac{3}{M^d}$*.*

Next, by invoking Lemma 4.1 (ii) of [143], we have that $\mathbf{C}_\beta$ *satisfies the RSC condition over the set given by Eq. A.2 with a constant given by:*

$$\kappa_0 = \frac{(1-\delta)\left(1 - 3\sqrt{\frac{1+\delta}{(r-1)(1-\delta)}}\right)^2}{\left(1 + \frac{9}{r-1}\right)}. \tag{A.8}$$

Hence, the negative log-likelihood satisfies the RSC with a constant given by $p_{\min}(1 - p_{\max})\sigma^2 N_\beta \kappa_0$*. Finally, by taking* $K \geq \frac{\log 2}{\log\left(\frac{1}{\beta}\right)}$*, we have that* $N_\beta \geq \frac{W}{2(1-\beta)}$*, which makes* κ *independent of* K *and* β*, given by:*

$$\kappa := \frac{p_{\min}(1 - p_{\max})\sigma^2 \kappa_0 W}{2}, \tag{A.9}$$

and $\beta \geq 1 - \frac{C'}{S^2 \log M}$ *with* $C' := \frac{W \delta^2}{64 B_0 \sigma^2 r^2 (d+2)}$*.*

Next, the result of Theorem 1 of [19] implies:

$$\|\widehat{\boldsymbol{\omega}} - \boldsymbol{\omega}\|_2 \leq \frac{2\gamma\sqrt{S}}{\kappa} + \sqrt{\frac{2\gamma \sigma_S(\boldsymbol{\omega})}{\kappa}}, \tag{A.10}$$

for $\gamma > 2\|\nabla \ell^\beta(\boldsymbol{\omega})\|_\infty$. We have, for $m \neq 1$,

$$\left(\nabla \ell^\beta(\boldsymbol{\omega})\right)_m = \sum_{i=1}^{K} \sum_{j=1}^{W} \beta^{K-i} s_{(i-1)W+j+m-M+1} \left(n_{(i-1)W+j+m-M+1} - \lambda_{(i-1)W+j+m-M+1} \Delta\right). \tag{A.11}$$

Now, let $\mathcal{F}_t$ be the σ-field generated by $s_{-M+1}, \cdots, s_t$, i.e., $\sigma(s_{-M+1}, \cdots, s_t)$. We have that

$$\begin{aligned}\mathbb{E}\left\{(n_t - \lambda_t\Delta)\, s_t\right\} &= \mathbb{E}\left\{\mathbb{E}\left\{(n_t - \lambda_t\Delta)\, s_t|\mathcal{F}_t\right\}\right\} \\ &= \mathbb{E}\left\{s_t\mathbb{E}\left\{(n_t - \lambda_t\Delta)\,|\mathcal{F}_t\right\}\right\} \\ &= \mathbb{E}\Big\{ s_t\mathbb{E}\big\{ \underbrace{(\lambda_t\Delta - \lambda_t\Delta)}_{=0}|\mathcal{F}_t\big\}\Big\} = 0.\end{aligned}$$

Hence for all m, $\mathbb{E}\left\{\left(\nabla\ell^\beta(\boldsymbol{\omega})\right)_m\right\} = 0$. Noting that $\left(\nabla\ell^\beta(\boldsymbol{\omega})\right)_m$ is a sum of martingale differences, we next invoke the following result for concentration of dependent random variables:

Proposition A.1 *Consider a sequence of σ-fields $\mathcal{F}_0 \subset \mathcal{F}_1 \subset \cdots$. Suppose that X_i is $\mathcal{F}_i$-measurable with $|X_i| \leq B_i$ for some constant B_i, $i = 1, 2, \cdots$ and $\mathbb{E}\{X_i|\mathcal{F}_{i-1}\} = 0$. Then for all $t > 0$,*

$$\mathbb{P}\left(\sum_{i=1}^{n} X_i \geq t\right) \leq \exp\left(-\frac{t^2}{2\sum_{i=1}^{n} B_i^2}\right).$$

Proof A.2 *This result is a special case of Theorem 2.5 of [142] for bounded and possibly dependent random variables, which generalizes Hoeffding's inequality.*

In our case, we can similarly show that $\mathbb{E}\big\{s_t(n_t - \lambda_t\Delta)|\mathcal{F}_{t-1}\big\} = \mathbb{E}\big\{s_t\mathbb{E}\big\{(n_t - \lambda_t\Delta)|\mathcal{F}_{t-1}, \mathcal{F}_t\big\}\big\} = 0$. Moreover, each summand is bounded by $2\beta^{K-i}B$. Hence, using the result of Proposition A.1, by taking $n = TW$ and $X_i = s_i(n_i - \lambda_i\Delta)$, we get:

$$\begin{aligned}\mathbb{P}\left(\left|\left(\nabla\ell^\beta(\boldsymbol{\omega})\right)_m\right| > tN_\beta\right) &\leq 2\exp\left(-\frac{t^2N_\beta^2}{8\sum_{i=1}^{K} W\beta^{2(K-i)}}\right) \\ &\leq 2\exp\left(-\frac{N_\beta t^2}{8}\right).\end{aligned}$$

Using the union bound, we have:

$$\mathbb{P}\left(\left\|\nabla \ell^{\beta}(\boldsymbol{\omega})\right\|_{\infty} > tN_{\beta}\right) \leq 2M \exp\left(-\frac{N_{\beta}t^2}{8}\right). \tag{A.12}$$

By choosing $t = \sqrt{\frac{8(d+1)\log M}{N_{\beta}}}$, we have that

$$\left\|\nabla \ell^{\beta}(\boldsymbol{\omega})\right\|_{\infty} < \sqrt{8(d+1)N_{\beta}\log M},$$

with probability at least $1 - \frac{2}{M^d}$.

Hence, for a fixed $\delta < 1$, $d > 0$, and $r > 1 + \frac{9(1+\delta)}{1-\delta}$, by taking $\beta \geq 1 - \frac{C'}{S^2 \log M}$ with $C' := \frac{W\delta^2}{64B_0\sigma^2 r^2(d+2)}$ and $\gamma = C''\sqrt{\frac{\log M}{1-\beta}}$ with $C'' := \sqrt{32(d+1)W}$, any maximizer $\widehat{\boldsymbol{\omega}}$ satisfies:

$$\|\widehat{\boldsymbol{\omega}} - \boldsymbol{\omega}\|_2 \leq C\sqrt{(1-\beta)S\log M} + \sqrt{C\sigma_S(\boldsymbol{\omega})}\sqrt[4]{(1-\beta)S\log M},$$

with probability at least $1 - \frac{3}{M^d} - \frac{2}{M^d}$, where C is given by

$$C := \frac{\sqrt{512(d+1)}\left(1 + \frac{9}{r-1}\right)}{\sqrt{W}p_{\min}(1-p_{\max})\sigma^2(1-\delta)\left(1 - 3\sqrt{\frac{1+\delta}{(r-1)(1-\delta)}}\right)^2}. \tag{A.13}$$

A.2 The Proximal Gradient Algorithm

In this appendix, we give an overview of the proximal gradient algorithm for minimization of convex functions. The corresponding algorithm for maximization of concave functions can be obtained by negating the objective functions. Consider the general optimization problem

$$\min_{\mathbf{x}} \; f(\mathbf{x}) + g(\mathbf{x}), \tag{A.14}$$

where functions $f(\mathbf{x}) : \mathbb{R}^M \to \mathbb{R}$ and $g(\mathbf{x}) : \mathbb{R}^M \to \mathbb{R} \cup \{\infty\}$ are assumed to be closed proper convex functions. Suppose that f is differentiable with a Lipschitz

continuous gradient ∇f with constant $L(\nabla f)$. The function g can be possibly non-smooth. A wide range of practical optimization problems can be cast in this form, particularly in the context of machine learning [145], where the objective function can be decomposed into a loss function and a regularization term.

The proximal gradient algorithm provides an iterative procedure for solving Eq. A.14 in the following form:

$$\mathbf{x}^{(\ell+1)} = \mathcal{P}_{\alpha^{(\ell)}g}\Big[\mathbf{x}^{(\ell)} - \alpha^{(\ell)}\nabla f(\mathbf{x}^{(\ell)})\Big], \tag{A.15}$$

where the parameter $\alpha^{(\ell)}$ is an appropriately chosen step size at iteration ℓ so that $\alpha^{(\ell)} < \frac{1}{L(\nabla f)}$, and the *proximal operator* $\mathcal{P}_{\alpha g}(.)$ of function g with parameter α is defined as

$$\mathcal{P}_{\alpha g}(\mathbf{x}) := \underset{\mathbf{u}}{\operatorname{argmin}} \Big\{ g(\mathbf{u}) + \frac{1}{2\alpha}\|\mathbf{u}-\mathbf{x}\|_2^2 \Big\}. \tag{A.16}$$

Among the several interpretations available for the proximal gradient method, we adopted a quadratic approximation-based model to derive the main iterative scheme in Eq. A.15. This interpretation [146,147], is based on the Majorization-Minimization algorithm (see [148] for a detailed discussion). In the approximation-based derivation, the ℓ-th iteration for solving the general problem of Eq. A.14 can be written in the following form:

$$\mathbf{x}^{(\ell+1)} = \underset{\mathbf{x}}{\operatorname{argmin}} \Big\{ \widehat{f}_\alpha(\mathbf{x}, \mathbf{x}^{(\ell)}) + g(\mathbf{x}) \Big\}, \tag{A.17}$$

where the original objective function f is replaced with a quadratically-regularized linear approximation around the previous iterate $\mathbf{x}^{(\ell)}$, given by

$$\widehat{f}_\alpha(\mathbf{x}, \mathbf{y}) := f(\mathbf{y}) + \nabla f(\mathbf{y})'(\mathbf{x}-\mathbf{y}) + \frac{1}{2\alpha}\|\mathbf{x}-\mathbf{y}\|_2^2,$$

where the quadratic term is referred to as the *trust region penalty*. Modulo constants, the objective function in Eq. A.17 can be rearranged to get the proximal gradient form

$$\begin{aligned}\mathbf{x}^{(\ell+1)} &= \underset{\mathbf{x}}{\operatorname{argmin}}\Big\{g(\mathbf{x}) + \frac{1}{2\alpha}\left\|\mathbf{x} - \mathbf{x}^{(\ell)} + \alpha\nabla f(\mathbf{x}^{(\ell)})\right\|_2^2\Big\} \\ &= \mathcal{P}_{\alpha g}\Big[\mathbf{x}^{(\ell)} - \alpha\nabla f(\mathbf{x}^{(\ell)})\Big]. \end{aligned} \tag{A.18}$$

The proximal operator often admits closed form expressions. As for ℓ_1- regularization, the proximal operator takes the simple form of the *soft thresholding shrinkage operator* $\mathcal{P}_{\alpha\|.\|_1} =: \mathscr{S}_\alpha$ whose ith component is given by

$$(\mathscr{S}_\alpha(x))_i := \operatorname{sgn}(x_i)(|x_i| - \alpha)_+,$$

with sgn denoting the standard signum function, and $(a)_+ := \max\{a, 0\}$. In this case, the proximal algorithm leads to a family of algorithms referred to as iterative shrinkage algorithms [20, 149, 150], where each iteration involves a simple gradient descent step followed by a shrinkage operation.

Finally, in our setting, the function f is taken to be the exponentially weighted log-likelihood $\ell^\beta(\cdot)$. Due to the smoothness of the logistic function, the Lipschitz constant for $\nabla\ell^\beta(\boldsymbol{\omega}_k)$ can be upper bounded by the trace of the Hessian $\mathbf{B}_k(\boldsymbol{\omega}_k)$ given in Eq. 3.14. Noting that the elements of $\boldsymbol{\Lambda}_i$ are at most equal to 1/4 yields $L(\nabla\ell^\beta(\boldsymbol{\omega}_k)) \leq \frac{1}{4}\sum_{i=1}^{k}\sum_{j=1}^{W}\beta^{k-i}x^2_{(i-1)W+j}$. Using assumption (1) of the proof in Appendix A.1 and an application of Hoeffding's inequality, we can show that the sum is concentrated around its mean given by $\frac{MW\sigma^2}{4(1-\beta)}$, for large enough k. Therefore, we choose the step size $\alpha = \frac{(1-\beta)}{cMW\sigma^2}$, for some constant $c \geq 1/4$.

A.3 Computation of Confidence Intervals

The ℓ_1-regularized ML estimate of Eq. 3.4 can be written in the following form

$$\widehat{\boldsymbol{\omega}}_k = \underset{\boldsymbol{\omega}_k}{\operatorname{argmax}} \left\{ \mathfrak{P}_\beta \ell(\boldsymbol{\omega}_k) - \gamma \|\boldsymbol{\omega}_k\|_1 \right\},$$

where $\ell(\boldsymbol{\omega}) := \log p\big(\mathbf{n}|\mathbf{X}, \boldsymbol{\omega}\big)$ denotes the log-likelihood function over a generic window with spiking vector $\mathbf{n}$, data matrix $\mathbf{X}$ and parameter vector $\boldsymbol{\omega}$, and the operator $\mathfrak{P}_\beta f(\mathbf{n}, \mathbf{X}, \boldsymbol{\omega})$ is defined for a function $f : \{0,1\}^W \times \mathbb{R}^{W \times M} \times \mathbb{R}^M \to \mathbb{R}$ as the empirical expectation exponentially weighted with a forgetting factor β:

$$\mathfrak{P}_\beta f(\boldsymbol{\omega}) := \sum_{i=1}^{k} \beta^{k-i} f\big(\mathbf{n}_i, \mathbf{X}_i; \boldsymbol{\omega}\big), \tag{A.19}$$

where we have suppressed the dependence of f on $\mathbf{n}$ and $\mathbf{X}$ on the left hand side for notational simplicity. Following the treatment of Theorem 3.1 of [142], the corresponding empirical gradient vector and Hessian are respectively given by:

$$\mathbf{g}_k(\boldsymbol{\omega}_k) := \mathfrak{P}_\beta \nabla \ell(\boldsymbol{\omega}_k) = \sum_{i=1}^{k} \beta^{k-i} \mathbf{X}_i' \boldsymbol{\varepsilon}_i(\boldsymbol{\omega}_k), \tag{A.20}$$

$$\mathbf{B}_k(\boldsymbol{\omega}_k) := \mathfrak{P}_\beta \nabla^2 \ell(\boldsymbol{\omega}_k) = -\sum_{i=1}^{k} \beta^{k-i}\, \mathbf{X}_i' \boldsymbol{\Lambda}_i(\boldsymbol{\omega}_k) \mathbf{X}_i. \tag{A.21}$$

The KKT conditions for the estimator $\widehat{\boldsymbol{\omega}}_k$ can be then written as:

$$\mathbf{g}_k(\widehat{\boldsymbol{\omega}}_k) - \gamma \widehat{\mathbf{s}}_k = 0, \quad \|\widehat{\mathbf{s}}_k\|_\infty \leq 1.$$

where $\widehat{\mathbf{s}}_k \in \partial \|\widehat{\boldsymbol{\omega}}_k\|_1$ is a subgradient vector from the subdifferential of the ℓ_1 norm, with components $(\widehat{\mathbf{s}}_k)_m = \operatorname{sgn}\left((\widehat{\boldsymbol{\omega}}_k)_m\right)$ for $(\widehat{\boldsymbol{\omega}}_k)_m \neq 0$ and $|(\widehat{\mathbf{s}}_k)_m| \leq 1$ otherwise, for $m = 1, 2, \ldots, M$. Substituting $\mathfrak{P}_\beta \ell(\boldsymbol{\omega}_k)$ by its quadratic approximation around

the true parameter vector $\boldsymbol{\omega}_k$, and inverting the corresponding KKT conditions, the *de-biased* estimator $\widehat{\mathbf{w}}_k$ can be obtained as:

$$\widehat{\mathbf{w}}_k := \widehat{\boldsymbol{\omega}}_k - \widehat{\boldsymbol{\Theta}}_k \mathbf{g}_k(\widehat{\boldsymbol{\omega}}_k), \tag{A.22}$$

where the matrix $\widehat{\boldsymbol{\Theta}}_k$ is the approximate inverse of Hessian matrix $\mathbf{B}_k(\widehat{\boldsymbol{\omega}}_k)$, and can be computed using the following node wise regression procedure [142]. To compute the m-th row of $\widehat{\boldsymbol{\Theta}}_k$, first the solution to the following LASSO problem is obtained:

$$\widehat{\boldsymbol{\psi}}_m := \underset{\boldsymbol{\psi} \in \mathbb{R}^{M-1}}{\operatorname{argmin}} \Big(- 2(\mathbf{B}_k)_{m,\backslash m} \boldsymbol{\psi} + \boldsymbol{\psi}' (\mathbf{B}_k)_{\backslash m, \backslash m} \boldsymbol{\psi} + 2\gamma_m \|\boldsymbol{\psi}\|_1 \Big), \tag{A.23}$$

where the dependence of $\mathbf{B}_k$ on $\widehat{\boldsymbol{\omega}}_k$ is suppressed for notational convenience, and the subscript notations are the same as those described in the footnote of Algorithm 3. Then, we define the vector $\mathbf{c} \in \mathbb{R}^M$ as:

$$(\mathbf{c})_m = 1, \quad (\mathbf{c})_{\backslash m} = -\widehat{\boldsymbol{\psi}}_m^{(L)}, \tag{A.24}$$

and the scaling constant τ_m^2 as

$$\tau_m^2 := (\mathbf{B}_k)_{m,m} - \widehat{\boldsymbol{\psi}}_m^{(L)} (\mathbf{B}_k)'_{m,\backslash m}. \tag{A.25}$$

Finally, the m-th row of $\widehat{\boldsymbol{\Theta}}_k$ is given by $(\widehat{\boldsymbol{\Theta}}_k)_m := \frac{1}{\tau_m^2} \mathbf{c}$. The variance and the confidence interval at a level of α for the m-th component of $\widehat{\boldsymbol{\omega}}_k$ can then be computed as given in lines 9 and the output of Algorithm 3 [142], where

$$\mathbf{G}_k(\boldsymbol{\omega}) := \mathfrak{P}_{\beta^2} \nabla \ell(\boldsymbol{\omega}) \nabla \ell'(\boldsymbol{\omega}) = \sum_{i=1}^{k} \beta^{2(k-i)} \mathbf{X}_i' \boldsymbol{\varepsilon}_i(\boldsymbol{\omega}) \boldsymbol{\varepsilon}_i(\boldsymbol{\omega})' \mathbf{X}_i. \tag{A.26}$$

Using Taylor expansion similar to that in the development of ℓ_1-PPF$_1$, the matrix $\mathbf{G}_k(\widehat{\boldsymbol{\omega}}_k)$ can be recursively updated as given in line 2 of Algorithm 3. Finally,

the node wise regression can be recursively computed using the SPARLS algorithm [21], which is given in lines 3–5 of Algorithm 3. The parameter γ_m can be chosen to be in the same order of γ in Eq. 3.4.

Appendix B: Proof of Theorem 4.1: Asymptotic Distributional Analysis of the Adaptive De-biased Deviance Statistic

In this Appendix, we provide the proof of Theorem 4.1, followed by a discussion of the results and their implications. Before presenting the proof of Theorem 4.1, we make a few technical assumptions, and introduce further notations.

1) We consider a scaling of $\gamma = \mathcal{O}(\sqrt{(1-\beta)\log M})$, where M denotes the model order ($M^{(F)}$ or $M^{(R)}$). This assumption leads to asymptotic consistency of ℓ_1-regularized ML estimation [33,87], similar to that used in Theorem 3.1 of chapter 3.

2) We assume that the stimuli $\{s_t\}_{t=1}^T$ form a Markovian random sequence. This assumption facilitates the limiting arguments used in our asymptotic analysis.

For a log-likelihood function $\ell(\boldsymbol{\omega})$ with parameter vector $\boldsymbol{\omega}$, we define:

$$\dot{\boldsymbol{\ell}}(\boldsymbol{\omega}) := \nabla_{\boldsymbol{\omega}} \ell(\boldsymbol{\omega}), \tag{B.1}$$

$$\ddot{\boldsymbol{\ell}}(\boldsymbol{\omega}) := \nabla^2_{\boldsymbol{\omega}} \ell(\boldsymbol{\omega}), \tag{B.2}$$

$$\boldsymbol{\mathcal{I}}(\boldsymbol{\omega}) := \mathbb{E}\left\{\dot{\boldsymbol{\ell}}(\boldsymbol{\omega})\dot{\boldsymbol{\ell}}'(\boldsymbol{\omega})\right\}, \tag{B.3}$$

where $\dot{\boldsymbol{\ell}}(\cdot)$ is the gradient of the log-likelihood with respect to the parameter vector $\boldsymbol{\omega}$, known as the *score* statistic, $\ddot{\boldsymbol{\ell}}(\cdot)$ is the Hessian of the log-likelihood, and $\boldsymbol{\mathcal{I}}(\cdot)$ denotes the Fisher information matrix as the covariance of the score vector, where

the expectation is over the realization of the process.

For simplicity of analysis, we consider a piece-wise constant model in which $\boldsymbol{\omega}_k$ is constant within observation windows indexed by $i = k-N, k-N+1, \ldots, k$ for some large $N = \mathcal{O}(\frac{1}{1-\beta})$, following the tradition of performance analysis of RLS-type algorithms [13]. Recall that the exponentially weighted log-likelihood at window k is given by:

$$\ell_k^{\beta}(\boldsymbol{\omega}_k) := (1-\beta) \sum_{i=k-N}^{k} \beta^{k-i} \ell_i(\boldsymbol{\omega}_k). \tag{B.4}$$

Let $\boldsymbol{\omega}_k$ and $\widehat{\boldsymbol{\omega}}_k$ denote the true and estimated parameter vectors of length M associated with a unit at window k, where M can take any of the two values $M^{(F)}$ and $M^{(R)}$ corresponding to full and reduced models, respectively. Suppose that the inverse Hessian exists at $\boldsymbol{\omega}_k$ for each time k, which we denote by $\boldsymbol{\Theta}_k := (\ddot{\boldsymbol{\ell}}_k^{\beta}(\boldsymbol{\omega}_k))^{-1}$ for notational convenience. Throughout the proof, we make use of the consistency results on ℓ_1-regularized exponentially-weighted maximum likelihood estimation, such as those discussed earlier in Chapter 3. These results imply that for β close enough to 1, we have $\|\widehat{\boldsymbol{\omega}}_k - \boldsymbol{\omega}_k\|_2 = \mathcal{O}(\sqrt{(1-\beta) S \log M})$, with a choice of $\gamma = \mathcal{O}(\sqrt{(1-\beta)\log M})$ for the regularization parameter.

The de-biased deviance $D_{k,\beta}(\widehat{\boldsymbol{\omega}}_k; \boldsymbol{\omega}_k)$ of Eq. 4.4 can be expressed in the following quadratic form:

$$D_{k,\beta}(\widehat{\boldsymbol{\omega}}_k; \boldsymbol{\omega}_k) = -\left(\frac{1+\beta}{1-\beta}\right) (\widehat{\mathbf{w}}_k - \boldsymbol{\omega}_k)' \, \ddot{\boldsymbol{\ell}}_k^{\beta}(\boldsymbol{\omega}_k) \, (\widehat{\mathbf{w}}_k - \boldsymbol{\omega}_k), \tag{B.5}$$

where

$$\widehat{\mathbf{w}}_k := \widehat{\boldsymbol{\omega}}_k - \boldsymbol{\Theta}_k \dot{\boldsymbol{\ell}}_k^{\beta}(\widehat{\boldsymbol{\omega}}_k). \tag{B.6}$$

By rearranging some terms, Eq. B.5 can be expressed as:

$$\left(\frac{1-\beta}{1+\beta}\right) D_{k,\beta}(\widehat{\boldsymbol{\omega}}_k;\boldsymbol{\omega}_k) = 2(\widehat{\boldsymbol{\omega}}_k-\boldsymbol{\omega}_k)'\dot{\boldsymbol{\ell}}_k^{\beta}(\widehat{\boldsymbol{\omega}}_k) - (\widehat{\boldsymbol{\omega}}_k-\boldsymbol{\omega}_k)'\ddot{\boldsymbol{\ell}}_k^{\beta}(\widehat{\boldsymbol{\omega}}_k)(\widehat{\boldsymbol{\omega}}_k-\boldsymbol{\omega}_k) - B_k + \Delta_1, \tag{B.7}$$

where $B_k := \dot{\boldsymbol{\ell}}_k^{\beta}(\widehat{\boldsymbol{\omega}}_k)'\boldsymbol{\Theta}_k\dot{\boldsymbol{\ell}}_k^{\beta}(\widehat{\boldsymbol{\omega}}_k)$ denotes the bias term due to ℓ_1-regularization, and Δ_1 denotes a remainder term given by:

$$\Delta_1 := (\widehat{\boldsymbol{\omega}}_k-\boldsymbol{\omega}_k)'\big(\ddot{\boldsymbol{\ell}}_k^{\beta}(\widehat{\boldsymbol{\omega}}_k) - \ddot{\boldsymbol{\ell}}_k^{\beta}(\boldsymbol{\omega}_k)\big)(\widehat{\boldsymbol{\omega}}_k-\boldsymbol{\omega}_k). \tag{B.8}$$

Next, we use the Taylor's series expansion as follows:

$$\ell_k^{\beta}(\boldsymbol{\omega}_k) = \ell_k^{\beta}(\widehat{\boldsymbol{\omega}}_k) + (\boldsymbol{\omega}_k-\widehat{\boldsymbol{\omega}}_k)'\dot{\boldsymbol{\ell}}_k^{\beta}(\widehat{\boldsymbol{\omega}}_k) + \frac{1}{2}(\boldsymbol{\omega}_k-\widehat{\boldsymbol{\omega}}_k)'\ddot{\boldsymbol{\ell}}_k^{\beta}(\widetilde{\boldsymbol{\omega}}_k)(\boldsymbol{\omega}_k-\widehat{\boldsymbol{\omega}}_k), \tag{B.9}$$

where $\widetilde{\boldsymbol{\omega}}_k := t\boldsymbol{\omega}_k + (1-t)\widehat{\boldsymbol{\omega}}_k$ is an intermediate vector for some $t \in (0,1)$, such that $\|\widetilde{\boldsymbol{\omega}}_k-\boldsymbol{\omega}_k\| < \|\widehat{\boldsymbol{\omega}}_k-\boldsymbol{\omega}_k\|$. Combining Eqs. B.7 and B.9, we get:

$$\left(\frac{1-\beta}{1+\beta}\right) D_{k,\beta}(\widehat{\boldsymbol{\omega}}_k;\boldsymbol{\omega}_k) = 2(\ell_k^{\beta}(\widehat{\boldsymbol{\omega}}_k)-\ell_k^{\beta}(\boldsymbol{\omega}_k)) - B_k + \Delta_2, \tag{B.10}$$

where the remainder term Δ_2 takes a similar form to Eq. B.8 with the Hessian evaluated at $\tilde{\boldsymbol{\omega}}_k$ instead. Using the Lipschitz property of the second-order derivative of the logistic function, boundedness assumption on the covariates ($\|\mathbf{X}\|_\infty = \mathcal{O}(K)$), and the consistency of $\widehat{\boldsymbol{\omega}}_k$, it can be proved that both remainder terms Δ_1 and Δ_2 are asymptotically negligible with a rate of $\|\widehat{\boldsymbol{\omega}}_k-\boldsymbol{\omega}_k\|^3 = o_{\mathbb{P}}((1-\beta)^{3/2})$ as $\beta \to 1$.

In order to adapt the treatment of Davidson and Lever [97] to our setting, we first consider a sequence of forgetting factors $\{\beta_j\}_{j=1}^{\infty}$ approaching unity, i.e., $\lim_{j\to\infty}\beta_j = 1$. Then, at window k, we test the null hypothesis $H_{0,k} : \boldsymbol{\omega}_k^0 = (\boldsymbol{\omega}_{0,k}, \mathbf{0})$ against a sequence of local alternatives $\{H_{1,k}^{\beta_j}\}_{j=1}^{\infty} = \{H_{1,k}^{\beta_j} : \boldsymbol{\omega}_k^{\beta_j} = (\boldsymbol{\omega}_{0,k}^*, \boldsymbol{\omega}_{1,k}^{\beta_j})\}$,

where $\boldsymbol{\omega}_{1,k}^{\beta_j} = \sqrt{\frac{1-\beta_j}{1+\beta_j}}\boldsymbol{\delta}_k$ corresponds to the unspecified sub-vector excluded in the reduced model for some constant vector $\boldsymbol{\delta}_k$ of dimension $M^{(d)}$.

Statistical inference under the sequence of local alternatives $\{H_{1,k}^{\beta_j}\}$ is carried out through testing local departures from null hypothesis to the limiting true parameter $\boldsymbol{\omega}_k^*$ at the rate of $\mathcal{O}\left(\sqrt{\frac{1-\beta_j}{1+\beta_j}}\right)$ as $\beta_j \to 1$. For notational convenience, we drop the dependence of β_j on the index j. It is understood that expressions involving limits of β are interpreted as the sequential limit. From the definition of $\widehat{\mathbf{w}}_k$ in Eq. B.6, it follows that:

$$\widehat{\mathbf{w}}_k - \boldsymbol{\omega}_k = \widehat{\boldsymbol{\omega}}_k - \boldsymbol{\omega}_k - \boldsymbol{\Theta}_k \dot{\boldsymbol{\ell}}_k^{\beta}(\widehat{\boldsymbol{\omega}}_k) = -\boldsymbol{\Theta}_k \dot{\boldsymbol{\ell}}_k^{\beta}(\boldsymbol{\omega}_k) + \boldsymbol{\Delta}, \tag{B.11}$$

where $\boldsymbol{\Delta} := (\mathbf{I} - \boldsymbol{\Theta}_k \ddot{\boldsymbol{\ell}}_k^{\beta}(\widetilde{\boldsymbol{\omega}}_k))(\widehat{\boldsymbol{\omega}}_k - \boldsymbol{\omega}_k)$, and we used:

$$\dot{\boldsymbol{\ell}}_k^{\beta}(\widehat{\boldsymbol{\omega}}_k) = \dot{\boldsymbol{\ell}}_k^{\beta}(\boldsymbol{\omega}_k) + \ddot{\boldsymbol{\ell}}_k^{\beta}(\widetilde{\boldsymbol{\omega}}_k)(\widehat{\boldsymbol{\omega}}_k - \boldsymbol{\omega}_k), \tag{B.12}$$

in Eq. B.11 which holds for some intermediate vector $\widetilde{\boldsymbol{\omega}}_k = t\boldsymbol{\omega}_k + (1-t)\widehat{\boldsymbol{\omega}}_k$ for some $t \in (0,1)$. It can be shown that $\boldsymbol{\Delta} = o_{\mathbb{P}}(1-\beta)$ is asymptotically negligible, following the aforementioned argument used for Δ_1 and Δ_2.

Next, we need to determine the asymptotic behavior of the Hessian $\ddot{\boldsymbol{\ell}}_k^{\beta}(\boldsymbol{\omega}_k)$ as $\beta \to 1$. Due to the dependencies of the covariates, the common law of large numbers (LLN) for i.i.d. random variables cannot be applied. Due to the logistic link used in defining the log-likelihood, the Hessian can be written as $(1-\beta)\mathbf{X}'\mathbf{WDX}$, where $\mathbf{W}$ is a diagonal bounded weighing matrix, $\mathbf{D}$ is a diagonal matrix containing the exponential weights, and $\mathbf{X}$ is the matrix of covariates [33]. Also, for finite M, $\{n_i^{(c)}\}_{c=1}^{C}$ form a 2^C-state Markov chain with ϕ-mixing property. Hence, the

version of LLN for bounded functions of ϕ-mixing random variables can be used to characterize the limit (e.g., [151] or Theorem 27.4 in [152]). Hence, as $\beta \to 1$:

$$\ddot{\boldsymbol{\ell}}_k^{\beta}(\boldsymbol{\omega}_k) \xrightarrow{p} \mathbb{E}[\ddot{\boldsymbol{\ell}}_i(\boldsymbol{\omega}_k)] = -\boldsymbol{\mathcal{I}}(\boldsymbol{\omega}_k), \tag{B.13}$$

where the second equality is obtained using the Fisher information equality.

Similarly, in order to characterize the asymptotic behavior of the score statistic, a version of the Central Limit Theorem (CLT) for dependent random variables is required. Note that the Lindeberg CLT for i.i.d. random variables does not apply, since the covariates are highly dependent. In the absence of the stimuli in the logistic model, i.e., $\mathbf{s}_i = \mathbf{0}, \forall i$, by invoking the aforementioned ϕ-mixing property of the equivalent 2^C-state Markov chain $\{n_i^{(c)}\}_{c=1}^{C}$, we use a version of the martingale CLT [151]. In the presence of stimuli, by the hypothesis that the stimuli are generated by a Markov process, we invoke stronger versions of the CLT for autoregressive models [153, 154]. Hence, the score statistic at the true parameter converges in distribution to a Gaussian random vector with zero mean and covariance given by the Fisher information matrix:

$$\sqrt{\frac{1+\beta}{1-\beta}}\, \dot{\boldsymbol{\ell}}_k^{\beta}(\boldsymbol{\omega}_k) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{\mathcal{I}}(\boldsymbol{\omega}_k)), \tag{B.14}$$

as $\beta \to 1$. Note that this result holds both under $H_{0,k}$ when $\boldsymbol{\omega}_k = \boldsymbol{\omega}_k^0$ is the true parameter vector, and for the sequence of alternatives $H_{1,k}^{\beta}$, where $\boldsymbol{\omega}_k = \boldsymbol{\omega}_k^{\beta}$ is the sequence of true parameters.

The asymptotic normality of $\widehat{\mathbf{w}}_k$ under $H_{0,k}$ follows by invoking the Slutsky's

theorem using Eqs. B.13 and B.14:

$$\sqrt{\frac{1+\beta}{1-\beta}}(\widehat{\mathbf{w}}_k - \boldsymbol{\omega}_k) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \boldsymbol{\mathcal{I}}(\boldsymbol{\omega}_k)^{-1}\right), \tag{B.15}$$

as $\beta \to 1$. Hence, under H_0, combining the asymptotic result on the Hessian in Eq. B.13, and the asymptotic normality of $\widehat{\mathbf{w}}_k$ in Eq. B.15 leads to the weak convergence of the adaptive de-biased deviance to a central chi-squared distribution with M degrees of freedom:

$$\left[D_{k,\beta}(\widehat{\boldsymbol{\omega}}_k; \boldsymbol{\omega}_k) \,|\, H_{0,k}\right] \xrightarrow{d} \chi^2(M), \tag{B.16}$$

as $\beta \to 1$. Following on the classical results [95, 96], it can be shown that the deviance difference of two nested full and reduced models asymptotically converges in distribution to a central chi-squared with $M^{(d)}$ degrees of freedom:

$$\left[D_{k,\beta}(\widehat{\boldsymbol{\omega}}_k^{\beta}; \widehat{\boldsymbol{\omega}}_k^{0}) \,\middle|\, H_{0,k}\right] \xrightarrow{d} \chi^2(M^{(d)}), \tag{B.17}$$

as $\beta \to 1$, where $M^{(d)}$ is the dimension of the specified sub-vector $\boldsymbol{\omega}_{1,k} = \mathbf{0}$ under the null hypothesis, i.e, the dimensionality difference of the two nested models. This establishes part (i) of the statement of Theorem 4.1.

As for part (ii), such an asymptotic result under the sequence of local alternative hypotheses will be slightly different, as the limiting Gaussian distributions are non-zero mean. To see this, we define the de-biased vector $\widehat{\mathbf{w}}_k^{\beta}$ associated with each local alternative $H_{1,k}^{\beta}$ at time step k as:

$$\widehat{\mathbf{w}}_k^{\beta} := \widehat{\boldsymbol{\omega}}_k^{\beta} - \boldsymbol{\Theta}_k^{*} \dot{\boldsymbol{\ell}}_k^{\beta}(\widehat{\boldsymbol{\omega}}_k^{\beta}), \tag{B.18}$$

where $\boldsymbol{\Theta}_k^* := \boldsymbol{\Theta}_k(\boldsymbol{\omega}_k^*)$. By similar arguments leading to Eq. B.11, it follows that:

$$\begin{aligned}\widehat{\mathbf{w}}_k^\beta - \boldsymbol{\omega}_k^* &= \widehat{\boldsymbol{\omega}}_k^\beta - \boldsymbol{\omega}_k^* - \boldsymbol{\Theta}_k^* \dot{\boldsymbol{\ell}}_k^\beta(\widehat{\boldsymbol{\omega}}_k^\beta) \\ &= -\boldsymbol{\Theta}_k^* \dot{\boldsymbol{\ell}}_k^\beta(\boldsymbol{\omega}_k^*) + o_{\mathbb{P}}(1-\beta) \qquad \text{(B.19)} \\ &= \boldsymbol{\omega}_k^\beta - \boldsymbol{\omega}_k^* - \boldsymbol{\Theta}_k^* \dot{\boldsymbol{\ell}}_k^\beta(\boldsymbol{\omega}_k^\beta) + o_{\mathbb{P}}(1-\beta), \qquad \text{(B.20)}\end{aligned}$$

where we have respectively used the following linear expansions around $\boldsymbol{\omega}_k^*$ in Eq. B.19 and B.20:

$$\dot{\boldsymbol{\ell}}_k^\beta(\widehat{\boldsymbol{\omega}}_k^\beta) = \dot{\boldsymbol{\ell}}_k^\beta(\boldsymbol{\omega}_k^*) + \ddot{\boldsymbol{\ell}}_k^\beta(\boldsymbol{\omega}_k^*)(\widehat{\boldsymbol{\omega}}_k^\beta - \boldsymbol{\omega}_k^*) + o_{\mathbb{P}}(1-\beta), \qquad \text{(B.21)}$$

$$\dot{\boldsymbol{\ell}}_k^\beta(\boldsymbol{\omega}_k^\beta) = \dot{\boldsymbol{\ell}}_k^\beta(\boldsymbol{\omega}_k^*) + \ddot{\boldsymbol{\ell}}_k^\beta(\boldsymbol{\omega}_k^*)(\boldsymbol{\omega}_k^\beta - \boldsymbol{\omega}_k^*) + o_{\mathbb{P}}(1-\beta). \qquad \text{(B.22)}$$

Using similar arguments leading to Eqs. B.13 and B.14, the asymptotic form of the Hessian and the asymptotic normality of the score function at the true parameter vector $\boldsymbol{\omega}_k^\beta$ under the sequence of local alternatives $H_{1,k}^\beta$ will follow:

$$\ddot{\boldsymbol{\ell}}_k^\beta(\boldsymbol{\omega}_k^\beta) \xrightarrow{p} -\boldsymbol{\mathcal{I}}(\boldsymbol{\omega}_k^*), \qquad \text{(B.23)}$$

$$\sqrt{\frac{1+\beta}{1-\beta}}\, \dot{\boldsymbol{\ell}}_k^\beta(\boldsymbol{\omega}_k^\beta) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{\mathcal{I}}(\boldsymbol{\omega}_k^*)). \qquad \text{(B.24)}$$

Hence, incorporating the asymptotics of Eqs. B.23 and B.24 into Eq. B.20, the de-biased estimate $\widehat{\mathbf{w}}_k^\beta$ under the sequence of local alternatives $H_{1,k}^\beta$ converges in distribution to a multivariate normal distribution:

$$\sqrt{\frac{1+\beta}{1-\beta}}(\widehat{\mathbf{w}}_k^\beta - \boldsymbol{\omega}_k^*) \xrightarrow{d} \mathcal{N}(\bar{\boldsymbol{\delta}}_k, \boldsymbol{\mathcal{I}}(\boldsymbol{\omega}_k^*)^{-1}), \qquad \text{(B.25)}$$

with non-zero asymptotic mean $\bar{\boldsymbol{\delta}}_k := [\mathbf{0}', \boldsymbol{\delta}_k']'$ as $\beta \to 1$. The asymptotic mean is obtained from the asymptotic rate of the *Pitman drift*, where the sequence of

true local parameter vectors $\{\boldsymbol{\omega}_k^\beta\}$ approach the limit $\boldsymbol{\omega}_k^*$ at a rate of $\|\boldsymbol{\omega}_k^\beta - \boldsymbol{\omega}_k^*\| = \mathcal{O}\left(\sqrt{\frac{1-\beta}{1+\beta}}\right)$.

Next, consider the decomposition of $\mathcal{I}(\boldsymbol{\omega}_k^*)$ into blocks corresponding to $\boldsymbol{\omega}_{0,k}$ and $\boldsymbol{\omega}_{1,k}$:

$$\mathcal{I}(\boldsymbol{\omega}_k^*) = \left[\begin{array}{c|c} \mathcal{I}_{0,0}(\boldsymbol{\omega}_k^*) & \mathcal{I}_{0,1}(\boldsymbol{\omega}_k^*) \\ \hline \mathcal{I}_{1,0}(\boldsymbol{\omega}_k^*) & \mathcal{I}_{1,1}(\boldsymbol{\omega}_k^*) \end{array}\right]. \tag{B.26}$$

By invoking a similar treatment as in the proof of Theorem 1 of [97] via the extension of Cochran's theorem to non-central chi-squared distribution [155, 156], and using the asymptotic result of Eq. B.25 in the quadratic forms of Eq. B.5 for both the reduced and full model estimates $(\widehat{\boldsymbol{\omega}}_k^0, \widehat{\boldsymbol{\omega}}_k^\beta)$, it can be shown that the deviance difference of two nested models converges in distribution to a non-central chi-squared distribution under the sequence of local alternatives $H_{1,k}^\beta$ as $\beta \to 1$:

$$\left[D_{k,\beta}(\widehat{\boldsymbol{\omega}}_k^\beta; \widehat{\boldsymbol{\omega}}_k^0) \,\middle|\, H_{1,k}^\beta\right] \xrightarrow{d} \chi^2(M^{(d)}, \nu_k), \tag{B.27}$$

where $M^{(d)}$ is the dimensionality difference of two nested models as before, and $\nu_k := \boldsymbol{\delta}_k' \bar{\boldsymbol{\mathcal{I}}}_{1,1}(\boldsymbol{\omega}_k^*)\boldsymbol{\delta}_k$ is the non-centrality parameter with $\bar{\boldsymbol{\mathcal{I}}}_{1,1}(\boldsymbol{\omega}_k^*) := \boldsymbol{\mathcal{I}}_{1,1}(\boldsymbol{\omega}_k^*) - \boldsymbol{\mathcal{I}}_{1,0}(\boldsymbol{\omega}_k^*)\boldsymbol{\mathcal{I}}_{0,0}^{-1}(\boldsymbol{\omega}_k^*)\boldsymbol{\mathcal{I}}_{0,1}(\boldsymbol{\omega}_k^*)$. This establishes part (ii) of the statement of Theorem 4.1.

Discussion of the Result of Theorem 4.1: Two remarks regarding the bias correction and implications of the result of Theorem 4.1 are in order:

Remark 1. The bias term B_k that emerged in the derivation of $D_{k,\beta}$ in Eq. B.7 can be estimated as $\widehat{B}_k = \dot{\boldsymbol{\ell}}_k^\beta(\widehat{\boldsymbol{\omega}}_k)' \widehat{\boldsymbol{\Theta}}_k \dot{\boldsymbol{\ell}}_k^\beta(\widehat{\boldsymbol{\omega}}_k)$, where $\widehat{\boldsymbol{\Theta}}_k = \left(\ddot{\boldsymbol{\ell}}_k^\beta(\widehat{\boldsymbol{\omega}}_k)\right)^{-1}$. Proof of the consistency of this estimate, i.e., $\widehat{B}_k \xrightarrow{p} B_k$ follows directly from the consistency of

the inverse Hessian $\widehat{\boldsymbol{\Theta}}_k \xrightarrow{p} \boldsymbol{\Theta}_k$. Since we assumed that the Hessian is invertible at true parameter $\boldsymbol{\omega}_k$, there exists a subsequence of the estimators $\{\widehat{\boldsymbol{\omega}}_k^{(\beta_\ell)}\}_\ell$, at which the Hessians are invertible, and approach the true inverse Hessian $\boldsymbol{\Theta}_k$, given that M is fixed. In the case that the Hessian $\ddot{\boldsymbol{\ell}}_k^\beta(\widehat{\boldsymbol{\omega}}_k)$ is not invertible, either due to the rank-deficiency at $\widehat{\boldsymbol{\omega}}_k$ for some k, or the case of infinitely growing dimensions $M^{(F)}$ and $M^{(R)}$ with *fixed* difference $M^{(d)}$, we adopt the approach taken in [87] and compute $\widehat{\boldsymbol{\Theta}}_k$ using the so-called *node-wise regression*, as discussed in Appendix A.3, for which similar asymptotic results have been proven, implying that $\|\widehat{\boldsymbol{\Theta}}_k - \boldsymbol{\Theta}_k\|_\infty = o_{\mathbb{P}}(1)$.

Remark 2. In the conventional asymptotic analysis of deviance, the true parameters $\{\boldsymbol{\omega}^N\}_{N=1}^\infty$ associated with the sequence of local alternatives H_1^N approach the limiting true parameter $\boldsymbol{\omega}^*$ at the rate of $\mathcal{O}\left(\frac{1}{\sqrt{N}}\right)$, where N is the number of observations. In our case, given a forgetting factor β, it follows from our asymptotic analysis that the true (cross-history) parameter $\boldsymbol{\omega}_{1,k}^\beta$ of order $\mathcal{O}\left(\sqrt{\frac{1-\beta}{1+\beta}}\right)$ associated with the alternative $H_{1,k}^\beta$ will lead to a non-trivial asymptotic distribution of the test statistic, i.e., a *non-central* chi-squared distribution. Hence, one expects that the underlying cross-history coefficients taking small values would still be detectable for β close enough to 1. In other words, the more number of observations we have for hypothesis testing, the easier it gets to distinguish between the null $H_0 : \boldsymbol{\omega}_{1,k} = \mathbf{0}$ and the alternative $H_1^\beta : \boldsymbol{\omega}_{1,k} = \boldsymbol{\omega}_1^\beta$. Therefore, we expect to detect causal links resulting from regression coefficients as small as $\mathcal{O}\left(\sqrt{\frac{1-\beta}{1+\beta}}\right)$, as stated in the theorem.

Bibliography

[1] L. M. Frank, G. B. Stanley, and E. N. Brown, "Hippocampal plasticity across multiple days of exposure to novel environments," *The Journal of neuroscience*, vol. 24, no. 35, pp. 7681–7689, 2004.

[2] D. A. Depireux, J. Z. Simon, D. J. Klein, and S. A. Shamma, "Spectro-temporal response field characterization with dynamic ripples in ferret primary auditory cortex," *Journal of neurophysiology*, vol. 85, no. 3, pp. 1220–1234, 2001.

[3] D. J. Daley and D. Vere-Jones, *An introduction to the theory of point processes: volume II: general theory and structure*. Springer Science & Business Media, 2007, vol. 2.

[4] Y. Ogata, "Statistical models for earthquake occurrences and residual analysis for point processes," *Journal of the American Statistical Association*, vol. 83, no. 401, pp. 9–27, 1988.

[5] D. Vere-Jones, "Stochastic models for earthquake occurrence," *Journal of the Royal Statistical Society. Series B*, pp. 1–62, 1970.

[6] E. N. Brown, R. E. Kass, and P. P. Mitra, "Multiple neural spike train data analysis: state-of-the-art and future challenges," *Nature neuroscience*, vol. 7, no. 5, pp. 456–461, 2004.

[7] E. N. Brown, D. P. Nguyen, L. M. Frank, M. A. Wilson, and V. Solo, "An analysis of neural receptive field plasticity by point process adaptive filtering," *Proceedings of the National Academy of Sciences*, vol. 98, no. 21, pp. 12 261–12 266, 2001.

[8] A. Smith and E. N. Brown, "Estimating a state-space model from point process observations," *Neural Comp.*, vol. 15, no. 5, pp. 965–991, 2003.

[9] L. Paninski, "Maximum likelihood estimation of cascade point-process neural encoding models," *Network: Comp. in Neural Systems*, vol. 15, no. 4, pp. 243–262, 2004.

[10] L. Paninski, J. Pillow, and J. Lewi, "Statistical models for neural encoding, decoding, and optimal stimulus design," *Progress in brain research*, vol. 165, pp. 493–507, 2007.

[11] J. W. Pillow, Y. Ahmadian, and L. Paninski, "Model-based decoding, information estimation, and change-point detection techniques for multineuron spike trains," *Neural Comp.*, vol. 23, no. 1, pp. 1–45, 2011.

[12] W. Truccolo, U. T. Eden, M. R. Fellows, J. P. Donoghue, and E. N. Brown, "A point process framework for relating neural spiking activity to spiking history, neural ensemble, and extrinsic covariate effects," *Journal of neurophysiology*, vol. 93, no. 2, pp. 1074–1089, 2005.

[13] S. S. Haykin, *Adaptive filter theory*. Pearson Education India, 2008.

[14] U. T. Eden, L. M. Frank, R. Barbieri, V. Solo, and E. N. Brown, "Dynamic analysis of neural encoding by point process adaptive filtering," *Neural comp.*, vol. 16, no. 5, pp. 971–998, 2004.

[15] D. L. Donoho, "Compressed sensing," *Information Theory, IEEE Trans. on*, vol. 52, no. 4, pp. 1289–1306, 2006.

[16] E. J. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *Information Theory, IEEE Transactions on*, vol. 52, no. 2, pp. 489–509, 2006.

[17] E. J. Candès *et al.*, "Compressive sampling," in *Proceedings of the International Congress of Mathematicians*, vol. 3. Madrid, Spain, 2006, pp. 1433–1452.

[18] E. J. Candès and M. B. Wakin, "An introduction to compressive sampling," *Signal Processing Magazine, IEEE*, vol. 25, no. 2, pp. 21–30, 2008.

[19] S. N. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu, "A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers," *Statistical Science*, vol. 27, no. 4, pp. 538–557, 2012.

[20] A. M. Bruckstein, D. L. Donoho, and M. Elad, "From sparse solutions of systems of equations to sparse modeling of signals and images," *SIAM review*, vol. 51, no. 1, pp. 34–81, 2009.

[21] B. Babadi, N. Kalouptsidis, and V. Tarokh, "SPARLS: The sparse RLS algorithm," *Signal Processing, IEEE Trans. on*, vol. 58, no. 8, pp. 4013–4025, 2010.

[22] N. Kalouptsidis, G. Mileounis, B. Babadi, and V. Tarokh, "Adaptive algorithms for sparse system identification," *Signal Processing*, vol. 91, no. 8, pp. 1910–1919, 2011.

[23] B. Dumitrescu, A. Onose, P. Helin, and I. Tăbuş, "Greedy sparse RLS," *Signal Processing, IEEE Trans. on*, vol. 60, no. 5, pp. 2194–2207, 2012.

[24] Y. C. Pati, R. Rezaiifar, and P. Krishnaprasad, "Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition," in *Proc. 27th Annu. Asilomar Conf. Signals, Systems and Computers*. IEEE, 1993, pp. 40–44.

[25] W. Dai and O. Milenkovic, "Subspace pursuit for compressive sensing signal reconstruction," *Information Theory, IEEE Trans. on*, vol. 55, no. 5, pp. 2230–2249, 2009.

[26] D. Needell and J. A. Tropp, "CoSaMP: Iterative signal recovery from incomplete and inaccurate samples," *Applied and Computational Harmonic Analysis*, vol. 26, no. 3, pp. 301–321, 2009.

[27] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of Roy. Stat. Soc. B.*, pp. 267–288, 1996.

[28] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM review*, vol. 43, no. 1, pp. 129–159, 2001.

[29] G. Mileounis, B. Babadi, N. Kalouptsidis, and V. Tarokh, "An adaptive greedy algorithm with application to nonlinear communications," *Signal Proc., IEEE Trans. on*, vol. 58, no. 6, pp. 2998–3007, 2010.

[30] J. Fritz, S. Shamma, M. Elhilali, and D. Klein, "Rapid task-related plasticity of spectrotemporal receptive fields in primary auditory cortex," *Nature neuroscience*, vol. 6, no. 11, pp. 1216–1223, 2003.

[31] A. Sheikhattar, J. B. Fritz, S. A. Shamma, and B. Babadi, "Adaptive sparse logistic regression with application to neuronal plasticity analysis," in *2015 49th Asilomar Conference on Signals, Systems and Computers*. IEEE, 2015, pp. 1551–1555.

[32] A. Sheikhattar and B. Babadi, "Real-time algorithms for sparse neuronal system identification," in *Engineering in Medicine and Biology Society (EMBC), 2016 IEEE 38th Annual International Conference of the*. IEEE, 2016, pp. 3410–3413.

[33] A. Sheikhattar, J. B. Fritz, S. A. Shamma, and B. Babadi, "Recursive sparse point process regression with application to spectrotemporal receptive field plasticity analysis," *IEEE Trans. on Signal Processing, in press. preprint: arXiv preprint arXiv:1507.04727*, 2016.

[34] B. A. Olshausen and D. J. Field, "Emergence of simple-cell receptive field properties by learning a sparse code for natural images," *Nature*, vol. 381, no. 6583, pp. 607–609, 1996.

[35] ——, "Sparse coding of sensory inputs," *Current opinion in neurobiology*, vol. 14, no. 4, pp. 481–487, 2004.

[36] O. Sporns and J. D. Zwi, "The small world of the cerebral cortex," *Neuroinformatics*, vol. 2, no. 2, pp. 145–162, 2004.

[37] S. Song, P. J. Sjöström, M. Reigl, S. Nelson, and D. B. Chklovskii, "Highly nonrandom features of synaptic connectivity in local cortical circuits," *PLoS biology*, vol. 3, no. 3, p. e68, 2005.

[38] M. Rehn and F. T. Sommer, "A network that uses few active neurons to code visual input predicts the diverse shapes of cortical receptive fields," *Journal of computational neuroscience*, vol. 22, no. 2, pp. 135–146, 2007.

[39] S. Druckmann, T. Hu, and D. B. Chklovskii, "A mechanistic model of early sensory processing based on subtracting sparse representations," in *Advances in Neural Information Processing Systems*, 2012, pp. 1988–1996.

[40] S. Ganguli and H. Sompolinsky, "Compressed sensing, sparsity, and dimensionality in neuronal information processing and data analysis," *Annual review of neuroscience*, vol. 35, pp. 485–508, 2012.

[41] B. Babadi and H. Sompolinsky, "Sparseness and expansion in sensory representations," *Neuron*, vol. 83, no. 5, pp. 1213–1226, 2014.

[42] M. D. Greicius, B. Krasnow, A. L. Reiss, and V. Menon, "Functional connectivity in the resting brain: a network analysis of the default mode hypothesis," *Proceedings of the National Academy of Sciences*, vol. 100, no. 1, pp. 253–258, 2003.

[43] J. Damoiseaux, S. Rombouts, F. Barkhof, P. Scheltens, C. Stam, S. M. Smith, and C. Beckmann, "Consistent resting-state networks across healthy subjects," *Proceedings of the national academy of sciences*, vol. 103, no. 37, pp. 13 848–13 853, 2006.

[44] P. Hagmann, L. Cammoun, X. Gigandet, R. Meuli, C. J. Honey, V. J. Wedeen, and O. Sporns, "Mapping the structural core of human cerebral cortex," *PLoS biology*, vol. 6, no. 7, p. e159, 2008.

[45] D. H. Perkel, G. L. Gerstein, and G. P. Moore, "Neuronal spike trains and stochastic point processes: Ii. simultaneous spike trains," *Biophysical journal*, vol. 7, no. 4, p. 419, 1967.

[46] G. L. Gerstein and D. H. Perkel, "Simultaneously recorded trains of action potentials: analysis and functional interpretation," *Science*, vol. 164, no. 3881, pp. 828–830, 1969.

[47] C. D. Brody, "Correlations without synchrony," *Neural computation*, vol. 11, no. 7, pp. 1537–1551, 1999.

[48] A. M. Aertsen and G. L. Gerstein, "Evaluation of neuronal connectivity: sensitivity of cross-correlation," *Brain research*, vol. 340, no. 2, pp. 341–354, 1985.

[49] C. Bernasconi and P. KoÈnig, "On the directionality of cortical interactions studied by structural analysis of electrophysiological recordings," *Biological cybernetics*, vol. 81, no. 3, pp. 199–210, 1999.

[50] M. Kamiński, M. Ding, W. A. Truccolo, and S. L. Bressler, "Evaluating causal relations in neural systems: Granger causality, directed transfer function and statistical assessment of significance," *Biological cybernetics*, vol. 85, no. 2, pp. 145–157, 2001.

[51] R. Goebel, A. Roebroeck, D.-S. Kim, and E. Formisano, "Investigating directed cortical interactions in time-resolved fmri data using vector autoregressive modeling and granger causality mapping," *Magnetic resonance imaging*, vol. 21, no. 10, pp. 1251–1261, 2003.

[52] A. Brovelli, M. Ding, A. Ledberg, Y. Chen, R. Nakamura, and S. L. Bressler, "Beta oscillations in a large-scale sensorimotor cortical network: directional influences revealed by granger causality," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 26, pp. 9849–9854, 2004.

[53] N. Wiener, "The theory of prediction," *Modern mathematics for engineers*, vol. 1, pp. 125–139, 1956.

[54] C. W. Granger, "Investigating causal relations by econometric models and cross-spectral methods," *Econometrica: Journal of the Econometric Society*, pp. 424–438, 1969.

[55] J. F. Geweke, "Measures of conditional linear dependence and feedback between time series," *Journal of the American Statistical Association*, vol. 79, no. 388, pp. 907–915, 1984.

[56] J. Geweke, "Measurement of linear dependence and feedback between multiple time series," *Journal of the American statistical association*, vol. 77, no. 378, pp. 304–313, 1982.

[57] M. Kaminski and K. J. Blinowska, "A new method of the description of the information flow in the brain structures," *Biological cybernetics*, vol. 65, no. 3, pp. 203–210, 1991.

[58] L. A. Baccalá and K. Sameshima, "Partial directed coherence: a new concept in neural structure determination," *Biological cybernetics*, vol. 84, no. 6, pp. 463–474, 2001.

[59] L. Sommerlade, M. Thiel, B. Platt, A. Plano, G. Riedel, C. Grebogi, J. Timmer, and B. Schelter, "Inference of granger causal time-dependent influences

in noisy multivariate time series," *Journal of neuroscience methods*, vol. 203, no. 1, pp. 173–185, 2012.

[60] T. Milde, L. Leistritz, L. Astolfi, W. H. Miltner, T. Weiss, F. Babiloni, and H. Witte, "A new kalman filter approach for the estimation of high-dimensional time-variant multivariate ar models and its application in analysis of laser-evoked brain potentials," *Neuroimage*, vol. 50, no. 3, pp. 960–969, 2010.

[61] M. Havlicek, J. Jan, M. Brazdil, and V. D. Calhoun, "Dynamic granger causality based on kalman filter for evaluation of functional network connectivity in fmri data," *Neuroimage*, vol. 53, no. 1, pp. 65–77, 2010.

[62] E. Möller, B. Schack, M. Arnold, and H. Witte, "Instantaneous multivariate eeg coherence analysis by means of adaptive high-dimensional autoregressive models," *Journal of neuroscience methods*, vol. 105, no. 2, pp. 143–158, 2001.

[63] W. Hesse, E. Möller, M. Arnold, and B. Schack, "The use of time-variant eeg granger causality for inspecting directed interdependencies of neural assemblies," *Journal of neuroscience methods*, vol. 124, no. 1, pp. 27–44, 2003.

[64] L. Astolfi, F. Cincotti, D. Mattia, F. D. V. Fallani, A. Tocci, A. Colosimo, S. Salinari, M. G. Marciani, W. Hesse, H. Witte *et al.*, "Tracking the time-varying cortical connectivity patterns by adaptive multivariate estimators," *IEEE Transactions on Biomedical Engineering*, vol. 55, no. 3, pp. 902–913, 2008.

[65] J. R. Sato, E. A. Junior, D. Y. Takahashi, M. de Maria Felix, M. J. Brammer, and P. A. Morettin, "A method to produce evolving functional connectivity maps during the course of an fmri experiment using wavelet-based time-varying granger causality," *Neuroimage*, vol. 31, no. 1, pp. 187–196, 2006.

[66] P. A. Valdés-Sosa, J. M. Sánchez-Bornot, A. Lage-Castellanos, M. Vega-Hernández, J. Bosch-Bayard, L. Melie-García, and E. Canales-Rodríguez, "Estimating brain functional connectivity with sparse multivariate autoregression," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 360, no. 1457, pp. 969–981, 2005.

[67] I. H. Stevenson, J. M. Rebesco, N. G. Hatsopoulos, Z. Haga, L. E. Miller, and K. P. Kording, "Bayesian inference of functional connectivity and network structure from spikes," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 17, no. 3, pp. 203–213, 2009.

[68] Z. Zhou, M. Ding, Y. Chen, P. Wright, Z. Lu, and Y. Liu, "Detecting directional influence in fmri connectivity analysis using pca based granger causality," *Brain research*, vol. 1289, pp. 22–29, 2009.

[69] M. Dhamala, G. Rangarajan, and M. Ding, "Analyzing information flow in brain networks with nonparametric granger causality," *Neuroimage*, vol. 41, no. 2, pp. 354–362, 2008.

[70] K. Sameshima and L. A. Baccalá, "Using partial directed coherence to describe neuronal ensemble interactions," *Journal of neuroscience methods*, vol. 94, no. 1, pp. 93–103, 1999.

[71] M. Krumin and S. Shoham, "Multivariate autoregressive modeling and granger causality analysis of multiple spike trains," *Computational intelligence and neuroscience*, vol. 2010, p. 10, 2010.

[72] M. Okatan, M. A. Wilson, and E. N. Brown, "Analyzing functional connectivity using a network likelihood model of ensemble neural spiking activity," *Neural computation*, vol. 17, no. 9, pp. 1927–1961, 2005.

[73] A. G. Nedungadi, G. Rangarajan, N. Jain, and M. Ding, "Analyzing multiple spike trains with nonparametric granger causality," *Journal of computational neuroscience*, vol. 27, no. 1, pp. 55–64, 2009.

[74] S. Kim, D. Putrino, S. Ghosh, and E. N. Brown, "A granger causality measure for point process models of ensemble neural spiking activity," *PLoS Comput Biol*, vol. 7, no. 3, p. e1001110, 2011.

[75] C. J. Quinn, T. P. Coleman, N. Kiyavash, and N. G. Hatsopoulos, "Estimating the directed information to infer causal relationships in ensemble neural spike train recordings," *Journal of computational neuroscience*, vol. 30, no. 1, pp. 17–44, 2011.

[76] S. Kim, C. J. Quinn, N. Kiyavash, and T. P. Coleman, "Dynamic and succinct statistical analysis of neuroscience data," *Proceedings of the IEEE*, vol. 102, no. 5, pp. 683–698, 2014.

[77] A. Sheikhattar and B. Babadi, "Dynamic estimation of causal influences in sparsely-interacting neuronal ensembles," in *Information Science and Systems (CISS), 2016 Annual Conference on.* IEEE, 2016, pp. 551–556.

[78] A. Sheikhattar, S. Miran, J. B. Fritz, S. A. Shamma, and B. Babadi, "Probing the functional circuitry underlying auditory attention via dynamic granger causality analysis," in *Signals, Systems and Computers, 2016 50th Asilomar Conference on.* IEEE, 2016, pp. 593–597.

[79] A. Sheikhattar, S. Miran, J. Liu, J. B. Fritz, S. A. Shamma, P. O. Kanold, and B. Babadi, "Extracting neuronal functional network dynamics via adaptive granger causality analysis," *Proceedings of the National Academy of Sciences*, vol. 115, no. 17, pp. E3869–E3878, 2018.

[80] N. A. Francis, D. E. Winkowski, A. Sheikhattar, K. Armengol, B. Babadi, and P. O. Kanold, "Small networks encode decision-making in primary auditory cortex," *Neuron*, vol. 97, no. 4, pp. 885–897, 2018.

[81] W. Truccolo, U. T. Eden, M. R. Fellows, J. P. Donoghue, and E. N. Brown, "A point process framework for relating neural spiking activity to spiking history, neural ensemble, and extrinsic covariate effects," *Journal of neurophysiology*, vol. 93, no. 2, pp. 1074–1089, 2005.

[82] Z. Chen, D. F. Putrino, S. Ghosh, R. Barbieri, and E. N. Brown, "Statistical inference for assessing functional connectivity of neuronal ensembles with sparse spiking data," *Neural Systems and Rehabilitation Engineering, IEEE Trans. on*, vol. 19, no. 2, pp. 121–135, 2011.

[83] E. J. Candès, Y. Plan *et al.*, "Near-ideal model selection by ℓ_1 minimization," *The Annals of Statistics*, vol. 37, no. 5A, pp. 2145–2177, 2009.

[84] E. J. Candes and M. A. Davenport, "How well can we estimate a sparse vector?" *Applied and Computational Harmonic Analysis*, vol. 34, no. 2, pp. 317–323, 2013.

[85] J. Haupt, W. U. Bajwa, G. Raz, and R. Nowak, "Toeplitz compressed sensing matrices with applications to sparse channel estimation," *Information Theory, IEEE Trans. on*, vol. 56, no. 11, pp. 5862–5875, 2010.

[86] A. Javanmard and A. Montanari, "Confidence intervals and hypothesis testing for high-dimensional regression," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 2869–2909, 2014.

[87] S. Van de Geer, P. Bühlmann, Y. Ritov, and R. Dezeure, "On asymptotically optimal confidence regions and tests for high-dimensional models," *The Annals of Stat.*, vol. 42, no. 3, pp. 1166–1202, 2014.

[88] C.-H. Zhang and S. S. Zhang, "Confidence intervals for low dimensional parameters in high dimensional linear models," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 76, no. 1, pp. 217–242, 2014.

[89] N. Meinshausen and P. Bühlmann, "High-dimensional graphs and variable selection with the lasso," *The Annals of Stat.*, pp. 1436–1462, 2006.

[90] J. Fan and R. Li, "Variable selection via nonconcave penalized likelihood and its oracle properties," *Journal of American stat. Assoc.*, vol. 96, no. 456, pp. 1348–1360, 2001.

[91] E. N. Brown, R. Barbieri, V. Ventura, R. E. Kass, and L. M. Frank, "The time-rescaling theorem and its application to neural spike train data analysis," *Neural comp.*, vol. 14, no. 2, pp. 325–346, 2002.

[92] R. Haslinger, G. Pipa, and E. Brown, "Discrete time rescaling theorem: determining goodness of fit for discrete time statistical models of neural spiking," *Neural comp.*, vol. 22, no. 10, pp. 2477–2506, 2010.

[93] J. Fritz, M. Elhilali, and S. Shamma, "Active listening: task-dependent plasticity of spectrotemporal receptive fields in primary auditory cortex," *Hearing research*, vol. 206, no. 1, pp. 159–176, 2005.

[94] N. Mesgarani, J. Fritz, and S. Shamma, "A computational model of rapid task-related plasticity of auditory cortical receptive fields," *Journal of computational neuroscience*, vol. 28, no. 1, pp. 19–27, 2010.

[95] S. S. Wilks, "The large-sample distribution of the likelihood ratio for testing composite hypotheses," *The Annals of Mathematical Statistics*, vol. 9, no. 1, pp. 60–62, 1938.

[96] A. Wald, "Tests of statistical hypotheses concerning several parameters when the number of observations is large," *Transactions of the American Mathematical society*, vol. 54, no. 3, pp. 426–482, 1943.

[97] R. R. Davidson and W. E. Lever, "The limiting distribution of the likelihood ratio statistic under a class of local alternatives," *Sankhyā: The Indian Journal of Statistics, Series A*, pp. 209–224, 1970.

[98] H. Peers, "Likelihood ratio and associated test criteria," *Biometrika*, vol. 58, no. 3, pp. 577–587, 1971.

[99] C. Bonferroni, "Teoria statistica delle classi e calcolo delle probabilita," *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commericiali di Firenze*, vol. 8, pp. 3–62, 1936.

[100] Y. Hochberg, "A sharper bonferroni procedure for multiple tests of significance," *Biometrika*, vol. 75, no. 4, pp. 800–802, 1988.

[101] Y. Benjamini and D. Yekutieli, "The control of the false discovery rate in multiple testing under dependency," *Annals of statistics*, pp. 1165–1188, 2001.

[102] D. Amos, "Computation of modified bessel functions and their ratios," *Mathematics of Computation*, vol. 28, no. 125, pp. 239–251, 1974.

[103] Á. Baricz, "Bounds for turánians of modified bessel functions," *Expositiones Mathematicae*, vol. 33, no. 2, pp. 223–251, 2015.

[104] R. H. Shumway and D. S. Stoffer, "An approach to time series smoothing and forecasting using the em algorithm," *Journal of time series analysis*, vol. 3, no. 4, pp. 253–264, 1982.

[105] A. Kazemipour, M. Wu, and B. Babadi, "Robust estimation of self-exciting point process models with application to neuronal modeling," *arXiv preprint arXiv:1507.03955*, 2015.

[106] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the royal statistical society. Series B (methodological)*, pp. 1–38, 1977.

[107] N. Parikh, S. Boyd *et al.*, "Proximal algorithms," *Foundations and Trends® in Optimization*, vol. 1, no. 3, pp. 127–239, 2014.

[108] S. L. Bressler and A. K. Seth, "Wiener–granger causality: a well established methodology," *Neuroimage*, vol. 58, no. 2, pp. 323–329, 2011.

[109] S. A. Glantz, *Primer of biostatistics*. McGraw Hill Professional, 2012.

[110] S. Guo, A. K. Seth, K. M. Kendrick, C. Zhou, and J. Feng, "Partial granger causality?eliminating exogenous inputs and latent variables," *Journal of neuroscience methods*, vol. 172, no. 1, pp. 79–93, 2008.

[111] E. A. Pnevmatikakis, D. Soudry, Y. Gao, T. A. Machado, J. Merel, D. Pfau, T. Reardon, Y. Mu, C. Lacefield, W. Yang *et al.*, "Simultaneous denoising, deconvolution, and demixing of calcium imaging data," *Neuron*, vol. 89, no. 2, pp. 285–299, 2016.

[112] T. Hromádka and A. M. Zador, "Representations in auditory cortex," *Current opinion in neurobiology*, vol. 19, no. 4, pp. 430–433, 2009.

[113] P. V. Watkins, J. P. Kao, and P. O. Kanold, "Spatial pattern of intra-laminar connectivity in supragranular mouse auditory cortex," *Frontiers in neural circuits*, vol. 8, p. 15, 2014.

[114] E. K. Miller and J. D. Cohen, "An integrative theory of prefrontal cortex function," *Annual review of neuroscience*, vol. 24, no. 1, pp. 167–202, 2001.

[115] J. I. Gold and M. N. Shadlen, "The neural basis of decision making," *Annu. Rev. Neurosci.*, vol. 30, pp. 535–574, 2007.

[116] T. J. Buschman and E. K. Miller, "Top-down versus bottom-up control of attention in the prefrontal and posterior parietal cortices," *science*, vol. 315, no. 5820, pp. 1860–1862, 2007.

[117] J. B. Fritz, S. V. David, S. Radtke-Schuller, P. Yin, and S. A. Shamma, "Adaptive, behaviorally gated, persistent encoding of task-relevant auditory information in ferret frontal cortex," *Nature neuroscience*, vol. 13, no. 8, pp. 1011–1019, 2010.

[118] C. D. Gilbert and W. Li, "Top-down influences on visual processing," *Nature Reviews Neuroscience*, vol. 14, no. 5, p. 350, 2013.

[119] V. Piëch, W. Li, G. N. Reeke, and C. D. Gilbert, "Network model of top-down influences on local gain and contextual interactions in visual cortex," *Proceedings of the National Academy of Sciences*, vol. 110, no. 43, pp. E4108–E4117, 2013.

[120] D. J. Klein, J. Z. Simon, D. A. Depireux, and S. A. Shamma, "Stimulus-invariant processing and spectrotemporal reverse correlation in primary auditory cortex," *Journal of computational neuroscience*, vol. 20, no. 2, pp. 111–136, 2006.

[121] S. Bandyopadhyay, S. A. Shamma, and P. O. Kanold, "Dichotomy of functional organization in the mouse auditory cortex," *Nature neuroscience*, vol. 13, no. 3, p. 361, 2010.

[122] G. Rothschild, I. Nelken, and A. Mizrahi, "Functional organization and population dynamics in the mouse primary auditory cortex," *Nature neuroscience*, vol. 13, no. 3, p. 353, 2010.

[123] X. Meng, D. E. Winkowski, J. P. Kao, and P. O. Kanold, "Sublaminar subdivision of mouse auditory cortex layer 2/3 based on functional translaminar connections," *Journal of Neuroscience*, vol. 37, no. 42, pp. 10 200–10 214, 2017.

[124] H. V. Oviedo, I. Bureau, K. Svoboda, and A. M. Zador, "The functional asymmetry of auditory cortex is reflected in the organization of local cortical circuits," *Nature neuroscience*, vol. 13, no. 11, p. 1413, 2010.

[125] M. M. Churchland, J. P. Cunningham, M. T. Kaufman, J. D. Foster, P. Nuyujukian, S. I. Ryu, and K. V. Shenoy, "Neural population dynamics during reaching," *Nature*, vol. 487, no. 7405, p. 51, 2012.

[126] M. B. Ahrens, J. M. Li, M. B. Orger, D. N. Robson, A. F. Schier, F. Engert, and R. Portugues, "Brain-wide neuronal dynamics during motor adaptation in zebrafish," *Nature*, vol. 485, no. 7399, pp. 471–477, 2012.

[127] G. Katona, G. Szalay, P. Maák, A. Kaszás, M. Veress, D. Hillier, B. Chiovini, E. S. Vizi, B. Roska, and B. Rózsa, "Fast two-photon in vivo imaging with three-dimensional random-access scanning in large tissue volumes," *Nature methods*, vol. 9, no. 2, p. 201, 2012.

[128] M. B. Ahrens, M. B. Orger, D. N. Robson, J. M. Li, and P. J. Keller, "Whole-brain functional imaging at cellular resolution using light-sheet microscopy," *Nature methods*, vol. 10, no. 5, pp. 413–420, 2013.

[129] T. W. Dunn, Y. Mu, S. Narayan, O. Randlett, E. A. Naumann, C.-T. Yang, A. F. Schier, J. Freeman, F. Engert, and M. B. Ahrens, "Brain-wide mapping of neural activity controlling zebrafish exploratory locomotion," *eLife*, vol. 5, p. e12741, 2016.

[130] W. Kristan and R. L. Calabrese, "Rhythmic swimming activity in neurones of the isolated nerve cord of the leech," *Journal of Experimental Biology*, vol. 65, no. 3, pp. 643–668, 1976.

[131] W. Singer, "Synchronization of cortical activity and its putative role in information processing and learning," *Annual review of physiology*, vol. 55, no. 1, pp. 349–374, 1993.

[132] A. Schnitzler and J. Gross, "Normal and pathological oscillatory communication in the brain," *Nature reviews neuroscience*, vol. 6, no. 4, p. 285, 2005.

[133] P. Fries, "A mechanism for cognitive dynamics: neuronal communication through neuronal coherence," *Trends in cognitive sciences*, vol. 9, no. 10, pp. 474–480, 2005.

[134] A. K. Engel, P. Fries, and W. Singer, "Dynamic predictions: oscillations and synchrony in top–down processing," *Nature Reviews Neuroscience*, vol. 2, no. 10, p. 704, 2001.

[135] R. W. Friedrich, C. J. Habermann, and G. Laurent, "Multiplexing using synchrony in the zebrafish olfactory bulb," *Nature neuroscience*, vol. 7, no. 8, p. 862, 2004.

[136] B. Babadi and E. N. Brown, "A review of multitaper spectral analysis," *IEEE Transactions on Biomedical Engineering*, vol. 61, no. 5, pp. 1555–1564, 2014.

[137] D. J. Thomson, "Spectrum estimation and harmonic analysis," *Proceedings of the IEEE*, vol. 70, no. 9, pp. 1055–1096, 1982.

[138] D. B. Percival and A. T. Walden, *Wavelet methods for time series analysis*. Cambridge university press, 2006, vol. 4.

[139] F. Gandolfo, C.-S. Li, B. Benda, C. P. Schioppa, and E. Bizzi, "Cortical correlates of learning in monkeys adapting to a new dynamical environment," *Proceedings of the National Academy of Sciences*, vol. 97, no. 5, pp. 2259–2263, 2000.

[140] S. F. Cooke and M. F. Bear, "Visual experience induces long-term potentiation in the primary visual cortex," *Journal of Neuroscience*, vol. 30, no. 48, pp. 16 304–16 313, 2010.

[141] R. F. Barber, E. J. Candès *et al.*, "Controlling the false discovery rate via knockoffs," *The Annals of Statistics*, vol. 43, no. 5, pp. 2055–2085, 2015.

[142] S. A. van de Geer, "On Hoeffding's inequality for dependent random variables," in *Empirical Process Techniques for Dependent Data*, H. Dehling and W. Philipp, Eds. Springer, 2001.

[143] P. J. Bickel, Y. Ritov, and A. B. Tsybakov, "Simultaneous analysis of Lasso and Dantzig selector," *The Annals of Stat.*, pp. 1705–1732, 2009.

[144] W. Hoeffding, "Probability inequalities for sums of bounded random variables," *Journal of the American statistical association*, vol. 58, no. 301, pp. 13–30, 1963.

[145] B. Scholkopf and A. J. Smola, *Learning with kernels: Support vector machines, regularization, optimization, and beyond*. MIT press, 2001.

[146] M. A. Figueiredo and R. D. Nowak, "A bound optimization approach to wavelet-based image deconvolution," in *Image Processing, 2005. ICIP 2005. IEEE International Conf. on*, vol. 2. IEEE, 2005, pp. II–782.

[147] M. A. Figueiredo, J. M. Bioucas-Dias, and R. D. Nowak, "Majorization–minimization algorithms for wavelet-based image restoration," *Image Processing, IEEE Trans. on*, vol. 16, no. 12, pp. 2980–2991, 2007.

[148] D. R. Hunter and K. Lange, "A tutorial on MM algorithms," *The American Statistician*, vol. 58, no. 1, pp. 30–37, 2004.

[149] M. A. Figueiredo and R. D. Nowak, "An EM algorithm for wavelet-based image restoration," *Image Processing, IEEE Trans. on*, vol. 12, no. 8, pp. 906–916, 2003.

[150] I. Daubechies, M. Defrise, and C. De Mol, "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint," *Communications on Pure and Applied Mathematics*, vol. 57, no. 11, pp. 1413–1457, 2004.

[151] M. J. Crowder, "Maximum likelihood estimation for dependent observations," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 45–53, 1976.

[152] P. Billingsley, *Probability and measure*. John Wiley & Sons, 2008.

[153] P. J. Bickel, Y. Ritov, and T. Ryden, "Asymptotic normality of the maximum-likelihood estimator for general hidden markov models," *Annals of Statistics*, pp. 1614–1635, 1998.

[154] R. Douc, E. Moulines, T. Ryden *et al.*, "Asymptotic properties of the maximum likelihood estimator in autoregressive models with markov regime," *The Annals of statistics*, vol. 32, no. 5, pp. 2254–2304, 2004.

[155] W. Tan, "On the distribution of quadratic forms in normal random variables," *Canadian Journal of Statistics*, vol. 5, no. 2, pp. 241–250, 1977.

[156] J. S. Chipman and M. Rao, "Projections, generalized inverses, and quadratic forms," *Journal of Mathematical Analysis and Applications*, vol. 9, no. 1, pp. 1–11, 1964.