UNIVERSITY OF CALIFORNIA,
IRVINE

# Analysis of Time Series with Applications to Electrophysiological Signals

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Statistics

by

Xu Gao

Dissertation Committee:
Professor Hernando Ombao, Chair
Professor Zhaoxia Yu
Professor Weining Shen
Professor Beth Lopour

2018

# DEDICATION

TO ERPA.

# TABLE OF CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

# LIST OF ALGORITHMS

# ACKNOWLEDGEMENTS

First and foremost, I, along with ERPA, deeply thank my thesis advisor, Dr. Hernando Ombao, for his encouragement, support and guidance. He is the "recruiter" that helps us get aboard the statistics family at UCI. He is the one that leads me to the fields of time series and brain imaging. He is always the one that raises me up and tells me never to give up. Some texts of this dissertation are reprints of the publications listed in the curriculum vitae. He is the one that builds the foundation of the dissertation. I would like to thank my committee members, Dr. Weining Shen, Zhaoxia Yu and Beth Lopour, for their patience, expertise and kindness. I am also grateful to Dr. Babak Shahbaba, Daniel Gillen and other faculties in the department. I would also like to thank my parents for their unlimited supports and caring. Finally, I am deeply grateful to ERPA, without whom I could never go anywhere.

# CURRICULUM VITAE

## Xu Gao

### EDUCATION

**Doctor of Philosophy in Statistics** **2018**
University of California, Irvine *Irvine, California*

**Master of Science in Mathematics** **2012**
University of Illinois at Chicago *Chicago, Illinois*

**Bachelor of Science in Mathematics** **2011**
University of Science and Technology of China *China*

### RESEARCH EXPERIENCE

**Graduate Research Assistant** **2014–2018**
University of California, Irvine *Irvine, California*

### TEACHING EXPERIENCE

**Teaching Assistant** **2014–2017**
University of California, Irvine *Irvine, California*

**Teaching Assistant** **2011–2014**
University of Illinois at Chicago *Chicago, Illinois*

### ACADEMIC REVIEWING

**NeuroImage** **2018**

**Biometrics** **2017**

### PUBLICATIONS

**Xu Gao**, Weining Shen, Babak Shahbaba, Norbert Fortin, Hernando Ombao. Evolutionary State-Space Model and Its Application to Time-Frequency Analysis of Local Field Potentials. *Under revision of Statistica Sinica*, 2018.

**Xu Gao**, Hernando Ombao, Babak Shahbaba. Modeling Binary Time Series using Gaussian Processes with Application to Predicting Sleep States. *Journal of Classification*, 2018.

**Xu Gao**, Hernando Ombao, Daniel Gillen. Fisher Information Matrix of Binary Time Series. *In submission*, 2017.

Peng Chen, Feiyang Sun, Zhenbo Wang, **Xu Gao**, Junfeng Jiao, Zhimin Tao. Built environment effects on bike crash frequency and risk in Beijing. *Journal of Safety Research*, 2017.

Peng Chen, Yuanjie Tu, **Xu Gao**, Jiawen Yang, Yang Tang. Examining Dockless Public Bike Usage (MoBike): A Generalized Additive Mixed Modelling Approach. *In submission*, 2017.

Qianshun Cheng, **Xu Gao**, Ryan Martin. Exact prior-free probabilistic inference on the heritability coefficient in a linear mixed model. *Electronic Journal of Statistics*, 2014.

Yin Zhao, **Xu Gao**, Daizhan Cheng. Semi-Tensor Product Approach to Boolean Functions. *Journal of the Graduate School of the Chinese Academy of Sciences*, 2012.

Wenjie Gao, **Xu Gao**. Research on the Comprehensive Evaluation Model Based on SEM for the Level of Modernization of Major Cities in Our Country. *Mathematics in practice and theory*, 2010.

Wenjie Gao, **Xu Gao**. A study on the analytic hierarchy process based on the gray scale (gahp). *Science and Technology Innovation Herald*, 2009.

# ABSTRACT OF THE DISSERTATION

Analysis of Time Series with Applications to Electrophysiological Signals

By

Xu Gao

Doctor of Philosophy in Statistics

University of California, Irvine, 2018

Professor Hernando Ombao, Chair

Time series analysis is widely discussed in fields such as finance, economy, brain imaging etc. Among all types of data, categorical and multivariate time series maintain both of challenges and promising applications. In this dissertation, we propose some statistical approaches to model binary and multivariate time series and thus provide alternative solutions of statistical inference and prediction.

We first focus on binary time series. Classical methods do not differentiate between exogenous and endogenous exploratory variable, which leads to invalid statistical inference. We develop a close form of the Fisher information matrix of logistic autoregressive model and demonstrate that it yields narrower confidence intervals while maintaining nominal type I error rate. We also propose a framework of predicting binary time series using Gaussian process. The approach comprises of a linear part that captures the effects from covariates and a stochastic process that characterizes the information not covered by the linear part. Both the simulation and the real data examples demonstrate the high predictive power and appropriate interpretability.

Next, we discuss on the problems of multivariate time series. In an illustrative example of

analyzing Local Field Potentials (LFPs) signals, existing methods such as Independent Component Analysis (ICA), Principal Component Analysis (PCA) have limitations in modeling spatial-temporal dependencies across trials (epochs). To address these issues, we introduce Evolutionary State Space Model (E-SSM) allowing the latent signals evolve during the experiment. By fixing the phase of the AR polynomial roots, the framework is able to model the evolution for pre-specified frequency bands. As the last part of this dissertation, we characterize multivariate time series as 2 - dimensional tensors. By introducing a penalized mixture matrix normal model, we are able to uncover the "latent" mean spatial-temporal structures across trials (epochs) and capture the sparsity simultaneously. Some theoretical results are established to show the consistency of the constrained maximum likelihood estimator.

# Chapter 1

# Introduction

Time series analysis serves as an important component of modern statistical analysis and has been widely developed during the past couple of decades (Brillinger, 1975; Brockwell and Davis, 1991; Fuller, 2009; Shumway and Stoffer, 2013). Among all the types of time series, categorical and multivariate valued data provide a large amounts of applications in the fields of finance, astronomy, electroencephalography etc. Throughout this dissertation, we shed light on some of the interesting topics on the inference and prediction of categorical and multivariate time series. By proposing some novel statistical frameworks, we provide competitive solutions against the drawbacks from the existing literature and could possibly benefit the communities of statistics and neuroscience. In this chapter, we present a brief overview of the problems covered in this dissertation.

## 1.1 Categorical Time Series

Categorical time series data is widely collected in many fields. Various models for categorical time series that take into account temporal correlation are discussed in Kedem and Yakowitz (1994), Kedem and Slud (1980), Diggle (2002) and Fahrmeir and Tutz (2013), among others. Keenan (1982a) developed a model with an underlying unobserved process that is Gaussian first-order autoregressive. For binary time series with a Markovian structure, Billingsley (1961), Meyn and Tweedie (2012a), Bonney (1987a), Fahrmeir and Kaufmann (1987), Kaufmann (1987), Keenan (1982a) and Muenz and Rubinstein (1985) developed an inferential procedure based on the conditional likelihood. A comprehensive modeling framework based on partial likelihood inference and generalized linear models was developed in Fokianos and Kedem (2003a) and Kedem and Fokianos (2005). In this dissertation, we mainly focus on the inference and prediction of binary time series.

### 1.1.1 Inference on Binary Time Series

In practice, standard software for fitting generalized linear models (GLMs) to binary time series use the past series values as "explanatory variables" in the conditional mean of the response for the regression (de Vries et al., 1998). This approach does not differentiate between explanatory variables that are exogenous to the time series data versus those that are endogenous, i.e., explanatory values that are past values of the time series. Thus, it does not properly take into account the auto-correlation structure in the data, leading to potentially undesirable consequences. In particular, the standard errors of regression parameter estimates that are derived using the Em-FI matrix also ignore the auto-correlation structure. In Fokianos and Kedem (1998a), the asymptotic conditional Fisher information (AFI) matrix was derived for the general case where the conditional distribution of a time

series depends on its own historical data as well as other covariates. While impressive in its generality, the primary limitation of this result is that it does not provide a closed form of the Fisher information matrix for specific models. In this dissertation, we derive the exact Fisher information matrix for a particular binary time series model and provide efficient inference on the parameters. This chapter is summarized in Gao et al. (2017).

### 1.1.2 Prediction of Binary Time Series

Various approaches have been proposed to model and predict binary time series. Caiado et al. (2006) introduced new measurements in classifying time series based on periodograms. Maharaj (2002) put forward a framework of comparing time series in frequency domain. Wavelet based clustering method was also introduced by Maharaj et al. (2010). Jacobs and Lewis (1978) proposed a discrete autoregressive-moving average (DARMA) model by utilizing probabilistic mixtures. A comprehensive modeling framework based on generalized linear models and partial likelihood inference have been developed in Fokianos and Kedem (2002) and Fokianos and Kedem (2003b). Fokianos and Kedem (1998b) extended the partial likelihood analysis to non-stationary categorical time series including stochastic time dependent covariates. With the Markovian structure, Meyn and Tweedie (2012b), Bonney (1987b) and Keenan (1982b) developed inferential procedures based on the conditional likelihood. These previous studies provide inference on binary time series. Their main drawback is that they involve massive computation for high dimensional integrals, which results in poor prediction accuracy. Lindquist and McKeague (2009) introduced a logistic regression model with functional predictors and extended it to generalized linear model. Their substantial work was superior in detecting sensitive and interpretable time points that were most predictive to the response. However, there are some drawbacks are: (1) the Brownian motion assumption is unlikely to be satisfied in practice because the covariates in this study hardly have the

property of increment independence; (2) the influence of covariates on responses is assumed to spread across the entire trajectory and hence implies the non-existence of "sensitive time points"; (3) prediction of the time series is not developed, which could be a serious limitation for this project since we are also interested in such predictions.

From the view of the machine learning community, typical classification methodologies such as decision tree, random forest and strategies such as boosting can also be used for predicting binary time series. Although such approaches are able to achieve predictions with high accuracy, the major drawback is that they give very little guidance of interpretation. In this dissertation, we develop a statistical model that can provide us simultaneously with convincing inference and interpretation at the same time produce prediction accuracy that is higher than that achieved by typical machine learning classification approaches. This chapter is summarized in Gao et al. (2017).

## 1.2 Multivariate Time Series

Multivariate time series has been of increased interest with its widespread applications in various fields such as brain imaging, finance, economy etc. Specifically, in the field of brain imaging, signals are collected from temporal and spatial domains with multiple trials (epochs). Classical univariate time series models fail in investigating the spatial-temporal dependence. This innegligible drawback motivates the development of multivariate time series models. Similar to univariate time series, most of methodologies derive from time and frequency domains. To name a few, time domain methods comprise of Vector Autoregressive (VAR), Vector Autoregressive Moving Average (VARMA), State Space Model, Vector Error Correction Model, Systems of Dynamic Simultaneous Equations etc (Reinsel, 1982; Lütkepohl, 2005). Frequency domain approaches include Dynamic Fourier Analysis and Wavelet, Spectral Ma-

trices Estimation, Factor Analysis, Coherence, Partial Directed Coherence etc (Shumway and Stoffer, 2013; Baccalá and Sameshima, 2001).

In practice, brain imaging signals such as Local Field Potentials (LFPs) are commonly characterized as multivariate time series with mixtures of different underlying brain oscillatory processes and there have been a number of approaches used to estimate these latent independent sources (Whitmore and Lin, 2016; Einevoll et al., 2007; Prado and Lopes, 2013). For example, data-adaptive methods such as independent components analysis (ICA) and principal components analysis (PCA) can provide estimates for the unobserved cortical sources. However, they usually do not take into account the spectral structure within underlying sources that could evolve over the course of the experiment given multiple epochs. Moreover, without any constraint on the structure of the sources, it is extremely difficult to pool information across the epochs in the experiment. Recently, Fiecas and Ombao (2016) studied the dynamics of LFPs during the course of experiment via Cramér representations. Their approach does not consider low-dimensional representations, which are indispensable to modeling these high dimensional multi-electrode LFPs. To overcome the drawbacks, we develop an evolutionary state space model framework in this dissertation. This chapter is summarized in Gao et al. (2016).

From an alternative perspective, spatial-temporal data (LFPs and other brain imaging signals) can be directly characterized as tensor data source. Signals obtained from multiple trials (epochs) can be observed as 3 dimensional tensors. Inspired by the framework of matrix normal (Dawid, 1981) and mixture models (Dutilleul, 1999), we propose a framework of penalized mixture matrix normal to investigate on the spatial-temporal dependency and evolution across various trials (epochs).

## 1.3 Outline and Contributions

The outline and contributions of this dissertation are listed as follows.

In Chapter 2, we derive the exact conditional Fisher information matrix of a general logistic autoregressive model with endogenous covariates for any series length $T$. Moreover, we also develop an iterative computational formula that allows for relatively easy implementation of the proposed estimator. Our simulation studies show that confidence intervals derived using the exact Fisher information matrix tend to be narrower than those utilizing the empirical Fisher information matrix while maintaining type I error rates at or below nominal levels. Further, we establish that the exact Fisher information matrix approaches, as $T$ tends to infinity, the asymptotic Fisher information matrix previously derived for binary time series data. The developed exact conditional Fisher information matrix is applied to time-series data on respiratory rate among a cohort of expectant mothers where it is found to provide narrower confidence intervals for functionals of scientific interest and lead to greater statistical power when compared to the empirical Fisher information matrix.

In Chapter 3, we develop a mixed effects model for binary time series with a stochastic component represented by a Gaussian process. The fixed component captures the effects of covariates on the binary-valued response. The Gaussian process captures the residual variations in the binary response that are not explained by covariates and past realizations. We develop a frequentist modeling framework that provides efficient inference and more accurate predictions. Results demonstrate the advantages of improved prediction rates over existing approaches such as logistic regression, generalized additive mixed model, models for ordinal data, gradient boosting, decision tree and random forest.

In Chapter 4, we propose an evolutionary state space model (E-SSM) for analyzing high

dimensional brain signals whose statistical properties evolve over the course of a non-spatial memory experiment. Under E-SSM, brain signals are modeled as mixtures of components (e.g., AR(2) process) with oscillatory activity at pre-defined frequency bands. To account for the potential non-stationarity of these components (since the brain responses could vary throughout the entire experiment), the parameters are allowed to vary over epochs. Compared with classical approaches such as independent component analysis and filtering, the proposed method accounts for the entire temporal correlation of the components and accommodates non-stationarity. For inference purpose, we propose a novel computational algorithm based upon using Kalman smoother, maximum likelihood and blocked resampling. The E-SSM model is applied to simulation studies and an application to a multi-epoch LFP signal data collected from a non-spatial (olfactory) sequence memory task study. The results confirm that our method captures the evolution of the power for different components across different phases in the experiment and identifies clusters of electrodes that behave similarly with respect to the decomposition of different sources.

In Chapter 5, we introduce a framework of mixture matrix normal to characterize the brain signals of multiple trials (epochs). By adding various regularization terms, the proposed model is able to identify "latent" spatial-temporal structures across hundreds of trials (epochs). We establish some theoretic results showing the consistency of the constraint optimizer. We also apply the proposed approach to two LFPs dataset from different experiments. The results outperform the existing method by producing more interpretable and stable clusters as well as mean "latent" spatial-temporal patterns across trials (epochs).

In Chapter 6, we conclude the findings in this dissertation and list some of the potential future works.

# Chapter 2

# Fisher Information Matrix of Binary Time Series

## 2.1 Introduction

de Vries et al. (1998) utilized a logistic autoregressive model (LAR/LARX) to predict the outcome of supervised exercise for intermittent claudication. The inference did not distinguish between covariates that were exogenous to the time series, versus covariates that were endogenous, yielding potentially invalid and/or inefficient statistical inference. In this chapter, we derive the *exact* conditional Fisher information (Ex-FI) matrix of a logistic autoregressive model for binary time series with arbitrary length $T$. We demonstrate that a correctly specified Ex-FI leads to more efficient inference for regression parameters as measured by typically narrower confidence intervals relative to those obtained using the empirical Fisher information (Em-FI) matrix (Dodge, 2006), while maintaining type I error rates at, or below, nominal levels. This model takes into account the correlation in binary time series.

Fokianos and Kedem (1998b) derived the the asymptotic conditional Fisher information (AFI) matrix where the conditional distribution of a time series depends on its own historical data as well as other covariates. The limitation is that the Ex-FI matrix has not been derived for finite $T$. Instead, only an asymptotic approximation based on the partial likelihood, which turned out to be equivalent to the Em-FI matrix for the LAR model, was provided. There are major consequences of these limitations. First, the result lacks the precise form of the Fisher information matrix to conduct inference on specific LAR coefficients and functionals of these coefficients (e.g., probability of $Y_t = 1$ given the past values of the binary series). Second, when $T$ is not sufficiently large, the discrepancy between the Ex-FI and Em-FI matrices could lead to poor power, incorrect significance level of tests, inefficient inference, and potentially misleading results from data analysis. Third, the large sample theory derived in Kedem and Fokianos (2005) is based on the crucial assumption that $\frac{1}{T}\sum_{t=1}^{T}\mathbb{I}(X_t \in A) \rightarrow \nu(A)$, where $\nu(.)$ is a probability measure, $A$ is a Borel set and $\mathbb{I}(.)$ is the indicator function. Even when $T$ is large, such assumption may not be easily met. In this way, using Em-FI rather than Ex-FI may be misleading since no large sample theory is guaranteed.

Motivated by these limitations, this chapter provides a derivation of the Ex-FI matrix of a LAR/LARX model for arbitrary finite $T$. While the derivation is non-trivial we provide a computationally tractable expression that can be easily implemented in an iterative manner. We report findings from simulation studies suggesting that the derived Ex-FI matrix yields superior results relative to the Em-FI for small to moderate sample sizes. In particular, when compared to using the Em-FI, inference based on the Ex-FI matrix produces narrower confidence intervals for a fixed significance level; close to expected false positive rate and higher power when conducting tests of hypotheses. The simulation studies also demonstrate that the Ex-FI matrix converges to the general AFI developed in Fokianos and Kedem (1998a) in the sense that the norm of the difference between the entries of the two matrices

converges to 0 when the length $T$ of the binary time series increases. Finally, we apply the developed Ex-FI matrix to time-series data on respiratory rate among a cohort of expectant mothers. Results show the similar pattern observed from simulations. Namely, the Ex-FI matrix is found to provide narrower confidence intervals for functionals of scientific interest (such as the probability or log odds) and produce more statistical power when compared to the Em-FI matrix.

The remainder of this chapter is organized as follows. In Section 2.2, we first derive the Ex-FI matrix of LAR/LARX model in general. We also propose a computation framework through functional iteration to obtain the Ex-FI matrix explicitly. At the end, we consider a special case when the order of LAR mode is 1 and calculate the analytic form of the Ex-FI matrix. In Section 2.3, we present some simulation results to compare the Ex-FI with Em-FI. Results show the benefit of using Ex-FI in terms of shorter confidence interval length and reasonable Type I error rate. Moreover, asymptotic behavior is also studied. In Section 2.4, we applied the Ex-FI matrix to time-series data on respiratory rate among expectant mothers. By comparing with the Em-FI, we conclude that using Ex-FI can produce greater power and shorter confidence intervals when conducting statistical inference.

## 2.2 Derivation of the Exact Conditional Fisher Information Matrix

### 2.2.1 Logistic Autoregressive Model of Order $p$ (LAR(p))

Consider a binary-valued correlated time series data $Y_t,\ t = 1, \ldots, T$ where the conditional distribution of $Y_t$ depends on the previous values via the conditional probability

$$P_t = \mathsf{P}(Y_t = 1 \mid y_{t-1}, y_{t-2}, \cdots, y_1) = \frac{\exp(\boldsymbol{y'_{-t}\beta})}{1 + \exp(\boldsymbol{y'_{-t}\beta})}, \tag{2.1}$$

where $\boldsymbol{y_{-t}} = (1, y_{t-1}, \cdots, y_{t-p})'$ is endogenous to the series and $\boldsymbol{\beta} = (\beta_0, \cdots, \beta_p)'$. The conditional log-likelihood function of $\boldsymbol{\beta}$ and the vector of conditional score functions are, respectively,

$$\begin{aligned} \ell(\boldsymbol{\beta} \mid \boldsymbol{Y}) &= \sum_{t=p+1}^{T} [Y_t(\boldsymbol{Y'_{-t}\beta}) - \log\{1 + \exp(\boldsymbol{Y'_{-t}\beta})\}] \\ \boldsymbol{U(\beta, Y)} &= \sum_{t=p+1}^{T} [\boldsymbol{Y_{-t}}\{Y_t - \frac{\exp(\boldsymbol{Y'_{-t}\beta})}{1 + \exp(\boldsymbol{Y'_{-t}\beta})}\}]. \end{aligned}$$

Then, it follows that

$$\frac{\partial^2}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta^T}} \ell(\boldsymbol{\beta} \mid \boldsymbol{Y}) = -\sum_{t=p+1}^{T} \left[ \frac{\exp(\boldsymbol{Y'_{-t}\beta})}{\{1 + \exp(\boldsymbol{Y'_{-t}\beta})\}^2} \boldsymbol{Y_{-t}} \boldsymbol{Y'_{-t}} \right]. \tag{2.2}$$

The Ex-FI matrix takes the form

$$I(\boldsymbol{\beta} \mid y_p, \cdots, y_1) = \sum_{t=p+1}^{T} \sum_{(\boldsymbol{y_{-t}})} \Big[ \frac{\exp(\boldsymbol{y'_{-t}\beta})}{\{1+\exp(\boldsymbol{y'_{-t}\beta})\}^2} \boldsymbol{y_{-t} y'_{-t}} \Big] Q_t(y_{t-1}, \cdots, y_{t-p}), \tag{2.3}$$

where the conditional joint probability of $Y_{t-1}, \ldots, Y_{t-p}$ is derived to be

$$\begin{aligned}
&Q_t(y_{t-1}, \cdots, y_{t-p}) \\
&= \mathsf{P}(Y_{t-1} = y_{t-1}, Y_{t-2} = y_{t-2}, \cdots, Y_{t-p} = y_{t-p} \mid y_p, y_{p-1}, \cdots, y_1) \\
&= \begin{cases}
1, & \text{if} \quad t = p+1 \\
\prod_{k=1}^{t-p-1} P_{t-k}^{y_{t-k}} (1-P_{t-k})^{1-y_{t-k}}, & \text{if} \quad p+2 \leqslant t \leqslant 2p+1 \\
\sum_{(y_{p+1}, \cdots, y_{t-p-1})} \prod_{k=1}^{t-p-1} P_{t-k}^{y_{t-k}} (1-P_{t-k})^{1-y_{t-k}}, & \text{if} \quad t \geqslant 2p+2.
\end{cases}
\end{aligned} \tag{2.4}$$

### 2.2.2 Logistic Autoregressive Model of Order $p$ with Endogenous Covariates (LARX(p))

Here we consider the case of additional endogenous covariate adjustment in the LAR(p) time series model. Specifically, consider a binary-valued correlated time series $Y_t,\ t = 1, \ldots, T$, where the conditional distribution of $Y_t$ depends on its previous values and endogenous covariates $X_t$ that relates to current time $t$ through the conditional probability

$$P_t = \mathsf{P}(Y_t = 1 \mid y_{t-1}, y_{t-2}, \cdots, y_1) = \frac{\exp(\boldsymbol{y'_{-t}\beta} + \boldsymbol{x'_t\alpha})}{1+\exp(\boldsymbol{y'_{-t}\beta} + \boldsymbol{x'_t\alpha})}.$$

where $\boldsymbol{x_t} = (x_{t1}, \cdots, x_{tl})'$, $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_l)'$ and all the other parameters follow the notation of the previous section. The conditional log-likelihood function of $\boldsymbol{\alpha}, \boldsymbol{\beta}$ and the vector of conditional score functions are respectively

$$\begin{aligned} \ell(\boldsymbol{\alpha}, \boldsymbol{\beta} \mid \boldsymbol{Y}) &= \sum_{t=p+1}^{T} [Y_t(\boldsymbol{X_t'\alpha} + \boldsymbol{Y_{-t}'\beta}) - \log\{1 + \exp(\boldsymbol{X_t'\alpha} + \boldsymbol{Y_{-t}'\beta})\}], \qquad (2.5) \\ \boldsymbol{U(\alpha, \beta, Y)} &= \sum_{t=p+1}^{T} \begin{pmatrix} \boldsymbol{X_t}\{Y_t - \frac{\exp(\boldsymbol{X_t'\alpha + Y_{-t}'\beta})}{1+\exp(\boldsymbol{X_t'\alpha + Y_{-t}'\beta})}\} \\ \boldsymbol{Y_{-t}}\{Y_t - \frac{\exp(\boldsymbol{X_t'\alpha + Y_{-t}'\beta})}{1+\exp(\boldsymbol{X_t'\alpha + Y_{-t}'\beta})}\} \end{pmatrix}. \end{aligned}$$

Then, it follows that the Hessian matrix is

$$\boldsymbol{H(\alpha, \beta \mid Y)} = -\sum_{t=p+1}^{T} \left[ \frac{\exp(\boldsymbol{X_t'\alpha} + \boldsymbol{Y_{-t}'\beta})}{\{1 + \exp(\boldsymbol{X_t'\alpha} + \boldsymbol{Y_{-t}'\beta})\}^2} \begin{pmatrix} \boldsymbol{X_t X_t'} & \boldsymbol{X_t Y_{-t}'} \\ \boldsymbol{Y_{-t} X_t'} & \boldsymbol{Y_{-t} Y_{-t}'} \end{pmatrix} \right]. \qquad (2.6)$$

In this case the Ex-FI matrix takes the form

$$I(\boldsymbol{\alpha}, \boldsymbol{\beta} \mid y_p, \cdots, y_1) =$$
$$\sum_{t=p+1}^{T} \sum_{\boldsymbol{(y_{-t})}} \Big[ \frac{\exp(\boldsymbol{X_t'\alpha} + \boldsymbol{Y_{-t}'\beta})}{\{1 + \exp(\boldsymbol{X_t'\alpha} + \boldsymbol{Y_{-t}'\beta})\}^2} \begin{pmatrix} \boldsymbol{X_t X_t'} & \boldsymbol{X_t Y_{-t}'} \\ \boldsymbol{Y_{-t} X_t'} & \boldsymbol{Y_{-t} Y_{-t}'} \end{pmatrix} \Big] Q_t(y_{t-1}, \cdots, y_{t-p}),$$

where $Q_t(y_{t-1}, \cdots, y_{t-p})$ is defined in Equation (2.4). In practice, examples of endogenous $X_t$ have been discussed in Davis et al. (2000). A particular example is $X_t = t/n$ if one believes that there is a linear temporal trend in the link function (e.g., log mean for the Poisson response in Davis et al. (2000) and the log odds for LARX(p)).

### 2.2.3 Computation through Functional Iteration

Since $(y_{t-1}, \cdots, y_{t-p})$ take values from $\{0,1\}^p$, computation of the Ex-FI matrix through direct calculation can be expensive. In this chapter, we propose an alternative approach to achieve the marginal probability mass function $Q_t(y_{t-1}, \cdots, y_{t-p})$ through functional iteration. We define

$$
\begin{aligned}
&Q_t(y_{t-1}, \cdots, y_{t-p}) \\
&= \begin{cases}
\sum\limits_{w \in \{0,1\}} \{\mathsf{P}(Y_{t-1} = 1 \mid y_{t-2}, \cdots, y_2, w)^{y_{t-1}} \mathsf{P}(Y_{t-1} = 0 \mid y_{t-2}, \cdots, y_2, w)^{(1-y_{t-1})} \\
\times Q_{t-1}(y_{t-2}, \cdots, y_{t-p}, w)\}, \quad \text{if} \quad p+2 \leqslant t \leqslant T \\
\mathbb{1}(y_p, \cdots, y_1), \quad \text{if} \quad t = p+1.
\end{cases}
\end{aligned}
$$

where $\mathbb{1}(y_p, \cdots, y_1)$ is the indicator function that takes value 1 when the realization is $(y_p, \cdots, y_1)$ and 0 otherwise. Then for any $T$ and order $p$, the marginal probability mass function $Q_t(y_{t-1}, \cdots, y_{t-p})$ can be obtained iteratively at low computational cost. The Ex-FI matrix can be achieved accordingly.

On the other hand, the Ex-FI matrix can also be obtained via iterated expectations. Specifically, for any $t \geqslant p+2$, we define

$$
\begin{aligned}
&f_k(\tilde{\boldsymbol{y}}_{-(\boldsymbol{t-k+1})}) \\
&= \begin{cases}
f_{k-1}(\tilde{\boldsymbol{y}}^{\boldsymbol{0}}_{-(\boldsymbol{t-k+1})}) + (f_{k-1}(\tilde{\boldsymbol{y}}^{\boldsymbol{1}}_{-(\boldsymbol{t-k+1})}) - f_{k-1}(\tilde{\boldsymbol{y}}^{\boldsymbol{0}}_{-(\boldsymbol{t-k+1})})) \frac{\exp(\boldsymbol{y}'_{-(\boldsymbol{t-k+1})}\boldsymbol{\beta})}{1+\exp(\boldsymbol{y}'_{-(\boldsymbol{t-k+1})}\boldsymbol{\beta})}, \\
\text{if} \quad 2 \leqslant k \leqslant t-p, \\
\frac{\exp(\boldsymbol{y}'_{-\boldsymbol{t}}\boldsymbol{\beta})}{\{1+\exp(\boldsymbol{y}'_{-\boldsymbol{t}}\boldsymbol{\beta})\}^2} \boldsymbol{y}_{-\boldsymbol{t}} \boldsymbol{y}'_{-\boldsymbol{t}}, \\
\text{if} \quad k = 1.
\end{cases}
\end{aligned}
$$

where $\tilde{\boldsymbol{y}}_{-\boldsymbol{t}} = (y_{t-1}, \cdots, y_{t-p})'$, $\tilde{\boldsymbol{y}}^{\boldsymbol{0}}_{-\boldsymbol{t}} = (0, y_{t-2}, \cdots, y_{t-p})'$ and $\tilde{\boldsymbol{y}}^{\boldsymbol{1}}_{-\boldsymbol{t}} = (1, y_{t-2}, \cdots, y_{t-p})'$. Then

for any particular $t_0$, by function iteration,

$$\mathbb{E}\left[\frac{\exp(\boldsymbol{Y'_{-t_0}}\boldsymbol{\beta})}{\{1+\exp(\boldsymbol{Y'_{-t_0}}\boldsymbol{\beta})\}^2}\boldsymbol{Y_{-t_0}}\boldsymbol{Y'_{-t_0}}\right] = f_{t_0-p}(\tilde{\boldsymbol{y}}_{-(\boldsymbol{p+1})}), \quad t_0 \geqslant p+1.$$

The Ex-FI matrix then takes the form $I(\boldsymbol{\beta} \mid y_p, \cdots, y_1) = \sum_{t_0=p+1}^{T} f_{t_0-p}(\tilde{\boldsymbol{y}}_{-(\boldsymbol{p+1})})$.

### 2.2.4 Special Case: Logistic Autoregressive Model of Order $p = 1$ (LAR(1))

Consider a binary-valued time series data $Y_t,\ t = 1, \ldots, T$, where the conditional distribution of $Y_t$ depends on its own immediate past value via the conditional probability

$$P_t = \mathsf{P}(Y_t = 1 \mid y_{t-1}, y_{t-2}, \cdots, y_1) = \frac{\exp(\beta_0 + \beta_1 y_{t-1})}{1+\exp(\beta_0 + \beta_1 y_{t-1})}.$$

Denote $p(y) = \frac{\exp(\beta_0+\beta_1 y)}{1+\exp(\beta_0+\beta_1 y)}$ and $v(y) = p(y)[1-p(y)]$. Then the corresponding Hessian matrix is derived to be

$$\frac{\partial^2}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta^T}}\ell(\boldsymbol{\beta} \mid \boldsymbol{Y}) = -\begin{pmatrix} \sum_{t=2}^{T} v(Y_{t-1}) & \sum_{t=2}^{T} v(Y_{t-1})Y_{t-1} \\ \sum_{t=2}^{T} v(Y_{t-1})Y_{t-1} & \sum_{t=2}^{T} v(Y_{t-1})Y_{t-1}^2 \end{pmatrix}. \tag{2.7}$$

Next, we will derive the conditional expectation with respect to each entry of the Hessian matrix in Equation (2.7). Due to the Markovian assumption, the conditional expectation

can be obtained through iterated expectations. For any particular $t > 2$, we have

$$\mathbb{E}\left[v(Y_{t-1}) \mid Y_1\right] = A_1 + A_2 \tag{2.8}$$

$$\mathbb{E}\left[v(Y_{t-1})Y_{t-1} \mid Y_1\right] = A_3 + A_4 \quad \text{where} \tag{2.9}$$

$$A_1 = \{v(1) - v(0)\}\{p(1) - p(0)\}^{t-3}\{p(Y_1) - p(0)/\{1 - p(1) + p(0)\}\}$$

$$A_2 = v(0) + p(0)\{v(1) - v(0)\}/\{1 - p(1) + p(0)\}$$

$$A_3 = v(1)\{p(1) - p(0)\}^{t-3}\left[p(Y_1) - p(0)/\{1 - p(1) + p(0)\}\right]$$

$$A_4 = v(1)p(0)/\{1 - p(1) + p(0)\}.$$

Denote the Ex-FI matrix to be $I(\boldsymbol{\beta} \mid Y_1)$. Its elements $I_{jk}, j = 1, 2; k = 1, 2$ are derived, respectively, as

$$\begin{aligned}
I_{11} &= \mathbb{E}[\sum_{t=3}^{T} v(Y_{t-1}) \mid Y_1] + v(Y_1) \\
&= \sum_{t=3}^{T}[\{v(1) - v(0)\}\{p(1) - p(0)\}^{t-3}\{p(Y_1) - \frac{p(0)}{1 - p(1) + p(0)}\} + \\
&\quad v(0) + p(0)\frac{v(1) - v(0)}{1 - p(1) + p(0)}] + v(Y_1) \\
&= \{v(1) - v(0)\}\{p(Y_1) - \frac{p(0)}{1 - p(1) + p(0)}\}\frac{1 - \{p(1) - p(0)\}^{T-2}}{1 - p(1) + p(0)} + \\
&\quad \frac{(T-2)\{p(0)v(1) + v(0) - v(0)p(1)\}}{1 - p(1) + p(0)} + v(Y_1)
\end{aligned}$$

$$
\begin{aligned}
I_{12} &= \mathbb{E}[\sum_{t=3}^{T} v(Y_{t-1})Y_{t-1} \mid Y_1] + v(Y_1)Y_1 \\
&= \sum_{t=3}^{T}[v(1)\{p(1)-p(0)\}^{t-3}\{p(Y_1) - \frac{p(0)}{1-p(1)+p(0)}\} + \frac{v(1)p(0)}{1-p(1)+p(0)}] + v(Y_1)Y_1 \\
&= v(1)\{p(Y_1) - \frac{p(0)}{1-p(1)+p(0)}\}\frac{1-\{p(1)-p(0)\}^{T-2}}{1-p(1)+p(0)} + \frac{(T-2)p(0)v(1)}{1-p(1)+p(0)} + v(Y_1)Y_1 \\
I_{22} &= \mathbb{E}[\sum_{t=3}^{T} v(Y_{t-1})Y_{t-1}^2 \mid Y_1] + v(Y_1)Y_1^2 \\
&= I_{12}.
\end{aligned}
$$

*Remark* 1. Note that $p(y)$ does not hold the constraint that $p(0) + p(1) = 1$, since

$$p(0) = \mathsf{P}(Y_t = 1 \mid Y_{t-1} = 0) \quad \text{and} \quad p(1) = \mathsf{P}(Y_t = 1 \mid Y_{t-1} = 1).$$

*Remark* 2. Evaluation and selection among different models could be a critical issue. In particular, selection of the order $p$ needs to be taken into serious consideration. Motivated by the work of Kedem and Fokianos (2005), we may select the optimal lag order $p$ using either the Akaike Information Criterion (AIC) or the Bayesian Information Criterion (BIC) which are defined to be $AIC(p) = -2\ell(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}} \mid \boldsymbol{Y}) + 2p$ and $BIC(p) = -2\ell(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}} \mid \boldsymbol{Y}) + p \log T$ respectively, where $(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}})$ is the maximum likelihood estimator of $(\boldsymbol{\alpha}, \boldsymbol{\beta})$.

## 2.3 Simulations

### 2.3.1 Evaluating Small Sample Performance

In this section, we compare the behavior of the newly derived Ex-FI and Em-FI in the context of inference for regression parameters under the LAR(1) model. Time series lengths are chosen as $T = 20, 50$ and 200 respectively, and $10,000$ simulations were generated under each of two parameters settings: $(\beta_0, \beta_1) = (0.1, 0.5)$ ("low ratio"), and $(\beta_0, \beta_1) = (0.1, 1)$ ("high ratio"). In this case, $\beta_1$ denotes the log odds ratio and $\beta_0$ denotes the log odds when the previous realization is 0. $\beta_1/\beta_0$ is a monotonic function of the log odds ratio of $Y_t = 1$. Particularly, large value (greater than 1) of $\beta_1/\beta_0$ implies the log odds of $Y_t = 1$ when $Y_{t-1} = 1$ is much higher compared to the log odds when $Y_{t-1} = 0$. For each scenario, we calculate the empirical type I error-rate for testing $H_0 : \beta_1 = 0$ at level .05, the average standard error of the point estimate of $\beta_1$, the observed standard deviation of the estimate of $\beta_1$ across simulations, and the empirical coverage probability of 95% confidence intervals (CIs).

Table 2.1 provides a summary of the conducted simulation study for various time series lengths. With respect to type I error, it can be seen that use of Ex-FI and Em-FI both result in conservative inference (lower than nominal type I error) for smaller values of $T$ and for high ratios. For the low ratio scenario, nominal type I error rates are achieved as time series lengths of $T = 50$. For time series lengths of $T = 200$ both variance estimators yield the desired type I error rates. As expected, similar patterns are observed with respect to the coverage probability of corresponding 95% confidence intervals. However, the benefit of using Ex-FI over Em-FI is observed when comparing the average standard error to the observed standard deviation of estimates of $\beta_1$ across simulations. Specifically, Em-FI tends to behave erratically for small sample sizes, yielding extremely large estimated standard

Table 2.1: Summary of simulation results for the LAR(1) model. Time series lengths are chosen as $T = 20, 50$ and 200. 10,000 simulations were generated under each of two parameters settings: $(\beta_0, \beta_1) = (0.1, 0.5)$ ("low ratio") and $(\beta_0, \beta_1) = (0.1, 1)$ ("high ratio"). For each scenario, we present the empirical type I error-rate for testing $H_0 : \beta_1 = 0$, the average standard error of the point estimator for $\beta_1$, the observed standard deviation of the regression parameter estimate of $\beta_1$ across simulations, and the empirical coverage probability of 95% confidence intervals.

| Length/Method | | Low Ratio ($(\beta_0, \beta_1) = (0.1, 0.5)$) | | | High Ratio ($(\beta_0, \beta_1) = (0.1, 1)$) | | |
|---|---|---|---|---|---|---|---|
| | | Type I Error | Standard Error* | Coverage Probability | Type I Error | Standard Error* | Coverage Probability |
| $T = 20$ | | | | | | | |
| | Ex-FI | 0.031 | **3.737(2.290)** | 0.969 | 0.008 | **7.868(3.015)** | 0.992 |
| | Em-FI | 0.030 | **32.87(2.290)** | 0.970 | 0.011 | **362.3(3.015)** | 0.989 |
| $T = 50$ | | | | | | | |
| | Ex-FI | 0.048 | **0.617(0.632)** | 0.952 | 0.039 | **1.065(1.074)** | 0.961 |
| | Em-FI | 0.044 | **0.956(0.632)** | 0.956 | 0.039 | **1.222(1.074)** | 0.961 |
| $T = 200$ | | | | | | | |
| | Ex-FI | 0.052 | **0.299(0.299)** | 0.948 | 0.051 | **0.332(0.325)** | 0.949 |
| | Em-FI | 0.052 | **0.297(0.299)** | 0.948 | 0.053 | **0.324(0.325)** | 0.947 |

*Standard error represents the average standard error of the point estimator for $\beta_1$ and, in parentheses, the actual observed standard deviation of the regression parameter estimate of $\beta_1$ across simulations.

error for some simulated datasets. This can be seen most notably in the high ratio scenario by observing that the average standard error computed using Em-FI is 362.3 compared to the actual observed standard deviation of the estimator across 10,000 simulations being only 3.015. In contrast, the average standard error computed using Ex-FI is only 7.868.

### 2.3.2 Evaluation of Confidence Interval Length

Here we consider the average length of derived 95% confidence intervals for $\beta_1$. Following the result that asymptotically, $\widehat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, I^{-1}(\boldsymbol{\beta}))$ for large values of $T$ (Fokianos and Kedem, 1998a), an approximate 95% confidence interval can be obtained using both Ex-FI and Em-FI. For each scenario of $\boldsymbol{\beta}$ described above, 1000 binary time series of lengths $T = 5, 6, \ldots, 250$ were generated. For each time series data, an approximate 95% confidence interval for $\beta_1$ was computed using both Ex-FI and Em-FI. We compared the two approaches by calculating

the relative difference of the lengths of the two confidence intervals. As expected from the average standard error values in Table 1, Fig.2.1 indicates that the confidence interval derived from Ex-FI behaves more efficiently on average than the confidence interval computed using Em-FI. It is noted that such substantial difference exists when $T < 200$ and tends to be roughly the same as $T$ goes beyond 200. Once again, it implies that one should be careful with the Em-FI when $T < 200$.

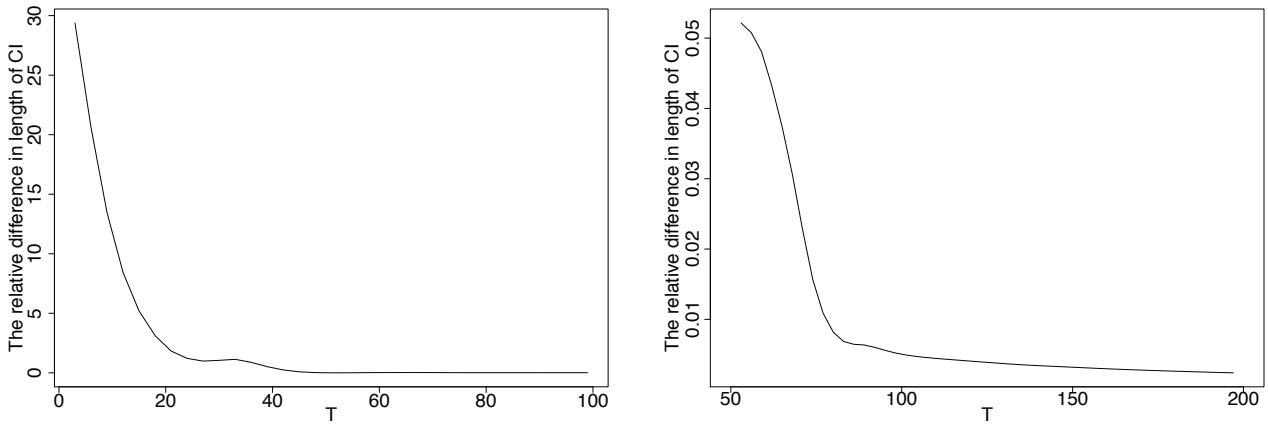



Figure 2.1: The average difference in lengths of confidence intervals derived from Ex-FI and Em-FI (length of $\mathrm{CI}^{\mathrm{empirical}}$−length of $\mathrm{CI}^{\mathrm{exact}}$)/length of $\mathrm{CI}^{\mathrm{exact}}$ computed from 1000 simulated time series with $\beta_1/\beta_0 = 10$. The lengths of time series, $T$ ranges from (a) 5 to 100 (left) and (b) 50 to 200 (right). The Em-FI matrix used here was identical to the one proposed in Fokianos and Kedem (1998a).

In Fig.2.2, $T$ was fixed to 60 and 100 and $\beta_1$ was allowed to vary while keeping $\beta_0 = 0.1$. Results clearly establish the advantage of Ex-FI over Em-FI especially as the true value of $\beta_1$ increases, i.e., the ratio increases.

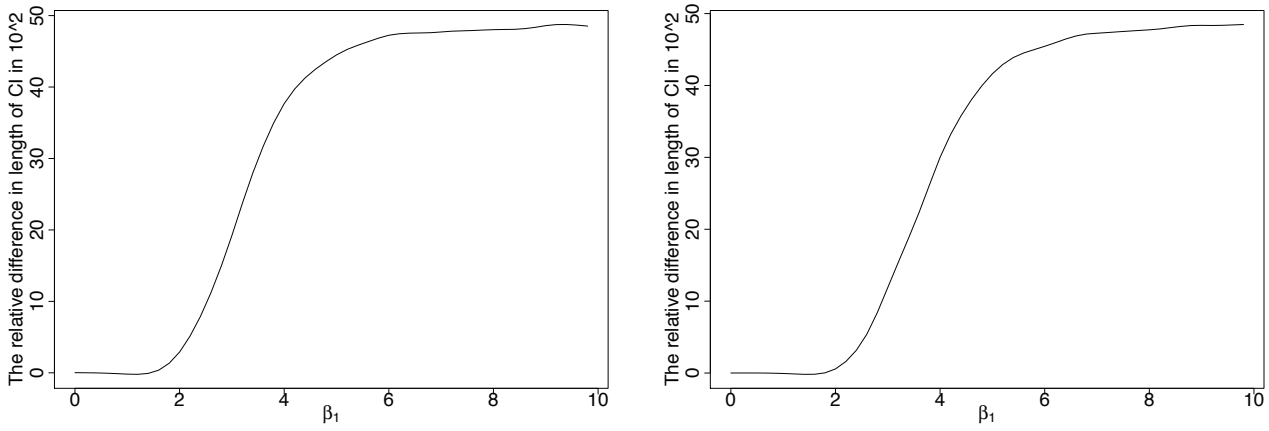



Figure 2.2: The average relative difference in length of confidence intervals, computed from 1000 simulated datasets, derived from Ex-FI and Em-FI (length of $\text{CI}^{\text{empirical}}-$length of $\text{CI}^{\text{exact}}$)/length of $\text{CI}^{\text{exact}}$, where the number of observations is taken to be (a) $T = 60$ (left) and (b) $T = 100$ (right). $\beta_0$ is fixed to be 0.1. The Em-FI matrix used was developed in Fokianos and Kedem (1998a).

### 2.3.3 Evaluating the Discrepancy between the Exact and Empirical Fisher Information

In this section, we discuss the results of simulations conducted to investigate the discrepancy between Ex-FI and Em-FI under the following scenarios: (i.) time series lengths $T$ ranging from $10-250$; (ii.) the ratio $\beta_0/\beta_1 \in \{5, 10\}$. Based on $1,000$ simulated time series under each scenario, the average Frobenius norm of the difference between the asymptotic covariance matrices (i.e. the inverse of Ex-FI and Em-FI), displayed in Fig.2.3, shows that when $T > 200$ any discrepancy between the two covariance matrices effectively vanishes. However, for $T < 200$, discrepancies do exist, primarily due to the instability of Em-FI for particular datasets. The result reiterates that caution needs to be taken when utilizing the Em-FI variance estimator for shorter time series, since this erratic behaviour could lead to significant errors in the estimated variances of regression parameters.

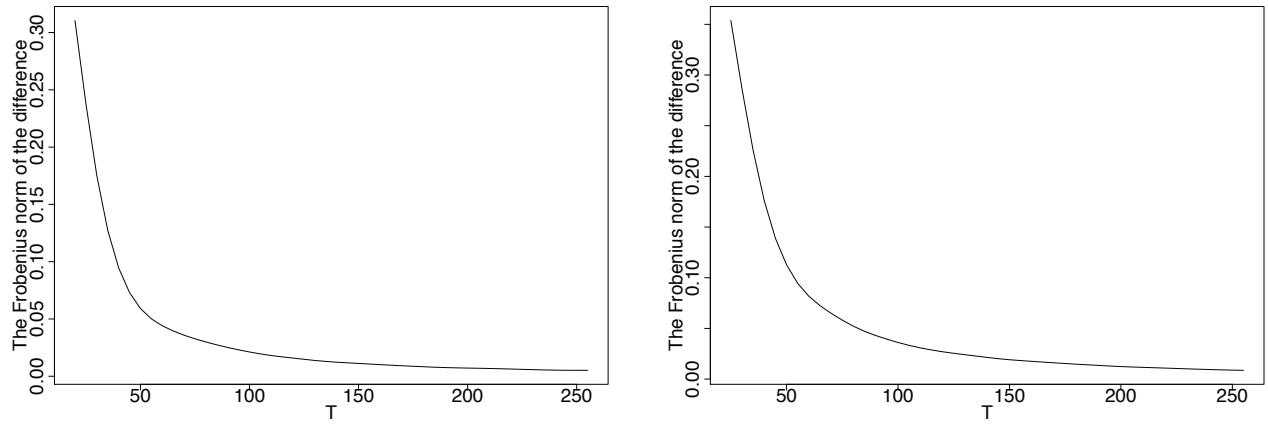



Figure 2.3: The average Frobenius norm of the difference between the inverse of the exact Fisher information (Ex-FI) and empirical Fisher information (Em-FI) (as developed in Fokianos and Kedem (1998a)) under the two parameter set up: (a) $\beta_1/\beta_0 = 5$ (left) and (b) $\beta_1/\beta_0 = 10$ (right). The average Frobenius norm was calculated from 1,000 simulated time series for varying time series lengths under each of the parameter set-up.

### 2.3.4 Evaluating the Convergence

We considered the asymptotic behavior of Ex-FI and compared it to the AFI proposed by Fokianos and Kedem (1998a) by computing the average Frobenius norm between the two matrices over $1,000$ simulated time series data. In Fig.2.4, it is clear that the discrepancy between these two matrices decays dramatically, which empirically indicates that the limiting behavior between the two estimators coincides. It should be emphasized that when $T < 200$, the difference is significant while as $T$ grows larger than 200, the discrepancy shrinks to small values around 0. Hence, utilizing the Em-FI when $T < 200$ may be problematic.

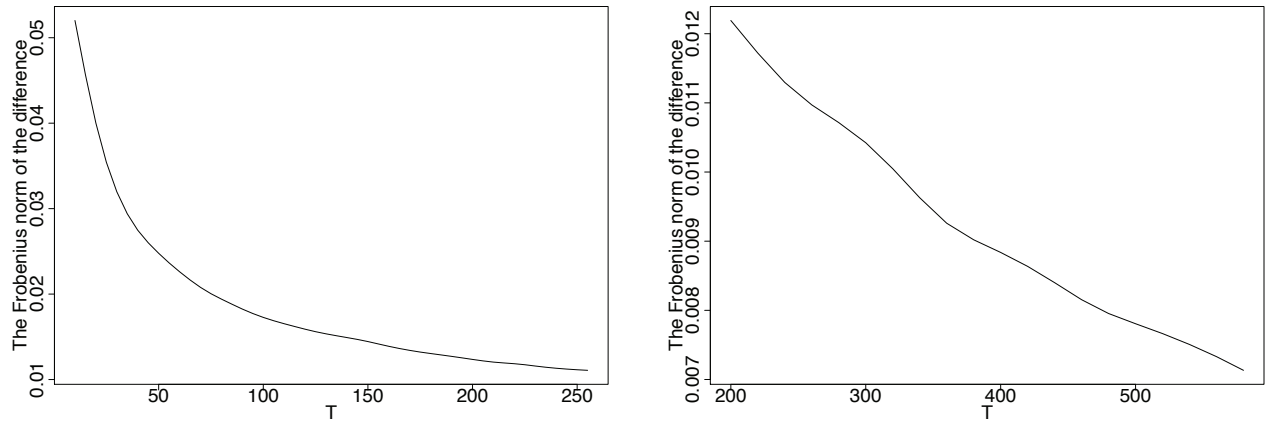



Figure 2.4: The average Frobenius norm of the difference in Ex-FI and AFI matrices (which is proposed in Fokianos and Kedem (1998a)) computed over 1000 simulated time series under the set-up $\beta_1/\beta_0 = 5$. The lengths of time series, (a) $T$ ranges from 5 to 250 (left) and (b) 250 to 550 (right).

## 2.4 Analysis of Binary Respiratory Time Series

### 2.4.1 Explanatory Analysis

In this section we consider time-series data on respiratory rate among a cohort of 113 expectant mothers. Briefly, the participants consist of a sub-sample of women from a larger cohort of women attending prenatal care at a university-based clinic in Pittsburgh, PA and participating in a prospective, longitudinal study from early gestation through birth (Entringer et al., 2015). Participants were asked to wear a heart and respiratory rate monitor for up to four consecutive days. In addition, each night prior to sleeping the participants were asked to fill out an electronic diary recording how stressful their day was on a scale from 1 to 10 ($X_i$), with 10 corresponding to the highest self-reported stress level. The study was approved by the local Institutional Review Board (IRB).

Of scientific interest is the potential association between self-reported stress and respiratory, or breath, rate measured as the number of breaths per 60 second period. For the purposes of illustration, we consider a participants breath rate averaged over one-hour intervals starting from midnight and running to midnight over the maximum of a 24 hour period. Empirical data suggests that a respiratory rate of over 20 breaths per minute is considered high for a healthy adult (Barrett et al., 2010). As such, the time series in this study are discretized into a binary response using a threshold of greater than 20 breaths/min. Accordingly, if we denote $Y_{it}$ as the average breath rate for subject $i$ at hour $t$, we define $Y_{it} = 1$ if the observed average respiratory rate is greater than 20 breaths/min, and 0 otherwise. To illustrate, Fig.2.5 presents the observed time series for a randomly sampled participant. Table 2.2 depicts the empirical transition table of respiratory rate across all subjects. It illustrates a strong association between the current realization of $Y_{it}$ and lagged values of $Y_{i,t-1}$ and $Y_{i,t-2}$. In this study, one scientific question of interest is whether or not a potential interaction exists between the lagged realizations $Y_{i,t-1}, Y_{i,t-2}$ and a participant's observed stress level $X_i$. Specifically, it is hypothesized that the association between lagged responses and current breath rate is lower among individuals reporting high stress due to the erratic breathing patterns that high stress situations can evoke. As such, we consider a LARX model including the lagged realization, an indicator for high stress ($1_{[X_i>7]}$), and their interaction. In this study, similar to the discussion in Holmes and Rahe (1967), a subject is considered to be in high stress if the scale exceeds 7.

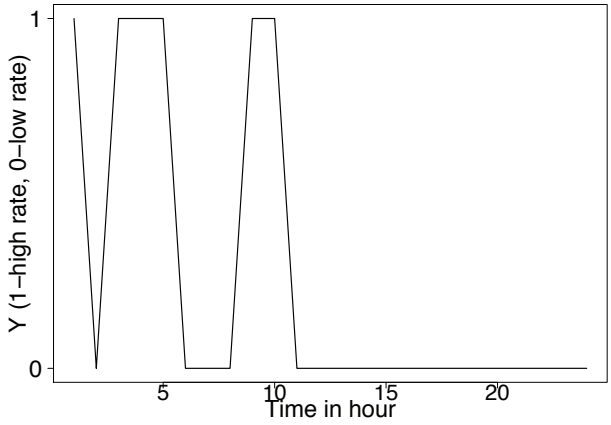



Figure 2.5: Binary respiratory time series.

Table 2.2: Empirical transition table of respiratory rate across all subjects

| Lagged respiratory rate | Current respiratory rate | |
|---|---|---|
| | $Y_{i,t} = 0$ | $Y_{i,t} = 1$ |
| $Y_{i,t-1} = 0$ | 0.865 | 0.046 |
| $Y_{i,t-1} = 1$ | 0.038 | 0.051 |
| $Y_{i,t-2} = 0$ | 0.855 | 0.047 |
| $Y_{i,t-2} = 1$ | 0.038 | 0.060 |

### 2.4.2 Fitting the LARX Model to the Respiratory Binary Time Series Data

We consider LARX(1) and LARX(2) models fitted across the 113 subjects with the same parameter. Stress level $1_{[X_i>7]}$ and the interactions between stress level and past values of the binarized respiratory rate $1_{[X_i>7]} \times Y_{it}$, $1_{[X_i>7]} \times Y_{i,t-1}$ were considered to be covariates. With the independence assumption across subjects, we fit a log likelihood function that is the sum of the log likelihood function (2.5) for each subject. Table 2.3 provides 95% confidence intervals for the functionals $\mathsf{P}(Y_{it} = 1 \mid Y_{i,t-1})$ and $\frac{\mathsf{P}(Y_{it}=1|Y_{i,t-1})}{\mathsf{P}(Y_{it}=0|Y_{i,t-1})}$ after fitting the LARX(1) model. It can be seen that the confidence intervals derived from Ex-FI are consistently

shorter than Em-FI. Specifically, when $Y_{i,t-1} = 1$ the confidence interval for $\mathsf{P}(Y_{it} = 1 \mid Y_{i,t-1})$ derived from Ex-FI excludes 0.5 (odds excludes 1), while the confidence interval resulting from the use of Em-FI includes 0.5 (odds includes 1). Under the LARX(2) model, the pattern is more obvious. From Table 2.4, it can be seen that comparing the confidence interval from Ex-FI to Em-FI, the average length of all the functionals are relatively smaller. In the most extreme case the Ex-FI derived confidence interval for the odds of high respiratory rate among high stress individuals is approximately 30% shorter (and excluding 1), when compared to the confidence interval derived using Em-FI. Using the Ex-FI approach, the lagged realizations are determined to be significantly associated with respiratory rate: expectant mothers with low stress level tend to have low rate if their previous realizations are low. In contrast, the wider Em-FI intervals do not rule out a odds of 1 associated with high prior respiratory state among high stress mothers.

Table 2.3: The 95% confidence intervals of functionals $\mathsf{P}(Y_{it} = 1 \mid Y_{i,t-1})$ (Prob) and $\frac{\mathsf{P}(Y_{it}=1|Y_{i,t-1})}{\mathsf{P}(Y_{it}=0|Y_{i,t-1})}$ (Odds) obtained by fitting the LARX(1) model with stress level and interaction between stress level and past values of the binarized respiratory rate.

| | | Low Stress ($1_{[X_i>7]} = 0$) | | High Stress ($1_{[X_i>7]} = 1$) | |
|---|---|---|---|---|---|
| Previous State/Method | | Prob | Odds | Prob | Odds |
| $Y_{i,t-1} = 0$ | | | | | |
| | Ex-FI | (0.042, 0.061) | (0.044, 0.065) | (0.027, 0.085) | (0.028, 0.093) |
| | Em-FI | (0.042, 0.061) | (0.044, 0.065) | (0.027, 0.085) | (0.028, 0.093) |
| $Y_{i,t-1} = 1$ | | | | | |
| | Ex-FI | **(0.505, 0.630)** | **(1.021, 1.701)** | (0.373, 0.731) | (0.594, 2.724) |
| | Em-FI | **(0.498, 0.635)** | **(0.998, 1.738)** | (0.366, 0.737) | (0.577, 2.802) |

Table 2.4: The 95% confidence intervals of functionals $\mathsf{P}(Y_{it} = 1 \mid Y_{i,t-1}, Y_{i,t-2})$ (Prob) and $\frac{\mathsf{P}(Y_{it}=1|Y_{i,t-1},Y_{i,t-2})}{\mathsf{P}(Y_{it}=0|Y_{i,t-1},Y_{i,t-2})}$ (Odds) obtained by fitting the LARX(2) model with stress level and interaction between stress level and past values of the binarized respiratory rate.

| | | Low Stress ($1_{[X_i>7]} = 0$) | | High Stress ($1_{[X_i>7]} = 1$) | |
|---|---|---|---|---|---|
| Previous State/Method | | Prob | Odds | Prob | Odds |
| $Y_{i,t-2} = 0, Y_{i,t-1} = 0$ | | | | | |
| | Ex-FI | (0.044, 0.064) | (0.046, 0.068) | (0.023, 0.081) | (0.023, 0.088) |
| | Em-FI | (0.044, 0.064) | (0.046, 0.068) | (0.023, 0.080) | (0.023, 0.087) |
| $Y_{i,t-2} = 1, Y_{i,t-1} = 0$ | | | | | |
| | Ex-FI | (0.394, 0.553) | (0.651, 1.241) | (0.349, 0.851) | (0.537, 5.701) |
| | Em-FI | (0.385, 0.563) | (0.626, 1.290) | (0.349, 0.851) | (0.537, 5.707) |
| $Y_{i,t-2} = 0, Y_{i,t-1} = 1$ | | | | | |
| | Ex-FI | (0.100, 0.201) | (0.117, 0.251) | (0.017, 0.230) | (0.017, 0.299) |
| | Em-FI | (0.100, 0.210) | (0.111, 0.265) | (0.015, 0.250) | (0.016, 0.332) |
| $Y_{i,t-2} = 1, Y_{i,t-1} = 1$ | | | | | |
| | Ex-FI | (0.670, 0.787) | (2.033, 3.696) | **(0.535, 0.869)** | **(1.151, 6.630)** |
| | Em-FI | (0.653, 0.800) | (1.878, 4.001) | **(0.469, 0.896)** | **(0.885, 8.621)** |

# Chapter 3

# Modeling Binary Time Series Using Gaussian Processes

## 3.1 Introduction

The goal of this chapter is motivated by developing statistical inference for studying changes in the sleep state (in particular, asleep versus awake) and the potential roles of covariates such as heart rate, respiration rate and body temperature on sleep states. A plot of the sleep states and the exogenous time series of heart rate and temperature, given in Figure (3.1), suggest a lead-lag depenence between sleep states and the exogenous time series. In this chapter, we develop a model that formally tests for these lead-lag dependence and predict future sleep states.

This work is inspired by Keenan (1982b) who developed a binary time series using a latent strictly stationary process. The focus here is to provide an accurate, interpretable, efficient yet computationally less demanding approach for estimation and prediction. When prior

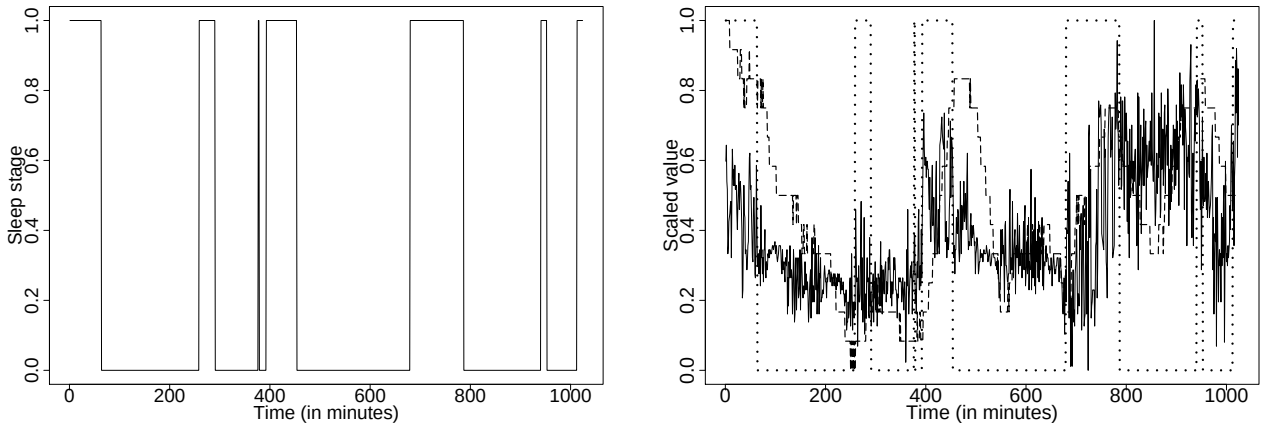



Figure 3.1: Left: sleep state. Right: sleep state plot (dotted line) overlaid by scaled heart rate (solid line) and body temperature (dashed line) time plots.

information indicates that a binary time series is determined by a process comprised with fixed and random components, we decompose the unobserved latent process into linear and stochastic effects with different covariates. On stage one, inference on the fixed effects is conducted using maximum likelihood estimation. On stage two, conditioned on the estimated fixed effect, a Gaussian process will be used to represent the random components. Predictions are obtained by combining inference on these two components. In addition, based on the results from these two stages, we use parametric bootstrap samplers from the estimated Gaussian process to obtain the final point and interval estimates of parameters.

Using the proposed procedure, we can identify the dependence of the endogenous time series (sleep state) on potential covariates (e.g., heart rate and body temperature) by providing the point and interval estimates of the coefficients from linear effects based on the results from the two-stage algorithm. Inference can be easily and directly performed by maximum likelihood using existing software. Moreover, results are easily interpretable under the framework of generalized linear model. On stage two, which is derived from Gaussian process classification strategy, we can predict the sleep state with high accuracy. Laplace approximation was

implemented to reduce the computation cost. This work is also inspired by Brillinger (1983) which, to the best of our knowledge, is the first to introduce this notion of a Gaussian random effect as random intercept in a logit model. Here, we generalize this by representing the random component as a stochastic process rather than just a scalar random variable.

The main advantages of our proposed approach, which we call the hybrid inference method for binary time series (HIBITS), are the following: (1) it accounts for the linear and non-linear stochastic effects of covariates and endogenous variables on sleep states; (2) it provides efficient point and interval estimates of the coefficients from the linear effects while maintaining type I error rates; (3) it produces more accurate predictions compared to other existing approaches; (4) it is easily implemented with low computational cost; and (5) unlike other classification approaches, it gives more straightforward interpretation of the results.

The remainder of this chapter is organized as follows. Section 3.2 is devoted to brief introduction of Gaussian process and its existing applications in regression and classification. In Section 3.3, we develop our proposed methodology and discuss the motivation and the technical derivation of the proposed HIBITS method. A complete algorithm that yields prediction and inference on the coefficients of covariates and endogenous variables is also provided. Model selection strategy is also developed to address application problems. Section 3.4 presents the simulation results that show the benefits of the proposed method over the existing methods in terms of the significant higher prediction accuracy and narrower confidence intervals. In Section 3.5, we apply our proposed model and inference procedure to identify predictors of sleep states and to predict future sleep states. The results are promising in terms of prediction accuracy at low computational cost and interpretability. Moreover, the proposed method can also be modified when there are missing values.

## 3.2 Background on Gaussian Processes in Binary Time Series

### 3.2.1 Gaussian Process and Regression Models

Gaussian process have been widely developed in spatial-temporal modeling (Williams and Rasmussen, 2006; Banerjee et al., 2008, 2014; Gelfand et al., 2005; Quick et al., 2013; Stein, 2012; Zhou et al., 2015; Vandenberg-Rodes and Shahbaba, 2015; Wang and Gelfand, 2014). It provides a framework that can capture the non-linear and stochastic components of exogenous and endogenous variables based on generalized linear models, which makes it useful for modeling binary time series and classification.

The definition of a Gaussian process is as follows.

**Definition 1.** A stochastic process is a Gaussian process if and only if for every finite set of indices $t_1, \cdots, t_k$ in the index set $T$, $\mathbf{x} = (x_{t_1}, \cdots, x_{t_k})^T$ is a multivariate Gaussian random variable.

We will write the Gaussian process $f(\mathbf{x})$ as $f(\mathbf{x}) \sim \mathcal{GP}(m(\mathbf{x}), K(\mathbf{x}, \mathbf{x}'))$, where $m(\mathbf{x}) = \mathbb{E}[f(\mathbf{x})]$ and $K(\mathbf{x}, \mathbf{x}') = \mathbb{E}[(f(\mathbf{x}) - m(\mathbf{x}))(f(\mathbf{x}') - m(\mathbf{x}'))]$. Let us now denote the observed data to be $\{(\mathbf{x_i}, y_i), i = 1, \cdots, n + n_*\}$, where $\mathbf{x_i} \in \mathbf{R^p}$ and $y_i$ is the response data. We split the dataset into $n$ training points and $n_*$ testing points. Let $(\mathbf{X}_*, \mathbf{y}_*)$ represent the testing datasets and $(\mathbf{X}, \mathbf{y})$ represent the training datasets respectively. Define $\mu = K(\mathbf{X}_*, \mathbf{X})K(\mathbf{X}, \mathbf{X})^{-1}\mathbf{y}, \Sigma = K(\mathbf{X}_*, \mathbf{X}_*) - K(\mathbf{X}_*, \mathbf{X})K(\mathbf{X}, \mathbf{X})^{-1}K(\mathbf{X}, \mathbf{X}_*)$. It follows that

$$\mathbf{y}_*|\mathbf{X}, \mathbf{X}_*, \mathbf{y} \sim N(\mu, \Sigma). \tag{3.1}$$

The distribution of the response $\mathbf{y}_*$ can be determined by Equation (3.1). Point estimates, interval estimates and sampling distribution of $\mathbf{y}_*$ can be derived accordingly.

*Remark* 3. On stage two of the proposed method (discussed in Section 3.3.2), results in Equation (3.1) will be utilized to achieve the distribution of the stochastic component which captures the variation in the binary time series beyond which are explained by the covariates.

### 3.2.2 Gaussian Process in Modeling Binary Time Series

#### Model Formulation

Denote the observed training data as $\{(\mathbf{x_i}, y_i), i = 1, \cdots, n\}$, where $y_i \in \{1, 0\}$ and $\mathbf{x_i} \in \mathbf{R^p}$. For our data in this chapter, $y_i$ denotes the sleep state at time point $i$ and $\mathbf{x_i}$ can be heart rate or body temperature at time point $i$. We define a latent Gaussian process indexed by $\mathbf{x}$ as $f(\mathbf{x})$. The relationship between $\mathbf{x_i}$ and $y_i$ is characterized by $\mathsf{P}(y_i = 1|\mathbf{x_i}) = t(f(\mathbf{x_i}))$, where $t$ is a link function that determines the relation between $\mathbf{x}$ and the probability of the sleep state. To name a few, $t$ can be a logit, probit or complementary log-log link functions (McCullagh, 1984).

#### Classification Method

For a given link, the inferential procedure will be divided into two steps. First, we compute the distribution of the latent process on the test data

$$p(f_*|\mathbf{X}, \mathbf{y}, \mathbf{X}_*) = \int p(f_*|\mathbf{X}, \mathbf{f}, \mathbf{X}_*)p(\mathbf{f}|\mathbf{X}, \mathbf{y})d\mathbf{f}, \tag{3.2}$$

where $p(\mathbf{f}|\mathbf{X},\mathbf{y}) = p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\mathbf{X})/p(\mathbf{y}|\mathbf{X})$. Then, we estimate the conditional probability of $y_* = 1$ by

$$p(y_* = 1 \mid \mathbf{X},\mathbf{y},\mathbf{X}_*) = \int t(f_*)p(f_*|\mathbf{X},\mathbf{y},\mathbf{X}_*)df_*, \tag{3.3}$$

which is approximately a weighted average of the probability of $y_* = 1$ over all possible realizations of predicted stochastic components that is a Gaussian process.

It should be pointed out that both of the two integrands in Equations (3.2) and (3.3) do not have closed forms. For Equation (3.3), following the argument in Williams and Rasmussen (2006), numerical tools such as Monte Carlo method can be used to obtain the approximate value of the integral given $p(f_*|\mathbf{X},\mathbf{y},\mathbf{X}_*)$. To obtain Equation (3.2), Williams and Barber (1998) introduced Laplace approximation for this problem. Minka (2001) proposed an alternative expectation propagation(EP). Besides these methods, a number of MCMC algorithms have also been considered. In the following section, we will follow the direct Laplace approximation.

From Equation (3.2), we can write the approximate distribution of $p(\mathbf{f}|\mathbf{X},\mathbf{y})$ as $N(\hat{\mathbf{f}}, \hat{I}^{-1})$, where $\hat{\mathbf{f}}$ is the MLE of the distribution and $\hat{I}$ is the observed Fisher information matrix. To find the value of $\hat{\mathbf{f}}$, Newton's method can be implemented, where in each iteration $\mathbf{f}^{\text{new}} = \mathbf{f}^{\text{old}} - \nabla^2 \log p(\mathbf{f}^{\text{old}}|\mathbf{X},\mathbf{y})^{-1}\nabla \log p(\mathbf{f}^{\text{old}}|\mathbf{X},\mathbf{y}) = (K^{-1}(\mathbf{X},\mathbf{X}) + W)^{-1}(W\mathbf{f}^{\text{old}} + \nabla \log p(\mathbf{y}|\mathbf{f}^{\text{old}}))$, where $W = -\nabla^2 \log p(\mathbf{y}|\mathbf{f}^{\text{old}})$ and $K(\mathbf{X},\mathbf{X})$ is the covariance matrix of $f(\mathbf{X})$. Thus, the distribution $p(\mathbf{f}|\mathbf{X},\mathbf{y})$ can be approximated by $N(\hat{\mathbf{f}}, (K^{-1}(\mathbf{X},\mathbf{X}) + W)^{-1})$.

Opper and Winther (1999) suggested the conditional expectation of $f_*$ could be obtained by $\mathbb{E}(f_*|\mathbf{X},\mathbf{y},\mathbf{X}_*) = K(\mathbf{X}_*,\mathbf{X})^T K^{-1}(\mathbf{X},\mathbf{X})\hat{\mathbf{f}} = K(\mathbf{X}_*,\mathbf{X})^T \nabla \log p(\mathbf{y}|\hat{\mathbf{f}})$. Following similar arguments, the conditional variance of $f_*$ can be obtained by $\mathbb{V}(f_*|\mathbf{X},\mathbf{y},\mathbf{x}_*) = K(\mathbf{X}_*,\mathbf{X}_*) - K(\mathbf{X}_*,\mathbf{X})^T(K^{-1}(\mathbf{X},\mathbf{X}) + W)^{-1}K(\mathbf{X}_*,\mathbf{X})$. Given the mean and variance, at the last step, the

probability of $y_* = 1$ can be approximated by $\int t(f_*)\hat{p}(f_*|\mathbf{X}, \mathbf{y}, \mathbf{X}_*)df_*$. It should be pointed out that the Gaussian process essentially captures information beyond those provided by past value of both endogenous and exogenous time series.

*Remark* 4. $\frac{\partial^2}{\partial \mathbf{f_i}^2} \log p(y_i|\mathbf{f_i})$ takes the following forms for the logit and probit links, respectively,

$$\begin{aligned}
\frac{\partial^2}{\partial \mathbf{f_i}^2} \log p(y_i|\mathbf{f_i}) &= -p(y_i = 1|\mathbf{f_i})p(y_i = 0|\mathbf{f_i}) \\
\frac{\partial^2}{\partial \mathbf{f_i}^2} \log p(y_i|\mathbf{f_i}) &= -\frac{\varphi(\mathbf{f_i})^2}{\Phi((2y_i - 1)\mathbf{f_i})^2} - \frac{(2y_i - 1)\mathbf{f_i}\varphi(\mathbf{f_i})}{\Phi((2y_i - 1)\mathbf{f_i})}
\end{aligned}$$

Here $\varphi(.)$ and $\Phi(.)$ are the normal probability density function and the cumulative distribution function, respectively.

## 3.3 HIBITS: The Hybrid Estimation Method for Modeling and Predicting Binary Time Series

Building on the established theoretical foundations of Gaussian processes, we now develop a novel two-stage inference and classification method. This section is organized as follows: in Section 3.3.1, we discuss the motivation of using the hybrid strategy in modeling sleep stage; followed by details of the two-stage hybrid method in Section 3.3.2; in Section 3.3.3, we discuss our model selection strategy; and in Section 3.3.4, we provide a method in providing point and interval estimates of the coefficients of the covariates and endogenous variables.

### 3.3.1 Motivation

The common approach is to use a Gaussian distribution with zero mean value as a random effect if the latent process yields, equally likely, positive and negative fluctuations around 0 (Kuss, 2006). Yet, when it comes to real data, this set up overlooks the linear structure between covariates and the actual response of interest. For instance, to model the binary sleep state, scientists believe that body temperature and heart rate should be involved as potential predictors. In Fokianos and Kedem (2002), a regression-based approach for modeling covariates is proposed. However, if we naively utilize the existing Gaussian distribution with zero mean function to model the data, the latent process equally produces positive and negative value fluctuating around 0 which can produce misleading results because it will render the effects of covariates (body temperature and heart rate) to be insignificant. In addition, incorporating those covariates in the covariance function is a reasonable approach to modeling the association. However, the interpretation is complicated. Much work has been done to overcome the aforementioned limitations. To name a few, Snelson et al. (2004) proposed an approach to transform data in agreement with the Gaussian process model. Their work generalized the Gaussian process by warping the observational space. Although the transformed data can be fitted by Gaussian process, it leads to difficulty in the interpretation of the transform. Another drawback is that the effects of particular covariates could be lost (or difficult to interpret). Cornford (1998) suggested a Gaussian process regression model with mean function $m(x) = \beta^{\mathbf{T}}\mathbf{x}$. Their work incorporates the effect of particular covariates. The main drawback is the computational burden that results from the choice of hyperparameter and MCMC sampler when it applies to classification problem. Building on the prior work, we develop a two-stage method that takes advantage of the strengths of the existing methods. It is able to model the linear association with particular covariates while maintaining computational efficiency.

### 3.3.2 The Proposed HIBITS Method

Consider the data $\{(\boldsymbol{x_{1,i}}, \boldsymbol{x_{2,i}}, y_i)\}$ where $y_i \in \{1, 0\}, \mathbf{x_{1,i}} \in \mathbf{R^p}, \mathbf{x_{2,i}} \in \mathbf{R^q}$. Here, $\mathbf{x_{1,i}}$ are the covariates in the fixed effects part and $\mathbf{x_{2,i}}$ are covariates in the stochastic part. Then, $\mathsf{P}(y_i = 1|\mathbf{x_{1,i}}, \mathbf{x_{2,i}}) = t(\eta(\mathbf{x_{1,i}}, \mathbf{x_{2,i}}))$. We now propose the systematic component of the generalized linear model to take the form

$$\eta(\mathbf{x_{1,i}}, \mathbf{x_{2,i}}) = \beta^{\mathbf{T}}\mathbf{x_{1,i}} + \mathbf{f}(\mathbf{x_{2,i}})$$

where $\beta^{\mathbf{T}} \in \mathbf{R^p}$ and $\mathbf{f}(\mathbf{x_{2,i}}) \sim \mathcal{GP}(\mathbf{0}, K(\mathbf{x_{2,i}}))$. The systematic component with fixed and random effects follow a linear mixed effect model with the first part capturing the fixed effect and the second part describing the randomness that is not covered by the first part. Note that $\eta(\mathbf{x_{1,i}}, \mathbf{x_{2,i}})$ does not include an intercept term on this stage. Following the same notation as previous sections, we denote $\mathbf{X}_d = (\mathbf{x}_{d,1}, \cdots, \mathbf{x}_{d,n}), d = 1, 2$ as the training dataset and $\mathbf{X}_{d*} = (\mathbf{x}_{d,n+1}, \cdots, \mathbf{x}_{d,n+n_*}), d = 1, 2$ as the testing subsets. The proposed inference method proceeds as follows.

**Stage 1.** Inference on the fixed effect.

The joint likelihood function $L(\beta|\mathbf{X_1}, \mathbf{X_2}, \mathbf{y}, \mathbf{f}(\mathbf{X_2}))$ can be written as

$$L(\beta|\mathbf{X_1}, \mathbf{X_2}, \mathbf{y}, \mathbf{f}(\mathbf{X_2})) = \prod_{i=1}^{n} t(\eta(\mathbf{x_{1,i}}, \mathbf{x_{2,i}}))^{y_i}(1 - t(\eta(\mathbf{x_{1,i}}, \mathbf{x_{2,i}})))^{1-y_i}. \tag{3.4}$$

On the first stage, we consider the latent Gaussian process $\mathbf{f}(\mathbf{X_2})$ fixed across time $i$. Numerical algorithms such as Newton-Raphson method can be used to obtain $\hat{\beta}$, the MLE of the joint likelihood function. In fact, in this stage, we regard the latent Gaussian process $\mathbf{f}(\mathbf{X_2})$ as the time-invariant intercept of the logistic regression, which is considered fixed but

unknown.

**Stage 2.** Inference on the stochastic components.

On the second stage, we make use of the result of inference on the fixed effect from Stage 1 and adjust the estimates by introducing the latent Gaussian process $\mathbf{f}(\mathbf{X_2})$. Conditional on $\hat{\beta}$, we define $\tilde{\eta}(\mathbf{x_{1,i}}, \mathbf{x_{2,i}}|\hat{\beta}) = \hat{\beta}^{\mathbf{T}}\mathbf{x_{1,i}}$, then it follows that

$$\mathsf{P}(y_i = 1|\mathbf{x_{1,i}}, \mathbf{x_{2,i}}, \hat{\beta}) = t(\tilde{\eta}(\mathbf{x_{1,i}}, \mathbf{x_{2,i}}|\hat{\beta}) + \mathbf{f}(\mathbf{x_{2i}})).$$

Here, we model the stochastic component $\mathbf{f}(\mathbf{X_2})$ as a Gaussian process with covariance function

$$Cov(\mathbf{f}(\mathbf{x_{2,i}}), \mathbf{f}(\mathbf{x_{2,j}})) = \lambda \exp(-\rho||\mathbf{x_{2,i}} - \mathbf{x_{2,j}}||^2) + \sigma^2\delta_{ij} \tag{3.5}$$

and $\delta_{ij}$ takes value 1 when $i = j$ and 0 otherwise. The parameters $\rho, \sigma$ and $\lambda$ are estimated by the strategy proposed by Section 3.3.3 and we will not specify any prior on those parameters. Since $\tilde{\eta}(\mathbf{x_{1,i}}, \mathbf{x_{2,i}}|\hat{\beta})$ is known, we can implement the strategy in Section 3.2.2 in dealing with the predictive probability from Equation (3.3). The complete hybrid method can be summarized in the following Algorithm 1.

*Remark* 5. The Hessian matrix $W$ is a diagonal matrix with the following elements for the logit and probit link respectively,

$$\begin{aligned} W_{ii} &= -p(y_i = 1|\hat{\beta}, \mathbf{f_i})p(y_i = 0|\hat{\beta}, \mathbf{f_i}), \\ W_{ii} &= -\frac{\varphi^2((2y_i - 1)(\hat{\beta}^{\mathbf{T}}\mathbf{x_{1,i}} + \mathbf{f_i}))(\hat{\beta}^{\mathbf{T}}\mathbf{x_{1,i}} + \mathbf{f_i})}{\Phi^2((2y_i - 1)(\hat{\beta}^{\mathbf{T}}\mathbf{x_{1,i}} + \mathbf{f_i}))} - \frac{(2y_i - 1)(\hat{\beta}^{\mathbf{T}}\mathbf{x_{1,i}} + \mathbf{f_i})\varphi(y_i(\hat{\beta}^{\mathbf{T}}\mathbf{x_{1,i}} + \mathbf{f_i}))}{\Phi((2y_i - 1)(\hat{\beta}^{\mathbf{T}}\mathbf{x_{1,i}} + \mathbf{f_i}))}. \end{aligned}$$

**Algorithm 1** The proposed binary hybrid method

**Stage 1.**

**Input**: $\mathbf{y}$, $K(\mathbf{X_2}, \mathbf{X_2})$(covariance matrix), $p(\mathbf{y}|\mathbf{X_1}, \mathbf{f})$ (the likelihood function)

**Compute** the MLE $\hat{\beta}$ of $L(\beta|\mathbf{X_1}, \mathbf{X_2}, \mathbf{y}, \mathbf{f}(\mathbf{X_2}))$ using Newton-Raphson method (see Equation (3.4)).

$\mathbf{f} := \mathbf{0}$ initialization

**While** (iter $<$ Max-iter)

**Repeat**

$W := -\nabla^2 \log p(\mathbf{y}|\hat{\beta}, \mathbf{f})$

$C := W * \mathbf{f} + \nabla \log p(\mathbf{y}|\hat{\beta}, \mathbf{f})$

$\mathbf{f} = (K^{-1}(\mathbf{X_2}, \mathbf{X_2}) + W)^{-1} * C$

**If** the difference of successive value of $\mathbf{f}$ is small enough, **break**

**else** continue this procedure.

**Return**: $\hat{\mathbf{f}} := \mathbf{f}$

**Stage 2.**

**Input**: $\mathbf{y}$, $\hat{\beta}$ (the estimates of coefficients of the fixed effect), $\hat{\mathbf{f}}$ (the mean of the Laplace approximation), $K(\mathbf{X_2}, \mathbf{X_2})$, $K(\mathbf{X_{2*}}, \mathbf{X_2})$, $K(\mathbf{X_{2*}}, \mathbf{X_{2*}})$(covariance matrix), $p(\mathbf{y}|\mathbf{X_1}, \mathbf{f})$(the likelihood function), $\mathbf{X_{1*}}, \mathbf{X_{2*}}$ (test input)

$W := -\nabla^2 \log p(\mathbf{y}|\hat{\beta}, \hat{\mathbf{f}})$

$\bar{\mathbf{f}}_* = K(\mathbf{X_{2*}}, \mathbf{X_2})^T \nabla \log p(\mathbf{y}|\hat{\beta}, \hat{\mathbf{f}})$

$\mathbf{v}_* = K(\mathbf{X_{2*}}, \mathbf{X_{2*}}) - K(\mathbf{X_{2*}}, \mathbf{X_2})^T W^{\frac{1}{2}} (I + W^{\frac{1}{2}} K(\mathbf{X_2}, \mathbf{X_2}) W^{\frac{1}{2}})^{-1} W^{\frac{1}{2}} K(\mathbf{X_{2*}}, \mathbf{X_2})$

$\bar{\pi}_* = \int t(\hat{\beta}^T \mathbf{X_{1*}} + z) N(z|\bar{\mathbf{f}}_*, \mathbf{v}_*) dz$

**Return**: $\bar{\pi}_*$ (the predictive probability of test input $\mathbf{X_{1*}}, \mathbf{X_{2*}}$)

In the implementation of this method, we conducted a model selection strategy on the covariance matrix $K$ based on maximum likelihood in Equation (3.6).

### 3.3.3 Model Selection

Strategies on model selection are also presented in two steps.

**Step 1.** In this study, we will use exploratory analysis to choose variables. Alternatively, we could use AIC or BIC focusing on the fixed effects. Using automatic variable selection strategies based on AIC or BIC, we can choose a model with a subset of predictors. AIC value is defined as AIC $= 2k - 2\log L$ and BIC is defined as BIC $= k\log n - 2\log L$, where $k$ is the number of parameters, $n$ is the number of observations and $L$ is the maximum value of likelihood.

**Step 2.** We select the parameters for the covariance matrix by maximum likelihood estimation. The strategy is inspired by the work of Williams and Rasmussen (2006). Our work is similar in terms of maximizing the marginal likelihood but differs in the way that the both fixed and random effects are involved.

We denote $\theta$ as the parameters in the covariance structure $Cov(\mathbf{y})$. The approximate log marginal likelihood is

$$\log q(\mathbf{y}|\mathbf{X_1}, \mathbf{X_2}, \theta) = -\frac{1}{2}\hat{\mathbf{f}}^T K^{-1}(\mathbf{X_1}, \mathbf{X_1})\hat{\mathbf{f}} + \log p(\mathbf{y}|\mathbf{X_1}, \hat{\mathbf{f}}) - \frac{1}{2}\log|B|, \tag{3.6}$$

where $B = I + W^{\frac{1}{2}}K(\mathbf{X_1}, \mathbf{X_1})W^{\frac{1}{2}}$ and $\hat{\mathbf{f}}$ is defined in Section 3.2.2. The strategy is to choose the value of $\theta$ that maximizes Equation (3.6). Note that the covariance matrix $K$ ($K(\mathbf{X_1}, \mathbf{X_1})$)and $\hat{\mathbf{f}}$ involve parameters $\theta$, the partial derivative of $\frac{\partial \log q(\mathbf{y}|\mathbf{X_1}, \mathbf{X_2}, \theta)}{\partial \theta_j}$ is therefore

$$\frac{\partial \log q(\mathbf{y}|\mathbf{X_1}, \mathbf{X_2}, \theta)}{\partial \theta_j} = A + B,$$

where $A$ and $B$ are defined as follows

$$A = \frac{1}{2}\hat{\mathbf{f}}^T K^{-1}\frac{\partial K}{\partial \theta_j}K^{-1}\hat{\mathbf{f}} - \frac{1}{2}tr((W^{-1}+K)^{-1}\frac{\partial K}{\partial \theta_j}),$$
$$B = \sum_{i=1}^{n} -\frac{1}{2}[(K^{-1}+W)^{-1}]_{ii}\frac{\partial^3}{\partial f_i^3}\log p(\mathbf{y}|\mathbf{X_1},\hat{\mathbf{f}})[(I+KW)^{-1}\frac{\partial K}{\partial \theta_j}\nabla \log p(\mathbf{y}|\mathbf{X_1},\hat{\mathbf{f}})]_i.$$

Newton-Raphson method or coordinate descent will be applied to optimize the log marginal likelihood in Equation (3.6).

In this study, the parameters $\theta$ from Equations (3.5) are $\rho, \sigma$ and $\lambda$. Through our simulation studies, we specify the parameters $\sigma$ and $\rho$ and apply the aforementioned strategy on estimating $\lambda$ for the following reasons: (1) it might lead to identifiability problem if we do not fix some of the parameters in this frequentist setting; (2) results do not show much difference if parameters $\sigma$ and $\rho$ are not fixed; (3) computation will be demanding if no parameter is fixed.

### 3.3.4 Inference on the Effects of Covariates

We propose to use bootstrap sampler to provide point and confidence intervals of the linear coefficients of the covariates $\mathbf{X_1}$. Our approach is based on resampling the stochastic component and maximum likelihood. The algorithm is summarized in Algorithm 2.

### 3.3.5 Summary

In summary, the proposed method on inference, prediction and model selection maintain the following strengths: (1.) it uses linear and non-linear stochastic components to model the effect of the covariates on the response; (2.) it provides point and interval estimates

**Algorithm 2** Inference on the linear effects

**Input**: $\mathbf{y}$, $\hat{K}$(the estimated covariance matrix derived from Section 3.3.3), $\hat{\beta}$ (the estimates of coefficients of the fixed effect derived from Stage 1 in Section 3.3.2)
**Procedure**:
$\tilde{\eta}(\mathbf{X_1}) := \hat{\beta}^T \mathbf{X_1}$
**While** (Iter < Max-iter)
**Repeat**
**Generate** $\mathbf{f^{(iter)}}(\mathbf{X_2})$ from $\mathcal{GP}$ with covariance function $\hat{K}$
$\eta^{(\mathbf{iter})}(\mathbf{X_1}, \mathbf{X_2}) := \tilde{\eta}(\mathbf{X_1}) + \mathbf{f^{(iter)}}(\mathbf{X_2})$
**Compute** the MLE $\hat{\beta}^{(\mathbf{iter})}$ of $L(\beta|\mathbf{X_1}, \mathbf{X_2}, \mathbf{y}, \mathbf{f}(\mathbf{X_2}))$ using Newton-Raphson method, where $L(\beta|\mathbf{X_1}, \mathbf{X_2}, \mathbf{y}, \mathbf{f}(\mathbf{X_2})) = \prod_{i=1}^{n} t(\eta^{(\mathbf{iter})}(\mathbf{x_{1,i}}, \mathbf{x_{2,i}}))^{y_i}(1 - t(\eta^{(\mathbf{iter})}(\mathbf{x_{1,i}}, \mathbf{x_{2,i}})))^{1-y_i}$
**End of while**
**Compute** $\hat{\beta}_* = \frac{1}{\text{Max-iter}} \sum_{i=1}^{\text{Max-iter}} \hat{\beta}^{(i)}$
$\hat{\beta}_{0.025}$ = 2.5-th percentile of $\{\hat{\beta}^{(i)}\}_{i=1}^{\text{Max-iter}}$
$\hat{\beta}_{0.975}$ = 97.5-th percentile of $\{\hat{\beta}^{(i)}\}_{i=1}^{\text{Max-iter}}$
**Return:** $\hat{\beta}_*$ (The point estimates of the parameters from linear effects); $(\hat{\beta}_{0.025}, \hat{\beta}_{0.975})$ (The 95% bootstrap confidence interval of the parameters from linear effects).

of the linear effects that are more efficient than the existing methods as demonstrated in Section 3.4; (3.) it is able to make accurate predictions as shown in Section 3.4; (4.) the computational cost is not demanding; (5.) similarly to generalized linear models, it provides results that are straightforward to interpret.

## 3.4 Simulations

In this section, simulations are implemented to test the performance of the proposed method. In Section 3.4.1, binary time series $\mathbf{y_i}$ are generated by the logit model. We compared the classification error rates derived from the proposed method with 6 other competing statistical and machine learning approaches, namely, the ordinal model, logistic regression, generalized additive mixed model, random forest, decision tree and gradient boosting. We also compute the point and confidence intervals of the coefficients of the covariates and endogenous variables in comparison with other existing methods. To test the robustness

of our method, in Section 3.4.2, we generate time series $\mathbf{y_i}$ from the probit model but use the logit model to fit the data. In Section 3.4.3, we utilize mixture kernels to generate the Gaussian process and then apply the proposed HIBITS method. Classification error rates, point estimates and confidence intervals are also utilized as measures for comparison.

### 3.4.1 Prediction and Inference Performance on Logit Model

To evaluate the prediction power and robustness of the proposed method, binary time series $y_i$ are generated under two scenarios:

- **Scenario 1 (with a stochastic process).**
  $\mathsf{P}(y_i = 1) = logit^{-1}(\beta_0 x_{1i} + \beta_1 y_{i-1} + \mathbf{f}(x_{2i}))$;
- **Scenario 2 (without a stochastic process).**
  $\mathsf{P}(y_i = 1) = logit^{-1}(\beta_0 x_{1i} + \beta_1 y_{i-1})$.

Here, $\mathbf{f}(\mathbf{x_2})$ follows Gaussian process with

$$Cov(\mathbf{f}(x_{2i}), \mathbf{f}(x_{2j})) = \lambda \exp(-\rho(x_{2i} - x_{2j})^2) + \sigma^2 \delta_{ij}$$

and $\delta_{ij}$ takes value 1 when $i = j$ and 0 otherwise. The parameter $\beta_1$ controls the strength of dependence on previous realizations $y_{i-1}$ and it denotes the log odds ratios of $y_{i-1} = 1$ versus $y_{i-1} = 0$. $\beta_0$ is the linear coefficients with respect to covariates at current time point. $\lambda$ is the parameter that determines the strength of dependence across adjacent time points. In this simulation, parameters $\beta = (\beta_0, \beta_1)$ and $\lambda$ vary in different scenarios. 1000 simulations are conducted in each scenario. Figure 3.2 shows plots of the simulated data. In this scenario,

$\beta = (0.5, 4), \lambda = 1, \rho = 1, \sigma = 0.1$. 500 sleep stages were generated.

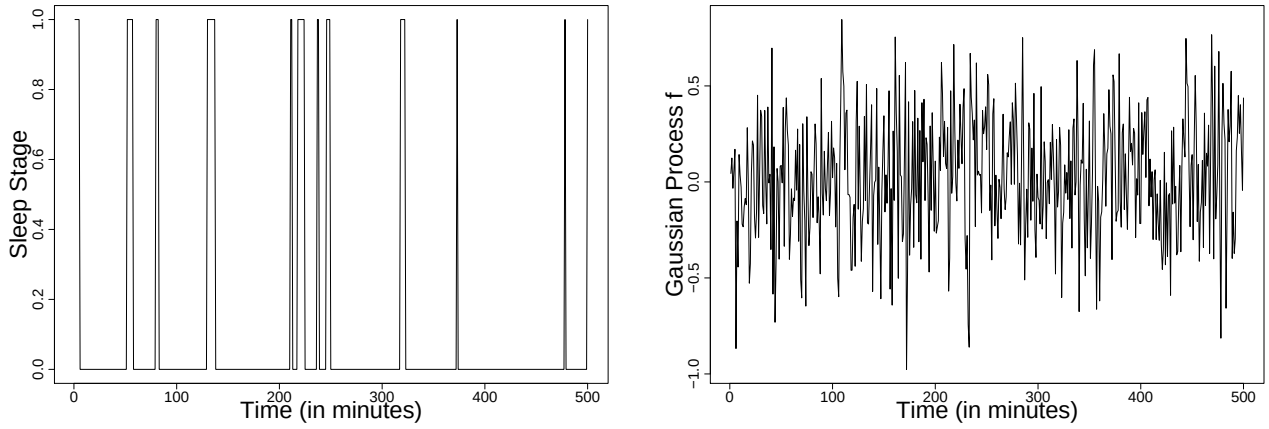



Figure 3.2: Plots of the generated sleep stage (left) and the simulated Gaussian process(right).

**Alternative Methods.** To evaluate the prediction power of the proposed method, we compare the classification error rates with other six competing approaches. In general, those approaches include regression and tree based classification strategies. Generalized linear model with logit link is fitted as the first competing method. Further, to respect the correlated structure of the binary time series, we consider the generalized additive mixed models (GAMMs) as the second regression based competing approach. In the work of Lin and Zhang (1999), linear structures of covariates are extended to be smooth functions. Following the notation in Section 3.3, the GAMM model is defined as

$$\eta(\mathbf{x_{1,i}}) = \beta_0 + f_1(x^1_{1,i}) + \cdots + f_p(x^p_{1,i}) + \mathbf{z_i^T b_i},$$

where $x^j_{1,i}$ denotes the $j^{th}$ component of vector $\mathbf{x_{1,i}}$, $f_j(.)$ is a centered twice-differentiable smooth function, the random effects $\mathbf{b}$ are assumed to be distributed as $N(0, D(\theta))$ and $\theta$ is the variance components. Lin and Zhang (1999) estimated nonparametric functions and

parameters by using smoothing splines and marginal quasi-likelihood. In this simulation, R package 'gamm4' was implemented to test the performance of GAMMs. We also considered the regression models for nominal and ordinal time series introduced by Fokianos and Kedem (2002). As is discussed in their concrete work, we implemented ordinal time series model in the simulation. It should be pointed out that due to the binary response, ordinal time series model is degenerated into logistic regression. Simulation results also suggest the equivalence of these two approaches. In addition, we compared our method to tree-based classification approaches. In general, we split the feature space (heart rate and previous sleep states in this study) into "subspaces" and fit simple models within each region. Following the derivation in Friedman et al. (2001), for a node $m$ denoting a region $R_m$ with $N_m$ observations, we let

$$\hat{p}_{mk} = \frac{1}{N_m} \sum_{\mathbf{x_1},\mathbf{i} \in R_m} \mathbf{1}(y_i = k),$$

where class $k$ is either 0 or 1 and $\mathbf{1}$ is the indicator function. We assign the observations in node $m$ to class $k(m) = \arg\max_k \hat{p}_{mk}$. Measures of node impurity, denoted as $Q_m(T)$, can be chosen as the misclassification error, Gini index and cross-entropy or deviance.

To further extend the decision tree approach, we also consider random forest and gradient boosting algorithms in the simulation. The essence of random forest is to average many noisy but asymptotically unbiased classifiers and hence reduce the variation. It requires bootstrapping samples and selection features from the training dataset. Since there exist only a few features in this model, the benefit from using random forest approach is mainly derived from the bootstrapping strategy. For each bootstrap training sample set, we grow a random forest tree $T_b, b = 1, \cdots, B$. The final output is the ensemble of trees and then predictions are made by majority vote. In addition to random forest, gradient boosting is another extension of decision tree based method. Similar to the general boosting methods, gradient boosting searches for a strategy to combine multiple weak classifiers in an iterative

manner. As discussed in Friedman (2001) and Friedman et al. (2001), the method generically starts from a model with constant value. At iteration $m(1 \leqslant m \leqslant M)$, suppose the classifier is denoted as $F_{m-1}(\mathbf{x_1}, \mathbf{x_2})$, we calculate pseudo-residuals by

$$r_{im} = -\Big[\frac{\partial L(y_i, F(\mathbf{x_{1,i}}, \mathbf{x_{2,i}}))}{\partial F(\mathbf{x_{1,i}}, \mathbf{x_{2,i}}))}\Big]_{F(\mathbf{x_{1,i}}, \mathbf{x_{2,i}}) = F_{m-1}(\mathbf{x_{1,i}}, \mathbf{x_{2,i}})},$$

where $L(y, F(x))$ is a loss function. Then, we fit a classifier $h_m(x)$ to the pseudo-residuals and implement a line search algorithm in solving the optimization problem

$$\gamma_m = \arg\min_{\gamma} \sum_{i=1}^{n} L(y_i, F_{m-1}(\mathbf{x_{1,i}}, \mathbf{x_{2,i}}) + \gamma h_m(\mathbf{x_{1,i}}, \mathbf{x_{2,i}})).$$

At the end of this iteration, we update the model by

$$F_m(\mathbf{x_{1,i}}, \mathbf{x_{2,i}}) = F_{m-1}(\mathbf{x_{1,i}}, \mathbf{x_{2,i}}) + \gamma_m h_m(\mathbf{x_{1,i}}, \mathbf{x_{2,i}}).$$

We keep repeating the full sweep until convergence. The final classifier is denoted as $F_M(\mathbf{x_{1,i}}, \mathbf{x_{2,i}})$.

**Model Evaluation.** To formally evaluate the performance of all the aforementioned approaches, we calculate the classification error rates under both scenarios. In particular, we fit the results in linear mixed effect model to account for the correlation among classification errors across different methods that result from the same simulated dataset. We consider the model

$$E_{ij} = \mu_i + z_j + \epsilon_{ij},$$

where $E_{ij}$ denotes the classification error rate of approach $i$ on dataset $j$; $\mu_i$ is the mean error rate of method $i$, which is well-defined by the law of large numbers. $z_j \overset{iid}{\sim} N(0, \sigma^2)$, $\epsilon_{ij} \overset{iid}{\sim}$

$N(0, \tau^2)$, $i = 1, \cdots, 6$ and $j = 1, \cdots, 1000$. We calculate the simultaneuous 95% Bonferroni confidence intervals of $(\mu_1 - \mu_i), i = 2, \cdots, 6$ to detect the difference in the mean error rates between the proposed method with all the other approaches. In particular, $\mu_1, \cdots, \mu_6$ denote the mean error rates of HIBITS, Ordinal model (logistic regression), GAMMs, Random forest, Gradient boosting and Decision tree respectively.

Table 3.1 provides a summary of the simulation studies for various parameters. It can be seen that for datasets with Gaussian process, there is statistically significant difference in comparison with the competing methods. In particular, the proposed HIBITS method produces significantly lower prediction error rates compared to existing methods. The advantage of the proposed approach is even more obvious when compared with gradient boosting and decision tree approaches. The results show that the proposed HIBITS method captures effective information from covariates $x_{1i}$, $\mathbf{y_{i-1}}$ and also the stochastic process. The covariate $\mathbf{y_{i-1}}$ serves as a significant predictor as we increase the ratio of $\beta_1$ over $\beta_0$.

For the datasets generated *without* the Gaussian process (Scenario 2) shown in Table 3.2, the accuracy prediction from the two-stage approach is significantly higher than some of the existing approaches such as decision tree and gradient boosting. Among all the other competitors, the proposed method behaves equally competitive. This shows the robustness of the proposed approach when data have no Gaussian process components. This is partly due to the strategy on choosing hyperparemeters. By controlling their values, the effects of Gaussian process will be adjusted to the data.

Table 3.1: Summary of simulation results. $\mu_1, \cdots, \mu_6$ denote the mean error rates of HIBITS, Ordinal model (logistic regression), GAMMs, Random forest, Gradient boosting and Decision tree respectively. 1000 simulated datasets were generated under the scenario: $\mathsf{P}(y_i = 1) = logit^{-1}(\beta_0 x_{1i} + \beta_1 y_{i-1} + \mathbf{f}(x_{2i}))$ ("Scenario 1"). We calculated the 95% Bonferroni-corrected confidence intervals of the prediction error difference from the testing dataset, $\mu_1 - \mu_i, i = 2, \cdots, 6$ that the classification error rate for the proposed method is lower than that for each of competing methods.

| Parameters$(\beta, \lambda)$ | Competing Method | Scenario 1 |
|---|---|---|
| | | 95% confidence interval of $\mu_1 - \mu_i, i = 2, \cdots, 7$ |
| $\beta = (0.5, 3), \lambda = 10$ | Ordinal model* | $(-0.052 \quad , \quad -0.032)$ |
| | GAMMs | $(-0.052 \quad , \quad -0.032)$ |
| | Random forest | $(-0.029 \quad , \quad -0.009)$ |
| | Gradient boosting | $(-0.075 \quad , \quad -0.055)$ |
| | Decision tree | $(-0.070 \quad , \quad -0.050)$ |
| $\beta = (0.5, 3), \lambda = 5$ | Ordinal model | $(-0.015 \quad , \quad -0.001)$ |
| | GAMMs | $(-0.017 \quad , \quad -0.001)$ |
| | Random forest | $(-0.017 \quad , \quad -0.002)$ |
| | Gradient boosting | $(-0.038 \quad , \quad -0.022)$ |
| | Decision tree | $(-0.048 \quad , \quad -0.032)$ |
| $\beta = (0.5, 3.5), \lambda = 10$ | Ordinal model | $(-0.036 \quad , \quad -0.013)$ |
| | GAMMs | $(-0.030 \quad , \quad -0.011)$ |
| | Random forest | $(-0.021 \quad , \quad -0.001)$ |
| | Gradient boosting | $(-0.046 \quad , \quad -0.028)$ |
| | Decision tree | $(-0.055 \quad , \quad -0.037)$ |
| $\beta = (0.5, 3.5), \lambda = 5$ | Ordinal model | $(-0.010 \quad , \quad -0.001)$ |
| | GAMMs | $(-0.011 \quad , \quad -0.001)$ |
| | Random forest | $(-0.015 \quad , \quad -0.001)$ |
| | Gradient boosting | $(-0.020 \quad , \quad -0.006)$ |
| | Decision tree | $(-0.035 \quad , \quad -0.021)$ |

* For binary time series, ordinal model is equivalent to logistic regression.

Table 3.2: Summary of simulation results. $\mu_1, \cdots, \mu_6$ denote the mean error rates of HIBITS, Ordinal model (logistic regression), GAMMs, Random forest, Gradient boosting and Decision tree respectively. 1000 simulated datasets were generated under the scenario: $\mathsf{P}(y_i = 1) = logit^{-1}(\beta_0 x_{1i} + \beta_1 y_{i-1})$ ("Scenario 2"). We calculated the 95% Bonferroni-corrected confidence intervals of the prediction error difference from the testing dataset, $\mu_1 - \mu_i, i = 2, \cdots, 6$ that the classification error rate for the proposed method is lower than that for each of competing methods.

| | | Scenario 2 |
|---|---|---|
| Parameters($\beta$) | Competing Method | 95% confidence interval of $\mu_1 - \mu_i, i = 2, \cdots, 7$ |
| $\beta = (0.5, 3)$ | Ordinal model | $(-0.006 \quad , \quad +0.010)$ |
| | GAMMs | $(-0.005 \quad , \quad +0.009)$ |
| | Random forest | $(-0.004 \quad , \quad +0.012)$ |
| | Gradient boosting | $(-0.022 \quad , \quad -0.001)$ |
| | Decision tree | $(-0.023 \quad , \quad -0.002)$ |
| $\beta = (0.5, 3.5)$ | Ordinal model | $(-0.003 \quad , \quad +0.010)$ |
| | GAMMs | $(-0.002 \quad , \quad +0.010)$ |
| | Random forest | $(-0.015 \quad , \quad -0.001)$ |
| | Gradient boosting | $(-0.020 \quad , \quad -0.006)$ |
| | Decision tree | $(-0.018 \quad , \quad -0.001)$ |

We also evaluate the performance of modeling the linear effects of covariates $x_{1i}$, $\mathbf{y_{i-1}}$ by comparing the 95% confidence intervals of $\beta_0$ and $\beta_1$ with the corresponding interval estimates from the other existing methods. Table 3.3 summarizes the results under the same scenarios in Table 3.1. It shows that compared with ordinal model, the proposed HIBITS method produces narrower confidence intervals of parameters $\beta_0$ while maintaining high capture rates of the true values. The length difference is obvious and it can gain almost 60% shorter confidence intervals in some scenario. It should be noted that using ordinal model, the capture rate of $\beta_1$ is extremely low while HIBITS method provides promising performance.

The same pattern can also be found in Table 3.4. Under Scenario 2, the benefits of using HIBITS is even more obvious in terms of shorter confidence interval length and high capture rate.

Table 3.3: Summary of simulation results. 1000 simulations were generated under the scenario: $\mathsf{P}(y_i = 1) = logit^{-1}(\beta_0 x_{1i} + \beta_1 y_{i-1} + \mathbf{f}(x_{2i}))$ ("Scenario 1" ). We present the 95% confidence intervals $\beta_0$ and $\beta_1$ from the training dataset.

| Parameters$(\beta, \lambda)$ | Method | Scenario 1 | |
|---|---|---|---|
| | | 95% confidence interval of | |
| | | $\beta_0$ | $\beta_1$ |
| $\beta = (0.5, 3), \lambda = 10$ | HIBITS method | $(0.113, 0.547)$ | $(1.385, 3.424)$ |
| | Ordinal model | $(-0.292, 0.586)$ | $(0.695, 2.473)$ |
| $\beta = (0.5, 3), \lambda = 5$ | HIBITS method | $(0.163, 0.572)$ | $(1.570, 3.700)$ |
| | Ordinal model | $(-0.267, 0.637)$ | $(0.850, 2.676)$ |
| $\beta = (0.5, 3.5), \lambda = 10$ | HIBITS method | $(0.092, 0.535)$ | $(1.628, 4.082)$ |
| | Ordinal model | $(-0.358, 0.625)$ | $(0.806, 2.593)$ |
| $\beta = (0.5, 3.5), \lambda = 5$ | HIBITS method | $(0.182, 0.582)$ | $(1.820, 3.985)$ |
| | Ordinal model | $(-0.286, 0.694)$ | $(0.991, 2.841)$ |

Table 3.4: Summary of simulation results. 1000 simulations were generated under the scenario: $\mathsf{P}(y_i = 1) = logit^{-1}(\beta_0 x_{1i} + \beta_1 y_{i-1})$ ("Scenario 2" ). We present the 95% confidence intervals $\beta_0$ and $\beta_1$ from the training dataset.

| Parameters$(\beta)$ | Method | Scenario 2 | |
|---|---|---|---|
| | | 95% confidence interval of | |
| | | $\beta_0$ | $\beta_1$ |
| $\beta = (0.5, 3)$ | HIBITS method | $(0.467, 0.600)$ | $(2.838, 3.173)$ |
| | Ordinal model | $(-0.177, 1.420)$ | $(1.677, 4.515)$ |
| $\beta = (0.5, 3.5)$ | HIBITS method | $(0.422, 0.556)$ | $(3.468, 3.771)$ |
| | Ordinal model | $(-0.081, 1.202)$ | $(2.275, 5.096)$ |

Overall, the proposed method outperforms competing approaches when comparing the results from data both with and without Gaussian process. Through the model selection strategy discussed in Section 3.3.3, the proposed approach can adjust the covariance matrix to the data, which in return produces lower prediction error rate and more efficients inference on covariates than existing methods.

### 3.4.2 Investigating Robustness of the Estimation Method

Our goal is to investigate robustness of the proposed model by applying the logistic-based model on data that are generated using a probit model. We generate binary time series $\mathbf{y_i}$ following the scenarios:

- **Scenario 3 (with a stochastic process).**
  $\mathsf{P}(y_i = 1) = \Phi(\beta_0 x_{1i} + \beta_1 y_{i-1} + \mathbf{f}(x_{2i}))$;
- **Scenario 4 (without a stochastic process).**
  $\mathsf{P}(y_i = 1) = \Phi(\beta_0 x_{1i} + \beta_1 y_{i-1})$. $\Phi(.)$ is the cumulative distribution function of standard normal distributions and $\mathbf{f}(\mathbf{x_2})$ is defined in the same manner as in Section 3.4.1.

Parameters $\beta = (\beta_0, \beta_1)$ and $\lambda$ vary in different scenarios. 1000 simulations are conducted in each scenario. We fit the same linear mixed effect model discussed in Section 3.4.1. Tables 3.5 and 3.6 show the summary of confidence intervals $\mu_1 - \mu_i, i = 2, \cdots, 6$. Similar to the results in Section 3.4.1, for dataset with Gaussian process, most of the confidence intervals do not cover 0. The negative values of the classification error rates imply remarkable benefits of using the proposed method over the other competing methods. Note that when comparing with the gradient boosting and decision tree approaches, the proposed method behaves significantly better in terms of extraordinary higher prediction accuracy. In Scenario 4, we

tested the proposed method on the dataset without Gaussian process. It is shown that although there is no significant difference in comparison with other competing methods, the proposed approach produces the same prediction power as other competing methods, which implies the robustness with regard to various link functions. In addition, Table 3.7 shows the 95% confidence intervals of the coefficients $\beta_0, \beta_1$ derived from the proposed method and the ordinal model. Similar to the results in Section 3.4.1, the proposed method yields much narrower confidence intervals while maintaining good properties of capturing true values.

Table 3.5: Summary of simulation results. $\mu_1, \cdots, \mu_6$ denote the mean error rates of HIBITS, Ordinal model (logistic regression), GAMMs, Random forest, Gradient boosting and Decision tree respectively. 1000 simulated datasets were generated under the scenario: $\mathsf{P}(y_i = 1) = \Phi(\beta_0 x_{1i} + \beta_1 y_{i-1} + \mathbf{f}(x_{2i}))$ ("Scenario 3"). We calculated the 95% Bonferroni-corrected confidence intervals of the prediction error difference from the testing dataset, $\mu_1 - \mu_i, i = 2, \cdots, 6$ that the classification error rate for the proposed method is lower than that for each of competing methods.

| Parameters($\beta, \lambda$) | Competing Method | Scenario 3 |
|---|---|---|
| | | 95% confidence interval of $\mu_1 - \mu_i, i = 2, \cdots, 7$ |
| $\beta = (0.5, 3), \lambda = 10$ | Ordinal model | $(-0.042 \quad , \quad -0.023)$ |
| | GAMMs | $(-0.041 \quad , \quad -0.022)$ |
| | Random forest | $(-0.025 \quad , \quad -0.006)$ |
| | Gradient boosting | $(-0.064 \quad , \quad -0.045)$ |
| | Decision tree | $(-0.060 \quad , \quad -0.041)$ |
| $\beta = (0.5, 3), \lambda = 5$ | Ordinal model | $(-0.015 \quad , \quad -0.002)$ |
| | GAMMs | $(-0.016 \quad , \quad -0.002)$ |
| | Random forest | $(-0.021 \quad , \quad -0.007)$ |
| | Gradient boosting | $(-0.024 \quad , \quad -0.009)$ |
| | Decision tree | $(-0.040 \quad , \quad -0.026)$ |
| $\beta = (0.5, 3.5), \lambda = 10$ | Ordinal model | $(-0.030 \quad , \quad -0.001)$ |
| | GAMMs | $(-0.029 \quad , \quad -0.007)$ |
| | Random forest | $(-0.006 \quad , \quad +0.026)$ |
| | Gradient boosting | $(-0.057 \quad , \quad -0.035)$ |
| | Decision tree | $(-0.024 \quad , \quad +0.002)$ |
| $\beta = (0.5, 3.5), \lambda = 5$ | Ordinal model | $(-0.014 \quad , \quad +0.001)$ |
| | GAMMs | $(-0.013 \quad , \quad +0.001)$ |
| | Random forest | $(-0.019 \quad , \quad -0.002)$ |
| | Gradient boosting | $(-0.120 \quad , \quad -0.008)$ |
| | Decision tree | $(-0.031 \quad , \quad -0.014)$ |

Table 3.6: Summary of simulation results. $\mu_1, \cdots, \mu_6$ denote the mean error rates of HIBITS, Ordinal model (logistic regression), GAMMs, Random forest, Gradient boosting and Decision tree respectively. 1000 simulated datasets were generated under the scenario: $\mathsf{P}(y_i = 1) = \Phi(\beta_0 x_{1i} + \beta_1 y_{i-1})$ ("Scenario 4" ). We calculated the 95% Bonferroni-corrected confidence intervals of the prediction error difference from the testing dataset, $\mu_1 - \mu_i, i = 2, \cdots, 6$ that the classification error rate for the proposed method is lower than that for each of competing methods.

| | | Scenario 4 |
|---|---|---|
| Parameters($\beta$) | Competing Method | 95% confidence interval of $\mu_1 - \mu_i, i = 2, \cdots, 7$ |
| $\beta = (0.5, 3)$ | Ordinal model | $(-0.003 \quad , \quad +0.015)$ |
| | GAMMs | $(-0.006 \quad , \quad +0.005)$ |
| | Random forest | $(-0.002 \quad , \quad +0.015)$ |
| | Gradient boosting | $(-0.012 \quad , \quad +0.009)$ |
| | Decision tree | $(-0.016 \quad , \quad +0.008)$ |
| $\beta = (0.5, 3.5)$ | Ordinal model | $(-0.003 \quad , \quad +0.007)$ |
| | GAMMs | $(-0.002 \quad , \quad +0.008)$ |
| | Random forest | $(-0.005 \quad , \quad +0.011)$ |
| | Gradient boosting | $(-0.010 \quad , \quad +0.016)$ |
| | Decision tree | $(-0.011 \quad , \quad +0.002)$ |

Table 3.7: Summary of simulation results. 1000 simulations were generated under the scenario: $\mathsf{P}(y_i = 1) = \Phi(\beta_0 x_{1i} + \beta_1 y_{i-1} + \mathbf{f}(x_{2i}))$ ("Scenario 3"). We present the 95% confidence intervals $\beta_0$ and $\beta_1$ from the training dataset.

| Parameters$(\beta, \lambda)$ | Method | Scenario 3 | |
|---|---|---|---|
| | | 95% confidence interval of | |
| | | $\beta_0$ | $\beta_1$ |
| $\beta = (0.5, 3), \lambda = 10$ | HIBITS method | $(0.129, 0.564)$ | $(1.529, 3.574)$ |
| | Ordinal model | $(-0.247, 0.564)$ | $(0.756, 2.549)$ |
| $\beta = (0.5, 3), \lambda = 5$ | HIBITS method | $(0.191, 0.502)$ | $(1.784, 3.956)$ |
| | Ordinal model | $(-0.273, 0.668)$ | $(0.966, 2.831)$ |
| $\beta = (0.5, 3.5), \lambda = 10$ | HIBITS method | $(0.129, 0.579)$ | $(1.766, 3.871)$ |
| | Ordinal model | $(-0.406, 0.734)$ | $(0.875, 2.678)$ |
| $\beta = (0.5, 3.5), \lambda = 5$ | HIBITS method | $(0.200, 0.509)$ | $(2.111, 4.310)$ |
| | Ordinal model | $(-0.239, 0.666)$ | $(1.156, 3.046)$ |

### 3.4.3 Investigating the Misspecification of the Covariance Function

The objective of this section is to study the effects of misspecification on the covariance function. We will assume the true covariance function follows mixtures of different kernels and apply the proposed HIBITS method to the generated dataset. In particular, we generate binary time series $y_i$ under the following scenario:

- **Scenario 5 (with a mixture covariance function).**

  $\mathsf{P}(y_i = 1) = logit^{-1}(\beta_0 x_{1i} + \beta_1 y_{i-1} + \mathbf{f}(x_{2i}))$.

Here, $\mathbf{f}(x_{2i}))$ follows Gaussian process with

$$Cov(\mathbf{f}(x_{2i}), \mathbf{f}(x_{2j})) = \eta\Big[\lambda \exp(-\rho(x_{2i} - x_{2j})^2) + \sigma^2\delta_{ij}\Big] + (1-\eta)\Big[\frac{1}{1+\tau(x_{2i}-x_{2j})^2}\Big].$$

Note that we assume the covariance function is a mixture of exponential and Cauchy kernels. This setting serves as an approach of modeling the long-term and short-term correlation on $\boldsymbol{x_2}$. By increasing the value of trade-off parameter $\eta$, the mixture kernel will weight more on the exponential kernel, which captures the short-term dependence. Table 3.8 summarizes the results of mean error rates under Scenario 5. It is shown that the proposed HIBITS is able to maintain significant lower error rates compared to the other competing methods when the trade-off parameter $\eta$ is 0.2. As we increase the value to be 0.8, HIBITS performs almost as good as all the other methods and significantly better than decision tree. Table 3.9 presents the confidence intervals in Scenario 5. Similar to the previous results, HIBITS is capable of yielding narrower intervals and high capture rates even when the trade-off parameter $\eta$ is large. In summary, through this section, simulation results show that the proposed HIBITS method is robust to the misspecification of covariance function. This is partly due to the fact that we are able to dynamically "learn" the hyperparameter through model selection. The fine-tuned covariance function could capture the long-term and short-term correlation from the generated dataset.

Table 3.8: Summary of simulation results. $\mu_1, \cdots, \mu_6$ denote the mean error rates of HIBITS, Ordinal model (logistic regression), GAMMs, Random forest, Gradient boosting and Decision tree respectively. 1000 simulated datasets were generated under the scenario: $\mathsf{P}(y_i = 1) = logit^{-1}(\beta_0 x_{1i} + \beta_1 y_{i-1} + \mathbf{f}(x_{2i}))$ ("Scenario 5"). We calculated the 95% Bonferroni-corrected confidence intervals of the prediction error difference from the testing dataset, $\mu_1 - \mu_i, i = 2, \cdots, 6$ that the classification error rate for the proposed method is lower than that for each of competing methods.

| Parameters($\beta$, $\eta$) | Competing Method | Scenario 5 |
|---|---|---|
| | | 95% confidence interval of $\mu_1 - \mu_i, i = 2, \cdots, 7$ |
| $\beta = (0.5, 3), \eta = 0.2$ | Ordinal model | $(-0.024 \quad , \quad -0.011)$ |
| | GAMMs | $(-0.023 \quad , \quad -0.010)$ |
| | Random forest | $(-0.022 \quad , \quad -0.002)$ |
| | Gradient boosting | $(-0.038 \quad , \quad -0.023)$ |
| | Decision tree | $(-0.061 \quad , \quad -0.037)$ |
| $\beta = (0.5, 3), \eta = 0.8$ | Ordinal model | $(-0.008 \quad , \quad +0.005)$ |
| | GAMMs | $(-0.008 \quad , \quad +0.007)$ |
| | Random forest | $(-0.005 \quad , \quad +0.001)$ |
| | Gradient boosting | $(-0.004 \quad , \quad +0.001)$ |
| | Decision tree | $(-0.016 \quad , \quad -0.002)$ |

Table 3.9: Summary of simulation results. 1000 simulations were generated under the scenario: $\mathsf{P}(y_i = 1) = logit^{-1}(\beta_0 x_{1i} + \beta_1 y_{i-1} + \mathbf{f}(x_{2i}))$ ("Scenario 5").We present the 95% confidence intervals $\beta_0$ and $\beta_1$ from the training dataset.

| Parameters$(\beta, \eta)$ | Method | Scenario 5 | |
|---|---|---|---|
| | | 95% confidence interval of | |
| | | $\beta_0$ | $\beta_1$ |
| $\beta = (0.5, 3), \eta = 0.2$ | HIBITS method | $(0.056, 0.505)$ | $(1.796, 3.306)$ |
| | Ordinal model | $(-0.232, 0.671)$ | $(0.829, 2.769)$ |
| $\beta = (0.5, 3), \eta = 0.8$ | HIBITS method | $(0.160, 0.702)$ | $(2.803, 3.309)$ |
| | Ordinal model | $(-0.333, 1.142)$ | $(1.186, 6.428)$ |

## 3.5 Analysis of the Sleep State Data

In this section, we apply our method to sleep state data. People spend one third of their lifetime on sleep. Studying and predicting sleep patterns is significant because our body requires sleep in much the same way as the need of eating and breathing. Moreover, disruptions in sleep are known to be associated with both psychiatric and chronic diseases. In what follows, we will analyze the sleep data obtained from an observational study with the goal predicting sleep states and identifying associations between sleep states and potential regulators such as temperature and heart rate.

### 3.5.1 Exploratory Analysis

The data were recorded from a four month old infant who was placed to bed at night. Heart rate ($H_i$, beats per minute at time $i$), temperature ($T_i$, in Celsius, at time $i$) and sleep stage ($S_i$ at time $i$) of length ($N = 1024$) were sampled every 30 seconds. Heart

rate was recorded automatically using a standard ECG (electrocardiogram) monitor. The infant's EEG (electroencephalogram) and EOG (electrooculogram) were also measured with a period of 30 seconds. The EEG captured brain waves including alpha (8 – 15 Hz), beta (16 – 31 Hz) and mu (8 – 12 Hz) rhythms; EOG recorded the eye movement. Sleep stage for each time point $i$ was determined by the sleep lab expert visually interpreting the EEG and EOG record (Nevsimalova and Sonka, 1997). It was classified as 4 categories: (1) quiet sleep, (2) indeterminate sleep, (3) active sleep and (4) awake (Benbadis, 2006). The sleep stage $S_i$ was measured as integers ranging from 1 to 4. In this section, following the work of Fokianos and Kedem (2002), sleep state is defined as a binary time series $Y_i$:

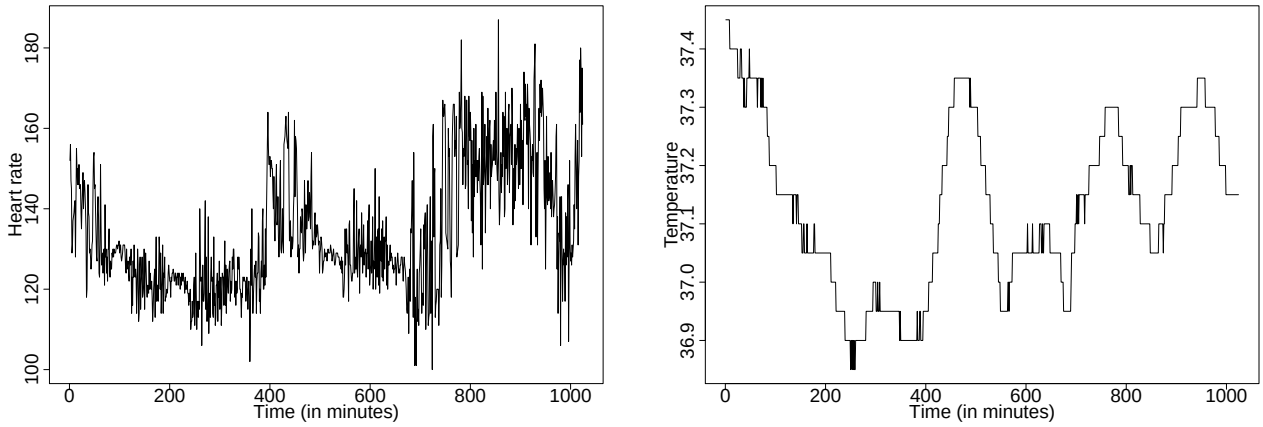
$$Y_i = \begin{cases} 1 & : \text{awake at time } i, \\ 0 & : \text{not awake at time } i. \end{cases}$$

where "not awake" stands for quiet sleep, indeterminate sleep or active sleep.

Time series plots of heart rate, temperature and sleep state are shown in Figure 3.1 and Figure 3.3. By comparing the heart rate, temperature with sleep state, we note that higher heart rate are likely to correspond to sleep state 1 (awake). While this pattern is clear for heart rate, no such pattern between temperature and sleep state can be detected by visual inspection. In addition, it can be seen that the current sleep state is highly related to previous states.

To further study the dependence of sleep state on the covariates temperature and heart rate, we conducted additional preliminary analysis. Particularly, we categorize heart rate and temperature (after taking the logarithm) into several levels and calculate the empirical log odds of awake over not-awake for each level. Figure 3.4 show the relationship between the empirical log odds and different levels of the underlined heart rate and temperature. We are able to identify a positive association between heart rate with current sleep states. The effect

of the lower heart rates are associated with higher probability of being asleep. Regarding temperature, one can hardly identify any definitive relationship using the log odds. Moreover, in Table 3.10, we report the empirical transition probability of sleep state. It shows that the current sleep state is highly dependent on the previous state. More specifically, there is a strong tendency for sleep to remain in its current state.




Heart rate time series

Temperature time series

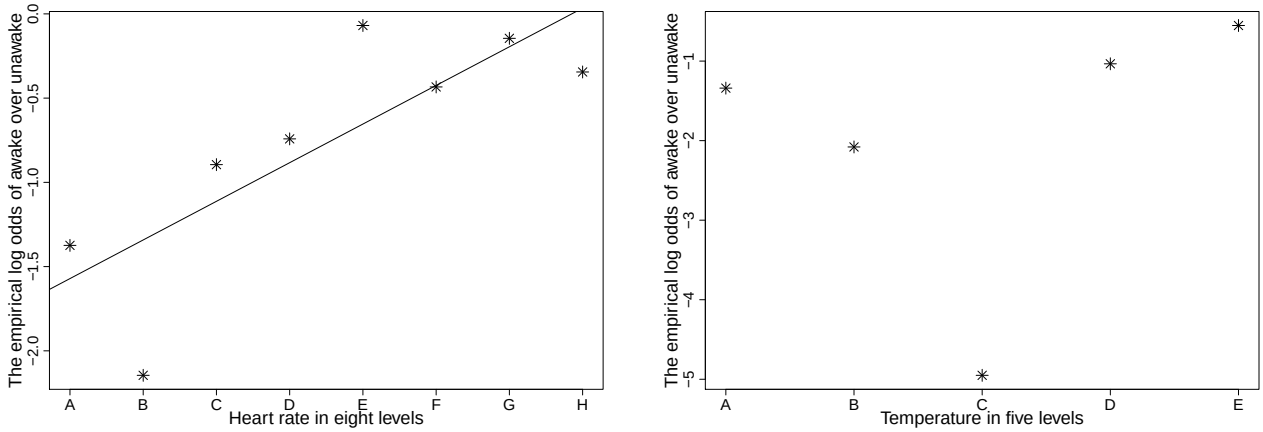
Figure 3.3: Left: heart rate (in beats per minute). Right: temperature (in Celsius).

Table 3.10: Empirical transition table of sleep state: when the current state is not awake, the sample probability of staying not wake in the next time point is 729/735 while the sample probability of being in the awake state at the next time point is 6/735. When the current state is awake, the sample probability of staying awake at the next time point is 282/288 while the sample probability of changing to a non-awake state at the next time point is 6/288.

| | $Y_{i-1} = 0$ | $Y_{i-1} = 1$ |
|---|---|---|
| $Y_i = 0$ | 729/1023 | 6/1023 |
| $Y_i = 1$ | 6/1023 | 282/1023 |

### 3.5.2 Modeling and Results

Following the exploratory analysis, $\log H_i$, $Y_{i-1}$ and time (in minutes) are suggested in the proposed model. Since there is strong effect of $\log H_i$, $Y_{i-1}$ on current sleep state, we in-




Empirical log odds versus heart rate Empirical log odds versus temperature

Figure 3.4: Scatterplots of empirical log odds versus heart rate and temperature. The left panel shows the empirical log odds over eight levels of heart rate. The right panel displays the same value versus temperature.

clude those two covariates as fixed effect components. Gaussian process on time domain is introduced to capture the nonlinear term.

We also applied our proposed binary hybrid approach to make the inference and prediction. Summaries of point and interval estimates are shown in Table 3.11. It is seen that compared with the ordinal model (logistic regression), the point estimates are similar. However, there is significantly large difference among the interval estimates. Using the proposed HIBITS method, we gain substantially narrower confidence intervals than ordinal model. The benefits are up to almost 90% shorter in length. From the proposed results, we find that one unit increment in heart rate at current time point will lead to 211.4% accretion of odds. Current odd of sleep state when previous sleep state is awake is estimated to be dramatically higher than that when previous state is not awake. To test the prediction power of this model, the proposed method was implemented with various training and testing data size. Numerical results are summarized in Table 3.12. It can be seen that the model produces around 99% prediction accuracy while ordinal model yields about 96%. As we decrease the ratio of

training over testing data size, the prediction accuracy remains stable. Time series plots of the real and predicted sleep state are presented in Figure 3.5. It can be shown that the proposed method produces high prediction accuracy and recover the same sleep state pattern as the real dataset. To check for the sensitivity of the proposed method to the estimated value of parameter $\lambda$, we compared the results from the data-adaptive estimate (0.730) against the following values $(1.730, 2.730)$. The data-adaptive estimate gave roughly the same prediction error but the confidence intervals were narrower.

Table 3.11: Summary of the sleep state analysis. The point and interval estimates from HIBITS method are obtained by Section 3.3.4. It can be seen that the widths of the confidence intervals from the HIBITS method are narrower than those of the classical ordinal model.

| Parameters$(\beta_0, \beta_1)$ | Method | Point estimate | 95% confidence intervals |
|---|---|---|---|
| $\beta_0$ | HIBITS method | 1.136 | $(1.000, 1.271)$ |
| | Ordinal model | 1.105 | $(0.101, 2.109)$ |
| $\beta_1$ | HIBITS method | 8.275 | $(8.124, 8.427)$ |
| | Ordinal model | 8.241 | $(6.669, 9.813)$ |

Table 3.12: Prediction accuracy with different training and testing data size.

| Training/Testing data size | Method | Prediction Accuracy |
|---|---|---|
| 600/400 | HIBITS method | 99.0% |
| | Ordinal model | 96.0% |
| 500/500 | HIBITS method | 99.2% |
| | Ordinal model | 96.1% |
| 400/600 | HIBITS method | 99.1% |
| | Ordinal model | 96.4% |

### 3.5.3 Discussion on Missing data

One advantage of the proposed prediction model is that it captures the information from its own past. Derived from the results, the odds when previous sleep state is awake is 4000-fold higher than that when the preceding state is not awake. However, if there are missing data or the observations are not collected successively, such information will be lost. This motivates

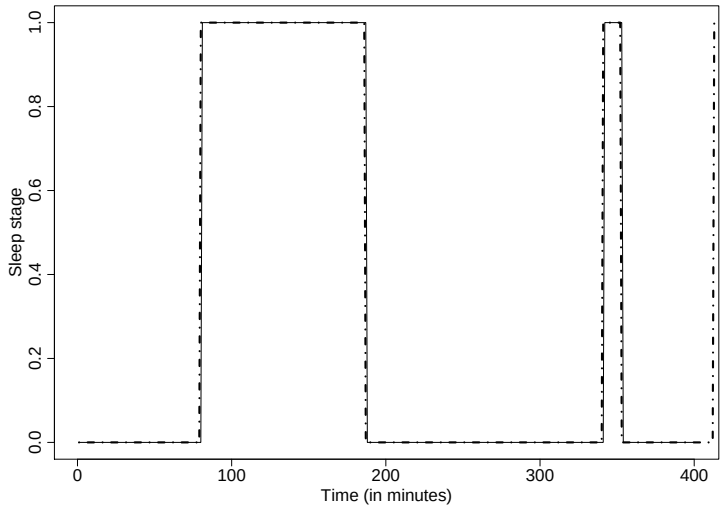



Figure 3.5: Predicted sleep state (solid line) overlaid with real data (dotted line) (training/testing data size 600/400).

us to adjust the model to fit such cases. In the adjusted model, we choose $\log H_i$ as fixed effects and still use Gaussian process on time domain. To test the prediction power, instead of fixing the training and testing dataset, we randomly pick those two pieces of data with fixed size. The proposed HIBITS method was implemented. Summaries of the test results can be found in Table 3.13. The tests were conducted 10 times with training and testing data of different sizes. From the results, it is clear that as the training data size becomes larger, the prediction accuracy increases at a reasonable rate. As the training data size reaches 600, the accuracy is promising.

Table 3.13: Prediction accuracy with different training and testing data size, *stands for the test number.

| Training/Testing data size | 1* | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | Average |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 400/100 | 0.76 | 0.92 | 0.95 | 0.93 | 0.80 | 0.64 | 0.84 | 0.89 | 0.90 | 0.93 | 0.86 |
| 500/100 | 0.99 | 0.89 | 0.90 | 0.98 | 0.91 | 0.88 | 0.89 | 0.93 | 0.95 | 0.91 | 0.92 |
| 600/100 | 0.95 | 0.97 | 0.99 | 0.91 | 0.93 | 0.96 | 0.97 | 0.90 | 0.92 | 0.93 | 0.94 |

To further study the performance of the proposed HIBITS method, we make the ratio of training over testing data size smaller. Particularly, we change the training and testing data size to be 700 and 300 respectively. The prediction accuracy is around 0.873. If we move further to change the training data size to be 800 and testing data size to be 200, the prediction accuracy is about 90%. All the results demonstrate that the proposed binary hybrid method produces promising prediction power when the dataset are not collected successively or partly missing. Moreover, it should be pointed out that the computation is not very demanding. The tests are conducted in R programming and the operation time is approximately 90 seconds for each test.

# Chapter 4

# Evolutionary State-Space Model

## 4.1 Introduction

The goal of this chapter is to develop a novel statistical model for investigating the evolution of a brain process duration of a learning experiment. To infer brain neuronal activity, electrophysiological recordings such as local field potentials (LFPs) and electroencephalograms (EEGs) are commonly used to indirectly measure electrical activity of neurons. In this chapter, we consider LFPs from multiple electrodes that capture the integration of membrane currents in a local region of cortex (Mitzdorf et al., 1985).

In practice, LFPs are the observed spatio-temporal signals at different tetrodes. In a motivating example, an olfactory (non-spatial) sequence memory experiment has been performed in a memory laboratory to study how neurons learn the sequential ordering of presented odors (Allen et al., 2016). In this study, LFP recordings in a rat are obtained from an implanted plate with 12 electrodes. One epoch corresponds to about 1 second in physical time. As shown in Figure 4.1, rats are trained to identify a sequence of odors while their LFP signals

from one electrode are recorded and plotted for the first 15 epochs. We further study the behavior of these LFPs by examining their spectra. In Figure 4.2, we plot the boxplots of the log periodograms across all the epochs from one electrode. These plots reveal that LFPs contain power at distinct bands: delta (0-4 Hertz), alpha (8-12 Hertz) and the high-beta low-gamma (30-35 Hertz) bands. As an exploratory step, we divide the entire experiment into three phases, early, middle, and late phases. In each phase, we compute the average periodogram (averaged across epochs) and present them on the left side of Figure 4.3. On the right side, we plot the relative periodogram (obtained by rescaling the periodogram so that the relative periodogram for each frequency sums up to 1) and find that the spectral power evolves during the course of experiment. During the early phase, power has a broad (rather than concentrated) spread across bands. However, at the late phase, power seems to be more concentrated at the lower beta band.

In summary, the preliminary results suggest that the spectra of the LFPs appear to change across the epochs in the experiment. Therefore, statistical models that are capable of describing LFP signals' evolution over the course of epochs are largely needed to help understand how the rat learns the sequence of the odor presentation.

As discussed in Chapter 1.2, existing approaches such as PCA, ICA have their critical limitations. Alternatively, we develop an evolutionary state space model (E-SSM) that explicitly captures the evolutionary behavior in high dimensional time series. The E-SSM shares a similar form with the classical state-space model (as in Shumway and Stoffer (2013)) but differs in that the parameters are varying across epochs and the mixing matrix is unknown and therefore has to be estimated. Moreover, E-SSM manages to capture the temporal correlation of each of the latent sources by characterizing them using second order autoregressive [AR(2)] processes. The reason for choosing AR(2) is due to its ability to capture the precise oscillatory behavior of these latent sources. In particular, by parameterizing these sources as

AR(2), we can easily constrain the power of each source to center at pre-specified frequency bands such as delta (0 - 4 Hertz), alpha (8 - 12 Hertz) and high-beta gamma (> 30 Hertz) bands, where the choice of these particular frequency bands is due to the standard convention in neuroscience based upon previous Electrophysiological data analysis (Deuschl et al., 1999). The use of AR(2) mixture here can be viewed as an analogy of Gaussian mixture models for classical density estimation problems. Compared to the classical methods such as ICA and PCA, the sources produced by E-SSM are more directly interpretable in terms of oscillatory properties.

The main contributions of this chapter are as follows: (1.) The proposed E-SSM model provides a rigorous framework in modeling brain activity, connectivity and their dynamic behavior during the course of experiments. In particular, our model accounts for the temporal evolution/dependence of the spectrum power for particular frequency bands across the entire experiment as well as the temporal structure among the latent sources. (2.) E-SSM gives interpretable results by modeling particular predominant frequency bands that are associated with various brain functional states through AR(2) processes. (3.) In theory, we show that the spectrum of arbitrary weakly stationary time series can be approximated by the spectrum of AR(2) mixtures, which gives a theoretical justification of the use of AR(2) mixtures. (4.) By applying the E-SSM model, one can easily conduct analysis on both of time and frequency domains and thus provide a complete characterization of the underlying brain process. (5.) Finally, the E-SSM model and the proposed estimation method, in general, are intuitive and can be implemented easily thanks to the existing theory and algorithm for state space model. However, the key difference is the generalization of the multiple epochs setting which allows pooling information across epochs and a flexible mixing matrix estimation step.

The rest of this chapter is organized as follows. In Section 4.2, we introduce the E-SSM method that models the variability across epochs while taking into account particular fre-

quency bands. In Section 4.3, we propose a hybrid iterative method that comprises of Kalman filter and blocked resampling for parameter estimation. We discuss the main differences between the proposed E-SSM and other existing approaches such as ICA and PCA in Section 4.4. In Section 4.5, we show that the proposed method is promising in reconstructing the latent source signals and their spectrum in simulation studies under both single-epoch and multiple-epoch scenarios. We then analyze LFPs dataset obtained from a non-spatial olfactory sequence memory study in Section 4.6.

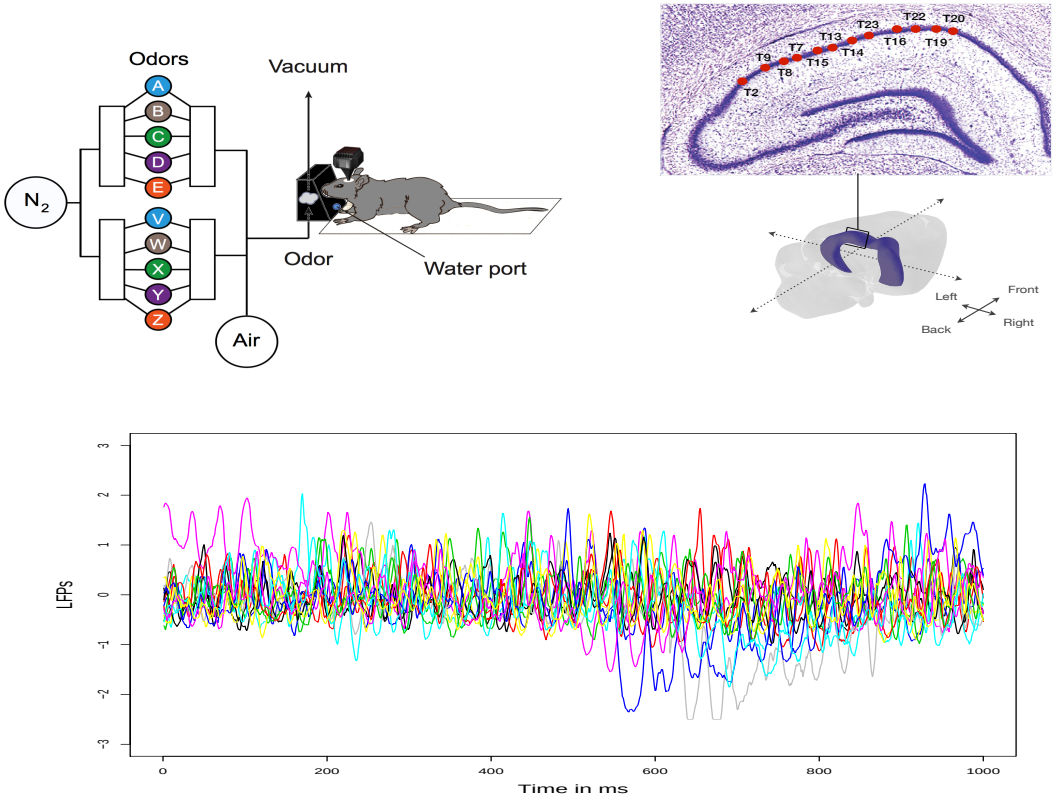



Figure 4.1: Top left: Apparatus and behavioral design for the olfaction (non-spatial) memory sequence experiment (Allen et al., 2016). Series of five odors were presented to rats from the same odor port. Top right: The spatial locations of electrodes implanted in the hippocampus region. Bottom: The overlaid time series LFPs plots of the first 15 epochs at electrode T22. Each epoch consists of 1 second recording (1000 milliseconds). The experiment and the data are reported in Allen et al. (2016).

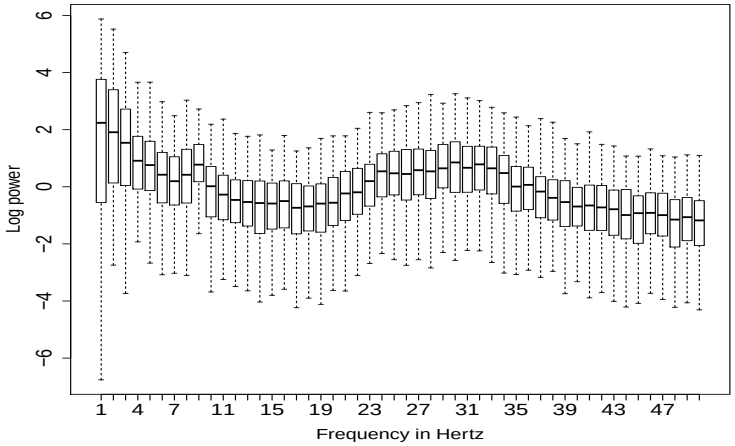



Figure 4.2: The log periodogram boxplots for each frequency obtained by all 247 epochs at electrode T22.

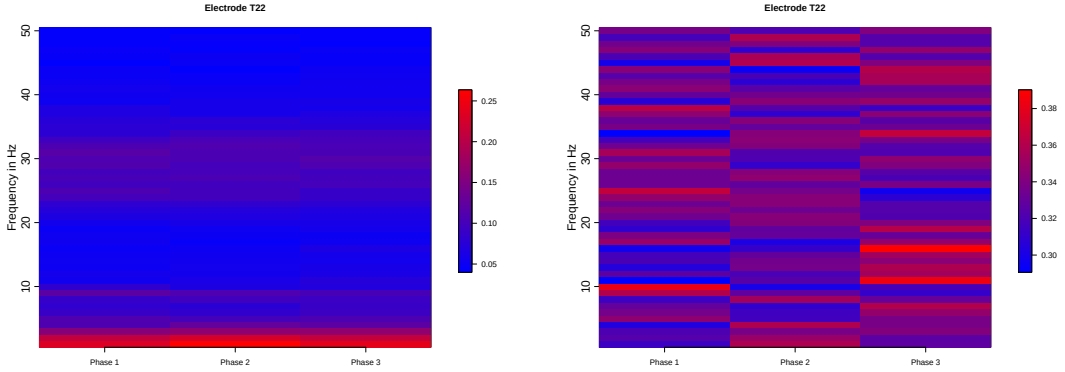



Figure 4.3: Left: The heatmap of the averaged periodogram among Phase 1 (epochs 1 - 80), Phase 2 (81 - 160) and Phase 3 (161 - 247) respectively at electrode T22. The original signals were rescaled to unit variance. Right: The heatmap of the relative periodogram (summing up to 1 for each frequency). Spectral power (decomposition of waveform) evolved across phases of the experiment.

## 4.2 Evolutionary State Space Model (E-SSM)

In this section, we will discuss models for inferring latent structures in LFPs and their evolution across epochs over the entire experiment. We shall first describe the model for a single epoch and then discuss the extension to treat multiple epochs.

### 4.2.1 State Space Model for a Single Epoch

Denote $t = 1, \cdots, T$ as the time points in a single-epoch and
$\boldsymbol{Y}_t = (Y_t(1), \cdots, Y_t(p))'$ as the observed LFPs where $p$ is the number of electrodes. For any fixed time point $t$, we assume that $\boldsymbol{Y}_t$ is a *mixture* of $q$ latent independent source signals $\boldsymbol{S}_t = (S_t(1), \cdots, S_t(q))'$, where $q$ is the number of spatial source signals. Then the model can be presented as $\boldsymbol{Y}_t = M\boldsymbol{S}_t + \boldsymbol{\epsilon}_t$, where $M$ is the mixing matrix, $\boldsymbol{\epsilon}_t = (\epsilon_t(1), \cdots, \epsilon_t(p))'$ is noise that follows $N(\boldsymbol{0}, \tau^2 \boldsymbol{I_p})$ and $\boldsymbol{I_p}$ is an identity matrix of dimension $p$. Each of the independent latent signals $S_t(l), l = 1, \cdots, q$ models the source that represents oscillatory activity at a set of pre-specified frequency bands (e.g., delta, alpha and gamma).

Modeling the source signals $\boldsymbol{S}_t$

One important parameterization in our model is to constrain the sources to have an AR(2) structure such that each represents a particular oscillator: delta ($\delta$: 0 - 4 Hertz), theta ($\theta$: 4 - 8 Hertz), alpha ($\alpha$: 8 - 12 Hertz), lower beta ($\beta$: 12 - 18 Hertz) and gamma ($\gamma$: $>$ 30 Hertz). Recall that an autoregressive operator of order 2 is defined by

$$\varphi(B) = 1 - \varphi_1 B - \varphi_2 B^2, \tag{4.1}$$

where $B$ is a backshift operator defined by $B^\ell S_t = S_{t-\ell}$, and $\varphi_1, \varphi_2$ are the corresponding

coefficients. It can be shown that the spectrum of an AR(2) process with noise level $\sigma_\omega$ is $f_S(\omega) = \frac{\sigma_w^2}{|1-\varphi_1 \exp(-2\pi i\omega)-\varphi_2 \exp(-4\pi i\omega)|^2}$. To illustrate its use in practice, we plot the spectrum of an AR(2) process with $\varphi_1 = 1.976, \varphi_2 = -0.980, \sigma_w = 0.1$ in Figure 4.4. It can be seen that there is a peak at frequency $\omega = 10$ Hertz, which means that the frequency band around $\omega = 10$ Hertz dominates the process and thus produces the most power. This property of AR(2) model makes it potentially useful for characterizing brain signals (such as LFPs) with oscillations at either broad or narrow frequency band.

We now explain the connection between the AR(2) coefficients and the spectrum (i.e., the location and spread of the peak). First, the process is causal when the roots of the polynomial in Equation (4.1) have magnitudes greater than 1. Furthermore, under causality, Jiru (2008) and Shumway and Stoffer (2013) demonstrate that when the roots of the polynomial in Equation (4.1) are complex-valued with magnitude greater than 1, then the spectrum attains a peak that is centered around the phase of the roots. Moreover, when the magnitude of the roots become larger than 1, the peak becomes less concentrated around the phase.

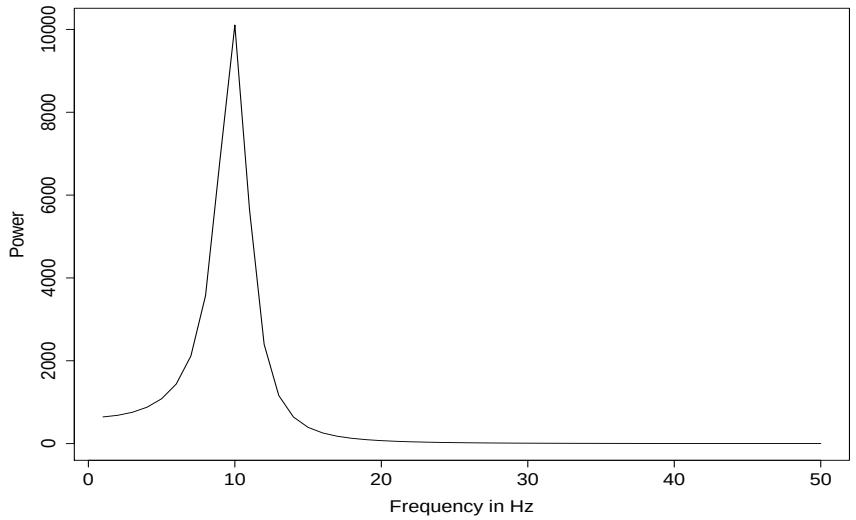



Figure 4.4: The theoretical spectra of an AR(2) process with power concentrated at the alpha band: $\varphi_1 = 1.976, \varphi_2 = -0.980, \sigma_w = 0.1$

Motivated by this result, we will fix the phase (or argument) of each of the AR(2) polynomial roots to model each of the particular bands obtained from previous study results. As noted, fixing the phase is consistent with neuroscience standard and thus will not be a constraint in practice. To model the evolution across epochs, we allow the modulus of the AR(2) polynomial roots to change among epochs. As a result, as the phase of the roots for each of the latent independent source signals is fixed, the AR(2) process is uniquely determined by the modulus and the variance. In practice, the value of modulus controls the spread of the spectrum curves. For an AR(2) process $S_t = \varphi_1 S_{t-1} + \varphi_2 S_{t-2} + w_t$, the modulus $\rho$ and phase $\psi$ of the roots of the polynomial have the relationship that $\varphi_1 = 2\rho^{-1}cos(\psi), \varphi_2 = -\rho^{-2}$. This result can be seen as an analogy of the use of Gaussian mixture model (or any location-scale mixture in general) for density estimation. In Section 4.2.3, we will further discuss the approximation property of the AR(2) mixture.

Generalized state-space model

Following the previous discussion, the latent independent spatial source signals are modeled as multivariate AR(2)s, $\boldsymbol{S}_t = \Phi_1 \boldsymbol{S}_{t-1} + \Phi_2 \boldsymbol{S}_{t-2} + \boldsymbol{\eta}_t$, where $\Phi_1 = \text{diag}(\varphi_{11}, \cdots, \varphi_{q1}), \Phi_2 = \text{diag}(\varphi_{12}, \cdots, \varphi_{q2}) \in \mathbb{R}^{q \times q}$ are diagonal matrices, and the noise $\boldsymbol{\eta}_t = (\eta_1(t), \cdots, \eta_q(t))' \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I_q})$. The final model can hence be viewed as a generalized state-space model:

$$\begin{aligned} \boldsymbol{Y}_t &= \widetilde{M} \boldsymbol{X}_t + \boldsymbol{\epsilon}_t, \\ \boldsymbol{X}_t &= \widetilde{\Phi} \boldsymbol{X}_{t-1} + \widetilde{\boldsymbol{\eta}}_t, \end{aligned} \tag{4.2}$$

where $\boldsymbol{X}_t = (\boldsymbol{S}_t', \boldsymbol{S}_{t-1}')'$, $\widetilde{M} = (M, \boldsymbol{0}) \in \mathbb{R}^{p * 2q}$, $\widetilde{\Phi} = \begin{bmatrix} \Phi_1 & \Phi_2 \\ \boldsymbol{I_q} & \boldsymbol{0} \end{bmatrix}$, and $\widetilde{\boldsymbol{\eta}}_t = (\boldsymbol{\eta}_t', \boldsymbol{0})'$. Note that the model in (4.2) is not a regular state-space model since the mixing matrix $\widetilde{M}$ is unknown. Moreover, following the aforementioned discussion, the coefficients of the autoregressive pro-

cesses are determined by the modulus $\boldsymbol{\rho} = (\rho_1, \cdots, \rho_q)$ and phase $\boldsymbol{\psi} = (\psi_1, \cdots, \psi_q)$ of the autoregressive polynomial roots. Since we are interested in particular frequency bands, we fix the phase $\boldsymbol{\psi}$ and the state equation in (4.2) is parameterized by $\boldsymbol{\rho}$ and $\sigma^2$.

### 4.2.2 Evolutionary State Space Model for Multiple Epochs

Next, we generalize the model in Section 4.2.1 to accommodate multiple epochs. We assume that across epochs, the mixing matrix $M$ is fixed and the latent independent autoregressive processes evolve through the modulus $\boldsymbol{\rho}$. This assumption implies that the cortical structure remains unchanged across epochs for each individual. We denote $r = 1, \cdots, R$ as the epochs in the experiment, then the model is given by

$$\begin{aligned} \boldsymbol{Y}_t^{(r)} &= \widetilde{M}\boldsymbol{X}_t^{(r)} + \boldsymbol{\epsilon}_t^{(r)}, \\ \boldsymbol{X}_t^{(r)} &= \widetilde{\Phi}^{(r)}\boldsymbol{X}_{t-1}^{(r)} + \widetilde{\boldsymbol{\eta}}_t^{(r)}, \end{aligned} \tag{4.3}$$

where the definition of $\boldsymbol{Y}_t^{(r)}, \widetilde{M}, \boldsymbol{X}_t^{(r)}, \widetilde{\Phi}^{(r)}, \boldsymbol{\epsilon}_t^{(r)}, \widetilde{\boldsymbol{\eta}}_t^{(r)}$ are similar as in Equation (4.2) except the additional superscript $r$ for each epoch $r$.

In the proposed model, we assume an autoregressive structure that evolves across epochs. This assumption is inspired by the preliminary analysis in Section 4.1 showing that the power spectrum evolves during the course of the experiment. Accordingly, the evolutionary spectrum of each latent source will be easily captured in an explicit form $f^{(r)}(\omega) = \frac{\sigma_w^{2(r)}}{|1-\varphi_1^{(r)}\exp(-2\pi i\omega)-\varphi_2^{(r)}\exp(-4\pi i\omega)|^2}$. We also assumed that the mixing matrix is invariant to epochs. This is due to the fact that the network structure of subjects is not changing across phases of experiments. To reiterate, non-stationarity will be captured by the AR(2) coefficients.

In the literature, there have been numerous discussions on the identifiability issues of state-space models (Hamilton, 1994). Indeed, for a general state-space model, the same representation can be obtained by applying an orthogonal transformation on matrices. Zhang and Hyvärinen (2011) proposed a non-Gaussian constraint to avoid the identifiability issue. In this chapter, to ensure the uniqueness of the solution, we require that each component of the latent independent source signals $\boldsymbol{S}(t)$ to have unit variance and the entries of $\widetilde{M}$ are positive.

### 4.2.3 Theoretical Results on AR(2) Decompositions

As we have discussed in the previous sections, we choose to represent individual sources by AR(2) models due to their ability to present each source signals at pre-specified frequency bands. In this section, we further justify this representation by showing that the spectrum of arbitrary weakly stationary process can be approximated by a linear mixture of AR(2) processes. In particular, the approximation error by the use of AR(2) mixture is asymptotically negligible as the number of mixtures goes towards infinity. Its proof is given in the Appendix.

**Theorem 4.1.** ***(AR(2) spectral decomposition theorem)*** *Let $Y_t$ be a weakly stationary time series with zero mean and spectrum $f_Y(\omega)$. Let $[\omega_0, \omega_1) \cup [\omega_1, \omega_2) \cup \cdots \cup [\omega_{M-1}, \omega_M]$ be a partition of the frequency domain $[0, \frac{1}{2}]$ such that*

$$\sup\{|\omega_1 - \omega_0|, \cdots, |\omega_M - \omega_{M-1}|\} \to 0 \quad as \quad M \to \infty. \tag{4.4}$$

*Denote $S_t^{(j)}, j = 1, \cdots, M$ as independent AR(2) processes with unit variance and spectrum of $f_{S^{(j)}}(\omega)$ such that the phase of its AR polynomial roots, denoted by $\psi^{(j)}$, satisfies $\psi^{(j)} \in [\omega_{j-1}, \omega_j)$. Consider a family of processes $\{Q_{t,M}\}_{M=1}^{\infty}$ defined by $Q_{t,M} = \sum_{j=1}^{M} a_j S_t^{(j)}$ with*

*non-negative coefficients* $\{a_j\}_{j=1}^{M}$ *for every* $M = 1, \ldots, \infty$. *Then we have*

$$\sup_{a_1,\ldots,a_M \geqslant 0} ||f_Y(\omega) - f_{Q_{t,M}}(\omega)||_2 \to 0, \quad as \quad M \to \infty, \tag{4.5}$$

*where* $f_{Q_{t,M}}(\omega)$ *is the spectrum of* $Q_{t,M}$.

## 4.3 Estimation Method for E-SSM

### 4.3.1 Estimating E-SSM for a Single Epoch

We first consider E-SSM for a single epoch. We propose an iterative algorithm that comprises of Kalman filter and least squares for parameter estimation purpose. We start with initial values $\widetilde{M} = \widetilde{M}_0$, $\boldsymbol{X}_0^0$ and $P_0^0$. The estimation procedure takes iterations between Algorithms 3 and 4 (shown below) until convergence.

**Algorithm 3** Kalman Filter and Maximum Likelihood

**procedure** GIVEN $\widetilde{M}, \boldsymbol{X}_0^0, P_0^0$, ESTIMATE $\boldsymbol{\rho}, \sigma^2, \tau^2$ BY KALMAN FILTER AND MAXIMUM LIKELIHOOD OF INNOVATIONS $\boldsymbol{\epsilon_t}$
  A.1 Kalman filter and Kalman gain step
  $\Phi_1 \leftarrow diag(2\rho_1^{-1}cos(\psi_1), \cdots, 2\rho_q^{-1}cos(\psi_q))$
  $\Phi_2 \leftarrow diag(-\rho_1^{-2}, \cdots, -\rho_q^{-2})$
  $\tilde{\Phi} \leftarrow \begin{bmatrix} \Phi_1 & \Phi_2 \\ \boldsymbol{I_q} & \boldsymbol{0} \end{bmatrix}$
  **for** $t = 0, \ldots, T$ **do**
    $\boldsymbol{X}_t^{t-1} \leftarrow \tilde{\Phi}\boldsymbol{X}_{t-1}^{t-1}$
    $P_t^{t-1} \leftarrow \tilde{\Phi}P_{t-1}^{t-1}\tilde{\Phi}' + \sigma^2 \begin{bmatrix} \boldsymbol{I_q} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix}$
    $K_t \leftarrow P_t^{t-1}\widetilde{M}'[\widetilde{M}P_t^{t-1}\widetilde{M}' + \tau^2\boldsymbol{I_p}]^{-1}$ ▷ The Kalman gain
    $\boldsymbol{X}_t^t \leftarrow \boldsymbol{X}_t^{t-1} + K_t(\boldsymbol{Y}_t - \widetilde{M}\boldsymbol{X}_t^{t-1})$
    $P_t^t \leftarrow (\boldsymbol{I_{2q}} - K_t\widetilde{M})P_t^{t-1}$
  A.2 Maximum likelihood estimation
  **for** $t = 0, \ldots, T$ **do**
    $\boldsymbol{\epsilon_t} \leftarrow \boldsymbol{Y}_t - \widetilde{M}\boldsymbol{X}_t^{t-1}$
    $\Sigma_t \leftarrow \widetilde{M}P_t^{t-1}\widetilde{M}' + \tau^2\boldsymbol{I_p}$
  $l_Y(\boldsymbol{\rho}, \sigma^2, \tau^2) \leftarrow \frac{1}{2}\sum_{t=1}^{T}\log|\Sigma_t| + \frac{1}{2}\sum_{t=1}^{T}\boldsymbol{\epsilon_t}'\Sigma_t^{-1}\boldsymbol{\epsilon_t}$ ▷ The negative loglikelihood
  $(\hat{\boldsymbol{\rho}}, \hat{\sigma}^2, \hat{\tau}^2) \leftarrow \underset{(\boldsymbol{\rho}, \sigma^2, \tau^2)}{\operatorname{argmin}} \; l_Y(\boldsymbol{\rho}, \sigma^2, \tau^2)$ ▷ Maximizing the likelihood of innovations
  **return** $\hat{\boldsymbol{\rho}}, \hat{\sigma}^2, \hat{\tau}^2$

In this study, since we are interested in the power of particular frequency bands, we will introduce box constraints to the modulus $\rho_1, \cdots, \rho_q$ to control the spread of the spectra curves. Hence in A.2 of Algorithm 3, we implement an optimization approach with box constraints on modulus $\rho_1, \cdots, \rho_q$ and no constraints on $\sigma^2, \tau^2$.

**Algorithm 4** Kalman Filter and Least Squares Estimation

**procedure** GIVEN THE CURRENT ESTIMATES OF $\boldsymbol{\rho}, \sigma^2, \tau^2$, WE CAN OBTAIN THE ESTIMATES OF $\widetilde{M}$ BY KALMAN FILTER AND LEAST SQUARES ESTIMATION.
  B.1 Kalman filter and Kalman gain step
  $\Phi_1 \leftarrow diag(2\rho_1^{-1}cos(\psi_1), \cdots, 2\rho_q^{-1}cos(\psi_q))$
  $\Phi_2 \leftarrow diag(-\rho_1^{-2}, \cdots, -\rho_q^{-2})$
  $\widetilde{\Phi} \leftarrow \begin{bmatrix} \Phi_1 & \Phi_2 \\ \boldsymbol{I_q} & \boldsymbol{0} \end{bmatrix}$
  **for** $t = 0, \ldots, T$ **do**
    $\boldsymbol{X}_t^{t-1} \leftarrow \widetilde{\Phi}\boldsymbol{X}_{t-1}^{t-1}$
    $P_t^{t-1} \leftarrow \widetilde{\Phi}P_{t-1}^{t-1}\widetilde{\Phi}' + \sigma^2 \begin{bmatrix} \boldsymbol{I_q} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix}$
    $K_t \leftarrow P_t^{t-1}\widetilde{M}'[\widetilde{M}P_t^{t-1}\widetilde{M}' + \tau^2\boldsymbol{I_p}]^{-1}$ ▷ The Kalman gain
    $\boldsymbol{X}_t^t \leftarrow \boldsymbol{X}_t^{t-1} + K_t(\boldsymbol{Y}_t - \widetilde{M}\boldsymbol{X}_t^{t-1})$
    $\boldsymbol{X}_t^t \leftarrow \boldsymbol{X}_t^t/\mathrm{sd}(\boldsymbol{X}_t^t)$ ▷ $\mathrm{sd}(\boldsymbol{X}_t^t)$ denotes the standard deviation of $\boldsymbol{X}_t^t$
    //Remark: We scale $\boldsymbol{X}_t^t$ to unit variance for identifiability issues discussed before.
    $P_t^t \leftarrow (\boldsymbol{I_{2q}} - K_t\widetilde{M})P_t^{t-1}$
  B.2 Least square estimation from Equation (4.2)
  $\boldsymbol{Y} \leftarrow (\boldsymbol{Y_1}, \cdots, \boldsymbol{Y_T})$ ▷ $\boldsymbol{Y} \in \mathbb{R}^{p*T}$
  $\boldsymbol{X} \leftarrow (\boldsymbol{X_1^1}, \cdots, \boldsymbol{X_T^T})$ ▷ $\boldsymbol{X} \in \mathbb{R}^{q*T}$
  **for** $w = 1, \ldots, p$ **do**
    $\widetilde{M_w} \leftarrow (\boldsymbol{X} * \boldsymbol{X'})^{-1} * \boldsymbol{X} * \boldsymbol{Y'_{(w)}}$ ▷ $\boldsymbol{Y}_{(w)}$ denotes the $w$th row of $\boldsymbol{Y}$
  $\widetilde{M} \leftarrow (\widetilde{M_1}, \cdots, \widetilde{M_w})'$
  **return** $\widetilde{M}$

### 4.3.2 Estimating E-SSM for Multiple Epochs

Now we extend the previous method to the multiple epoch setting in Equation (4.3). The major challenge lies in pooling information from different epochs in estimating the epoch-invariant mixing matrix. To solve this problem, we propose a blocked resampling based approach. The key idea can be summarized as follows: we first divide the epochs into blocks; then for each block we estimate the corresponding mixing matrix and the epoch-specific AR(2) parameters. These blocks retain the temporal sequence of the epochs and the final estimate at a previous epoch serves as the initial estimate of mixing matrix at

the current epoch. The final estimates of the mixing matrix obtained from each block are averaged to produce the estimate for the common mixing matrix. Moving on to the next step, given the estimated mixing matrix, we follow Algorithm 3 to obtain estimates of the epoch-specific AR(2) parameters. The iterative approach is summarized below.

**II.A** We fix the length of the blocked resampling sampler as $l$. We draw the starting epoch index $s$ from the set $\{1, 2, \cdots, R - l + 1\}$. Then at current iteration, the blocked resampling sampler is $(\{\boldsymbol{Y}_t^{(s)}\}_{t=1}^T, \cdots, \{\boldsymbol{Y}_t^{(s+l-1)}\}_{t=1}^T)$.

**A.1**. Starting with epoch $s$, we implement the approach for single epoch in Section 4.2.1 on $\{\boldsymbol{Y}_t^{(s)}\}_{t=1}^T$to obtain estimates $\widetilde{M}^{(s)}$.

**A.2**. Staring with epoch $s+1$ and the initial value $\widetilde{M}^{(s)}$, we repeat A.1 to obtain estimates $\widetilde{M}^{(s+1)}$.

**A.3**. We repeat A.2 until the last epoch $s + l - 1$. We denote the final estimates $\widetilde{M}^{(s+l-1)}$ as the ultimate estimates of resampling sampler $(\{\boldsymbol{Y}_t^{(s)}\}_{t=1}^T, \cdots, \{\boldsymbol{Y}_t^{(s+l-1)}\}_{t=1}^T)$. The pipeline of the procedure is summarized below.

$$\begin{bmatrix} \boldsymbol{Y}_1^{(s)} \\ \boldsymbol{Y}_2^{(s)} \\ \cdots \\ \boldsymbol{Y}_T^{(s)} \end{bmatrix} \rightarrow \widetilde{M}^{(s)} \rightarrow \begin{bmatrix} \boldsymbol{Y}_1^{(s+1)} \\ \boldsymbol{Y}_2^{(s+1)} \\ \cdots \\ \boldsymbol{Y}_T^{(s+1)} \end{bmatrix} \rightarrow \qquad \widetilde{M}^{(s+1)} \cdots \rightarrow \begin{bmatrix} \boldsymbol{Y}_1^{(s+l-1)} \\ \boldsymbol{Y}_2^{(s+l-1)} \\ \cdots \\ \boldsymbol{Y}_T^{(s+l-1)} \end{bmatrix} \rightarrow \widetilde{M}^{(s+l-1)}$$

**II.B**. Repeat II.A until a sufficient number of resampling estimates is obtained. Compute the average of those estimates, defined by $\widetilde{M}_g$, as the global estimate of $\widetilde{M}$.

**II.C**. Plug the global estimate $\widetilde{M}_g$ into every single epoch. Following Algorithm 3 for single epoch discussed in Section 4.2.1, we obtain the estimates of $\boldsymbol{\rho}^{(r)}, \sigma^{2(r)}, \tau^{2(r)}, r = 1, \cdots, R$.

The over-all work flow is given in Figure A.10. Note that since the mixing matrix $\widetilde{M}$ are the same across epochs, we use the blocked resampling strategy to get the global estimates sequentially. Given that estimate, we proceed to make inference on every single epoch.

## 4.4 A Comparison to Existing Methods

We discuss a few major differences between our method and the existing state-of-art approaches including ICA and classical state-space models.

ICA has been widely used in single/between-subject electrophysiological exploratory analysis. For example, Makarova et al. (2011) proposed an ICA method to segregate pathways with partially overlapped synaptic territories from hippocampal LFPs. To investigate the variability across different subjects or subgroups, Guo (2011) proposed a general group probabilistic ICA (pICA) framework to accommodate cross-subject structure in multi-subject spatial-temporal brain signals. Although these methods work well under certain settings, there is still plenty of room for improvement in modeling electrophysiological signals. First, they do not have a mechanism for capturing how the parameters (and spectral properties) of the latent source signals evolve across epochs over the entire experiment. Most of the existing methods are based on concatenating the signals from different epochs and estimating parameters as though these signals are realizations of the *same* underlying process. However, since the "reconstructed" latent sources vary across epochs, there is no rigorous framework for modeling how these parameters could change across epochs. As demonstrated in our exploratory analysis, Figures 4.2 and 4.3 show that the power of LFP signals changes quite drastically from the middle phase to the late phase of the experiment. Simply lumping together signals that are generated from different underlying source processes could yield misleading results. Second, the existing methods do not take into account the temporal

structure of the latent sources. In fact, these sources are estimated for each time point independently of the other time points. Third, the current ICA methods for source modeling may not produce interpretable results from spectral analysis of electrophysiological signals. In fact, brain researchers have observed association between power at different frequency bands and brain functional states (Michel et al., 1992). Thus, it is necessary to develop a framework that accounts for the evolution of the power at these frequency bands over many epochs. Lastly, there are limitations in the connection between time and frequency domain analysis. Methods from time and frequency domain are developed almost exclusively from each other, which is counter-intuitive since these two approaches ought to be used concurrently in order to give a complete characterization of brain processes.

## 4.5 Simulation Studies

### 4.5.1 Results on Single Epoch Analysis

In this section, we evaluate the proposed E-SSM on single epoch data. For the latent independent source signals, we assume that there are three AR(2) stationary processes. Each of them corresponds to delta ($\delta$: 0 - 4 Hertz), alpha ($\alpha$: 8 - 12 Hertz), lower beta ($\beta$: 12 - 18 Hertz) frequency bands respectively. We randomly generate a positive "mixing" matrix $M$ and fix the number of electrodes of the observational brain signals to be 20. In summary, following the notation in Section 4.2.1, we have: $p = 20, T = 1000, q = 3, \tau^2 = 1, \sigma^2 = .1, (\rho_1, \psi_1) = (1.0012, 2), (\rho_2, \psi_2) = (1.0012, 8), (\rho_3, \psi_3) = (1.0012, 15)$.

We implement the proposed method in Section 4.2.1 and evaluated its performance. Figure A.1 shows the periodograms of the true and reconstructed signals. As we can see, the estimated source signals share exactly the same shape as the true signals.

### 4.5.2 Results on Multiple Epoch Analysis

We then evaluate the performance of the proposed method for multiple epochs. We choose 20 electrodes and 3 latent independent AR(2) processes. To model the evolution across epochs, we allow the modulus $(\rho_1^{(r)}, \rho_2^{(r)}, \rho_3^{(r)})$ increase from $(1.001, 1.001, 1.001)$ with an increment of 0.00005 as the epoch $r$ propagates. All the remaining parameters are the same as in Section 4.5.1. Figure A.2 shows the heatmap of periodogram from electrode 1 as epochs evolve.

We implement the method in this scenario and find the results satisfactory. Figure A.3 shows the periodograms of the true and estimated signals from the three underlying AR(2) processes. For the delta, alpha, and lower beta bands, we can see the peaks at the corresponding dominating frequency from the true and estimated signals. As the epochs evolve, we find that both of the true and estimated periodograms spread out around the dominating frequency. Our results show that the pattern of the periodograms from the reconstructed AR(2) process is consistent with that of the true AR(2) process.

### 4.5.3 Results for Settings Derived from the Data

Here we simulate the data using parameter setting from the motivating sequence memory study example. We use the estimated modulus $(\hat{\rho}_1^{(r)}, \hat{\rho}_2^{(r)}, \hat{\rho}_3^{(r)})$, variances $(\hat{\sigma}^{2(r)}, \hat{\tau}^{2(r)})$ and mixing matrix $\widetilde{M}$ to generate signals across 12 electrodes among 247 epochs. To evaluate the performance of E-SSM, we also apply the classical state space model (SSM) estimation methods as a benchmark in comparison with E-SSM. Specifically, we fit SSM for each single epoch and take average to obtain parameters estimates. Note that this is the approach that most of the existing methods will follow when analyzing signals with multiple epochs.

We compare mean of sum of square errors (MSE) of the parameters obtained from E-SSM and the benchmark. From Table 4.1, it is clear that E-SSM successfully captures the evolution of parameters compared to classical state space models. Among all the frequency bands, the benefits are dramatic. These results highlight the advantages of using E-SSM when signals are comprised of multiple epochs. Meanwhile, it also indicates the potential loss of information if we naively average over all the epochs when conducting analysis.

Table 4.1: Mean of sum of square errors obtained from E-SSM and SSM (benchmark)

| Parameters | E-SSM | SSM |
|---|---|---|
| $\widetilde{\Phi}$ (delta band) | $\mathbf{3.33 \times 10^{-5}}$ | $7.27 \times 10^{-5}$ |
| $\widetilde{\Phi}$ (alpha band) | $\mathbf{1.41 \times 10^{-5}}$ | $3.23 \times 10^{-5}$ |
| $\widetilde{\Phi}$ (gamma band) | $\mathbf{1.69 \times 10^{-5}}$ | $8.07 \times 10^{-5}$ |
| $\tau^2$ | $\mathbf{9.31 \times 10^{-6}}$ | $2.03 \times 10^{-4}$ |
| $\sigma^2$ | $\mathbf{1.93 \times 10^{-1}}$ | $1.93 \times 10^{-1}$ |

### 4.5.4 Sensitivity Analysis

We also conduct sensitivity analysis for the proposed E-SSM in Section 4.2.2 via simulation studies. We generate 5 latent independent source signals ( AR(2) processes) corresponding to delta ($\delta$: 0 - 4 Hertz), theta ($\theta$: 4 - 8 Hertz), alpha ($\alpha$: 8 - 12 Hertz), lower beta ($\beta$: 12 - 18 Hertz) and gamma ($\gamma$: $>$ 32 Hertz). To generate the observed signals, we *only* choose 3 latent independent AR(2) processes (delta, theta and lower beta bands) and 20 electrodes. Similar to Section 4.5.2, we allow the modulus $(\rho_1^{(r)}, \rho_2^{(r)}, \rho_3^{(r)})$ to increase from $(1.001, 1.001, 1.001)$ with an increment of 0.00005 as the epoch $r$ propagates. All the remaining parameters are the same as in Section 4.5.1. To evaluate the robustness of the proposed method, we also fit *FIVE* frequency bands into the observed signals.

Figure A.4 shows the periodogram of the generated signals from electrode 1. We fit the proposed model with *FIVE* frequency bands. Figure A.5 shows the true mixing matrix (left) and its estimation (right). From the true matrix, we can observe zero columns corresponding to "alpha" and "gamma" bands that indicate the observed signals are generated only by the three remaining bands (delta, theta and lower beta bands). From the estimation result, it is clear that the columns of "alpha" and "gamma" bands are roughly zero, which shows the proposed E-SSM successfully capture the three latent sources (delta, theta and lower beta bands) while neglecting the impacts from alpha and gamma bands. Figure A.6 shows the periodograms of the true and estimated signals from the three underlying AR(2) processes. Similar to the results in Section 4.5.2, we can see the pattern of the periodograms from the reconstructed AR(2) process is consistent with that of the true AR(2) process.

## 4.6 Analysis of LFPs Data from Olfaction Sequence Memory Study

### 4.6.1 Data Description

The LFP dataset was obtained from an experiment searching for direct evidence of coding for the memory of sequential relationships among non-spatial events (Allen et al., 2016). During the course of the experiment, rats were provided with series of five odors. All the odors were delivered in the same odor port. In each session, each rat was presented the same sequence multiple times. Each odor presentation was initiated by a nose poke and rats were required to correctly identify whether the odor was presented in the correct or incorrect sequence position (by holding their nose in the port until the signal or withdrawing before the signal, respectively). During the experiment, as rats performed the tasks, LFPs were recorded in

the CA1 pyramidal layer of the dorsal hippocampus. In total, 22 tetrodes were implanted but LFPs were only analyzed from electrodes that exhibited task-critical single-cell activity (12 in this case). The LFPs dataset in this study comprise of 12 electrodes and 247 epochs. Each epoch is recorded over 1 second, aligned to port entry, sampled at 1000 Hertz and thus has $T = 1000$ time points.

### 4.6.2 Exploratory Analysis

In our exploratory analysis, we are interested in two key goals: (1.) to determine how the original high-dimensional signals can be sufficiently represented by lower dimensional summary signals; and (2.) to assess if and how the spectral properties of the LFP signals evolve across epochs during the experiment.

To address the first question, we note the assertion in other studies (e.g., Makarova et al. (2014)) that the natural geometry of these neuronal assemblies gives rise to possible spatial segregation. This suggests that it is plausible to represent LFP data by lower dimensional summaries. In this nonspatial sequence memory study, we observe similar pattern across all the 12 electrodes. In Figure 4.5, although the power varies within each electrode, the synchrony of pattern across electrodes is still critical. For example, electrodes T13 and T14 behave almost identically. Electrodes T7, T8 and T9 also follow the same pattern during the course of experiment. Moreover, as part of this exploratory analysis, we implemented spectral principal component analysis (Brillinger, 1964). This approach is widely used in the exploratory analysis of brain imaging data (Wang et al., 2016). Figure A.7 presents the boxplots of the percentage of variability accounted by the first one and the first three components respectively. It can be shown that 3 components (mixture of delta, alpha and gamma bands) account for roughly 92% of the variability with the first component accounting

for 70%. All these findings validate the assumption that the original LFPs can be projected into low dimensional source signals without substantial loss of information. In this chapter, we will build on this preliminary analyses by giving a more specific characterization of these signal summaries or components using the AR(2) process.

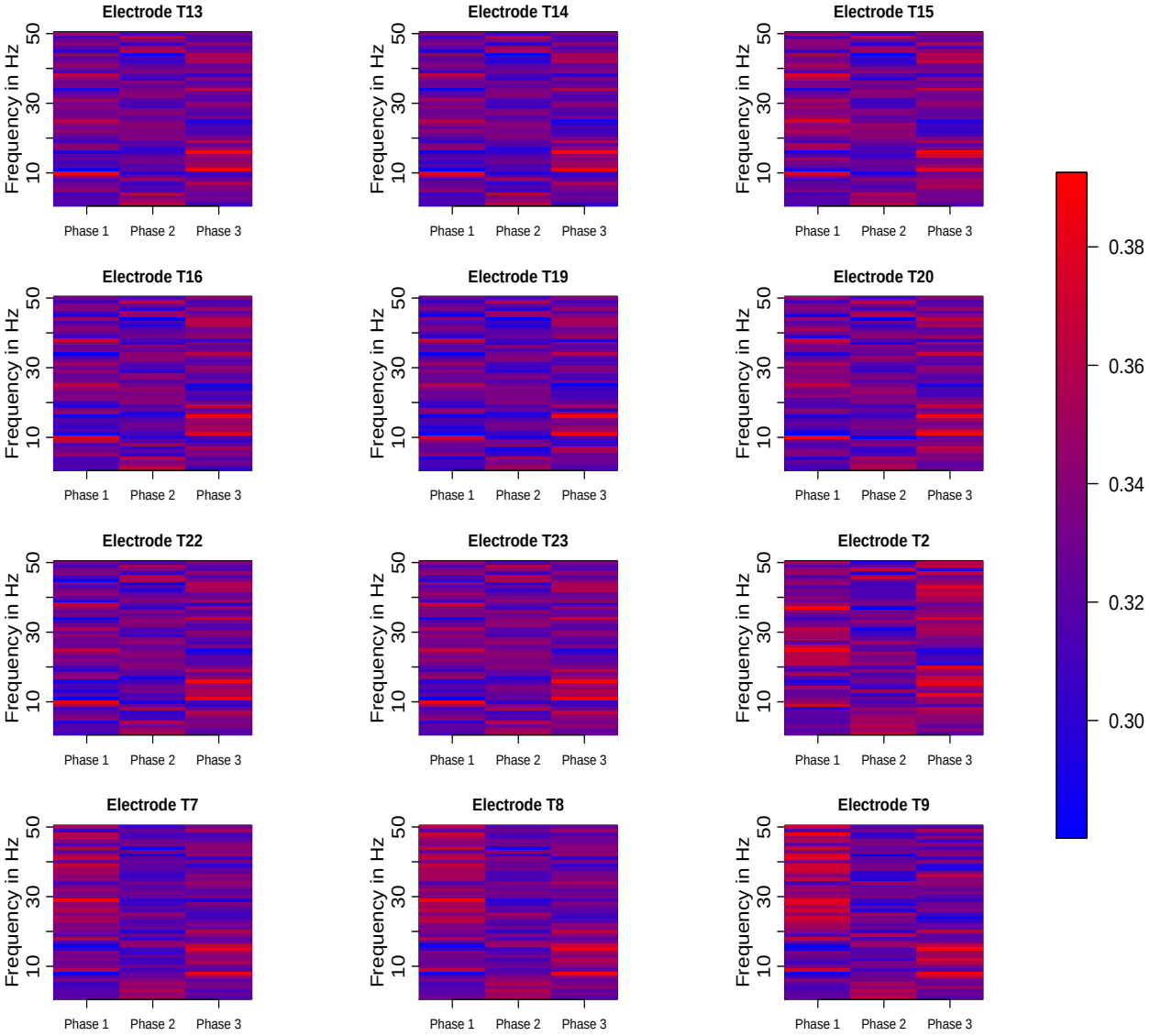



Figure 4.5: The evolution of the relative periodogram (summing up to 1 for each frequency) across the duration of experiment. Each plot displays the estimated power spectrum during the 3 phases: Phase 1 (epoch 1 - 80), Phase 2 (epoch 81 - 160) and Phase 3 (epoch 161 - 247). Frequency bands around particular hertz are present, which can be modeled as AR(2).

To gain insights into addressing the second question, we examined the LFP traceplots of the first 15 epochs at electrode T22 (Figure 4.1). It is clear that signals across various electrodes are more highly synchronized as time evolves. Similarly, from the log periodogram boxplots in Figure 4.2 across all the frequencies, we notice that the powers are quite spread out, especially at lower frequencies and the two peaks around delta and slow gamma bands. The heatmap in Figure 4.3 demonstrates the dynamics from early, middle, and late stages of the whole session. Figure 4.5 shows the evolving of the power across all the electrodes particularly on delta, alpha, and gamma bands. It shows that higher frequency bands dominate in early stage, while lower frequency bands capture more power during the evolution of experiment. In Figure 4.6, an interesting pattern emerges: the burst of gamma activity on Phase 1 of the epochs is not replicated at other phases. One possible interpretation is that odor sequence (on which the animals have had extensive training) is re-encoded early in each session, which requires high frequency (gamma) activity, but later in the session, gamma activity is regulated and other lower frequencies (delta and alpha) become more prominent. Promoted by all these results, a further study is necessary to uncover the latent lower dimensional source signals that drive the observed LFPs.

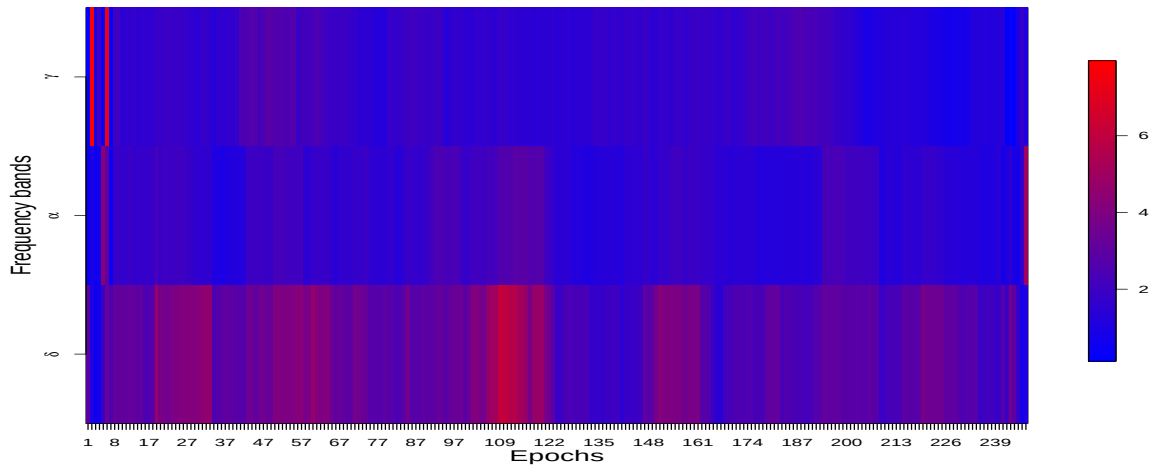



Figure 4.6: The evolution of power spectrum among delta (0-4 Hertz), alpha (8-12 Hertz) and gamma (30-35 Hertz) bands. Each band was averaged over all the electrodes.

### 4.6.3 Results and Discussion

We applied our proposed E-SSM method to this study. Figure A.8 shows time series plots of modulus (root magnitudes) corresponding to each of the three frequency bands as epochs evolve. In this plot, we could clearly identify the evolution of each individual module and a strong temporal dependence. Figure 4.7 displays the power of three latent source signals evolving during the period of experiment. We observe that the delta band captures the most power among all bands and is persistent across all phases. Gamma band power narrows down slightly towards the late phase. The alpha band attains its maximum power during the early phase and diminishes quickly in the middle stage and obtains more power in the end. There appear to be discontinuities in the delta, alpha and gamma power across the entire experiment. One interpretation to these results from the E-SSM analysis is that these on-off patterns could be just random variation. Another is that these are actual resetting of neuronal responses. This phenomenon of phase resetting in neurons is also observed in many

biological oscillators. In fact, it is believed that phase resetting plays a role in promoting neural synchrony in various brain pathways. In either case, it is imperative to be cautious about blindly assuming that the neuronal process behaves identically across epochs. Doing so could produce misleading results.

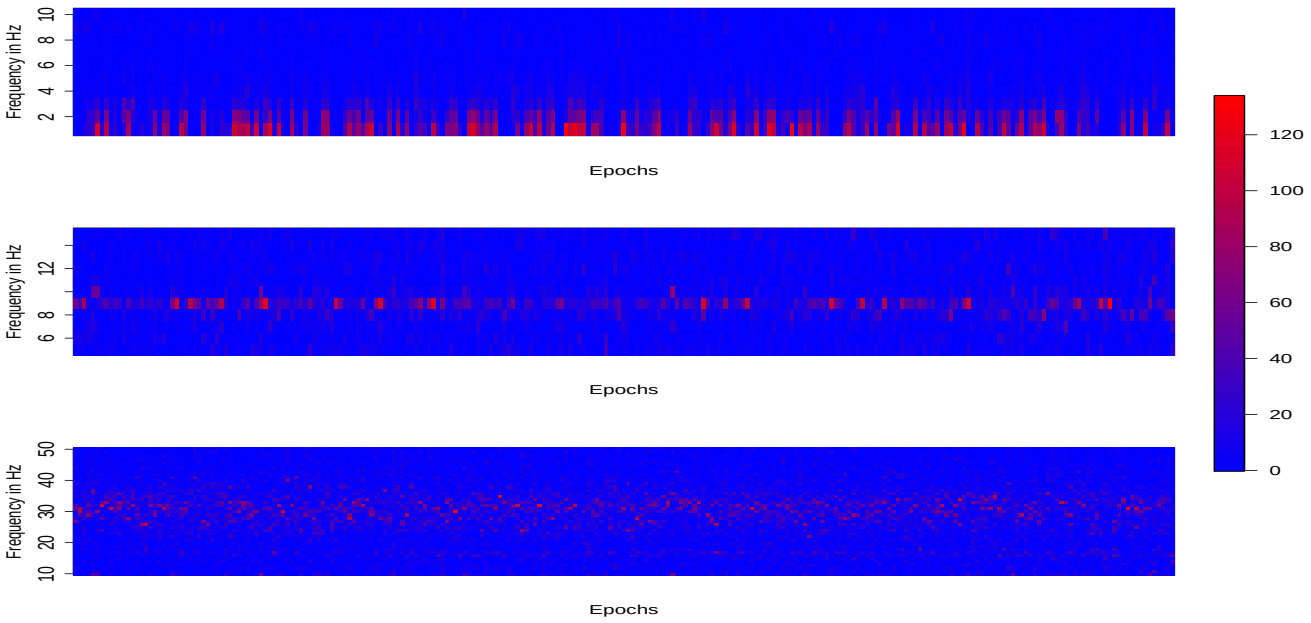



Figure 4.7: The periodograms of estimated latent AR(2) processes corresponding to delta (top), alpha (middle) and gamma (bottom) frequency band.

We also study the mixing matrix to investigate how electrodes are associated across the three frequency bands. From Figure 4.8, at delta band, electrodes T13, T14, T16, T19, T22, T23 are likely to be linked in terms of large power. Electrodes T15, T2, T7, T8 and T9 share the lowest power. At the alpha band, electrodes T16, T22 and T23 maintain the most power in contrast with electrodes T15, T2, T7-9 that obtain the lowest power. This pattern of association may result from the anatomical connections. Similarly, at gamma band, electrodes are connected in the same way as alpha band. We also used a cluster analysis on the entries of "mixing" matrix to understand the connection among electrodes. Similar

to the results shown in Figure 4.8, we are able to identify the same pattern in Figure 4.9, through the visualization of cluster analysis. At delta band, electrodes T13, T14, T16, T19, T20, T22, T23 share the same pattern while T3, T7-9 are in the same cluster. Clusters at the alpha and gamma bands are roughly identical, which coincide with the results in Figure 4.8. To the best of our knowledge, this approach (i.e., clustering of electrodes or nodes) has not be used previously for this kind of analysis. This has the potential for future explorations on synchrony among neuronal populations. Finally, we note here that the specific parametric AR(2) structure in our E-SSM has facilitated ease of interpretation of the oscillatory activity of these sources.

Model validation and diagnostics were done using sample auto-correlations (ACF) and partial auto-correlations (PACF) calculated from the residuals. Figure A.9 shows an example of those values obtained from a representative electrode. We could easily observe the uncorrelated structure among the residuals. A p-value of 0.75 based on the Ljung-Box test also provides some evidence to suggest white noise residuals and thus conclude that the proposed E-SSM fits this LFP data well.

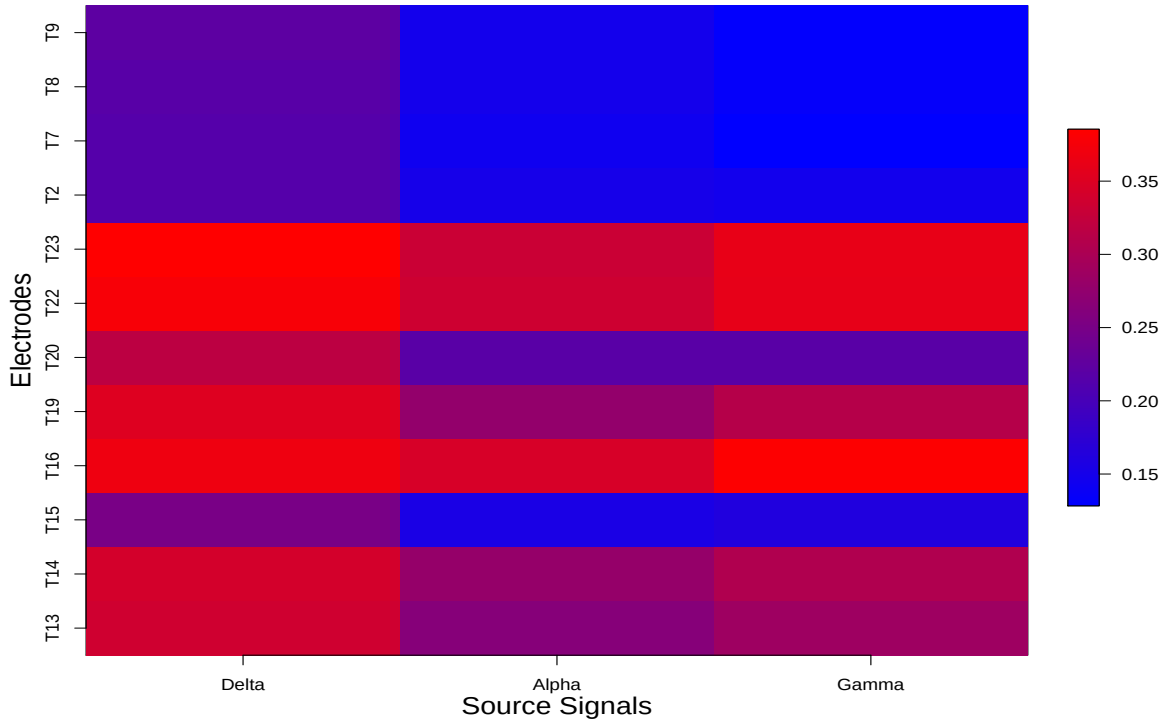
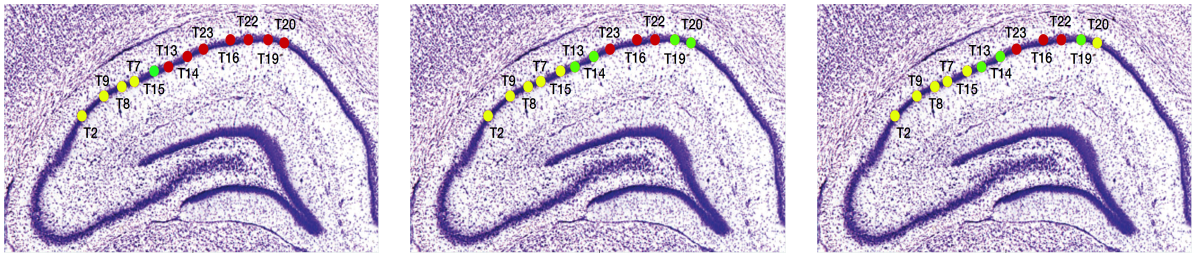



Figure 4.8: The estimated mixing matrix. Darker color represents heavier weights given by the latent processes (delta, alpha, gamma) on the LFPs.




delta frequency band alpha frequency band gamma frequency band

Figure 4.9: Cluster analysis results among all the three frequency bands. Same color indicates the same cluster.

# Chapter 5

# Penalized Probabilistic Matrix Data Clustering

## 5.1 Introduction

In this chapter, the goal is to provide a novel framework of analyzing matrix-valued data and apply it to electrophysiological signals – Local Field Potentials (LFPs). The LFPs essentially capture the integration of membrane currents across local regions of cortex (Mitzdorf et al., 1985).

In a motivating example of this chapter, researchers conducted an olfactory (non-spatial) sequence memory experiment to uncover the neuron learning process on the sequential ordering of odors (Allen et al., 2016). 12 electrodes were implanted into a rat's brain and LFPs were recorded. The entire experiment consists of 5 odors ABCDE with each corresponding to one epoch. As shown in Figure 4.1, rats were trained to identify odors denoted by ABCDE. 12 electrodes were implanted according to the schematic plot on the right. Preliminary anal-

ysis have been conducted to understand the association between the LFPs signals and the particular odor. Figure 5.1 presents the smoothed LFPs across 12 electrodes by different sequence odors and the mean signal. It can be found indisputably that the mean patterns vary dramatically across different odor, which motivates the study of analyzing "latent" structures. To take one step further, if we compare the signals among different electrodes within each odor, strong spatial dependence can be easily detected. It shows that roughly two "paradigm" can be found across electrodes especially in Sequence A, B and D. Typical cluster analysis can be done by directly lumping the signals over electrodes as vectors. However, the spatial dependence pattern would be accidently ignored in this case. This innegligible drawback inspires us to develop a statistical strategy directly on the "matrices" that respect the "row-wise" and "column-wise" dependence simultaneuously. From the literature of statistics and machine learning communities, a large amount of approaches are only applicable to vectors. As shown from the motivating example, such approaches have a few limitations: 1) Spatial and temporal correlation are not easily captured simultaneously; 2) It would be computation demanding when analyzing high-dimensional signals; 3) We would lose the interpretability from the results obtained by the manipulated "vectors". To address those issues, we propose a probabilistic model directly on the matrix-valued signals. Inspired by the work of Dawid (1981) and Dutilleul (1999), the framework is built upon a mixture matrix normal model. For the purpose of clustering signals, the advantages of using such distribution are its interpretability, conceptual and computational easiness. To account for the structures such as sparsity or low-rank, we also introduce flexible regularization terms (e.g. $\ell_1, \ell_2$ and nuclear norm). We have successfully demonstrate that by adding those penalties, the proposed approach outperform over the existing cluster method and also prevent overfitting the training data. On the foundation of the results from Fan and Li (2001), we also prove the strongly consistency of the proposed estimator.

The rest of the chapter is organized as follows. In Section 5.2, we mainly state some background knowledge of matrix normal distribution and the estimation method. In Section 5.3, we introduce the proposed penalized mixture matrix normal model and its estimation approach based on modified Expectation Maximization (EM) and one-step-late algorithms. In Section 5.4, we provide some theoretic results on the consistency of the (penalized) estimators in a restricted parameter space. In Sections 5.5, 5.6 and 5.7, we present some simulation results and apply the proposed method to two LFPs dataset obtained from odor sequence and stroke experiments.

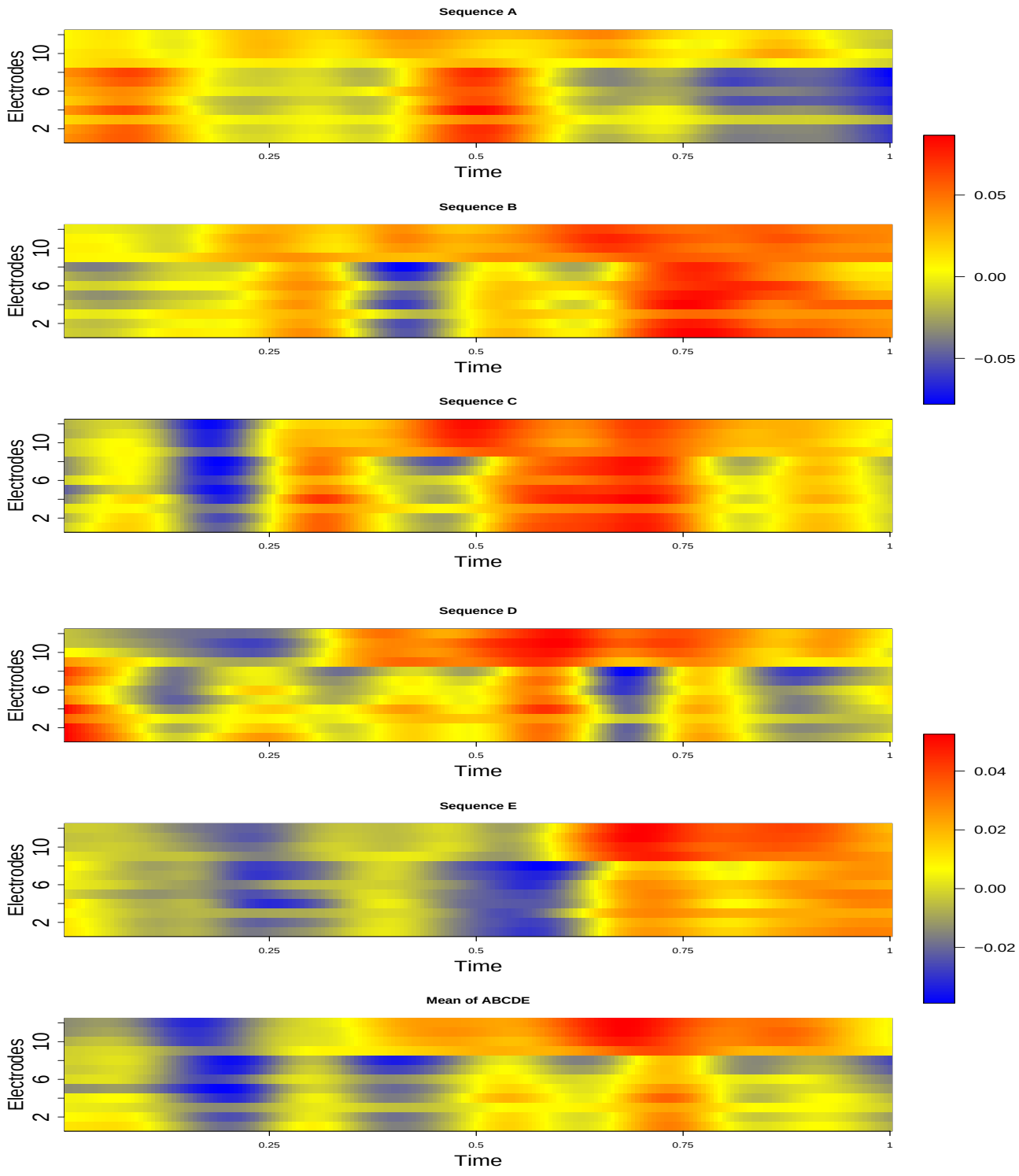



Figure 5.1: The mean LFPs across different odors.

## 5.2 Background on Matrix Normal Distribution

In this section, we mainly focus on a brief review of matrix normal distribution. In the field of modeling image or spatial-temporal data, it is natural to obtain a sequence of matrix valued observations $Y_1, Y_2, \cdots, Y_n$ with dimension $r \times p$. For example, in the case spatial-temporal data, $p, r$ denotes the spatial and temporal attributes respectively. As an extension of vector-valued data, covariance structures regarding "spatial" and "temporal" need to be considered simultaneously. Following the convention of multivariate normal distribution for vectors, $r \times p$ matrix normal distribution $MN_{r,p}(M, U, V)$ is defined as

$$f(Y|M, U, V) = \frac{\exp(-\frac{1}{2}\text{tr}(V^{-1}(Y - M)^T U^{-1}(Y - M))}{(2\pi)^{rp/2}|V|^{r/2}|U|^{p/2}}, \tag{5.1}$$

where $M \in \mathbb{R}^{r\times p}$, $U \in \mathbb{R}^{r\times r}$, $V \in \mathbb{R}^{p\times p}$ and matrices $U$ and $V$ are treated as between and within covariance matrices. With some algebraic manipulations, it can be shown that $Y \sim MN_{r,p}(M, U, V)$ if and only if

$$vec(Y) \sim N(vec(M), V \otimes U), \tag{5.2}$$

where $vec$ is vectorization operation and $\otimes$ is the Kronecker product. It should be pointed that not all the multivariate normal random variable of dimension $r \times p$ is able to convert into matrix normal distribution. Only particular covariance matrices of dimension $rp$ that follow the form in (5.2) has its corresponding matrix normal representation (Dutilleul, 1999). Such pattern is defined as "separable" (Cressie, 2015). In the application of electrophysiological data analysis, traditional statistical methods such as state space model (Gao et al., 2016), vector autoregressive model (Derado et al., 2010) all meet the "separable" assumption. Moreover, Reinsel (1982) showed it lead to efficient inference when incorporating such structure into analysis.

On Estimating the Parameters

Suppose that $Y_1, Y_2, \cdots, Y_n$ are i.i.d random samples from matrix normal distribution $MN_{r,p}(M, U, V)$, the log-likelihood is given by

$$\ell(M, U, V) = -\frac{npr}{2}\log 2\pi - \frac{nr}{2}\log|V| - \frac{np}{2}\log|U| - \frac{1}{2}\sum_{i=1}^{n}\mathrm{tr}(V^{-1}(Y_i - M)^T U^{-1}(Y_i - M)). \quad (5.3)$$

After some matrix derivatives manipulation, the maximum likelihood estimator (MLE) yields

$$\begin{aligned}
\hat{M} &= \sum_{i=1}^{n} Y_i = \bar{Y} \\
\hat{U} &= \frac{1}{np}\sum_{i=1}^{n}(Y_i - \bar{Y})\hat{V}^{-1}(Y_i - \bar{Y})' \\
\hat{V} &= \frac{1}{nr}\sum_{i=1}^{n}(Y_i - \bar{Y})'\hat{U}^{-1}(Y_i - \bar{Y})
\end{aligned} \quad (5.4)$$

It is obvious that there are some identifiability issues since one can simply replace $\hat{U}, \hat{V}$ by $c\hat{U}, \frac{1}{c}\hat{V}$ to satisfy Equations (5.4) (Dutilleul, 1999). However, the Kronecker product $\hat{U} \otimes \hat{V}$ will remain invariant and we will mainly focus on the mean parameter $M$ throughout this study.

There is no close form for $\hat{U}, \hat{V}$. Alternatively, one can utilize iterative algorithms to achieve those values numerically. The algorithm is summarized as follows. Note that this approach is also used as an update step in Section 5.3.

*Remark* 6. Note that $||.||$ denotes the frobenius norm. $\mathrm{diag}(1, \cdots, 1)$ represents the identity matrix of dimension $r$.

**Algorithm 5** The MLE of covariance matrices

**Input**: $\mathbf{Y} = \{Y_1, Y_2, \cdots, Y_n\}$, $\tau$ (tolerance level), Max-iter
**Initializing**: iter $= 0$, $U_0 = \text{diag}(1, \cdots, 1)$, $V_0 = \frac{1}{nr}\sum_{i=1}^{n}(Y_i - \bar{Y})'{U_0}^{-1}(Y_i - \bar{Y})$
$U_1 = \frac{1}{np}\sum_{i=1}^{n}(Y_i - \bar{Y}){V_0}^{-1}(Y_i - \bar{Y})'$, $V_1 = \frac{1}{nr}\sum_{i=1}^{n}(Y_i - \bar{Y})'{U_1}^{-1}(Y_i - \bar{Y})$
**While** (iter $<$ Max-iter or $||U_1 - U_0|| > \tau$ or $||V_1 - V_0|| > \tau$)
**Repeat**
$U_0 := U_1$
$V_0 := V_1$
$U_1 = \frac{1}{np}\sum_{i=1}^{n}(Y_i - \bar{Y}){V_0}^{-1}(Y_i - \bar{Y})'$
$V_1 = \frac{1}{nr}\sum_{i=1}^{n}(Y_i - \bar{Y})'{U_1}^{-1}(Y_i - \bar{Y})$
iter := iter $+ 1$
**Return**: $\hat{U} := U_1$, $\hat{V} := V_1$

## 5.3 Penalized Mixture Matrix Normal Clustering

### 5.3.1 Mixture Matrix Normal Models

Suppose the observed matrix-valued data $Y_1, \cdots, Y_n$ are obtained from a population with $k$ "regimes". The probability density function is essentially a mixture of matrix normal densities. For simplicity, if we write $\Theta_j = (M_j, U_j, V_j)$, and the prior association densities as $\pi_j, j = 1, \cdots, k$, then the marginal density function of $Y_i$ can be written as

$$f(Y_i|\Theta_1, \cdots, \Theta_k, \pi_1, \cdots, \pi_k) = \sum_{j=1}^{k} \pi_j f(Y_i|\Theta_j), \tag{5.5}$$

where $f(Y_i|\Theta_j)$ is shown in Equation (5.1) and $\sum_{j=1}^{k} \pi_j = 1$. The observed log-likelihood yields

$$\ell_{obs}(\Theta_1, \cdots, \Theta_k, \pi_1, \cdots, \pi_k) = \sum_{i=1}^{n} \log\{\sum_{j=1}^{k} \pi_j f(Y_i|\Theta_j)\}. \tag{5.6}$$

On Estimating the Parameters

Expectation Maximization (EM) algorithm (Dempster et al., 1977) can be efficiently used to provide estimations of parameter. In general, it is an iterative approach consisting of expectation and maximization steps.

In E-step, a posterior probability of observation $Y_i$ derives from $j-th$ cluster is calculated by Bayes Theorem that

$$\alpha_{ij} = \frac{\pi_j f(Y_i|\Theta_j)}{\sum_{l=1}^{k} \pi_l f(Y_i|\Theta_l)}. \tag{5.7}$$

In M-step, optimal values are obtained by solving the non-constraint optimization problem that

$$\hat{\Theta} = \arg\max_{\Theta} \sum_{i=1}^{n} \sum_{j=1}^{k} \alpha_{ij} \log\{\pi_j f(Y_i|\Theta_j)\}$$

After some matrix derivatives and algebra manipulations, we can obtain the explicit solutions that

$$\begin{aligned}
\hat{\pi}_j &= \frac{\sum_{i=1}^{n} \alpha_{ij}}{n} \\
\hat{M}_j &= \frac{\sum_{i=1}^{n} \alpha_{ij} Y_i}{\sum_{i=1}^{n} \alpha_{ij}} \\
\hat{U}_j &= \frac{\sum_{i=1}^{n} \alpha_{ij} (Y_i - \hat{M}_j) \hat{V}_j^{-1} (Y_i - \hat{M}_j)'}{p \sum_{i=1}^{n} \alpha_{ij}} \\
\hat{V}_j &= \frac{\sum_{i=1}^{n} \alpha_{ij} (Y_i - \hat{M}_j)' \hat{U}_j^{-1} (Y_i - \hat{M}_j)}{r \sum_{i=1}^{n} \alpha_{ij}}
\end{aligned} \tag{5.8}$$

Note that $\hat{U}_j, \hat{V}_j$ can be obtained numerically using the similar method to Algorithm 5.

### 5.3.2 Penalized Mixture Matrix Normal Models

It is quite common that we have some prior information on parameters $\Theta$. This could originate from the sparsity, rank, smoothness or a prior probability density on parameters (Green, 1990). To this end, it is natural to add a regularization term to the likelihood and alternatively, maximum penalize likelihood estimate should be obtained. Specifically, we penalized log-likelihood follows

$$Q(\lambda, \Theta_1, \cdots, \Theta_k, \pi_1, \cdots, \pi_k) = \sum_{i=1}^{n} \log\{\sum_{j=1}^{k} \pi_j f(Y_i|\Theta_j)\} - \lambda P(\Theta), \tag{5.9}$$

where $P(.)$ is some penalized function. Examples can be logarithm of probability density functions, $\ell_1, \ell_2$ norms, nuclear norm etc.

On Estimating the Parameters

Similar to the approach in Section 5.3.1, we propose a modified EM algorithm to estimate the parameters. The E-step can be easily achieved by Equation (5.7). The M-step boils down to the optimization problem that

$$\hat{\Theta} = \arg\max_{\Theta} \sum_{i=1}^{n} \sum_{j=1}^{k} \alpha_{ij} \log\{\pi_j f(Y_i|\Theta_j)\} - \lambda P(\Theta). \tag{5.10}$$

In contrast to the case without penalty, the solution $\hat{\Theta}$ may not have an explicit form. Lange (1995) proposed a gradient method related to EM algorithm. It replaces the M-step by conducting one iteration of Newton's method. Theoretic results on the convergence were also discussed. As an alternative approach, other methods including surrogate functions (Lange et al., 2000), overrelaxed EM algorithm (Yu, 2012) were introduced to this issue.

Throughout this article, we mainly focus on three types of penalties: $\ell_1, \ell_2$ and nuclear

norm. Pan and Shen (2007) introduced $\ell_1$ penalty to the mean parameters in the setting of mixture univariate normal models. An explicit form of the M-step is derived using sub-gradient. Green (1990) developed "one-step-late" (OSL) algorithm that can be applied to more general case. Inspire by the aforementioned results, we developed a sub-gradient update for $\ell_1$ norm and OSL step for $\ell_2$ and nuclear norms.

In the case of $\ell_1$ norm penalty, the update of $M_j$ is the optimal value that maximizes

$$\sum_{i=1}^{n}\sum_{j=1}^{k}\alpha_{ij}\log\{\pi_j f(Y_i|\Theta_j)\} - \lambda\sum_{j=1}^{k}||M_j||_1.$$

Following a similar derivation by Pan and Shen (2007), the update step of $M_j$ has the form that

$$\hat{M_j} = \text{sign}(\tilde{M_j})(|\tilde{M_j}| - \frac{\lambda}{\sum_{i=1}^{n}\alpha_{i,j}}U_i\mathbf{1}_{r\times p}V_i)_+, \tag{5.11}$$

where $\tilde{M_j} = \frac{\sum_{i=1}^{n}\alpha_{i,j}Y_i}{\sum_{i=1}^{n}\alpha_{i,j}}$ is the update for $M_i$ without penalty, $B_+ = max(B,0)$, $\mathbf{1}_{r\times p}$ is a matrix of all 1's. sign() and $(.)_+$ are all component-wise operators.

In the case of $\ell_2$ norm penalty, the objective function is derived to be

$$Q_{\ell_2}(\pi,\Theta) = \sum_{i=1}^{n}\sum_{j=1}^{k}\alpha_{ij}\log\{\pi_j f(Y_i|\Theta_j)\} - \lambda\sum_{j=1}^{k}||M_j||_2.$$

After matrix derivative manipulations, we have

$$\frac{\partial Q_{\ell_2}(\pi,\Theta)}{\partial M_j} = U_j^{-1}\sum_{i=1}^{n}\alpha_{i,j}(Y_i - M_j)V_j^{-1} - 2\lambda M_j,$$

The update step of $M_i$ follows the form

$$\hat{M}_j = \tilde{M}_j - \frac{2\lambda}{\sum_{i=1}^{n} \alpha_{ij}} U_j M_j V_j, \tag{5.12}$$

where $U_j, M_j, V_j$ are the update from the previous step.

For the case of *nuclear* norm penalty, similar derivation yields

$$\hat{M}_j = \tilde{M}_j - \frac{\lambda}{\sum_{i=1}^{n} \alpha_{ij}} U_j \Phi_j \Omega_j' V_j, \tag{5.13}$$

where $M_j$ has the singular value decomposition $M_j = \Phi_j \Lambda_j \Omega_j'$.

As a summary, the proposed estimation approach involves algorithms of initialization and alternating from E-step and M-step. Details are presented as follows

**I. (Initialization)** We start with vectorizing the original matrix-valued observations $Y_1, \cdots, Y_n$ and applied k means to achieve the initial cluster membership values, written as $S_1, \cdots, S_k$, where $S_j = \{i \mid Y_i \text{ in j-th cluster}\}$. Note that we can relax this step by randomly assign clusters to those observations. Then for each cluster, the initial value of $\Theta_i$ can be obtained following the same manner in Section 5.2. $\pi_j$ can be directly estimated by $\hat{\pi}_j = \frac{|S_j|}{n}$

**II. (E-step)** We update the posterior membership by

$$\alpha_{ij} = \frac{\pi_j f(Y_i|\Theta_j)}{\sum_{l=1}^{k} \pi_l f(Y_i|\Theta_l)}.$$

**III. (M-step)** The mean parameter $M_j$ with respect to various penalties is updated by the Equations (5.11), (5.12) and (5.13) respectively. Updates for $\pi_j, U_j, V_j$ follows Equations (5.8) and Algorithm 5 is also utilized.

**IV. (Stopping criteria)** The iterative approach will alternate by **I.** and **II.** until certain iterations have been reached or the frobenius norm change of the mean parameter $M_j$ is small enough.

On Choosing the Number of Clusters

A key question in the proposed method is to determine the number of clusters. Inspired by the approach proposed by Smyth (2000), we introduce cross validated penalized likelihood (CVPL) as the key measure. Without loss of generality, let us denote $f(.), f_k(.)$ as the "true" and $k$ mixture probability density functions, $\Psi, \Psi_k$ as the corresponding parameters. We split the dataset $\boldsymbol{Y} = \{Y_1, \cdots, Y_n\}$ into training and testing groups denoted by $\boldsymbol{Y}_{train}, \boldsymbol{Y}_{test}$. If we write the averaged penalized negative log-likelihood as

$$\ell_k = -\frac{1}{N_{test}} \left(\ell_{obs}(\Psi_k(\boldsymbol{Y}_{train})|\boldsymbol{Y}_{test}) - \lambda P(\Psi_k)\right) \tag{5.14}$$

It can be shown directly that

$$E(\ell_k) = \int \log \frac{f(\boldsymbol{Y})}{\tilde{f}_k(\boldsymbol{Y})} f(\boldsymbol{Y}) d\boldsymbol{Y} + C, \tag{5.15}$$

where $\tilde{f}_k(\boldsymbol{Y}) = \exp\{\log f_k(\boldsymbol{Y}) - \lambda P(\Psi_k)\}$. It shows that the expectation of $\ell_k$ is the Kullbak-Leibler (KL) distance between $f(.)$ and the exponential penalized k mixture likelihood up to some constant. Derived from this result, we propose CVPL to determine the optimal number of clusters.

## 5.4 Theory

In this section, we first show some theoretic results on the consistency of the maximum likelihood estimator without regularizations. In order to guarantee a constrained (global) maximum likelihood formulation, we define the constrained parameter space $\Psi^{d_1,d_2}$ as

$$\begin{aligned}\Psi^{d_1,d_2} = \{&\pi_1, \cdots, \pi_k \in \mathbb{R},\ M_1, \cdots, M_k \in \mathbb{R}^{r*p},\ V_1 \otimes U_1, \cdots, V_k \otimes U_k \in \mathbb{R}^{rp*rp}: \\ &\min_{1 \leqslant h \neq j \leqslant k} \rho(U_h U_j^{-1}) \geqslant d_1 > 0, \min_{1 \leqslant h' \neq j' \leqslant k} \rho(V_{h'} V_{j'}^{-1}) \geqslant d_2 > 0,\ \sum_{i=1}^{k} \pi_i = 1,\ \pi_l > 0, \\ &\rho(U_l) > 0,\ \rho(V_l) > 0 \text{ for } l = 1, \cdots, k\},\end{aligned} \tag{5.16}$$

where $d_1, d_2 \in (0, 1]$, $\rho(.)$ denotes the minimum eigenvalue.

**Theorem 1.** *Let $Y_1, \cdots, Y_n$ be random samples from a mixture matrix normal distribution* (5.5)*, then for $d_1, d_2 \in (0, 1]$, there exists a constrained global maximizer $\hat{\psi}^n$ of the log-likelihood* (5.6) *over $\Psi^{d_1,d_2}$. Moreover, $\hat{\psi}^n$ is also strongly consistent in $\Psi^{d_1,d_2}$.*

*Proof.* First, we state the fact that

$$\min_{1 \leqslant h \neq j \leqslant k} \rho(\Sigma_h \Sigma_j^{-1}) \geqslant \min_{1 \leqslant h \neq j \leqslant k} \rho(V_h V_j^{-1}) * \min_{1 \leqslant h' \neq j' \leqslant k} \rho(U_{h'} U_{j'}^{-1}), \tag{5.17}$$

where $\Sigma_h = V_h \otimes U_h$.

Actually, it follows directly from the property that

$$\begin{aligned}\rho(\Sigma_h \Sigma_j^{-1}) &= \rho\big[(V_h \otimes U_h)(V_j \otimes V_h)^{-1}\big] \\ &= \rho\big[(V_h V_j^{-1}) \otimes (U_h U_j^{-1})\big] \\ &= \rho(V_h V_j^{-1}) * \rho(U_h U_j^{-1}),\end{aligned}$$

where the equalities follow the results in Schacke (2004). We denote the parameter space $\widetilde{\Psi}^d$ as

$$\begin{aligned}\widetilde{\Psi}^d = \{&\pi_1, \cdots, \pi_k,\ M_1, \cdots, M_k,\ V_1 \otimes U_1, \cdots, V_k \otimes U_k : \min_{1 \leqslant h \neq j \leqslant k} \rho(\Sigma_h \Sigma_j^{-1}) \geqslant d > 0, \\ &\sum_{i=1}^{k} \pi_i = 1,\ d_1 d_2 = d,\ \pi_l > 0,\ \rho(\Sigma_l) > 0 \text{ for } l = 1, \cdots, k\},\end{aligned} \tag{5.18}$$

then due to the definition (5.2) and results in (Hathaway, 1985), there exists a global constraint maximizer of (5.6) $\hat{\psi}^n$ over $\widetilde{\Psi}^d$ so that $\ell_{obs}(\hat{\psi}^n) = \sup_{\widetilde{\Psi}^d} \ell_{obs}(\psi)$ and there exists a compact set $S \in \widetilde{\Psi}^d$ such that $\hat{\psi}^n \in S$ and $\sup_S \ell_{obs}(\psi) = \sup_{\widetilde{\Psi}^d} \ell_{obs}(\psi)$. Moreover, the fact (5.17) implies that $\sup_{\widetilde{\Psi}^d} \ell_{obs}(\psi) \geqslant \sup_{\Psi^{d_1,d_2}} \ell_{obs}(\psi)$ for any $d_1, d_2$. Due to the boundedness of $S$, it can be shown by contradiction that there exist $d_1, d_2$ so that $S \in \Psi^{d_1,d_2}$. Thus, we have that $\sup_S \ell_{obs}(\psi) = \sup_{\widetilde{\Psi}^d} \ell_{obs}(\psi) \geqslant \sup_{\Psi^{d_1,d_2}} \ell_{obs}(\psi) \geqslant \sup_S \ell_{obs}(\psi)$, which completes the proof of the first part. To show the strongly consistency, the same argument can be utilized as in Hathaway (1985) with the fact of definition (5.2). □

*Remark* 7. Note that the preceding results hold for unidentifiable case resulting from Hathaway (1985).

*Remark* 8. The condition in (5.16) is not easy to check in practice. One might bound all the eigenvalues within an interval $(a, b)$ for numerical stability.

Next, we will show that under wild conditions, there also exists a root-n consistent penalized likelihood estimator of (5.9). We first define the parameter space denoted as $\bar{\Psi}^{d_1,d_2}$ where

$$\begin{aligned}\bar{\Psi}^{d_1,d_2} = \{&\pi_1, \cdots, \pi_k, M_1, \cdots, M_k, V_1 \otimes U_1, \cdots, V_k \otimes U_k \in \Psi^{d_1,d_2} : \frac{\sigma_i(U_h)}{\sigma_i(V_h)} = c_h \\ &\text{for } i = 1, \cdots, \min\{r, p\}, h = 1, \cdots, k\},\end{aligned} \tag{5.19}$$

where $\sigma_i(U_h)$ denotes the $i$th eigenvalue of matrix $U_h$ and $c_h$ is a positive constant.

We state the condition (A) as

(A) Let $\beta = \left(vec(\psi^1)', \cdots, vec(\psi^k)'\right)'$, where $\psi^i$ denotes the parameters $(\pi_i, M_i, V_i, U_i)$ in $i$th component. The Fisher information matrix $I(\beta)$ is finite in the parameter space $\bar{\Psi}^{d_1,d_2}$ and positive definite at $\beta = vec(\psi_0)$.

**Theorem 2.** *Let $Y_1, \cdots, Y_n$ be random samples from a mixture matrix normal distribution* (5.5)*, in the case of $\ell_1$ and $\ell_2$ norm penalties, under condition (A), if $\lambda = O_p(n^\eta)$, $0 < \eta \leqslant \frac{1}{2}$, then there exists a local maximizer $\hat{\zeta}$ of the penalized likelihood* (5.9) *such that $||\hat{\zeta} - \psi_0|| = O_p(n^{-1/2})$ in the parameter space $\bar{\Psi}^{d_1,d_2}$, where $\psi_0$ is the true parameter in $\bar{\Psi}^{d_1,d_2}$.*

*Proof.* The proof can be directly adapted from the argument of Theorem 1 proposed by Fan and Li (2001). It suffices to check the conditions in their proof. For the first condition, all the assumptions are trial except the identifiability issue. Actually, since $\sigma_i(V_h \otimes U_h) = \sigma_{i'}(V_h)\sigma_{i''}(U_h)$, by fixing the ratio of eigenvalues as shown in (5.19), there exists a unique eigenvalue pair of $\sigma_{i'}(V_h), \sigma_{i''}(U_h)$ for a given value of $\sigma_i(V_h \otimes U_h)$. Thus $V_h \otimes U_h = V_h' \otimes U_h'$ implies $V_h = V_h'$ and $U_h = U_h'$. The identifiability property then directly follow given the results from Yakowitz and Spragins (1968). For the second condition, our assumption (A) directly implies that. For the last condition, it holds from the compactness of the parameter space $\bar{\Psi}^{d_1,d_2}$. □

## 5.5 Simulations

### 5.5.1 Results on Choosing the Number of Clusters

In this section, we evaluate the effectiveness of the proposed cross validated penalized likelihood (CVPL) in different scenarios. We generate two clusters of signal that follow matrix normal distribution with mean structures shown in Figure 5.2. The row-wise and column-wise covariance matrices follow an autoregressive setting where $cov\{Y_{k_1,l_1}, Y_{k_2,l_2}\} = 0.9^{|k_1-k_2|+|l_1-l_2|}, 1 \leqslant k_i \leqslant r, 1 \leqslant l_i \leqslant p$. The proportion for both of the clusters is equal. In Scenario I, we set the number of signals $n = 100$ with $r = p = 60$. In Scenario II, we let $n = 50,\ r = p = 30$. 200 simulations were conducted for each of the two cases.

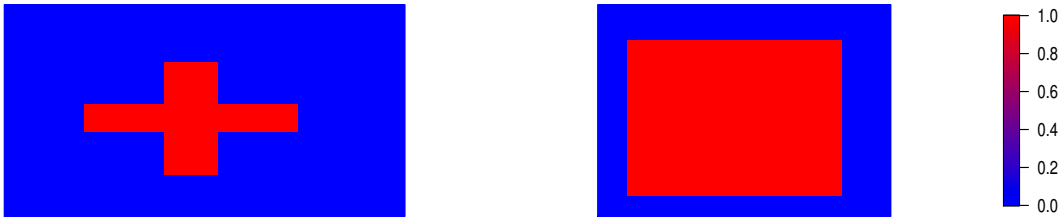



Figure 5.2: The mean structure of the two clusters.

We applied the proposed method to the simulated dataset. $L_1, L_2$ and Nuclear penalties were all implemented. As is shown in Table 5.1, among all the penalties, $\lambda$ and sample sizes, the proposed CVPL values suggest the true number of cluster. Comparing $L_1$ with $L_2$ penalty in Scenario I, the outperformance of $k = 2$ among all the other clusters are higher with $L_1$ penalty, which results from the sparsity of the two mean structures. When the sample size decreases as in Scenario II, such pattern becomes less obvious. It shows that the

smaller dimension of images attenuates the discrepancy between $L1$ and $L_2$ regularizations. In the setting of Nuclear regularization, the proposed CVPL value leads to the true number of clusters, which is due to the low rank of mean structures.

Table 5.1: The cross validated penalized likelihood (CVPL) values obtained from different number of clusters and penalties under two scenarios.

| Penalty | $\lambda$ | CVPL (Scenario I) | | | CVPL (Scenario II) | | |
|---|---|---|---|---|---|---|---|
| | | $k=2$ | $k=3$ | $k=4$ | $k=2$ | $k=3$ | $k=4$ |
| L1 | 0.5 | 2.345* | 2.337 | 2.333 | 0.458* | 0.453 | 0.451 |
| | 1 | 2.344* | 2.336 | 2.330 | 0.457* | 0.455 | 0.452 |
| | 1.5 | 2.341* | 2.337 | 2.332 | 0.458* | 0.457 | 0.455 |
| L2 | 0.5 | 2.351* | 2.349 | 2.344 | 0.462* | 0.449 | 0.431 |
| | 1 | 2.352* | 2.350 | 2.345 | 0.450* | 0.434 | 0.419 |
| | 1.5 | 2.352* | 2.349 | 2.344 | 0.446* | 0.429 | 0.413 |
| Nuclear | 0.5 | 2.351* | 2.348 | 2.343 | 0.461* | 0.456 | 0.452 |
| | 1 | 2.351* | 2.348 | 2.344 | 0.461* | 0.457 | 0.452 |
| | 1.5 | 2.353* | 2.349 | 2.345 | 0.460* | 0.456 | 0.454 |

* The highest values across different scenarios ($\times 10^5$)

### 5.5.2 Results on Comparing with K-Means

This section is contributed to compare the proposed approach with K means. Similar to Section 5.5.1, we generated signals using the same mean and covariance structures. In Scenario III, the sample size is set to be 50 and the dimension of images $20 * 20$. In Scenario IV, we increase the sample size to 100 and the dimension to $60 * 60$. To compare the results obtained from the two underlying approaches, we calculate the adjusted random index (Milligan and Cooper, 1986) and accuracy. We repeat the procedure 200 times for this simulation study.

Results are summarized in Table 5.2. In Scenario III where the size is relatively low, the benefit of the proposed method is critical compared to K means. The ARI and accuracy values are almost double of the results obtained from K means. When it comes to larger sample size, which is presented as Scenario IV, the gain is also apparent. Among all the regularizations, the $L1$ penalty performs superiously due to the sparsity of the generated signals.

Table 5.2: The adjusted random index (ARI) and accuracy obtained from the proposed method and K means under Scenario III and IV.

| Penalty | $\lambda$ | ARI (Scenario III) | | Accuracy | | ARI (Scenario IV) | | Accuracy | |
|---|---|---|---|---|---|---|---|---|---|
| | | our method | kmeans | our method | kmeans | our method | kmeans | our method | kmeans |
| L1 | 0 | 0.867 | 0.513 | 0.882 | 0.626 | 0.644 | 0.517 | 0.696 | 0.607 |
| | 0.5 | 0.924 | | 0.938 | | 0.691 | | 0.744 | |
| | 1 | 0.962 | | 0.980 | | 0.781 | | 0.822 | |
| | 1.5 | 0.966 | | 0.985 | | 0.788 | | 0.824 | |
| L2 | 0.5 | 0.879 | 0.514 | 0.892 | 0.623 | 0.632 | 0.518 | 0.687 | 0.607 |
| | 1 | 0.907 | | 0.918 | | 0.665 | | 0.715 | |
| | 1.5 | 0.868 | | 0.881 | | 0.788 | | 0.824 | |
| Nuclear | 0.5 | 0.898 | 0.515 | 0.909 | 0.623 | 0.645 | 0.516 | 0.697 | 0.607 |
| | 1 | 0.860 | | 0.876 | | 0.660 | | 0.710 | |
| | 1.5 | 0.884 | | 0.897 | | 0.636 | | 0.687 | |

## 5.6 Analysis of Odor Memory Data

In this section, we focus on analyzing a LFP dataset from a memory coding experiment on non-spatial events (Allen et al., 2016). Rats were trained to identify a series of five odors during the experiment. Each of the odors was presented through an odor port. In most of the cases, those five odors were in the same sequence ("*in-sequence*" odors) while

there were some violations ("*out-sequence*" odors). For example, odor sequence $ABCDE$ is an "in-sequence" odor yet $ABBDE$ is an "out-sequence" odor. Rats were required to poke and hold their nose in the port to correctly identify whether the odors were "in" or "out" sequence. Throughout the experiment, spike and LFP data were collected. 22 electrodes were implanted in the CA1 pyramidal layer of the dorsal hippocampus, among which we only focus on 12 electrodes exhibiting task-critical single-cell activity. The whole LFP dataset contains 247 trials with a sampling rate 1000 Hertz and $T = 2000$ time points. Figure 5.3 exposes a snapshot of the LFP signals across 12 electrodes.

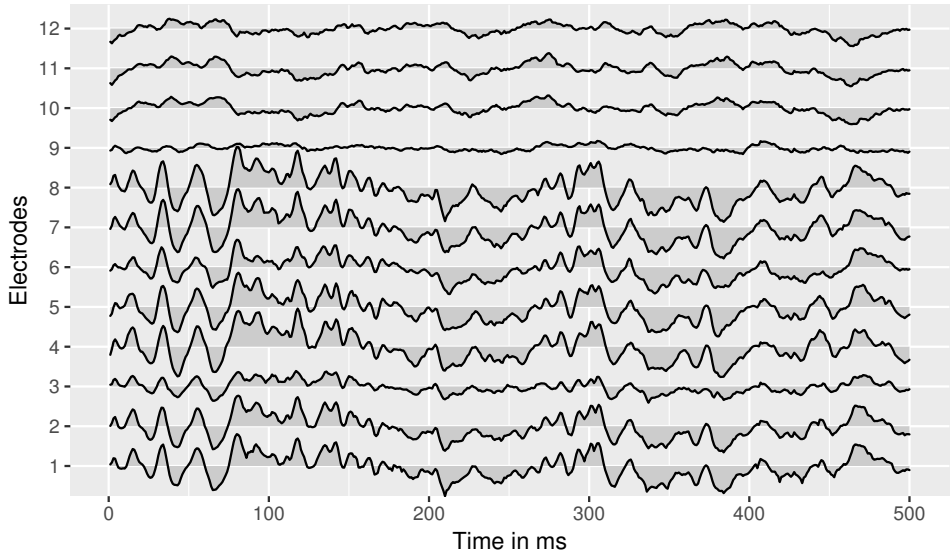



Figure 5.3: Time series plot of LFP signals across 12 electrodes in trial 1. The plot only presents the first 500 time points.

### 5.6.1 Time Domain Analysis on Imaging Clustering

We applied the proposed clustering method to the LFP dataset with 247 trials to identify underlying patterns. As an initial step, we focus on time domain to uncover the association between raw multi-channel signals with "in-sequence" or "out-sequence" patterns. We

implemented the proposed method to the raw LFP signals across all the 247 trials.

Table 5.3: The cross validated penalized likelihood values and the adjusted random index obtained across different number of clusters among all the three penalties.

| Penalty | $\lambda$ | CVPL | | | ARI | |
|---|---|---|---|---|---|---|
| | | $k=2$ | $k=3$ | $k=4$ | our method | K means |
| L1 | 0 | 1.290* | 1.285 | 1.281 | 0.768 | 0.499 |
| | 0.5 | 1.253 | 1.253* | 1.246 | 0.786 | |
| | 1 | 1.243* | 1.206 | 1.204 | 0.768 | |
| | 1.5 | 1.249* | 1.234 | 1.218 | 0.780 | |
| L2 | 0.5 | 1.302* | 1.107 | 1.240 | 0.768 | 0.510 |
| | 1 | 1.301* | 1.027 | 1.202 | 0.774 | |
| | 1.5 | 1.298* | 1.189 | 1.235 | 0.756 | |
| Nuclear | 0.5 | 1.309* | 1.299 | 1.274 | 0.756 | 0.498 |
| | 1 | 1.299* | 1.287 | 1.277 | 0.733 | |
| | 1.5 | 1.290* | 1.286 | 1.214 | 0.711 | |

* The highest CVPL value ($\times 10^5$).

Table 5.3 summarizes the cross validated penalized likelihood values among different number of clusters and penalties. It is obvious that 2 clusters are mostly suggested especially in the case of L2 or nuclear norm regularization. These findings motivate us to further investigate the cluster results with respect to the "in/out sequence" patterns. Table 5.3 shows such association. The adjusted random index was related to the true label of "in/out sequence" patterns. Comparing to K means, the proposed method outperforms in detecting the latent structure representing "in" or "out" sequences. Filter the LFPs by all the "in-sequence" signals.

As a further step, researchers are also interested in understanding how LFP signals are related to rat's correctness in this experiment. Due to the small size of "out" sequence trials, we only focus on those "in" sequence trials. In this way, we are able to investigate on the "sensitivity" (true positive rate) of the experiment.

Table 5.4: The cross validated penalized likelihood values obtained across different number of clusters on all the "in-sequence" trials.

| Penalty | $\lambda$ | CVPL | | | | ARI | |
|---|---|---|---|---|---|---|---|
| | | $k=2$ | $k=3$ | $k=4$ | $k=5$ | our method | K means |
| L1 | 0 | 1.135* | 1.135* | 1.126 | 1.131 | 0.762 | 0.506 |
| | 0.5 | 1.103* | 1.076 | 1.084 | 1.094* | 0.783 | |
| | 1 | 1.099* | 1.070 | 1.077 | 1.136* | 0.783 | |
| | 1.5 | 1.107* | 1.1078 | 1.118* | 1.068 | 0.609 | |
| L2 | 0.5 | 1.142* | 1.139 | 0.885 | 1.144* | 0.769 | 0.499 |
| | 1 | 1.139* | 1.016 | 1.101* | 0.986 | 0.743 | |
| | 1.5 | 1.150* | 0.865 | 1.016 | 1.061* | 0.762 | |
| Nuclear | 0.5 | 1.159* | 1.125 | 1.119 | 1.126* | 0.769 | 0.498 |
| | 1 | 1.153* | 1.116 | 1.136* | 1.105 | 0.756 | |
| | 1.5 | 1.141* | 1.142* | 1.036 | 1.123 | 0.783 | |

* The top two CVPL values ($\times 10^5$).

Table 5.4 shows the cross validated penalized likelihood obtained from the proposed approach. Among all the regularizations and $\lambda$ values, $k = 2$ stands out among all the possible clusters. These results inspire us to further study the consistency between cluster results and the "correctness" of this experiment. Table 5.4 also presents the adjusted random index in relation to the "correctness" labels. Compared to K means, our proposed approach is able to successfully identify the rat's "correctness" on identifying "in/out" sequences. It is worth

mentioning that in addition to 2 clusters, Table 5.4 also suggests 5 clusters. These results indicate our approach can possibly identify the five different odors. We will shed light on this direction in the next section.

### 5.6.2 Time Frequency Clustering Analysis

We will continue to uncover the latent structure carried from the LFP dataset. Allen et al. (2016) suggests two oscillatory bands (Theta: 4 - 12 Hertz and Slow Gamma: 20 - 40 Hertz) yield strong power and playing significant roles in detecting the "in/out" sequences. Figure 5.4 shows the time frequency plot on Theta and Slow Gamma bands. Although these two bands enjoy the most power, low frequency theta band apparently obtains much more than slow gamma bands. It has been shown that slow gamma bands were strongly modulated by the "in" and "out" pattern Allen et al. (2016). In this study, to take one step further, we applied the proposed method to the spectrum of Theta and Slow Gamma bands separately.

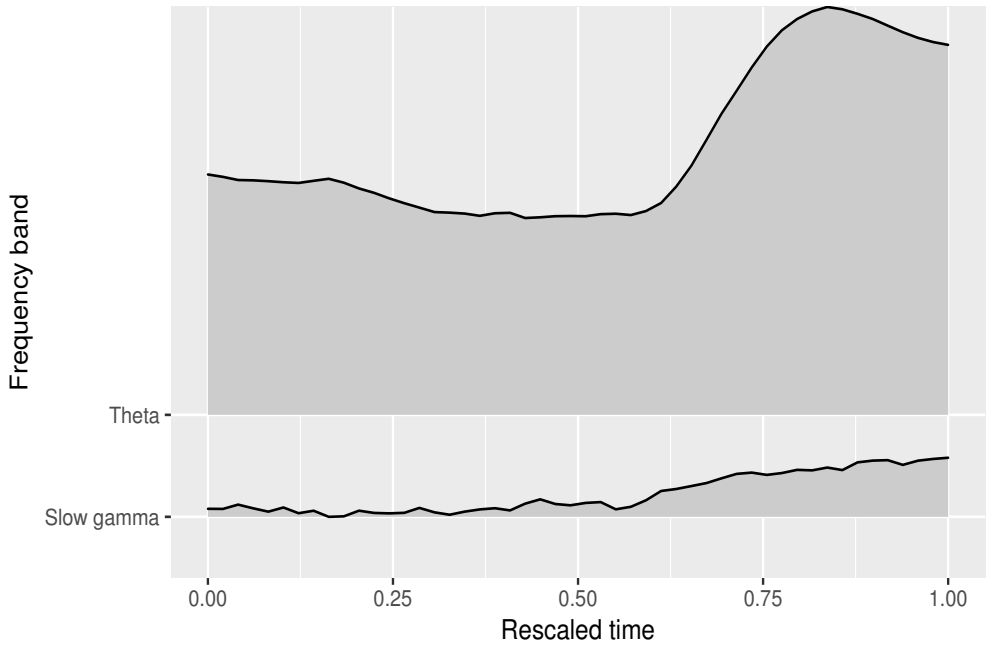



Figure 5.4: The time frequency plot of Theta and Slow Gamma bands over the "in-sequence" trials.

Table 5.5 presents the results after implementing the proposed method to the spectrum on Theta band. It can be easily found that for each regularization setting, 4 or 5 clusters are highly suggested. We further compare the 5 cluster results with the true odor sequence. As is shown in Table 5.5, the consistency is strong especially when comparing wit K means. Our approach provides some evidence indicating the association between the low frequency band (Theta) and the odor sequence.

Table 5.5: The cross validated penalized likelihood obtained from the "in-sequence" trials. The spectrum are from Theta band.

| Penalty | $\lambda$ | CVPL | | | | ARI | |
|---|---|---|---|---|---|---|---|
| | | $k=2$ | $k=3$ | $k=4$ | $k=5$ | our method | K means |
| L1 | 0 | 11.001 | 11.300* | 11.198* | 11.172 | 0.712 | 0.679 |
| | 0.5 | 8.516 | 8.975* | 8.849 | 8.997* | 0.692 | |
| | 1 | 8.650 | 8.632 | 8.725* | 8.745* | 0.703 | |
| | 1.5 | 8.571 | 8.705* | 8.556 | 8.701* | 0.709 | |
| L2 | 0.5 | 8.965* | 8.881* | 8.671 | 7.277 | 0.693 | 0.672 |
| | 1 | 8.719* | 8.388 | 8.544* | 7.616 | 0.686 | |
| | 1.5 | 8.650 | 8.632 | 8.825* | 8.745* | 0.682 | |
| Nuclear | 0.5 | 9.034 | 9.196* | 9.183 | 9.259* | 0.707 | 0.671 |
| | 1 | 9.013 | 9.166 | 9.255* | 9.263* | 0.714 | |
| | 1.5 | 8.571 | 9.040* | 8.995* | 8.969 | 0.712 | |

* The top two highest values ($\times 10^3$).

Further, we concentrate on the Slow Gamma band. Allen et al. (2016) has established the conclusion that slow gamma band strongly aligned with the "in/out" pattern. In this part, we applied the proposed method to all the "in-sequence" trials to uncover latent patterns. Table 5.6 summarizes the cross validated penalized likelihood values among different clusters. 2 clusters are being recommended in most of the cases. We later compare the cluster result with the "correctness" labels. In the case of nuclear norm regularization, the adjusted random index (0.5733) is almost 20% higher than K means (0.497).

Table 5.6: The cross validated penalized likelihood obtained from the "in-sequence" trials. The spectrum are from Slow Gamma band.

| Penalty | $\lambda$ | CVPL | | | |
|---|---|---|---|---|---|
| | | $k=2$ | $k=3$ | $k=4$ | $k=5$ |
| L1 | 0.5 | 8.470* | 8.395 | 8.064 | 7.993 |
| | 1 | 8.129* | 8.023 | 7.507 | 7.312 |
| | 1.5 | 7.689* | 7.641 | 7.215 | 6.765 |
| L2 | 0.5 | 8.360* | 7.933 | 7.980 | 7.660 |
| | 1 | 7.977* | 7.755 | 5.744 | 6.977 |
| | 1.5 | 7.696 | 7.754* | 6.584 | 6.502 |
| Nuclear | 0.5 | 8.687 | 8.785* | 8.531 | 8.373 |
| | 1 | 8.686* | 8.416 | 8.532 | 8.183 |
| | 1.5 | 8.534* | 8.438 | 8.324 | 7.981 |

* The highest values ($\times 10^3$).

## 5.7 Analysis of Rat Stroke Data

In this section, we apply the proposed approach to another LFPs dataset from a rat stroke experiment. In this study, LFPs were recorded before and after the stroke. 32 electrodes were implanted with 4 layers shown in Figure 5.5. Throughout this section, we work on the signals of 5 minutes before and after the stroke. The sampling rate is 1000 Hertz and each epoch is 1 second long. One of the scientific interests from this experiment is to identify the "latent" patterns that lead to before and after stroke.

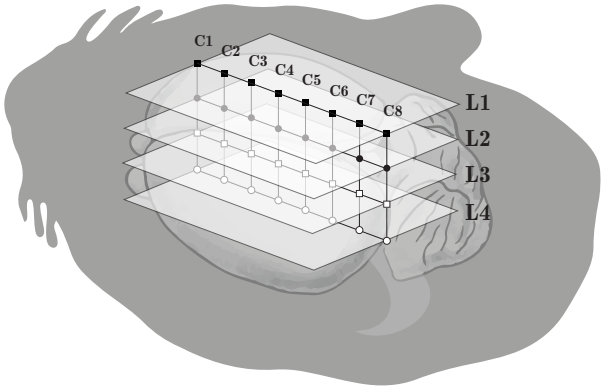



Figure 5.5: The schematic diagram of electrodes implanted in rat brain.

As preliminary analysis, we implemented time frequency analysis on this dataset. Figure 5.6 shows the log power spectra of two typical channels. These results were obtained by averaging all the trials before and after stroke separately. Most of the channels behaves "smoothly" within each epoch and there exists small discrepancy before and after stroke. However, just like the case of Channel 10, some channels presents innegligible dynamics and obvious difference between and after stroke. These findings shows that it is not optimal to average over or vectorize all the channels when we do cluster analysis to identify the "latent" pattern before and after stroke.

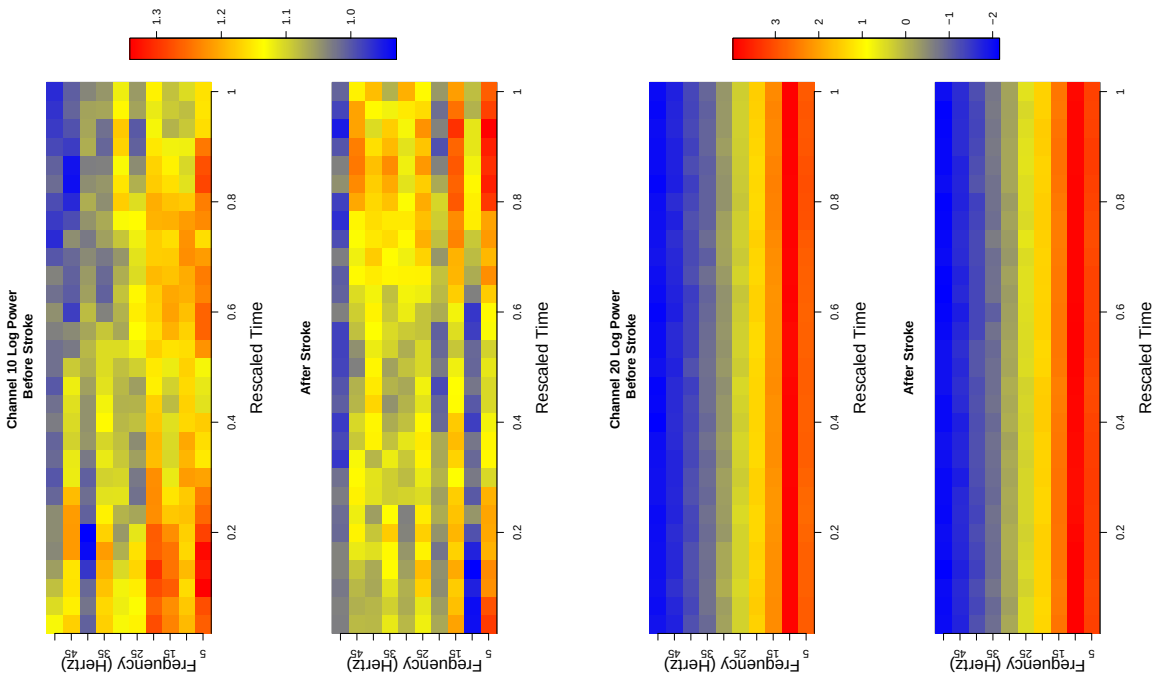



Figure 5.6: The time frequency plot of Channel 10 and 20 among all the 600 trials before and after the stroke.

To deepen the preliminary findings and motivate our proposed approach, we also study the dynamics across all the 32 channels before and after stroke. Figure 5.7 is the time frequency plot of Beta and Slow Gamma frequency bands across the channels. The log power spectra were obtained by averaging over the trials. Among the plots before and after stroke, we observe strong dependence across channels both for the two bands. This demonstrates the importance of introducing regularization terms into the mixture normal model. Comparing the plots before and after stroke, local discrepancy is easily identified. Such difference will be easily ignored if we just naively vectorize the original signals when doing cluster analysis.

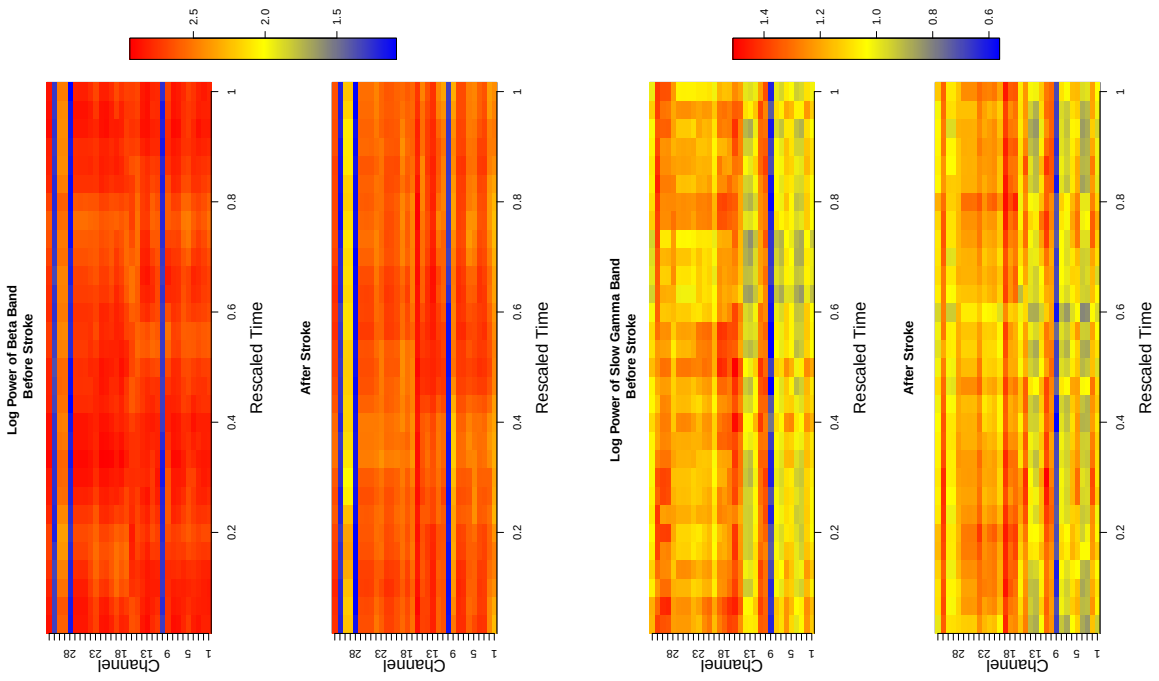



Figure 5.7: The time frequency plot of particular frequency bands among all the channels before and after stroke.

We applied the proposed approach to the time frequency images across all the trials before and after stroke. Table 5.7 shows the cross validated penalized likelihood values across different number of clusters and regularizations. With only one exception, all the scenarios suggest 2 clusters. As the next step, we compare the 2 cluster results with the index related to "stroke" or "normal". Table 5.8 summarizes the adjusted random index values (ARI). In comparing with K means results, the proposed approach outperforms in identifying "stroke" or "normal" sequences. Note that as by introducing regularizations, the proposed method is able to improve the results by 80%. In particular, Slow Gamma bands performs perfectly (ARI 1.000) when adding nuclear norm term with $\lambda = 2$. This result is almost double the case without penalty (ARI 0.507). Similar pattern can also be found in Beta band case. These findings are consistent with the conjecture in preliminary analysis.

Table 5.7: The cross validated penalized likelihood obtained from all the trials. The log power spectra are from Beta and Slow Gamma bands.

| Penalty | $\lambda$ | CVPL (Slow Gamma) | | | | CVPL (Beta) | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | $k=2$ | $k=3$ | $k=4$ | $k=5$ | $k=2$ | $k=3$ | $k=4$ | $k=5$ |
| L1 | 0 | 2.941 | 2.964* | 2.764 | 2.822 | 4.645* | 4.627 | 4.526 | 4.598 |
| | 1 | 2.472* | 2.031 | 1.513 | 0.4213 | 3.98* | 3.268 | 3.594 | 1.676 |
| | 2 | 2.106* | 1.370 | 1.288 | 0.621 | 4.167* | 3.373 | 3.227 | 3.277 |
| L2 | 0.5 | 2.688* | 2.474 | 2.306 | 2.184 | 4.245* | 4.179 | 4.036 | 3.424 |
| | 1 | 2.484* | 2.188 | 1.787 | 1.895 | 4.063* | 3.557 | 3.429 | 3.329 |
| | 2 | 2.338* | 2.163 | 1.539 | 1.733 | 4.024* | 3.699 | 2.972 | 3.206 |
| Nuclear | 0.5 | 2.806* | 2.627 | 2.502 | 2.303 | 4.464* | 4.299 | 4.130 | 3.963 |
| | 1 | 2.556* | 2.362 | 1.946 | 1.720 | 4.191* | 3.977 | 3.618 | 3.371 |
| | 2 | 2.748* | 1.689 | 1.257 | 0.684 | 3.687* | 3.274 | 2.795 | 2.262 |

* The highest values over different frequency bands ($\times 10^4$)

Table 5.8: The adjusted random index in relation to "Stroke". The spectrum are from Slow Gamma and Beta bands.

| Penalty | $\lambda$ | ARI (Slow Gamma) | | ARI (Beta) | |
|---|---|---|---|---|---|
| | | our method | kmeans | our method | kmeans |
| L1 | 0 | 0.507 | 0.751 | 0.887 | 0.716 |
| | 0.5 | 0.981 | | 0.942 | |
| | 1 | 0.961 | | 0.914 | |
| | 2 | 0.951 | | 0.861 | |
| L2 | 0.5 | 0.951 | 0.751 | 0.941 | 0.716 |
| | 1 | 0.951 | | 0.878 | |
| | 2 | 0.961 | | 0.787 | |
| Nuclear | 0.5 | 0.951 | 0.751 | 0.941 | 0.715 |
| | 1 | 0.960 | | 0.942 | |
| | 2 | 1.000 | | 0.951 | |

# Chapter 6

# Conclusions and Future Directions

In this dissertation, we discuss on some interesting problems in categorical and multivariate time series with applications to electrophysiological signals.

- In Chapter 2, we demonstrate that applying the Em-FI matrix to highly correlated data may lead to undesirable consequences in inference. Such consequences include longer average confidence interval widths and potentially misleading inferential results. To overcome these limitation, we derive the exact form and an iterative computation formula of the conditional Fisher information matrix for the general logistic autoregressive model with (without) endogenous covariates (LAR($p$)/LARX($p$)). Simulation studies based on the LAR($p$)/LARX($p$) model demonstrate the advantages of Ex-FI over Em-FI in terms of small sample stability, leading to narrower confidence intervals on average while maintaining false positive rates at or below nominal levels. Numerically, we establish the convergence of the exact conditional Fisher information and studied the asymptotic behavior as $T$ grows large. Consequently, analysis of the respiratory binary time series data suggests that using Ex-FI may result in greater statistical

power when making inference. In summary, the Ex-FI matrix is recommended over the Em-FI as it provides greater stability for small time series and equivalent large sample inference. While the derivation of the Ex-FI is non-trivial it is computationally tractable because it can be obtained iteratively. The result is a stable estimator that is easily implementable and more stable, particularly for sample sizes less than 200. As future work, it is of great interest to achieve theoretic results on quantifying the loss of using the Em-FI.

- In Chapter 3, the proposed hybrid inference method for binary time series (HIBITS) produces efficient inference and promising predictions with a relatively low computational cost. Compared to existing methods, our proposed approach has the following advantages: on one hand, by involving known covariates as fixed effect components, we make use of the information indicating the association between the response and covariates. On the second stage, a Gaussian process captures the information beyond what provided by those covariates of both endogenous and exogenous time series. On the other hand, as indicated in the simulation, the proposed method is robust compared to existing methods. The proposed model selection strategy allows the model to fit the data even though not enough information is captured by the fixed effect components. The strategies in providing point and interval estimates, in addition, allows researchers to gain more informative conclusions in the association between response and covariates. These advantages make our model easy to interpret. In summary, the proposed HIBITS method, serving as an approach with high prediction power, efficient inference capability and direct interpretability, provides a promising methodology in modeling and predicting sleep states and other binary time series.

  As a future direction, we could extend the proposed HIBITS to model general categorical responses (e.g. the 4 sleep stages). In the case of nominal categorical outcomes, we

could follow the similar framework of multinomial logit model discussed by Fokianos and Kedem (2002). Specifically, the link function could be extended to softmax function where fixed and random effects can be imposed on the systematic component, which is a natural generalization of our proposed HIBITS. On the other hand, if the outcomes are ordinal categorical time series, one can impose a threshold mechanism on the systematic component of the model. Following the scenario of proportional odds models in Fokianos and Kedem (2003b), the HIBITS method can be extended by incorporating fixed and random effects.

- In Chapter 4, we propose an evolutionary state space model (E-SSM) that allows the latent source signals to evolve across epochs. Although the results reported in this chapter are quite promising, nevertheless, modeling the evolution/dynamics across epochs still remains a challenge in general. For example, we ignored the subject specific random effects in the current chapter, which should be taken into account in a future work. Other future directions include incorporating different experimental conditions to improve the efficiency of statistical inference.

- In Chapter 5, we develop a mixture matrix normal model with various regularizations on the mean spatial-temporal structure. In theory, we show the consistency of the constraint penalized maximum likelihood estimator. Simulation studies prove that the proposed approach outperforms over the existing methods in uncovering the "latent" clusters. In the LFPs dataset study, the model produce stable and interpretable results in understanding the sparse spatial-temporal structure among different trials (epochs). This also demonstrates the LFPs are evolving smoothly during the experiment. As a further step, it would be promising to generalize the proposed model to characterize tensors with more than 3 dimensions. Another direction could be to incorporate the subject-specific effects into the framework.

# Bibliography


Allen, T. A., D. M. Salz, S. McKenzie, and N. J. Fortin (2016). Nonspatial sequence coding in ca1 neurons. *Journal of Neuroscience 36*(5), 1547–1563.

Baccalá, L. A. and K. Sameshima (2001). Partial directed coherence: a new concept in neural structure determination. *Biological cybernetics 84*(6), 463–474.

Banerjee, S., B. P. Carlin, and A. E. Gelfand (2014). *Hierarchical modeling and analysis for spatial data*. Crc Press.

Banerjee, S., A. E. Gelfand, A. O. Finley, and H. Sang (2008). Gaussian predictive process models for large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 70*(4), 825–848.

Barrett, K. E. et al. (2010). Ganong's review of medical physiology.

Benbadis, S. R. (2006). Introduction to sleep electroencephalography. *Sleep: A Comprehensive Handbook. USA: John Wiley & Sons*, 989–1024.

Billingsley, P. (1961). Statistical inference for markov processes.

Bonney, G. E. (1987a). Logistic regression for dependent binary observations. *Biometrics*, 951–973.

Bonney, G. E. (1987b). Logistic regression for dependent binary observations. *Biometrics*, 951–973.

Brillinger, D. (1964). A frequency approach to the techniques of principal components, factor analysis and canonical variates in the case of stationary time series. In *Invited Paper, Royal Statistical Society Conference, Cardiff Wales.(Available at http://stat-www.berkeley.edu/users/brill/papers.html)*.

Brillinger, D. R. (1975). The identification of point process systems. *The Annals of Probability*, 909–924.

Brillinger, D. R. (1983). A generalized linear model with gaussian regressor variables. In *A Festschrift for Erich L. Lehmann*, pp. 97 – 114. Pacific Grove, CA: Wadsworth.

Brockwell, P. and R. Davis (1991). Time series: data analysis and theory. *Springer, New York*.

Caiado, J., N. Crato, and D. Peña (2006). A periodogram-based metric for time series classification. *Computational Statistics & Data Analysis 50*(10), 2668–2684.

Cornford, D. (1998). Non-zero mean gaussian process prior wind field models.

Cressie, N. (2015). *Statistics for spatial data*. John Wiley & Sons.

Davis, R. A., W. T. Dunsmuir, and Y. Wang (2000). On autocorrelation in a poisson regression model. *Biometrika 87*(3), 491–505.

Dawid, A. P. (1981). Some matrix-variate distribution theory: notational considerations and a bayesian application. *Biometrika 68*(1), 265–274.

de Vries, S. O., V. Fidler, W. D. Kuipers, and M. G. Hunink (1998). Fitting multistate transition models with autoregressive logistic regression: supervised exercise in intermittent claudication. *Medical decision making 18*(1), 52–60.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, 1–38.

Derado, G., F. D. Bowman, and C. D. Kilts (2010). Modeling the spatial and temporal dependence in fmri data. *Biometrics 66*(3), 949–957.

Deuschl, G., A. Eisen, et al. (1999). *Recommendations for the practice of clinical neurophysiology: guidelines of the International Federation of Clinical Neurophysiology*. Elsevier.

Diggle, P. (2002). *Analysis of longitudinal data*. Oxford University Press.

Dodge, Y. (2006). *The Oxford dictionary of statistical terms*. Oxford University Press on Demand.

Dutilleul, P. (1999). The mle algorithm for the matrix normal distribution. *Journal of statistical computation and simulation 64*(2), 105–123.

Einevoll, G. T., K. H. Pettersen, A. Devor, I. Ulbert, E. Halgren, and A. M. Dale (2007). Laminar population analysis: estimating firing rates and evoked synaptic activity from multielectrode recordings in rat barrel cortex. *Journal of neurophysiology 97*(3), 2174–2190.

Entringer, S., E. S. Epel, J. Lin, E. H. Blackburn, C. Buss, B. Shahbaba, D. L. Gillen, R. Venkataramanan, H. N. Simhan, and P. D. Wadhwa (2015). Maternal folate concentration in early pregnancy and newborn telomere length. *Annals of Nutrition and Metabolism 66*(4), 202–208.

Fahrmeir, L. and H. Kaufmann (1987). Regression models for non-stationary categorical time series. *Journal of time series Analysis 8*(2), 147–160.

Fahrmeir, L. and G. Tutz (2013). *Multivariate statistical modelling based on generalized linear models*. Springer Science & Business Media.

Fan, J. and R. Li (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association 96*(456), 1348–1360.

Fiecas, M. and H. Ombao (2016). Modeling the evolution of dynamic brain processes during an associative learning experiment. *Journal of the American Statistical Association 111*(516), 1440–1453.

Fokianos, K. and B. Kedem (1998a). Prediction and classification of non-stationary categorical time series. *Journal of multivariate analysis 67*(2), 277–296.

Fokianos, K. and B. Kedem (1998b). Prediction and classification of non-stationary categorical time series. *Journal of multivariate analysis 67*(2), 277–296.

Fokianos, K. and B. Kedem (2002). Regression model for time series analysis.

Fokianos, K. and B. Kedem (2003a). Regression theory for categorical time series. *Statistical science*, 357–376.

Fokianos, K. and B. Kedem (2003b). Regression theory for categorical time series. *Statistical science*, 357–376.

Friedman, J., T. Hastie, and R. Tibshirani (2001). *The elements of statistical learning*, Volume 1. Springer series in statistics Springer, Berlin.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189–1232.

Fuller, W. A. (2009). *Introduction to statistical time series*, Volume 428. John Wiley & Sons.

Gao, X., H. Ombao, and D. Gillen (2017). Fisher information matrix of binary time series. *arXiv preprint arXiv:1711.05483*.

Gao, X., B. Shahbaba, N. Fortin, and H. Ombao (2016). Evolutionary state-space model and its application to time-frequency analysis of local field potentials. *arXiv preprint arXiv:1610.07271*.

Gao, X., B. Shahbaba, and H. Ombao (2017). Modeling binary time series using gaussian processes with application to predicting sleep states. *arXiv preprint arXiv:1711.05466*.

Gelfand, A. E., A. Kottas, and S. N. MacEachern (2005). Bayesian nonparametric spatial modeling with dirichlet process mixing. *Journal of the American Statistical Association 100*(471), 1021–1035.

Green, P. J. (1990). On use of the em for penalized likelihood estimation. *Journal of the Royal Statistical Society. Series B (Methodological)*, 443–452.

Guo, Y. (2011). A general probabilistic model for group independent component analysis and its estimation methods. *Biometrics 67*(4), 1532–1542.

Hamilton, J. D. (1994). *Time series analysis*, Volume 2. Princeton university press Princeton.

Hathaway, R. J. (1985). A constrained formulation of maximum-likelihood estimation for normal mixture distributions. *The Annals of Statistics*, 795–800.

Holmes, T. H. and R. H. Rahe (1967). The social readjustment rating scale. *Journal of psychosomatic research 11*(2), 213–218.

Jacobs, P. A. and P. A. Lewis (1978). Discrete time series generated by mixtures ii: asymptotic properties. *Journal of the Royal Statistical Society. Series B (Methodological)*, 222–228.

Jiru, A. R. (2008). *Relationships between spectral peak frequencies of a causal AR (P) process and arguments of roots of the associated ar polynomial*. Ph. D. thesis, San Jose State University.

Kaufmann, H. (1987). Regression models for nonstationary categorical time series: asymptotic estimation theory. *The Annals of Statistics*, 79–98.

Kedem, B. and K. Fokianos (2005). *Regression models for time series analysis*, Volume 488. John Wiley & Sons.

Kedem, B. and E. Slud (1980). Binary time series. Technical report, MARYLAND UNIV COLLEGE PARK DEPT OF MATHEMATICS.

Kedem, B. and S. Yakowitz (1994). *Time series analysis by higher order crossings*. IEEE press New York.

Keenan, D. M. (1982a). A time series analysis of binary data. *Journal of the American Statistical Association 77*(380), 816–821.

Keenan, D. M. (1982b). A time series analysis of binary data. *Journal of the American Statistical Association 77*(380), 816–821.

Kuss, M. (2006). *Gaussian process models for robust regression, classification, and reinforcement learning*. Ph. D. thesis, TU Darmstadt.

Lange, K. (1995). A gradient algorithm locally equivalent to the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 425–437.

Lange, K., D. R. Hunter, and I. Yang (2000). Optimization transfer using surrogate objective functions. *Journal of computational and graphical statistics 9*(1), 1–20.

Lin, X. and D. Zhang (1999). Inference in generalized additive mixed modelsby using smoothing splines. *Journal of the royal statistical society: Series b (statistical methodology) 61*(2), 381–400.

Lindquist, M. A. and I. McKeague (2009). Logistic regression with brownian-like predictors. *Journal of the American Statistical Association 104*, 1575–1585.

Lütkepohl, H. (2005). *New introduction to multiple time series analysis*. Springer Science & Business Media.

Maharaj, E. A. (2002). Comparison of non-stationary time series in the frequency domain. *Computational Statistics & Data Analysis 40*(1), 131–141.

Maharaj, E. A., P. DUrso, and D. U. Galagedera (2010). Wavelet-based fuzzy clustering of time series. *Journal of classification 27*(2), 231–275.

Makarova, J., J. M. Ibarz, V. A. Makarov, N. Benito, and O. Herreras (2011). Parallel readout of pathway-specific inputs to laminated brain structures. *Frontiers in systems neuroscience 5*, 77.

Makarova, J., T. Ortuño, A. Korovaichuk, J. Cudeiro, V. A. Makarov, C. Rivadulla, and O. Herreras (2014). Can pathway-specific lfps be obtained in cytoarchitectonically complex structures? *Frontiers in systems neuroscience 8*, 66.

McCullagh, P. (1984). Generalized linear models. *European Journal of Operational Research 16*(3), 285–292.

Meyn, S. P. and R. L. Tweedie (2012a). *Markov chains and stochastic stability*. Springer Science & Business Media.

Meyn, S. P. and R. L. Tweedie (2012b). *Markov chains and stochastic stability*. Springer Science & Business Media.

Michel, C., D. Lehmann, B. Henggeler, and D. Brandeis (1992). Localization of the sources of eeg delta, theta, alpha and beta frequency bands using the fft dipole approximation. *Electroencephalography and clinical neurophysiology 82*(1), 38–44.

Milligan, G. W. and M. C. Cooper (1986). A study of the comparability of external criteria for hierarchical cluster analysis. *Multivariate Behavioral Research 21*(4), 441–458.

Minka, T. P. (2001). *A family of algorithms for approximate Bayesian inference*. Ph. D. thesis, Massachusetts Institute of Technology.

Mitzdorf, U. et al. (1985). Current source-density method and application in cat cerebral cortex: investigation of evoked potentials and eeg phenomena.

Muenz, L. R. and L. V. Rubinstein (1985). Markov models for covariate dependence of binary sequences. *Biometrics*, 91–101.

Nevsimalova, S. and K. Sonka (1997). Poruchy spanku a bdeni. *Maxdorf/Jessenius, Parha*.

Opper, M. and O. Winther (1999). Gaussian processes for classification: Mean field algorithms, submitted to neural computation.

Pan, W. and X. Shen (2007). Penalized model-based clustering with application to variable selection. *Journal of Machine Learning Research 8*(May), 1145–1164.

Prado, R. and H. F. Lopes (2013). Sequential parameter learning and filtering in structured autoregressive state-space models. *Statistics and Computing*, 1–15.

Quick, H., S. Banerjee, B. P. Carlin, et al. (2013). Modeling temporal gradients in regionally aggregated california asthma hospitalization data. *The Annals of Applied Statistics 7*(1), 154–176.

Reinsel, G. (1982). Multivariate repeated-measurement or growth curve models with multivariate random-effects covariance structure. *Journal of the American Statistical Association 77*(377), 190–195.

Schacke, K. (2004). On the kronecker product. *Master's thesis, University of Waterloo*.

Shumway, R. H. and D. S. Stoffer (2013). *Time series analysis and its applications*. Springer Science & Business Media.

Smyth, P. (2000). Model selection for probabilistic clustering using cross-validated likelihood. *Statistics and computing 10*(1), 63–72.

Snelson, E., C. E. Rasmussen, and Z. Ghahramani (2004). Warped gaussian processes. *Advances in neural information processing systems 16*, 337–344.

Stein, M. L. (2012). *Interpolation of spatial data: some theory for kriging*. Springer Science & Business Media.

Vandenberg-Rodes, A. and B. Shahbaba (2015). Dependent mat\'ern processes for multivariate time series. *arXiv preprint arXiv:1502.03466*.

Wang, F. and A. E. Gelfand (2014). Modeling space and space-time directional data using projected gaussian processes. *Journal of the American Statistical Association 109*(508), 1565–1580.

Wang, Y., C.-M. Ting, and H. Ombao (2016). Exploratory analysis of high dimensional time series with applications to multichannel electroencephalograms. *arXiv preprint arXiv:1610.07684*.

Whitmore, N. W. and S.-C. Lin (2016). Unmasking local activity within local field potentials (lfps) by removing distal electrical signals using independent component analysis. *NeuroImage 132*, 79–92.

Williams, C. K. and D. Barber (1998). Bayesian classification with gaussian processes. *Pattern Analysis and Machine Intelligence, IEEE Transactions on 20*(12), 1342–1351.

Williams, C. K. and C. E. Rasmussen (2006). Gaussian processes for machine learning. *the MIT Press 2*(3), 4.

Yakowitz, S. J. and J. D. Spragins (1968). On the identifiability of finite mixtures. *The Annals of Mathematical Statistics*, 209–214.

Yu, Y. (2012). Monotonically overrelaxed em algorithms. *Journal of Computational and Graphical Statistics 21*(2), 518–537.

Zhang, K. and A. Hyvärinen (2011). A general linear non-gaussian state-space model: Identifiability, identification, and applications. In *JMLR Workshop and Conference Proc., Asian Conf. on Machine Learning*, pp. 113–128.

Zhou, B., D. E. Moorman, S. Behseta, H. Ombao, and B. Shahbaba (2015). A dynamic bayesian model for characterizing cross-neuronal interactions during decision making. *Journal of the American Statistical Association* (just-accepted), 1–44.

# Appendix A

# Some Theoretic Results and Supplementary Figures of Chapter 4

We will briefly prove the main theorem (AR(2) spectral decomposition theorem) and provide some figures regarding simulation results and LFP analysis.

## A.1 Proof of AR(2) Spectral Decomposition Theorem

We first present a lemma that gives us an explicit form of the autocovariance function of an AR(2) process. Such results will be helpful for proving the main theorem.

**Lemma 1.** *Given a (weakly) stationary zero mean AR(2) process $S_t$, the autocovariance function $\gamma_S(h)$ takes the form*

$$\gamma_S(h) = A_1(\rho e^{\psi i})^{-h} + A_2(\rho e^{-\psi i})^{-h}, \tag{A.1}$$

*where* $A_1, A_2$ *can be determined by solving the linear equation* $A_1+A_2 = \frac{(1-\varphi_2)\sigma_w^2}{(1+\varphi_2)(1-\varphi_1-\varphi_2)(1+\varphi_1-\varphi_2)}$ *and* $A_1(\rho e^{\psi i})^{-1} + A_2(\rho e^{-\psi i})^{-1} = \frac{\varphi_1\sigma_w^2}{(1+\varphi_2)(1-\varphi_1-\varphi_2)(1+\varphi_1-\varphi_2)}$.

The proof is due to the fact that $\gamma_S(h) = \varphi_1\gamma_S(h-1) + \varphi_2\gamma_S(h-2)$.

To prove Theorem 1, we first show that for any fixed $M$, $\{f_{S^{(j)}}(\omega)\}_{j=1}^M$ are linearly independent.

In fact, suppose there exists some constants $b_1, \cdots, b_M$ such that $\sum_{j=1}^M b_j f_{S^{(j)}}(\omega) = 0$, then we must have $\sum_{j=1}^M b_j \sum_{h=-\infty}^{\infty} \gamma_{S^{(j)}}(h)e^{2\pi i\omega h} = \sum_{h=-\infty}^{\infty}\sum_{j=1}^M b_j\gamma_{S^{(j)}}(h)e^{2\pi i\omega h} = 0$. As a direct result from Fourier theorem, we have $\sum_{j=1}^M b_j\gamma_{S^{(j)}}(h) = 0$ for any $h$. Thus for any positive integer $H$, $b_1, \cdots, b_M$ are solutions of the linear equation

$$\Gamma\mathbf{b} = 0, \tag{A.2}$$

where $\Gamma = \begin{bmatrix} \gamma_{S^{(1)}}(0) & \gamma_{S^{(2)}}(0) & \dots & \gamma_{S^{(M)}}(0) \\ \gamma_{S^{(1)}}(1) & \gamma_{S^{(2)}}(1) & \dots & \gamma_{S^{(M)}}(1) \\ \vdots & \vdots & \vdots & \vdots \\ \gamma_{S^{(1)}}(H) & \gamma_{S^{(2)}}(H) & \dots & \gamma_{S^{(M)}}(H) \end{bmatrix}_{(H+1)*M}$ and $\mathbf{b} = (b_1, \cdots, b_M)'$. From Lemma 1, it is easy to show that $\gamma_{S^{(j)}}(h) = (\rho^{(j)})^{-h}(A_1^{(j)} + A_2^{(j)})cos(h\psi^{(j)})$. Note that due to the condition that $\max_M\{|\omega_1 - \omega_0|, \cdots, |\omega_M - \omega_{M-1}|\} \to 0$ and $A_1^{(j)}, A_2^{(j)}$ are nonlinear functions of $j$, we have $rank(\Gamma) = min\{H+1, M\}$. It implies $\mathbf{b} = \mathbf{0}$ and $\{f_{S^{(j)}}(\omega)\}_{j=1}^M$ are linearly independent. Then we can implement the Gram-Schmidt process on the family of functions $\{f_{S^{(j)}}(\omega)\}_{j=1}^{\infty}$ to obtain a family of orthonormal functions $\{\tilde{f}_{S^{(j)}}(\omega)\}_{j=1}^{\infty}$ in $L^2(0, \frac{1}{2})$. It follows that for any nonnegative coefficients $a_1, \cdots, a_M$, there exist $\tilde{a}_1, \cdots, \tilde{a}_M$ such that $||f_Y(\omega) - \sum_{j=1}^M a_j^2 f_{S^{(j)}}(\omega)||_2 = ||f_Y(\omega) - \sum_{i=1}^M \tilde{a}_j^2\tilde{f}_{S^{(j)}}(\omega)||_2$. If we can show $\{\tilde{f}_{S^{(j)}}(\omega)\}_{j=1}^{\infty}$ is also complete in $L^2(0, \frac{1}{2})$, by Parseval equality, we can obtain that $||f_Y(\omega) - \sum_{j=1}^M \tilde{a}_j^2\tilde{f}_{S^{(j)}}(\omega)||_2 \to 0$ as $M \to \infty$ and equivalently, $||f_Y(\omega) - f_{\hat{Q}_{t,M}}(\omega)||_2 \to 0$ as $M \to \infty$.

To show that $\{\tilde{f}_{S^{(j)}}(\omega)\}_{j=1}^{\infty}$ is complete in $L^2(0, \frac{1}{2})$, it suffices to show $\{f_{S^{(j)}}(\omega)\}_{j=1}^{\infty}$ is complete. Let us define $\mathbb{B} = \{f_{S^{(j)}}(\omega)\}_{j=1}^{\infty}$. For any function $g(\omega)$ in $L^2(0, \frac{1}{2})$, if $g(\omega) \perp \mathbb{B}$, then we have $\int_0^{\frac{1}{2}} g(\omega) f_{S^{(j)}}(\omega) d\omega = 0$ for any $j$. It is equivalent to $\sum_{h=\infty}^{\infty} \int_0^{\frac{1}{2}} g(\omega) \gamma_{S^{(j)}}(h) e^{2\pi i \omega h} d\omega = \sum_{h=-\infty}^{\infty} \gamma_g(h) \gamma_{S^{(j)}}(h) = 0$ for any $j$. It boils down to the problem of solving for the linear equation $\Gamma' \gamma = 0$, where $\Gamma$ is defined in Equation (A.2) and $\gamma = (\gamma_g(0), \cdots, \gamma_g(H))$ for any $M$ and $H$. We have proved that $\Gamma$ is of full row rank and thus $\gamma_g(h) = 0$ for any $h$. Thus $\{f_{S^{(j)}}(\omega)\}_{j=1}^{\infty}$ is complete in $L^2(0, \frac{1}{2})$. □

## A.2 Figures of Chapter 4

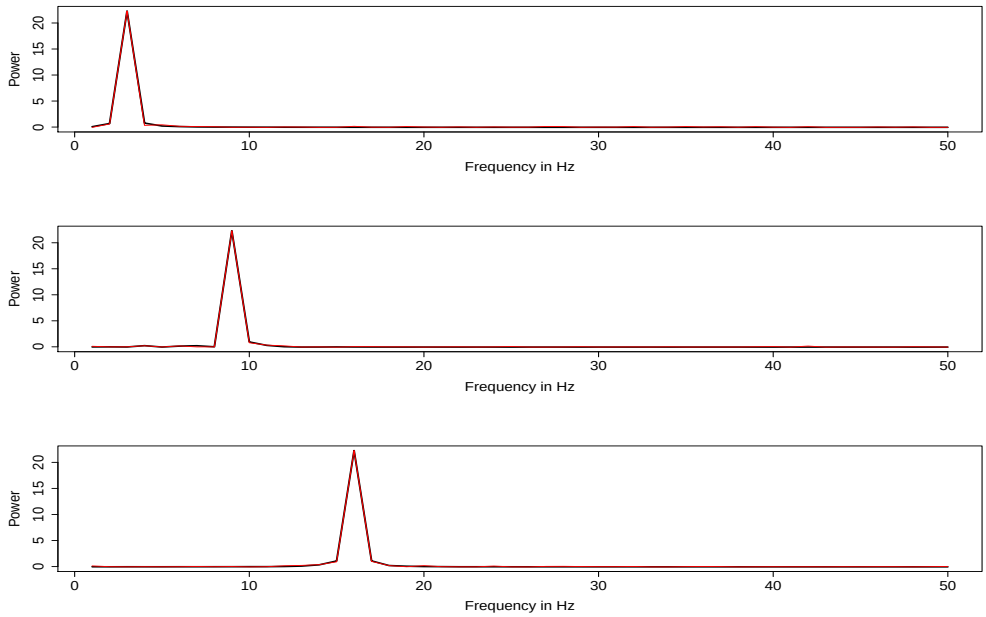



Figure A.1: The periodograms of the true (black) and estimated (red) latent processes.

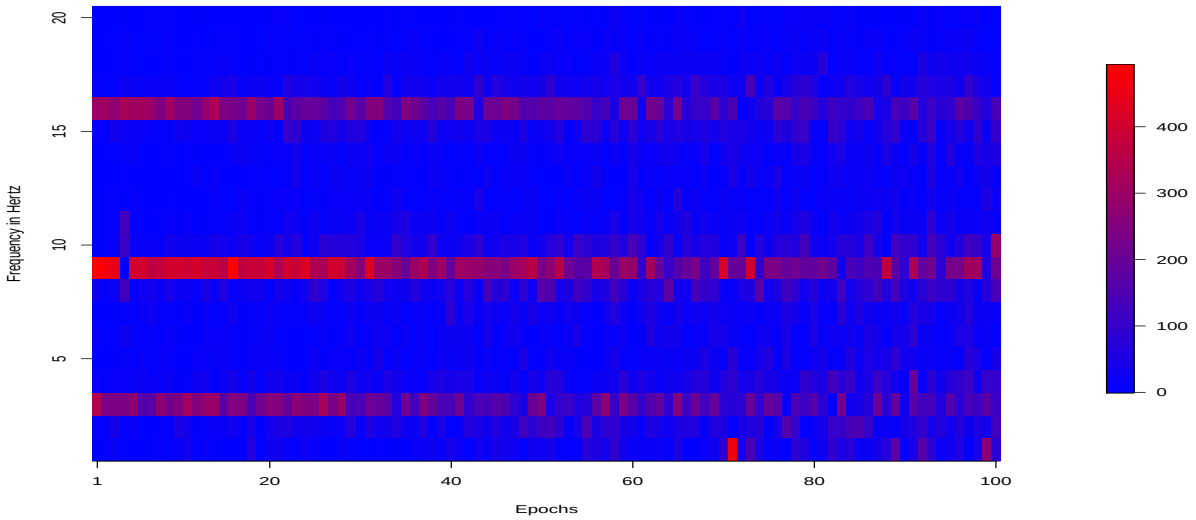



Figure A.2: The periodogram of generated signals from electrode 1 computed over all 100 epochs. From the heat map, we are observing the powers are evolving across epochs. At early stage, three dominating frequency bands can be identified clearly. As epoch evolves, such pattern is getting less clear.

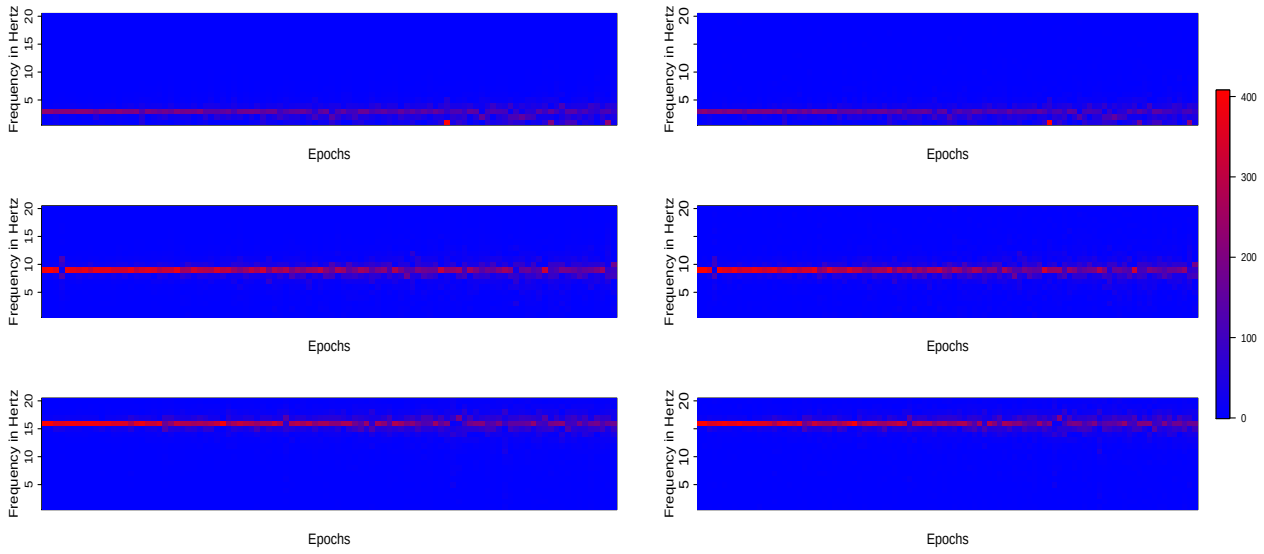



Figure A.3: The periodograms of the true (left) and estimated (right) latent AR(2) processes corresponding to delta (top), alpha (middle) and beta (bottom) frequency band.

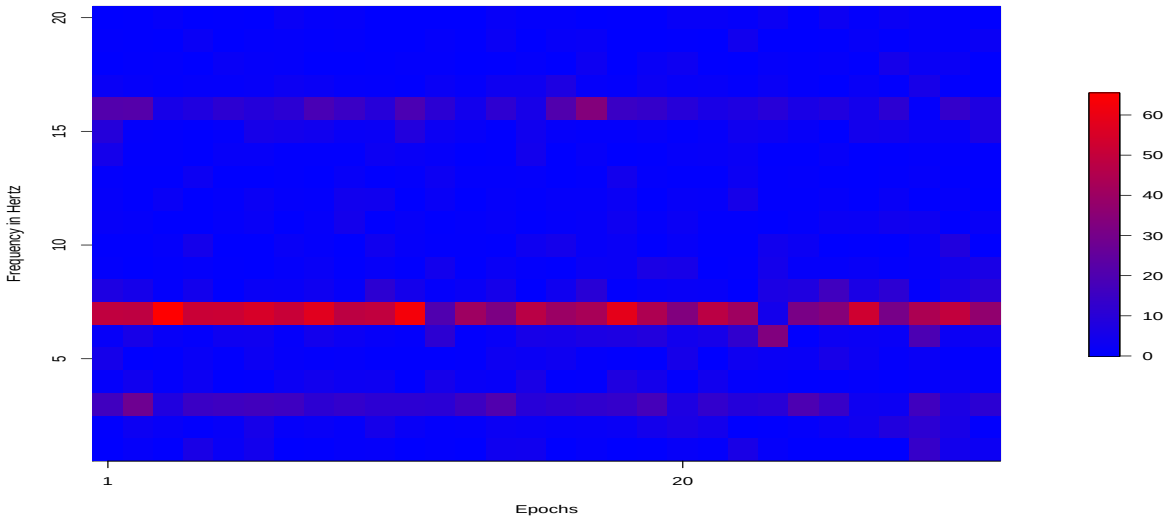



Figure A.4: The periodogram of generated signals from electrode 1 computed over all 30 epochs.

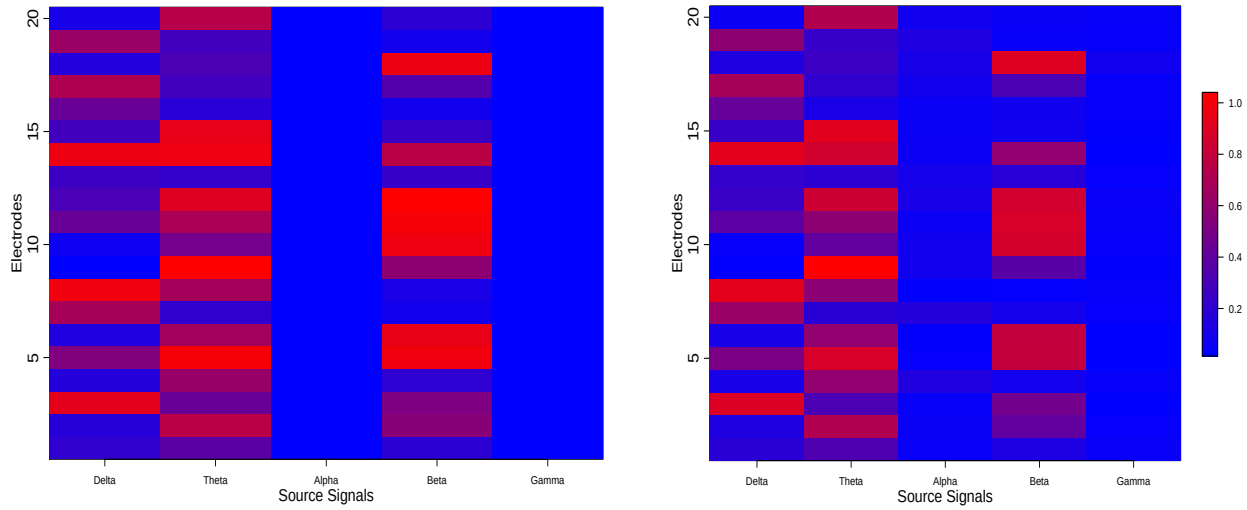



Figure A.5: The true mixing matrix (left) and estimated mixing matrix (right). Darker color indicates heavier weight given by the corresponding latent processes. Columns corresponding to "alpha" and "gamma" bands are zero in the true mixing matrix (left). In the estimated mixing matrix (right), those two columns are also close to zero.

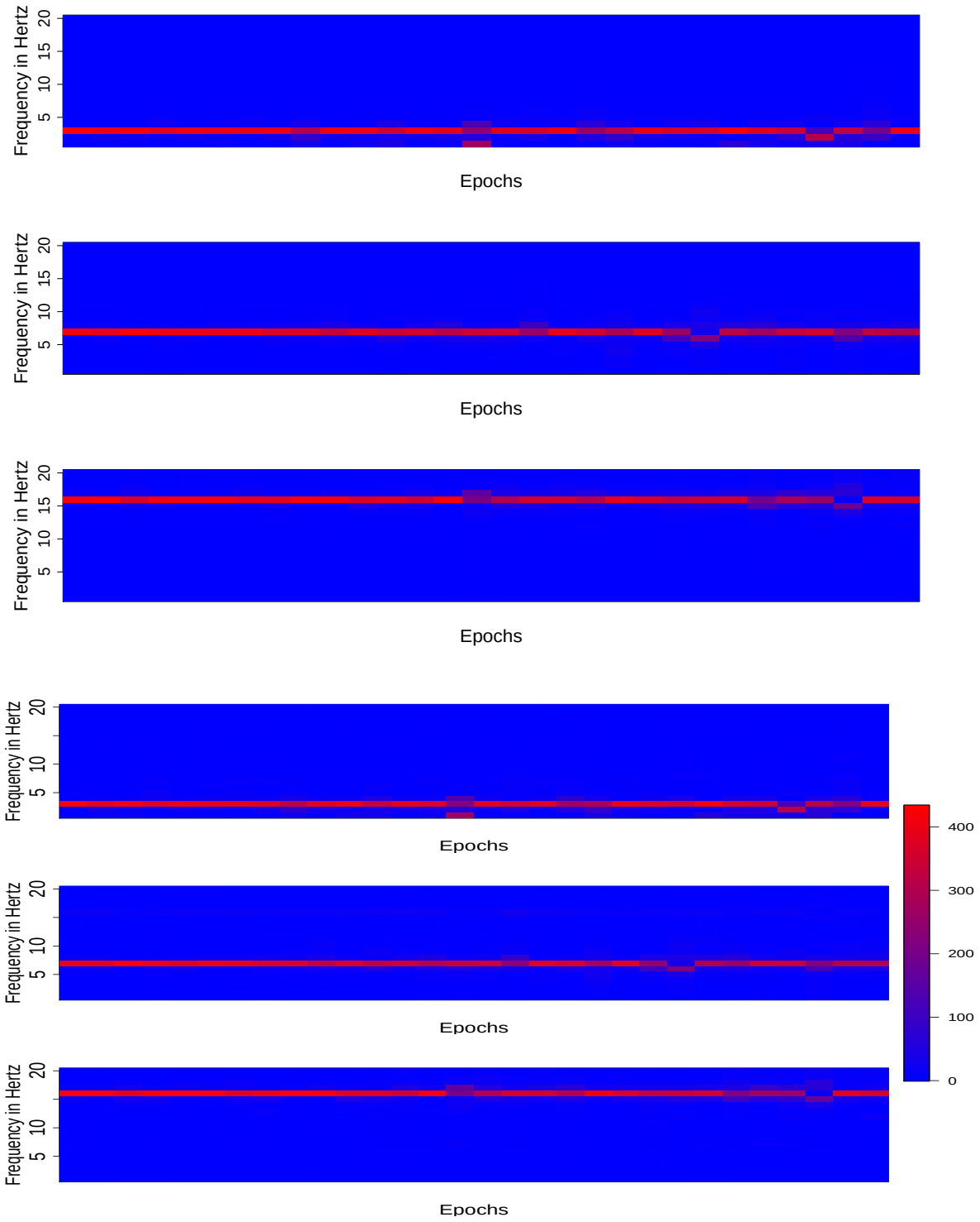



Figure A.6: The periodograms of the true (left) and estimated (right) latent AR(2) processes corresponding to delta (top), theta (middle) and beta (bottom) frequency band.

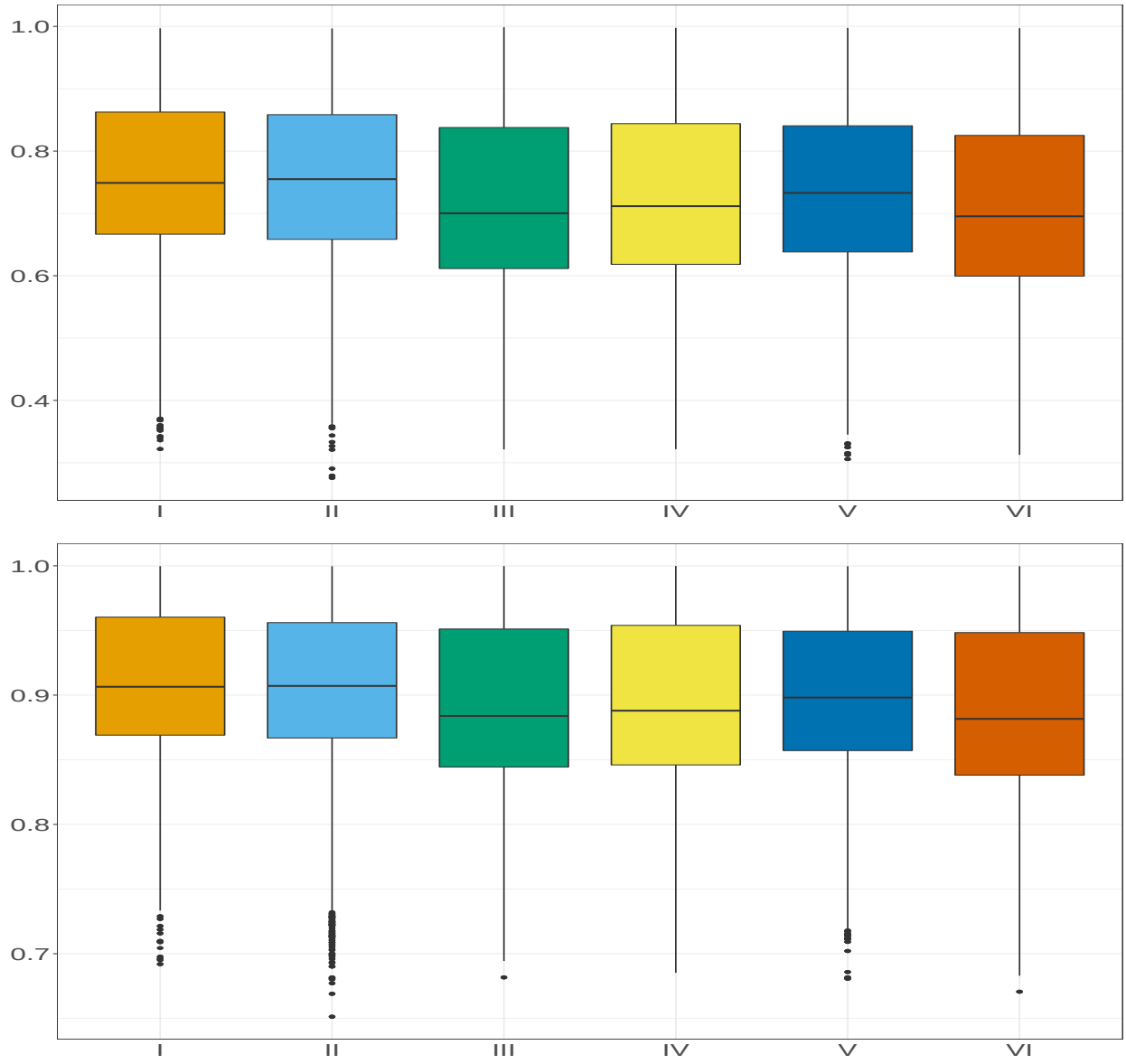



Figure A.7: The boxplots of variance accounted by different components across different stages during the experiment. The results were obtained by conducting principal component analysis on frequency domain. Epochs in the entire experiment have been classified as 6 stages with each consisting of 40 epochs (Stage I: 1-40, II: 41-80, III: 81-120, IV: 121-160, V: 161-200, VI: 201-247). The first component is shown on top and the first three cumulative components is at the bottom. We could observe that about 90% of variance can be explained by three components.

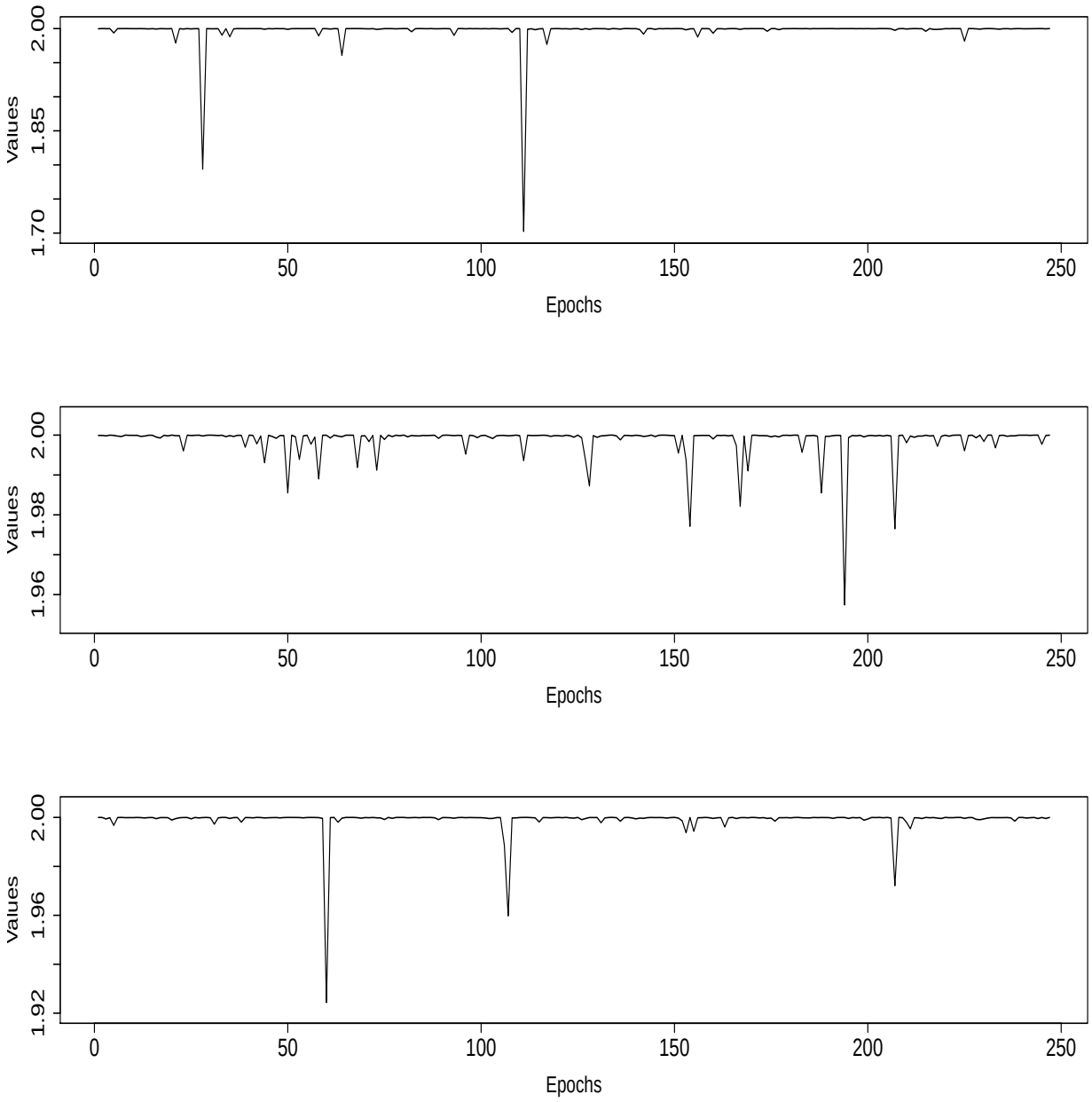



Figure A.8: The time series plots of modulus corresponding to delta (above), alpha (middle) and gamma (bottom) frequency bands.

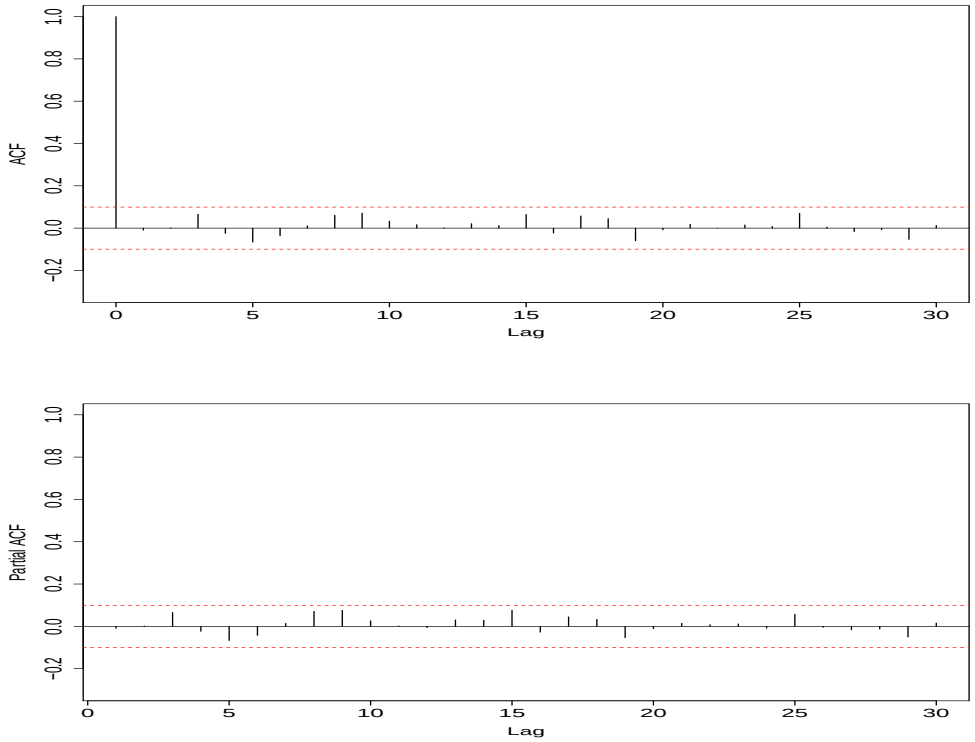



Figure A.9: Top: Auto-correlation function (ACF) of the residual plots from electrode 1. Bottom: Partial auto-correlation function (PACF) of the residual plots from electrode 1. The dashed lines indicate the threshold for non-zero correlation. These plots, along with the Ljung-Box test for white noise ($p - value \approx 0.75$) suggest that the residuals are white noise and hence the E-SSM model fits the data well. These same plots were observed in all the other electrodes but we do not report them here due to space constraints.

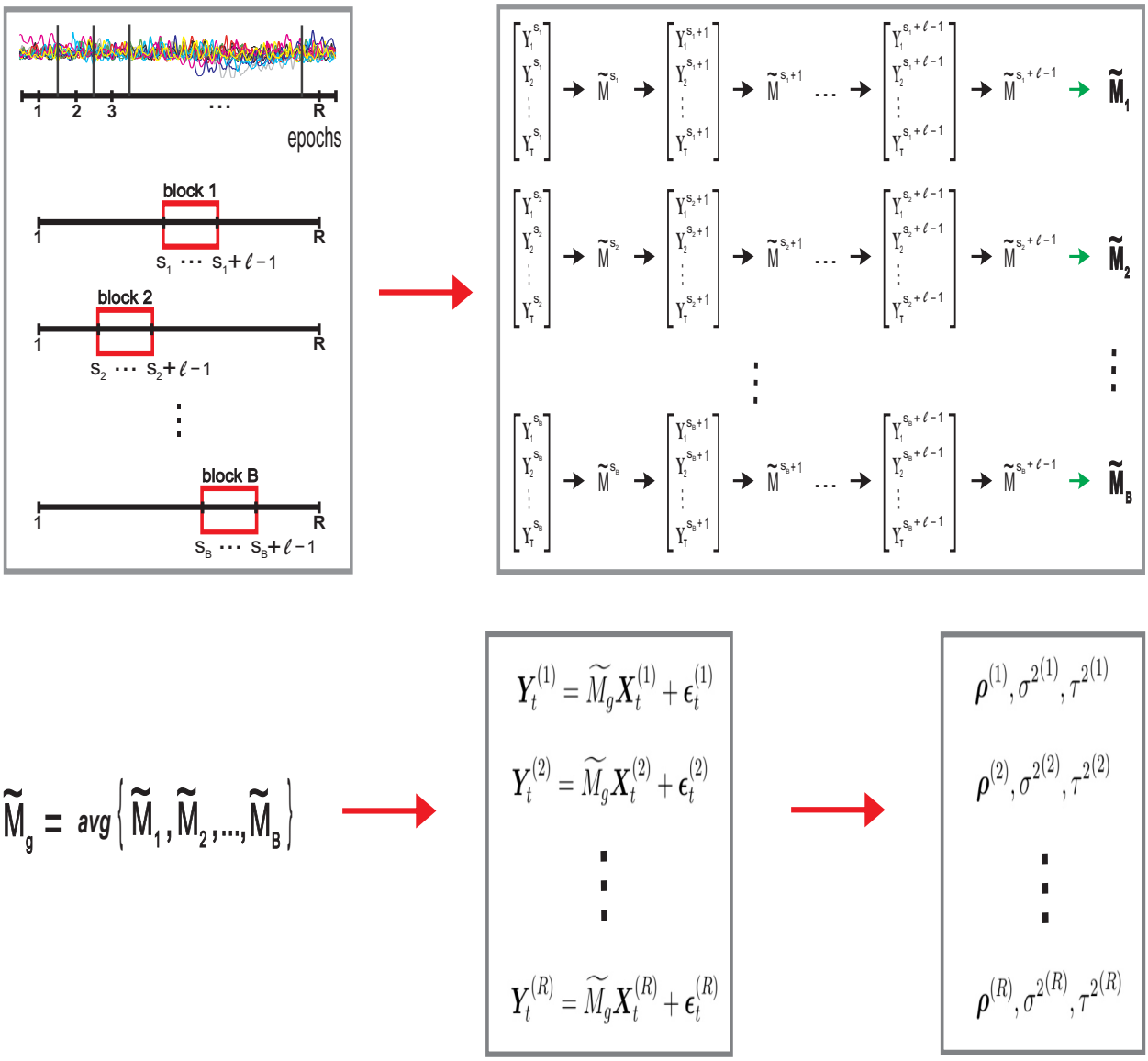



Figure A.10: Schematic illustration of the estimation methods that summarize II.A, II.B and II.C in Chapter 4.3.