CARNEGIE MELLON UNIVERSITY

DIETRICH COLLEGE OF HUMANITIES AND SOCIAL SCIENCES
DISSERTATION

Submitted in Partial Fulfillment of the Requirements
For the Degree of DOCTOR OF PHILOSOPHY

Title: “Source-Space Analyses in MEG/EEG and Applications to Explore Spatio-temporal Neural Dynamics in Human Vision”

Presented by: Ying Yang

Accepted by: The Center for the Neural Basis of Cognition
February 17, 2017

Thesis Committee:
Robert Kass (co-chair)
Michael Tarr (co-chair)
Geoffrey Gordon
Pulkit Grover
Matti Hamalainen (external)

Source-Space Analyses in MEG/EEG and Applications to Explore Spatio-temporal Neural Dynamics in Human Vision

Ying Yang

Center for the Neural basis of Cognition,
Dietrich College of Humanities and Social Sciences
and
Machine Learning Department,
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Thesis Committee:
Robert Kass (co-chair),
Michael Tarr (co-chair),
Geoffrey Gordon,
Pulkit Grover,
Matti Hamalainen

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy.

Copyright © 2017 Ying Yang

Keywords: Spatio-temporal dynamics, source localization, MEG, EEG, human vision, convolutional neural networks, information flow in the visual cortex

Abstract

Human cognition involves dynamic neural activities in distributed brain areas. For studying such neural mechanisms, magnetoencephalography (MEG) and electroencephalography (EEG) are two important techniques, as they non-invasively detect neural activities with a high temporal resolution. Recordings by MEG/EEG sensors can be approximated as a linear transformation of the neural activities in the brain space (i.e., the source space). However, we only have a limited number sensors compared with the many possible locations in the brain space; therefore it is challenging to estimate the source neural activities from the sensor recordings, in that we need to solve the underdetermined inverse problem of the linear transformation. Moreover, estimating source activities is typically an intermediate step, whereas the ultimate goal is to understand what information is coded and how information flows in the brain. This requires further statistical analysis of source activities. For example, to study what information is coded in different brain regions and temporal stages, we often regress neural activities on some external covariates; to study dynamic interactions between brain regions, we often quantify the statistical dependence among the activities in those regions through "connectivity" analysis.

Traditionally, these analyses are done in two steps: Step 1, solve the linear problem under some regularization or prior assumptions, (e.g., each source location being independent); Step 2, do the regression or connectivity analysis. However, biases induced in the regularization in Step 1 can not be adapted in Step 2 and thus may yield inaccurate regression or connectivity results. To tackle this issue, we present novel one-step methods of regression or connectivity analysis in the source space, where we explicitly modeled the dependence of source activities on the external covariates (in the regression analysis) or the cross-region dependence (in the connectivity analysis), jointly with the source-to-sensor linear transformation. In simulations, we observed better performance by our models than by commonly used two-step approaches, when our model assumptions are reasonably satisfied.

Besides the methodological contribution, we also applied our methods in a real MEG/EEG experiment, studying the spatio-temporal neural dynamics in the visual cortex. The human visual cortex is hypothesized to have a hierarchical organization, where low-level regions extract low-level features such as local edges, and high-level regions extract semantic features such as object categories. However, details about the spatio-temporal dynamics are less understood. Here, using both the two-step and our one-step regression models in the source space, we correlated neural responses to naturalistic scene images with the low-level and high-level features extracted from a well-trained convolutional neural network. Additionally, we also studied the interaction between regions along the hierarchy using the two-step and our one-step connectivity models. The results from the two-step and the one-step methods were generally consistent; however, the one-step methods demonstrated some intriguing advantages

in the regression analysis, and slightly different patterns in the connectivity analysis. In the consistent results, we not only observed an early-to-late shift from low-level to high-level features, which support feedforward information flow along the hierarchy, but also some novel evidence indicating non-feedforward information flow (e.g., top-down feedback). These results can help us better understand the neural computation in the visual cortex.

Finally, we compared the empirical sensitivity between MEG and EEG in this experiment, in detecting dependence between neural responses and visual features. Our results show that the less costly EEG was able to achieve comparable sensitivity with that in MEG when the number of observations was about twice of that in MEG. These results can help researchers empirically choose between MEG and EEG when planning their experiments with limited budgets.

Acknowledgments

In this long journey of pursuing my PhD, I was so fortunate to have guidance, support and help from many people. I would like to thank my great advisors, Mike Tarr and Rob Kass. I still remember the moment I first heard Mike's lecture on computational models for understanding human vision. It was an enthusiastic moment, when I first saw how desires to decipher human visual computation could be pursued in an elegant and concrete way. Indeed, it was Mike's boundless encouragement and insights that drove me through the frustrations and excitements in this journey. He was not only an academic model for me, but also a great friend, who had a huge impact on my philosophy of science and life. I have also learned a lot from Rob, who always kindly shared his expertise and insights on how to improve statistical models to rigorously answer scientific questions. Rob was also extremely resourceful and supportive, especially in helping me to combine statistics and machine learning with neuroimaging and cognitive neuroscience. The work in this thesis will not be possible without Mike and Rob's guidance and help.

I would also like to thank my committee members, Geoff Gordon, Matti Hamalainen and Pulkit Grover, for their time and their insightful advice and feedback on this thesis. I thank Marlene Behrmann, David Plaut, Avniel Ghuman and Tim Verstynen , who taught me a lot in psychology and cognitive science during the discussions in the "Viscog" group meetings. I was also very lucky to have collaborated with Elissa Aminoff, Will Bishop, John Pyles and Yang Xu, who never hesitated in sharing their expertise and thoughts. In addition, a huge thank-you goes to Tarrlab's research assistants, Carol Jew, Kevin Tan and Austin Marcus, for their help in data collection and being awesome officemates. I have also received a lot of help from Erika Laine, Shawn Walls, Scott Kurdilla, Michael Ward and Deborah Viszlay on collecting the MEG and fMRI data, and from Abhinav Gupta and Xinlei Chen for preparing the stimulus images.

I am also very grateful to spend these years with many smart and energetic fellow students, labmates and colleagues: Pengcheng Zhou, Natalie Klein, Mariya Toneva, Daniel Leeds, Yuanning Li, Yimeng Zhang, Praveen Venkatesh, Lingxue Zhang, Li-Yun Chang, Juliet Shafto, Amanda Robinson, Mark Vida, Qiong Zhang, Tina Liu, Charles Wu, Yuan Wang, Ut Na Sio and many others in the "Viscog" group, the "Neurostats" group, the Center for the Neural Basis of Cognition and the Machine Learning Department. I am grateful to know and learn from them and get their support and help in many ways.

I would like to thank my unique PhD program—Program in Neural Computation in the Center for the Neural Basis of Cognition, which created this fantastic interdisciplinary environment for me to interact with people in different fields and to embrace the sparkles when machine learning, statistics, math, engineering and neuroscience meet each other. This program also gave me the opportunity to be a joint student in the machine learning PhD program, which truly facilitated the combination of hard-core machine learning with hard-core neuroscience. My sincere thanks go to the director and staff in the Neural Computation program and the Machine Learning Department. In addition, I would like to thank the staff in the Psychology Department for their support, and the Global Communication Center at CMU for their help in writing the thesis.

Finally, I thank my dear parents, who went through the tough years of being apart from me but always gave me unconditional support and love. I also thank my boyfriend, Yifei Ma, who helped me in many respects and created and shared with me the great memories during our PhD years.

This work in this thesis was partly supported by the National Science Foundation Grant 1439237, the National Institute of Mental Health Grant RO1 MH64537, the grant from National Institute of Health for the Multimodel Neuroimaging Training Program (MNTP) (5R90DA023420-08,5R90DA023420-09), funding from the Richard King Mellon Foundation and the Henry L. Hillman Presidential Fellowship at Carnegie Mellon University.

Additional information

The Python implementation of the novel methods in thesis can be found at:
github.com/YingYang/STFT_R_git_repo
github.com/YingYang/MEEG_connectivity
Questions or comments about the thesis can be sent to
ying.yang.cnbc.cmu@gmail.com.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Understanding the neural mechanisms of human cognition is one of the major goals of neuroscience. Many sophisticated cognitive processes involve highly dynamic neural activities distributed across many areas in the brain. For example, object and scene recognition can be accurately accomplished in less than a second; such proficiency is achieved via information flow across multiple areas in the visual cortex. To understand the neural computation in human cognition, we need to analyze the joint spatio-temporal neural activities, by examining what information is coded at different temporal stages and spatial locations in the brain, and how activities at different locations interact with each other to implement the information flow. Such analysis requires non-invasive techniques that record neural activities with good temporal and spatial resolutions.

Among the currently available techniques, magnetoencephalography (MEG) and electroencephalography (EEG) provide a high temporal resolution, which is necessary to capture fast neural dynamics. Unlike other popular techniques, such as functional magnetic resonance imaging (fMRI), which measures slowly changing blood oxygen levels that are indirectly related to neural activities, MEG and EEG measure magnetic field changes or scalp voltage changes that are directly induced by neural electrical activities. Due to the short response time of the electromagnetic mechanism, MEG/EEG can achieve a temporal resolution at the millisecond level, which is much better than the temporal resolution of fMRI. However, the spatial resolution of MEG/EEG is limited by the *source localization* problem—estimating the neural activities in the brain from the MEG/EEG sensor recordings. If we know the true neural activities distributed in the brain space (termed as the *source space*), the sensor recordings can be approximated as a pre-computed linear transformation of the source activities. Solving the inverse of this linear problem is the key step in source localization. However, because there are only a limited number of sensors compared with the many possible locations in the source space, the linear problem is underdetermined, in that infinitely many source activity patterns can yield the same sensor recordings. Therefore it is challenging to achieve a good spatial resolution in MEG/EEG.

Nevertheless, a good temporal resolution is crucial for studying the spatio-temporal neural activities during fast cognitive processes, such as high-level vision (e.g., object and scene recognition). Therefore MEG and EEG are more useful tools than fMRI for this purpose. In this context, improving the spatial inference in MEG/EEG becomes important. Many previous methods of source localization solve the inverse problem and obtain estimates of source-space neural activities using different constraints, (e.g., constraints that encourage low variance, spatial sparsity, or spatio-temporal contiguity). However, estimating the source activities is not necessarily the final goal; researchers typically do further statistical analyses to answer their scientific questions. For example, to understand whether some brain region is involved in coding certain information, one can define external covariates to describe the information (e.g., whether a visual scene is indoor or outdoor), and then test whether the neural response in this region is correlated with the covariates; additionally, to gain insights about how information flows, one can estimate the statistical dependence among the neural activities in different regions (also known as "functional connectivity"). In these scenarios, researchers traditionally use a two-step approach: Step 1, applying some source localization method to estimate the source activities, and Step 2, running further analysis on these estimates, such as regression analysis against the covariates, or connectivity analysis across regions, in our two examples above.

The aforementioned two-step approach is intuitive, easy to implement, and could have given accurate results if the source estimates were correct. However, the source localization problem is inherently underdetermined. In the commonly used methods for Step 1, the constraints often represent limited prior assumptions about the source activities, such as different source locations being independent or only locally correlated without long-range cross-region dependence. These assumptions create biases in the estimates of source activities, which can not be adjusted according to the analysis in Step 2. To tackle this issue, we promote an alternative one-step approach. In this approach, we model the source activities according to the statistical analysis of interest (e.g. assuming dependence on external covariates or assuming cross-region dependence), and incorporate such models into the source localization problem, so we can directly fit these models from the sensor data in one step. Although we still add constraints to the model and thus create biases, these constraints can be chosen specifically according to the statistical analysis. Hence the one-step approach provides more flexibility and may give better results than the two-step approach. Browsing the previous literature, we found only a few publications exploiting this one-step approach [Gramfort et al., 2012; David et al., 2006; Fukushima et al., 2015], each of which has specific assumptions that only apply in limited cases. In this thesis, we present new one-step methods for source-space regression and connectivity analysis in MEG/EEG, which are based on arguably more general assumptions and can be applied in a wide range of scenarios.

Moreover, we present real applications of these methods in studying the spatio-temporal neural dynamics in high-level vision. By presenting naturalistic images of scenes to human participants, and analyzing their dynamic neural responses in the source space, we were able to get some new insights on what information is coded in different brain areas at different time stages, and how information flows in the visual cortex.

The thesis is organized in the following way. In Chapter 2 we briefly introduce some background about MEG and EEG and review previous source localization methods. Next, we present three major components of this thesis—our methodological contributions including models for one-step source-space analyses, our scientific findings on the spatio-temporal dynamics in the visual cortex, and finally an empirical comparison between MEG and EEG.

First, we introduce our methodological contribution, including one-step models for two types of analyses, (1) regressing source activities on external covariates, which is used to understand what information is coded in the spatio-temporal neural activities (Chapter 3), and (2) estimating dependence of source activities across given regions of interest, which is used to characterize functional connectivity in the brain (Chapter 4). We demonstrate the advantage of these methods using simulations or real-world applications.

Next, in Chapter 5, we present our scientific findings by applying the source-space analyses—both the traditional two-step methods and our novel one-step methods—to explore neural dynamics in the visual cortex. High-level vision, such as object and scene recognition, happens in a fast and sophisticated manner, and the underlying neural computation has been of great interest in both neuroscience and computer vision. It has been hypothesized that the visual cortex is organized hierarchically from posterior to anterior parts. In this hypothesis, different areas code features of visual inputs from low-level (e.g., local edges) to high-level (e.g., semantic labels such as "animate" and "inanimate"), and information may flow both in a bottom-up feedforward direction and a top-down feedback direction. However, besides such a vague description, details about the spatio-temporal dynamics are less known. Here, we recorded neural responses using MEG and EEG while human participants viewed naturalistic images of scenes, and then we characterized the dependence between the spatio-temporal neural responses and the different levels of visual features of the images, using source-space regression. Additionally, we also analyzed functional connectivity between different regions along the hierarchy. Our results not only demonstrated a clear pattern of feedforward information flow, but also gave novel evidence of non-feedforward dynamics. These results can provide insights to understand the neural computation.

Finally, although MEG and EEG are similar in many ways, they differ in their mechanisms, instantiation, and costs. As an additional empirical contribution, in Chapter 6 we provide a comparison between MEG and EEG, in detecting the dependence between the neural representations of scene images and the visual or semantic features of the images. MEG, which exploits expensive superconducting devices, gave slightly better signals than EEG. However, EEG was able to achieve comparable sensitivity to MEG, possibly with a larger number of observations but still much lower cost. These findings may help researchers in selecting between MEG and EEG according to their budgets. At the end, we summarize the conclusions and discuss future directions in Chapter 7.

Mathematical notations.

Before moving to the following chapters, we introduce general mathematical notations through out this thesis. Letters in the bold style (e.g. $\boldsymbol{y}$, $\boldsymbol{G}$ and $\boldsymbol{Q}$) denote vectors or matrices, and letters in the regular italic style denote scalars. The identity matrix is always denoted by $\boldsymbol{I}$. The fields of real numbers and complex numbers are denoted by $\mathbb{R}$ and $\mathbb{C}$ respectively. Accordingly, a real-valued matrix $\boldsymbol{G}$ of size $m \times n$ can be described as $\boldsymbol{G} \in \mathbb{R}^{m \times n}$. The transpose of a real matrix $\boldsymbol{G}$ is denoted by $\boldsymbol{G}'$, and the Hermitian transpose of a complex matrix $\boldsymbol{\Phi}$ is $\boldsymbol{\Phi}^H$. For any real square matrix $\boldsymbol{Q}$, $\det(\boldsymbol{Q})$ and $\mathrm{trace}(\boldsymbol{Q})$ are used to denote the determinant and trace of $\boldsymbol{Q}$.

When indexing elements of vectors and matrices, we use the following notations. The ith element in a vector $\boldsymbol{y}$ is written as $\boldsymbol{y}[i]$. Similarly, the entry in the ith row and jth column of a matrix $\boldsymbol{G}$ is written as $\boldsymbol{G}[i, j]$; the ith row and the jth column are written as $\boldsymbol{G}[i, :]$ and $\boldsymbol{G}[:, j]$ respectively. Following conventional definitions of norms, we define the L_2 norm of a vector $\boldsymbol{y}$ as $\|\boldsymbol{y}\|_2 = \sqrt{\sum_i (\boldsymbol{y}[i])^2}$, and the L_1 norm as $\|\boldsymbol{y}\|_1 = \sum_i |\boldsymbol{y}[i]|$. We also define the Frobenius norm of a matrix $\boldsymbol{G}$ as $\|\boldsymbol{G}\|_F = \sqrt{\sum_{i,j} (\boldsymbol{G}[i, j])^2}$. Finally, the notation $\mathcal{N}(\boldsymbol{y}, \boldsymbol{Q})$ denotes a normal (or Gaussian) distribution with mean $\boldsymbol{y}$ and covariance $\boldsymbol{Q}$.

Chapter 2

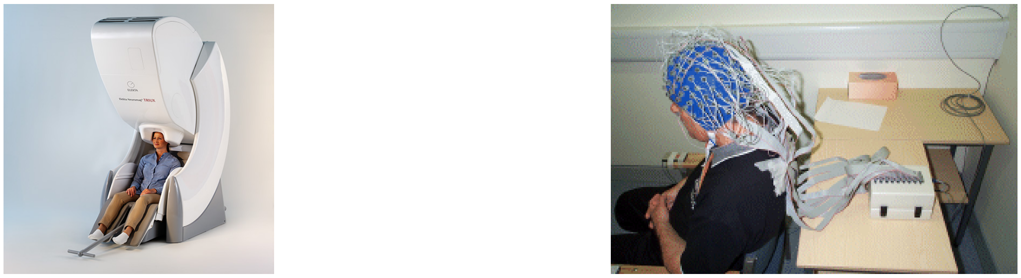
Background and related work on source localization in MEG/EEG

Magnetoencephalography (MEG) and electroencephalography (EEG) [1] are two widely used non-invasive techniques to record neural activities in the brain. The typical settings of MEG and EEG are shown in Figure 2.1. A MEG system typically has a helmet that includes a few hundred sensors. The participant sits or lies with their head located in the helmet. There are usually two types of sensors inside the helmet—the magnetometers, which measure the strength of the magnetic fields near the scalp, and the gradiometers, which measure the planar or axial gradients of the magnetic fields. An EEG system typically includes dozens of electrodes that are attached to a cap placed on the scalp, with conductive gel filled between each electrode and the scalp. The system measures voltages at different locations of the scalp in relation to some reference (e.g., some electrodes attached behind ears). In this chapter, we introduce how MEG and EEG work biophysically, and also briefly review related work on the *source localization* problem, which infers the neural activities in the brain space from MEG/EEG recordings.

2.1 Signal sources of MEG and EEG recordings

MEG and EEG are based on similar principles; the sensors measure either magnetic field changes outside the scalp (in MEG), or voltage changes on the scalp (in EEG), both of which are induced by electric neural activities in the brain. The major contributing activities are from the pyramidal cells in the cerebral cortex [Hamalainen et al., 1993]. As illustrated in Figure 2.2, pyramidal cells have long apical dendrites, which align perpendicularly to the cortical surface. These cells receive

[1] By "EEG", we only refer to the electroencephalography that measures scalp voltages in a non-invasive way. Another technique, electrocorticography, is often called "intracranial EEG". This technique measures voltages on exposed cortical surfaces in an invasive way. Although it shares some similarities with EEG, we will not discuss it in this thesis.



(a) MEG (Elekta, www.elekta.com) (b) EEG (Biosemi, www.biosemi.com)

Figure 2.1: Typical settings of MEG and EEG

ion influxes through synapses from other input neurons, which can be excitatory and inhibitory. Depending on the spatial distributions of the synapses on the dendrites and the cell body, as well as different activation patterns of input neurons at a given time point, the concentrations of positive and negative ions (or the density of electric charges) along the apical dendrite can be different, creating an electric current [Buzsáki et al., 2012]. Because the apical dendrites in nearby pyramidal cells align approximately in the same direction, when their activities are synchronized to some degree, the integration of these currents gets large enough to generate both magnetic field changes and scalp voltage changes that are detectable in MEG and EEG. These cortical synaptic currents are the primary sources of MEG/EEG signals. Due to the fast electromagnetic mechanism, the MEG and EEG recordings reflect the neural activities in an almost instantaneous way. Hence, MEG and EEG have a high temporal resolution (as high as the millisecond level), and they are well-suited for studying temporal dynamics of cortical activities.

Besides synaptic currents, action potentials along the axons can also induce magnetic field changes or scalp voltage changes. However, their effects decay with distance faster than the effects of synaptic currents, and therefore action potentials contribute much less to the sensed signals in MEG or EEG [Hamalainen et al., 1993]. Subcortical structures and the hippocampus, which are deeper and further away from the sensors, are also thought to contribute less to MEG/EEG recordings, although Quraan et al. [2011] demonstrated that in optimal conditions given strong signals, one may extract hippocampal activities from MEG recordings.

2.2 The forward model

At an arbitrary time point t, assume we know the electric current (denoted by $\vec{\boldsymbol{j}}_t(\vec{\boldsymbol{r}})$) at any position (denoted by a vector $\vec{\boldsymbol{r}}$) in the head. The magnetic field or the scalp voltage detected by one sensor can be modeled as an integration (i.e., a linearly weighted combination) of the currents at all positions, computed using Maxwell's equations under a reasonable head model that describes the shape, the electrical conductivity and the permeability of various tissues [Hamalainen et al., 1993; Mosher et al., 1999]. Formally, assume we have n sensors. Let an $n \times 1$ vector $\boldsymbol{y}_t$ denote the

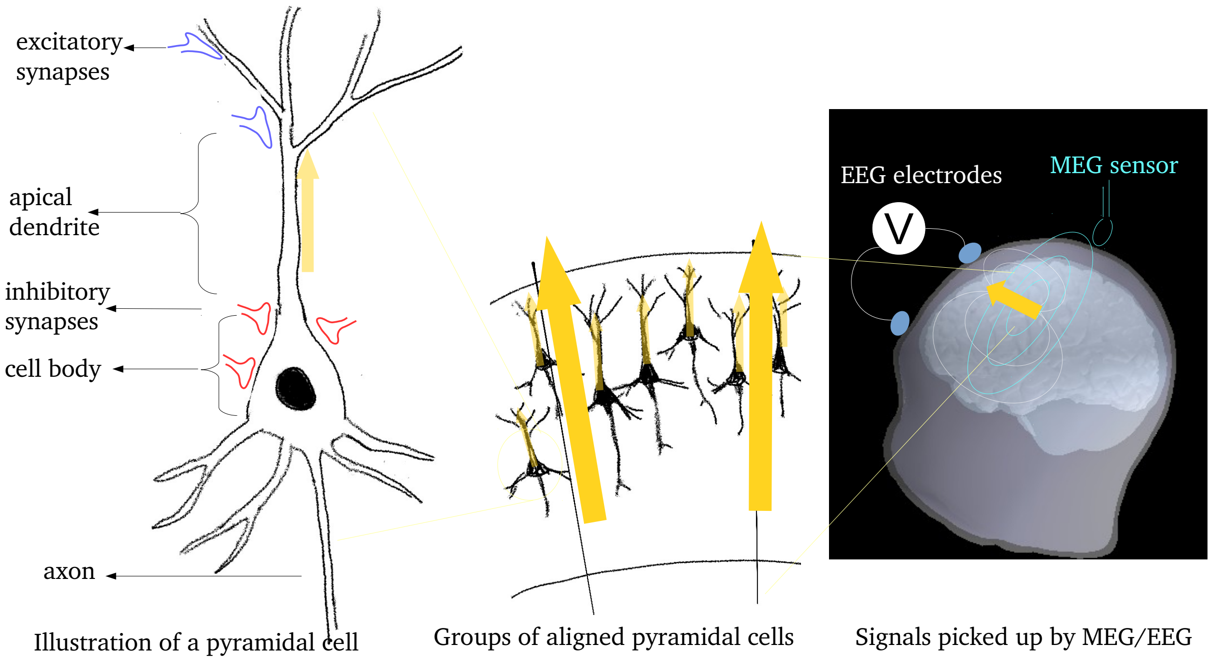

excitatory synapses
apical dendrite
inhibitory synapses
cell body
axon
Illustration of a pyramidal cell
Groups of aligned pyramidal cells
EEG electrodes
MEG sensor
V
Signals picked up by MEG/EEG

Figure 2.2: Illustration of the biophysical basis of MEG and EEG

readings from the n sensors at the time point t, and let $\boldsymbol{y}_t[i]$ denote the reading from the ith sensor ($i = 1, 2, \cdots, n$). From now on, we term this n-dimensional space the *sensor space*. We can decompose the reading from the ith sensor as a signal generated by the source current integration, plus some independent noise $\boldsymbol{e}_t[i]$, due to sensor fluctuations or other external noise. (Accordingly, let the $n \times 1$ vector $\boldsymbol{e}_t$ denote the noise in the n sensors.) Then we have the following *forward model*

$$\boldsymbol{y}_t[i] = \int G(\vec{\boldsymbol{r}}, i)\vec{\boldsymbol{j}}_t(\vec{\boldsymbol{r}})d\vec{\boldsymbol{r}} + \boldsymbol{e}_t[i] \quad \text{for } i = 1, 2, \cdots, n. \tag{2.1}$$

The weights (denoted by $G(\vec{\boldsymbol{r}}, i)$), describing the contribution of the current $\vec{\boldsymbol{j}}(\vec{\boldsymbol{r}})$ at any given location $\vec{\boldsymbol{r}}$ to the ith sensor, are often termed the *lead fields*. It is worth noting that the lead fields can be assumed static over time, given the head movement is accounted for in MEG and the EEG electrodes are attached to the scalp firmly. The computation of the lead fields is known as *forward modeling*, where various head models with different levels of simplification can be used. For example, in a spherically symmetric head model, there are multiple spherical layers describing the boundaries of different areas (the inner brain tissue plus the cerebrospinal fluid, the skull and the scalp skin), each assumed to have a constant electrical conductivity; more realistically, the geometry of the scalp and the skull of an individual participant can be constructed from structural MRI, and the lead fields at discrete locations can be solved numerically using the boundary element methods [Mosher et al., 1999].

Since the signal sources in both MEG and EEG are the neural activities in the brain, the two modalities have a lot of similarity in forward modeling. For example, they both have the following two properties of the lead fields that are worth mentioning. First, two currents can have similar contributions to the readings of sensors if they have close positions and similar orientations. Such spatial correlations, along with the number and placement of the sensors, affect the spatial resolution of localizing the source neural activities (see below, Section 2.3). Secondly, the magnitudes of the lead fields decrease as the distance between the currents and the sensors increases; therefore, the shallower currents that are closer to the scalp contribute more to the sensor readings than the deeper currents in general. Despite the aforementioned similarity, there are also some differences in the lead fields of MEG and EEG. For example, if we assume the head is a spherical volume conductor, the MEG sensors can only detect the effects induced by tangential currents, not radial currents, whereas the EEG sensors can detect the effects of both types of currents. In more realistic head models based on the structural MRI, similar issues still exist. Additionally, the magnitudes of lead fields in MEG fall off with distance more quickly than those in EEG, and as a result, MEG is less sensitive to deeper brain activities than EEG [Hamalainen et al., 1993; Malmivuo, 2012]. Moreover, the lead fields of EEG can be inaccurate if incorrect conductivity values of scalps and skulls are used in computing forward models, while the lead fields in EEG are less affected [Hamalainen et al., 1993].

2.3 Source localization—solving the inverse problem

To estimate the source currents in the brain from MEG/EEG recordings, we need to solve the inverse problem of the forward model. In other words, we need to estimate the source currents given the sensor readings. There are only a few hundred sensors, but infinitely many patterns of source currents can generate similar sensor readings, so the problem is inherently underdetermined. In order to get a unique solution, additional assumptions and constraints are needed. However, there are no perfect constraints that universally work for every research question. This challenging inverse problem has been an important research topic in the field. Below, we briefly review some prominent approaches in the literature.

Since the lead fields are assumed to be static, solving the inverse problem can be simplified as solving for the instantaneous source currents at any time t. In each of the approach reviewed below, we will start with methods for such cases, and then extend to methods that consider the source currents as time series and exploit temporal constraints. It is also worth noting that for simplification and identifiability, most approaches use an electric current dipole to represent the continuously distributed currents in a unit area of the cortical surface or a unit volume in the brain.

2.3.1 Equivalent dipole fitting

The first approach uses only a few "equivalent current dipoles" to approximate the integrated currents that contribute the most to the sensor readings. In an instantaneous setting at time t, each dipole has six parameters to be solved: the three coordinates representing the position and the dipole moment strength projected to three orthogonal basis directions. A further extension by Scherg [1990] uses dipoles that have static positions but time-varying strengths. The orientations can be fixed, as well as time-varying, if the dipoles are represented with three orthogonal components. The number of free parameters is often smaller than the number of sensors, and can be pre-defined using a principal component analysis of the sensor readings. The locations and (time-varying) strengths of the dipoles are obtained by minimizing the sum of squared differences between the sensor readings predicted from the dipoles and the observations. This step can often be done empirically by searching over a spatial grid for the positions that explain the sensor variance well, and then obtaining the (time-varying) dipole strengths with a least square regression. The equivalent dipole approach, although intuitive, is mostly suited when we are interested in source currents that are far apart and independent of each other. When several source regions have highly correlated activities, the resulting equivalent dipoles can be misleading. Moreover, we often want to study the distributed activities in large cortical areas, in which case the few equivalent dipoles might not be the most reasonable representation.

2.3.2 Spatial filtering

A second approach is spatial filtering, where for any given location, the strengths of the current dipole in three orthogonal basis directions are obtained by applying a linear spatial filter on the sensor readings. Such filtering is repeated at each location on a discrete grid covering the volume of the head, with the assumption that the source currents at different locations are linearly independent. The filter for each location is designed to keep the contribution to sensor readings from this location as much as possible, but to suppress contributions from all other locations. One commonly used method is the linearly constrained minimum variance (LCMV) beamforming [Van Veen et al., 1997], which optimizes the filter by minimizing filter output power (represented as the variance of the filtered results), subject to a constraint that the projection of the theoretical true signal at this location onto the filtered result is an identity matrix. This approach can also be applied on the time-frequency components of the MEG/EEG data, exploiting the possible oscillatory structure of the time series [Dalal et al., 2008]. However, the independent assumption in the spatial filtering approach is often wrong. In such cases, the source currents at different locations are estimated in separate models, not jointly in one model, making it difficult to discuss the dependence or connectivity of source activities between locations.

2.3.3 Probabilistic modeling of distributed sources

In the third approach, the continuous head space is discretized as either a 3-dimensional grid covering the volume of the head, or a 2-dimensional mesh covering the cortical surfaces of the left and right hemispheres, including the major contributing sources to the MEG/EEG signals. At each point on the grid or the mesh, the continuously distributed currents within the unit volume or area are represented by a current dipole. The orientation can be freely variable if the current dipole is represented by three orthogonal components; however, the orientation is often assumed to be fixed, perpendicular to the cortical surface, following the aligned direction of the local pyramidal cells. We call the points on the grid or the mesh the *source points* hereafter. With an average spacing of 5 to 7 mm, there are usually $m \sim 10^3$ to 10^4 source points. In the fixed-orientation case, the source current dipoles at each time point can be characterized by a real-valued m-dimensional vector, and we call this m-dimensional space the *source space*. Similarly, in the free-orientation case, the source space is $3m$-dimensional. With the discretization, the lead fields can be written as a matrix $\boldsymbol{G}$ of size $n \times m$ in the fixed-orientation case, or of size $n \times 3m$ in the free-orientation case. We call this matrix the *forward matrix*. Let an $m \times 1$ (or $3m \times 1$) vector $\boldsymbol{J}_t$ denote the source current dipoles at time point t. The forward model in (2.1) can be written as

$$\boldsymbol{y}_t = \boldsymbol{G}\boldsymbol{J}_t + \boldsymbol{e}_t. \tag{2.2}$$

That is, the $n \times 1$ sensor reading at time t, $\boldsymbol{y}_t$, is modeled as a linear projection of the source current dipoles $\boldsymbol{J}_t$ plus the $n \times 1$ sensor noise $\boldsymbol{e}_t$. The noise term $\boldsymbol{e}_t$ is usually modeled as a multivariate Gaussian variable with zero mean and covariance matrix $\boldsymbol{Q}_e$, (i.e., $\boldsymbol{e}_t \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{Q}_e)$), independent of $\boldsymbol{J}_t$. The noise covariance $\boldsymbol{Q}_e$ can be estimated from empty-room data, or recordings when the participant is at resting state [2].

Because $n \ll m$, to solve for $\boldsymbol{J}_t$, some constraints or an appropriate prior distribution need to be assumed. A common prior is a Gaussian distribution with zero mean and an $m \times m$ covariance matrix $\boldsymbol{Q}_J$ (see Figure 2.3a for the graphical model). In this case, we have

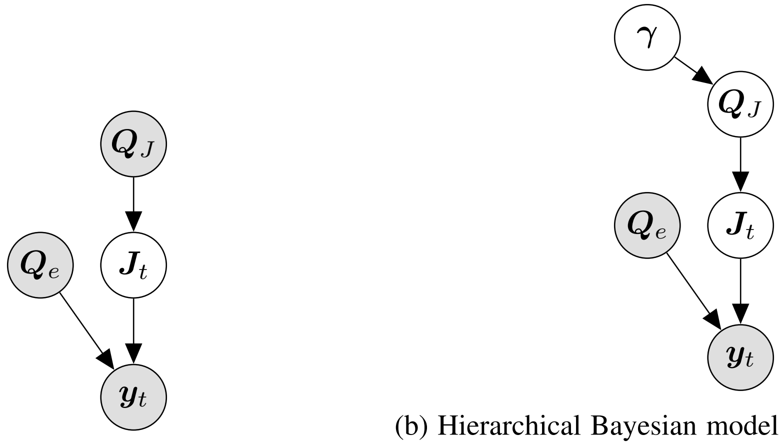
$$\begin{aligned} \boldsymbol{J}_t | \boldsymbol{Q}_J &\sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{Q}_J) \\ \boldsymbol{y}_t | \boldsymbol{J}_t, \boldsymbol{Q}_e &\sim \mathcal{N}(\boldsymbol{G}\boldsymbol{J}_t, \boldsymbol{Q}_e) \end{aligned} \tag{2.3}$$

Let $f(\cdot)$ denote the probability density function in general. According to the Bayes' rule,

$$f(\boldsymbol{J}_t | \boldsymbol{y}_t, \boldsymbol{Q}_J, \boldsymbol{Q}_e) \propto f(\boldsymbol{y}_t | \boldsymbol{J}_t, \boldsymbol{Q}_e) f(\boldsymbol{J}_t | \boldsymbol{Q}_J)$$

If $\boldsymbol{Q}_J$ is given, the estimate of $\boldsymbol{J}_t$ can by easily solved by maximizing the posterior probability

[2] From here on, the notation of many variables will be consistent throughout the thesis. The readers can always refer to Table 2.1 to look up the meanings of the variables.



(a) The simple Bayesian model

(b) Hierarchical Bayesian model with a hyperprior of $\boldsymbol{Q}_J$

Figure 2.3: Graphical models for the Bayesian methods

density $f(\boldsymbol{J}_t|\boldsymbol{y}_t, \boldsymbol{Q}_J, \boldsymbol{Q}_e)$. This maximum a posteriori estimator $\hat{\boldsymbol{J}}_t$ is

$$\begin{aligned}\hat{\boldsymbol{J}}_t =&\arg\min_{\boldsymbol{J}_t}((\boldsymbol{y}_t - \boldsymbol{G}\boldsymbol{J}_t)'\boldsymbol{Q}_e^{-1}(\boldsymbol{y}_t - \boldsymbol{G}\boldsymbol{J}_t) + \boldsymbol{J}_t'\boldsymbol{Q}_J^{-1}\boldsymbol{J}_t)\\ =&(\boldsymbol{G}'\boldsymbol{Q}_e^{-1}\boldsymbol{G} + \boldsymbol{Q}_J^{-1})^{-1}\boldsymbol{G}'\boldsymbol{Q}_e^{-1}\boldsymbol{y}_t\\ =&\boldsymbol{Q}_J\boldsymbol{G}'(\boldsymbol{G}\boldsymbol{Q}_J\boldsymbol{G}' + \boldsymbol{Q}_e)^{-1}\boldsymbol{y}_t \end{aligned} \tag{2.4}$$

where the Woodbury matrix identity $(\boldsymbol{G}'\boldsymbol{Q}_e^{-1}\boldsymbol{G}+\boldsymbol{Q}_J^{-1})^{-1} = \boldsymbol{Q}_J - \boldsymbol{Q}_J\boldsymbol{G}'(\boldsymbol{G}\boldsymbol{Q}_J\boldsymbol{G}' + \boldsymbol{Q}_e)^{-1}\boldsymbol{G}\boldsymbol{Q}_J$ is used to get (2.4). Note if we define $\boldsymbol{M} = \boldsymbol{Q}_J\boldsymbol{G}'(\boldsymbol{G}\boldsymbol{Q}_J\boldsymbol{G}' + \boldsymbol{Q}_e)^{-1}$, then $\boldsymbol{J}_t$ is obtained by applying the time-invariant linear operator $\boldsymbol{M}$ on $\boldsymbol{y}_t$ (i.e. $\hat{\boldsymbol{J}}_t = \boldsymbol{M}\boldsymbol{y}_t$). Such an estimate of $\boldsymbol{J}_t$ is computationally simple and also flexible to incorporate different assumptions. Many commonly used methods are in this framework, and we list a few below.

The first example is the minimum norm estimate (MNE, Hamalainen and Ilmoniemi [1994]), which assumes that $\boldsymbol{Q}_J$ is an identity matrix multiplied by a positive scalar $1/\lambda$. In the prior, each source current dipole is independent of others, and all source points share the same variance $1/\lambda$. Plugging $\boldsymbol{Q}_J = 1/\lambda\boldsymbol{I}$ in (2.4), we have a maximum a posteriori estimate $\hat{\boldsymbol{J}}_t = \boldsymbol{G}(\boldsymbol{G}'\boldsymbol{G} + \lambda\boldsymbol{Q}_e)^{-1}\boldsymbol{y}_t$. If $\boldsymbol{y}_t$ and $\boldsymbol{G}$ can be pre-whitened, (i.e., left multiplied by $\boldsymbol{Q}_e^{-1/2}$), then we can assume the new $\boldsymbol{Q}_e = \boldsymbol{I}$ and we can estimate $\boldsymbol{J}_t$ as

$$\arg\min_{\boldsymbol{J}_t}(\|\boldsymbol{y}_t - \boldsymbol{G}\boldsymbol{J}_t\|_2^2 + \lambda\|\boldsymbol{J}_t\|_2^2) = (\boldsymbol{G}'\boldsymbol{G} + \lambda\boldsymbol{I})^{-1}\boldsymbol{G}'\boldsymbol{y}_t = \boldsymbol{G}'(\boldsymbol{G}\boldsymbol{G}' + \lambda\boldsymbol{I})^{-1}\boldsymbol{y}_t$$

Thus the problem is reduced to a ridge regression or an L_2-norm regularized regression in the statistics literature. However, there is one issue with the MNE method—the source points that are closer to the scalp are often emphasized because their corresponding columns in $\boldsymbol{G}$ have larger norms. To alleviate such biases, some extensions of the MNE method, including the "dynamic statistical parametric mapping" (dSPM, Dale et al. [2000]) and the "standardized low resolution electromagnetic tomography" (sLORETA Pascual-Marqui [2002]), normalize the estimated source current dipoles to unitless statistics, allowing easier comparisons among source points. The dSPM method assumes a null hypothesis $\boldsymbol{y}_t = \boldsymbol{e}_t$, and computes a covariance matrix of $\hat{\boldsymbol{J}}_t$ under this null hypothesis ($\boldsymbol{M}\boldsymbol{Q}_e\boldsymbol{M}'$). It then divides $\hat{\boldsymbol{J}}_t$ by the marginal standard deviations—the square root of the diagonal entries in this covariance matrix—yielding the Z-statistics for testing the null hypothesis (i.e., $\boldsymbol{y}_t = \boldsymbol{e}_t$). The sLORETA method follows a similar idea but estimates the covariance of $\hat{\boldsymbol{J}}_t$ in a slightly different way.

A second example in the framework of (2.4) is the "low resolution electromagnetic tomography" (LORETA, Pascual-Marqui et al. [1994]), where $\boldsymbol{Q}_J$ is equivalent to the inverse of a Laplacian matrix, which expresses spatial correlations between adjacent source points. In this way, the model encourages spatially smooth solutions.

Thirdly, prior knowledge from functional magnetic resonance imaging (fMRI) can also be embedded in $\boldsymbol{Q}_J$. For example, in the fMRI-weighted-minimum-norm estimation (fMNE, Liu and He [2008]; Ahlfors and Simpson [2004]), weights determined by fMRI results are combined into the variance of the source points (i.e., the diagonal of $\boldsymbol{Q}_J$), to obtain spatial results that are consistent with prior knowledge.

Besides the examples above, which use pre-determined $\boldsymbol{Q}_J$, more flexible hierarchical Bayesian methods have been recently proposed (Figure 2.3b), where $\boldsymbol{Q}_J$ is also a random variable with a prior distribution. In the work by Mattout et al. [2006], Henson et al. [2011] and Wipf and Nagarajan [2009], $\boldsymbol{Q}_J$ is modeled as a linear combination of a set of basis covariance matrices, $\boldsymbol{C}_i$s, (i.e., $\boldsymbol{Q}_J = \sum_i \gamma_i \boldsymbol{C}_i, \quad \gamma_i > 0$). The $\boldsymbol{C}_i$s are usually used to capture the local covariance in spatially grouped patches of source points. Additional $\boldsymbol{C}_i$s characterizing priors from fMRI or anatomical knowledge can also be included. The hyperparameter $\boldsymbol{\gamma} = (\gamma_1, \cdots, \gamma_i, \cdots)$ can be assumed to have a flat prior and can be estimated by maximizing the marginal likelihood. Such a setting, known as the Gamma-MAP [Wipf and Nagarajan, 2009]—equivalent to automatic relevance determination [Wipf and Nagarajan, 2008] in a more general sense—can often give solutions where only a sparse set of patches are non-zero, but the estimates are locally smooth within the patches. Such patterns provide good interpretability in many cases. The hyperparameter $\boldsymbol{\gamma}$ can also be assumed to have some other prior distribution, allowing researchers to flexibly incorporate other assumptions; in these cases, the hyperparameter can be estimated via variational or sampling methods.

A second family of methods directly encourage sparsity by penalizing the L_1 norms of the source current dipoles, or a sum of L_2 norms of grouped source current dipoles. An early implementation, known as the minimum current estimate (MCE), was introduced by Matsuura and Okabe [1995]. Later, following the development of the regression models that induce structured sparsity in the statistical literature (e.g., the "lasso" [Tibshirani, 1996] and "group lasso" [Friedman et al., 2010], etc), more methods along this line have been proposed [Gramfort et al., 2012, 2013; Babadi et al., 2014]. These methods are favorable not only because of the interpretability of the structured sparsity, but also because the solutions can be estimated with efficient convex optimization algorithms or greedy methods. It is worth noting that although the implementation of these sparsity-inducing methods might not be Bayesian, there are equivalent prior distributions (e.g., Laplace distribution) that can formulate the methods into a Bayesian framework. Therefore we still generally categorize these methods as probabilistic models.

Many of the methods above (MNE, LORETA, fMNE) are applied to each instantaneous time point separately. For the hierarchical Bayesian models where $\boldsymbol{Q_J}$ needs to be estimated, data at different time points are often treated as independently and identically distributed (i.i.d.) samples as well. Further improvements can be made by modeling the temporal smoothness or oscillatory structure in the source neural activities. On the one hand, autoregressive models and Kalman smoothing, which characterize local spatial and temporal dependence [Baillet and Garnero, 1997; Galka et al., 2004; Lamus et al., 2012] have been shown to achieve better estimates than the instantaneous models, although the temporal stationary assumption in these methods is not necessarily satisfied in practice. On the other hand, Gramfort et al. [2013] have introduced a time-frequency decomposition of the source activities using the short-time Fourier transform, which is tolerant of non-stationary data; along with a sparsity-inducing penalty, their method yields source solutions that are both spatially sparse and temporally smooth with easily interpretable waveforms.

In additional to what we have reviewed above, recent work on source localization methods has also been focused on integrating multiple brain imaging modalities together (e.g., combining MEG and EEG with fMRI). Although some methods above do use fMRI results as priors, they are asymmetric in the sense that the fMRI priors are not modeled with uncertainty. Ou et al. [2010] and Henson et al. [2011] both proposed to model the MEG/EEG and fMRI recordings (of the same task) symmetrically with Bayesian graphical models, allowing integration of the strengths while still flexibly accounting for the uncertainty in each modality.

So far, we have introduced how MEG and EEG work, and reviewed a variety of methods for source localization in MEG and EEG. As a short summary, given the underdetermined nature of the source localization problem, one has to exploit additional assumptions to solve it. No single set of assumptions can work uniformly well in recovering the true source activities under all possible circumstances, and therefore when talking about which methods perform well, we should always note whether the assumptions are empirically met in the context of applications. Among the three approaches we have seen above, the third one, probabilistic modeling of distributed sources, allows formulating the problem with concepts in statistical machine learning, and gives a lot of flexibility. Hence we use this framework in this thesis. It is worth mentioning that there are some practical issues in this framework. First, in defining the source space, we need to choose between a mesh covering the cortical surfaces and a grid covering the volume of the head. In the former case, where the orientations of the current dipoles are usually assumed to be perpendicular to the local surface, we can focus on main sources of MEG and EEG signals—the activities of cortical pyramidal cells. In the latter case, where the current dipoles have free orientations, we can possibly capture source currents due to other neural activities. Secondly, even for a source space defined on the mesh, we can still allow free orientations of current dipoles by adding two tangential basis directions besides the main direction that is perpendicular to the cortical surface. Sometimes, we can assume a "looseorientation", where we penalize the magnitudes of components of source current dipoles along the tangential directions more than those along the main directions. Thirdly, as mentioned above, shallower source points, which are closer to the scalp, correspond to the columns in the forward matrix $\boldsymbol{G}$ that have larger norms, resulting in a bias towards the shallower sources; therefore, sometimes, a "depth compensation" can be applied, where we can choose a real value $0 < a_{depth} < 1$ and divide each column $\boldsymbol{G}[:, j]$ $(j = 1 \cdots, m)$ by $\|\boldsymbol{G}[:, j]\|_2^{a_{depth}}$. However, all these choices above are up to the user, depending on their assumptions. In the scientific applications in this thesis, we are mainly interested in cortical activities, (i.e., activities of pyramidal cells perpendicular to the cortical surface), so we define the source space as the mesh on the cortical surfaces and assume fixed orientations of the current dipoles. In some cases, especially when analyzing real data other than simulated data, we apply the depth compensation ($a_{depth} = 0.8$). See the individual chapters for more details.

The source localization methods aforementioned have focused on reconstructing the source activities per se, leaving statistical analyses in the source space as the next step. Yet the actual source-space statistical analyses—such as testing whether neural activities are correlated with behavior measurements in a task or analyzing dependence between regions—can also affect how we constrain the source localization problem. In the following two chapters, we present our novel methodological contribution—one-step source-space analyses with source localization embedded, which is based on some ideas reviewed above.

Table 2.1: Notation list

Symbol	Explanation
n	number of sensors in MEG or EEG
m	number of source points in the source space
$\boldsymbol{y}_t$	$n \times 1$ vector, the sensor readings at time t
$\boldsymbol{J}_t$	$m \times 1$ vector, the source current dipoles at time t
$\boldsymbol{e}_t$	$n \times 1$ vector, the sensor noise at time t
$\boldsymbol{G}$	$n \times m$ matrix, the forward matrix that projects source current dipoles to the sensor space
T	number of time points to be considered (in a trial)
$\boldsymbol{Y}$	$n \times T$ matrix, the sensor readings at T timepoints (in a trial), $\boldsymbol{y}_t = \boldsymbol{Y}[:, t]$
$\boldsymbol{J}$	$m \times T$ matrix, the source current dipoles at T time points (in a trial), $\boldsymbol{J}_t = \boldsymbol{J}[:, t]$
$\boldsymbol{E}$	$n \times T$ matrix, the sensor noise at T timepoints (in a trial), $\boldsymbol{e}_t = \boldsymbol{E}[:, t]$
q	number of trials to be considered
$.^{(r)}$	the superscript indicates that the variable is from the rth trial ($r \in \{1, 2, \cdots, q\}$), e.g. $\boldsymbol{Y}^{(r)}$ and $\boldsymbol{y}_t^{(r)}$

Chapter 3

One-step regression analysis in the source space: a short-time Fourier transform regression model

3.1 Motivation

MEG and EEG experiments are often designed to study the neural basis of perceptual or cognitive functions. These studies are usually organized in trials, each of which contains a stimulus and/or a behavioral response in a relevant task. Usually, there are two types of designs. In the first type, the trials are organized into different conditions (e.g., an experimental condition, such as reading words, and a control condition, such as reading non-word strings of letters), and each trial is correspondingly assigned a categorical label indicating which condition it is in. In these cases, the goal is to identify the neural activities that are different across the conditions (e.g., the activities that are only involved in the experimental condition but not in the control condition), and one can apply a two-sample t-test or analysis of variance for the purpose. In the second type, there can be one or more continuous variables associated with each trial (for example, certain features of the presented stimulus, or behavioral metrics such as reaction time), and the goal is to identify neural activities that are correlated with these variables. In this context, one can regress the neural activities on the variables and examine how well these variables explain or predict the neural data. In fact, the analyses in the two cases above can both be unified into a regression framework, because in the first type of design, if we represent the categorical condition labels as dummy variables, a t-test (or analysis of variance) is the equivalent to a linear regression model. Hereafter, we mainly focus on the regression framework and use the term *covariates* or *regressors* to denote either the categorical labels or the continuous variables. Although many sophisticated regression models are available, linear regression is the most commonly used because it is simple and easily interpretable, so below we focus on linear regression analysis of MEG and EEG data.

When analyzing MEG and EEG data using linear regression, we are interested in spatio-temporal effects, that is, in what temporal windows and in what regions in the brain, the neural activity is correlated with, or is well explained by the covariates. If we could observe the neural activity in the source space (i.e., the electric current dipole) at each location and each time point, then we could run an individual linear regression model for each time point and location, testing if a significant amount of variance was explained by the covariates. However, as discussed in Chapter 2, we need to solve the source localization problem to estimate the current dipoles. Traditionally, one can use a two-step approach: 1. applying some source localization method (e.g., the minimum norm estimate or MNE) to estimate the source current dipoles in each trial; 2. linearly regressing the estimates at each location and time point on the covariates.

However, this two-step approach is limited for the following reason. To regularize the underdetermined source localization problem, one has to add some constraints in Step 1, (e.g., penalizing the squared L_2 norm of the source current dipoles as in MNE), but these constraints can not be customized for the regression in Step 2. For example, if we want the regression coefficients for certain covariates to be spatially sparse, the regularization in Step 1 can not directly implement such constraints.

To formally describe this, let us assume we have q trials in total, and let a $q \times p$ matrix $\boldsymbol{X}$ denote the p-dimensional covariates. Using the notation in Chapter 2, we focus on a single time point t for simplicity. Let $\boldsymbol{y}_t^{(r)}$ be the $n \times 1$ sensor recordings in the rth trial, $\boldsymbol{J}_t^{(r)}$ be the corresponding $m \times 1$ source current dipoles, $\boldsymbol{G}$ be the $n \times m$ forward matrix, and $\boldsymbol{e}_t^{(r)}$ be the $n \times 1$ Gaussian sensor noise, with zero mean and covariance $\boldsymbol{Q}_e$ (i.e. $\boldsymbol{e}_t^{(r)} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{Q}_e)$). We further assume the ith source current dipole $\boldsymbol{J}_t^{(r)}[i]$ has a linear relationship with the p-dimensional covariates $\boldsymbol{X}[r,:]$, characterized by the regression coefficients $\boldsymbol{A}_t[i,:]$. Then the regression coefficients at all source points at time t can be written as an $m \times p$ matrix $\boldsymbol{A}_t$. The residuals not explained by the covariates are represented as $m \times 1$ source-space noise $\boldsymbol{w}_t^{(r)} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{Q}_{J_n})$, with an $m \times m$ source noise covariance matrix $\boldsymbol{Q}_{J_n}$. In the two-step framework, we have the following equations:

$$\text{Step 1: } \boldsymbol{y}_t^{(r)} = \boldsymbol{G}\boldsymbol{J}_t^{(r)} + \boldsymbol{e}_t^{(r)} \tag{3.1}$$

$$\text{Step 2: } \boldsymbol{J}_t^{(r)} = \underset{m\times p}{\boldsymbol{A}_t}\, \underset{p\times 1}{\boldsymbol{X}[r,:]'} + \underset{m\times 1}{\boldsymbol{w}_t^{(r)}} \tag{3.2}$$

If we use the MNE method for source localization, then with the penalization parameter λ, we have

$$\hat{\boldsymbol{J}}_t^{(r)} = \arg\min_{\boldsymbol{J}_t^{(r)}} \left((\boldsymbol{y}_t^{(r)} - \boldsymbol{G}\boldsymbol{J}_t^{(r)})'\boldsymbol{Q}_e^{-1}(\boldsymbol{y}_t^{(r)} - \boldsymbol{G}\boldsymbol{J}_t^{(r)}) + \lambda\|\boldsymbol{J}_t^{(r)}\|_2^2 \right) = \boldsymbol{G}(\boldsymbol{G}'\boldsymbol{G} + \lambda\boldsymbol{Q}_e)^{-1}\boldsymbol{y}_t^{(r)} \tag{3.3}$$

$$\hat{\boldsymbol{A}}_t = \arg\min_{\boldsymbol{A}_t} \left(\sum_{r=1}^{q} (\hat{\boldsymbol{J}}_t^{(r)} - \boldsymbol{A}_t\boldsymbol{X}[r,:]')'\boldsymbol{Q}_{J_n}^{-1}(\hat{\boldsymbol{J}}_t^{(r)} - \boldsymbol{A}_t\boldsymbol{X}[r,:]') + \text{penalty}(\boldsymbol{A}_t) \right) \tag{3.4}$$

Note that the penalty term that reflects our constraints on $\boldsymbol{A}_t$ in Equation (3.4) has no effect on the source localization step in Equation (3.3). The constraints in Step 1 (e.g., the L_2 penalty in Equation (3.3)) may penalize $\hat{\boldsymbol{J}}_t^{(r)}$ and consequently penalize $\hat{\boldsymbol{A}}_t$ in a different way from we want by using penalty$(\boldsymbol{A}_t)$ alone. In other words, the fixed constraints in Step 1 may introduce additional biases other than what we intend to introduce. Alternatively, if we plug Equation (3.2) into Equation (3.1), we have

$$\boldsymbol{y}_t^{(r)} = \boldsymbol{G}\boldsymbol{A}_t\boldsymbol{X}[r,:]' + \boldsymbol{G}\boldsymbol{w}_t^{(r)} + \boldsymbol{e}_t^{(r)}$$

Noticing that the covariance of the term $(\boldsymbol{G}\boldsymbol{w}_t^{(r)} + \boldsymbol{e}_t^{(r)})$ is $(\boldsymbol{Q}_e + \boldsymbol{G}\boldsymbol{Q}_{J_n}\boldsymbol{G}')$, we can directly solve for $\boldsymbol{A}_t$ in one step:

$$\min_{\boldsymbol{A}_t}\left(\sum_{r=1}^{q}(\boldsymbol{y}_t^{(r)} - \boldsymbol{G}\boldsymbol{A}_t\boldsymbol{X}[r,:]')'(\boldsymbol{Q}_e + \boldsymbol{G}\boldsymbol{Q}_{J_n}\boldsymbol{G}')^{-1}(\boldsymbol{y}_t^{(r)} - \boldsymbol{G}\boldsymbol{A}_t\boldsymbol{X}[r,:]') + \text{penalty}(\boldsymbol{A_t})\right).$$

In this one-step approach, the constraints in penalty$(\boldsymbol{A_t})$ are directly applied. It is also worth noting that in practical applications of the two-step approach, the source localization methods in Step 1 are usually the probabilistic methods that assume a fixed prior of the source current dipoles and do not introduce sparsity (e.g., MNE), so that they can be applied uniformly to all trials. Other methods that have sparsity-inducing constraints or hyper-parameters determined by the data, may result in different non-zero structures or hyper-parameters for each trial. Therefore, it is difficult to directly plug those methods in the two-step approach.

Given the limitations of the two-step approach, we propose to use the alternative one-step approach, which models the regression and the source localization problem jointly and allows flexible penalties directly on the regression coefficients. We speculate that this approach can provide more flexibility and give better results. In fact, there is some previous exploration along this line in comparing MEG/EEG data between different task conditions (i.e., with categorical covariates) [Gramfort et al., 2012]. Here we introduce a more general implementation of the one-step approach, built on a recent source localization model by Gramfort et al. [2013].

By using the short-time Fourier transform (STFT) to represent the time series of source current dipoles, and by further applying a structured sparsity-inducing penalty on the time-frequency components, Gramfort et al. [2013] were able to reconstruct source time series, which were sparse in the spatial domain and the time-frequency domain. These reconstructed source time series are generally smooth but still have relatively sharp peaks corresponding to the evoked responses—it is easy to interpret these “clean” waveforms in terms of timing and magnitudes of different peaks. Based on this previous work, we propose a new short-time Fourier transform regression model (STFT-R). Our model further represents the trial-by-trial variations of each time-frequency component as a linear combination of the trial-by-trial covariates; it also imposes a hierarchical sparsity-inducing penalty on the regression coefficients, instead of on the time-frequency components per se as in Gramfort et al. [2013]. In this case, our STFT-R model yields regression coefficients that are sparse in the source space and in the time-frequency domain and are likely to have similar “clean” waveforms. Additionally, when there are regions of interest (ROIs) that we want to emphasize, the hierarchical constraints can group the source points within each ROIs, and emphasize the ROIs by using smaller penalties on these groups.

Below we give a formal description of the STFT-R and corresponding optimization algorithms, and then we compare the STFT-R with a common two-step approach that uses MNE in both simulations and applications on real data.

3.2 Model

In this section, we introduce our STFT-R model and the related algorithms to solve for the regression coefficients. The code in Python is available at `github.com/YingYang/STFT_R_git_repo`.

3.2.1 Short-time Fourier transform (STFT)

Our approach builds on the short-time Fourier transform (STFT) implemented in Gramfort et al. [2013]. Given a time series $\boldsymbol{a} = (a_1, a_2, \cdots, a_t, \cdots, a_T)'$, where $t = 1, \cdots, T$, a time step τ_0 and a window size T_0, we define the STFT as

$$b(\boldsymbol{a}, \tau, \omega_{\mathfrak{l}}) = \sum_{t=1}^{T} a_t K(t-\tau) e^{(-\mathtt{i}\omega_{\mathfrak{l}} t)}$$

for $\omega_{\mathfrak{l}} = 2\pi\mathfrak{l}/T_0, \mathfrak{l} = 0, 1, \cdots, T_0/2$ and $\tau = \tau_0, 2\tau_0, \cdots n_0\tau_0$ where $n_0 = T/\tau_0$. Here $\mathtt{i}$ denotes the unit imaginary number (i.e., $\mathtt{i}^2 = -1$). The real and imaginary parts of the basis functions (i.e., $\{K(t-\tau)e^{(-\mathtt{i}\omega_{\mathfrak{l}} t)}\}_{t=1}^{T}$, for each τ and $\omega_{\mathfrak{l}}$) are sinusoidal functions of frequency $\omega_{\mathfrak{l}}$ multiplied by a window function $K(t-\tau)$ (e.g. half-cycle sine window, $K(t-\tau) = sin(\pi(t-\tau+T_0/2)/T_0)$) centered at time τ. Together, we have $s = (T_0/2+1) \times n_0$ basis functions, and correspondingly s time-frequency components (i.e., the $b(\boldsymbol{a}, \tau, \omega_{\mathfrak{l}})$s). Concatenating the s components into a complex-valued vector $\boldsymbol{b} \in \mathbb{C}^{s\times 1}$, and writing the basis functions $\{K(t-\tau)e^{(-\mathtt{i}\omega_{\mathfrak{l}} t)}\}_{t=1}^{T}$ as columns in an $T \times s$ complex-valued matrix $\boldsymbol{\Phi} \in \mathbb{C}^{T\times s}$, we have

$$\underset{1\times s}{\boldsymbol{b}'} = \underset{1\times T}{\boldsymbol{a}'} \underset{T\times s}{\boldsymbol{\Phi}}$$

With appropriate constants in the basis function, if we use $\boldsymbol{\Phi}^H$ to denote the Hermitian (or conjugate) transpose of $\boldsymbol{\Phi}$, then $\boldsymbol{\Phi}\boldsymbol{\Phi}^H = \boldsymbol{I}$. Then the inverse short-time Fourier transform can be written as

$$\underset{1\times T}{\boldsymbol{a}'} = \underset{1\times s}{\boldsymbol{b}'} \underset{s\times T}{\boldsymbol{\Phi}^H}.$$

3.2.2 The short-time Fourier transform regression model (STFT-R)

In accordance with our notation in Table 2.1, let us assume we have n sensors, m source points, T time points in each trial, and q trials together. Let $\boldsymbol{Y}^{(r)} \in \mathbb{R}^{n\times T}$ be the sensor time series we observe in the rth trial, and $\boldsymbol{G} \in \mathbb{R}^{n \times m}$ be the forward matrix that projects the m-dimensional source activity to the n-dimensional sensor space. As discussed in Section 3.2.1, let $\boldsymbol{\Phi}^H \in \mathbb{C}^{s \times T}$ be s pre-defined STFT basis functions at different frequencies and time points. Let the $q \times p$ matrix $\boldsymbol{X}$ denote the p-dimensional covariates in the q trials, where $\boldsymbol{X}[r,k]$ is the kth covariate in the rth trial, $r = 1, \cdots, q$ and $k = 1, \cdots, p$. Note that $\boldsymbol{X}$ may include an additional all-one column to represent the intercept, and besides this possible all-one column, we assume other columns have zero means. Let the $m \times T$ matrix $\boldsymbol{J}^{(r)}$ be the time series of m source points in the rth trial. We assume that the time series of the ith source point is represented by s time-frequency components: $\boldsymbol{b}_i^{(r)} \in \mathbb{C}^{1 \times s}$ (i.e., $\boldsymbol{J}^{(r)}[i,:] = \boldsymbol{b}_i^{(r)} \boldsymbol{\Phi}^H$). To embed the linear regression, we further assume that each time-frequency component is a linear combination of the p covariates; that is, the jth time-frequency component in the rth trial is $\boldsymbol{b}_i^{(r)}[j] = \sum_{k=1}^{p} \boldsymbol{X}[r,k]\boldsymbol{Z}[i,j,k]$, where $\boldsymbol{Z} \in \mathbb{C}^{m \times s \times p}$ is a complex tensor denoting all the linear regression coefficients, and $\boldsymbol{Z}[i,j,k]$ denotes the coefficient relating the jth time-frequency component of the source time series at the ith source point to the kth column of the regressors. Therefore, the STFT-R model reads

$$\underset{n \times T}{\boldsymbol{Y}^{(r)}} = \underset{n \times m}{\boldsymbol{G}} \left(\sum_{k=1}^{p} \boldsymbol{X}[r,k] \underset{m \times s}{\boldsymbol{Z}[:,:,k]} \right) \underset{s \times T}{\boldsymbol{\Phi}^{\boldsymbol{H}}} + \underset{n \times T}{\boldsymbol{E}^{(r)}} \qquad r = 1, \cdots, q.$$

We assume the sensor error $\boldsymbol{E}^{(r)}$ has independently and identically distributed (i.i.d.) Gaussian entries in each trial, with a zero mean and a constant variance. Also, it is worth noting that we assume the time-frequency components (i.e., the $\boldsymbol{b}_i^{(r)}[j]$s) can be perfectly explained by the covariates with no residuals. These assumptions (i.i.d. sensor noise and perfect regression) are made mainly for tractability, but they do not necessarily hold in real data. Such a model mismatch may cause additional errors. However, we may be able to alleviate such effects. Note that the residuals in the source space are essentially projected to the sensor space and can be absorbed in the $\boldsymbol{E}^{(r)}$ term. If the combined sensor and source noise in the $\boldsymbol{E}^{(r)}$ term can be assumed temporally independent (or "white"), we only need to consider the dependence across sensors. We can use the baseline sensor recordings (before the stimulus onset), which include both the sensor noise and the projection of the source-space noise, to estimate the noise covariance across sensors and pre-whiten the model; in this way, $\boldsymbol{E}^{(r)}$ in the new, whitened model will have i.i.d. entries. If there is strong temporal dependence in the combined sensor and source noise, the model mismatch problem may have a larger effect. This is a general issue for a lot of source localization methods, which we will discuss later in Chapter 7.

To solve for the regression coefficients $\boldsymbol{Z}$, we minimize the sum of squared prediction errors across the q trials, with a hierarchical penalty Ω on $\boldsymbol{Z}$:

$$\min_{\boldsymbol{Z}} \left(\sum_{r=1}^{q} \frac{1}{2} \| \boldsymbol{Y}^{(r)} - \boldsymbol{G}(\sum_{k=1}^{p} \boldsymbol{X}[r,k]\boldsymbol{Z}[:,:,k]) \boldsymbol{\Phi}^{\boldsymbol{H}} \|_F^2 + \Omega(\boldsymbol{Z}) \right) \tag{3.5}$$

where $\|\cdot\|_F$ is the Frobenius norm and

$$\Omega(\boldsymbol{Z}) = \alpha \sum_{l} w_l \sqrt{\sum_{i \in \mathcal{A}_l} \sum_{j=1}^{s} \sum_{k=1}^{p} |\boldsymbol{Z}[i,j,k]|^2} \tag{3.6}$$

$$+ \beta \sum_{i=1}^{m} \sum_{j=1}^{s} \sqrt{\sum_{k=1}^{p} |\boldsymbol{Z}[i,j,k]|^2} \tag{3.7}$$

$$+ \gamma \sum_{i=1}^{m} \sum_{j=1}^{s} \sum_{k=1}^{p} \sqrt{|\boldsymbol{Z}[i,j,k]|^2}. \tag{3.8}$$

where $|\cdot|$ denote the absolute value of a complex number. The penalty $\Omega(\boldsymbol{Z})$ involves three terms corresponding to the L_2 norms of entries in $\boldsymbol{Z}$ in three levels of nested groups; each of the group includes the entries of $\boldsymbol{Z}$ under a square root symbol, and α, β and γ are tuning parameters for each level. Because the L_2 norm of a group is not differentiable when all entries in the group is zero, this penalty can create nested non-zero patterns in the solution—if a larger group is completely zero, all subgroups are zero, too. Figure 3.1 illustrates the three levels of groups on the tensor $\boldsymbol{Z}$, and a possible non-zero pattern.

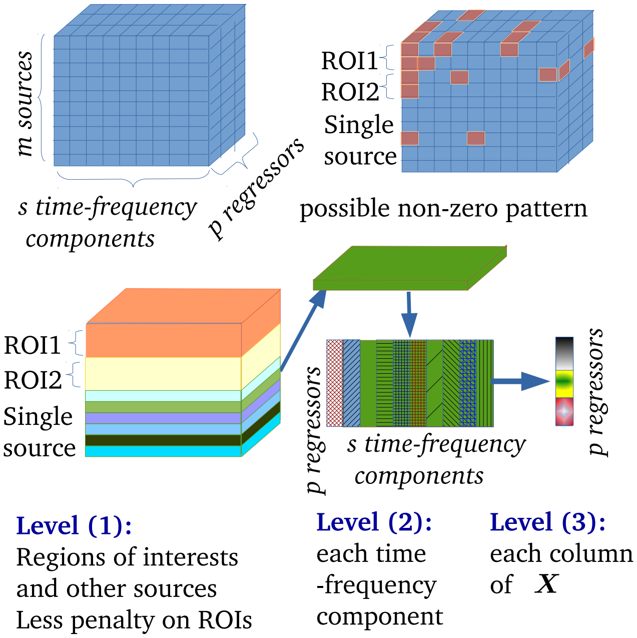

m sources
s time-frequency components
p regressors
ROI1
ROI2
Single source
possible non-zero pattern
ROI1
ROI2
Single source
p regressors
s time-frequency components
p regressors
Level (1):
Regions of interests and other sources
Less penalty on ROIs
Level (2):
each time -frequency component
Level (3):
each column of $\boldsymbol{X}$

Figure 3.1: Illustration of the three levels in hierarchical penalty $\Omega(\boldsymbol{Z})$

Suppose we are given several pre-defined regions of interest (ROIs). On the first level in (3.6), each group under the square root symbol either consists of coefficients for all source points within one ROI, or coefficients for one single source point outside the ROIs. Counting them together, we have N_α groups, denoted by $\mathcal{A}_l, l = 1, \cdots, N_\alpha$, where N_α is the number of ROIs plus the number of source points outside any ROI. Such a structure encourages the source signals outside the ROIs to be spatially sparse. By specifying the weights (w_ls) for the N_α groups, we can also make the penalty on the coefficients for source points within the ROIs smaller than that on the coefficients for source points outside the ROIs. If no ROI is pre-defined, coefficients corresponding to each source point form a single group at this level ($N_\alpha = m$) and we expect to see spatially sparse source points to show non-zero coefficients in the results. On the second level, for each source point i, the term (3.7) groups the p regression coefficients for the jth time-frequency component under the square root symbol, inducing sparsity in the time-frequency domain. Finally, on the third level, (3.8) adds a penalty on the absolute value of each Z_{ijk} to encourage sparsity on the p covariates individually, for each time-frequency component of each source point.

3.2.3 Solving the STFT-R

3.2.3.1 The FISTA algorithm

As in the work by Gramfort et al. [2013], we also use the fast iterative shrinkage-thresholding algorithm (FISTA, Beck and Teboulle [2009]) to solve (3.5). Let $\boldsymbol{z}$ be a vector obtained by concatenating all entries in $\boldsymbol{Z}$, and let $\boldsymbol{z}_1$ be a vector of the same size as $\boldsymbol{z}$. We also reorganize all the nested groups in the penalty $\Omega(\boldsymbol{Z}) = \Omega(\boldsymbol{z})$ into an ordered list, $\{g_1, g_2, \cdots, g_N\}$, where each g_h ($h = 1, 2, \cdots, N$) denotes the set entries in one group, assuming there are N groups in total. This ordered list is obtained by listing all the third level groups first, then the second level groups and finally the first level groups, such that given h_1 and $h_2 \in \{1, 2, 3, \cdots, N\}$, if $h_1 < h_2$, then $g_{h_1} \subset g_{h_2}$ or $g_{h_1} \cap g_{h_2} = \emptyset$. We further denote the penalization parameter on each group g_h by λ_h, for $h = 1, 2, \cdots, N$. For example, $\lambda_h = \alpha w_l$ if g_h is the lth group on the first level, $\lambda_h = \beta$ if g_h is on the second level, and $\lambda_h = \gamma$ if g_h is on the third level. Let $\boldsymbol{z}|_{g_h}$ be the elements of $\boldsymbol{z}$ in the group g_h. In this context, we have $\Omega(\boldsymbol{z}) = \sum_{h=1}^{N} \lambda_h \|\boldsymbol{z}|_{g_h}\|_2$. In each iteration of the FISTA algorithm, we need the proximal operator associated with the hierarchical penalty:

$$\text{Prox}(\boldsymbol{z}_1) = \arg\min_{\boldsymbol{z}} (\frac{1}{2}\|\boldsymbol{z} - \boldsymbol{z}_1\|^2 + \Omega(\boldsymbol{z})) = \arg\min_{\boldsymbol{z}} (\frac{1}{2}\|\boldsymbol{z} - \boldsymbol{z}_1\|^2 + \sum_{h=1}^{N} \lambda_h \|\boldsymbol{z}|_{g_h}\|_2) \tag{3.9}$$

As proved in Jenatton et al. [2011], (3.9) can be solved by composing the proximal operators for the L_2 norm penalty on each g_h, following the order in the list; that is, initialize $\boldsymbol{z} \leftarrow \boldsymbol{z}_1$, for $h = 1, \cdots N$ in the ordered list,

$$\boldsymbol{z}|_{g_h} \leftarrow \begin{cases} \boldsymbol{z}|_{g_h}(1 - \lambda_h / \|\boldsymbol{z}|_{g_h}\|_2) & \text{if } \|\boldsymbol{z}|_{g_h}\|_2 > \lambda_h \\ 0 & \text{otherwise} \end{cases}$$

If we define the sum-of-squares term in Equation (3.5) as

$$\mathfrak{f}(\boldsymbol{z}) = \frac{1}{2}\sum_{r=1}^{q} \|\boldsymbol{Y}^{(r)} - \boldsymbol{G}\left(\sum_{k=1}^{p}\boldsymbol{X}[r,k]\boldsymbol{Z}[:,:,k]\right)\boldsymbol{\Phi^H}\|_F^2$$

the gradient of $\mathfrak{f}(\boldsymbol{z})$ can be computed as

$$\frac{\partial \mathfrak{f}}{\partial \boldsymbol{Z}[:,:,k]} = -\boldsymbol{G}^T\sum_{r=1}^{q} \boldsymbol{X}[r,k]\boldsymbol{Y}^{(r)}\boldsymbol{\Phi} + \boldsymbol{G}^T\boldsymbol{G}(\sum_{r=1}^{q} \boldsymbol{X}[r,k]\sum_{k_1=1}^{p} \boldsymbol{Z}[:,:,k_1]\boldsymbol{X}[r,k_1])\boldsymbol{\Phi}^H\boldsymbol{\Phi}.$$

Let $\boldsymbol{z}_0$ and $\boldsymbol{z}_1$ be auxiliary variables of the same shape as $\boldsymbol{z}$. Given a positive number $L > 0$, which determines the step size, we define the proximal operator with respect to L and $\boldsymbol{z}_1 - \frac{1}{L}\nabla\mathfrak{f}(\boldsymbol{z}_1)$ as

$$\text{Prox}_L(\boldsymbol{z}_1) = \arg\min_{\boldsymbol{z}_0}(\frac{1}{2}\|\boldsymbol{z}_0 - (\boldsymbol{z}_1 - \frac{1}{L}\nabla\mathfrak{f}(\boldsymbol{z}_1))\|^2 + \frac{1}{L}\Omega(\boldsymbol{z}_0))$$

If we fix the step size, L is set to the Lipschitz constant of the gradient $\nabla\mathfrak{f}$, computed with the power iteration method used in Gramfort et al. [2013]. In this case, the FISTA algorithm is described in Algorithm 1, where ζ and ζ_0 are constants for accelerating convergence, and $\leftarrow$ denotes the operation of assigning the value on the right to the left variable.

Algorithm 1: The FISTA algorithm given the Lipschitz constant L

```
Data: $L$, $\mathfrak{f}(\boldsymbol{z})$, $\Omega(\boldsymbol{z}) = \Omega(\boldsymbol{Z})$,
$\boldsymbol{z}_{ini}$: initial value of $\boldsymbol{z}$
Result: the optimal solution $\boldsymbol{z}$
initialization: $\boldsymbol{z}_0 \leftarrow \boldsymbol{z}_{ini}$; $\boldsymbol{z}_1 \leftarrow \boldsymbol{z}_0$; $\boldsymbol{z} \leftarrow \boldsymbol{z}_0$; $\zeta \leftarrow 1$; $\zeta_0 \leftarrow 1$;
while the difference of $\boldsymbol{z}$ in two iterations is larger than a pre-defined threshold do
    $\boldsymbol{z}_0 \leftarrow \boldsymbol{z}$ ;
    Compute $\nabla\mathfrak{f}(\boldsymbol{z}_1)$;
    Apply the proximal operator $\boldsymbol{z} \leftarrow \text{Prox}_L(\boldsymbol{z}_1)$;
    $\zeta_0 \leftarrow \zeta$;
    $\zeta \leftarrow \frac{1+\sqrt{4\zeta_0^2+1}}{2}$;
    $\boldsymbol{z}_1 \leftarrow \boldsymbol{z} + \frac{\zeta_0 - 1}{\zeta}(\boldsymbol{z} - \boldsymbol{z_0})$;
end
```

If we do not fix the step size, we can also use back-tracking [Beck and Teboulle, 2009] without computing the Lipschitz constant (see Algorithm 2).

Algorithm 2: The FISTA algorithm with back-tracking

```
**Data**: constant $\eta > 1$, $\mathfrak{f}(\boldsymbol{z})$, $\Omega(\boldsymbol{z}) = \Omega(\boldsymbol{Z})$
$L_0$:initial value of $L$ ,
$\boldsymbol{z}_{ini}$: initial value of $\boldsymbol{z}$
**Result**: the optimal solution $\boldsymbol{z}$
initialization: $\boldsymbol{z_0} \leftarrow \boldsymbol{z}_{ini}$; $\boldsymbol{z}_1 \leftarrow \boldsymbol{z_0}$; $\boldsymbol{z} \leftarrow \boldsymbol{z_0}$; $\zeta \leftarrow 1$; $\zeta_0 \leftarrow 1$; $L \leftarrow L_0$;
**while** *the difference of $\boldsymbol{z}$ in two iterations is larger than a pre-defined threshold* **do**
    $\boldsymbol{z_0} \leftarrow \boldsymbol{z}$;
    Compute $\nabla\mathfrak{f}(\boldsymbol{z}_1)$ and $\mathfrak{f}(\boldsymbol{z}_1)$;
    $\boldsymbol{z} \leftarrow \mathrm{Prox}_L(\boldsymbol{z}_1)$;
    Compute diff $\leftarrow \mathfrak{f}(\boldsymbol{z}) - \mathfrak{f}(\boldsymbol{z}_1) - \nabla\mathfrak{f}(\boldsymbol{z_1})'(\boldsymbol{z} - \boldsymbol{z}_1) - \frac{1}{2}L\|\boldsymbol{z} - \boldsymbol{z}_1\|^2$;
    **while** diff $> 0$ **do**
        $L \leftarrow \eta L$ ;
        $\boldsymbol{z} \leftarrow \mathrm{Prox}_L(\boldsymbol{z}_1)$;
        diff $\leftarrow \mathfrak{f}(\boldsymbol{z}) - \mathfrak{f}(\boldsymbol{z}_1) - \nabla\mathfrak{f}(\boldsymbol{z}_1)'(\boldsymbol{z} - \boldsymbol{z}_1) - \frac{1}{2}L\|\boldsymbol{z} - \boldsymbol{z}_1\|^2$
    **end**
    $\zeta_0 \leftarrow \zeta$ ;
    $\zeta \leftarrow \frac{1+\sqrt{4\zeta_0^2+1}}{2}$;
    $\boldsymbol{z}_1 \leftarrow \boldsymbol{z} + \frac{\zeta_0 - 1}{\zeta}(\boldsymbol{z} - \boldsymbol{z_0})$;
**end**
```

3.2.3.2 Active-set strategy

In practice, it may be time-consuming to solve the original problem in (3.5). Thus we derive a greedy active-set strategy (Algorithm 3), based on the one introduced by Bach et al. [2011]. Consider a subset $\mathcal{B}$ of the N_α groups on the first level ($\mathcal{B} \subset \{1, \cdots, N_\alpha\}$). We start with a union of source points in the groups in $\mathcal{B}$, denoted as $\mathcal{J} = \cup_{l \in \mathcal{B}} \mathcal{A}_l$, and then we compute the solution to a sub-problem that is constrained on $\mathcal{J}$ (i.e., the sub-problem contains only the source points in $\mathcal{J}$ and the columns of $\boldsymbol{G}$ corresponding to $\mathcal{J}$). Next, we examine whether it is optimal for the original problem, which contains all source points, by checking whether the Karush-Kuhn-Tucker (KKT) conditions are met. If yes (within some tolerance), we accept the solution; otherwise, we greedily add K groups to $\mathcal{J}$ and repeat the procedure (see Algorithm 3).

Algorithm 3: Active-set strategy

```
initialization: choose initial $\mathcal{J}$ and initial value of $\boldsymbol{Z}$;
compute the violation of the KKT conditions for each $\mathcal{A}_l \notin \mathcal{J}$;
while the total violation of the KKT conditions is larger than a pre-defined threshold do
    add to $\mathcal{J}$ $K$ of the remaining groups that have the largest violation;
    compute a solution to the sub-problem constrained on $\mathcal{J}$ using FISTA;
    compute the violation of the KKT conditions for each $\mathcal{A}_l \not\subset \mathcal{J}$;
end
```

Below we derive the KKT conditions and describe how to quantify the violation of the KKT conditions. Since the term $\mathfrak{f}(\boldsymbol{z})$ is essentially a sum of squared errors of a linear problem, we can rewrite it as $\mathfrak{f}(\boldsymbol{z}) = \frac{1}{2}||\boldsymbol{c} - \boldsymbol{B}\boldsymbol{z}||^2$. Here $\boldsymbol{z}$ again is a vector obtained by concatenating all entries in $\boldsymbol{Z}$ and $\boldsymbol{c}$ is obtained by concatenating all entries in $\boldsymbol{Y}^{(1)}, \cdots, \boldsymbol{Y}^{(q)}$; $\boldsymbol{B}$ is a linear operator, such that the entries in $\boldsymbol{B}\boldsymbol{z}$ are equal to the concatenated entries in $\boldsymbol{G}(\sum_{k=1}^{p} \boldsymbol{X}[r,k]\boldsymbol{Z}[:,:,k])\boldsymbol{\Phi}^{\boldsymbol{H}}, r = 1, \cdots, q$. Note that although $\boldsymbol{z}$ and $\boldsymbol{B}$ have complex values, we can further rewrite the problem into a real-valued problem by rearranging the real and imaginary parts of $\boldsymbol{z}$ and $\boldsymbol{B}$. Without loss of generality, we derive the KKT conditions assuming $\boldsymbol{z}$, $\boldsymbol{c}$ and $\boldsymbol{B}$ are real-valued vectors and matrix. Again we use $\{g_1, \cdots, g_h, \cdots, g_N\}$ to denote our ordered group set, and we use λ_h to denote the corresponding penalty parameter on group g_h. We also define diagonal matrices $\boldsymbol{D}_h$ such that

$$\boldsymbol{D}_h[l,l] = \begin{cases} 1 & \text{if } l \in g_h \\ 0 & \text{otherwise} \end{cases} \quad \forall h$$

Therefore, the non-zero elements of $\boldsymbol{D}_h\boldsymbol{z}$ is equal to $\boldsymbol{z}|_{g_h}$. With the new notation, we recast the original problem into a standard formulation:

$$\min_{\boldsymbol{z}}(\frac{1}{2}\|\boldsymbol{c} - \boldsymbol{B}\boldsymbol{z}\|_2^2 + \sum_h \lambda_h \|\boldsymbol{D}_h\boldsymbol{z}\|_2) \tag{3.10}$$

To better describe the KKT conditions, we introduce some auxiliary variables: $\mathfrak{u} = \boldsymbol{B}\boldsymbol{z}$ and $\mathfrak{v}_h = \boldsymbol{D}_h\boldsymbol{z}$. Then (3.10) is equivalent to

$$\min_{\boldsymbol{z},\mathfrak{u},\mathfrak{v}_h\forall h}(\frac{1}{2}\|\boldsymbol{c} - \mathfrak{u}\|_2^2 + \sum_h \lambda_h \|\mathfrak{v}_h\|_2)$$
$$\text{subject to } \mathfrak{u} = \boldsymbol{B}\boldsymbol{z}, \quad \mathfrak{v}_h = \boldsymbol{D}_h\boldsymbol{z}, \quad h = 1, 2, \cdots, N.$$

The corresponding Lagrange function is

$$\text{Lagrange}(\boldsymbol{z}, \mathfrak{u}, \mathfrak{v}_h, \boldsymbol{\mu}, \boldsymbol{\xi}_h) = \frac{1}{2}\|\boldsymbol{c} - \mathfrak{u}\|_2^2 + \sum_h \lambda_h \|\mathfrak{v}_h\|_2 + \boldsymbol{\mu}'(\boldsymbol{B}\boldsymbol{z} - \mathfrak{u}) + \sum_h \boldsymbol{\xi}_h'(\boldsymbol{D}_h\boldsymbol{z} - \mathfrak{v}_h)$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\xi}_h$s are Lagrange multipliers, which have the same dimension as $\boldsymbol{z}$. At the optimum, the following KKT conditions hold

$$\frac{\partial}{\partial \boldsymbol{\mathfrak{u}}}\text{Lagrange} = \boldsymbol{\mathfrak{u}} - \boldsymbol{c} - \boldsymbol{\mu} = 0 \tag{3.11}$$

$$\frac{\partial}{\partial \boldsymbol{z}}\text{Lagrange} = \boldsymbol{B}'\boldsymbol{\mu} + \sum_h \boldsymbol{D}_h \boldsymbol{\xi}_h = 0 \tag{3.12}$$

$$\frac{\partial}{\partial \boldsymbol{\mathfrak{v}}_h}\text{Lagrange} = \lambda_h \partial \|\boldsymbol{\mathfrak{v}}_h\|_2 - \boldsymbol{\xi}_h \ni 0, \forall h \tag{3.13}$$

where $\partial \| \cdot \|_2$ is the subgradient of the L_2 norm. From (3.11) we have $\boldsymbol{\mu} = \boldsymbol{\mathfrak{u}} - \boldsymbol{c}$, then (3.12) becomes $\boldsymbol{B}'(\boldsymbol{\mathfrak{u}} - \boldsymbol{c}) + \sum_h \boldsymbol{D}_h \boldsymbol{\xi}_h = 0$. Plugging $\boldsymbol{\mathfrak{u}} = \boldsymbol{B}\boldsymbol{z}$ in, we can see that the first term $\boldsymbol{B}'(\boldsymbol{\mathfrak{u}} - \boldsymbol{c}) = \boldsymbol{B}'(\boldsymbol{B}\boldsymbol{z} - \boldsymbol{c})$ is the gradient of $\mathfrak{f}(\boldsymbol{z}) = \frac{1}{2}\|\boldsymbol{c} - \boldsymbol{B}\boldsymbol{z}\|_2^2$. For a solution $\boldsymbol{z}^\dagger$, once we plug in $\boldsymbol{\mathfrak{v}}_h = \boldsymbol{D}_h \boldsymbol{z}^\dagger$, the KKT conditions become

$$\nabla \mathfrak{f}(\boldsymbol{z})_{\boldsymbol{z}=\boldsymbol{z}^\dagger} + \sum_h \boldsymbol{D}_h \boldsymbol{\xi}_h = 0 \tag{3.14}$$

$$\lambda_h \partial \|\boldsymbol{D}_h \boldsymbol{z}^\dagger\|_2 - \boldsymbol{\xi}_h \ni 0, \forall h \tag{3.15}$$

In (3.15), we have the following according to the definition of subgradients

$$\boldsymbol{\xi}_h = \lambda_h \frac{\boldsymbol{D}_h \boldsymbol{z}^\dagger}{\|\boldsymbol{D}_h \boldsymbol{z}^\dagger\|_2} \text{ if } \|\boldsymbol{D}_h \boldsymbol{z}^\dagger\|_2 > 0$$
$$\|\boldsymbol{\xi}_h\|_2 \leq \lambda_h \text{ if } \|\boldsymbol{D}_h \boldsymbol{z}^\dagger\|_2 = 0$$

Therefore we can determine whether (3.14) and (3.15) hold by solving the following problem.

$$\min_{\boldsymbol{\xi}_h} \frac{1}{2}\|\nabla f(\boldsymbol{z})_{\boldsymbol{z}=\boldsymbol{z}^\dagger} + \sum_h \boldsymbol{D}_h \boldsymbol{\xi}_h\|_2^2$$
$$\text{subject to } \boldsymbol{\xi}_h = \lambda_h \frac{\boldsymbol{D}_h \boldsymbol{z}^\dagger}{\|\boldsymbol{D}_h \boldsymbol{z}^\dagger\|_2} \text{ if } \|\boldsymbol{D}_h \boldsymbol{z}^\dagger\|_2 > 0$$
$$\|\boldsymbol{\xi}_h\|_2 \leq \lambda_h \text{ if } \|\boldsymbol{D}_h \boldsymbol{z}^\dagger\|_2 = 0$$

which is a standard group lasso problem with no overlap. For h where $||\boldsymbol{D}_h \boldsymbol{z}^\dagger||_2 > 0$, we plug in $\boldsymbol{\xi}_h = \lambda_h \frac{\boldsymbol{D}_h \boldsymbol{z}^\dagger}{\|\boldsymbol{D}_h \boldsymbol{z}^\dagger\|_2}$, and only solve for $\boldsymbol{\xi}_h$ where $||\boldsymbol{D}_h \boldsymbol{z}^\dagger||_2 = 0$. We can use the coordinate descent algorithm by focusing on each $\boldsymbol{\xi}_h$ at each iteration [1]. We define $\frac{1}{2}\|\nabla \mathfrak{f}(\boldsymbol{z})_{\boldsymbol{z}=\boldsymbol{z}^\dagger} + \sum_h \boldsymbol{D}_h \boldsymbol{\xi}_h\|_2^2$ at the optimum as a measure of violation of the KKT conditions.

[1] It is worth noting that although $\boldsymbol{\xi}_h$ has the same length as $\boldsymbol{z}$, only the elements in the group g_h contribute to the problem, so we assume all the elements of $\boldsymbol{\xi}_h$ outside the group g_h are zero. For each h where $\|\boldsymbol{D}_h \boldsymbol{z}^\dagger\|_2 = 0$, solving for $\boldsymbol{\xi}_h$ is the same as solving

$$\min_{\boldsymbol{\xi}_h} ||\boldsymbol{c}_0 + \boldsymbol{D}_h \boldsymbol{\xi}_h||_2^2 \quad \text{such that} \quad ||\boldsymbol{\xi}_h||_2^2 \leq \lambda_h^2$$

where $\boldsymbol{c}_0$ includes the terms related to the gradient and related to other hs. Let the Lagrange multiplier of this particular problem be μ, we have the Lagrange function $||\boldsymbol{c}_0 + \boldsymbol{D}_h \boldsymbol{\xi}_h||_2^2 + \mu(\|\boldsymbol{\xi}_h\|_2^2 - \lambda_h^2)$. Taking the gradient of this Lagrange

Next, we consider how to add new groups to $\mathcal{J}$ greedily in the iteration of the active-set strategy. For $\mathcal{A}_l \not\subset \mathcal{J}$, we use the sum of squares in $(\nabla \mathfrak{f}(\boldsymbol{z})_{\boldsymbol{z}=\boldsymbol{z}^\dagger} + \sum_h \boldsymbol{D}_h \boldsymbol{\xi}_h)$ constrained to only the entries corresponding to $\mathcal{A}_l$ at the optimum as a measurement of violation and add the K groups that have the largest violation to $\mathcal{J}$ for the next iteration.

3.2.4 Additional L_2 regularization and bootstrapping

To obtain the standard deviations of the regression coefficients in $\boldsymbol{Z}$, we can use bootstrapping. However, regular bootstrapping does not work well for sparsity-inducing problems, due to the non-differentiability at zero. Therefore, in addition to the original solution with the hierarchical sparsity-inducing penalty ($\Omega(\boldsymbol{Z})$), we can also compute an additional L_2-regularized solution of $\boldsymbol{Z}$, where the zero entries of $\boldsymbol{Z}$ learned in the sparsity-inducing solution are constrained to be zero, and the penalty is the sum of squares of the non-zero entries in $\boldsymbol{Z}$. The tuning parameters in both the sparsity-inducing and the L_2-regularized cases can be selected by minimizing cross-validated prediction errors of the sensor data.

We use a data-splitting bootstrapping procedure. The data in all trials can be split into two halves; using the first half of the trials, we obtain the sparse solution in the STFT-R model, and using the second half of the trials, we compute an L_2-regularized solution constrained on the non-zero entries learned in the first half of the data. Next, we can do bootstrapping using the second half of the data either in a non-parametric way or parametric way. In the non-parametric version, we can sample the trials and the corresponding rows in $\boldsymbol{X}$ with replacement. In the parametric version, we can plug in the learned L_2-regularized estimate of $\boldsymbol{Z}$ to obtain residual sensor time series of each trial in the second half of the data ($\boldsymbol{R}^{(r)} = \boldsymbol{Y}^{(r)} - \boldsymbol{G}(\sum_{k=1}^p \boldsymbol{X}[r,k]\boldsymbol{Z}[:,:,k])\boldsymbol{\Phi}^{\boldsymbol{H}}$). We rescale the residuals in each trial by multiplying $1/(1-h_r)^{0.5}$ where $h_r = \boldsymbol{X}[r,:](\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}[r,:]'$ [Stine, 1985]. Then we add re-sampled residuals back to $\boldsymbol{G}(\sum_{k=1}^p \boldsymbol{X}[k,r]\boldsymbol{Z}[:,:,k])\boldsymbol{\Phi}^{\boldsymbol{H}}$ to obtain the bootstrapped samples of sensor data. For each bootstrapped sample, we estimate the L_2-regularized solution constrained on the non-zero entries again. For each bootstrapped sample, the best L_2 tuning parameter is again determined by cross-validation.

3.3 Results

In this section, we compare the STFT-R model and a two-step source-space regression method using the MNE. In this two-step method, we first obtain the source-space estimates using the MNE

function, we need $\boldsymbol{\xi}_h$ to satisfy

$$\boldsymbol{\xi}_h|_{g_h} = (\boldsymbol{D}_{\boldsymbol{h}}^T \boldsymbol{D}_{\boldsymbol{h}}(1+\mu))^{-1}(-\boldsymbol{D}_{\boldsymbol{h}}^T \boldsymbol{c}_0) = (-\boldsymbol{D}_{\boldsymbol{h}}\boldsymbol{c}_0)/(1+\mu)$$

if $\|(-\boldsymbol{D}_{\boldsymbol{h}}\boldsymbol{c})\|_2 < \lambda_h$, then $\mu = 0$, else, $\|(-\boldsymbol{D}_{\boldsymbol{h}}\boldsymbol{c})/(1+\mu)\|_2 = \lambda_h$; in other words, we shrink the $-\boldsymbol{D}_{\boldsymbol{h}}^T\boldsymbol{b}$ to satisfy the constraint.

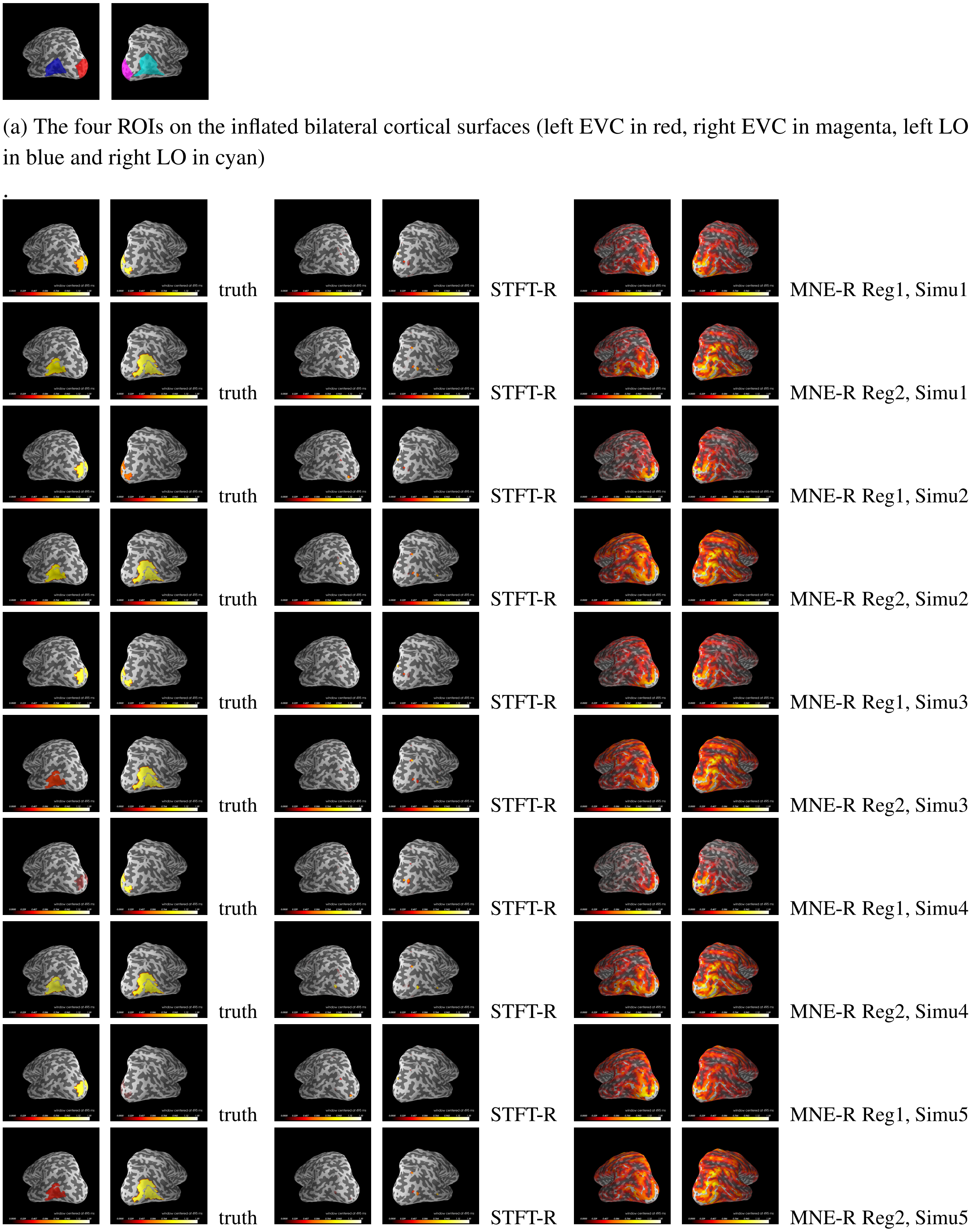
and apply the STFT on the estimated activity (i.e., the estimated time series) for each source point and each trial; in the second step, we regress the STFT components across trials on the external covariates (or regressors) to obtain the regression coefficients. Below, we denote this two-step method by MNE-R. We applied both the STFT-R and the MNE-R and compare their performance in learning the regression coefficients on simulated data and real data.

3.3.1 Simulations

The simulations were implemented using the `MNE-python` package in `Python` [Gramfort et al., 2014]. This package provided a sample MEG dataset with $n = 306$ sensors, along with the source space of a real human participant, which consisted of $m = 7498$ source points that were perpendicular to the local cortical surface. The $n \times m$ forward matrix was also pre-computed in the data set. Using the source space and the forward matrix, we simulated a simplified situation of our application in studying the visual cortex in Chapter 5. We used four regions of interest (ROIs): the left and right pericalcarine areas, which were in the posterior end of the brain and overlapped with the early visual cortex, and two regions in the left and right lateral occipital areas (LO), which were at a higher level than the early visual cortex in the posterior-to-anterior hierarchy of the visual cortex, according to a well established hypothesis of the hierarchical organization of the visual cortex [DiCarlo and Cox, 2007] (see Chapter 5 for details). Figure 3.2a shows the locations of the four regions on the inflated bilateral cortical surfaces in lateral views (left EVC in red, right EVC in magenta, left LO in blue and right LO in cyan).

We used two regressors, each of which independently followed a standard normal distribution. At each time point, the activity in a source point within an ROI was generated as a linear combination of the two regressors plus noise. The source points within the same ROI shared the same regression coefficients. With $T = 100$ time points and a time step of 10 ms, the regression coefficients for each regressor at all time points composed a time series; these time series of regression coefficients were generated by applying an inverse STFT on two time-frequency components. The real parts of the time-frequency components were random samples from a standard normal distribution multiplied by 5, whereas the imaginary parts were 0. The EVC only had non-zero regression coefficients for Regressor 1, concentrated within 300 to 500 ms, and the LO only had non-zero regression coefficients for Regressor 2, concentrated within 500 to 700 ms, as shown in the first columns of Figure 3.3, which plot the true time series of regression coefficients in each ROI and for each regressor. The regression coefficients for source points outside of any ROIs were set to zero.

For each source point in each trial, besides the time series obtained by linearly combining the regressors at this trial using the time series of the regression coefficients, we also added an independently sampled noise time series—a T-dimensional Gaussian variable with a zero mean and a covariance matrix with a radial basis function kernel. The marginal variance of such noise was set to 0.1. The unit of the source-space activity was nanoampere meter. Finally, the 306-dimensional sensor noise (independent at each time point) was added, using a sensor noise covariance matrix



(a) The four ROIs on the inflated bilateral cortical surfaces (left EVC in red, right EVC in magenta, left LO in blue and right LO in cyan)

.

(b) The spatial patterns of the magnitudes of the true regression coefficients (the first two columns), and of the estimates by the STFT-R model (the middle two columns) and by the MNE-R model (the last two columns). Each row corresponds to one regressor (Reg1 or Reg2) in one simulation (Simu1, Simu2, etc)

Figure 3.2: The ROIs and the spatial patterns of the magnitudes of the regression coefficients

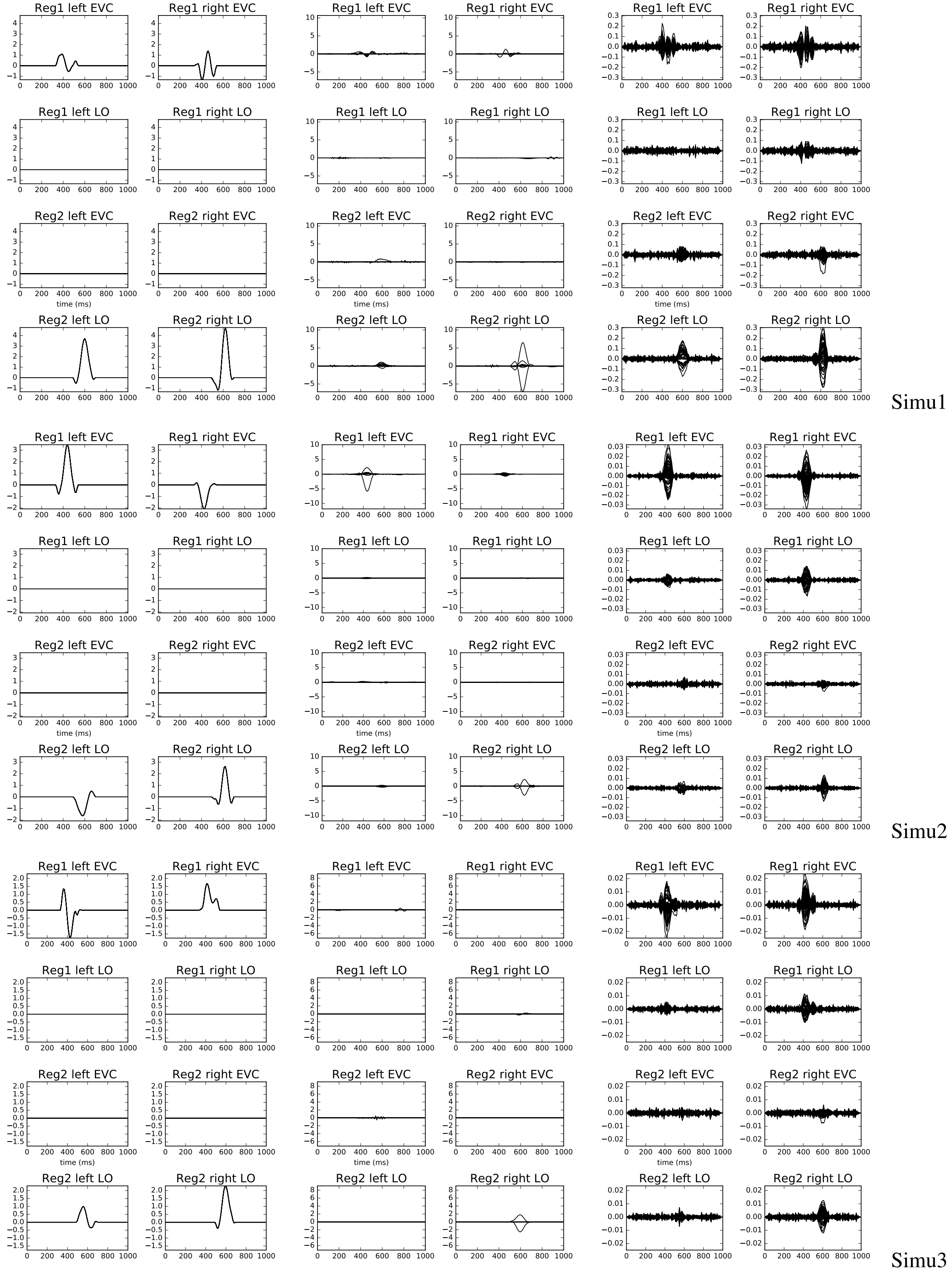


Figure 3.3: Time series of the regression coefficients within each ROI. The three columns from left to right show the time series of the true regression coefficients and the estimates by the STFT-R and the MNE-R respectively. In each plot, each curve corresponded to one source point.

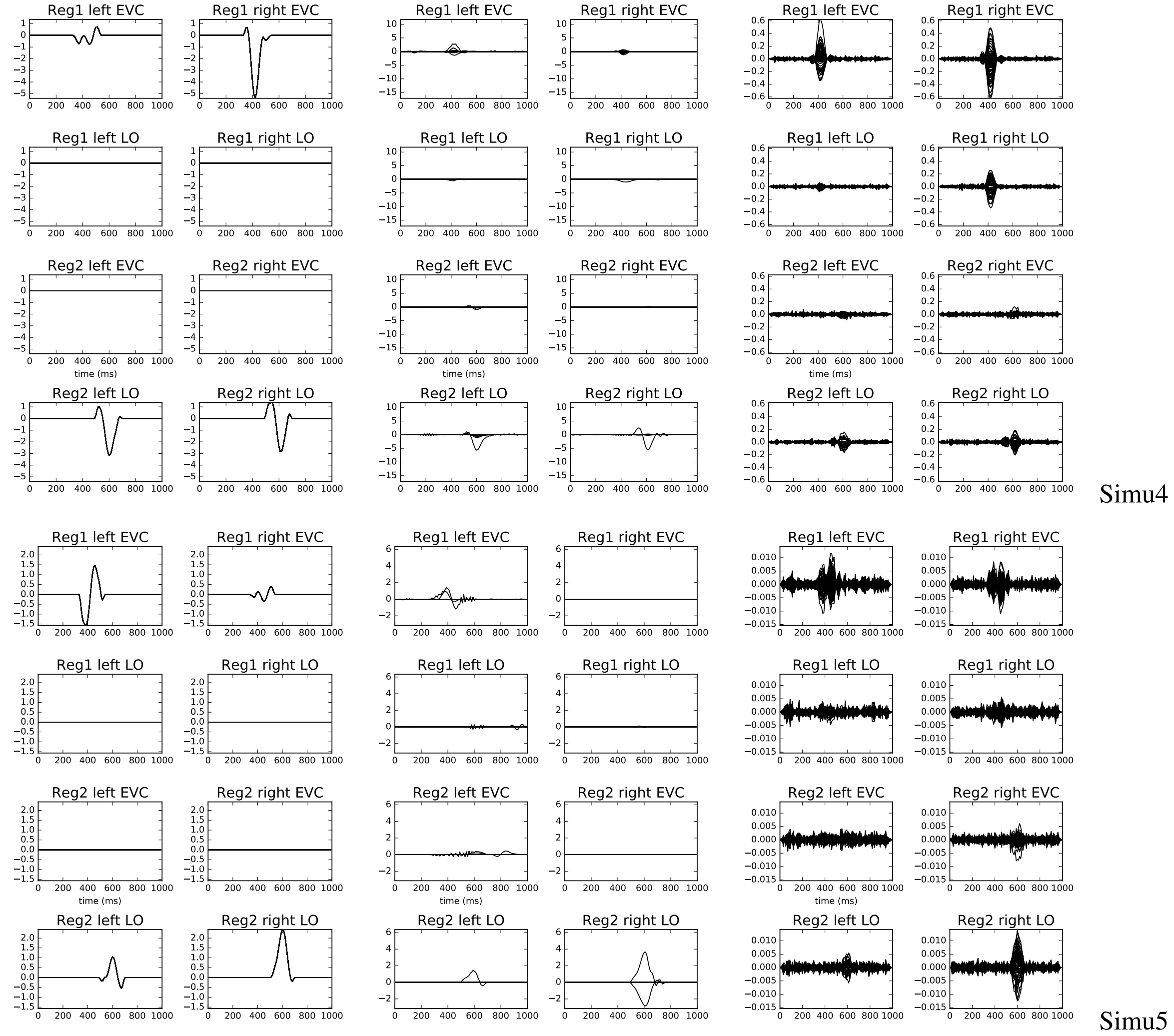


Figure 3.3: Time series of the regression coefficients within each ROI(continued)

from the sample dataset. We simulated $q = 50$ trials and applied the STFT-R and the MNE-R models on these data to learn the regression coefficients. For the STFT-R, on the first level, the regression coefficients in the time-frequency domain at each source point composed a single group, and the weights of all groups at this level were equal. A wide range of tuning parameters were provided for selection via cross-validation in both models. After learning the regression coefficients in the time-frequency domain, we transformed them back to the time domain. We did not compute further L_2 regularized regression coefficients after applying the STFT-R model. No depth compensation was applied in both models. The window size was 160 ms and the step size was $\tau_0 = 40$ ms for the STFT.

We ran five independent simulations. We show the spatial patterns of the regression coefficients in Figure 3.2b and the time series of estimated regression coefficients for the source points within each ROI in Figure 3.3. To visualize the spatial patterns of the magnitudes of regression coefficients for the truth and the estimates by the two models, we took the sum of the absolute values of the regression coefficients across all the T time points for each source point and each regressor, obtaining an $m \times 1$ map for each regressor. We divided the m values by their maximum to rescale the range of the map to 0 to 1. Figure 3.2b shows the results for each regressor in the 5 simulations, on inflated bilateral cortical surfaces in lateral views, where the first two columns show the patterns from the underlying truth, and the second two columns show the patterns by the STFT-R model and the last two columns show the patterns by the MNE-R model. The STFT-R model yielded sparse patterns. Compared with the true patterns, we can see that, in some cases, the identified source points with non-zero coefficients were within the true ROIs (e.g., the identified source points within the right LO for Regressor 2 in Simulation 1); however, there were also cases where the identified source points were outside of (although not far from) the true ROIs (e.g., the identified source points near the left LO for Regressor 2 in Simulation 1). In contrast, the patterns by the MNE-R covered a large area of the visual cortex including the ROIs; from the patterns for Regressor 1 to the patterns for Regressor 2, in the majority of the simulations, there appeared to be a shift from the posterior end to the lateral sides, corresponding to the relative locations of the EVCs and the LOs.

Figure 3.3 shows the time series of the regression coefficients in the 5 simulations. The three columns from left to right show the time series of the true regression coefficients and the estimates by the STFT-R and the MNE-R respectively. Each subplot corresponds to one ROI and one regressor; each curve corresponds to one source point. Note that in the first column, all source points within an ROI shared the same regression coefficients, so the curves merged into one curve. In the second column, the non-zero curves correspond to the sparse source points with non-zero regression coefficients. Note that each source point represents a current dipole perpendicular to the local cortical surface; the positive and negative signs only indicate the directions of the current dipoles. Due to the folding of the cortical surface, the source points within one ROI could have opposite orientations and the contribution to sensors from source points that have opposite directions may cancel. As a result, the reconstructed source activity (or the corresponding regression coefficients) by the STFT-R or the MNE-R could have opposite signs to the truth, as seen in the second and the third columns of Figure 3.3. Comparing the results by the STFT-R and the truth, we can see that the high-frequency components that did not explain the sensor data well (i.e., noise) were attenuated. In addition, during the periods in which we would expect the regression coefficients to be zero, the STFT-R method typically recovered zero coefficients. In contrast, the results by the MNE-R method were generally "spiky". In this sense, the results by the STFT-R are better in the sense that they look smooth, concentrated and "clean".

It is worth noting that in the results by both methods, the regression coefficients in one ROI "leaked" into another ROI or to source points outside of the ROIs. For example, the STFT-R model failed to recover the non-zero regression coefficients in the EVC in Simulation 3, because it falsely localized the regression effects to source points that were not in the ROIs. In another example, the MNE-R results in the right LO for Regressor 1 was not zero but very similar to those in the right EVC. This "spatial leak" can also be seen in the spatial patterns in Figure 3.2b—in the results by the MNE-R, the recovered regression effects dispersed widely; in the results by the STFT-R, although the patterns were sparse, there were also such leaks. Such effects are likely to be caused by the spatial correlations between the columns in the forward matrix. In this sense, there may be a fundamental limit of spatial resolution that affects both the STFT-R and the MNE-R models.

3.3.2 An application on real data: neural correlates of face-learning

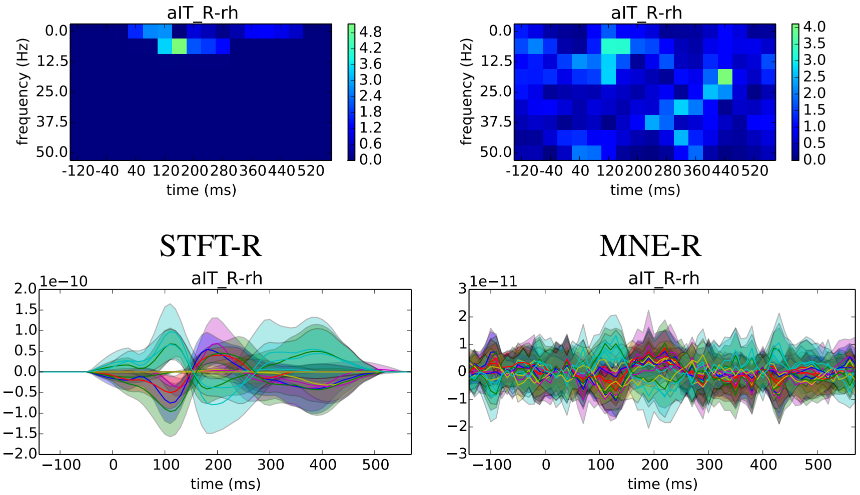
We applied the STFT-R and the MNE-R methods on MEG data from a face-learning experiment [Xu, 2013], to look for the spatio-temporal neural correlates of face-learning. In this experiment, the participants learned to distinguish two categories of computer-generated faces, which varied in their eye-size and mouth-width. In each trial, one face exemplar from one category was presented, and the participants were instructed to report the category label. After their response, feedback was provided; in this way, the participants learned the diagnostic features to categorize the faces in an online manner. For each participant, 364 distinct exemplars in each of the two categories were used, each presented once, resulting in 728 trials. In addition, an independent MEG experiment was run to localize several face-sensitive regions, which are involved in face processing according to the literature [Ishai, 2008; Pyles et al., 2013; Nestor et al., 2008, 2011]. These regions were used as the ROIs that we emphasized in learning the STFT-R model, and we also focused on the dependence between neural activity and behavioral learning within these regions, estimated by both methods. Nine participants showed significant learning effects—the behavioral accuracy went from 50% (i.e., chance) to at least 70%. We obtained the behavioral learning curves (as a function of trial index) by modeling the behavioral responses (i.e., "correct" or not) using logistic regression, for each face category and each participant.

We applied the STFT-R and the MNE-R to regress the spatio-temporal neural activity on the behavioral learning curve for each face category and each participant (the window size of STFT was 160 ms and the step size was $\tau_0 = 40$ ms); the regression coefficients (i.e., the "slope" coefficients) represented the linear dependence of neural activity on the behavioral learning curve and thus were of our main interest. In the regression, an additional all-one column was used in the regressors; the regression coefficient corresponding to this column represented the "intercept" or mean activity across the learning session. In the first level of grouping in the STFT-R, the source points within each ROI corresponded to one group, and the source points outside of the ROIs each corresponded to an individual group. The first-level penalties on the groups corresponding to the face-sensitive ROIs were set to zero, whereas the penalties on all other single-source-point groups were α. We provided a wide range of candidate tuning parameters (α, β and γ for the STFT-R and λ for the MNE-R) and selected the best ones via cross-validation.

Figure 3.4 shows the regression results in one ROI for one participant. The results were estimated on trials for one face category. The upper row shows the linear dependence of neural activity on the learning curve in the time-frequency domain, quantified by the sum of squared regression coefficients across their real and imaginary parts and across the source points within the ROI, divided by the bootstrapped standard deviation of the sum. The STFT-R pattern was sparse and concentrated on the lower frequency, whereas the MNE-R pattern was dispersed. After learning the regression coefficients in the time-frequency domain, we also transformed them back to the temporal domain using an inverse STFT. These time series of slope coefficients are shown in the lower row of Figure 3.4. Each curve with one color represents one source point [2]. The shaded bands are 95% confidence intervals at each time point, where an asymptotic normal distribution was assumed without further corrections for multiple comparisons at multiple time points. The time series of the slope coefficients by the STFT-R were smooth due to the sparsity in the time-frequency domain. Moreover, we could see the clear peaks, concentrated on the time windows near 100 ms, 200 ms and in 300 to 500 ms. In contrast, the results by the MNE-R did not demonstrate such "clean" patterns, due to the spiky noise at all time points.

For each participant, we ran a permutation test to examine whether the sum of squared slope coefficients within a face-sensitive ROI at each time point was significantly different from zero. The p-values of the permutation tests for trials in the two face categories were combined using Fisher's method [Fisher, 1925]. We use the negative logarithm of the p-values with base 10 to quantify the significance of correlation between the neural activity and the behavioral learning curve. The results by the STFT-R and the MNE-R were qualitatively consistent, but they did show some difference. In Figure 3.5, we show the $-\log_{10}(p\text{-value})$s in two ROIs where the difference were easily seen. Each row represents one participant; a dark blue row indicates that the ROI was not identified in the participant. In both the ROIs listed, the results by STFT-R were smooth; the patterns indicated a window near 200 ms where the activities in the ROIs were correlated with behavioral learning in most participants. In contrast, the results by the MNE-R were less smooth and the temporal correlation patterns appeared more noisy.

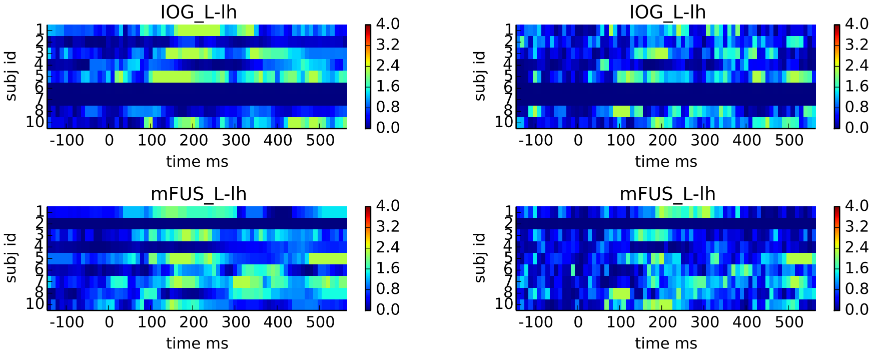
[2] However, the colors in the results by the two methods did not necessarily match.



STFT-R MNE-R

STFT-R MNE-R

Figure 3.4: Regression against the behavioral learning curve in a face-sensitive ROI in the right anterior temporal lobe (aIT) for one example participant. Upper: the summarized regression effects in the time-frequency domain. Lower: the time series of regression coefficients in each source point in the ROI.



STFT-R MNE-R

Figure 3.5: Regression of the source data in the ROIs: $-\log_{10}(p\text{-value})$ of the permutation tests, combined across two categories using Fisher's method in two example ROIs (left inferior occipital gyrus (IOG) and left mid-fusiform gyrus (mFUS)). Each row represents one participant; a dark blue row indicates that the ROI was not identified in the participant.

3.4 Discussion

In this chapter, we introduced a one-step source-space regression model, which represented the source current dipoles as sparse time-frequency components that were linearly dependent on external covariates. As a result of the time-frequency sparsity in the STFT-R, when the regression coefficients were transformed back to the time domain, the high-frequency components that did not explain the sensor data well (i.e., noise) were attenuated. In addition, during the periods in which we would expect the regression coefficients to be zero (e.g., the baseline time window before the stimulus onset), the method typically recovered zero coefficients. The resulting waveforms were generally smooth, concentrated in the time windows of interest, and thus "clean" and easy to interpret. In contrast, it is difficult to impose such sparsity-inducing constraints on the regression coefficients if we use a two-step approach, because, during the source localization step for each trial, it is hard to guarantee that the non-zero patterns in different trials are the same. If we use a constraint to force the non-zero patterns to be the same, biases due to this constraint may affect the resulting regression coefficients in the second step, and moreover, such a model is already close to a one-step method where we directly constrain the regression coefficients.

However, there are some limitations of the STFT-R model. First, we assume that the source-space neural activity has a perfect linear relationship with the external covariates (or regressors) without modeling additional residuals in the source space. In practice, there may be a model mismatch, where this assumption does not hold. However, as mentioned above, if such residuals exist, they are projected to the sensor space and can be absorbed in the sensor error term. We can alleviate this model mismatch problem in the following way—first we estimate the covariance structure of the errors in the sensor space in the baseline time window, which includes the original sensor errors and the projected source-space errors, and secondly, we use this covariance structure for pre-whitening. As a result, the errors in the sensor space in the pre-whitened data can be viewed as roughly i.i.d.. In addition, in most cases, we obtain the final scientific conclusions via statistical tests such as a permutation test. The model-mismatch problem is also present in the permuted results (or the control cases) as in the intact results; therefore our scientific conclusions are still likely to be correct, despite some possible loss of statistical power due to the model mismatch. Secondly, we assumed the sensor errors to be temporally independent; this assumption may also be violated in real data. However, this model mismatch is a general issue in many existing source localization methods, which also assume temporal independence in errors. We defer the related discussion in Chapter 7. Nevertheless, similarly to the first issue, such a model mismatch is unlikely to yield large errors in the final statistical conclusions, because it affects the permuted counterparts (or the control condition) in a similar manner. Thirdly, in the STFT-R, we use a univariate regression framework, where the activity at each source point is regressed on the external covariates. This univariate regression framework is limited in the sense that joint relationship among source points is not modeled. In future work, it will be intriguing to develop one-step models that relate the joint activity of source points in each local brain region with the external covariates.

Finally, we discuss a fundamental issue in both the STFT-R model and the MNE-R model—the limitation of spatial resolution due to the underdetermined nature of the source localization problem and the forward matrix. In simulations, we observed that sometimes there were "leaked" regression effects in the estimates by the MNE-R, which were not in the ROIs where the true regression coefficients were in. For the STFT-R, because of the spatial sparsity constraints, only a few source points showed the regression effects, and they could be outside of the true ROIs as well. Such spatial localization errors rise from the nature of the forward matrix, in which different columns corresponding to different locations are possibly correlated; the limited number of sensors also adds additional difficulty in spatially localizing the true sources. In other words, there may be some fundamental limits in recovering the spatial locations of source activity (or regression effects), and without additional information such as spatial priors, both the one-step and two-step models can suffer from these limits. In some domains where the researchers have reliable knowledge about the source locations, adding spatial priors may vastly improve the localization; in the STFT-R, we have built in the option to emphasize certain ROIs, which allows incorporating spatial priors, yet, it is restricted in the sense that we have to use ROIs with hard boundaries. In future one-step methods, we can further improve the specification of spatial priors to allow more flexibility.

Chapter 4

One-step cross-region functional connectivity analysis in the MEG/EEG source space

4.1 Motivation

Neurons in different locations of the brain are often anatomically connected, and the joint dynamic activities across brain regions are believed to underlie various perceptual and cognitive functions. Recently, the neuroimaging community has focused on *functional connectivity* [Friston, 2011], which describes the spatio-temporal statistical dependence across brain regions. Unlike anatomical connectivity, which describes physical connections through axons and dendrites of neurons, *functional connectivity* is more of a statistical model, obtained from observational data, and does not necessarily reflect causal interactions. Nevertheless, functional connectivity is a good way of describing the joint distribution of neural activities; it also serves as a neural marker of different cognitive states and mental diseases [Carhart-Harris et al., 2015; Tian et al., 2006]. More importantly, if we model functional connectivity with built-in assumptions, for example, using dynamic systems to describe how the activity in one region predicts later activities in other regions (e.g., using the concept of "effective connectivity" by Friston [2011]), we may be able to make candidate hypotheses about how information flows in the brain (or about causal interactions of activities across regions), which can be further tested with experimental manipulations or interventions of the neural system in animal studies.

Featuring a high temporal resolution at the millisecond level, MEG and EEG are well-suited tools to record neural activities in human brains and estimate functional connectivity. However, the challenging source localization problem needs to be solved. Typically researchers use a two-step approach—first applying some source localization methods with generic assumptions (as described in Chapter 2), and secondly learning the statistical dependence among the regional activities that are derived from the source estimates. Typically in the second step, some regions of interest (ROIs) are defined, each covering some number of source points, and the representative activity for each ROI is abstracted as a(n) (weighted) average of the estimated source current dipoles in the ROI. Afterward, a variety of functional connectivity models can be used—for example, correlations between pairs of regions, Gaussian graphical models that describe the dependence structure across ROIs [Cribben et al., 2012], phase-locking values of the oscillatory time series in pairs of ROIs at certain frequency bands [Lachaux et al., 1999], and Granger causality models that estimate how well the history in one ROI predicts the activity in another ROI [Granger, 1988; Gow et al., 2008].

This two-step framework is easy to implement and allows flexible choices of functional connectivity models in the second step; therefore, it is widely used. However, the typical regularization or priors in the source localization step do not model the cross-region dependence that we are interested in. For example, the popular L_2-norm-based regularization such as the minimum norm estimate (MNE [Hamalainen and Ilmoniemi, 1994]) assumes an independent prior distribution for each source point; other models by [Pascual-Marqui et al., 1994; Galka et al., 2004; Mattout et al., 2006; Long et al., 2011; Lamus et al., 2012] do consider temporal dependence of source current dipoles as well as local (neighboring) spatial dependence, but they do not model the possibly long-range dependence between regions. Biases due to these assumptions in the first step can not be adjusted in the second step and thus may result in additional errors in the connectivity analysis.

An alternative framework is to solve the source localization problem and estimate functional connectivity jointly in one step. Some previous methods have explored this direction. For example, the dynamical causal modeling method (DCM [David et al., 2006]) assumes that there is a single source current dipole in each of the pre-defined ROIs, and that only these current dipoles in the defined ROIs contribute to the signals in the sensor data. In the DCM, the neural activities in the ROIs are described by a complicated nonlinear dynamical system constrained by neurophysiological knowledge, with time-invariant parameters describing how the current activities in the ROIs are statistically dependent on those in the previous time step. Another method by Fukushima et al. [2015] does not use pre-defined ROIs, but builds a time-invariant multivariate autoregressive model to characterize the dependence across all source points. The coefficients in this high-dimensional multivariate autoregressive model are regularized by a sparsity-inducing prior and more importantly constrained by structural white matter connectivity. In simulation studies, Fukushima et al. [2015] demonstrated that their one-step model better reconstructed the source activities and the dependence structure than several popular two-step methods, when their model assumptions reasonably held.

Both the DCM and the method by Fukushima et al. [2015] are pioneering explorations demonstrating the possible advantage of the one-step framework. However, in order to regularize the problem, they each have their own specific assumptions (e.g., complicated nonlinear models, time-invariant dependence, the availability of accurate structural connectivity constraints, etc.), which may not always be satisfied in practice. In this chapter, we propose novel one-step models that estimate functional connectivity across pre-defined ROIs directly from the sensor recordings, using different assumptions and formulations from those by David et al. [2006] and Fukushima et al. [2015]. We assume that we are given p ROIs, each encompassing multiple source points, and that each ROI has a time-varying representative activity. We also assume that at each time point t, the current dipole of each source point within an ROI is independently distributed as a Gaussian variable, with a mean determined by the ROI activity and a variance that is shared across all source points within the ROI. In addition, the source current dipoles outside any ROIs are also modeled as i.i.d. Gaussian variables with a zero mean and a shared variance. Along with the forward projection from the source space to the sensor space, we have a hierarchical probabilistic model linking the ROI activities and the sensor observations at each time point. By specifying the dependence model across ROIs with some parameters, and eliminating the source current dipoles (i.e., integrating over all possible values of source current dipoles), we can estimate the dependence parameters directly from the sensor data.

In Section 4.2, we first introduce a simplistic model at a fixed time point t, where the activities in the p ROIs are represented as a multivariate Gaussian variable. In this case, the functional connectivity across the ROIs is expressed in the covariance matrix of the Gaussian variable, which is estimated from the sensor observations in multiple trials. Next, we extend the simplistic model further, considering the joint spatio-temporal activity of the p ROIs at all time points in a trial. Again, we model such a spatial-temporal activity as a large multivariate Gaussian variable, assuming its covariance is a Kronecker product of a spatial covariance matrix and a temporal covariance matrix. After estimating the two matrices, we mainly focus on the spatial covariance matrix, which reflects the cross-region functional connectivity. The first model above focuses on the functional connectivity at a single time point but does not consider temporal dependence; the second model includes a description of temporal dependence, but it is mainly designed to characterize stationary neural activities. In some cases, the functional connectivity can be highly dynamic (i.e., time-varying). Thus in Section 4.3, we introduce a state-space model that represents the ROI activities with a time-varying autoregressive model of order 1, which can capture non-stationary neural activities and describe time-varying functional connectivity. The Python code of the three models above can be found in `github.com/YingYang/MEEG_connectivity`.

Using simulations, we demonstrate that our models outperformed the two-step methods that exploit the commonly used MNE, in cases where our model assumptions hold. However, in the scope of this thesis, we have not included empirical comparisons with the other existing one-step methods [David et al., 2006; Fukushima et al., 2015]. The assumptions in our models and in the other existing one-step models differ in terms of whether there are pre-defined ROIs, whether source points outside of the ROIs are modeled, whether the dependence structure is time-varying, and so on. In fact, the assumptions in each method are more general than the others in some ways but less general in other ways. Here, we do not claim that our model assumptions are universally applicable, but we hope our models, as new methodological tools in the one-step framework, can widen the range of options for researchers to choose from.

4.2 One-step models of cross-region covariance

4.2.1 Cross-region covariance at a fixed time point

In this subsection, we simplify the problem by only focusing on the functional connectivity across p given regions of interest (ROIs) at a fixed time point t, which is modeled as the covariance matrix of the activity in the p ROIs across independently and identically distributed (i.i.d.) trials. More specifically, we assume that each ROI contains some number of source points, and that within each ROI, there is a one-dimensional latent variable representing the mean activity across the source points within this ROI [1]. Given this latent variable, the individual source current dipoles within the ROI can be modeled as this mean activity of the ROI plus some additional random noise, which we call *source-space noise* hereafter. Both the ROI latent variables and the source-space noise are projected into the sensor space by the forward model in MEG or EEG, and the goal of our one-step analysis is to estimate the cross-region covariance of the p-latent variables directly from the multi-trial sensor observations.

4.2.1.1 Model formulation

We assume there are q i.i.d. observations at time t from q trials, and there are n sensors and m source points covering the bilateral cortical surfaces. Let $\boldsymbol{y}_t \in \mathbb{R}^n$ denote the readings form the n sensors at time t in an arbitrary trial. Correspondingly, let $\boldsymbol{J}_t \in \mathbb{R}^m$ denote the electric current dipoles at the m source points, among which each current dipole is perpendicular to the local cortical surface. Let $\boldsymbol{u}_t \in \mathbb{R}^p$ denote the p latent variables of the p ROIs at time t in the corresponding trial (i.e., $\boldsymbol{u}_t[i]$ corresponds to the ith ROI ($i = 1, \cdots, p$)). We assume the following distribution of $\boldsymbol{u}_t$

$$\textbf{ROI model:} \quad \boldsymbol{u}_t \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{Q}_t). \tag{4.1}$$

Here the zero-mean assumption can be met practically if we subtract off the sample mean across trials from the sensor readings. The covariance matrix $\boldsymbol{Q}_t$ represents the functional connectivity among the regions at time t; therefore it is the main parameter of interest.

Let $\mathcal{A}_i$ be the set of indices of the source points in the ith ROI. For any $l \in \mathcal{A}_i$, the electric current dipole of the lth source point at time t in the trial, denoted by $\boldsymbol{J}_t[l]$, is modeled as the mean activity across the source points in the ROI (i.e., the latent variable in the ROI), $\boldsymbol{u}_t[i]$, plus a noise term, $\boldsymbol{w}_t[l]$, which is independent of $\boldsymbol{u}_t[i]$. Here, the vector $\boldsymbol{w}_t \in \mathbb{R}^m$ denotes the source-space noise at all source points at time t. In addition, we assume that for all $l \in \mathcal{A}_i$, $\boldsymbol{w}_t[l] \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma_i^2)$; that is, the source-space noise is independently distributed as a Gaussian variable for each source point

[1] The latent variable can also be extended to a more general case; see the model formulation below.

within the ROI, with a zero mean and a variance term σ_i^2, which is shared within the ROI.

$$\boldsymbol{J}_t[l] = \boldsymbol{u}_t[i] + \boldsymbol{w}_t[l], \quad \boldsymbol{w}_t[l] \sim \mathcal{N}(0, \sigma_i^2), \quad \text{for } l \in \mathcal{A}_i, \tag{4.2}$$

Additionally, we denote the indices of sources points outside of any ROIs by $\mathcal{A}_0 = \{l, \forall l \notin \cup_{i=1}^{p} \mathcal{A}_i\}$. For each of these source points, we assume the electric current dipole at time t only includes the i.i.d. source-space noise term, with a zero mean and a variance σ_0^2, which is shared by all the source points in $\mathcal{A}_0$.

$$\boldsymbol{J}_t[l] = \boldsymbol{w}_t[l], \quad \boldsymbol{w}_t[l] \sim \mathcal{N}(0, \sigma_0^2), \quad \text{for } l \in \mathcal{A}_0. \tag{4.3}$$

An illustration of the model is shown in Figure 4.1.

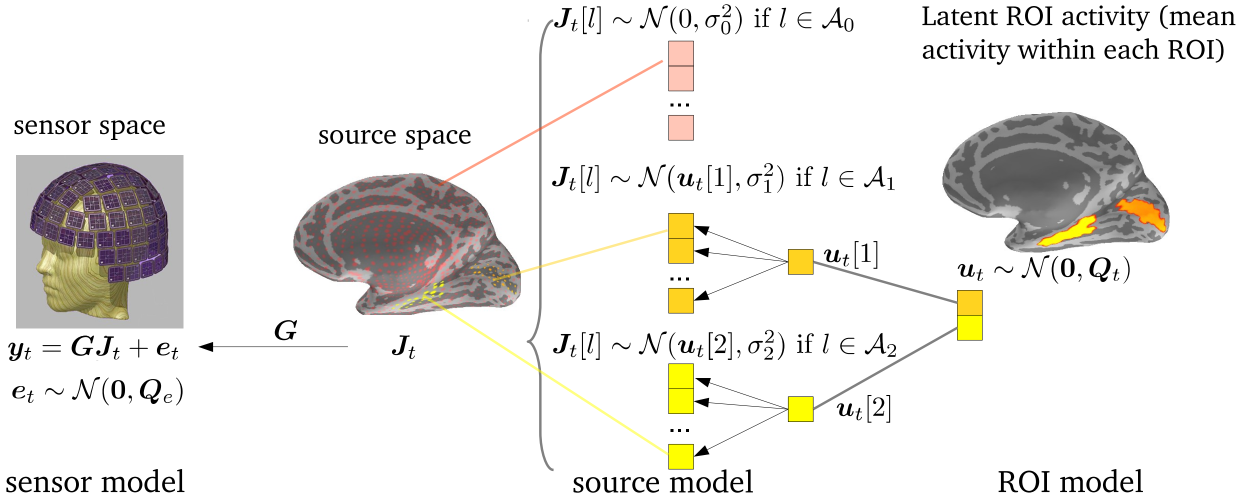

$\boldsymbol{J}_t[l] \sim \mathcal{N}(0, \sigma_0^2)$ if $l \in \mathcal{A}_0$
Latent ROI activity (mean activity within each ROI)
sensor space
source space
$\boldsymbol{J}_t[l] \sim \mathcal{N}(\boldsymbol{u}_t[1], \sigma_1^2)$ if $l \in \mathcal{A}_1$
$\boldsymbol{u}_t[1]$
$\boldsymbol{u}_t \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{Q}_t)$
$\boldsymbol{G}$
$\boldsymbol{y}_t = \boldsymbol{G}\boldsymbol{J}_t + \boldsymbol{e}_t$
$\boldsymbol{J}_t$
$\boldsymbol{J}_t[l] \sim \mathcal{N}(\boldsymbol{u}_t[2], \sigma_2^2)$ if $l \in \mathcal{A}_2$
$\boldsymbol{e}_t \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{Q}_e)$
$\boldsymbol{u}_t[2]$
sensor model
source model
ROI model

Figure 4.1: Illustration of the one-step cross-region covariance analysis at time point t, where the latent activity ($\boldsymbol{u}_t[i]$) represents the mean activity across the source points within the ROI ($\mathcal{A}_i$). Here, for simplicity, we include two ROIs as an example.

Writing Equations (4.2) and (4.3) in a concise way, we have

$$\textbf{source model:} \quad \boldsymbol{J}_t = \boldsymbol{L}\boldsymbol{u}_t + \boldsymbol{w}_t, \quad \boldsymbol{w}_t \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{Q}_{J_n}) \tag{4.4}$$

where $\boldsymbol{Q}_{J_n}$ is an $m \times m$ diagonal matrix and $\boldsymbol{Q}_{J_n}[l,l] = \sigma_i^2$ if $l \in \mathcal{A}_i$, for $l = 1, 2, \cdots, m$ and $i = 0, 1, 2, \cdots, p$. The matrix $\boldsymbol{L}$ is of size $m \times p$, where $\boldsymbol{L}[l,i]$ represents the contribution of the latent variable in the ith ROI to the lth source point. That is, if the lth source point is in the ith ROI, $\boldsymbol{L}[l,i] = 1$, otherwise $\boldsymbol{L}[l,i] = 0$. In a simple toy example, if there are $p = 2$ ROIs and $m = 7$ source points, and $\mathcal{A}_1 = \{1,2,3\}; \mathcal{A}_2 = \{4,6\}; \mathcal{A}_0 = \{5,7\}$, then we have

$$\boldsymbol{L} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 0 \\ 0 & 1 \\ 0 & 0 \end{pmatrix}$$

and $\boldsymbol{Q}_{J_n} = \text{diag}\{\sigma_1^2, \sigma_1^2, \sigma_1^2, \sigma_2^2, \sigma_0^2, \sigma_2^2, \sigma_0^2\}$.

We can also extend this model to a more general case, where $\boldsymbol{L}[l,i]$s for all $l \in \mathcal{A}_i, l = 1, \cdots, m$ for each $i = 1, \cdots, p$ can take any real values (even negative values) besides 1. In this case, each source current dipole in an ROI is distributed with the mean $(\boldsymbol{L}[l,i]\boldsymbol{u}_t[i])$. Such an extension can more flexibly describe the contributions of the latent ROI activity to the source current dipoles, especially because of the following reason. Here the source space is defined as a mesh covering the cortical surfaces, where the current dipole at each source point is perpendicular to the local surface. Due to the folding of the cortical surfaces, the current dipoles within an ROI can have opposite orientations (e.g., on each side of a gyrus as shown in Figure 4.2). Now, if we define the latent ROI activity as the average of the vector-valued current dipoles, where we do vector summation including the orientations, instead of summing the scalar-valued dipoles constrained to their orientations (as in the case with 0/1 entries in $\boldsymbol{L}$), the inner products between the latent ROI current dipole (the blue arrow in Figure 4.2) and the individual source current dipoles (the black arrows in Figure 4.2) can vary across source points; for some source points, the inner products can be negative. In this case, the contribution of the latent ROI activity to different source points can be expressed in the real-valued entries of $\boldsymbol{L}$ (i.e., $\boldsymbol{L}[l,i]$s for all $l \in \mathcal{A}_i, l = 1, \cdots, m$ for each $i = 1, \cdots, p$), and these entries can be learned from the data. Note that we still fix the entries in $\boldsymbol{L}$ that correspond to the source points outside of each ROI to zero—$\boldsymbol{L}[l,i] = 0$ if $l \notin \mathcal{A}_i, l = 1, \cdots, m$ for each $i = 1, \cdots, p$—as we assume that each source point can only belong to at most one ROI and the ROIs do not overlap.

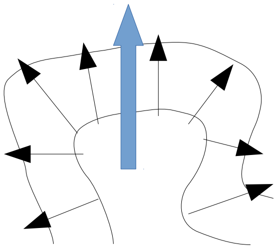


Figure 4.2: Illustration of the latent ROI activity (blue) and the current dipoles of individual source points (black) in the extended case.

Finally, given $\boldsymbol{J}_t$, the n-dimensional sensor readings in the trial follow the forward model below.

$$\textbf{sensor model:} \quad \boldsymbol{y}_t = \boldsymbol{G}\boldsymbol{J}_t + \boldsymbol{e}_t \tag{4.5}$$

where $\boldsymbol{G} \in \mathbb{R}^{n \times m}$ is the forward matrix and $\boldsymbol{e}_t \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{Q}_e)$ is the sensor noise, assuming the noise covariance matrix $\boldsymbol{Q}_e$ is pre-computed from empty-room recordings or baseline recordings (while the participant is not doing relevant tasks).

As illustrated in Figure 4.1, putting together what we have described so far, we have the following hierarchical model relating the p-dimensional ROI latent activity $\boldsymbol{u}_t$ to the sensor readings $\boldsymbol{y}_t$.

$$\begin{aligned} \boldsymbol{u}_t &\sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{Q}_t) \\ (\boldsymbol{J}_t | \boldsymbol{u}_t) &\sim \mathcal{N}(\boldsymbol{L}\boldsymbol{u}_t, \boldsymbol{Q}_{J_n}) \\ (\boldsymbol{y}_t | \boldsymbol{J}_t) &\sim \mathcal{N}(\boldsymbol{G}\boldsymbol{J}_t, \boldsymbol{Q}_e) \end{aligned} \tag{4.6}$$

Let $f(\cdot)$ denote the probability density function in general; we have

$$f(\boldsymbol{y}_t|\boldsymbol{u}_t) \propto \int f(\boldsymbol{y}_t|\boldsymbol{J}_t) f(\boldsymbol{J}_t|\boldsymbol{u}_t) d\boldsymbol{J}_t,$$

by integrating $\boldsymbol{J}_t$ out, we have

$$\begin{aligned} \boldsymbol{u}_t &\sim \mathcal{N}(0, \boldsymbol{Q}_t) \\ (\boldsymbol{y}_t | \boldsymbol{u}_t) &\sim \mathcal{N}(\boldsymbol{G}\boldsymbol{L}\boldsymbol{u}_t, \boldsymbol{Q}_e + \boldsymbol{G}\boldsymbol{Q}_{J_n}\boldsymbol{G}'). \end{aligned} \tag{4.7}$$

In other words, the conditional distribution of $(\boldsymbol{y}_t|\boldsymbol{u}_t)$ can be described as

$$\begin{aligned} \boldsymbol{y}_t &= \boldsymbol{G}\boldsymbol{L}\boldsymbol{u}_t + \boldsymbol{\eta}_t \\ \boldsymbol{\eta}_t &\sim \mathcal{N}(0, \boldsymbol{Q}_e + \boldsymbol{G}\boldsymbol{Q}_{J_n}\boldsymbol{G}'). \end{aligned} \tag{4.8}$$

Note that the covariance of the new noise $\boldsymbol{\eta}_t$ in the sensor space is a sum of the sensor noise covariance $\boldsymbol{Q}_e$ and the projection of source-space noise covariance $\boldsymbol{G}\boldsymbol{Q}_{J_n}\boldsymbol{G}'$.

Similarly, by further integrating $\boldsymbol{u}_t$ out, we have the marginal distribution of $\boldsymbol{y}_t$ below.

$$\begin{aligned} f(\boldsymbol{y}_t) &= \int f(\boldsymbol{y}_t|\boldsymbol{u}_t) f(\boldsymbol{u}_t) d\boldsymbol{u}_t \\ \boldsymbol{y}_t &\sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{Q}_e + \boldsymbol{G}\boldsymbol{Q}_{J_n}\boldsymbol{G}' + \boldsymbol{G}\boldsymbol{L}\boldsymbol{Q}_t\boldsymbol{L}'\boldsymbol{G}') \end{aligned} \tag{4.9}$$

which shows a direct relationship between the cross-ROI covariance $\boldsymbol{Q}_t$ and the sensor readings $\boldsymbol{y}_t$. This provides us a way of solving for $\boldsymbol{Q}_t$ using maximum-likelihood estimation.

4.2.1.2 Parameter estimation

We have been using $\boldsymbol{u}_t$, $\boldsymbol{J}_t$ and $\boldsymbol{y}_t$ to denote the corresponding variables in any arbitrary trial. In the text below, we add the superscript $\cdot^{(r)}$ (e.g., $\boldsymbol{y}_t^{(r)}$), when it is necessary to refer to these variables specifically in the rth trial ($r = 1, \cdots, q$).

In our model above, $\boldsymbol{Q}_t$ and $\boldsymbol{Q}_{J_n}$ (or σ_i^2, $i = 0, 1, \cdots, p$) are unknown parameters to be estimated, and we are mainly interested in $\boldsymbol{Q}_t$. The sensor readings across all trials (i.e., $\boldsymbol{y}_t^{(r)}$s), $\boldsymbol{G}$ and $\boldsymbol{Q}_e$ are given. The weight matrix $\boldsymbol{L}$ can be a fixed "0/1" matrix as described, which is known, if the ROI latent variable represents the mean activity across the source points within each ROI (i.e., the mean of the scalar activities); in the extension where $\boldsymbol{L}[l, i]$s for all $l \in \mathcal{A}_i, l = 1, \cdots, m$ for each $i = 1, \cdots, p$ can be any real values, we can treat $\boldsymbol{L}$ as an unknown parameter. In the latter case, the problem can be unidentifiable without any penalization. We set $\boldsymbol{L}[l,i] = 0$ if $l \notin \mathcal{A}_i, l = 1, \cdots, m$ for each $i = 1, \cdots, p$, and also add additional priors on the non-zero entries in $\boldsymbol{L}$ to encourage spatial smoothness. Optional priors of $\boldsymbol{Q}_t$ and $\boldsymbol{Q}_{J_n}$ can be added to further regularize the problem. Below we discuss how to solve for the unknown parameters (with the priors).

According to Equation (4.9), we can solve for the parameters $\boldsymbol{Q}_t$, $\boldsymbol{Q}_{J_n}$ and $\boldsymbol{L}$ by minimizing the negative log-likelihood of $\{\boldsymbol{y}_t^{(r)}\}_{r=1}^q$, which we denote by "nllh", plus the penalty functions corresponding to the negative logarithms of the probability density functions of the priors on $\boldsymbol{Q}_t$, σ_is and $\boldsymbol{L}$, which we denote by "Pen$(\boldsymbol{Q}_t)$, Pen(σ_i) and Pen$(\boldsymbol{L})$". Therefore, the objective function to be minimized is

$$\text{obj} = \text{nllh} + \text{Pen}(\boldsymbol{Q}_t) + \sum_{i=0}^{p} \text{Pen}(\sigma_i) + \text{Pen}(\boldsymbol{L}).$$

We use coordinate descent to alternate among the optimization for $\boldsymbol{Q}_t$, $\boldsymbol{Q}_{J_n}$ and $\boldsymbol{L}$. That is, we optimize each $\boldsymbol{\theta} \in \{\boldsymbol{Q}_t, \boldsymbol{Q}_{J_n}, \boldsymbol{L}\}$ in each iteration while fixing the other two parameters, until the change of the objective function is small enough. Each $\boldsymbol{\theta}$ is optimized with the gradient descent algorithm with backtracking (Algorithm 4).

Algorithm 4: The gradient descent algorithm to minimize the objective function obj$(\boldsymbol{\theta})$ for each $\boldsymbol{\theta}$ with backtracking

```
Data: the objective function obj($\boldsymbol{\theta}$), the gradient $\nabla\text{obj} = \frac{\partial \text{obj}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$, an initial value of $\boldsymbol{\theta}$, a
      positive scalar $L_0$ that determines the step size, a scalar $\eta > 1$
Result: $\boldsymbol{\theta}$ that yields the local minimum of obj($\boldsymbol{\theta}$)
initialization: set $\boldsymbol{\theta}$ to the initial value ;
while the difference of the objective between two iterations is not small enough do
    $L \leftarrow L_0$;
    while $obj(\boldsymbol{\theta} - \frac{1}{L}\nabla obj) > obj(\boldsymbol{\theta}) - \frac{1}{2L}\|\nabla obj\|_2^2$ do
        $L \leftarrow \eta L$;
    end
    $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \frac{1}{L}\nabla\text{obj}$;
end
```

Below we present the formulation negative log-likelihood and the corresponding gradients. The negative log-likelihood of the multi-trial sensor data is

$$\begin{aligned} \text{nllh} &= q \log\det(\boldsymbol{Q}_{yt}) + \sum_{r=1}^{q} (\boldsymbol{y}_t^{(r)})' \boldsymbol{Q}_{yt}^{-1} (\boldsymbol{y}_t^{(r)}) = q \log\det(\boldsymbol{Q}_{yt}) + \text{trace}(\tilde{\mathbf{Y}}_t' \tilde{\mathbf{Y}}_t \boldsymbol{Q}_{yt}^{-1}) \\ \boldsymbol{Q}_{yt} &= \boldsymbol{Q}_e + \boldsymbol{G}\boldsymbol{Q}_{J_n}\boldsymbol{G}' + \boldsymbol{G}\boldsymbol{L}\boldsymbol{Q}_t\boldsymbol{L}'\boldsymbol{G}' \qquad (4.10) \\ &= \boldsymbol{Q}_e + \sum_{i=0}^{p} \sigma_i^2 (\boldsymbol{G}[:, l \in \mathcal{A}_i] \boldsymbol{G}[:, l \in \mathcal{A}_i])' + \boldsymbol{G}\boldsymbol{L}\boldsymbol{\Gamma}_t\boldsymbol{\Gamma}_t'\boldsymbol{L}'\boldsymbol{G}' \end{aligned}$$

where $\tilde{\boldsymbol{Y}}_t \in \mathbb{R}^{q \times n}$ denotes the sensor data of all q trials at time t (i.e., $\tilde{\boldsymbol{Y}}_t[r,:] = (\boldsymbol{y}_t^{(r)})'$). In addition, since $\boldsymbol{Q}_t$ must be a positive definite matrix, we re-parametrize it using the Cholesky decomposition $\boldsymbol{Q}_t = \boldsymbol{\Gamma}_t \boldsymbol{\Gamma}_t'$ where $\boldsymbol{\Gamma}_t$ is a full rank $p \times p$ lower triangular matrix. The $\boldsymbol{G}[:, l \in \mathcal{A}_i]$ term in Equation (4.10) denotes the columns of $\boldsymbol{G}$ corresponding to the source points in $\mathcal{A}_i$ for $i = 0, 1, \cdots, p$. The gradient of the negative log-likelihood with respect to $\boldsymbol{\Gamma}_t$, $\boldsymbol{L}$ and σ_is are listed below (Equation (4.11)), and the derivations can be found in the Appendix (Section 4.5.1.2).

$$\begin{aligned}
\frac{\partial \text{llh}}{\partial \boldsymbol{Q}_{yt}} &= q\boldsymbol{Q}_{yt}^{-1} - \boldsymbol{Q}_{yt}^{-1}\tilde{\boldsymbol{Y}}_t'\tilde{\boldsymbol{Y}}_t\boldsymbol{Q}_{yt}^{-1} \\
\frac{\partial \text{llh}}{\partial \boldsymbol{\Gamma}_t} &= 2\boldsymbol{L}'\boldsymbol{G}'\frac{\partial \text{llh}}{\partial \boldsymbol{Q}_{yt}}\boldsymbol{G}\boldsymbol{L}\boldsymbol{\Gamma}_{\boldsymbol{t}} \\
\frac{\partial \text{llh}}{\partial \boldsymbol{L}} &= 2\boldsymbol{G}'\frac{\partial \text{llh}}{\partial \boldsymbol{Q}_{yt}}\boldsymbol{G}\boldsymbol{L}\boldsymbol{\Gamma}_t\boldsymbol{\Gamma}_t' \\
\frac{\partial \text{llh}}{\partial \sigma_i} &= 2\sigma_i \text{trace}\left((\frac{\partial \text{llh}}{\partial \boldsymbol{Q}_{yt}})'(\boldsymbol{G}[:, l \in \mathcal{A}_i]\boldsymbol{G}[:, l \in \mathcal{A}_i]') \right)
\end{aligned} \tag{4.11}$$

We defer the formulation of Pen$(\boldsymbol{L})$ and the optional penalty functions (Pen$(\boldsymbol{Q}_t)$ and Pen(σ_i)), as well as the corresponding gradients to the appendix (Section 4.5.1.3).

4.2.2 Cross-region covariance with Kronecker structure

In this section, we again assume that the $p \times 1$ vector $\boldsymbol{u}_t$ represents the latent activity of the p ROIs at time t. However, we consider all of the T time points within a trial. Let $\boldsymbol{U} \in \mathbb{R}^{p \times T}$ denote the ROI latent activity in an arbitrary trial, where $\boldsymbol{U} = (\boldsymbol{u}_1, \boldsymbol{u}_2, \cdots, \boldsymbol{u}_T)$. We concatenate the columns of $\boldsymbol{U}$ into a large (pT)-dimensional vector $\boldsymbol{u}$, and focus on its $pT \times pT$ covariance matrix $\boldsymbol{Q}_u$ across trials, which reflects the spatio-temporal dependence across the p ROIs. We further assume that $\boldsymbol{Q}_u$ can be decomposed as a Kronecker product of a $T \times T$ temporal covariance matrix ($\boldsymbol{Q}_T$), and a static $p \times p$ spatial covariance matrix ($\boldsymbol{Q}_S$) (i.e., $\boldsymbol{Q}_u = \boldsymbol{Q}_T \otimes \boldsymbol{Q}_S$, where $\otimes$ denotes the Kronecker product operator). Note that this assumption supposes that the spatial-temporal dependence among the p ROIs can be represented by simply multiplying entries in the spatial covariance matrix and the temporal covariance matrix, and the spatial correlation structure across ROIs is stationary over time. The spatial covariance $\boldsymbol{Q}_S$ is of main interest. By modeling the temporal dependence, we may utilize multiple time points instead of a single time point to better estimate the spatial covariance. In practice, if the dependence between the ROIs is time-varying, we can segment the time windows into short and relatively stationary windows before applying this model.

Let vec$(\cdot)$ denote the operation of concatenating columns in a matrix; then we have $\boldsymbol{u} = \text{vec}(\boldsymbol{U})$. As mentioned above, we assume vec$(\boldsymbol{U}) \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{Q}_u)$. Let $\boldsymbol{J} \in \mathbb{R}^{m \times T}$ be the m source current dipoles at all T time points in the trial, and let $\boldsymbol{Y} \in \mathbb{R}^{n \times T}$ be corresponding sensor time series. Given $\boldsymbol{U}$, we assume a similar model as Equation (4.4), where vec$(\boldsymbol{J})$ has a Gaussian distribution around the mean vec$(\boldsymbol{L}\boldsymbol{U})$, and with a covariance matrix $\boldsymbol{Q}_{JT} \otimes \boldsymbol{Q}_{J_n}$. Here $\boldsymbol{Q}_{JT}$ is a $T \times T$ temporal covariance matrix of the source-space noise, and $\boldsymbol{Q}_{J_n}$ is defined as in Section 4.2.1. Together with the forward model in Equation (4.5), we have

$$\begin{aligned} \text{vec}(\boldsymbol{U}) &\sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{Q}_T \otimes \boldsymbol{Q}_S) \\ \text{vec}(\boldsymbol{J})|\text{vec}(\boldsymbol{U}) &\sim \mathcal{N}(\text{vec}(\boldsymbol{L}\boldsymbol{U}), \boldsymbol{Q}_{JT} \otimes \boldsymbol{Q}_{J_n}) \\ \text{vec}(\boldsymbol{Y})|\text{vec}(\boldsymbol{J}) &\sim \mathcal{N}(\text{vec}(\boldsymbol{G}\boldsymbol{J}), \underset{T\times T}{\boldsymbol{I}} \otimes \boldsymbol{Q}_e). \end{aligned} \tag{4.12}$$

To further reduce the complexity of the model, we assume $\boldsymbol{Q}_{JT} = \boldsymbol{Q}_T$. That is, the temporal covariance of the source-space noise is the same as the temporal covariance of the ROI latent variables. Also, after pre-whitening (left multiplying $\boldsymbol{Q}_e^{-1/2}$ with $\boldsymbol{Y}$, $\boldsymbol{G}$ and the sensor noise), we can assume the new spatial covariance of the sensor noise $\boldsymbol{Q}_e$ is transformed into an identity matrix $\boldsymbol{I}_{n\times n}$. Below, let us still use $\boldsymbol{G}$ and $\boldsymbol{Y}$ to denote the pre-whitened forward matrix and sensor data. Under these assumptions, we eliminate $\boldsymbol{J}$ and $\boldsymbol{U}$ and obtain the following marginal distribution (see the derivations in the appendix in Section 4.5.2.1).

$$\begin{aligned} \text{vec}(\boldsymbol{Y}) &\sim \mathcal{N}(\boldsymbol{0}, \underset{T\times T}{\boldsymbol{I}} \otimes \underset{n\times n}{\boldsymbol{I}} + (\underset{T\times T}{\boldsymbol{I}} \otimes \boldsymbol{G})(\boldsymbol{Q}_T \otimes \boldsymbol{Q}_{J_n})(\underset{T\times T}{\boldsymbol{I}} \otimes \boldsymbol{G})' + (\underset{T\times T}{\boldsymbol{I}} \otimes \boldsymbol{G}\boldsymbol{L})(\boldsymbol{Q}_T \otimes \boldsymbol{Q}_S)(\underset{T\times T}{\boldsymbol{I}} \otimes \boldsymbol{G}\boldsymbol{L})' \\ &= \mathcal{N}(\boldsymbol{0}, (\underset{nT\times nT}{\boldsymbol{I}} + \boldsymbol{Q}_T \otimes (\boldsymbol{G}\boldsymbol{Q}_{J_n}\boldsymbol{G}' + \boldsymbol{G}\boldsymbol{L}\boldsymbol{Q}_S\boldsymbol{L}'\boldsymbol{G}')) \end{aligned} \tag{4.13}$$

Again, we can estimate $\boldsymbol{Q}_T$ and $\boldsymbol{Q}_S$, as well as $\boldsymbol{Q}_{J_n}$ and $\boldsymbol{L}$, by maximizing the marginal log-likelihood of the trial-by-trial sensor readings plus optional penalty terms representing priors on the parameters. We denote the sensor observation at all T time points in the rth trial by $\boldsymbol{Y}^{(r)}$ hereafter ($r = 1, \cdots, q$). Let $\boldsymbol{Q}_Y = (\boldsymbol{I} + \boldsymbol{Q}_T \otimes (\boldsymbol{G}\boldsymbol{Q}_{J_n}\boldsymbol{G}' + \boldsymbol{G}\boldsymbol{L}\boldsymbol{Q}_S\boldsymbol{L}'\boldsymbol{G}')$. Then the negative log-likelihood, denoted by nllh$_{kron}$, is

$$\text{nllh}_{kron} = q \log\det(\boldsymbol{Q}_Y) + \sum_{r=1}^{q} \text{vec}(\boldsymbol{Y}^{(r)})'\boldsymbol{Q}_Y^{-1}\text{vec}(\boldsymbol{Y}^{(r)})) \tag{4.14}$$

The same priors in Section 4.2.1 can be used [2]. Again, we can solve for the unknown parameters using the coordinate descent method—alternating among the optimization for $\boldsymbol{Q}_T$, $\boldsymbol{Q}_S$, $\boldsymbol{Q}_{J_n}$ and $\boldsymbol{L}$ with the gradient descend algorithm (Algorithm 4). Re-parametrization of $\boldsymbol{Q}_T$ and $\boldsymbol{Q}_S$ with Cholesky decompositions also applies here. The log-likelihood function and its gradients with respect to each parameter can be effectively computed utilizing the Kronecker structure. We defer the details to the Appendix in Section 4.5.2.2.

4.2.3 Comparison with a two-step approach using the MNE in simulations

Using simulated data, we compared our one-step models with a two-step method using the commonly-used source localization method—the minimum norm estimate (MNE [Hamalainen and Ilmoniemi,

[2] We note that the Kronecker structure of $\boldsymbol{Q}_u$ is non-identifiable in the sense that $c\boldsymbol{Q}_T \otimes 1/c\boldsymbol{Q}_S = \boldsymbol{Q}_T \otimes \boldsymbol{Q}_S$, for any non-zero real scalar c. We can use the priors defined for $\boldsymbol{Q}_t$ in Section 4.2.1 to regularize $\boldsymbol{Q}_T$ and $\boldsymbol{Q}_S$ to alleviate the effect of this issue.

1994]). In Step 1, we obtained the minimum-norm estimate of $\boldsymbol{J}_t$ for each time point in each trial. Given $\boldsymbol{G}, \boldsymbol{Q}_e$, a prior $\boldsymbol{J}_t \sim \mathcal{N}(\boldsymbol{0}, (1/\lambda)\boldsymbol{I}), \lambda > 0$ and the corresponding $\boldsymbol{y}_t$, the estimate was $\boldsymbol{J}_t \leftarrow \boldsymbol{G}'(\boldsymbol{G}\boldsymbol{G}' + \lambda \boldsymbol{Q}_e)^{-1}\boldsymbol{y}_t$. After $\boldsymbol{J}_t$ was estimated, we computed $\boldsymbol{u}_t$ that represented the latent activity in individual ROIs, according to Equation (4.4). Because we simulated the data, we knew the true values of $\boldsymbol{L}$, so we plugged in the true $\boldsymbol{L}$ to estimate the least square solution of $\boldsymbol{u}_t$, assuming $\boldsymbol{Q}_{J_n}$ was proportional to an identity matrix (no heteroskedasticity). This procedure was done for each of the q trials. In Step 2, if we focused on a single time point t, we computed the covariance of the estimated $\boldsymbol{u}_t$ across trials, as an estimate of $\boldsymbol{Q}_t$; if we used the Kronecker assumption, based on the estimated $\boldsymbol{u}_t$ of all T time points across all trials, we obtained the maximum-likelihood estimator of $\boldsymbol{Q}_T$ and $\boldsymbol{Q}_S$ (see the appendix in Section 4.5.2.4 for details). We label the two-step procedure *mneTrueL* for the case with a single time point t, and *mneTrueLKronecker* for the case with the Kronecker assumption.

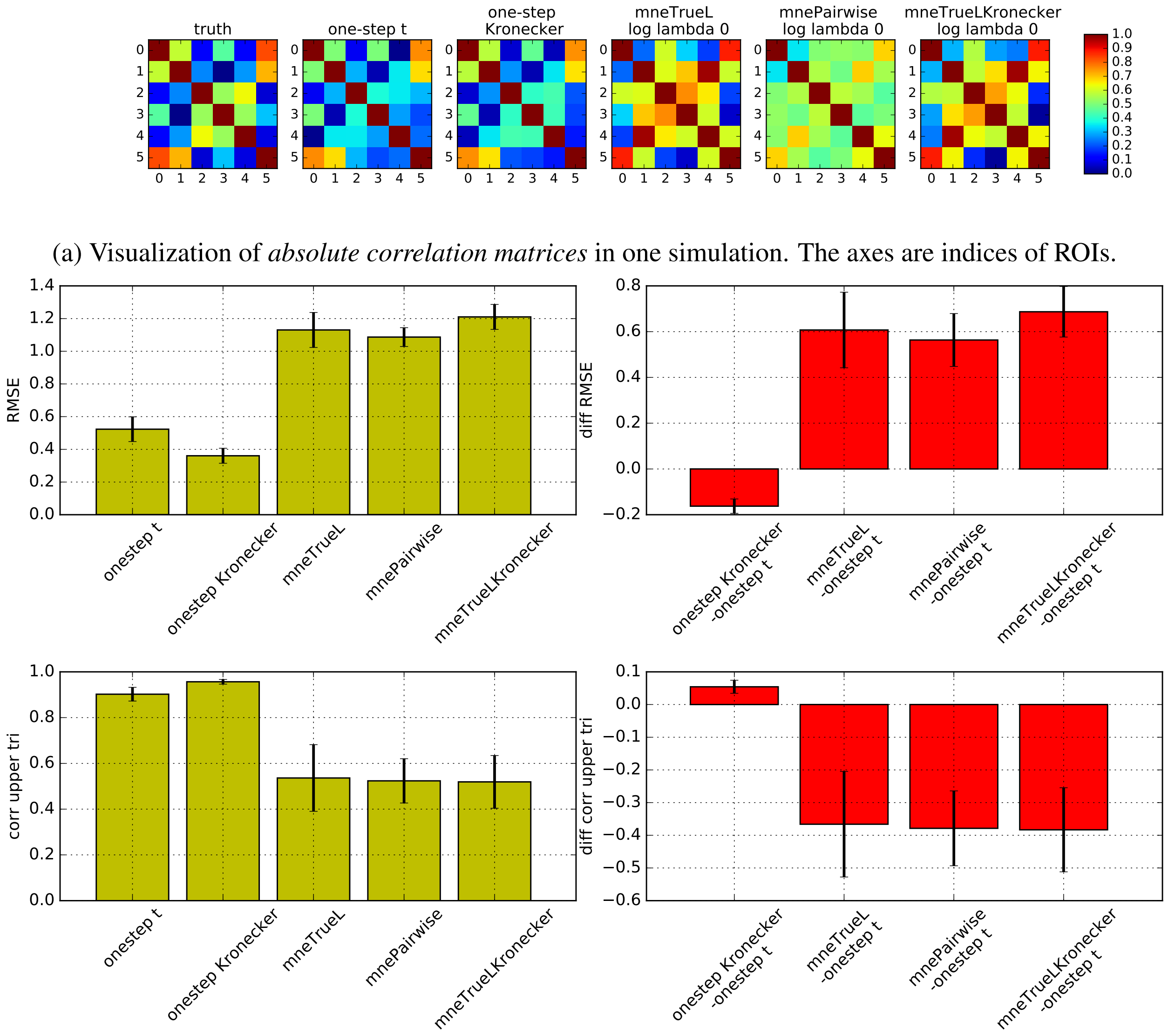
We evaluated how well the one-step and the two-step methods were able to estimate the spatial connectivity—how similar the estimated $\boldsymbol{Q}_t$ (in the single-time-point case) or $\boldsymbol{Q}_S$ (in the Kronecker-structured covariance case) was to the true spatial covariance. The MNE inherently shrinks the estimates of $\boldsymbol{J}_t$ (and thus $\boldsymbol{u}_t$) toward zero, so the resulting $\boldsymbol{Q}_t$ or $\boldsymbol{Q}_S$ might have a different scale from the truth. Therefore in the evaluation, we looked at the correlation matrix derived from $\boldsymbol{Q}_t$ or $\boldsymbol{Q}_S$, such that the marginal variance of the latent activity in each ROI was normalized to 1. The positive or negative signs in $\boldsymbol{y}_t$ only represented the directions of source current dipoles, so both positive and negative correlations between two ROIs were meaningful. Hence, we took the absolute values of the entries in the correlation matrix, which we term the *absolute correlation matrix* hereafter. In addition, typically we would not know the true $\boldsymbol{L}$ in practice. In the single-time-point case, we could compute the absolute values of pairwise of correlations between all source points in two different ROIs and then take the mean across the pairs, as a summary of the correlations between the ROIs. We also computed the *absolute correlation matrix* in this way; the results are labeled *mnePairwise* below.

We simulated 306-dimensional MEG sensor data according to our model assumptions, using the forward model of a real brain. The source space included $m \approx 8000$ source points covering the cortical surfaces in the two hemispheres, and each source point was assigned an orientation perpendicular to the local cortical surface. Six ROIs were used ($p = 6$), including the bilateral orbitofrontal regions, the bilateral parahippocampal regions and the bilateral lateral occipital regions in the ventral visual cortex. We randomly sampled the positive definite spatial covariance matrix $\boldsymbol{Q}_S = \boldsymbol{V}_S \boldsymbol{D}_S \boldsymbol{V}_S'$, where entries in the $p \times p$ matrix $\boldsymbol{V}_S$ were i.i.d. samples from a standard normal distribution, and entries in the diagonal $p \times p$ matrix $\boldsymbol{D}_S$ were i.i.d. samples from a Gamma distribution (shape = 0.5, scale = 1.0). We set $T = 5$, and defined $\boldsymbol{Q}_T[i, j] = \exp(-0.1(i-j)^2)$ if $i \neq j$ and $\boldsymbol{Q}_T[i, i] = 1.01$. We sampled $q = 320$ trials of $\text{vec}(\boldsymbol{U})$, with a zero mean and the spatiotemporal covariance $\boldsymbol{Q}_T \otimes \boldsymbol{Q}_S$. For each time point t, we sampled $\boldsymbol{J}_t$ according to Equation (4.4), where σ_i^2s in $\boldsymbol{Q}_{J_n}$ were i.i.d samples from a Gamma distribution (shape = 2, scale = 1). The relevant (non-zero) entries in each column of the true $\boldsymbol{L}$ matrix ($\boldsymbol{L}[l \in \mathcal{A}_i, i]$ for $l = 1, \cdots, m$ and each $i = 1, \cdots, p$) were either 1 or -1, depending on whether the local orientation of the source point had a positive or negative dot product with the "principal" orientation in the ROI, which was the first singular vector of all orientations in that ROI. The unit of the source activity was nanoamphere meter (nAm). Finally, the sensor data was generated according to the forward model, where the sensor noise covariance matrix $\boldsymbol{Q}_e$ was from a real data set in MNE-Python. The forward matrix $\boldsymbol{G}$ in the simulations was not normalized in any way to reduce depth bias in both the one-step and the two-step methods.

When applying our one-step model, only a prior on $\boldsymbol{L}$ mentioned above was used (i.e., we did not use priors on other parameters). See the appendix in Section 4.5.1.3 for details on the prior. Multiple values of the tuning parameter λ were used in the two-step methods with the MNE. For the one-step model at a fixed time point, as well as the two-step methods with the MNE at a fixed time point (*mneTrueL* and *mnePairwise*), only data in the middle time point ($t = 3$) were used. For the one-step Kronecker-structured covariance model and the two-step method *mneTrueLKronecker*, data in all $T = 5$ time points were used. The initial value of $\boldsymbol{Q}_S$ in the one-step Kronecker model was set as the estimate of $\boldsymbol{Q}_t$ by the one-step single-time-point model.

Figure 4.3a shows the *absolute correlation matrices* obtained using our one-step methods, both for a fixed time point (*one-step t*) and under the Kronecker assumption (*one-step Kronecker*), as well as the *absolute correlation matrices* by the two-step MNE methods (*mneTrueL*, *mnePairwise* and *mneTrueLKronecker*) in one example simulation. For the latter, we only show the results for one λ value, but the results with other λ values were visually similar. The results by the one-step methods were visually more similar to the truth than those by the two-step MNE methods.

To further quantify the similarity to the truth, we used two metrics. The first was the root mean squared error (RMSE) of the *absolute correlation matrix* compared to the truth; the second was the correlation of the concatenated upper triangular entries of the *absolute correlation matrix* (excluding the diagonal) between the estimates and the truth, which reflected how similar to the truth the pairwise correlation pattern across ROIs was. The left column in Figure 4.3b shows the averaged metrics across 5 independent simulations, where the error bars show the standard errors. For *mneTrueL*, *mnePairwise* and *mneTrueLKronecker*, where multiple values of λ were used, only the best metric across the different tuning parameters in each simulation was used to compute the mean. The mean across simulations of the paired differences in the metrics between the two-step methods and the one-step method in the single-time point case are also plotted in the right column. So is the mean of the paired differences between the two versions of one-step methods (the Kronecker version versus the single-time-point version). The one-step methods yielded better metrics (smaller RMSE and larger correlation with the truth) than the two-step methods in these simulations. Compared with the one-step method for a fixed time point, the one-step method with the Kronecker-structured covariance gave slightly better performance, by utilizing data from multiple time points.



(a) Visualization of *absolute correlation matrices* in one simulation. The axes are indices of ROIs.

(b) Evaluation of the performance. In the left column, *RMSE* denotes the root mean squared error of the *absolute correlation matrices* compared with the truth, and *corr upper tri* denotes the correlation of the upper triangular entries of the *absolute correlation matrices* with the truth. Differences between all the other methods and the one-step single-time-point method are shown in the right column. The error bars denote standard errors across 5 independent simulations.

Figure 4.3: Simulation results: estimating the cross-region correlation matrix.

4.2.4 Analytical comparison with the two-step approach using the MNE

Here, we aim to give an intuitive explanation on why the one-step approach was better. For simplicity, we consider the cross-region covariance at one single time point, and assume the covariance of sensor noise $\boldsymbol{Q}_e = \boldsymbol{I}$ after pre-whitening. Let us also assume the latent activity of each ROI represents the mean activity across source points in the ROI, (i.e. $\boldsymbol{L}$ is a known 0/1 matrix), and the true source activity is generated according to our model (Figure 4.1 and Equation (4.6)). Then the marginal covariance of sensor data at time t across trials is $\text{cov}(\boldsymbol{y}_t) = \boldsymbol{I} + \boldsymbol{G}\boldsymbol{Q}_{J_n}\boldsymbol{G}' + \boldsymbol{G}\boldsymbol{L}\boldsymbol{Q}_t\boldsymbol{L}'\boldsymbol{G}'$ according to Equation (4.9).

If we use the two-step MNE method with a penalization parameter λ, then the estimated source activity in the first step is $\hat{\boldsymbol{J}}_t = (\boldsymbol{G}'\boldsymbol{G} + \lambda\boldsymbol{I})^{-1}\boldsymbol{G}'\boldsymbol{y}_t$. If $\boldsymbol{L}$ is known, we can further estimate $\boldsymbol{u}_t$ using the least square, $\hat{\boldsymbol{u}}_t = (\boldsymbol{L}'\boldsymbol{L})^{-1}\boldsymbol{L}'\hat{\boldsymbol{J}}_t = (\boldsymbol{L}'\boldsymbol{L})^{-1}\boldsymbol{L}'(\boldsymbol{G}'\boldsymbol{G} + \lambda\boldsymbol{I})^{-1}\boldsymbol{G}'\boldsymbol{y}_t$. In the second step, we compute the empirical covariance of $\hat{\boldsymbol{u}}_t$, which concentrates on the theoretical covariance below

$$\begin{aligned}
\text{cov}(\hat{\boldsymbol{u}}_t) &= (\boldsymbol{L}'\boldsymbol{L})^{-1}\boldsymbol{L}'(\boldsymbol{G}'\boldsymbol{G} + \lambda\boldsymbol{I})^{-1}\boldsymbol{G}'\text{cov}(\boldsymbol{y}_t)\boldsymbol{G}(\boldsymbol{G}'\boldsymbol{G} + \lambda\boldsymbol{I})^{-1}\boldsymbol{L}(\boldsymbol{L}'\boldsymbol{L})^{-1}\\
&= (\boldsymbol{L}'\boldsymbol{L})^{-1}\boldsymbol{L}'(\boldsymbol{G}'\boldsymbol{G} + \lambda\boldsymbol{I})^{-1}\boldsymbol{G}'(\boldsymbol{I} + \boldsymbol{G}\boldsymbol{Q}_{J_n}\boldsymbol{G}' + \boldsymbol{G}\boldsymbol{L}\boldsymbol{Q}_t\boldsymbol{L}'\boldsymbol{G}')\boldsymbol{G}(\boldsymbol{G}'\boldsymbol{G} + \lambda\boldsymbol{I})^{-1}\boldsymbol{L}(\boldsymbol{L}'\boldsymbol{L})^{-1}\\
&= (\boldsymbol{L}'\boldsymbol{L})^{-1}\boldsymbol{L}'(\boldsymbol{G}'\boldsymbol{G} + \lambda\boldsymbol{I})^{-1}\boldsymbol{G}'\boldsymbol{G}(\boldsymbol{G}'\boldsymbol{G} + \lambda\boldsymbol{I})^{-1}\boldsymbol{L}(\boldsymbol{L}'\boldsymbol{L})^{-1} \qquad (4.15)\\
&+ (\boldsymbol{L}'\boldsymbol{L})^{-1}\boldsymbol{L}'(\boldsymbol{G}'\boldsymbol{G} + \lambda\boldsymbol{I})^{-1}\boldsymbol{G}'\boldsymbol{G}\boldsymbol{Q}_{J_n}\boldsymbol{G}'\boldsymbol{G}(\boldsymbol{G}'\boldsymbol{G} + \lambda\boldsymbol{I})^{-1}\boldsymbol{L}(\boldsymbol{L}'\boldsymbol{L})^{-1} \qquad (4.16)\\
&+ (\boldsymbol{L}'\boldsymbol{L})^{-1}\boldsymbol{L}'(\boldsymbol{G}'\boldsymbol{G} + \lambda\boldsymbol{I})^{-1}\boldsymbol{G}'\boldsymbol{G}\boldsymbol{L}\boldsymbol{Q}_t\boldsymbol{L}'\boldsymbol{G}'\boldsymbol{G}(\boldsymbol{G}'\boldsymbol{G} + \lambda\boldsymbol{I})^{-1}\boldsymbol{L}(\boldsymbol{L}'\boldsymbol{L})^{-1} \qquad (4.17)
\end{aligned}$$

Let $\boldsymbol{G} = \boldsymbol{U}_G\boldsymbol{D}_G\boldsymbol{V}_G$ be the singular value decomposition of $\boldsymbol{G}$; that is, $\boldsymbol{U}_G$ has orthonormal columns, $\boldsymbol{V}_G$ has orthonormal rows, and $\boldsymbol{D}_G$ is a diagonal matrix, whose dimension is the rank of $\boldsymbol{G}$. Then $(\boldsymbol{G}'\boldsymbol{G} + \lambda\boldsymbol{I})^{-1}(\boldsymbol{G}'\boldsymbol{G}) = \boldsymbol{V}_G'\tilde{\boldsymbol{D}}_G\boldsymbol{V}$, where the entries in $\boldsymbol{D}_G$ are $\tilde{\boldsymbol{D}}_G[i,i] = \frac{(\boldsymbol{D}_G[i,i])^2}{(\boldsymbol{D}_G[i,i])^2+\lambda}$. When λ is small, $\tilde{\boldsymbol{D}}_G \approx \boldsymbol{I}$ and $(\boldsymbol{G}'\boldsymbol{G} + \lambda\boldsymbol{I})^{-1}(\boldsymbol{G}'\boldsymbol{G}) \approx \boldsymbol{I}$. In this case, the last term in $\text{cov}(\hat{\boldsymbol{u}}_t)$ (i.e., (4.17)) is approximately $\boldsymbol{Q}_t$, yet the first two terms ((4.15) and (4.16)) are biases due to the sensor noise covariance $\boldsymbol{Q}_e = \boldsymbol{I}$ and the source noise covariance $\boldsymbol{Q}_{J_n}$, which do not go to zero even with an infinite number of observations. When λ is relatively large, the third term (4.17) is a shrunken estimator of $\boldsymbol{Q}_t$ depending on $\boldsymbol{G}$ and $\boldsymbol{L}$, and the first two terms, although also shrunken, can still make $\text{cov}(\hat{\boldsymbol{u}}_t)$ relatively far from the true $\boldsymbol{Q}_t$.

In contrast, if we use no additional priors, our one-step approach solves for $\boldsymbol{Q}_t$ by maximizing the marginal likelihood of $\boldsymbol{y}_t$. If the number free parameters is not too large compared with the number of sensors and if the global optimum is found, the one-step solution is a statistically consistent estimator of $\boldsymbol{Q}_t$, and its efficiency is guaranteed by the Cramer-Rao bound [Wasserman, 2010]. Therefore, in our simulations, where the data was generated according to our source model, it was not surprising that the one-step methods gave better results. The similar arguments apply to the Kronecker-structured covariance case.

If our model assumptions about the latent ROI activity and the source activity are not met, the advantage of the one-step methods may be reduced. Some violations of the assumptions may have minor effects on the performance. For example, if all important ROIs are included, then we expect our one-step models to give reasonable results even if Equation (4.4) is not exactly correct. Given that $\boldsymbol{L}$ and σ_i^2 are allowed to vary, the models should be able to fit the data flexibly. In fact, in the simulations in Section 4.2.3, the true $\boldsymbol{L}$, which contained +1 and -1s, was hard to learn given our prior on $\boldsymbol{L}$, yet in our results, although the estimated entries in $\boldsymbol{L}$ was only weakly correlated with the truth (instead of being close to the truth), the one-step method still performed better than the two-step methods in terms of estimating the correlation structure across ROIs.

However, we do think there are other assumptions that are critical for the good performance of the one-step models. One of such assumptions is the correctness of ROI specification. If some ROIs that generate the true source activity is not modeled, our one-step models may have difficulty in even estimating the marginal connectivity of the included subset of ROIs. Assume there are p ROIs that contribute to the data, but only the first $p_1 < p$ ROIs are included in fitting the one-step models, and the remaining $p_2 = p - p_1$ ROIs are missing. For simplicity, let us consider the single-time-point case. Let $\boldsymbol{u}_{t1}$ and $\boldsymbol{u}_{t2}$ be the latent activity in the included p_1 ROIs and the missing p_2 ROIs respectively. Similarly, we partition the covariance of $\boldsymbol{u}_t$ in the following way.

$$\boldsymbol{Q}_t = \begin{pmatrix} \boldsymbol{Q}_{t11} & \boldsymbol{Q}_{t12} \\ \boldsymbol{Q}_{t21} & \boldsymbol{Q}_{t22} \end{pmatrix}$$

Let $\boldsymbol{L}_1$ and $\boldsymbol{L}_2$ be the corresponding columns in $\boldsymbol{L}$ for the p_1 and p_2 ROIs (i.e., $\underset{m\times p}{\boldsymbol{L}} = (\underset{m\times p_1}{\boldsymbol{L}_1}, \underset{m\times p_2}{\boldsymbol{L}_2})$). The covariance of the sensor data is

$$\begin{aligned}
\text{cov}(\boldsymbol{y}_t) &= \boldsymbol{Q}_e + \boldsymbol{G}\boldsymbol{Q}_{J_n}\boldsymbol{G}' + \boldsymbol{G}\boldsymbol{L}_1\boldsymbol{Q}_t\, \boldsymbol{L}'\boldsymbol{G}' \\
&= \boldsymbol{Q}_e + \sigma_0^2\boldsymbol{G}[:, l \in \mathcal{A}_0]\boldsymbol{G}[:, l \in \mathcal{A}_0]' + \sum_{i=1}^{p_1} \sigma_i^2\boldsymbol{G}[:, l \in \mathcal{A}_i]\boldsymbol{G}[:, l \in \mathcal{A}_i]' \\
&+ \sum_{i=p_1+1}^{p} \sigma_i^2\boldsymbol{G}[:, l \in \mathcal{A}_i]\boldsymbol{G}[:, l \in \mathcal{A}_i]' \qquad (4.18) \\
&+ \boldsymbol{G}\boldsymbol{L}_1\boldsymbol{Q}_{t11}\boldsymbol{L}_1'\boldsymbol{G}' \\
&+ \boldsymbol{G}(\boldsymbol{L}_1\boldsymbol{Q}_{t12}\boldsymbol{L}_2' + \boldsymbol{L}_2\boldsymbol{Q}_{t21}\boldsymbol{L}_1' + \boldsymbol{L}_2\boldsymbol{Q}_{t22}\boldsymbol{L}_2')\boldsymbol{G}' \qquad (4.19)
\end{aligned}$$

If we only model the p_1 ROIs, we replace the term (4.18) with $\sum_{i=p+1}^{p} \sigma_0^2\boldsymbol{G}[:, l \in \mathcal{A}_i]\boldsymbol{G}[:, l \in \mathcal{A}_i]'$, which might not cause large errors if $\sigma_0^2 \approx \sigma_2^2$. However, we also miss the term (4.19). Missing this term makes the marginal likelihood of sensor data incorrect, and the estimated covariance matrix across the p_1 included ROIs may be different from the true marginal covariance. In fact, only when $\boldsymbol{Q}_{t12}$ and $\boldsymbol{Q}_{t22}$ are close to zero, can the one-step results be close to the marginal covariance across the p_1 ROIs. In contrast, the two-step methods are less susceptible to missing ROIs, because they do not rely on the pre-defined ROIs in the source localization step.

4.3 Cross-region dynamic connectivity: a state-space model

In many cases, we want to estimate the time-lagged dependence in the connectivity analysis, that is, whether previous activity in a region can predict later activities in other regions. Such time-lagged connectivity can help us to generate hypotheses about how information flows in a dynamic way. However, the two one-step models we just introduced, which represent the connectivity with the covariance matrix across ROIs, are not designed to estimate time-lagged connectivity. The one-step model in the single-time-point case (in Section 4.2.1) considers the covariance at a fixed time point, ignoring time-lagged dependence. The second model we introduced (in Section 4.2.2) considers the multidimensional activities in all ROIs at all T time points in a trial. In this case, if the joint spatio-temporal covariance matrix can be estimated well with minimal constraints, it can reflect time-lagged dependence; however, for tractability, we assumed a Kronecker structure of the covariance, which limits the model's ability to describe time-lagged dependence. Moreover, in many cases, the brain activity during perceptual or cognitive tasks is not stationary, and the spatio-temporal dependence across brain regions may vary with time. However, the two models introduced in Section 4.2.1 and 4.2.2 are not designed to model time-varying connectivity. To tackle these issues, in this section, we introduce a state-space model, where the state variable is the ROI latent activity $\boldsymbol{u}_t$. This state variable is represented by a time-varying linear dynamical system, which can flexibly model non-stationary time-lagged connectivity. For simplicity, we assume that the ROI latent activity $\boldsymbol{u}_t$ represents the mean current dipoles of source points within individual ROIs. In other words, we assume that the $m \times p$ matrix $\boldsymbol{L}$ is a pre-defined "0/1" matrix that represents whether each source point is in or outside of each ROI, and we do not allow it to change when fitting the model.

4.3.1 Model formulation

Assuming that across q i.i.d. trials, the ROI activity (i.e., the state variable) $\boldsymbol{u}_t$ has a zero mean. This assumption is practically met if we subtract the mean across trials from the sensor observations. We use a time-varying autoregressive model of order 1 to describe the dynamics of the state variable $\boldsymbol{u}_t$, assuming at each time point $t = 0, 1, \cdots T$ [3].

$$\textbf{ROI model:} \quad \begin{aligned} &\boldsymbol{u}_0 \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{Q}_0) \\ &\boldsymbol{u}_t = \boldsymbol{A}_t \boldsymbol{u}_{t-1} + \boldsymbol{\epsilon}_t, \quad \boldsymbol{\epsilon}_t \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{Q}_\epsilon), \quad \text{for } t = 1, \cdots, T. \end{aligned} \tag{4.20}$$

Here, $\boldsymbol{Q}_0$ is a $p \times p$ covariance matrix of $\boldsymbol{u}_t$ at $t = 0$, and the $p \times p$ matrices $\boldsymbol{A}_t$s are the time-varying autoregressive coefficients, which describe lagged dependence across ROIs. The p-dimensional error term $\boldsymbol{\epsilon}_t$ is independent of the past, with a mean zero and a covariance matrix $\boldsymbol{Q}_\epsilon$.

At each time point t, the source model is the same as the one described in Equation (4.2) and Equation (4.3), or more concisely in Equation (4.4) (i.e., $\boldsymbol{J}_t = \boldsymbol{L}\boldsymbol{u}_t + \boldsymbol{w}_t$ and $\boldsymbol{w}_t \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{Q}_{J_n})$); given $\boldsymbol{J}_t$, the sensor data $\boldsymbol{y}_t$ is determined by the forward model ($\boldsymbol{y}_t = \boldsymbol{G}\boldsymbol{J}_t + \boldsymbol{e}_t$ where $\boldsymbol{e}_t \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{Q}_e)$). Together, as shown in Figure 4.4, our one-step state-space model includes the following conditional

[3] Here, slightly differently from the previous notation, we assume there are $T + 1$ time points in a trial and the time index t starts at 0.

distributions

$$\begin{gathered}
\boldsymbol{u}_0 \sim \mathcal{N}(\mathbf{0}, \boldsymbol{Q}_0) \\
(\boldsymbol{u}_t | \boldsymbol{u}_{t-1}) \sim \mathcal{N}(\boldsymbol{A}_t \boldsymbol{u}_{t-1}, \boldsymbol{Q}_\epsilon) \\
(\boldsymbol{J}_t | \boldsymbol{u}_t) \sim \mathcal{N}(\boldsymbol{L}\boldsymbol{u}_t, \boldsymbol{Q}_{J_n}) \\
(\boldsymbol{y}_t | \boldsymbol{J}_t) \sim \mathcal{N}(\boldsymbol{G}\boldsymbol{J}_t, \boldsymbol{Q}_e)
\end{gathered}$$

and by eliminating $\boldsymbol{J}_t$, we have the following time-varying state-space model

$$\begin{aligned}
&\boldsymbol{u}_0 \sim \mathcal{N}(\mathbf{0}, \boldsymbol{Q}_0) \\
&\boldsymbol{u}_t = \boldsymbol{A}_t \boldsymbol{u}_{t-1} + \boldsymbol{\epsilon}_t, \quad \boldsymbol{\epsilon}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{Q}_\epsilon) \\
&\boldsymbol{y}_t = \boldsymbol{G}\boldsymbol{L}\boldsymbol{u}_t + \boldsymbol{\eta}_t, \quad \boldsymbol{\eta}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{Q}_\eta) \quad \boldsymbol{Q}_\eta = \boldsymbol{Q}_e + \boldsymbol{G}\boldsymbol{Q}_{J_n}\boldsymbol{G}' + \boldsymbol{G}\boldsymbol{L}\boldsymbol{Q}_\epsilon\boldsymbol{L}'\boldsymbol{G}'.
\end{aligned} \tag{4.21}$$

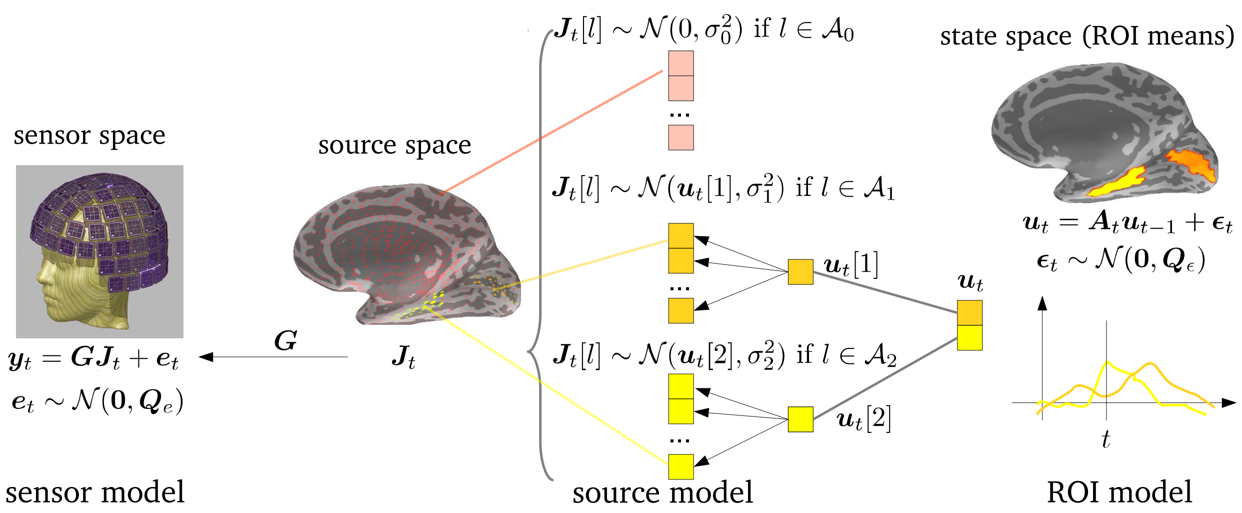

sensor space
source space
state space (ROI means)
$\boldsymbol{J}_t[l] \sim \mathcal{N}(0, \sigma_0^2)$ if $l \in \mathcal{A}_0$
$\boldsymbol{J}_t[l] \sim \mathcal{N}(\boldsymbol{u}_t[1], \sigma_1^2)$ if $l \in \mathcal{A}_1$
$\boldsymbol{J}_t[l] \sim \mathcal{N}(\boldsymbol{u}_t[2], \sigma_2^2)$ if $l \in \mathcal{A}_2$
$\boldsymbol{u}_t = \boldsymbol{A}_t \boldsymbol{u}_{t-1} + \boldsymbol{\epsilon}_t$
$\boldsymbol{\epsilon}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{Q}_\epsilon)$
$\boldsymbol{u}_t[1]$
$\boldsymbol{u}_t$
$\boldsymbol{u}_t[2]$
$\boldsymbol{G}$
$\boldsymbol{J}_t$
$\boldsymbol{y}_t = \boldsymbol{G}\boldsymbol{J}_t + \boldsymbol{e}_t$
$\boldsymbol{e}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{Q}_e)$
t
sensor model
source model
ROI model

Figure 4.4: Illustration of the state-space model of cross-region dynamic connectivity. There are two ROIs in this example.

4.3.2 Fitting the parameters using the expectation-maximization (EM) algorithm

In our time-varying state-space model (Equation (4.21)), $\boldsymbol{Q}_e$, $\boldsymbol{G}$ and $\boldsymbol{L}$ are given and fixed. We observe the sensor data from q i.i.d. trials, denoted by $\{\boldsymbol{y}_t^{(r)}\}_{t=0,r=1}^{T,q}$, where the superscript $\cdot^{(r)}$ corresponds to the rth trial. Given these, we need to estimate the unknown parameters, denoted by $\boldsymbol{\theta} = \{\{\boldsymbol{A}_t\}_{t=1}^T, \boldsymbol{Q}_0, \boldsymbol{Q}_\epsilon, \{\sigma_i^2\}_{i=0}^p\}$, where $\{\sigma_i^2\}_{i=0}^p$ determines $\boldsymbol{Q}_{J_n}$ (i.e., $\boldsymbol{Q}_{J_n}[l,l] = \sigma_i^2$ if $l \in \mathcal{A}_i$). Among the unknown parameters, we are mainly interested in $\{\boldsymbol{A}_t\}_{t=1}^T$, which describes the spatio-temporal dependence.

Let $f(\cdot)$ denote probability density functions in general. We can obtain $\boldsymbol{\theta}$ that maximizes the log-likelihood of the sensor data—$\arg\max_{\boldsymbol{\theta}} \log f\{\boldsymbol{y}_t^{(r)}\}_{t=0,r=1}^{T,q}; \boldsymbol{\theta}$, where $\{\boldsymbol{y}_t^{(r)}\}_{t=0,r=1}^{T,q}$ denotes the sensor time series across in all q trials. We can also add additional priors on $\boldsymbol{\theta}$ to regularize the problem. Here, because there can be many parameters in $\{\boldsymbol{A}_t\}_{t=1}^T$ when T is large, we add a prior

$$f(\boldsymbol{\theta}) = f(\{\boldsymbol{A}_t\}_{t=1}^T) \propto \exp(-(\lambda_0 \sum_{t=1}^T \|\boldsymbol{A}_t\|_F^2 + \lambda_1 \sum_{t=2}^T \|\boldsymbol{A}_t - \boldsymbol{A}_{t-1}\|_F^2)),$$

where $\|\cdot\|_F$ is the Frobenius norm of a matrix. This prior shrinks the entries of in $\boldsymbol{A}_t$s and also encourages temporal smoothness of $\boldsymbol{A}_t$s. The hyper parameters λ_0 and λ_1 can be either pre-specified or empirically chosen to achieve the largest cross-validated log-likelihood of the sensor data. Now our goal of parameter estimation is to maximize the objective function

$$\log f(\{\boldsymbol{y}_t^{(r)}\}_{t=0,r=1}^{T,q}; \boldsymbol{\theta}) + \log f(\boldsymbol{\theta}).$$

The state-space frame work allows us to exploit the commonly-used expectation-maximization (EM) algorithm [Shumway and Stoffer, 1982], which includes an E-step and an M-step in each iteration. The E-M algorithm uses the expectation of

$$\log f(\{\boldsymbol{y}_t^{(r)}\}_{t=0,r=1}^{T,q}, \{\boldsymbol{u}_t^{(r)}\}_{t=0,r=1}^{T,q}; \boldsymbol{\theta}) + \log f(\boldsymbol{\theta})$$

under the posterior distribution of $\{\boldsymbol{u}_t^{(r)}\}_{t=0,r=1}^{T,q}$ given $\{\boldsymbol{y}_t^{(r)}\}_{t=0,r=1}^{T,q}$, which is easy to compute, to approximate the objective function. Below we briefly describe the procedure. More details are in the appendix in Section 4.5.3.1.

In the E-step, given an estimate of the parameters (denoted by $\tilde{\boldsymbol{\theta}}$), we use the forward and backward steps in the Kalman smoothing algorithm [Shumway and Stoffer, 1982] to obtain the posterior mean of $\boldsymbol{u}_t$, $\boldsymbol{u}_{t|T}^{(r)} \stackrel{\text{def}}{=} \mathbb{E}(\boldsymbol{u}_t^{(r)} | \{\boldsymbol{y}_\tau^{(r)}\}_{\tau=0}^T)$, the posterior covariance of $\boldsymbol{u}_t$, $\boldsymbol{P}_{t|T}^{(r)} \stackrel{\text{def}}{=} \text{cov}(\boldsymbol{u}_t^{(r)} | \{\boldsymbol{y}_\tau^{(r)}\}_{\tau=0}^T)$, and the posterior cross covariance of $\boldsymbol{u}_t$ and $\boldsymbol{u}_{t-1}$, $\boldsymbol{P}_{(t,t-1)|T}^{(r)} \stackrel{\text{def}}{=} \text{cov}(\boldsymbol{u}_t^{(r)}, \boldsymbol{u}_{t-1}^{(r)} | \{\boldsymbol{y}_\tau^{(r)}\}_{\tau=0}^T)$, for each t in each trial r. Here $\mathbb{E}(\cdot)$ and $\text{cov}(\cdot)$ denote the expectation and the covariance. More details are in Appendix and [Shumway and Stoffer, 1982].

In the M-step, we maximize the expectation of $\log f(\{\boldsymbol{y}_t^{(r)}\}_{t=0,r=1}^{T,q}, \{\boldsymbol{u}_t^{(r)}\}_{t=0,r=1}^{T,q}; \boldsymbol{\theta}) + \log f(\boldsymbol{\theta})$, with respect to the posterior distribution $\tilde{f} \stackrel{\text{def}}{=} f(\{\boldsymbol{u}_t^{(r)}\}_{t=0,r=1}^{T,q} | \{\boldsymbol{y}_t^{(r)}\}_{t=0,r=1}^{T,q}; \tilde{\boldsymbol{\theta}})$. Let $\text{trace}(\cdot)$ and $\det(\cdot)$ denote the trace and the determinant of a matrix. Given results in the E-step based on $\tilde{\boldsymbol{\theta}}$, the

M-step is equivalent to minimizing three objectives separately

$$
\begin{aligned}
&\min_{\boldsymbol{\theta}}(-\mathbb{E}_{\tilde{f}}(\log f(\{\boldsymbol{y}_t^{(r)}\}_{t=0,r=1}^{T,q}, \{\boldsymbol{u}_t^{(r)}\}_{t=0,r=1}^{T,q}; \boldsymbol{\theta})) - \log f(\boldsymbol{\theta})) \\
&\equiv \min_{\boldsymbol{Q}_0} \mathcal{L}_1 + \min_{\boldsymbol{Q}_\epsilon, \{\boldsymbol{A}_t\}_{t=1}^T} \mathcal{L}_2 + \min_{\{\sigma_i^2\}_{i=0}^p} \mathcal{L}_3.
\end{aligned}
\tag{4.22}
$$

$$
\mathcal{L}_1(\boldsymbol{Q}_0) = q \log \det(\boldsymbol{Q}_0) + \text{trace}(\boldsymbol{Q}_0^{-1}\boldsymbol{B}_0) \quad \text{where } \boldsymbol{B}_0 = \sum\nolimits_{r=1}^q (\boldsymbol{P}_{0|T}^{(r)} + \boldsymbol{u}_{0|T}^{(r)}(\boldsymbol{u}_{0|T}^{(r)})') \tag{4.23}
$$

$$
\begin{aligned}
&\mathcal{L}_2(\boldsymbol{Q}_\epsilon, \{\boldsymbol{A}_t\}_{t=1}^T) = qT \log \det(\boldsymbol{Q}_\epsilon) + \text{trace}(\boldsymbol{Q}_\epsilon^{-1} \sum\nolimits_{t=1}^T (\boldsymbol{B}_{1t} - \boldsymbol{A}_t\boldsymbol{B}_{2t}' - \boldsymbol{B}_{2t}\boldsymbol{A}_t' + \boldsymbol{A}_t\boldsymbol{B}_{3t}\boldsymbol{A}_t')) \\
&+ \log f(\{\boldsymbol{A}_t\}_{t=1}^T)
\end{aligned}
\tag{4.24}
$$

$$
\begin{aligned}
&\text{where } \boldsymbol{B}_{1t} = \sum\nolimits_{r=1}^q (\boldsymbol{P}_{t|T}^{(r)} + \boldsymbol{u}_{t|T}^{(r)}(\boldsymbol{u}_{t|T}^{(r)})'), \quad \boldsymbol{B}_{2t} = \sum\nolimits_{r=1}^q (\boldsymbol{P}_{(t,t-1)|T}^{(r)} + \boldsymbol{u}_{t|T}^{(r)}(\boldsymbol{u}_{(t-1)|T}^{(r)})',) \\
&\boldsymbol{B}_{3t} = \sum\nolimits_{r=1}^q (\boldsymbol{P}_{(t-1)|T}^{(r)} + \boldsymbol{u}_{(t-1)|T}^{(r)}(\boldsymbol{u}_{(t-1)|T}^{(r)})')
\end{aligned}
$$

$$
\mathcal{L}_3(\{\sigma_i^2\}_{i=0}^p) = q(T+1) \log \det(\boldsymbol{Q}_\eta) + \text{trace}(\boldsymbol{Q}_\eta^{-1}\boldsymbol{B}_4), \quad \text{where } \boldsymbol{Q}_\eta = \boldsymbol{Q}_e + \boldsymbol{G}\boldsymbol{Q}_{J_n}\boldsymbol{G}', \tag{4.25}
$$

$$
\text{and } \boldsymbol{B}_4 = \sum\nolimits_{r=1}^q \sum\nolimits_{t=0}^T [(\boldsymbol{y}_t^{(r)} - \boldsymbol{C}\boldsymbol{u}_{t|T}^{(r)})(\boldsymbol{y}_t^{(r)} - \boldsymbol{C}\boldsymbol{u}_{t|T}^{(r)})' + \boldsymbol{C}\boldsymbol{P}_{t|T}^{(r)}\boldsymbol{C}')]
$$

The optimization for the three separate objectives is relatively easy.

- For $\mathcal{L}_1$, the analytical solution is $\boldsymbol{Q}_0 \leftarrow (1/q)(\boldsymbol{B_0})$.
- For $\mathcal{L}_2$, optimization for $\{\boldsymbol{A}_t\}_{t=1}^T$ and $\boldsymbol{Q}_\epsilon$ can be done in alternations. Given $\{\boldsymbol{A}_t\}_{t=1}^T$, $\boldsymbol{Q}_\epsilon$ has the analytical solution $\boldsymbol{Q}_\epsilon \leftarrow 1/(qT)\sum_{t=1}^T(\boldsymbol{B}_{1t} - \boldsymbol{A}_t\boldsymbol{B}_{2t}' - \boldsymbol{B}_{2t}\boldsymbol{A}_t' + \boldsymbol{A}_t\boldsymbol{B}_{3t}\boldsymbol{A}_t')$. Given $\boldsymbol{Q}_\epsilon$, we use gradient descent with back-tracking line search [Boyd and Vandenberghe, 2004] to solve for $\{\boldsymbol{A}_t\}_{t=1}^T$, where the gradients are $\frac{\partial \mathcal{L}_2}{\partial \boldsymbol{A}_t} = 2\boldsymbol{Q}_\epsilon^{-1}(-\boldsymbol{B}_{2t} + \boldsymbol{A}_t\boldsymbol{B}_{3t}) + 2\boldsymbol{D}_t$, $\boldsymbol{D}_t = \lambda_1(2\boldsymbol{A}_t - \boldsymbol{A}_{t+1} - \boldsymbol{A}_{t-1}) + \lambda_0\boldsymbol{A}_t$ for $t = 2, \cdots, T-1$, $\boldsymbol{D}_t = \lambda_1(\boldsymbol{A}_1 - \boldsymbol{A}_2) + \lambda_0\boldsymbol{A}_1$ for $t = 1$, and $\boldsymbol{D}_t = \lambda_1(\boldsymbol{A}_t - \boldsymbol{A}_{T-1}) + \lambda_0\boldsymbol{A}_t$ for $t = T$.
- For $\mathcal{L}_3$, we can also use gradient descent to solve for σ_i, with the gradient $\frac{\partial \mathcal{L}_3}{\partial \sigma_i} = \text{trace}((\frac{\partial \mathcal{L}_3}{\partial \boldsymbol{Q}_\eta})'\frac{\partial \boldsymbol{Q}_\eta}{\partial \sigma_i})$, where $\frac{\partial \mathcal{L}_3}{\partial \boldsymbol{Q}_\eta} = \boldsymbol{Q}_\eta^{-1} - \boldsymbol{Q}_\eta^{-1}\boldsymbol{B}_4\boldsymbol{Q}_\eta^{-1}$ and $\frac{\partial \boldsymbol{Q}_\eta}{\partial \sigma_i} = 2\sigma_i\boldsymbol{G}[:, l \in \mathcal{A}_i]\boldsymbol{G}[:, l \in \mathcal{A}_i]'$. Here $\boldsymbol{G}[:, l \in \mathcal{A}_i]$ denotes the columns in $\boldsymbol{G}$ corresponding to source points in the ith region.

Because the E-M algorithm only guarantees to find a local optimum, we use multiple initializations, and select the solution that yields the best objective function $\log f(\{\boldsymbol{y}_t^{(r)}\}_{t=0,r=1}^{T,q}) + \log f(\boldsymbol{\theta})$ (see the appendix on computing $\log f(\{\boldsymbol{y}_t^{(r)}\}_{t=0,r=1}^{T,q}; \boldsymbol{\theta})$).

4.3.3 Simulation

We compared this one-step state-space model with a two-step procedure using the MNE. In the first step, we obtained the MNE estimates of $\boldsymbol{J}_t$; then we averaged the MNE estimates for the source points within each ROI, at each time point and in each trial respectively, and treated the averages as an estimate of the ROI means ($\boldsymbol{u}_t$). In Step 2, according to the autoregressive model in Equation (4.20), we estimated $\boldsymbol{Q}_0$, $\{\boldsymbol{A}_t\}_{t=1}^T$ and $\boldsymbol{Q}_\epsilon$ by maximizing the sum of log-likelihood

($\log f(\{\boldsymbol{u}_t^{(r)}\}_{t=0,r=1}^{T,q})$) and the logarithm of the prior ($\log f(\{\boldsymbol{A}_t\}_{t=1}^{T})$); the maximization is very similar to the optimization for $\mathcal{L}_2$ in the M-step. Details are deferred to the appendix (Section 4.5.3.2).

We simulated MEG sensor data according to our model assumptions. The source space was defined as $m \approx 5000$ source points covering the bilateral cortical surfaces of a real brain, with 6.2 mm spacing on average, and $n = 306$ sensors were used. The sensor noise covariance matrix $\boldsymbol{Q}_e$ was estimated from real data. Two bilaterally merged ROIs were used: the pericalcarine area (ROI 1) and the parahippocampal gyri (ROI 2) (see Figure 4.5a). We generated the autoregressive coefficients for $T = 20$ time points, where for each $\boldsymbol{A}_t$, the diagonal entries were set to 0.5, and the off-diagonal entries were generated as a Morlet function multiplied by a random scaler drawn uniformly from the interval $(-1, 1)$ (see Figure 4.5b for an example). The covariances $\boldsymbol{Q}_0$ and $\boldsymbol{Q}_\epsilon$ were random positive definite matrices, whose diagonal entries were a constant a. The variances of source space noise $\{\sigma_i^2\}_{i=0}^{p}$ were randomly drawn from a Gamma distribution (shape= 2 and scale= 1). We used two different values, $a = 2$ and $a = 5$, where the relative strength of the ROI means compared with the variances of source-space noise ($\{\sigma_i^2\}_{i=0}^{p}$) were different. Each simulation had $q = 200$ trials, and 5 independent simulations for each a value were generated. The unit of the source activity was nanoampere meter (nAm).

When running the two-step method with the MNE for each simulation, a wide range of values of the penalization parameter (λ) were used. When fitting the state-space model, multiple initializations were used, including one of the two-step MNE estimates. In the prior of $\{\boldsymbol{A}_t\}_{t=1}^{T}$, we set $\lambda_0 = 0$ and $\lambda_1 = 0.1$.

For the fitted parameters $\{\boldsymbol{A}_t\}_{t=1}^{T}$ and $\boldsymbol{Q}_\epsilon$, we defined the relative error as the Frobenius norm of the difference between the estimate and the true parameter, divided by the Frobenius norm of the true parameter (e.g., for the true $\boldsymbol{Q}_\epsilon$ and the estimate $\hat{\boldsymbol{Q}}_\epsilon$, the relative error was $\|\hat{\boldsymbol{Q}}_\epsilon - \boldsymbol{Q}_\epsilon\|_F / \|\boldsymbol{Q}_\epsilon\|_F$). For different estimates by the two-step MNE method with different values of λ, the smallest relative error was selected for comparison. Figure 4.5c and 4.5d show the relative errors and paired differences in relative errors between the two methods, averaged across the independent simulations, for $\{\boldsymbol{A}_t\}_{t=1}^{T}$ and $\boldsymbol{Q}_\epsilon$. In these simulations, the state-space model yielded smaller estimation errors than the two-step method with MNE.

It is worth noting that the in the two-step method with the MNE, although the estimated $\boldsymbol{u}_t$ could be correlated with the true $\boldsymbol{u}_t$ in the simulations, the magnitudes were much smaller. For example, Figure 4.5f shows the true $\boldsymbol{u}_t$ (blue) in one trial in ROI 2 in one simulation ($\alpha = 5$), as well as the estimated $\boldsymbol{u}_t$ by the two-step MNE method (red), and the posterior estimate of $\boldsymbol{u}_t$ by the state-space model (green). The estimates from the two-step method with MNE were very close to zero. This could be due to the shrinkage effect of the L_2 penalty with a relatively large penalization parameter λ. However, using a smaller λ resulted in noisy estimations of $\boldsymbol{u}_t$, which were not correlated with the true $\boldsymbol{u}_t$.

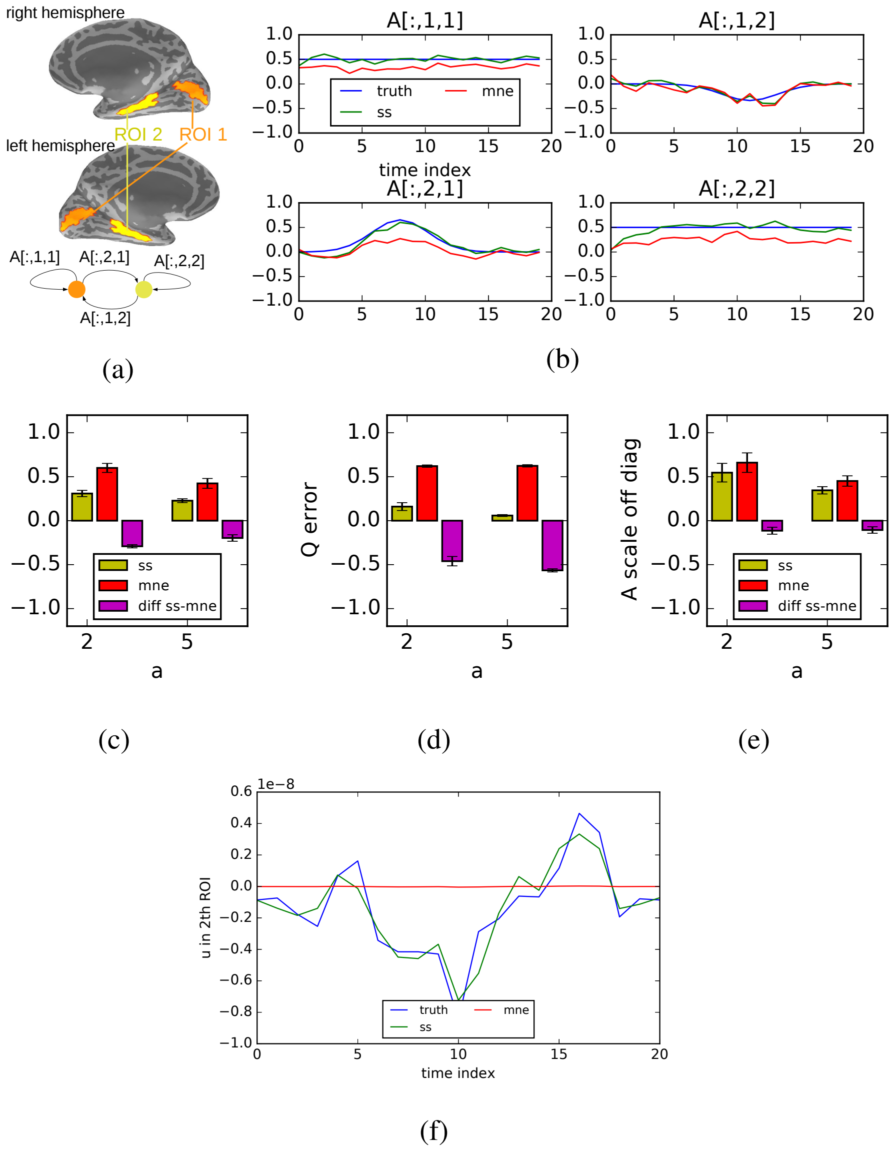


Figure 4.5: Simulation results. (a), Illustration of the two ROIs. Here $\boldsymbol{A}_t[i_1, i_2]$ represents the dependence of activity in ROI i_1 on the one-step-back activity on ROI i_2, for $i_1, i_2 = 1, 2$. (b), The autoregressive coefficients $\{\boldsymbol{A}_t\}_{t=1}^T$ of $T = 20$ time points in one example simulation ($a = 5$). Here $\boldsymbol{A}[:, i_1, i_2]$ indicates the time-varying coefficients in $\boldsymbol{A}_t[i_1, i_2]$, for $i_1, i_2 = 1, 2$. (The legend: *truth* (blue), true values; *ss* (green), estimates by the state-space model; *mne* (red), estimates by the two-step MNE method.) (c) and (d), Comparison of the state-space model (*ss*) with the two-step MNE method (*mne*) in relative errors of $\{\boldsymbol{A}_t\}_{t=1}^T$ (c) and $\boldsymbol{Q}_\epsilon$ (d). The yellow and the red bars show the averaged relative errors across the simulations, by the one-step state-space method and the two-step method with the MNE. The magenta bar shows the averaged paired difference between the two methods . The error bars show standard errors across the simulations. (e), The relative errors of AR coefficients after scaling, averaged within the off-diagonal entries and averaged across the simulations. (f), Visualization of $\boldsymbol{u}_t$ in ROI 2 in one trial from one simulation. (The legend: *truth* (blue), true values; *ss* (green), estimates by the state-space model; *mne* (red), estimates by the two-step MNE method.)

Although the autoregressive (AR) coefficients ($\boldsymbol{A}_t$s) should not be very sensitive to the magnitudes of $\boldsymbol{u}_t$, when the shrinkage effect is different for each ROI, the off-diagonal entries in $\boldsymbol{A}_t$s, which describe the dependence between different ROIs, might be affected. In contrast, the diagonal entries, which describe the dependence of the ROI activity on the one-step-back history of itself, are unlikely to be affected by the magnitudes of the estimates. To give the two-step method with MNE some advantage to correct for the shrinkage, we also computed a different measurement of relative error. Let $\boldsymbol{A}_t$ denote the true AR coefficients. For each pair of ROIs (i, j), we multiply the estimate ($\hat{\boldsymbol{A}}_t$) with a scalar ρ, such that $\sum_{t=1}^{T}(\boldsymbol{A}_t[i,j] - \rho\hat{\boldsymbol{A}}_t[i,j])^2$ was minimized, and then we defined the relative error as

$$\sqrt{\frac{\sum_{t=1}^{T}(\boldsymbol{A}_t[i,j] - \alpha\hat{\boldsymbol{A}}_t[i,j])^2}{\sum_{t=1}^{T}(\boldsymbol{A}_t[i,j])^2}}.$$

We took the average of the relative error after such scaling, for the off-diagonal (i, j) pairs (i.e. $i \neq j$) (Figure 4.5e), and still observed that the state-space model yielded smaller errors than the two-step method with the MNE.

4.4 Discussion

In this chapter, we have presented novel one-step models to directly estimate functional connectivity across pre-defined ROIs from sensor data. Compared with a two-step method using the MNE and the same connectivity formulation (either through the covariance matrix or an autoregressive model), the one-step models gave better results than the two-step method on simulated data, where the assumptions about the relationship between the ROI activity and source activity were reasonably met.

It is worth noting that there are several practical issues about the one-step models. First, since the optimization problem is not necessarily convex, the algorithms introduced here only guarantee local optima; in this context, it is important to use good initialization. For example, we can use the solutions in the two-step methods in the initialization or use multiple random initial values. Secondly, the optimization problem is more difficult than that in the two-step approach; as a result, our current implementation of the one-step models was computationally slower than that of the two-step method; we look forward to speed up the algorithms with parallel computing in future work.

One limitation of the work here is that we did not compare with other one-step models [David et al., 2006; Fukushima et al., 2015]. In future work, we plan to perform more comprehensive empirical evaluations of the available one-step methods. Another issue is there can be violations of our model assumptions in practice. First, given the ROI means, the noise on source points could be spatially and temporally correlated, rather than independently distributed. Secondly, as discussed in Section 4.2.4, if we fail to include an important ROI, the connectivity estimates may be inaccurate—the estimates may not even be equivalent to the estimates when this ROI is marginalized out, due to the underdetermined nature of source localization. Thirdly, the assumption that source points within an ROI share a common mean is typically correct for small ROIs but could be less accurate for larger ROIs, where the diverse activities of many source points might not be well-represented by a one-dimensional latent activity. That being said, as long as the activity in different source points within the ROI is not fully canceled, positive dependence effects of the kind identified by our models would still be meaningful in the sense that they reflect some cross-region dependence. To deal with the last two issues, one may divide the entire source space into sufficiently small, non-overlapping ROIs, when applying our state-space model. In such cases, the number of parameters can be large, and some sparsity-inducing regularization (such as the one in [Fukushima et al., 2015]) can be applied. In ongoing and future work, we plan to explore this idea and also address the effects of potential assumption violations.

4.5 Appendix

4.5.1 Appendix of Section 4.2.1

4.5.1.1 Eliminating the source current dipoles

Here we briefly derive the steps to get Equation (4.7) from Equation (4.6).

$$
\begin{aligned}
f(\boldsymbol{y}_t|\boldsymbol{u}_t) &\propto \int f(\boldsymbol{y}_t|\boldsymbol{J}_t) f(\boldsymbol{J}_t|\boldsymbol{u}_t) d\boldsymbol{J}_t \\
&\propto \int \exp(-\frac{1}{2}((\boldsymbol{y}_t - \boldsymbol{G}\boldsymbol{J}_t)'\boldsymbol{Q}_e^{-1}(\boldsymbol{y}_t - \boldsymbol{G}\boldsymbol{J}_t) + (\boldsymbol{J}_t - \boldsymbol{L}\boldsymbol{u}_t)'\boldsymbol{Q}_{J_n}^{-1}(\boldsymbol{J}_t - \boldsymbol{L}\boldsymbol{u}_t)))d\boldsymbol{J}_t
\end{aligned}
$$

Since the conditional distributions ($f(\boldsymbol{y}_t|\boldsymbol{J}_t)$ and $f(\boldsymbol{y}_t|\boldsymbol{J}_t)$) are Gaussian distributions, the resulting $f(\boldsymbol{y}_t|\boldsymbol{u}_t)$ is also a Gaussian distribution, and we only need to find the mean and covariance terms for $\boldsymbol{y}_t$. To do this, we consider the quadratic terms within the “exp”. We first complete the square for $\boldsymbol{J}_t$ first, and then drop the terms related to $\boldsymbol{J}_t$, for which the integral is a constant irrelevant to $\boldsymbol{y}_t$. Finally by completing the square for $\boldsymbol{y}_t$, we have the covariance term $(\boldsymbol{Q}_e + \boldsymbol{G}\boldsymbol{Q}_{J_n}\boldsymbol{G}')$ and the mean term $\boldsymbol{G}\boldsymbol{L}\boldsymbol{u}_t$.

$$
\begin{aligned}
&(\boldsymbol{y}_t - \boldsymbol{G}\boldsymbol{J}_t)'\boldsymbol{Q}_e^{-1}(\boldsymbol{y}_t - \boldsymbol{G}\boldsymbol{J}_t) + (\boldsymbol{J}_t - \boldsymbol{L}\boldsymbol{u}_t)'\boldsymbol{Q}_{J_n}^{-1}(\boldsymbol{J}_t - \boldsymbol{L}\boldsymbol{u}_t) \\
=&\boldsymbol{J}_t'(\boldsymbol{G}'\boldsymbol{Q}_e^{-1}\boldsymbol{G} + \boldsymbol{Q}_{J_n}^{-1})\boldsymbol{J}_t - 2(\boldsymbol{y}_t'\boldsymbol{Q}_e^{-1}\boldsymbol{G} + \boldsymbol{u}_t'\boldsymbol{L}'\boldsymbol{Q}_{J_n}^{-1})\boldsymbol{J}_t + \boldsymbol{y}_t'\boldsymbol{Q}_e^{-1}\boldsymbol{y}_t + \text{a term irrelvent to } \boldsymbol{y}_t \text{ or } \boldsymbol{J}_t \\
=&\boldsymbol{y}_t'\boldsymbol{Q}_e^{-1}\boldsymbol{y}_t - (\boldsymbol{G}'\boldsymbol{Q}_e^{-1}\boldsymbol{y}_t + \boldsymbol{Q}_{J_n}^{-1}\boldsymbol{L}\boldsymbol{u}_t)'(\boldsymbol{G}'\boldsymbol{Q}_e^{-1}\boldsymbol{G} + \boldsymbol{Q}_{J_n}^{-1})^{-1}(\boldsymbol{G}'\boldsymbol{Q}_e^{-1}\boldsymbol{y}_t + \boldsymbol{Q}_{J_n}^{-1}\boldsymbol{L}\boldsymbol{u}_t) \\
&+(\boldsymbol{J}_t - \boldsymbol{\mu}_0)'(\boldsymbol{G}'\boldsymbol{Q}_e^{-1}\boldsymbol{G} + \boldsymbol{Q}_{J_n}^{-1})(\boldsymbol{J}_t - \boldsymbol{\mu}_0)(\text{can be integrated out}) + \text{a term irrelvent to } \boldsymbol{y}_t
\end{aligned}
$$

where $\boldsymbol{u}_0 = (\boldsymbol{G}'\boldsymbol{Q}_e^{-1}\boldsymbol{G} + \boldsymbol{Q}_{J_n}^{-1})^{-1}(\boldsymbol{G}'\boldsymbol{Q}_e^{-1}\boldsymbol{y}_t + \boldsymbol{Q}_{J_n}^{-1}\boldsymbol{L}\boldsymbol{u}_t)$. Consider the first term that is relevant to $\boldsymbol{y}_t$

$$\begin{aligned}
&\boldsymbol{y}_t'\boldsymbol{Q}_e^{-1}\boldsymbol{y}_t - (\boldsymbol{G}'\boldsymbol{Q}_e^{-1}\boldsymbol{y}_t + \boldsymbol{Q}_{J_n}^{-1}\boldsymbol{L}\boldsymbol{u}_t)'(\boldsymbol{G}'\boldsymbol{Q}_e^{-1}\boldsymbol{G} + \boldsymbol{Q}_{J_n}^{-1})^{-1}(\boldsymbol{G}'\boldsymbol{Q}_e^{-1}\boldsymbol{y}_t + \boldsymbol{Q}_{J_n}^{-1}\boldsymbol{L}\boldsymbol{u}_t)\\
=&\boldsymbol{y}_t'(\boldsymbol{Q}_e^{-1} - \boldsymbol{Q}_e^{-1}\boldsymbol{G}'(\boldsymbol{G}'\boldsymbol{Q}_e^{-1}\boldsymbol{G} + \boldsymbol{Q}_{J_n}^{-1})^{-1}\boldsymbol{G}\boldsymbol{Q}_e^{-1})\boldsymbol{y}_t + \boldsymbol{y}_t'\boldsymbol{Q}_e^{-1}\boldsymbol{G}(\boldsymbol{G}'\boldsymbol{Q}_e^{-1}\boldsymbol{G} + \boldsymbol{Q}_{J_n}^{-1})^{-1}\boldsymbol{Q}_{J_n}^{-1}\boldsymbol{L}\boldsymbol{u}_t\\
&+ \text{a term irrelevant to } \boldsymbol{y}_t\\
=&\boldsymbol{y}_t'(\boldsymbol{Q}_e + \boldsymbol{G}\boldsymbol{Q}_{J_n}\boldsymbol{G}')^{-1}\boldsymbol{y}_t + \boldsymbol{y}_t'\boldsymbol{Q}_e^{-1}\boldsymbol{G}(\boldsymbol{G}'\boldsymbol{Q}_e^{-1}\boldsymbol{G} + \boldsymbol{Q}_{J_n}^{-1})^{-1}\boldsymbol{Q}_{J_n}^{-1}\boldsymbol{L}\boldsymbol{u}_t \quad \text{(using matrix inversion lemma)}\\
&+ \text{a term irrelevant to } \boldsymbol{y}_t\\
=&(\boldsymbol{y}_t - \boldsymbol{\mu}_1)'(\boldsymbol{Q}_e + \boldsymbol{G}\boldsymbol{Q}_{J_n}\boldsymbol{G}')^{-1}(\boldsymbol{y}_t - \boldsymbol{\mu}_1)' + \text{ a term irrelvent to } \boldsymbol{y}_t
\end{aligned}$$

The covariance term is $(\boldsymbol{Q}_e + \boldsymbol{G}\boldsymbol{Q}_{J_n}\boldsymbol{G}')$; the mean term is

$$\begin{aligned}
&\boldsymbol{\mu}_1 = (\boldsymbol{Q}_e + \boldsymbol{G}\boldsymbol{Q}_{J_n}\boldsymbol{G}')\boldsymbol{Q}_e^{-1}\boldsymbol{G}(\boldsymbol{G}'\boldsymbol{Q}_e^{-1}\boldsymbol{G} + \boldsymbol{Q}_{J_n}^{-1})^{-1}\boldsymbol{Q}_{J_n}^{-1}\boldsymbol{L}\boldsymbol{u}_t\\
=&(\boldsymbol{Q}_e + \boldsymbol{G}\boldsymbol{Q}_{J_n}\boldsymbol{G}')\boldsymbol{Q}_e^{-1}\boldsymbol{G}(\boldsymbol{Q}_{J_n} - \boldsymbol{Q}_{J_n}\boldsymbol{G}'(\boldsymbol{G}\boldsymbol{Q}_{J_n}\boldsymbol{G}' + \boldsymbol{Q}_e)^{-1}\boldsymbol{G}\boldsymbol{Q}_{J_n})\boldsymbol{Q}_{J_n}^{-1}\boldsymbol{L}\boldsymbol{u}_t\\
&(\text{ expanding } (\boldsymbol{G}'\boldsymbol{Q}_e^{-1}\boldsymbol{G} + \boldsymbol{Q}_{J_n}^{-1})^{-1} \text{ using matrix inversion lamma})\\
=&(\boldsymbol{Q}_e + \boldsymbol{G}\boldsymbol{Q}_{J_n}\boldsymbol{G}')\left(\boldsymbol{Q}_e^{-1}\boldsymbol{G}\boldsymbol{L}\boldsymbol{u}_t - \boldsymbol{Q}_e^{-1}(\boldsymbol{G}\boldsymbol{Q}_{J_n}\boldsymbol{G}')(\boldsymbol{G}\boldsymbol{Q}_{J_n}\boldsymbol{G}' + \boldsymbol{Q}_e)^{-1}\boldsymbol{G}\boldsymbol{L}\boldsymbol{u}_t\right)\\
=&(\boldsymbol{Q}_e + \boldsymbol{G}\boldsymbol{Q}_{J_n}\boldsymbol{G}')\boldsymbol{Q}_e^{-1}(\boldsymbol{I} - (\boldsymbol{G}\boldsymbol{Q}_{J_n}\boldsymbol{G}')(\boldsymbol{Q}_e + \boldsymbol{G}\boldsymbol{Q}_{J_n}\boldsymbol{G}')^{-1})\boldsymbol{G}\boldsymbol{L}\boldsymbol{u}_t\\
=&(\boldsymbol{Q}_e + \boldsymbol{G}\boldsymbol{Q}_{J_n}\boldsymbol{G}')\boldsymbol{Q}_e^{-1}((\boldsymbol{Q}_e + \boldsymbol{G}\boldsymbol{Q}_{J_n}\boldsymbol{G}') - (\boldsymbol{G}\boldsymbol{Q}_{J_n}\boldsymbol{G}'))(\boldsymbol{Q}_e + \boldsymbol{G}\boldsymbol{Q}_{J_n}\boldsymbol{G}')^{-1}\boldsymbol{G}\boldsymbol{L}\boldsymbol{u}_t\\
&(\text{ by plugging in } \boldsymbol{I} = (\boldsymbol{Q}_e + \boldsymbol{G}\boldsymbol{Q}_{J_n}\boldsymbol{G}')(\boldsymbol{Q}_e + \boldsymbol{G}\boldsymbol{Q}_{J_n}\boldsymbol{G}')^{-1})\\
=&(\boldsymbol{Q}_e + \boldsymbol{G}\boldsymbol{Q}_{J_n}\boldsymbol{G}')\boldsymbol{Q}_e^{-1}\boldsymbol{Q}_e(\boldsymbol{Q}_e + \boldsymbol{G}\boldsymbol{Q}_{J_n}\boldsymbol{G}')^{-1}\boldsymbol{G}\boldsymbol{L}\boldsymbol{u}_t\\
=&\boldsymbol{G}\boldsymbol{L}\boldsymbol{u}_t
\end{aligned}$$

Similar derivations also work to obtain Equation (4.9) from Equation (4.7).

4.5.1.2 Gradients of the negative log-likelihood function

Given the negative log-likelihood function defined in Equation (4.10), we prove the correctness of the gradients in Equation (4.11). The first equation on $\frac{\partial \text{llh}}{\partial \boldsymbol{Q}_{yt}}$ is easily obtained according to "the matrix cookbook" [Petersen and Pedersen, 2008]. Next we consider $\frac{\partial \text{llh}}{\partial \boldsymbol{\Gamma}_t}$. Let i, j, k, l, s and w be indices of rows or columns of matrices. Using the chain rule of matrix derivatives in "the matrix cookbook" [Petersen and Pedersen, 2008], we have

$$\frac{\partial \text{llh}}{\partial \boldsymbol{\Gamma}_t[s,w]} = \text{trace}\left((\frac{\partial \text{llh}}{\partial \boldsymbol{Q}_{yt}})'(\frac{\partial \boldsymbol{Q}_{yt}}{\partial \boldsymbol{\Gamma}_t[s,w]}) \right) = \sum_{i,j} \left(\frac{\partial \text{llh}}{\partial \boldsymbol{Q}_{yt}} \right)[i,j] \frac{\partial \boldsymbol{Q}_{yt}[i,j]}{\partial \boldsymbol{\Gamma}_t[s,w]}$$

Let $\boldsymbol{C} = \boldsymbol{GL}$; then $\boldsymbol{Q}_{yt} = \boldsymbol{C}\boldsymbol{\Gamma}_t\boldsymbol{\Gamma}_t'\boldsymbol{C}' +$ terms irrelvent to $\boldsymbol{\Gamma}_t$.

$$\begin{aligned}
\boldsymbol{Q}_{yt}[i,j] &= \sum_k(\sum_l \boldsymbol{C}[i,l](\sum_h \boldsymbol{\Gamma}_t[l,h]\boldsymbol{\Gamma}_t[k,h])\boldsymbol{C}[j,k]) + \text{terms irrelvent to } \boldsymbol{\Gamma}_t \\
&= \sum_k\sum_l\sum_h \boldsymbol{C}[j,k]\boldsymbol{C}[i,l]\boldsymbol{\Gamma}_t[l,h]\boldsymbol{\Gamma}_t[k,h] + \text{terms irrelvent to } \boldsymbol{\Gamma}_t \\
\frac{\partial \boldsymbol{Q}_{yt}[i,j]}{\partial \boldsymbol{\Gamma}_t[s,w]} &= 2\boldsymbol{C}[j,s]\boldsymbol{C}[i,s]\boldsymbol{\Gamma}_t[s,w] + \sum_{k \neq s}\boldsymbol{C}[j,k]\boldsymbol{C}[i,s]\boldsymbol{\Gamma}_t[k,w] + \sum_{l \neq s}\boldsymbol{C}[j,s]\boldsymbol{C}[i,l]\boldsymbol{\Gamma}_t[l,w] \\
&= \sum_k \boldsymbol{C}[j,k]\boldsymbol{C}[i,s]\boldsymbol{\Gamma}_t[k,w] + \sum_l \boldsymbol{C}[j,s]\boldsymbol{C}[i,l]\boldsymbol{\Gamma}_t[l,w] \\
&\text{(merging the first term into the last two terms)} \\
&= \sum_k \boldsymbol{C}[j,k]\boldsymbol{C}[i,s]\boldsymbol{\Gamma}_t[k,w] + \sum_k \boldsymbol{C}[j,s]\boldsymbol{C}[i,k]\boldsymbol{\Gamma}_t[k,w] \\
&\text{(replacing } l \text{ with } k \text{ in the second term)}
\end{aligned}$$

According to the chain rule, we have

$$\begin{aligned}
\frac{\partial \text{nllh}}{\partial \boldsymbol{\Gamma}_t[s,w]} &= \sum_{ij}\left(\frac{\partial \text{llh}}{\partial \boldsymbol{Q}_{yt}}\right)[i,j]\frac{\partial \boldsymbol{Q}_Y[i,j]}{\partial \boldsymbol{\Gamma}_t[s,w]} \\
&= \sum_{ij}\left(\frac{\partial \text{llh}}{\partial \boldsymbol{Q}_{yt}}\right)[i,j]\sum_k \boldsymbol{C}[j,k]\boldsymbol{C}[i,s]\boldsymbol{\Gamma}_t[k,w] + \sum_{ij}\left(\frac{\partial \text{llh}}{\partial \boldsymbol{Q}_{yt}}\right)[i,j]\sum_k \boldsymbol{C}[j,s]\boldsymbol{C}[i,k]\boldsymbol{\Gamma}_t[k,w] \\
&= \sum_{ij}\left(\frac{\partial \text{llh}}{\partial \boldsymbol{Q}_{yt}}\right)[i,j]\sum_k \boldsymbol{C}[j,k]\boldsymbol{C}[i,s]\boldsymbol{\Gamma}_t[k,w] + \sum_{ji}\left(\frac{\partial \text{llh}}{\partial \boldsymbol{Q}_{yt}}\right)[j,i]\sum_k \boldsymbol{C}[i,s]\boldsymbol{C}[j,k]\boldsymbol{\Gamma}_t[k,w] \\
&\text{(swapping } i \text{ and } j \text{ in the second term)} \\
&= 2\sum_{ij}\left(\frac{\partial \text{llh}}{\partial \boldsymbol{Q}_{yt}}\right)[i,j]\sum_k \boldsymbol{C}[j,k]\boldsymbol{C}[i,s]\boldsymbol{\Gamma}_t[k,w] \\
&\text{(merging the two terms, noticing that } \left(\frac{\partial \text{llh}}{\partial \boldsymbol{Q}_{yt}}\right) \text{ is symmetric)}
\end{aligned}$$

This term is equal to the (s,w) entry in $2\boldsymbol{C}'\left(\frac{\partial \text{llh}}{\partial \boldsymbol{Q}_{yt}}\right)\boldsymbol{C}\boldsymbol{\Gamma}_t$, which is $2\sum_k\sum_i\sum_j \boldsymbol{C}[i,s]\left(\frac{\partial \text{llh}}{\partial \boldsymbol{Q}_{yt}}\right)[i,j]\boldsymbol{C}[j,k]\boldsymbol{\Gamma}_t[k,w$ so we can verify that the following in Equation (4.11) holds

$$\frac{\partial \text{llh}}{\partial \boldsymbol{\Gamma}_t} = 2\boldsymbol{C}'\left(\frac{\partial \text{llh}}{\partial \boldsymbol{Q}_{yt}}\right)\boldsymbol{C}\boldsymbol{\Gamma}_t = 2\boldsymbol{L}'\boldsymbol{G}'\left(\frac{\partial \text{llh}}{\partial \boldsymbol{Q}_{yt}}\right)\boldsymbol{GL\Gamma_t}$$

We can verify the gradient with respect to $\boldsymbol{L}$ in Equation (4.11) similarly. Let i, j, l, k, h, r, w and

s be indices of rows and columns of matrices. We also note that $\boldsymbol{Q}_t = \boldsymbol{\Gamma}_t \boldsymbol{\Gamma}_t'$.

$$\boldsymbol{Q}_{yt}[i,j] = \sum_l \sum_k \sum_h \boldsymbol{G}[i,k]\boldsymbol{L}[k,l]\boldsymbol{Q}_t[l,h] \sum_r \boldsymbol{G}[j,r]\boldsymbol{L}[r,h] + \text{terms irrelevant to } \boldsymbol{L}$$

$$\begin{aligned}
\frac{\partial \boldsymbol{Q}_{yt}[i,j]}{\partial \boldsymbol{L}[s,w]} &= 2\boldsymbol{G}[i,s]\boldsymbol{Q}_t[w,w]\boldsymbol{G}[j,s]\boldsymbol{L}[s,w] \\
&+ \sum_{r\neq s \& h\neq w} \boldsymbol{G}[i,s]\boldsymbol{Q}_t[w,h]\boldsymbol{G}[j,r]\boldsymbol{L}[r,h] + \sum_{k\neq s \& l\neq w} \boldsymbol{G}[i,k]\boldsymbol{Q}_t[l,w]\boldsymbol{G}[j,s]\boldsymbol{L}[k,l] \\
&= \sum_{rh} \boldsymbol{G}[i,s]\boldsymbol{Q}_t[w,h]\boldsymbol{G}[j,r]\boldsymbol{L}[r,h] + \sum_{kl} \boldsymbol{G}[i,k]\boldsymbol{Q}_t[l,w]\boldsymbol{G}[j,s]\boldsymbol{L}[k,l] \\
&\text{(merging the first term into the last two terms)} \\
&= \sum_{rh} \boldsymbol{G}[i,s]\boldsymbol{Q}_t[w,h]\boldsymbol{G}[j,r]\boldsymbol{L}[r,h] + \sum_{rh} \boldsymbol{G}[i,r]\boldsymbol{Q}_t[h,w]\boldsymbol{G}[j,s]\boldsymbol{L}[r,h] \\
&\text{(replacing } (k,l) \text{ with } (r,h) \text{ in the second term)}
\end{aligned}$$

$$\begin{aligned}
\frac{\partial \text{llh}}{\partial \boldsymbol{L}[s,w]} &= \sum_{ij} \left(\frac{\partial \text{llh}}{\partial \boldsymbol{Q}_{yt}}\right)[i,j] \sum_{rh} \boldsymbol{G}[i,s]\boldsymbol{Q}_t[w,h]\boldsymbol{G}[j,r]\boldsymbol{L}[r,h] \\
&+ \sum_{ij} \left(\frac{\partial \text{llh}}{\partial \boldsymbol{Q}_{yt}}\right)[i,j] \sum_{rh} \boldsymbol{G}[i,r]\boldsymbol{Q}_t[h,w]\boldsymbol{G}[j,s]\boldsymbol{L}[r,h] \\
&= \sum_{ij} \left(\frac{\partial \text{llh}}{\partial \boldsymbol{Q}_{yt}}\right)[i,j] \sum_{rh} \boldsymbol{G}[i,s]\boldsymbol{Q}_t[w,h]\boldsymbol{G}[j,r]\boldsymbol{L}[r,h] \\
&+ \sum_{ji} \left(\frac{\partial \text{llh}}{\partial \boldsymbol{Q}_{yt}}\right)[j,i] \sum_{rh} \boldsymbol{G}[j,r]\boldsymbol{Q}_t[h,w]\boldsymbol{G}[i,s]\boldsymbol{L}[r,h] \\
&\text{(swapping } i \text{ and } j \text{ in the second term; noticing that } \left(\frac{\partial \text{llh}}{\partial \boldsymbol{Q}_{yt}}\right) \text{ and } \boldsymbol{Q}_t \text{ are symmetric)} \\
&= 2\sum_{ij} \left(\frac{\partial \text{llh}}{\partial \boldsymbol{Q}_{yt}}\right)[i,j] \sum_{rh} \boldsymbol{G}[i,s]\boldsymbol{Q}_t[w,h]\boldsymbol{G}[j,r]\boldsymbol{L}[r,h]
\end{aligned}$$

This term is equal to the (s,w) entry in $2\boldsymbol{G}'\left(\frac{\partial \text{llh}}{\partial \boldsymbol{Q}_{yt}}\right)\boldsymbol{G}\boldsymbol{L}\boldsymbol{Q}_t$, which is

$$2\sum_{i,j,r,h} \boldsymbol{G}[i,s]\left(\frac{\partial \text{llh}}{\partial \boldsymbol{Q}_{yt}}\right)[i,j]\boldsymbol{G}[j,r]\boldsymbol{L}[r,h]\boldsymbol{Q}_t[h,w],$$

so we can verify that the following in Equation (4.11) holds

$$\frac{\partial \text{llh}}{\partial \boldsymbol{L}} = 2\boldsymbol{G}'\left(\frac{\partial \text{llh}}{\partial \boldsymbol{Q}_{yt}}\right)\boldsymbol{G}\boldsymbol{L}\boldsymbol{\Gamma}_t\boldsymbol{\Gamma}_t' = 2\boldsymbol{G}'\left(\frac{\partial \text{llh}}{\partial \boldsymbol{Q}_{yt}}\right)\boldsymbol{G}\boldsymbol{L}\boldsymbol{Q}_t$$

Note that in the actual updates during gradient descent, there are entries in $\boldsymbol{L}$ that are fixed as zero ($\boldsymbol{L}[l,i] = 0$ if $l \notin \mathcal{A}_i, l = 1,\cdots,m$ for each $i = 1,\cdots,p$). We only update $\boldsymbol{L}[l,i]$s for all $l \in \mathcal{A}_i, l = 1,\cdots,m$ for each $i = 1,\cdots,p$. Finally, the gradient with respect to σ_i in Equation (4.11) is obtained by directly applying the chain rule.

4.5.1.3 The penalty functions due to the priors on the parameters and the corresponding gradients

Here, we describe the penalties due to the priors on $\boldsymbol{Q}_t$ (or $\boldsymbol{\Gamma}_t$), $\boldsymbol{Q}_{J_n}$ (or each σ_i^2), and the non-zero entries of $\boldsymbol{L}$ (i.e., $\boldsymbol{L}[l,i]$s for all $l \in \mathcal{A}_i, l = 1, \cdots, m$ for each $i = 1, \cdots, p$), as well as the corresponding gradient terms.

1. We can use a Wishart prior on $\boldsymbol{Q}_t$ ($\boldsymbol{Q}_t \sim \mathcal{W}(\boldsymbol{V}_0, \nu_0)$), where the probability density function is $f(\boldsymbol{Q}_t) \propto \det(\boldsymbol{Q}_t)^{(\nu_0 - p - 1)/2} \exp(-\frac{1}{2}\text{trace}(\boldsymbol{V}_0^{-1}\boldsymbol{Q}_t))$. Then we have

$$\begin{aligned}\text{Pen}(\boldsymbol{\Gamma}_t) &= -(\nu_0 - p - 1)\log\det(\boldsymbol{\Gamma}_t\boldsymbol{\Gamma}_t') + \text{trace}(\boldsymbol{V}_0^{-1}\boldsymbol{\Gamma}_t\boldsymbol{\Gamma}_t'),\\ \frac{\partial \text{Pen}(\boldsymbol{\Gamma}_t)}{\partial \boldsymbol{\Gamma}_t} &= 2\left(-(\nu_0 - p - 1)(\boldsymbol{\Gamma}_t\boldsymbol{\Gamma}_t')^{-1} + \boldsymbol{V}_0^{-1}\right)\boldsymbol{\Gamma}_t.\end{aligned}$$

   We can also exploit an inverse Wishart prior on $\boldsymbol{Q}_t$ ($\boldsymbol{Q}_t \sim \mathcal{W}^{-1}(\boldsymbol{V}_1, \nu_1)$), which is the conjugate prior for Gaussian covariance matrices. The probability density function is $f(\boldsymbol{Q}_t) \propto \det(\boldsymbol{Q}_t)^{-(\nu_1 + p + 1)/2} \exp(-\frac{1}{2}\text{trace}(\boldsymbol{V}_1\boldsymbol{Q}_t^{-1}))$. Then we have

$$\begin{aligned}\text{Pen}(\boldsymbol{\Gamma}_t) &= (\nu_1 + p + 1)\log\det(\boldsymbol{\Gamma}_t\boldsymbol{\Gamma}_t') + \text{trace}(\boldsymbol{V}_1(\boldsymbol{\Gamma}_t\boldsymbol{\Gamma}_t')^{-1}),\\ \frac{\partial \text{Pen}(\boldsymbol{\Gamma}_t)}{\partial \boldsymbol{\Gamma}_t} &= 2\left((\nu_1 + p + 1)\boldsymbol{I} - (\boldsymbol{\Gamma}_t\boldsymbol{\Gamma}_t')^{-1}\boldsymbol{V}_1\right)(\boldsymbol{\Gamma}_t\boldsymbol{\Gamma}_t')^{-1}\boldsymbol{\Gamma}_t.\end{aligned}$$

2. For each σ_i^2 in $\boldsymbol{Q}_{J_n}$, we can apply either a Gamma prior or an inverse Gamma prior. In a Gamma prior with the parameters α_0 and β_0, the probability density function is $f(\sigma_i^2) \propto (\sigma_i^2)^{\alpha_0 - 1}\exp(-\beta_0\sigma_i^2)$. The penalty and the corresponding gradient are

$$\text{pen}(\sigma_i^2) = 2(2(1-\alpha_0)\log\sigma_i + \beta\sigma_i^2), \quad \frac{d\text{pen}(\sigma_i^2)}{d\sigma_i} = 4((1-\alpha_0)/\sigma_j + \beta_0\sigma_j).$$

   In an inverse Gamma prior with the parameters α_1 and β_1, the probability density function is $f(\sigma_i^2) \propto (\sigma_i^2)^{-\alpha_1 - 1}\exp(-\beta_1/\sigma_i^2)$. The penalty and the corresponding gradient are

$$\text{pen}(\sigma_i^2) = 2(2(\alpha_1 + 1)\log\sigma_i + \beta_1/\sigma_i^2), \quad \frac{d\text{pen}(\sigma_i^2)}{d\sigma_i} = 4((\alpha_1 + 1)/\sigma_j - \beta_1/\sigma_j^3).$$

3. In each of the ith column in $\boldsymbol{L}$, which corresponds to the ith ROI, we assume the non-zero entries, denoted by $\boldsymbol{L}[l \in \mathcal{A}_i, i]$, has a Gaussian prior with a zero mean and a covariance matrix $\boldsymbol{Q}_{L_i}$, (i.e., $\boldsymbol{L}[l \in \mathcal{A}_i, i] \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{Q}_{L_i})$). The prior covariance matrix $\boldsymbol{Q}_{L_i}$ encourages correlations of source points that are spatially close and share similar orientations. The entry in the l_1th row and the l_2th column in $\boldsymbol{Q}_{L_i}$ is

$$\boldsymbol{Q}_{L_i}[l_1, l_2] = b_0\left(\vec{n}_{l_1}'\vec{n}_{l_2}\right)\exp(-a_0\|\vec{r}_{l_1} - \vec{r}_{l_2}\|_2^2)$$

where $\vec{r}_{l_1}$ and $\vec{r}_{l_2}$ are the coordinates of the source points in the source space corresponding to the l_1th and l_2th elements in $\boldsymbol{L}[l \in \mathcal{A}_i, i]$, and $\vec{n}_{l_1}$ and $\vec{n}_{l_2}$ are the unit vectors pointing to the corresponding orientations. Here a_0 and b_0 are pre-defined hyper-parameters.

$$\text{Pen}(\boldsymbol{L}) = \sum_{i=1}^{p} \boldsymbol{L}[l \in \mathcal{A}_i, i]' \boldsymbol{Q}_{L_i}^{-1} \boldsymbol{L}[l \in \mathcal{A}_i, i] \quad \frac{\partial \text{Pen}(\boldsymbol{L})}{\partial \boldsymbol{L}[l \in \mathcal{A}_i, i]} = 2\boldsymbol{Q}_{L_i}^{-1} \boldsymbol{L}[l \in \mathcal{A}_i, i]$$

4.5.2 Appendix of Section 4.2.2

4.5.2.1 Derivation of the marginal distribution of sensor data in the Kronecker-structured covariance case

Here we drive the steps to get the marginal distribution of $\boldsymbol{Y}$ in Equation (4.13) from the conditional distributions in Equation (4.12). According to the definition of the Kronecker product, we have

$$\text{vec}(\boldsymbol{GJ}) = \begin{pmatrix} \sum_{k=1}^{m} \boldsymbol{G}[1,k]\boldsymbol{J}[k,1] \\ \sum_{k=1}^{m} \boldsymbol{G}[2,k]\boldsymbol{J}[k,1] \\ \cdots \\ \sum_{k=1}^{m} \boldsymbol{G}[n,k]\boldsymbol{J}[k,1] \\ \sum_{k=1}^{m} \boldsymbol{G}[1,k]\boldsymbol{J}[k,2] \\ \sum_{k=1}^{m} \boldsymbol{G}[2,k]\boldsymbol{J}[k,2] \\ \cdots \\ \sum_{k=1}^{m} \boldsymbol{G}[n,k]\boldsymbol{J}[k,2] \\ \cdots \end{pmatrix}$$

$$= \begin{pmatrix} \boldsymbol{G}[1,1] & \boldsymbol{G}[1,2] & \cdots & \boldsymbol{G}[1,m] & 0 & 0 & \cdots & 0 & \cdots \\ \boldsymbol{G}[2,1] & \boldsymbol{G}[2,2] & \cdots & \boldsymbol{G}[2,m] & 0 & 0 & \cdots & 0 & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \\ \boldsymbol{G}[n,1] & \boldsymbol{G}[n,2] & \cdots & \boldsymbol{G}[n,m] & 0 & 0 & \cdots & 0 & \cdots \\ 0 & 0 & \cdots & 0 & \boldsymbol{G}[1,1] & \boldsymbol{G}[1,2] & \cdots & \boldsymbol{G}[1,m] & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \end{pmatrix} \begin{pmatrix} \boldsymbol{J}[1,1] \\ \boldsymbol{J}[2,1] \\ \cdots \\ \boldsymbol{J}[n,1] \\ \boldsymbol{J}[1,2] \\ \boldsymbol{J}[2,2] \\ \cdots \\ \boldsymbol{J}[n,2] \\ \cdots \end{pmatrix}$$

$$= (\boldsymbol{I} \otimes \boldsymbol{G})\text{vec}(\boldsymbol{J})$$

Similarly, $\text{vec}(\boldsymbol{LU}) = (\boldsymbol{I} \otimes \boldsymbol{L})\text{vec}(\boldsymbol{U})$. Applying the similar observations as those in Equations (4.6), (4.7) and (4.9) on $\text{vec}(\boldsymbol{U})$, $\text{vec}(\boldsymbol{J})$ and $\text{vec}(\boldsymbol{Y})$, we can eliminate $\boldsymbol{J}$.

$$\text{vec}(\boldsymbol{Y})|\text{vec}(\boldsymbol{U}) \sim \mathcal{N}(\text{vec}(\boldsymbol{GLU}), \boldsymbol{I} \otimes \boldsymbol{Q}_e + (\boldsymbol{I} \otimes \boldsymbol{G})(\boldsymbol{Q}_{JT} \otimes \boldsymbol{Q}_{J_n})(\boldsymbol{I} \otimes \boldsymbol{G})')$$

Similarly, by eliminating $\boldsymbol{U}$, we have

$$\begin{aligned}\mathrm{vec}(\boldsymbol{Y}) &\sim \mathcal{N}(\boldsymbol{0}, (\boldsymbol{I}\otimes \boldsymbol{Q}_e) + (\boldsymbol{I}\otimes\boldsymbol{G})(\boldsymbol{Q}_{JT}\otimes\boldsymbol{Q}_{J_n})(\boldsymbol{I}\otimes\boldsymbol{G})' + (\boldsymbol{I}\otimes\boldsymbol{GL})(\boldsymbol{Q}_T\otimes\boldsymbol{Q}_S)(\boldsymbol{I}\otimes\boldsymbol{GL})')\\ &= \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}+\boldsymbol{Q}_T\otimes(\boldsymbol{G}\boldsymbol{Q}_{J_n}\boldsymbol{G}'+\boldsymbol{GL}\boldsymbol{Q}_S\boldsymbol{L}'\boldsymbol{G}'))\end{aligned}$$

where we assume $\boldsymbol{Q}_e = \boldsymbol{I}$ and $\boldsymbol{Q}_{JT} = \boldsymbol{Q}_T$.

4.5.2.2 Evaluation of the log-likelihood and the gradients for the Kronecker-structured covariance model

We derive the evaluation of the terms in the negative log-likelihood (nlln$_{kron}$ in Equation (4.14)) and the corresponding gradients, using the formulation in Stegle et al. [2011]. Let $\tilde{\boldsymbol{Q}} = \boldsymbol{G}\boldsymbol{Q}_{J_n}\boldsymbol{G}' + \boldsymbol{GL}\boldsymbol{Q}_S\boldsymbol{L}'\boldsymbol{G}'$. Let the singular value decomposition of $\boldsymbol{Q}_T$ and $\tilde{\boldsymbol{Q}}$ be $\boldsymbol{Q}_T = \boldsymbol{V}_T\boldsymbol{D}_T\boldsymbol{V}_T'$, $\tilde{\boldsymbol{Q}} = \boldsymbol{V}_{\tilde{\boldsymbol{Q}}}\boldsymbol{D}_{\tilde{\boldsymbol{Q}}}\boldsymbol{V}_{\tilde{\boldsymbol{Q}}}'$, where $\boldsymbol{V}_T'\boldsymbol{V}_T = \boldsymbol{V}_T\boldsymbol{V}_T' = \boldsymbol{I}$ and $\boldsymbol{V}_{\tilde{\boldsymbol{Q}}}'\boldsymbol{V}_{\tilde{\boldsymbol{Q}}} = \boldsymbol{V}_{\tilde{\boldsymbol{Q}}}\boldsymbol{V}_{\tilde{\boldsymbol{Q}}}' = \boldsymbol{I}$. We also note that $\boldsymbol{Q}_Y = \boldsymbol{I} + \boldsymbol{Q}_T\otimes\tilde{\boldsymbol{Q}}$. In the derivation below, we use the following equations a lot.

$$\boldsymbol{Q}_Y = (\boldsymbol{I}+\boldsymbol{Q}_T\otimes\tilde{\boldsymbol{Q}}) = (\boldsymbol{V}_T\otimes\boldsymbol{V}_{\tilde{\boldsymbol{Q}}})(\boldsymbol{D}_T\otimes\boldsymbol{D}_{\tilde{\boldsymbol{Q}}}+\boldsymbol{I})(\boldsymbol{V}_T'\otimes\boldsymbol{V}_{\tilde{\boldsymbol{Q}}}') \tag{4.26}$$

$$\boldsymbol{Q}_Y^{-1} = (\boldsymbol{I}+\boldsymbol{Q}_T\otimes\tilde{\boldsymbol{Q}})^{-1} = (\boldsymbol{V}_T\otimes\boldsymbol{V}_{\tilde{\boldsymbol{Q}}})(\boldsymbol{D}_T\otimes\boldsymbol{D}_{\tilde{\boldsymbol{Q}}}+\boldsymbol{I})^{-1}(\boldsymbol{V}_T'\otimes\boldsymbol{V}_{\tilde{\boldsymbol{Q}}}') \tag{4.27}$$

The first term in nllh$_{kron}$ is proportional to the determinant of $\boldsymbol{Q}_Y$, which can be computed in the following way.

$$\begin{aligned}\log\det(\boldsymbol{Q}_Y) &= \log\det(\boldsymbol{I}+\boldsymbol{Q}_T\otimes\tilde{\boldsymbol{Q}})\\ &= \log\det((\boldsymbol{V}_T\otimes\boldsymbol{V}_{\tilde{\boldsymbol{Q}}})(\boldsymbol{D}_T\otimes\boldsymbol{D}_{\tilde{\boldsymbol{Q}}}+\boldsymbol{I})(\boldsymbol{V}_T'\otimes\boldsymbol{V}_{\tilde{\boldsymbol{Q}}}')\ \text{(Equation (4.26))}\\ &= \log\det(\boldsymbol{D}_T\otimes\boldsymbol{D}_{\tilde{\boldsymbol{Q}}}+\boldsymbol{I})\\ &= \sum_{i,j}\log(\boldsymbol{D}_T[i,i]\boldsymbol{D}_{\tilde{\boldsymbol{Q}}}[j,j]+1)\ \text{(because } \boldsymbol{D}_T \text{ and } \boldsymbol{D}_{\tilde{\boldsymbol{Q}}} \text{ are diagonal matrix)}\end{aligned}$$

The second term in nllh$_{kron}$, the quadratic form, can be obtained in the following way, where $\boldsymbol{Y}^{(r)}$ denotes that sensor data in the rth trial.

$$\begin{aligned}&\sum_{r=1}^{q}\mathrm{vec}(\boldsymbol{Y}^{(r)})'\boldsymbol{Q}_Y^{-1}\mathrm{vec}(\boldsymbol{Y}^{(r)})\\ =&\sum_{r=1}^{q}\mathrm{vec}(\boldsymbol{Y}^{(r)})'(\boldsymbol{V}_T\otimes\boldsymbol{V}_{\tilde{\boldsymbol{Q}}})(\boldsymbol{D}_T\otimes\boldsymbol{D}_{\tilde{\boldsymbol{Q}}}+\boldsymbol{I})^{-1}(\boldsymbol{V}_T'\otimes\boldsymbol{V}_{\tilde{\boldsymbol{Q}}}')\mathrm{vec}(\boldsymbol{Y}^{(r)})\quad \text{(plugging in Equation (4.27))}\\ =&\sum_{r=1}^{q}\mathrm{vec}(\boldsymbol{V}_{\tilde{\boldsymbol{Q}}}'\boldsymbol{Y}^{(r)}\boldsymbol{V}_T)'(\boldsymbol{D}_T\otimes\boldsymbol{D}_{\tilde{\boldsymbol{Q}}}+\boldsymbol{I})^{-1}\mathrm{vec}(\boldsymbol{V}_{\tilde{\boldsymbol{Q}}}'\boldsymbol{Y}^{(r)}\boldsymbol{V}_T)\quad \text{(where } \mathrm{vec}(\boldsymbol{V}_{\tilde{\boldsymbol{Q}}}'\boldsymbol{Y}^{(r)}\boldsymbol{V}_T) = (\boldsymbol{V}_T'\otimes\boldsymbol{V}_{\tilde{\boldsymbol{Q}}}')\mathrm{vec}(\boldsymbol{Y}^{(r)})\text{)}\\ =&\sum_{r=1}^{q}\sum_{i,j}((\boldsymbol{V}_{\tilde{\boldsymbol{Q}}}'\boldsymbol{Y}^{(r)}\boldsymbol{V}_T)[i,j])^2(\boldsymbol{D}_T[i,i]\boldsymbol{D}_{\tilde{\boldsymbol{Q}}}[j,j]+1)^{-1}\\ =&\sum_{i,j}(\boldsymbol{D}_T[i,i]\boldsymbol{D}_{\tilde{\boldsymbol{Q}}}[j,j]+1)^{-1}\sum_{r=1}^{q}((\boldsymbol{V}_{\tilde{\boldsymbol{Q}}}'\boldsymbol{Y}^{(r)}\boldsymbol{V}_T)[i,j])^2\end{aligned}$$

Next, we derive the gradients of the two terms in nllh$_{kron}$ with respect to the parameters. We reparametrize $\boldsymbol{Q}_T$ and $\boldsymbol{Q}_S$ with the Cholesky decomposition (i.e., $\boldsymbol{Q}_T = \boldsymbol{\Gamma}_T\boldsymbol{\Gamma}_T'$ and $\boldsymbol{Q}_S = \boldsymbol{\Gamma}_S\boldsymbol{\Gamma}_S'$). The gradient of the determinant term with respect to the (i, j) entry in $\boldsymbol{\Gamma}_T$ is

$$
\begin{aligned}
&\frac{\partial \log\det(\boldsymbol{Q}_Y)}{\partial \boldsymbol{\Gamma}_T[i,j]} = \text{trace}\left(\boldsymbol{Q}_Y^{-1}\frac{\partial \boldsymbol{Q}_Y}{\partial \boldsymbol{\Gamma}_T[i,j]}\right)\\
=&\text{trace}\left((\boldsymbol{I}+\boldsymbol{Q}_T\otimes\tilde{\boldsymbol{Q}})^{-1}\frac{\partial}{\partial \boldsymbol{\Gamma}_T[i,j]}(\boldsymbol{I}+\boldsymbol{Q}_T\otimes\tilde{\boldsymbol{Q}})\right)\\
=&\text{trace}\left((\boldsymbol{V}_T\otimes\boldsymbol{V}_{\tilde{\boldsymbol{Q}}})(\boldsymbol{D}_T\otimes\boldsymbol{D}_{\tilde{\boldsymbol{Q}}}+\boldsymbol{I})^{-1}(\boldsymbol{V}_T'\otimes\boldsymbol{V}_{\tilde{\boldsymbol{Q}}}')(\frac{\partial \boldsymbol{Q}_T}{\partial \boldsymbol{\Gamma}_T[i,j]}\otimes\tilde{\boldsymbol{Q}})\right) \text{ (plugging in Equation (4.27))}\\
=&\text{trace}\left((\boldsymbol{D}_T\otimes\boldsymbol{D}_{\tilde{\boldsymbol{Q}}}+\boldsymbol{I})^{-1}(\boldsymbol{V}_T'\otimes\boldsymbol{V}_{\tilde{\boldsymbol{Q}}}')(\frac{\partial \boldsymbol{Q}_T}{\partial \boldsymbol{\Gamma}_T[i,j]}\otimes\tilde{\boldsymbol{Q}})(\boldsymbol{V}_T\otimes\boldsymbol{V}_{\tilde{\boldsymbol{Q}}})\right) \text{ (rotating the terms within trace)}\\
=&\text{trace}\left((\boldsymbol{D}_T\otimes\boldsymbol{D}_{\tilde{\boldsymbol{Q}}}+\boldsymbol{I})^{-1}(\boldsymbol{V}_T'\frac{\partial \boldsymbol{Q}_T}{\partial \boldsymbol{\Gamma}_T[i,j]}\boldsymbol{V}_T)\otimes(\boldsymbol{V}_{\tilde{\boldsymbol{Q}}}'\tilde{\boldsymbol{Q}}\boldsymbol{V}_{\tilde{\boldsymbol{Q}}})\right)\\
=&\text{trace}\left((\boldsymbol{D}_T\otimes\boldsymbol{D}_{\tilde{\boldsymbol{Q}}}+\boldsymbol{I})^{-1}(\boldsymbol{V}_T'\frac{\partial \boldsymbol{Q}_T}{\partial \boldsymbol{\Gamma}_T[i,j]}\boldsymbol{V}_T)\otimes\boldsymbol{D}_{\tilde{\boldsymbol{Q}}})\right) \text{ (using } \boldsymbol{V}_{\tilde{\boldsymbol{Q}}}'\tilde{\boldsymbol{Q}}\boldsymbol{V}_{\tilde{\boldsymbol{Q}}}=\boldsymbol{D}_{\tilde{\boldsymbol{Q}}})\\
=&\sum_{lh}1/(\boldsymbol{D}_T[l,l]\boldsymbol{D}_{\tilde{\boldsymbol{Q}}}[h,h]+1)(\boldsymbol{V}_T'\frac{\partial \boldsymbol{Q}_T}{\partial \boldsymbol{\Gamma}_T[i,j]}\boldsymbol{V}_T)[l,l]\boldsymbol{D}_{\tilde{\boldsymbol{Q}}}[h,h]
\end{aligned}
$$

where $l = 1\cdots, T$ and $h = 1\cdots, n$. The term $\frac{\partial \boldsymbol{Q}_T}{\partial \boldsymbol{\Gamma}_T[i,j]}$ is computed for each (i, j) in the following way. Generally, for a positive definite matrix $\boldsymbol{Q}$ and its Cholesky decomposition $\boldsymbol{Q} = \boldsymbol{\Gamma}\boldsymbol{\Gamma}'$, $\boldsymbol{Q}[s, w] = \sum_{k\le min(s,w)} \boldsymbol{\Gamma}[s, k]\boldsymbol{\Gamma}[w, k]$, since $\boldsymbol{\Gamma}$ is a lower triangular matrix.

$$
\frac{\partial \boldsymbol{Q}[s,w]}{\partial \boldsymbol{\Gamma}[i,j]} = \begin{cases} 0 & \text{if } i\neq s, i\neq w\\ \boldsymbol{\Gamma}[w,j] & \text{if } i=s\neq w, j\le min(s,w)\\ \boldsymbol{\Gamma}[s,j] & \text{if } i=w\neq s, j\le min(s,w)\\ 2\boldsymbol{\Gamma}[w,j] & \text{if } i=s=w, j\le min(s,w)\end{cases}
$$

We compute $\frac{\partial \boldsymbol{Q}_T}{\partial \boldsymbol{\Gamma}_T[i,j]}$ accordingly.

The gradient of the quadratic form term with respect to the (i, j) entry in $\boldsymbol{\Gamma}_T$ is

$$
\begin{aligned}
&\frac{\partial}{\partial \boldsymbol{\Gamma}_T[i,j]}(\sum_{r=1}^{q} \text{vec}(\boldsymbol{Y}^{(r)})' \boldsymbol{Q}_Y^{-1} \text{vec}(\boldsymbol{Y}^{(r)}) \\
=& -\sum_{r=1}^{q} \text{vec}(\boldsymbol{Y}^{(r)})'(\boldsymbol{I} + \boldsymbol{Q}_T \otimes \tilde{\boldsymbol{Q}})^{-1} \frac{\partial (\boldsymbol{I} + \boldsymbol{Q}_T \otimes \tilde{\boldsymbol{Q}})}{\partial \boldsymbol{\Gamma}_T[i,j]} (\boldsymbol{I} + \boldsymbol{Q}_T \otimes \tilde{\boldsymbol{Q}})^{-1} \text{vec}(\boldsymbol{Y}^{(r)}) \\
=& -\sum_{r=1}^{q} \text{vec}(\boldsymbol{Y}^{(r)})'(\boldsymbol{V}_T \otimes \boldsymbol{V}_{\tilde{\boldsymbol{Q}}})(\boldsymbol{D}_T \otimes \boldsymbol{D}_{\tilde{\boldsymbol{Q}}} + \boldsymbol{I})^{-1}(\boldsymbol{V}_T' \otimes \boldsymbol{V}_{\tilde{\boldsymbol{Q}}}')(\frac{\partial \boldsymbol{Q}_T}{\partial \boldsymbol{\Gamma}_T[i,j]} \otimes \tilde{\boldsymbol{Q}}) \\
& (\boldsymbol{V}_T \otimes \boldsymbol{V}_{\tilde{\boldsymbol{Q}}})(\boldsymbol{D}_T \otimes \boldsymbol{D}_{\tilde{\boldsymbol{Q}}} + \boldsymbol{I})^{-1}(\boldsymbol{V}_T' \otimes \boldsymbol{V}_{\tilde{\boldsymbol{Q}}}') \text{vec}(\boldsymbol{Y}^{(r)}) \\
=& -\sum_{r=1}^{q} \text{vec}(\boldsymbol{V}_{\tilde{\boldsymbol{Q}}}' \boldsymbol{Y}^{(r)} \boldsymbol{V}_T)'(\boldsymbol{D}_T \otimes \boldsymbol{D}_{\tilde{\boldsymbol{Q}}} + \boldsymbol{I})^{-1}(\boldsymbol{V}_T' \frac{\partial \boldsymbol{Q}_T}{\partial \boldsymbol{\Gamma}_T[i,j]} \boldsymbol{V}_T \otimes \boldsymbol{V}_{\tilde{\boldsymbol{Q}}}' \tilde{\boldsymbol{Q}} \boldsymbol{V}_{\tilde{\boldsymbol{Q}}})(\boldsymbol{D}_T \otimes \boldsymbol{D}_{\tilde{\boldsymbol{Q}}} + \boldsymbol{I})^{-1} \text{vec}(\boldsymbol{V}_{\tilde{\boldsymbol{Q}}}' \boldsymbol{Y}^{(r)} \boldsymbol{V}_T) \\
=& -\sum_{r=1}^{q} \text{vec}(\tilde{\boldsymbol{Y}}^{(r)})'(\boldsymbol{V}_T' \frac{\partial \boldsymbol{Q}_T}{\partial \boldsymbol{\Gamma}_T[i,j]} \boldsymbol{V}_T \otimes \boldsymbol{D}_{\tilde{\boldsymbol{Q}}}) \text{vec}(\tilde{\boldsymbol{Y}}^{(r)}) \\
& \quad (\text{where } \text{vec}(\tilde{\boldsymbol{Y}}^{(r)}) = (\boldsymbol{D}_T \otimes \boldsymbol{D}_{\tilde{\boldsymbol{Q}}} + \boldsymbol{I})^{-1} \text{vec}(\boldsymbol{V}_{\tilde{\boldsymbol{Q}}}' \boldsymbol{Y}^{(r)} \boldsymbol{V}_T))
\end{aligned}
$$

The term $\text{vec}(\tilde{\boldsymbol{Y}}^{(r)}) = (\boldsymbol{D}_T \otimes \boldsymbol{D}_{\tilde{\boldsymbol{Q}}} + \boldsymbol{I})^{-1} \text{vec}(\boldsymbol{V}_{\tilde{\boldsymbol{Q}}}' \boldsymbol{Y}^{(r)} \boldsymbol{V}_T)$ can be easily computed because $(\boldsymbol{D}_T \otimes \boldsymbol{D}_{\tilde{\boldsymbol{Q}}} + \boldsymbol{I})$ is diagonal. We also use the following trick to further simplify the computation.

$$
(\boldsymbol{V}_T' \frac{\partial \boldsymbol{Q}_T}{\partial \boldsymbol{\Gamma}_T[i,j]} \boldsymbol{V}_T \otimes \boldsymbol{D}_{\tilde{\boldsymbol{Q}}}) \text{vec}(\tilde{\boldsymbol{Y}}^{(r)}) = \text{vec}(\boldsymbol{D}_{\tilde{\boldsymbol{Q}}} \tilde{\boldsymbol{Y}}^{(r)} \boldsymbol{V}_T' \frac{\partial \boldsymbol{Q}_T}{\partial \boldsymbol{\Gamma}_T[i,j]} \boldsymbol{V}_T)
$$

The gradient with respect to other unknown parameters $\boldsymbol{\Gamma}_S$, $\boldsymbol{L}$ and $\sigma_0, \sigma_1, \cdots, \sigma_p$ can be derived in a similar way. Let θ denote any entry in $\boldsymbol{\Gamma}_S$ or $\boldsymbol{L}$ of interest, or any $\sigma_i, i = 0, 1, \cdots, p$.

$$
\begin{aligned}
& \frac{\partial \log \det(\boldsymbol{Q}_Y)}{\partial \theta} = \text{trace}\left((\boldsymbol{D}_T \otimes \boldsymbol{D}_{\tilde{\boldsymbol{Q}}} + \boldsymbol{I})^{-1}(\boldsymbol{D}_T \otimes \boldsymbol{V}_{\tilde{\boldsymbol{Q}}}' \frac{\partial \tilde{\boldsymbol{Q}}}{\partial \boldsymbol{\theta}} \boldsymbol{V}_{\tilde{\boldsymbol{Q}}})\right) \\
=& \sum_{lh} 1/(\boldsymbol{D}_T[l,l] \boldsymbol{D}_{\tilde{\boldsymbol{Q}}}[h,h] + 1) \boldsymbol{D}_T[l,l](\boldsymbol{V}_{\tilde{\boldsymbol{Q}}}' \frac{\partial \tilde{\boldsymbol{Q}}}{\partial \theta} \boldsymbol{V}_{\tilde{\boldsymbol{Q}}})[h,h] \\
& \frac{\partial}{\partial \theta}(\sum_{r=1}^{q} \text{vec}(\boldsymbol{Y}^{(r)})' \boldsymbol{Q}_Y^{-1} \text{vec}(\boldsymbol{Y}^{(r)})) = -\sum_{r=1}^{q} \text{vec}(\tilde{\boldsymbol{Y}}^{(r)})'(\boldsymbol{D}_T \otimes (\boldsymbol{V}_{\tilde{\boldsymbol{Q}}}' \frac{\partial \tilde{\boldsymbol{Q}}}{\partial \theta} \boldsymbol{V}_{\tilde{\boldsymbol{Q}}}) \text{vec}(\tilde{\boldsymbol{Y}}^{(r)})
\end{aligned}
$$

What remains is the derivation of $\frac{\partial \tilde{\boldsymbol{Q}}}{\partial \theta}$. When $\theta = \sigma_i, i = 0, 1, \cdots, p$,

$$
\frac{\partial \tilde{\boldsymbol{Q}}}{\partial \theta} = \frac{\partial \boldsymbol{G} \boldsymbol{Q}_{J_n} \boldsymbol{G}'}{\partial \sigma_i} = \frac{\partial}{\partial \sigma_i}(\sigma_i^2 \boldsymbol{G}[:, l \in \mathcal{A}_i] \boldsymbol{G}[:, l \in \mathcal{A}_i]') = 2\sigma_j \boldsymbol{G}[:, l \in \mathcal{A}_i] \boldsymbol{G}[:, l \in \mathcal{A}_i]'
$$

Notice that $\tilde{\boldsymbol{Q}}$ has the same form as $\boldsymbol{Q}_{yt}$ in Section 4.5.1.2; therefore when θ is an entry in $\boldsymbol{\Gamma}_S$ or $\boldsymbol{L}$, we can use what we derived in Section 4.5.1.2.

When $\theta = \boldsymbol{\Gamma}_S[s,w]$, $\boldsymbol{C} = \boldsymbol{G}\boldsymbol{L}$, we have

$$\begin{aligned}
\frac{\partial \tilde{\boldsymbol{Q}}[i,j]}{\partial \boldsymbol{\Gamma}_S[s,w]} = \frac{\partial(\boldsymbol{C}\boldsymbol{\Gamma}_S\boldsymbol{\Gamma}_S'\boldsymbol{C}')[i,j]}{\partial \boldsymbol{\Gamma}_S[s,w]} &= \sum_k \boldsymbol{C}[j,k]\boldsymbol{C}[i,s]\boldsymbol{\Gamma}_S[k,w] + \sum_k \boldsymbol{C}[j,s]\boldsymbol{C}[i,k]\boldsymbol{\Gamma}_S[k,w] \\
&= \boldsymbol{C}[i,s]\sum_k \boldsymbol{C}[j,k]\boldsymbol{\Gamma}_S[k,w] + \boldsymbol{C}[j,s]\sum_k \boldsymbol{C}[i,k]\boldsymbol{\Gamma}_S[k,w] \\
&= \boldsymbol{C}[i,s](\boldsymbol{C}\boldsymbol{\Gamma}_S)[j,w] + (\boldsymbol{C}\boldsymbol{\Gamma}_S)[i,w]\boldsymbol{C}[j,s]
\end{aligned}$$

In matrix form, we have

$$\begin{aligned}
\frac{\partial \tilde{\boldsymbol{Q}}}{\partial \boldsymbol{\Gamma}_S[s,w]} &= \begin{pmatrix} \boldsymbol{C}[1,s] \\ \boldsymbol{C}[2,s] \\ \cdots \\ \boldsymbol{C}[n,s] \end{pmatrix} ((\boldsymbol{C}\boldsymbol{\Gamma}_S)[1,w], (\boldsymbol{C}\boldsymbol{\Gamma}_S)[2,w], \cdots, (\boldsymbol{C}\boldsymbol{\Gamma}_S)[n,w]) \\
&+ \begin{pmatrix} (\boldsymbol{C}\boldsymbol{\Gamma}_S)[1,w] \\ (\boldsymbol{C}\boldsymbol{\Gamma}_S)[2,w] \\ \cdots \\ (\boldsymbol{C}\boldsymbol{\Gamma}_S)[n,w] \end{pmatrix} (\boldsymbol{C}[1,s], \boldsymbol{C}[2,s], \cdots, \boldsymbol{C}[n,s]) \\
&= \boldsymbol{C}[:,s]((\boldsymbol{C}\boldsymbol{\Gamma}_S)[:,w])' + ((\boldsymbol{C}\boldsymbol{\Gamma}_S)[:,w])((\boldsymbol{C}[:,s])'
\end{aligned}$$

When $\theta = \boldsymbol{L}[s,w]$,

$$\begin{aligned}
\frac{\partial \tilde{\boldsymbol{Q}}[i,j]}{\partial \boldsymbol{L}[s,w]} = \frac{\partial(\boldsymbol{G}\boldsymbol{L}\boldsymbol{Q}_S\boldsymbol{L}'\boldsymbol{G}')[i,j]}{\partial \boldsymbol{L}[s,w]} &= \sum_{rh} \boldsymbol{G}[i,s]\boldsymbol{Q}_S[w,h]\boldsymbol{G}[j,r]\boldsymbol{L}[r,h] + \sum_{rh} \boldsymbol{G}[i,r]\boldsymbol{Q}_S[h,w]\boldsymbol{G}[j,s]\boldsymbol{L}[r,h] \\
&= \boldsymbol{G}[i,s]\sum_h(\sum_r \boldsymbol{G}[j,r]\boldsymbol{L}[r,h])\boldsymbol{Q}_S[w,h] + \boldsymbol{G}[j,s]\sum_h(\sum_r \boldsymbol{G}[i,r]\boldsymbol{L}[r,h])\boldsymbol{Q}_S[h,w] \\
&= \boldsymbol{G}[i,s](\boldsymbol{G}\boldsymbol{L}\boldsymbol{Q}_S)[j,w] + (\boldsymbol{G}\boldsymbol{L}\boldsymbol{Q}_S)[i,w]\boldsymbol{G}[j,s] \\
\frac{\partial \tilde{\boldsymbol{Q}}}{\partial \boldsymbol{L}[s,w]} &= \boldsymbol{G}[:,s]((\boldsymbol{G}\boldsymbol{L}\boldsymbol{Q}_S)[:,w])' + ((\boldsymbol{G}\boldsymbol{L}\boldsymbol{Q}_S)[:,w])(\boldsymbol{G}[:,s])'
\end{aligned}$$

4.5.2.3 Sampling from Kronecker-structured spatio-temporal covariance

Let $\boldsymbol{U}$ be a random $p \times T$ matrix, where $\text{vec}(\boldsymbol{U})$ has a zero mean and a Kronecker-structured covariance matrix $\boldsymbol{Q}_T \otimes \boldsymbol{Q}_S$. An easy way to generate such samples of $\boldsymbol{U}$ is

$$\begin{aligned}
\boldsymbol{U}_0 &\sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I} \otimes \boldsymbol{I}) \\
\boldsymbol{U} &= \boldsymbol{\Gamma}_S\boldsymbol{U}_0\boldsymbol{\Gamma}_T'
\end{aligned}$$

where $\boldsymbol{\Gamma}_T$ and $\boldsymbol{\Gamma}_S$ are from the Cholesky decomposition of $\boldsymbol{Q}_T$ and $\boldsymbol{Q}_S$ ($\boldsymbol{Q}_T = \boldsymbol{\Gamma}_T\boldsymbol{\Gamma}_T'$ and $\boldsymbol{Q}_S = \boldsymbol{\Gamma}_S\boldsymbol{\Gamma}_S'$).

4.5.2.4 Maximum-likelihood estimate of the Kronecker-structured spatio-temporal covariance

Given $p \times T$ multidimensional time series in q i.i.d trials, $\boldsymbol{U}^{(r)}, r = 1, \cdots, q$, which are sampled from the distribution above (i.e., vec$(\boldsymbol{U}^{(r)})$s have a zero mean and a Kronecker-structured covariance, $\boldsymbol{Q}_T \otimes \boldsymbol{Q}_S$), we alternate between the following steps to obtain the maximum-likelihood estimates of $\hat{\boldsymbol{Q}_T}$ and $\hat{\boldsymbol{Q}_S}$.

$$\hat{\boldsymbol{Q}_T} \leftarrow 1/(qp) \sum_{r=1}^{q} (\boldsymbol{U}^{(r)})' \hat{\boldsymbol{Q}_S}^{-1} \boldsymbol{U}^{(r)}$$

$$\hat{\boldsymbol{Q}_S} \leftarrow 1/(qT) \sum_{r=1}^{q} \boldsymbol{U}^{(r)} \hat{\boldsymbol{Q}_T}^{-1} (\boldsymbol{U}^{(r)})'$$

4.5.3 Appendix of Section 4.3

4.5.3.1 Details on the E-M algorithm

Let $f(\cdot)$ denote probability density functions in general. We aim to obtain the set of parameters $\boldsymbol{\theta} = \{\{\boldsymbol{A}_t\}_{t=1}^{T}, \boldsymbol{Q}_0, \boldsymbol{Q}_\epsilon, \{\sigma_i^2\}_{i=0}^{p}\}$ that maximizes the objective function $\log f(\{\boldsymbol{y}_t^{(r)}\}_{t=0,r=1}^{T,q}; \boldsymbol{\theta}) + \log f(\boldsymbol{\theta})$, where $f(\boldsymbol{\theta})$ denotes the prior on $\boldsymbol{\theta}$. In our case $\log f(\{\boldsymbol{y}_t^{(r)}\}_{t=0,r=1}^{T,q}, \{\boldsymbol{u}_t^{(r)}\}_{t=0,r=1}^{T,q}; \boldsymbol{\theta})$ is much easier to compute than $\log f(\{\boldsymbol{y}_t^{(r)}\}_{t=0,r=1}^{T,q}; \boldsymbol{\theta})$, and the E-M algorithm Shumway and Stoffer [1982] utilizes this property. Below, we briefly introduce how it works. Let $\tilde{\boldsymbol{\theta}}$ denote an estimate of $\boldsymbol{\theta}$. For a more succinct notation, let $\boldsymbol{u}^\dagger \overset{\text{def}}{=} \{\boldsymbol{u}_t^{(r)}\}_{t=0,r=1}^{T,q}$ and $\boldsymbol{y}^\dagger \overset{\text{def}}{=} \{\boldsymbol{y}_t^{(r)}\}_{t=0,r=1}^{T,q}$. Let $\tilde{f}(\boldsymbol{u}^\dagger) = f(\boldsymbol{u}^\dagger | \boldsymbol{y}^\dagger; \tilde{\boldsymbol{\theta}})$ be the posterior distribution of $\boldsymbol{u}^\dagger$ conditioned on observations $\boldsymbol{y}^\dagger$, based on the current estimate $\tilde{\boldsymbol{\theta}}$.

$$\begin{aligned}
&\log f(\boldsymbol{y}^\dagger; \boldsymbol{\theta}) + \log f(\boldsymbol{\theta}) \\
= &\int \tilde{f}(\boldsymbol{u}^\dagger) \log f(\boldsymbol{y}^\dagger; \boldsymbol{\theta}) d\boldsymbol{u}^\dagger + \log f(\boldsymbol{\theta}) = \mathbb{E}_{\tilde{f}}(\log f(\boldsymbol{y}^\dagger; \boldsymbol{\theta})) + \log f(\boldsymbol{\theta}) \\
= &\mathbb{E}_{\tilde{f}}(\log \frac{f(\boldsymbol{y}^\dagger, \boldsymbol{u}^\dagger; \boldsymbol{\theta})}{f(\boldsymbol{u}^\dagger | \boldsymbol{y}^\dagger; \boldsymbol{\theta})}) + \log f(\boldsymbol{\theta}) \\
= &\mathbb{E}_{\tilde{f}}(\log f(\boldsymbol{y}^\dagger, \boldsymbol{u}^\dagger; \boldsymbol{\theta})) + \log f(\boldsymbol{\theta}) - \mathbb{E}_{\tilde{f}}(\log f(\boldsymbol{u}^\dagger | \boldsymbol{y}^\dagger; \boldsymbol{\theta}))
\end{aligned}$$

This also holds for $\boldsymbol{\theta} = \tilde{\boldsymbol{\theta}}$.

$$\log f(\boldsymbol{y}^\dagger; \tilde{\boldsymbol{\theta}}) + \log f(\tilde{\boldsymbol{\theta}}) = \mathbb{E}_{\tilde{f}}(\log f(\boldsymbol{y}^\dagger, \boldsymbol{u}^\dagger; \tilde{\boldsymbol{\theta}})) + \log f(\tilde{\boldsymbol{\theta}}) - \mathbb{E}_{\tilde{f}}(\log \tilde{f}(\boldsymbol{u}^\dagger))$$

Now consider the difference

$$(\log f(\boldsymbol{y}^\dagger; \boldsymbol{\theta}) + \log f(\boldsymbol{\theta})) - (\log f(\boldsymbol{y}^\dagger; \tilde{\boldsymbol{\theta}}) + \log f(\tilde{\boldsymbol{\theta}})) \tag{4.28}$$

$$= \mathbb{E}_{\tilde{f}}(\log f(\boldsymbol{y}^\dagger, \boldsymbol{u}^\dagger; \boldsymbol{\theta})) + \log f(\boldsymbol{\theta}) \tag{4.29}$$

$$- \mathbb{E}_{\tilde{f}}(\log f(\boldsymbol{y}^\dagger, \boldsymbol{u}^\dagger; \tilde{\boldsymbol{\theta}})) - \log f(\tilde{\boldsymbol{\theta}}) \tag{4.30}$$

$$+ \mathbb{E}_{\tilde{f}}(\log \frac{\tilde{f}(\boldsymbol{u}^\dagger)}{f(\boldsymbol{u}^\dagger | \boldsymbol{y}^\dagger; \boldsymbol{\theta})}) \tag{4.31}$$

In each iteration, given $\tilde{\boldsymbol{\theta}}$, we select the new $\boldsymbol{\theta}$ that maximize the term (4.29). Because term (4.30) is fixed given $\tilde{\boldsymbol{\theta}}$, and term (4.31) is the non-negative Kullback-Leibler distance between $\tilde{f}(\boldsymbol{u}^\dagger)$ and $f(\boldsymbol{u}^\dagger | \boldsymbol{y}^\dagger; \boldsymbol{\theta})$, we are essentially maximizing a lower bound of the difference term (4.28). Therefore, when it converges, term (4.31) goes to zero, and we reach a local maximum of the objective function.

In each iteration, there are two steps: an E-step to compute $\tilde{f}$ or the expression for term (4.29) given the current $\tilde{\boldsymbol{\theta}}$, and an M-step to maximize the term (4.29) and update $\boldsymbol{\theta}$.

If we only have a prior on $\{\boldsymbol{A}_t\}_{t=1}^T$, (i.e., $f(\boldsymbol{\theta}) = f(\{\boldsymbol{A}_t\}_{t=1}^T) \propto \exp(-(\lambda_0 \sum_{t=1}^T \|\boldsymbol{A}_t\|_F^2 + \lambda_1 \sum_{t=2}^T \|\boldsymbol{A}_t - \boldsymbol{A}_{t-1}\|_F^2)))$, then term (4.30) has the following form up to some constant

$$\mathbb{E}_{\tilde{f}}(\log f(\boldsymbol{y}^\dagger, \boldsymbol{u}^\dagger; \boldsymbol{\theta})) + \log f(\boldsymbol{\theta}) \tag{4.32}$$

$$= -\frac{1}{2}(q \log \det(\boldsymbol{Q}_0) + \mathbb{E}_{\tilde{f}}(\sum_{r=1}^{q} \boldsymbol{u}_0^{(r)\prime} \boldsymbol{Q}_0^{-1} \boldsymbol{u}_0^{(r)}) \tag{4.33}$$

$$- \frac{1}{2}(qT \log \det(\boldsymbol{Q}_\epsilon) + \mathbb{E}_{\tilde{f}}(\sum_{t=1}^{T} \sum_{r=1}^{q} (\boldsymbol{u}_t^{(r)} - \boldsymbol{A}_t \boldsymbol{u}_{t-1}^{(r)})' \boldsymbol{Q}_\epsilon^{-1} (\boldsymbol{u}_t^{(r)} - \boldsymbol{A}_t \boldsymbol{u}_{t-1}^{(r)})) + \log f(\{\boldsymbol{A}_t\}_{t=1}^T) \tag{4.34}$$

$$- \frac{1}{2}(q(T+1) \log \det(\boldsymbol{Q}_\eta) + \mathbb{E}_{\tilde{f}}(\sum_{t=0}^{T} \sum_{r=1}^{q} (\boldsymbol{y}_t^{(r)} - \boldsymbol{C} \boldsymbol{u}_t^{(r)})' \boldsymbol{Q}_\eta^{-1} (\boldsymbol{y}_t^{(r)} - \boldsymbol{C} \boldsymbol{u}_t^{(r)})). \tag{4.35}$$

where $'$ denotes the transpose of a column vector or a matrix, and $\det(\cdot)$ denotes the determinant of a square matrix. To evaluate (4.33), (4.34), and (4.35), we need to compute the posterior mean and covariance of $\boldsymbol{u}_t$, as well as the cross covariance of $\boldsymbol{u}_t$ and $\boldsymbol{u}_{t-1}$ at each t for each trial ($r = 1, 2, \cdots, q$) given $\tilde{\boldsymbol{\theta}}$:

$$\boldsymbol{u}_{t|T}^{(r)} \overset{def}{=} \mathbb{E}(\boldsymbol{u}_t^{(r)} | \{\boldsymbol{y}_\tau^{(r)}\}_{\tau=0}^T),$$

$$\boldsymbol{P}_{t|T}^{(r)} \overset{def}{=} \text{cov}(\boldsymbol{u}_t^{(r)} | \{\boldsymbol{y}_\tau^{(r)}\}_{\tau=0}^T),$$

$$\boldsymbol{P}_{(t,t-1)|T}^{(r)} \overset{def}{=} \text{cov}(\boldsymbol{u}_t^{(r)}, \boldsymbol{u}_{t-1}^{(r)} | \{\boldsymbol{y}_\tau^{(r)}\}_{\tau=0}^T)$$

To compute these values, we use the forward and backward steps in the Kalman smoothing algorithm Shumway and Stoffer [1982]. For simplicity, we drop the superscript $^{(r)}$ and the $\tilde{}$ symbol on

$\tilde{\boldsymbol{\theta}}$. We define the following terms

$$\begin{aligned}
\boldsymbol{u}_{t|s} &\overset{\text{def}}{=} \mathbb{E}(\boldsymbol{u}_t|\boldsymbol{y}_0, \boldsymbol{y}_1, \cdots, \boldsymbol{y}_s), \\
\boldsymbol{P}_{t|s} &\overset{\text{def}}{=} \text{cov}(\boldsymbol{u}_t|\boldsymbol{y}_0.\boldsymbol{y}_1, \cdots, \boldsymbol{y}_s), \\
\boldsymbol{P}_{(t,t-1)|s} &\overset{\text{def}}{=} \text{cov}(\boldsymbol{u}_t, \boldsymbol{u}_{t-1}|\boldsymbol{y}_0, \boldsymbol{y}_1, \cdots, \boldsymbol{y}_s).
\end{aligned}$$

In the forward step, we set $\boldsymbol{u}_{0|0} = \boldsymbol{Q}_0\boldsymbol{C}'(\boldsymbol{C}\boldsymbol{Q}_0\boldsymbol{C}' + \boldsymbol{Q}_\eta)^{-1}\boldsymbol{y}_0$ and $\boldsymbol{P}_{0|0} = \boldsymbol{Q}_0 - \boldsymbol{Q}_0\boldsymbol{C}'(\boldsymbol{C}\boldsymbol{Q}_0\boldsymbol{C}' + \boldsymbol{Q}_\eta)^{-1}\boldsymbol{C}\boldsymbol{Q}_0$. For $t = 1, 2, \cdots, T$, we have

$$\begin{aligned}
\boldsymbol{u}_{t|(t-1)} &= \boldsymbol{A}_t\boldsymbol{u}_{(t-1)|(t-1)} \\
\boldsymbol{P}_{t|(t-1)} &= \boldsymbol{A}_t\boldsymbol{P}_{(t-1)|(t-1)}\boldsymbol{A}_t' + \boldsymbol{Q} \\
\boldsymbol{K}_t &\overset{\text{def}}{=} \boldsymbol{P}_{t|(t-1)}\boldsymbol{C}'(\boldsymbol{C}\boldsymbol{P}_{t|(t-1)}\boldsymbol{C}' + \boldsymbol{Q}_\eta)^{-1} \\
\boldsymbol{u}_{t|t} &= \boldsymbol{u}_{t|(t-1)} + \boldsymbol{K}_t(\boldsymbol{y}_t - \boldsymbol{C}\boldsymbol{u}_{t|(t-1)}) \\
\boldsymbol{P}_{t|t} &= \boldsymbol{P}_{t|(t-1)} - \boldsymbol{K}_t\boldsymbol{C}\boldsymbol{P}_{t|(t-1)}
\end{aligned}$$

In the backward step, for $t = T, T-1, \cdots, 1$

$$\begin{aligned}
\boldsymbol{H}_{t-1} &\overset{\text{def}}{=} \boldsymbol{P}_{(t-1)|(t-1)}\boldsymbol{A}_t'\boldsymbol{P}_{t|(t-1)}^{-1} \\
\boldsymbol{u}_{(t-1)|T} &= \boldsymbol{u}_{(t-1)|(t-1)} + \boldsymbol{H}_{t-1}(\boldsymbol{u}_{t|T} - \boldsymbol{A}_t\boldsymbol{u}_{(t-1)|(t-1)}) \\
\boldsymbol{P}_{(t-1)|T} &= \boldsymbol{P}_{(t-1)|(t-1)} + \boldsymbol{H}_{t-1}(\boldsymbol{P}_{t|T} - \boldsymbol{P}_{t|(t-1)})\boldsymbol{H}_{t-1}'
\end{aligned}$$

and with $\boldsymbol{P}_{(T,T-1)|T} = (\boldsymbol{I} - \boldsymbol{K}_T\boldsymbol{C})\boldsymbol{A}_T\boldsymbol{P}_{(T-1)|(T-1)}$, for $t = T-1, T-2, \cdots, 2$, we have

$$\boldsymbol{P}_{(t-1,t-2)|T} = \boldsymbol{P}_{(t-1)|(t-1)}\boldsymbol{H}_{t-2}' + \boldsymbol{H}_{t-1}(\boldsymbol{P}_{(t,t-1)|T} - \boldsymbol{A}_t\boldsymbol{P}_{(t-1)|(t-1)})\boldsymbol{H}_{t-2}'.$$

If we denote the terms (4.33), (4.34), and (4.35) with $\mathcal{L}_1$, $\mathcal{L}_2$ and $\mathcal{L}_3$ respectively, then using the posterior statistics above, we have

$$\begin{aligned}
\mathcal{L}_1 &= q\log\det(\boldsymbol{Q}_0) + \mathbb{E}_{\tilde{f}}(\sum_{r=1}^{q}(\boldsymbol{u}_0^{(r)'}\boldsymbol{Q}_0^{-1}\boldsymbol{u}_0^{(r)})) = q\log\det(\boldsymbol{Q}_0) + \text{trace}(\boldsymbol{Q}_0^{-1}\boldsymbol{B}_0) \\
\mathcal{L}_2 &= qT\log\det(\boldsymbol{Q}_\epsilon) + \mathbb{E}_{\tilde{f}}(\sum_{t=1}^{T}\sum_{r=1}^{q}(\boldsymbol{u}_t^{(r)} - \boldsymbol{A}_t\boldsymbol{u}_{t-1}^{(r)})'\boldsymbol{Q}_\epsilon^{-1}(\boldsymbol{u}_t^{(r)} - \boldsymbol{A}_t\boldsymbol{u}_{t-1}^{(r)}))] - \log f(\{\boldsymbol{A}_t\}_{t=1}^{T}) \\
&= qT\log\det(\boldsymbol{Q}_\epsilon) + \text{trace}(\boldsymbol{Q}_\epsilon^{-1}\sum_{t=1}^{T}(\boldsymbol{B}_{1t} - \boldsymbol{A}_t\boldsymbol{B}_{2t}' - \boldsymbol{B}_{2t}\boldsymbol{A}_t' + \boldsymbol{A}_t\boldsymbol{B}_{3t}\boldsymbol{A}_t')) \\
&+ \lambda_0\sum_{t=1}^{T}\|\boldsymbol{A}_t\|_F^2 + \lambda_1\sum_{t=2}^{T}\|\boldsymbol{A}_t - \boldsymbol{A}_{t-1}\|_F^2 \\
\mathcal{L}_3 &= q(T+1)\log\det(\boldsymbol{Q}_\eta) + \mathbb{E}_{\tilde{f}}(\sum_{t=0}^{T}\sum_{r=1}^{q}(\boldsymbol{y}_t(r) - \boldsymbol{C}\boldsymbol{u}_t^{(r)})'\boldsymbol{Q}_\eta^{-1}(\boldsymbol{y}_t(r) - \boldsymbol{C}\boldsymbol{u}_t^{(r)})') \\
&= q(T+1)\log\det(\boldsymbol{Q}_\eta) + \text{trace}(\boldsymbol{Q}_\eta^{-1}\boldsymbol{B}_4)
\end{aligned}$$

where

$$
\begin{aligned}
\boldsymbol{B}_0 &= \sum_{r=1}^{q}(\boldsymbol{P}_{0|T}^{(r)} + \boldsymbol{u}_{0|T}^{(r)}(\boldsymbol{u}_{0|T}^{(r)})') \\
\boldsymbol{B}_{1t} &= \sum_{r=1}^{q}(\boldsymbol{P}_{t|T}^{(r)} + \boldsymbol{u}_{t|T}^{(r)}(\boldsymbol{u}_{t|T}^{(r)})') \\
\boldsymbol{B}_{2t} &= \sum_{r=1}^{q}(\boldsymbol{P}_{(t,t-1)|T}^{(r)} + \boldsymbol{u}_{t|T}^{(r)}(\boldsymbol{u}_{(t-1)|T}^{(r)})') \\
\boldsymbol{B}_{3t} &= \sum_{r=1}^{q}(\boldsymbol{P}_{(t-1)|T}^{(r)} + \boldsymbol{u}_{(t-1)|T}^{(r)}(\boldsymbol{u}_{(t-1)|T}^{(r)})') \\
\boldsymbol{B}_4 &= \sum_{r=1}^{q}\sum_{t=0}^{T}[(\boldsymbol{y}_t^{(r)} - \boldsymbol{C}\boldsymbol{u}_{t|T}^{(r)})(\boldsymbol{y}_t^{(r)} - \boldsymbol{C}\boldsymbol{u}_{t|T}^{(r)})' + \boldsymbol{C}\boldsymbol{P}_{t|T}^{(r)}\boldsymbol{C}')]
\end{aligned}
$$

In the M-step, each term was optimized either using the analytical solution or gradient descent with back-tracking.

4.5.3.2 Solving for dynamic connectivity parameters when the state-variables are observed

Given $\{\boldsymbol{u}_t^{(r)}\}_{t=0,r=1}^{T,q}$, we solve for $\boldsymbol{Q}_0$, $\boldsymbol{Q}_\epsilon$ and $\{\boldsymbol{A}_t\}_{t=1}^{T}$ by maximizing the log-likelihood plus the logarithm of the prior, which is equivalent to minimizing

$$
\begin{aligned}
& -\log f(\{\boldsymbol{u}_t^{(r)}\}_{t=0,r=1}^{T,q}) - \log f(\{\boldsymbol{A}_t\}_{t=1}^{T}) \\
& \propto q \log \det(\boldsymbol{Q}_0) + \text{trace}(\boldsymbol{Q}_0^{-1}\sum_{r=1}^{q}\boldsymbol{u}_0^{(r)}\boldsymbol{u}_0^{(r)'}) \\
& + qT \log\det(\boldsymbol{Q}_\epsilon) + \text{trace}(\boldsymbol{Q}_\epsilon^{-1}\sum_{t=1}^{T}\sum_{r=1}^{q}(\boldsymbol{u}_t^{(r)} - \boldsymbol{A}_t\boldsymbol{u}_{t-1}^{(r)})(\boldsymbol{u}_t^{(r)} - \boldsymbol{A}_t\boldsymbol{u}_{t-1}^{(r)})') \\
& + \lambda_0\sum_{t=1}^{T}\|\boldsymbol{A}_t\|_F^2 + \lambda_1\sum_{t=2}^{T}\|\boldsymbol{A}_t - \boldsymbol{A}_{t-1}\|_F^2
\end{aligned}
$$

where the optimization procedure is similar to that in the M-step. $\boldsymbol{Q}_0$ has an analytical solution $\boldsymbol{Q}_0 \leftarrow (1/q)\sum_{r=1}^{q}\boldsymbol{u}_0^{(r)}(\boldsymbol{u}_0^{(r)})'$, and $\{\boldsymbol{A}_t\}_{t=1}^{T}$ and $\boldsymbol{Q}_\epsilon$ can be updated in alternations. Given $\{\boldsymbol{A}_t\}_{t=1}^{T}$, $\boldsymbol{Q}_\epsilon$ has an analytical solution $\boldsymbol{Q}_\epsilon \leftarrow 1/(qT)\sum_{t=1}^{T}\sum_{r=1}^{q}(\boldsymbol{u}_t^{(r)} - \boldsymbol{A}_t\boldsymbol{u}_{t-1}^{(r)}))(\boldsymbol{u}_t^{(r)} - \boldsymbol{A}_t\boldsymbol{u}_{t-1}^{(r)})'$, and given $\boldsymbol{Q}_\epsilon$, $\{\boldsymbol{A}_t\}_{t=1}^{T}$ can be solved by gradient descent with backtracking line search, where the gradient is

$$
2\boldsymbol{Q}_\epsilon^{-1}(-\sum_{r=1}^{q}\boldsymbol{u}_t^{(r)}(\boldsymbol{u}_{t-1}^{(r)})' + \boldsymbol{A}_t\sum_{r=1}^{q}\boldsymbol{u}_{t-1}^{(r)}(\boldsymbol{u}_{t-1}^{(r)})') + 2\boldsymbol{D}_t
$$

and

$$\boldsymbol{D}_t = \begin{cases} \lambda_1(2\boldsymbol{A}_t - \boldsymbol{A}_{t+1} - \boldsymbol{A}_{t-1}) + \lambda_0 \boldsymbol{A}_t & \text{for } t = 2, \cdots, T-1; \\ \lambda_1(\boldsymbol{A}_1 - \boldsymbol{A}_2) + \lambda_0 \boldsymbol{A}_1 & \text{for } t = 1; \\ \lambda_1(\boldsymbol{A}_T - \boldsymbol{A}_{T-1}) + \lambda_0 \boldsymbol{A}_T & \text{for } t = T; \end{cases}$$

4.5.3.3 Computing the spatio-temporal covariance of the ROI mean activity and evaluating the marginal log-likelihood of sensor data

According to the auto-regressive model, given $\{\boldsymbol{A}_t\}_{t=1}^T, \boldsymbol{Q}_0, \boldsymbol{Q}_\epsilon$, we have

$$\boldsymbol{u}_t = \boldsymbol{A}_t \boldsymbol{u}_{t-1} + \boldsymbol{\epsilon}_t = (\prod_{\tau=t}^{1} \boldsymbol{A}_\tau)\boldsymbol{u}_0 + \boldsymbol{\epsilon}_t + \sum_{j=1}^{t-1}(\prod_{\tau=t}^{t-j+1} \boldsymbol{A}_\tau)\boldsymbol{\epsilon}_{t-j}$$

and the marginal covariance of $\boldsymbol{u}_t$ for $t = 1, \cdots, T$ is

$$\text{cov}(\boldsymbol{u}_t) = (\prod_{\tau=t}^{1} \boldsymbol{A}_\tau)\boldsymbol{Q}_0(\prod_{\tau=t}^{1} \boldsymbol{A}_\tau)' + \boldsymbol{Q}_\epsilon + \sum_{i=1}^{t-1}(\prod_{\tau=t}^{t-j+1} \boldsymbol{A}_\tau)\boldsymbol{Q}_\epsilon(\prod_{\tau=t}^{t-j+1} \boldsymbol{A}_\tau)'$$

where the in the product $\prod_{\tau=t}^{t-j+1}$, τ decreases from t to $t-j+1$. To compute the marginal covariance, it is convenient to first compute matrices $\tilde{\boldsymbol{A}}_{j,k} = \prod_{\tau=k}^{j} \boldsymbol{A}_\tau, j \leq k$, and then we have

$$\text{cov}(\boldsymbol{u}_t) = \tilde{\boldsymbol{A}}_{1,t}\boldsymbol{Q}_0\tilde{\boldsymbol{A}}'_{1,t} + \boldsymbol{Q}_\epsilon + \sum_{j=1}^{t-1} \tilde{\boldsymbol{A}}_{(t-j+1),t}\boldsymbol{Q}_\epsilon\tilde{\boldsymbol{A}}'_{(t-j+1),t}.$$

Let $\boldsymbol{U}_{p\times(T+1)} = [\boldsymbol{u}_0, \boldsymbol{u}_1, \cdots, \boldsymbol{u}_T]$ (where $\boldsymbol{u}_t$ is of size $p \times 1$); let vec($\boldsymbol{U}$) be the concatenation of the columns of $\boldsymbol{U}$. Let the $p(T+1)\times p(T+1)$ matrix $\boldsymbol{\Sigma}$ denote the covariance matrix of (vec($\boldsymbol{U}$)). This covariance can be computed as

$$\begin{aligned} &\boldsymbol{\Sigma}[tp+1:(t+1)p, (t+h)p+1:(t+h+1)p] \\ =&\text{cov}(\boldsymbol{u}_t, \boldsymbol{u}_{t+h}) \\ =&\text{cov}(\boldsymbol{u}_t, \boldsymbol{A}_{t+h}\boldsymbol{A}_{t+h-1}\cdots\boldsymbol{A}_{t+1}\boldsymbol{u}_t) \\ =&\text{cov}(\boldsymbol{u}_t)(\prod_{\tau=t+h}^{t+1} \boldsymbol{A}_\tau)' \\ =&\text{cov}(\boldsymbol{u}_t)\tilde{\boldsymbol{A}}'_{t+1,t+h} \end{aligned}$$

where by $\boldsymbol{\Sigma}[tp+1:(t+1)p, (t+h)p+1:(t+h+1)p]$, we mean the sub-matrix composed of the consecutive rows from $tp+1$ to $(t+1)p$ and the consecutive columns from $(t+h)p+1$ to $(t+h+1)p$ (indices are 1-based). If needed, we can also compute the marginal correlation between the mean of i_1th ROI at time t_1 and the mean of i_2th ROI at time t_2:

$$\text{correlation}(\boldsymbol{u}_{t_1}[i_1], \boldsymbol{u}_{t_2}[i_2]) = \frac{\boldsymbol{\Sigma}[t_1p+i_1, t_2p+i_2]}{\sqrt{\boldsymbol{\Sigma}[t_1p+i_1, t_1p+i_1]\boldsymbol{\Sigma}[t_2p+i_2, t_2p+i_2])}} \tag{4.36}$$

Next, we discuss how to evaluate the objective function that contains the marginal likelihood of the observed multi-trial sensor data and a prior term for the parameters (i.e., $\log f(\{\boldsymbol{y}_t^{(r)}\}_{t=0,r=1}^{T,q}) + \log f(\boldsymbol{\theta})$). Since $\log f(\boldsymbol{\theta})$ is easy to compute, we mainly focus on evaluating $-\log f(\{\boldsymbol{y}_t^{(r)}\}_{t=0,r=1}^{T,q})$. Let the $n \times (T+1)$ matrix $\boldsymbol{Y} = (\boldsymbol{y}_0, \boldsymbol{y}_1, \cdots, \boldsymbol{y}_T)$ denote the sensor time series in an arbitrary trial. Let vec$(\boldsymbol{Y})$ be a vector obtained by concatenating the columns in $\boldsymbol{Y}$.

Let $\tilde{\boldsymbol{C}}$ be an $nT \times pT$ block-diagonal matrix, where $\tilde{\boldsymbol{C}}[tn+1:(t+1)n, tp+1:(t+1)p] = \boldsymbol{C}$, for $t = 0, 1, \cdots, T$; $\tilde{\boldsymbol{C}}[tn+1:(t+1)n, tp+1:(t+1)p]$ denotes the intersection of the sub-rows (rows $tn+1$ to $(t+1)n$) and the sub-columns (columns $tp+1$ to $(t+1)p$). Similarly, let $\tilde{\boldsymbol{R}}$ be an $nT \times nT$ block diagonal matrix, where $\tilde{\boldsymbol{R}}[tn+1:(t+1)n, tn+1:(t+1)n] = \boldsymbol{Q}_\eta$, for $t = 0, 1, \cdots, T$. Then we have

$$\text{cov}(\text{vec}(\boldsymbol{Y})) = \tilde{\boldsymbol{C}}\boldsymbol{\Sigma}\tilde{\boldsymbol{C}}' + \tilde{\boldsymbol{R}}$$

and

$$-\log f(\{\boldsymbol{y}_t^{(r)}\}_{t=0,r=1}^{T,q}) \propto q \log\det(\text{cov}(\text{vec}(\boldsymbol{Y}))) + \sum_{r=1}^{q} \text{trace}(\text{cov}(\text{vec}(\boldsymbol{Y}))^{-1}\text{vec}(\boldsymbol{Y}^{(r)})\text{vec}(\boldsymbol{Y}^{(r)})')$$
$$+ \text{ some constant irrelevant to the parameters}$$

We use the following equations to facilitate the computation

$$\begin{aligned} \text{cov}(\text{vec}(\boldsymbol{Y}))^{-1} &= \tilde{\boldsymbol{R}}^{-1} - \tilde{\boldsymbol{R}}^{-1}\tilde{\boldsymbol{C}}(\boldsymbol{\Sigma}^{-1} + \tilde{\boldsymbol{C}}'\tilde{\boldsymbol{R}}^{-1}\tilde{\boldsymbol{C}})^{-1}\tilde{\boldsymbol{C}}'\tilde{\boldsymbol{R}}^{-1} \\ \log\det(\text{cov}(\text{vec}(\boldsymbol{Y}))) &= \log\det(\boldsymbol{\Sigma}^{-1} + \tilde{\boldsymbol{C}}'\tilde{\boldsymbol{R}}^{-1}\tilde{\boldsymbol{C}}) + \log\det(\boldsymbol{\Sigma}) + \log\det(\tilde{\boldsymbol{R}}) \end{aligned}$$

Chapter 5

Exploring the spatio-temporal neural dynamics in the human visual cortex

[1]

5.1 Introduction

Humans can effortlessly recognize objects and understand scenes. What computational processes in the visual cortex result in such proficiency? Decades of research has indicated that the visual cortex includes multiple functional brain areas, which have a topological mapping (i.e., "retinotopy") of the visual field [Engel et al., 1997]. Moreover, these areas are hypothesized to process visual inputs in a hierarchical manner [DiCarlo and Cox, 2007]. As illustrated in Figure 5.1, at a low level, neurons in the primary visual cortex (V1) at the posterior end of the brain have small receptive fields and extract low-level information such as local edge orientations [Hubel and Wiesel, 1968]. Neurons in the next level (V2) further process the V1 outputs and are sensitive to angles and junctions [Ito and Komatsu, 2004], as well as textures [Freeman et al., 2013]. Such a structure continues to next levels and diverges into two pathways [Mishkin et al., 1983]: a dorsal pathway towards the parietal cortex, related to motion perception, visuospatial processing, attention and action, and a ventral pathway towards the inferior temporal cortex, hypothesized as the neural basis of recognizing "what" we see (e.g., object/face/scene recognition). In the middle and anterior regions in the ventral pathway, neurons in the inferior temporal cortex show selectivity to complex shapes, invariance to scales and lighting, and encode semantic information of the visual input [Tanaka, 1996]. For example, in humans, previous fMRI studies have identified an object-selective functional area in the ventral pathway, the lateral occipital complex [Grill-Spector et al., 1999], which have greater responses to objects than to other visual stimuli (such as meaningless

[1] The work in this chapter was in collaboration with Dr. Elissa Aminoff, who co-designed and co-ran the experiments, ran the fMRI analysis, and gave important advice and help on the scientific interpretations.

patterns) and encode information to distinguish different categories of objects. Similarly, in the two pathways, at the levels that are higher than the early visual cortex, several scene-selective regions have been identified, including the parahippocampal/lingual region (known as the parahippocampal place area, "PPA"), the retrosplenial complex ("RSC"), and the occipital place area (also referred as the transverse occipital sulcus, "TOS") [Epstein et al., 1999; Epstein, 2008]. These regions have greater responses to scenes than to other stimuli and encode scene categories and other relevant semantic information (e.g., associations of objects in scenes [Aminoff and Tarr, 2015], physical sizes of scenes [Park et al., 2015], etc.).

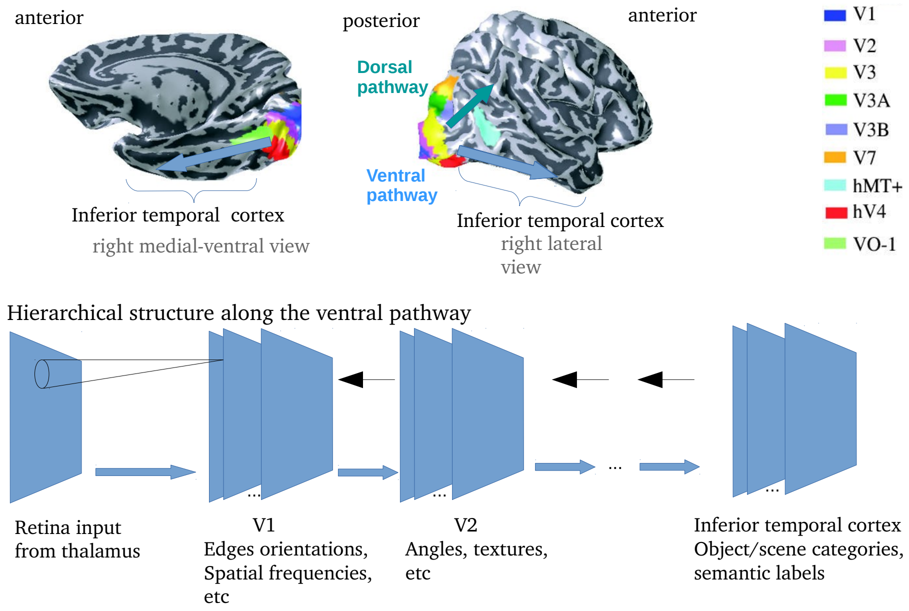


Figure 5.1: Illustration of the hierarchical organization of human visual cortex (with the mapping of various visual areas from Wandell et al. [2005]).

According to this hypothesis of hierarchical organization, from the posterior to the anterior parts along the ventral pathway, neurons at each level of the hierarchy receive inputs from previous levels and extract progressively higher-level information. Such a feedforward mechanism may be the key computation to recognizing what we see. On the other hand, there are also top-down anatomical connections between the areas along the inverse direction of the hierarchy, as well as connections from the frontal and parietal lobes [Felleman and Van Essen, 1991]. These connections indicate that there can be non-feedforward dynamics, such as top-down feedback, which can facilitate fast recognition. Besides the studies cited above, there is extensive experimental evidence supporting the hypothesis of hierarchical organization, from electrophysiology studies in primates and neuroimaging studies in humans (e.g., [Yamins et al., 2014; Cichy et al., 2016b; Aminoff et al., 2015]). However, the hypothesis does not specify detailed dynamics in the visual cortex. For example, due to complex connections among regions, information may flow along the feedforward or feedback directions at different temporal stages, or even flow in a recurrent manner. Indeed, details of such dynamics can be the key to proficient recognition.

In order to better understand the information flow and thus build better computational models of the visual cortex, we need to answer two questions on the joint spatio-temporal neural activities: *(1) what kind of information is extracted at different temporal stages and different brain locations and (2) how neural activities in different areas interact with each other dynamically*. To answer these questions, we need to record dynamic neural activity in a non-invasive manner, especially while humans are processing naturalistic visual scenes as they do in their daily life. In particular, MEG and EEG provide millisecond-level temporal resolution, which is necessary for capturing fast neural dynamics in object and scene recognition. In addition, using the source localization techniques and the source-space analysis methods we developed, we can also get intermediate spatial resolution from MEG and EEG. Therefore, to answer the two questions, we use MEG and EEG to record dynamic neural responses, while human participants process a large number of naturalistic scenes.

With regard to the first question—*what kind of information is extracted at different temporal stages and spatial locations*, a straightforward analysis is to regress the spatio-temporal neural activity on different candidate features that describe different levels of information. However, defining the candidate features is non-trivial. Recently, in computer vision, a type of artificial neural network model, known as convolutional neural networks (CNNs), has demonstrated a great capability of learning diagnostic features for categorizing objects or scenes [Krizhevsky et al., 2012]. These CNN models are inspired by the feedforward structure in the hypothesis of the hierarchical organization of the visual cortex. They typically have multiple layers with increasing receptive field sizes; the first layer takes raw image inputs and the last layer outputs object or scene category labels. The connections between layers are learned by minimizing labeling error on a large amount of image data, and the layered structure inherently provides an operational definition of low-level and high-level information. Features extracted in these CNNs have been shown to share significant similarity with neural representations of objects and scenes [Yamins et al., 2014; Khaligh-Razavi and Kriegeskorte, 2014; Cichy et al., 2016c,a].

Based on these observations, we answer Question (1) by using low- and high-level features from a well-trained CNN and linearly regressing the neural responses at different locations and time points on these features. Such a regression analysis can examine how well the neural responses are explained by the features and yield a spatial-temporal profile of the linear dependence (which we also call "correlation profile" hereafter) between neural responses and different features. Although some recent MEG studies have also correlated neural activities with CNN features or other computer vision features, these studies either focused only on recognizing isolated objects, restricted to individual objects on blank backgrounds [Clarke et al., 2014; Cichy et al., 2016c], or focused on specific properties of scenes, restricted to a small set of scene stimuli [Cichy et al., 2016a]. In addition, the majority of these studies focused on the temporal patterns, but not the spatial patterns (except Clarke et al. [2014]). To overcome these limitations, we recorded neural responses to a relatively large number of naturalistic scenes in daily life and applied source-space regression analysis models (including our novel STFT-R model), aiming to obtain detailed correlation profiles between spatio-temporal neural activities and low- and high-level CNN features. Such profiles, which have not been documented so far to our knowledge, will be helpful for understanding the information flow during scene and object recognition. In addition, because the low- and high-level features from CNNs are defined operationally, features at different levels may share some components in common, resulting in some difficulty in interpretation. Uniquely, we decomposed the low- and high-level CNN features into three lower-dimensional groups, respectively representing (i) the low-level features that are relevant to high-level features, (ii) the low-level features that are roughly orthogonal to high-level features, and (iii) the high-level features that are roughly orthogonal to low-level features. By specifically comparing the spatio-temporal correlation profiles of these three groups, we not only observed evidence for feedforward processing in the visual cortex, but also found interesting evidence for non-feedforward processing, which could reflect top-down feedback.

Question (2)—*how neural activities in different areas interact with each other dynamically*—is more directly related to information flow in the brain. To address this question, we quantify the functional connectivity, that is, statistical dependence, between different regions of interest. Although functional connectivity is observational and does not directly translate into causal interactions, analysis of functional connectivity, especially time-lagged functional connectivity, may reveal directed interactions between regions at important time windows, which are helpful in generating candidate hypothesis of information flow for further causal tests (e.g., interventions of brain areas in animal studies).

Previous work has suggested that functional connectivity between scene-selective regions and other brain areas may play an important role in scene processing [Kveraga et al., 2011]. More specifically, a scene includes multiple objects that are contextually associated; strong contextual objects (e.g., oven) may elicit neural responses related to the corresponding scenes (e.g., kitchen). Functional MRI studies have shown that the scene-selective regions, and the medial prefrontal cortex (mPFC), which is involved in episodic memory, have greater responses to strong-contextual objects (e.g., oven) than to weak-contextual objects (e.g., pen) [Bar and Aminoff, 2003]; these regions together are hypothesized as a connected "contextual network" by Bar and Aminoff [2003]. Kveraga et al. [2011] observed higher synchronization of neural activity among regions in the contextual network and the early visual cortex, while human participants processed strong-contextual objects than while they processed weak-contextual objects. This result indicates that interactions among these regions (e.g., top-down influences) may facilitate associative processing when humans recognize scenes. To further explore whether this speculation is true, we analyzed functional connectivity while our participants processed the naturalistic scene stimuli. Considering the underdetermined nature of the source localization problem, we restricted the analysis to certain functional regions of interest, including the object/scene-selective regions, the early visual cortex, and regions in or near the mPFC in the "contextual network" dubbed by Bar and Aminoff [2003]. Using a time-lagged autoregressive model in Chapter 4, we specifically studied whether the earlier activity in one region was able to predict later activities in other regions, among the aforementioned regions in the ventral visual pathway and the mPFC. Our results demonstrated leading and lagged linear dependence of neural activities among some of these regions and provide insights on directions of information flow during naturalistic scene processing.

5.2 Materials and methods

5.2.1 Participants

Eighteen healthy adults (8 females, 1 left-handed, aged 18-30) participated in the MEG and fMRI sessions, and fifteen of them (6 females, 1 left-handed, aged 20-30) participated in a following EEG session. All participants gave written informed consent and were financially compensated for their participation. All procedures followed the principles in the Declaration of Helsinki and were approved by the Institutional Review Boards of Carnegie Mellon University and the University of Pittsburgh.

5.2.2 Stimuli

The stimulus set presented in the MEG and EEG sessions included color images (photographs) of 181 scene categories (e.g.,“airport”,“beach”,“coffee shop”, etc). There were 2 exemplar images in each category, resulting in 362 images in total. These images were obtained from the dataset of the “Never Ending Imaging Learner” [Chen et al., 2013] [2], which were collected from the Internet. The scene images varied in aspect ratios; the longer dimension between the width and the hight was set to 500 pixels, and the images were placed in the center of a 600×600 bounding box with gray value $= 135$ (out of 255). Figure 5.2 shows two example images. The images (with the gray bounding boxes) were presented roughly at a visual angle of $10^\circ \times 10^\circ$ in MEG and $12^\circ \times 12^\circ$ in EEG.

In the fMRI session (see below), to define the scene-selective regions in the cortex, we also ran a functional localizer experiment, in which a separate set of images were used. These images included 60 color images from each of the three conditions: scenes, objects and phase-scrambled pictures of the scenes. The scene images included outdoor and indoor scenes, which did not overlap with the 362 stimuli images mentioned above. The objects used were weak contextual objects [Bar and Aminoff, 2003]. The phase-scrambled pictures served as control stimuli of scenes; they were generated by doing a Fourier transform of the scene images, scrambling the phases, and then doing an inverse Fourier transform back to the pixel space. The images were presented at a visual angle of $5.5^\circ \times 5.5^\circ$.

[2] `www.neil-kb.com`

5.2.3 Experimental procedure

The experimental procedures in the MEG, EEG and MRI sessions were all implemented using `Psychtoolbox 3`[3] in `MATLAB`. The trial structure in the MEG/EEG sessions were described in Figure 5.2. Before the stimulus presentation, the participants were asked to fixate their eyes on a white "+" symbol (spanning 80 pixels) at the center of the screen. The gray value of the screen was 180 (out of 255). We term this screen the fixation screen hereafter. Then one stimulus image was presented at the center of the screen for 200 ms, which was short enough to reduce artifacts due to saccades during stimulus presentation. The stimulus image was then followed by the fixation screen lasting for a random duration until the onset of the next stimulus. This duration was uniformly sampled from a $(1600, 2100)$ ms interval independently in each trial. The participants were asked to perform a "one-back" task —to respond when the current stimulus was the same as the last presented one, by pressing a button on a glove pad using the right index finger (MEG) or by pressing the "space" key on a keyboard (EEG). The participants were given 1500 ms to respond after the stimulus onset.

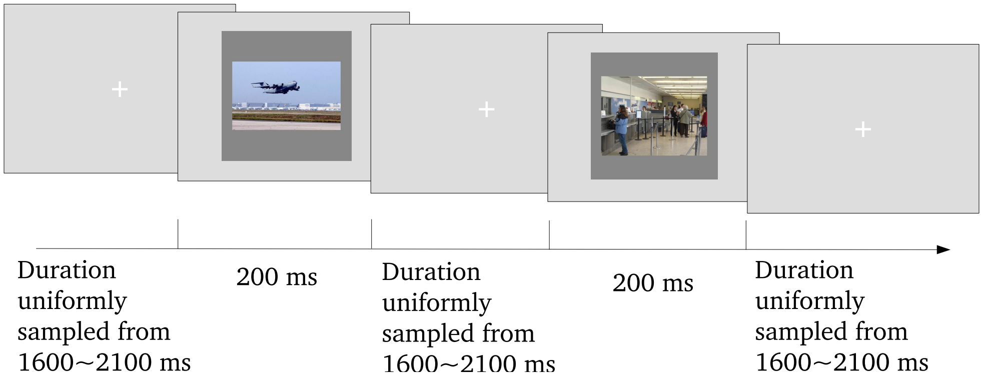

+
+
+
Duration uniformly sampled from 1600~2100 ms
200 ms
Duration uniformly sampled from 1600~2100 ms
200 ms
Duration uniformly sampled from 1600~2100 ms

Figure 5.2: Illustration of the trial structure in the MEG/EEG sessions

Each MEG session included 6 to 12 runs; each run included 181 images, and every two runs covered all 362 images. In each run, 10% of the images were immediately repeated once, and we only analyzed data in the first presentation of each image. In total, there were 3 to 6 repetitions of each image for most of the participants in the MEG session. For one participant, due to a problem with acquisition, we only presented 2 repetitions of some images, and 4 repetitions of the remaining images. Each EEG session included 6 runs; each run included all 362 images, and there was a self-paced break between every 40 trials. Similarly, 10% of the images were immediately repeated once, and there were 6 repetitions of each image for each participant.

In the fMRI session, a functional localizer experiment was run to independently define scene- and object- selective regions in the brain. Most participants went through two runs of the localizer experiment; yet two participants went through one run due to time limits. Each run started and

[3] `http://psychtoolbox.org/`

ended with a 12-second time window, during which a black fixation cross ("+") was presented on a gray background. Between the starting and ending fixation windows, there were twelve 14-second to 16-second blocks, each of which had stimuli in one condition (either scenes or weak-contextual objects or phase-scrambled scenes). There were four blocks per condition and three conditions in total and there was a 10-second fixation time window between each two consecutive blocks. In each block, 14 stimuli were presented in a row, with an 800 ms presentation duration and a 200 ms inter-stimulus interval, with the exception that the first stimulus in each of the block other than the first block was presented for 2800 ms, yielding 16-second blocks [4]. Among the 14 stimuli, 12 were unique images, and 2 were immediate repetitions of the previous image. Again, the participants were instructed to do a one-back task — pressing a button on a glove pad with the right index finger when they detected an immediate repetition.

5.2.4 Data acquisition

MEG and EEG. MEG data was collected using a 306-channel whole-head MEG system (Elekta Neuromag, Helsinki, Finland) at the Brain Mapping Center at the University of Pittsburgh. The MEG system had 102 triplets, each consisting of a magnetometer and two perpendicular gradiometers. The recordings were acquired at 1 kHz, high-pass filtered at 0.1 Hz and low-pass filtered at 330 Hz. Four head position indicator (HPI) coils were placed on the scalp of each participant to record the position of the head in relation to the MEG helmet. Empty room MEG recordings were also collected in the same session, and used to estimate the covariance matrix of the sensor noise. About a hundred points describing the shape of the head and the coordinates of the HPI coils on the scalp were collected using a digitization device; these coordinates were later used in aligning the head position in the MEG session with the structural MRI scan.

EEG data was collected using a 128-channel whole-head system (ActiveTwo, Biosemi, Amsterdam, Netherlands) at the EEG laboratory of the Psychology Department at Carnegie Mellon University. The recordings were acquired at 512 Hz. The coordinates of the electrode holes on the EEG cap, which the electrodes were plugged in, were collected using a digitization device. For a subset of the participants, additional 100 coordinates describing the shape of the head with the cap on were also collected using the same device. These coordinates were later used in aligning the head position in the EEG session with the structural MRI scan.

For both MEG and EEG, electrooculogram (EOG) was monitored by recording muscle activity above and below one eye and lateral to both eyes; electrocardiography (ECG) was recorded by placing additional electrodes above the chest. The EOG and ECG recordings captured eye blinks

[4] The long presentation was due to a timing issue in the customized image presentation program. However, we do not think this issue would have changed our results in defining the object/scene-selective regions, because in this block design, the main effect is the difference of neural responses in different conditions, which should be robust to variations in presentation durations of individual stimuli.

and heartbeats, so that these artifacts could be removed from the MEG or EEG recordings afterwards.

MRI. Magnetic resonance imaging (MRI) data was collected on a 3T Siemens Verio MR scanner at the Scientific Imaging and Brain Research Center at Carnegie Mellon University using 32-channel head coil. For each participant, a high resolution structural MRI scan was first acquired (T1-weighted MPRAGE sequence, 1 mm × 1 mm × 1 mm, 176 sagittal slices, TR = 2300 ms, TE = 1970 ms, flip angle = 9°, GRAPPA = 2, field of view = 256). In addition, functional MRI (fMRI) data was also collected for the functional localizer experiment ($T2^*$ -weighted echo-planar imaging multiband pulse sequence, 69 slices aligned to the AC/PC, in-plane resolution 2 mm × 2 mm, 2 mm slice thickness, no gap, TR = 2000 ms, TE = 30 ms, flip angle = 79 °, multi-band acceleration factor = 3, field of view 192 mm, phase encoding direction $A \gg P$, ascending acquisition). A fieldmap scan was also acquired to correct for distortion effects using the same slice prescription as the echo-planar imaging scans (69 slices aligned to the AC/PC, in-plane resolution 2 mm × 2 mm, 2 mm slice thickness, no gap, TR = 724 ms, TE1 = 5 ms; TE2 = 7.46 ms, flip angle = 70°, field of view 192 mm, phase encoding direction $A \gg P$, interleaved acquisition).

5.2.5 Preprocessing of MEG/EEG data

Preprocessing of the raw MEG or EEG recordings included steps in the following pipeline. All steps were implemented using the MNE-python[Gramfort et al., 2014] package in `Python`.

1. *Filtering*. The raw recordings (including MEG empty room recordings) were filtered with a $1 - 110$ Hz bandpass filter, which removed low-frequency drifts and higher frequency noise (such as the oscillations generated by the head position indicator coils during head tracking). Then the recordings were further filtered by a notch filter centered at 60 Hz to remove the interference of the power line.

2. *Removing artifacts due to eye blinks and heartbeats*. Independent component analysis (ICA) was used to decompose the recordings into multiple components, and when reconstructing the data from the components, the ones highly correlated with eye blinks and heartbeats (recorded by EOG and ECG) were removed.

3. *Interpolation of bad channels*. By manually inspecting the raw data, bad channels that had extremely high or low variances or very frequent square-wave-like artifacts were marked, and data in these channels was interpolated using neighbouring good channels.

4. *Obtaining trial-by-trial data*. In both MEG and EEG sessions, trial-by-trial recordings (also referred as "epochs") were obtained by segmenting the data from -100 ms to 900 ms with respect to the "trigger" onset (the "stimulus onset" recorded in the acquisition system, defined as 0 ms). For each trial, each channel, the mean across time points in the baseline window (-100 to 0 ms) was subtracted from the recording at each time point. It is worth noting that the timing here was recorded by the data acquisition devices. Yet the image

presentation devices had additional delays. In MEG, the projector used for image presentation had a constant delay for about 40 ms, and in EEG the LCD screen had a delay for roughly about 20 ms. We shifted the time points back for 40 ms in MEG, and 20 ms for EEG, such that the timing in MEG and EEG were roughly aligned. For MEG data, a signal space projection (SSP) was applied to the epochs. The SSP constructed a low-dimensional linear subspace characterizing the empty room noise (via principal component analysis), and removed from the experimental MEG recordings the projection onto this subspace, so that neural signals orthogonal to the principal components of empty room noise remained. The EEG epochs were re-referenced to the average across all 128 channels, that is, the averaged readings across channels was subtracted off from each channel at each time point.

5. *Obtaining averaged neural responses to each image for each participant.* To reduce the computational cost for some of the analyses below, we down-sampled the trial-by-trial data to 100 Hz sampling rate. The trials corresponding to the second presentation in immediate repetitions might have lower signal strength due to adaptation, therefore they were removed from further analysis. To remove outlier trials that had extreme large variations in each session for each participant, we computed the difference between the maximum and minimum of the recordings for each channel in each trial, and discarded the trials where the difference was larger than 15 standard deviations plus the mean across all trials for at least one channel. Finally, the data in the remaining trials that corresponded to the same image was averaged within each session for each participant.

6. *Removing data explained by nuisance covariates.* Although our stimulus images were wrapped in the same 600×600 pixel boxes, the widths and heights of the images varied. These nuisance covariates that were irrelevant to the image contents could explain a significant amount of variance in the MEG/EEG recordings. To alleviate such effects, we regressed the MEG/EEG data against four covariates —image width, image height, area (width $\times$ hight) and aspect ratio (width $/$ height) —respectively at each time point for each sensor in each participant. An all-one column was added in the regressors to remove the mean response across all images as well. For each regression, the residuals were kept as new sensor data to be analyzed.

5.2.6 Forward modeling

For each participant, based on the T1-weighted structural MRI scan, the outer skin surface of the scalp and the inner and outer surfaces of the skull were computed using the watershed algorithm [Ségonne et al., 2004] implemented in the `Freesurfer` software [Fischl et al., 2002] and the `MNE` software. Additionally, the cortical surfaces that segmented the gray and white matter were also computed using `Freesurfer`. The source space was defined as about 8000 distributed "dipoles" (or source points) that pointed perpendicular to the cortical surface of both hemispheres. The average spacing between source points was 4.9 mm, yielding 24 mm^2 of cortical surface area per source point. Source points that were within 2.5 mm of the inner skull surface were excluded.

The digitized points that described the shape of the head in MEG were used for co-registration with the structural MRI. In the co-registration, we solved for a rigid-body transformation that minimized the sum of squared distances from the digitized points to the scalp surface, using an interface implemented in `MNE-C`. Note that the optimization problem was not necessarily convex, therefore no global minimum could be guaranteed. Yet, by manually adjusting initial values, the solution for each participant looked reasonably good in our visual inspection. For EEG sessions, similar distance minimization did not work well. This was possibly because the digitization device could not penetrate the EEG cap to reach the scalp, leaving a few millimeters between the digitized points and the scalp surface, such that many possible alignments could result in similar sums of squared distances. In this case, an alignment that looked most visually reasonable was manually selected for each participant. Due to the difference in the co-registration step between the MEG and EEG sessions, the source-space analysis of our EEG data might be less reliable than that of the MEG data.

For each participant, the forward matrix for each run in the MEG session and for the whole EEG session was computed using the boundary element model implemented in `MNE-C`, after transforming the MEG sensor locations or EEG electrode locations into the structural MRI space, based on the alignment in the co-registration step above. For the MEG session, the forward matrices across all runs were averaged for each participant.

5.2.7 Source-space regression analysis

To characterize how much the spatio-temporal neural activity was correlated with CNN features, we regressed the neural responses in the source space against the features of the images, using both the conventional two-step regression analysis, and our novel short-time Fourier transform regression (STFT-R) model in Chapter 3. In both cases, the noise covariance matrix $\boldsymbol{Q}_e$ was estimated from sensor recordings within the baseline time windows (-140 to -90 ms in MEG and -120 to -70 ms in EEG) for each participant. For the two-step analysis, we used the dSPM source localization method implemented in `MNE-python` to obtain unit-less source current dipole estimates for each image. The penalization parameter was set to 1.0. For each participant, after the dSPM solutions were obtained for each image, an ordinary least square regression was run for each time point and each source point. The coefficient of determination (or R-squared), indicating the proportion of variance explained by the regressors (CNN features) was used as the summarizing statistics to characterize the correlation between the neural responses and the CNN features.

In the one-step analysis, we applied the short-time Fourier transform regression method [Yang et al., 2014] introduced in Chapter 3, with the hierarchical sparsity-inducing penalty on the regression coefficients, which related time-frequency components of source activity to the regressors. There were three levels of nested grouping in the sparsity-inducing penalty: the source point level, the time-frequency level and the level of individual dimensions of the regressors. On the first level, the regression coefficients for each source point formed a single group, on the second level, each time-frequency component corresponded to a group, and on the third level, each dimension of the regressors corresponded to one group as well. The penalization parameters of the penalty on the three levels (i.e. α, β and γ in Chapter 3) were selected among a range of values via two-fold cross validations for each participant respectively. The time step of the STFT was set to $\tau_0 = 40$ ms, and the window length was 160 ms. Neural responses to all 362 images were used, and after fitting the STFT-R models, the regression coefficients for each source point in the time-frequency domain were transformed to the time domain for further statistical analysis.

5.2.8 Preprocessing of fMRI data and definition of regions of interest (ROIs)

The fMRI localizer data were preprocessed in `SPM12` [5]. The preprocessing included an unwarp transform to correct for geometric distortions using the fieldmap scan, a frame-by-frame transform to correct for head motion combined with a transform to align with the structural MRI, and finally spatial smoothing with an isotropic Gaussian kernel (where the full width at half maximum was 4 mm). The data in all localizer runs for each participant were concatenated, high pass filtered with 0.0078125 Hz (a 128-second period), and then analyzed using a "general linear model" in a block design. In this model, the time series at each voxel was essentially linearly regressed against a design matrix, which included, in separate columns, the pre-defined canonical hemodynamic response function convolved with the square-wave-like indicators of blocks for each stimulus condition (scenes, weak contextual objects, and phase scrambled scenes). The design matrix also included extra columns corresponding to nuisance covariates (e.g., the time series of parameters in motion correction). An autoregressive model of order 1 (AR(1)) was used to account for the temporal correlations of the residuals.

Scene/object selective regions were defined using the `MarsBaR` toolbox [6]. For any voxel, let β_{scene} denote the regression coefficient for the scene condition, β_{object} for the weak-contextual object condition, and finally β_{scramble} for the phase-scrambled scene condition. The t-statistics of the difference ($\beta_{\text{scene}} - 1/2(\beta_{\text{object}} + \beta_{\text{scramble}})$) was computed as the estimated difference divided by the estimated standard deviation of the difference. Then the voxels where the t-statistics was above a threshold were selected as scene-selective voxels. A customized threshold was set for each participant, such that clusters of contiguous voxels above the threshold were identified within or in the proximity of the parahippocampal gyrus, the retrosplenial cortex, and the transverse occipital sulcus in each hemisphere. These clusters were labelled as the scene-selective ROIs, which were the parahippocampal place area (PPA), the retrosplenial cortex (RSC) and the transverse occipital sulcus (TOS). For the majority of the participants, the threshold was equal to the value where the family-wise error rate was controlled at 0.05. For some individuals, if the threshold was too stringent, we relaxed the threshold to a smaller value. Similarly, object-selective clusters in the

[5] `http://www.fil.ion.ucl.ac.uk/spm/software/spm12/`

[6] `http://marsbar.sourceforge.net/index.html`

lateral occipital cortex and the fusiform area (the lateral occipital complex or LOC) were also identified in both hemispheres for each participant respectively, where the difference of interest was $\beta_{\text{object}} - \beta_{\text{scramble}}$.

The SUMA software was used to project these clusters of voxels to sets of vertices on the cortical surfaces, so that they could be labelled in the source space in MEG/EEG. The mapped sets of vertices were then manually examined, and the ones that had fuzzy boundaries or that were anatomically off were corrected. Each remaining contiguous set was defined as one region of interest (ROI). The ROIs that did not cover at least 10 source points were dilated until they could cover 10 source points. Finally, the vertices in the LOC that were also in one of the scene-selective ROIs (PPA, RSC or TOS) were removed from the LOC. See Figure 5.14 for the locations of the ROIs in each participant.

Additionally, some anatomical ROIs were defined based on the parcellation of the structural MRI by Freesurfer. These ROIs included the pericalcarine areas that covered the early visual cortex (EVC) [7], and the medial orbitofrontal cortex (mOFC), which was roughly in or near the location of the mPFC in the "contextual network" introduced by Bar and Aminoff [2003]. Here, we use the mOFC as an approximation of the mPFC. Moreover, we also defined a larger ventral visual ROI, as a union of the following regions in the "aparc" parcellation of anatomical ROIs in Freesurfer: the entorhinal, fusiform, inferiortemporal, lateraloccipitical, lingual and parahippocampal regions and the temporal pole. We obtained the coordinates along the anterior and posterior axis of the source points in this ventral visual ROI, segmented the range of the coordinates into 3 sub-ranges of equal lengths, and finally partitioned the source points into 3 groups corresponding to these sub-ranges. Hence, we had 3 sub-ROIs in the ventral visual cortex, named as Vent 1, Vent 2 and Vent 3 from the posterior side to the anterior side, as shown in Figure 5.3. Finally, each pair of the corresponding bilateral regions were merged into one ROI, resulting in the bilateral PPA, RSC, TOS, LOC, EVC, mOFC and Vent 1, 2 and 3.

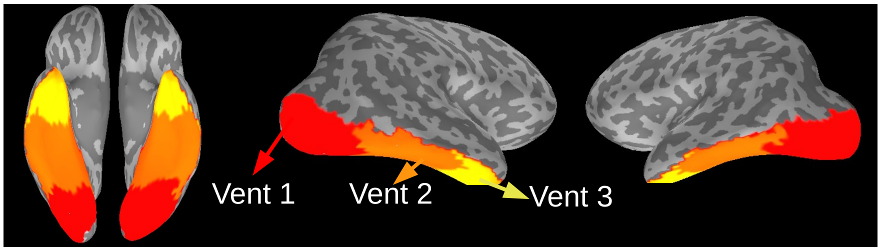


Figure 5.3: Illustration of the three ROIs in the ventral visual cortex from one example participant.

5.2.9 Extracting features of the images

We used a convolutional neural network called Alexnet [Krizhevsky et al., 2012] implemented in the Caffe [Jia et al., 2014] software to extract features. This convolutional neural network

[7] the pericalcarine areas included mostly V1 but might also include some part of V2

was trained to classify images into 1000 object categories with 1.2×10^6 training samples and 5×10^4 validation samples in the `ImageNet` database [Deng et al., 2009; Russakovsky et al., 2015]. Figure 5.4 shows the 8-layer architecture of `Alexnet`. The first 5 layers had convolutional units. Each unit in these layers applied a dot product of a "kernel" weight matrix with the inputs within a receptive field—for example, in Layer 1, each unit's receptive field was 11×11 pixels of the RGB channels of the raw image. Within each of the convolutional layers, there were a number of sub-layers; for example, Layer 1 had $48 \times 2 = 96$ sub-layers. Units in each sub-layer shared a "kernel" weight matrix, and therefore in such a convolutional architecture, the number of parameters was much smaller than that of an all-to-all connected neural network with the same number of units, yielding a more tractable model to train. After the convolution operation (dot product), each unit then applied a rectified linear function $\mathfrak{f}(x) = \max(0, x)$ on the dot product to generate the output. In Layer 1, 2, and 5, an additional normalization step was applied, where in each location, the output of the unit in each sub-layer unit was normalized by a function of the sum of squares of the responses in its neighbor sub-layers, including itself; a max-pooling operation was also added after the normalization in these layers (see Krizhevsky et al. [2012] for more details). Layer 6 and 7 were fully connected layers, each consisting of $2048 \times 2 = 4096$ units, and Layer 8 was the final output layer with 1000 units, corresponding to the 1000 object categories. We downloaded a version of pretrained `Alexnet` from the `Caffe` "model zoo", and resized our 600×600 images (with the gray box) to the input size, then collected responses of all the units in each layer as our features. For Layer 1, 2 and 5, we used the responses before normalization and max-pooling.

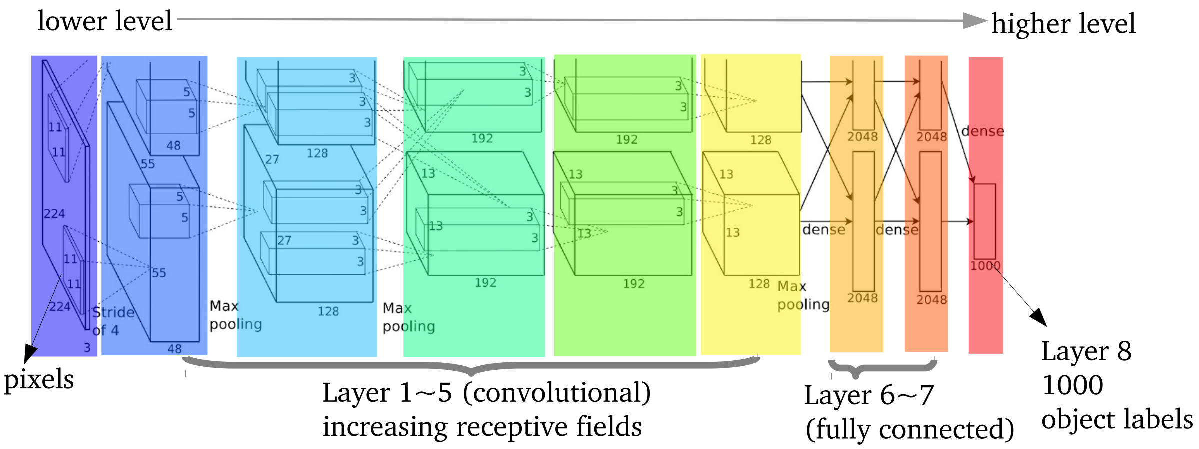


Figure 5.4: Structure of `Alexnet`. Note that the network was distributed on two graphics processing units (GPUs), illustrated by the upper and lower rows for each Layer. In terms of extracted responses of the units, we concatenated the responses of all the units in both the upper and lower parts together.

The 8-layer structure of `Alexnet` is feedforward; it naturally provides a progressive shift from low-level to high-level features. Here we chose Layer 1 and Layer 7 as representatives of low-level features and high-level (object-category-related) features. Layer 1 had the smallest receptive field and the convolutional "kernel" weight matrices were similar to 2-D Gaussian functions and Gabor filters (see visualization in Krizhevsky et al. [2012]). Layer 7 was the last fully-connected hidden layer before the output layer, which should represent important semantic information about the images that facilitates object recognition. Here, we did not use Layer 8 because the 1000 category labels in ImageNet might not align with the basic organization of object categories in humans.

We only had 362 images in the stimulus set, which was limited for analyzing the features per se (e.g., comparing different layers). Therefore we created a set of extra images, from which we also extracted features from the same layers. These extra images were from the same 181 scene categories in the same dataset, including 6 exemplars per category, different from the stimulus images ($6 \times 181 = 1086$ images in total); they had similar sizes (longest side = 500 pixels) as the stimulus images, and were also centered in the same 600×600 bounding boxes.

It is also worth noting that the nuisance covariates due to various widths and heights of the images could also be correlated with the extracted CNN features. Therefore we regressed out the width, height, area (width × hight) and aspect ratio (width / height) for the features extracted in each unit across the union of the stimulus set and extra image set. Again, an all-one column was included in the regression, which removed the mean across all images.

We also extracted simpler low-level features—the `local contrast` features—in the following way. For each image resized to the input size of `Alexnet`, the patch within each 11×11 receptive field in Layer 1 was converted to gray values (by averaging the values of the RGB channels), and the contrast in the patch was defined as $(x_{\max} - x_{\min})/(x_{\max} + x_{\min})$, where $x_{\min}$ and $x_{\max}$ were the minimum and maximum of the gray values in the patch. The contrast values were concatenated across all receptive fields, yielding a $55 \times 55 = 3025$-dimensional vector for each image. When using these local contrast features as regressors in our analysis of neural data, linear projections onto the nuisance covariates related to the widths and heights of the images were removed.

We observe that the low-level and high-level layers in `AlexNet` are both highly dependent on the local contrast, which would naturally elicit strong neural responses in the visual cortex (see Results in Section 5.3 below). To account for possible confounding effects, we regressed out the first 160 principal components of local contrast features (explaining 90% of the variance in the local contrast features) from the intact features in all layers in `AlexNet`, and took the residuals as new features of interest. We term these new features local-contrast-reduced features hereafter. Note that the local contrast features across both stimulus images and extra images were used in the PCA to obtain the 160 principal components, and similarly the features in each layer across all images went through the aforementioned regression.

5.2.10 Characterizing the common linear space of two feature sets

From Layer 1 to Layer 7, the units in `AlexNet` non-linearly transform the raw pixel inputs to informative features related to object categories. However, this does not necessarily mean the features in Layer 1 and Layer 7 are completely orthogonal. There may still be some linear dependence between the low-level Layer 1 and the much higher-level Layer 7. In this case, when we compared the correlation of neural activity with the two layers, it would be more insightful to estimate how much the correlation was due to the linearly dependent components between the two layers. Hence we exploited the following canonical correlation analysis to extract the common linear components between Layer 1 and Layer 7. Let matrices $\boldsymbol{X}_1$ $(q \times p_1)$ and $\boldsymbol{X}_2$ $(q \times p_2)$ denote two sets of features of the same q images, extracted from p_1 units in Layer 1 and p_2 units in Layer 7 of `Alexnet`. We assumed that the rows in $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ had zero means, which was empirically satisfied by subtracting the sample mean across q images. To characterize the common linear space of the two feature sets, we used canonical correlation analysis to find linear projections of each feature set that maximized the Pearson correlations of the projected features of the two sets.

Because there were more units in the layers of `AlexNet` than the number of images $(q < p_1, p_2)$, we first used principal component analysis (PCA) to reduce the dimensions of both feature sets to $p < p_1, p_2$. With the assumption that both $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ had zero mean, the PCA was implemented using singular value decompositions,

$$\underset{q\times p_1}{\boldsymbol{X}_1} \approx \underset{q\times p}{\boldsymbol{U}_1} \quad \underset{p\times p}{\boldsymbol{D}_1} \quad \underset{p\times p_1}{\boldsymbol{V}_1^T}, \qquad \underset{q\times p_2}{\boldsymbol{X}_2} \approx \underset{q\times p}{\boldsymbol{U}_2} \quad \underset{p\times p}{\boldsymbol{D}_2} \quad \underset{p\times p_2}{\boldsymbol{V}_1^T}$$

where $\boldsymbol{D}_1$ and $\boldsymbol{D}_2$ were diagonal matrices, and $\boldsymbol{U}_1$, $\boldsymbol{U}_2$, $\boldsymbol{V}_1$ and $\boldsymbol{V}_2$ had orthonormal columns.

We used the projections of $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ onto the p orthogonal dimensions $\tilde{\boldsymbol{X}}_1 = \boldsymbol{U}_1\boldsymbol{D}_1$ and $\tilde{\boldsymbol{X}}_2 = \boldsymbol{U}_2\boldsymbol{D}_2$ in the canonical correlation analysis, where linear weights $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$ (both $p \times p$) were obtained as follows.

$$\begin{aligned}
&\underset{\boldsymbol{W}_1[:,i],\boldsymbol{W}_2[:,i]}{\arg\max}\ \text{corr}(\boldsymbol{X}_1^\dagger[:,i], \boldsymbol{X}_2^\dagger[:,i]) \text{ where } \boldsymbol{X}_1^\dagger = \tilde{\boldsymbol{X}}_1\boldsymbol{W}_1, \boldsymbol{X}_2^\dagger = \tilde{\boldsymbol{X}}_2\boldsymbol{W}_2\\
&\text{subject to corr}(\boldsymbol{W}_1[:,i], \boldsymbol{W}_1[:,j]) = 0 \text{ if } i \neq j\\
&\qquad\qquad\ \text{corr}(\boldsymbol{W}_2[:,i], \boldsymbol{W}_2[:,j]) = 0 \text{ if } i \neq j\\
&\qquad\qquad\ ||\boldsymbol{X}_1^\dagger[:,i]||_2 = ||\boldsymbol{X}_2^\dagger[:,i]||_2 = 1\\
&\qquad\qquad\ i, j = 1, \cdots, p
\end{aligned}$$

The ith columns of $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$, ($\boldsymbol{W}_1[:,i]$ and $\boldsymbol{W}_2[:,i]$) were the weights that linearly combined columns in $\tilde{\boldsymbol{X}}_1$ and $\tilde{\boldsymbol{X}}_2$, such that the combinations $\boldsymbol{X}_1^\dagger[:,i]$ and $\boldsymbol{X}_2^\dagger[:,i]$ had the highest correlation. Different columns in $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$ projected $\tilde{\boldsymbol{X}}_1$ and $\tilde{\boldsymbol{X}}_2$ onto orthogonal components. This optimization problem was solved using the `canoncorr` function in `MATLAB`, which implemented eigen decompositions of ${\boldsymbol{\Sigma}_{11}}^{-1/2}\boldsymbol{\Sigma}_{12}{\boldsymbol{\Sigma}_{22}}^{-1}\boldsymbol{\Sigma}_{21}{\boldsymbol{\Sigma}_{11}}^{-1/2}$, where $\boldsymbol{\Sigma}_{12}$ and $\boldsymbol{\Sigma}_{21}$ were cross covariances of $\tilde{\boldsymbol{X}}_1$ and $\tilde{\boldsymbol{X}}_2$, and $\boldsymbol{\Sigma}_{11}$ and $\boldsymbol{\Sigma}_{22}$ were covariances of $\tilde{\boldsymbol{X}}_1$ and $\tilde{\boldsymbol{X}}_2$ respectively.

The solution yielded p orthogonal components in $\boldsymbol{X}_1^\dagger$ and $\boldsymbol{X}_2^\dagger$, where the correlations between the corresponding components in the two feature sets decreased from the 1st to the pth components. To determine how many components to include in further analysis, we used cross validation error in predicting one feature set from the other as a measurement of goodness. Assume we used the top p_c components in the prediction, for each $i = 1, \cdots, p_c$, we learned a linear regression $\boldsymbol{X}_2^\dagger[:, i] \approx a_i \boldsymbol{X}_1^\dagger[:, i] + b_i$. Given new observations $\tilde{\boldsymbol{X}}_1^{\text{new}}$, we predicted $\tilde{\boldsymbol{X}}_2^{\text{new}}$ using the following pipeline.

$$\tilde{\boldsymbol{X}}_1^{\text{new}} \xrightarrow{\boldsymbol{X}_1^\dagger = \tilde{\boldsymbol{X}}_1 \boldsymbol{W}_1} (\boldsymbol{X}_1^\dagger)^{\text{new}}[:, 1:p_c] \xrightarrow{\text{linear mapping}} \text{prediction } \hat{\boldsymbol{X}}_2^\dagger[:, 1:p_c] \xrightarrow{\text{least square}} \text{prediction } \hat{\tilde{\boldsymbol{X}}}_2$$

The last arrow above involved solving the following least square problem

$$\arg\min_{\hat{\tilde{\boldsymbol{X}}}_2} \|\hat{\tilde{\boldsymbol{X}}}_2 \boldsymbol{W}_2[:, 1:p_c] - \hat{\boldsymbol{X}}_2^\dagger[:, 1:p_c]\|_F^2$$

where $\|\cdot\|_F$ was the Frobenius norm. Setting this objective function's gradient to zero, we got

$$\hat{\tilde{\boldsymbol{X}}}_2 = (\hat{\boldsymbol{X}}_2^\dagger[:, 1:p_c]\boldsymbol{W}_2[:, 1:p_c]')(\boldsymbol{W}_2[:, 1:p_c](\boldsymbol{W}_2[:, 1:p_c])')^{-1} = \hat{\boldsymbol{X}}_2^\dagger[:, 1:p_c]\boldsymbol{V}_{\boldsymbol{W}_2}\text{diag}(1/\boldsymbol{D}_{\boldsymbol{W}_2})\boldsymbol{U}'_{\boldsymbol{W}_2}$$

where $\boldsymbol{U}_{\boldsymbol{W}_2}$, $\boldsymbol{D}_{\boldsymbol{W}_2}$ and $\boldsymbol{V}'_{\boldsymbol{W}_2}$ were obtained from the singular value decomposition of $\boldsymbol{W}_2[:, 1:p_c]$. The prediction error was quantified as the squared Frobenius norm of the difference between the prediction and the true $\tilde{\boldsymbol{X}}_2^{\text{new}}$, divided by the squared Frobenius norm of the true value (i.e. the error was $\|\hat{\tilde{\boldsymbol{X}}}_2 - \tilde{\boldsymbol{X}}_2^{\text{new}}\|_F^2 / \|\tilde{\boldsymbol{X}}_2^{\text{new}}\|_F^2$). Similarly, a symmetric procedure could be applied to predict $\tilde{\boldsymbol{X}}_1$ from $\tilde{\boldsymbol{X}}_1$.

Noticing that selecting p_c using only the 362 stimulus images might be restricted, we used the 1086 extra images for this purpose (see Figure 5.5). We obtained $\tilde{\boldsymbol{X}}_1$ and $\tilde{\boldsymbol{X}}_2$ from the union of the extra images and stimulus images, and then we ran a leave-one-exemplar-out-in-each-category (6-fold) cross validation only on the rows corresponding to the extra images, predicting $\tilde{\boldsymbol{X}}_1$ from $\tilde{\boldsymbol{X}}_2$ and vice versa. The cross validation error was computed for different values of p_c and used for selecting the best value (see Results in Section 5.3).

After selecting the best p_c, we applied $\boldsymbol{W}_1[:, 1:p_c]$ and $\boldsymbol{W}_2[:, 1:p_c]$ that were learned from the rows corresponding to the extra images ($\tilde{\boldsymbol{X}}_1^{\text{extra}}$ and $\tilde{\boldsymbol{X}}_2^{\text{extra}}$ in Figure 5.5) on the rows corresponding to the stimulus images ($\tilde{\boldsymbol{X}}_1^{\text{stim}}$ and $\tilde{\boldsymbol{X}}_2^{\text{stim}}$), obtaining the projections $(\boldsymbol{X}_1^\dagger)^{\text{stim}}[:, 1:p_c]$ and $(\boldsymbol{X}_2^\dagger)^{\text{stim}}[:, 1:p_c]$. Let $\boldsymbol{X}_{1,2}^\dagger$ ($362 \times 2p_c$) be the union of the columns in $(\boldsymbol{X}_1^\dagger)^{\text{stim}}[:, 1:p_c]$ and $(\boldsymbol{X}_2^\dagger)^{\text{stim}}[:, 1:p_c]$; then there were p_c correlated pairs in these columns. We reduced the dimension to p_c using PCA (singular value decomposition assuming zero mean), obtaining

$$\underset{362 \times 2p_c}{\boldsymbol{X}_{1,2}^\dagger} \approx \underset{362 \times p_c}{\boldsymbol{U}_{1,2}} \; \underset{p_c \times p_c}{\boldsymbol{D}_{1,2}} \; \underset{p_c \times 2p_c}{\boldsymbol{V}'_{1,2}},$$

where we call $\boldsymbol{X}_c \overset{def}{=} \boldsymbol{U}_{1,2}$ the *common components*.

Finally we regressed $\boldsymbol{X}_c$ out from $\tilde{\boldsymbol{X}}_1^{\text{stim}}$ and $\tilde{\boldsymbol{X}}_2^{\text{stim}}$ separately, and used PCA (singular value decomposition if assuming zero mean) to extract the p_c-dimensional projections of the residual space respectively for Layer 1 and Layer 7. For example, we have the following for Layer 1

$$\underset{362 \times p_c}{\boldsymbol{U}_{r1}} \; \underset{p_c \times p_c}{\boldsymbol{D}_{r1}} \; \underset{p_c \times p_c}{\boldsymbol{V}'_{r1}} \approx (\boldsymbol{I} - \boldsymbol{X}_c(\boldsymbol{X}'_c\boldsymbol{X}_c)^{-1}\boldsymbol{X}'_c)\tilde{\boldsymbol{X}}_1^{\text{stim}}$$

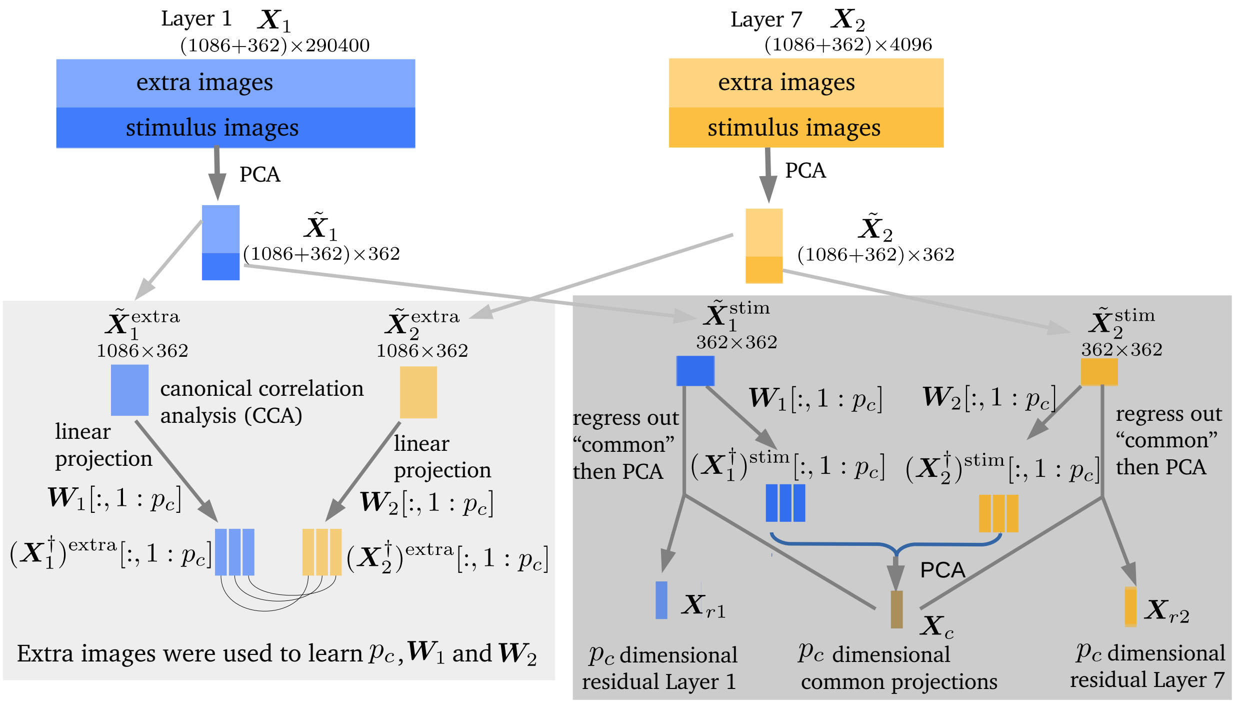

Layer 1 $\boldsymbol{X}_1$
(1086+362)×290400
extra images
stimulus images
PCA
$\tilde{\boldsymbol{X}}_1$
(1086+362)×362
Layer 7 $\boldsymbol{X}_2$
(1086+362)×4096
extra images
stimulus images
PCA
$\tilde{\boldsymbol{X}}_2$
(1086+362)×362
$\tilde{\boldsymbol{X}}_1^{\text{extra}}$
1086×362
$\tilde{\boldsymbol{X}}_2^{\text{extra}}$
1086×362
canonical correlation analysis (CCA)
linear projection
$\boldsymbol{W}_1[:, 1 : p_c]$
linear projection
$\boldsymbol{W}_2[:, 1 : p_c]$
$(\boldsymbol{X}_1^{\dagger})^{\text{extra}}[:, 1 : p_c]$
$(\boldsymbol{X}_2^{\dagger})^{\text{extra}}[:, 1 : p_c]$
Extra images were used to learn p_c, $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$
$\tilde{\boldsymbol{X}}_1^{\text{stim}}$
362×362
$\tilde{\boldsymbol{X}}_2^{\text{stim}}$
362×362
$\boldsymbol{W}_1[:, 1 : p_c]$
$\boldsymbol{W}_2[:, 1 : p_c]$
regress out “common” then PCA
$(\boldsymbol{X}_1^{\dagger})^{\text{stim}}[:, 1 : p_c]$
$(\boldsymbol{X}_2^{\dagger})^{\text{stim}}[:, 1 : p_c]$
regress out “common” then PCA
PCA
$\boldsymbol{X}_{r1}$
$\boldsymbol{X}_c$
$\boldsymbol{X}_{r2}$
p_c dimensional residual Layer 1
p_c dimensional common projections
p_c dimensional residual Layer 7

Figure 5.5: Illustration of separating the common and residual spaces of Layer 1 and Layer 7 features. The superscript $^{\text{extra}}$ and $^{\text{stim}}$ denote the rows corresponding to the extra images and stimulus images respectively.

and we call $\boldsymbol{X}_{r1} \overset{def}{=} \boldsymbol{U}_{r1}$ *residual Layer 1*. Similarly, we obtained *residual Layer 7* ($\boldsymbol{X}_{r2} \overset{def}{=} \boldsymbol{U}_{r2}$, where $\boldsymbol{U}_{r2}\boldsymbol{D}_{r2}\boldsymbol{V}_{r2}' \approx (\boldsymbol{I} - \boldsymbol{X}_c(\boldsymbol{X}_c'\boldsymbol{X}_c)^{-1}\boldsymbol{X}_c')\tilde{\boldsymbol{X}}_2^{\text{stim}}$).

Given how the *common components*, the *residual Layer 1* and the *residual Layer 7* were created, we have the following intuitive interpretation. The *common components* are the linearly correlated components between the low-level Layer 1 features and the high-level Layer 7 features. Since the features in Layer 7 are highly relevant to the object-category labels, the *common components* represent the “object-category-relevant” low-level features. For example, these features may include informative edges that formed the physical boundary of an object. In contrast, the *residual Layer 1* represents components in Layer 1 features that were roughly orthogonal to higher-level semantic information that is relevant to object categorization. The *residual Layer 7* represents linear components in Layer 7 that were roughly orthogonal to features in Layer 1; in other words, it represents unique high-level features besides the low-level features.

We used the following measurement to quantify how much variance of the original Layer 1 and Layer 7 features of the stimulus images ($\boldsymbol{X}_1^{\text{stim}}$ and $\boldsymbol{X}_2^{\text{stim}}$) were explained by the *common components* ($\boldsymbol{X}_c$), *residual Layer 1* ($\boldsymbol{X}_{r1}$) and *residual Layer 7* ($\boldsymbol{X}_{r2}$).

$$\text{variance proportion} = 1 - \frac{\|(\boldsymbol{I} - \boldsymbol{X}_*(\boldsymbol{X}_*'\boldsymbol{X}_*)^{-1}\boldsymbol{X}_*')\boldsymbol{X}_0\|_F^2}{\|\boldsymbol{X}_0\|_F^2} \tag{5.1}$$

$$\boldsymbol{X}_0 = \boldsymbol{X}_1^{\text{stim}} \text{ or } \boldsymbol{X}_2^{\text{stim}}, \quad \boldsymbol{X}_* = \boldsymbol{X}_c \text{ or } \boldsymbol{X}_{r1} \text{ or } \boldsymbol{X}_{r2}$$

assuming both $\boldsymbol{X}_0$ and $\boldsymbol{X}_*$ had zero means.

5.2.11 Confidence intervals and statistical tests

Percentile confidence intervals After obtaining the statistics representing the regression effects (e.g., R-squared, or sum of squared coefficients) for each time point, we obtain the group-averaged time series across participants, for which the confidence intervals were obtained through bootstrapping. We randomly re-sampled the statistics time series at participant level with replacement, and used the $(\alpha_0, 1 - \alpha_0/2)$ percentile confidence intervals of the bootstrapped sample [Wasserman, 2010]. The significance level α_0 here was defined as $0.05/T/n_{stat}$, where T was the number of time points in the time series, and n_{stat} was the number of statistics time series considered (Bonferroni correction).

Permutation-excursion tests When examining whether a time series of statistics is significantly different from the null hypothesis in some time windows, it is necessary to correct for multiple comparisons across different time points. Here, permutation-excursion tests [Maris and Oostenveld, 2007; Xu et al., 2011], were used to control the family-wise error rate and obtain a global p-value for the time windows. In a one-sided test that examines whether some statistics were significantly larger than the null, we first identify clusters of continuous time points where the statistics were above a threshold, and then took the sum within each of these clusters. Similarly, in each permutation, the statistics of permuted data were thresholded, and summed within each of the detected clusters. The global p-value for a cluster in the original, non-permuted case, was then defined as the proportion of permutations where the largest summed statistics among all detected clusters was greater than the summed statistics in the cluster from the non-permuted case. Specifically, to test whether the mean time series of some statistics (e.g., R-squared in regression) across participants were significantly larger than that in the baseline time window before the stimulus onset, we first took the difference between the original time series and the temporal mean of the statistics within the baseline time window, for each participant separately. Then across participants at each time point, we used the t-statistics defined in the Student's t-tests to examine if the group means of these differences were significantly above zero in any time windows. Here the each permutation was implemented by assigning a random sign to the difference time series for each subject. This test, which we refer to as *permutation-excursion t-test* hereafter, was implemented in `MNE-python`, where the number of permutations was set to 1024. The threshold of the t-statistics was equivalent to an uncorrected p-value ≤ 0.05.

Other corrections of multiple comparisons Besides the permutation-excursion tests above, in some cases, where the possible dependence structure was less easy to describe than that in adjacent time points, we used other methods to correct for multiple comparisons, including the Bonferroni criterion and the Benjamini-Hochberg procedure [Benjamini and Hochberg, 1995] to control for the false discovery rate.

Combining p-values of tests across individual participants In some cases (e.g., the connectivity analyses below), we ran statistical tests and obtained the p-value for each individual participant. To combine these individual p-values and obtain a group-level p-value, we used the binomial test [Darlington and Hayes, 2000]. Suppose we control the p-values for each participant at a level of 0.05. Under the null hypothesis (i.e., in all of the participants, the null hypotheses of the individual tests are true), the p-values of each individual participant should be uniformly distributed, and the probability of observing a p-value < 0.05 is 0.05. So we counted the number of participants who showed a p-value smaller than 0.05, denoted as n_{pos}. Then we computed the group-level p-value of the binomial test as the tail probability of Binomial$(n_{parti}, n_{pos}, 0.05)$, where n_{parti} was the total number of participants. Note that a significantly small group-level p-value indicates that at least for one participant, the alternative hypothesis was likely true.

5.2.12 Source-space connectivity analysis

In the connectivity analysis, we focused on six ROIs, including the pericalcarine areas that mainly covered the early visual cortex (EVC), an object-selective ROI (the lateral occipital complex, LOC), three scene-selective ROIs (PPA, RSC and TOS), and the medial orbitofrontal cortex (mOFC), which was roughly in or near the location of the mPFC in the "contextual network" introduced by Bar and Aminoff [2003]. Each ROI was the union of the corresponding regions in the left and right hemispheres (see Figure 5.14). We used the time-varying autoregressive (AR) model in Section 4.3 to describe the connectivity (i.e., the leading or lagged dependence across ROIs in the neural responses to the scene images). Besides our one-step state-space model, we also used a two-step approach, where the dSPM estimation of source activity was obtained in the first step, and then the time-varying AR coefficients were fitted on the mean activity across source points within each ROI. The penalization parameter in the dSPM was set to the default value for a single trial in the MNE-python software. The regularization parameters for the AR model was set to $\lambda_0 = 0$ and $\lambda_1 = 1.0$ for both methods (see Section 4.3 for details). The models were fitted for each participant individually, and standard deviations of the AR coefficients were obtained by bootstrapping the observations for 362 images (sampling the 362 observations with replacement). The sampling indices were the same across participants. For the EM algorithm in our one-step model, we mainly used the AR coefficients by the two-step method for initialization.

5.3 Results

As mentioned in the Introduction (Section 5.1), we aim to analyze the spatio-temporal neural activity to address two questions: (1) what kind of information is extracted at different temporal stages and different areas and (2) how neural activities in different areas interact with each other dynamically.

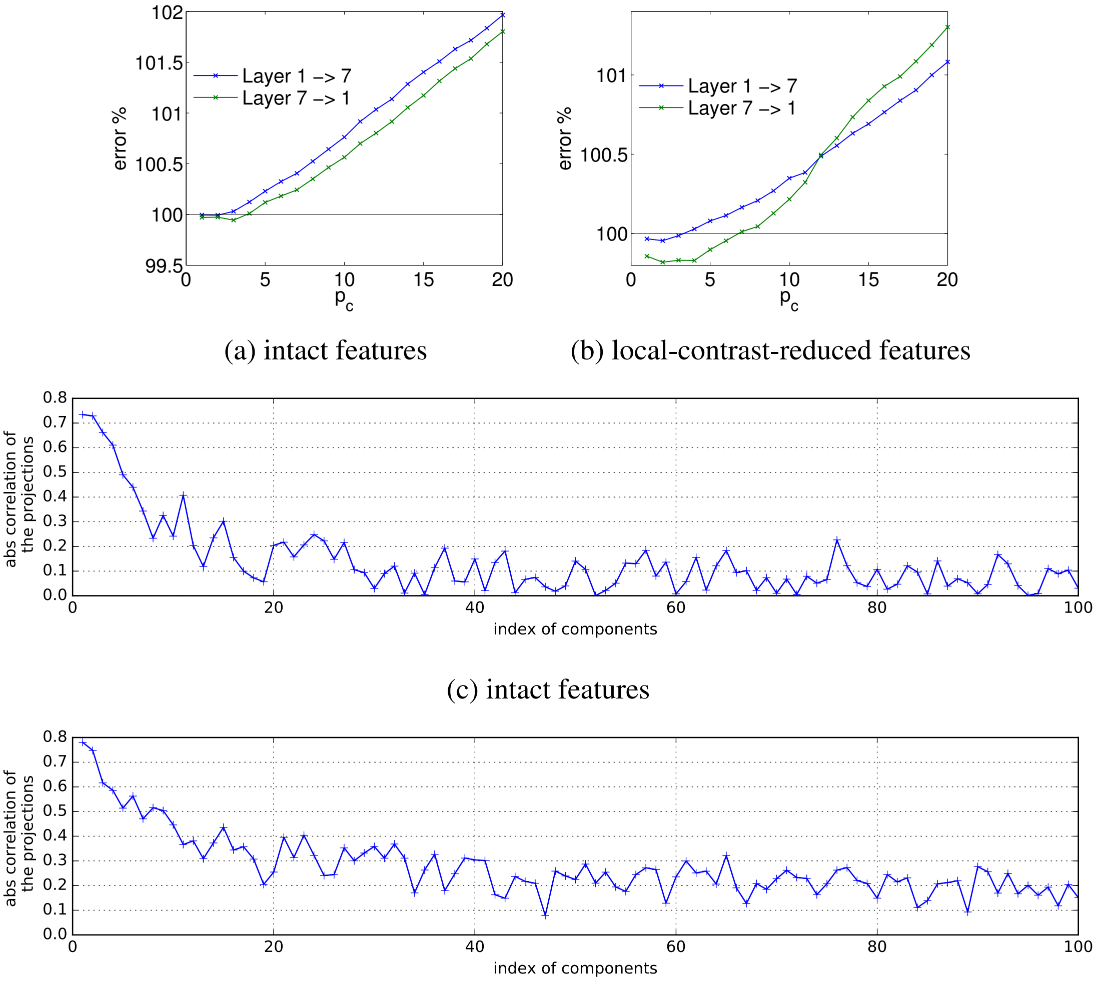
First, we present results related to Question (1). We exploited features in different layers of

AlexNet, which represented low-level and high-level information in the visual stimuli, and we regressed the neural activities against these features. Before showing the regression results, we first present analyses of the features from the low-level Layer 1 and the high-level Layer 7 per se. In these analysis, we extracted the linearly dependent components between the two, as well as residual components in each layer, yielding three nearly orthogonal groups of regressors. These groups can be intuitively interpreted as the low-level features that are relevant to high-level features, the low-level features that are roughly orthogonal to high-level features and the high-level features that are roughly orthogonal to low-level features. Using these groups as regressors, we computed time series of statistics that described the linear dependence between neural activities and these groups, which we call "correlation profiles" hereafter. We first present the temporal correlation profile in the sensor space and then present the spatio-temporal correlation profiles in the source space from both a traditional two-step method and our novel one-step STFT-R regression method. The results from both the MEG and EEG analysis were similar to each other, yet the source-space results in EEG appeared more noisy than in MEG, possibly because there were fewer sensors in EEG and the locations of EEG sensors in relation to the scalp were less reliable (see Data acquisition in the Methods section). In the following text, we mainly focus on the MEG results.

Secondly, we address Question (2) by presenting analyses of time-lagged functional connectivity across six regions of interest (ROIs), which are important in scene processing. These regions included the early visual cortex, the object/scene-selective regions and the medial orbital frontal cortex, which was roughly in or near the location of the mPFC in the "contextual network" introduced by Bar and Aminoff [2003]. We quantified the connectivity using a time-varying autoregressive model of order 1 with both a two-step approach and our novel one-step state-space model in Chapter 4; then we ran statistical analyses of the autoregressive coefficient at each time point for each pair of ROIs. Because the source localization in our MEG data was better than that in our EEG data, we only ran the analysis on the MEG data.

5.3.1 Canonical correlation analysis of features in Layer 1 and Layer 7

In this subsection, we present our analyses of the features from the low-level Layer 1 and the high-level Layer 7. To characterize the common linear space between the two feature sets, we used canonical correlation analysis to find linear projections of each feature set that maximized the Pearson correlations of the projected features of the two sets. To determine the number of CCA components (p_c) needed, we computed the cross-validated errors when we used Layer 1 to predict Layer 7 (and used Layer 7 to predict Layer 1) though the CCA components among the extra images. Figure 5.6a shows the prediction errors for both prediction directions. When $p_c > 3$, the prediction errors were greater than 100%. In these cases, the Frobenius norm of the difference between the predicted values and the true values was even larger than the Frobenius norm of the true values; in other words, Layer 1 and Layer 7 did a poor job in predicting each other with $p_c > 3$ CCA components.



(a) intact features

(b) local-contrast-reduced features

(c) intact features

(d) local-contrast-reduced features

Figure 5.6: Canonical correlation analysis of features in Layer 1 and Layer 7. (a) and (b): cross-validated errors of predicting the Layer 7 features using the Layer 1 features (blue) and vice versa (green), with different p_cs, computed among the extra images. (a) was obtained from the intact features; (b) was obtained from the *local-contrast-reduced features*. (c) and (d): absolute values of correlations between each pair of projections of the features of the stimulus images. The projections were obtained by applying $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$, which were fitted from $\tilde{\boldsymbol{X}}_1^{\text{extra}}$ and $\tilde{\boldsymbol{X}}_1^{\text{extra}}$, on $\tilde{\boldsymbol{X}}_1^{\text{stim}}$ and $\tilde{\boldsymbol{X}}_1^{\text{stim}}$. (c) was obtained from the intact features; (d) was obtained from the *local-contrast-reduced features*.

In additional to the prediction error, we also examined the Pearson correlation of the CCA components for the stimulus images, which did not overlap with the extra images used in learning the CCA. These correlations can be seen as the "validation correlation" in this data-split case. In other words, after learning the CCA projections ($\boldsymbol{W}_1$ and $\boldsymbol{W}_2$) from the features in Layer 1 and 7 of the extra images, we applied $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$ on the features of the stimulus images, and computed the correlation of the projections for each corresponding CCA component. Noticing that both positive and negative correlations were meaningful, we took the absolute values of the correlations.

Figure 5.6c shows the absolute values of the correlations; the first 4 components had correlations higher than 0.6, and the first 6 components had correlations higher than 0.4.

Considering the results in Figure 5.6a and 5.6c, we chose $p_c = 3$ and $p_c = 6$ in computing the following three groups of features for the stimulus images: the *common components* as the first p_c principal components of the union of the p_c CCA projections of the two layers, and the *residual Layer 1* and *residual Layer 7* as the first p_c principal components that were orthogonal to the *common components* in the two layers (see the Methods part in Section 5.2.10). We further examined how much variance each group could explain in the original features from Layer 1 and Layer 7. The results are listed in Table 5.1 (first two blocks and first two columns). The proportions of explained variance were all relatively small but reasonable considering the large dimensions of the original features ($\geq 10^3$). Not surprisingly, the *residual Layer 1* explained more variance of Layer 1 than of Layer 7 and the *residual Layer 7* explained more variance of Layer 7 than of Layer 1.

Table 5.1: Proportion of variance (%) of different features explained by the *common components*, *residual Layer 1* and *residual Layer 7*.

$p_c = 3$	Layer 1 (intact)	Layer 7 (intact)	Local contrast	Layer 1 (local-contrast-reduced features)	Layer 7 (local-contrast-reduced features)
residual Layer 1	5.8	2.3	8.7	2.8	1.5
residual Layer 7	1.4	10.1	2.0	1.2	8.9
common components	3.9	7.1	10.7	2.6	4.3
$p_c = 6$					
residual Layer 1	8.1	2.7	9.7	4.4	2.0
residual Layer 7	1.9	15.8	1.5	1.9	14.1
common components	6.6	10.8	18.6	4.5	6.8
$p_c = 6$					
residual Layer 1(local-contrast-reduced features)	5.4	2.5	1.4	6.5	2.7
residual Layer 7(local-contrast-reduced features)	1.7	13.3	1.0	1.9	15.6
common components(local-contrast-reduced features)	4.6	8.1	1.3	5.6	9.5

It is worth noting that neurons in the lower-level visual cortex can have strong responses to high-contrast stimuli. In order to describe visual features due to contrast, we extracted the *local contrast* features, which represented the gray-value contrasts within the local receptive fields of Layer 1

[8]. Although the local contrast features defined in this manner are mainly low-level features, they may also contain some high-level information such as a (blurred) contour of an object. Therefore, we also computed how much variance of the local contrast features could be explained by the three groups—the *common components*, *residual Layer 1* and *residual Layer 7* (shown in Table 5.1). Interestingly, the *common components* explained more variance of the *local contrast* than the *residual Layer 1* (when $p_c = 6$, the proportion by the *common components* was almost twice as that by *residual Layer 1*). This could be because image patches with high local contrast usually contain important information about boundaries of objects and scenes in naturalistic photos and thus these high-contrast patches contributed a lot to the *common components*, which were correlated with both low-level and high-level features. In our regression analysis of the MEG and EEG data, we observed that the local contrast had a very strong effect in explaining the MEG/EEG data (see results in Section 5.3.2); this effect may confound our comparisons between the neural correlations with the *common components*, the *residual Layer 1* and the *residual Layer 7*. To reduce the confounding effect, we regressed out the first 160 principal components of the *local contrast features* (explaining 90% of the variance in the *local contrast features*) from the intact features in each layer of `AlexNet`, and took the residuals as new features of interest [9]. We term these new features *local-contrast-reduced features* hereafter.

Using the *local-contrast-reduced features*, we again plotted the cross-validated prediction errors (Figure 5.6b). The prediction errors increased with p_c when $p_c > 10$. The smallest prediction error occurred at $p_c = 2$ when we used Layer 1 to predict Layer 7, and at $p_c = 4$ when we used Layer 7 to predict Layer 1. Nevertheless, the error was smaller than 100% in at least one prediction direction when $p_c \leq 6$. Consistent with these results, the absolute values of validation correlations of the CCA components (in Figure 5.6d) were greater than 0.5 when $p_c \leq 6$. In terms of choosing p_c, which determined the dimensions of the *common components*, the *residual Layer 1* and the *residual Layer 7*, we needed to choose $p_c \leq 6$ according to the cross-validation results; however, using a very small number made little sense because visual information in naturalistic images is usually rich and possibly requires a higher-dimensional description. As a result of the trade-off, we used $p_c = 6$ in our further analyses. The variance explained by the newly computed *common components*, *residual Layer 1* and *residual Layer 7* based on the *local-contrast-reduced features* is also listed in Table 5.1 (last block, last three columns). As designed, all of the three new groups explained little variance in the local contrast features; the *residual Layer 1* explained little variance of Layer 7 and the *residual Layer 7* explained little variance of Layer 1. The *residual Layer 1* explained slightly higher variance in Layer 1 than the *common components* did, but the proportions were comparable (6.5% and 5.6%); the *residual Layer 7* explained more variance in Layer 7 than the *common components* did (15.6% and 9.5%).

[8] The nuisance covariates related to the widths and heights of the images were regressed out from the local contrast features here.

[9] The PCA and regression were run using the data across both the stimulus images and the extra images.

5.3.2 Sensor-space regression

In this subsection, to examine whether the features extracted from different layers in `AlexNet` could explain the MEG and EEG data in the sensor space, we ran an ordinary least square regression analysis at each sensor and each time point for each participant. In this way, we were able to compare temporal profiles describing the linear dependence (which is equivalent to correlations) between different features and the neural activities recorded in MEG and EEG. In the regression, neural responses to all 362 images were used. Because the `AlexNet` features were high-dimensional compared with the number of observations, we needed to avoid overfitting with either dimension reduction or other regularization. Here, our main goal was to test whether a significant amount of variance was explained by each layer; for computational simplicity, we used principal component analysis (PCA) to reduce the dimensionality and included the first 10 principal components as regressors [10].

We used the first 10 principal components of both the intact features and the *local-contrast-reduced features* from each layer as the regressors [11], and we quantified the correlation between neural recordings and the regressors in each layer as the proportion of variance explained (i.e., R-squared). For visualizing the overall effects, the R-squared were averaged across all sensors at each time point for each participant, in the MEG and EEG sessions respectively. Figure 5.7 shows the results, where each curve represents the R-squared for each layer, averaged across all sensors and further averaged across participants. The transparent bands show the confidence intervals obtained by bootstrapping the observed R-squared time series at the participant level. The *permutation-excursion t-tests* were used to test whether the averaged R-squared across sensors were greater than the temporal average of that in the baseline time window (-140 to -40 ms in MEG, -120 to -20 ms in EEG), during which the MEG/EEG recordings should be independent of the stimulus images and thus the regressors. The significant time windows were identified where the p-values of the *permutation-excursion t-tests* were smaller than $0.05/8$. These significant windows are marked by the colored segments under the curves in Figure 5.7. Note that these p-values were already corrected for the multiple comparisons at different time points; the denominator 8 was used as a correction for the 8 tests corresponding to the 8 layers according to the Bonferroni criterion.

In Figure 5.7, the left column shows the results from using the intact features ((a) in MEG and (c) in EEG), and the right column shows the results from using the *local-contrast-reduced features* ((b) in MEG and (d) in EEG). In both columns, we identified significant time windows (from 80 ms to at least 400 ms) for Layer 1 through Layer 7. In these windows, the variance explained by the features was significantly greater than that in the baseline time windows. These results

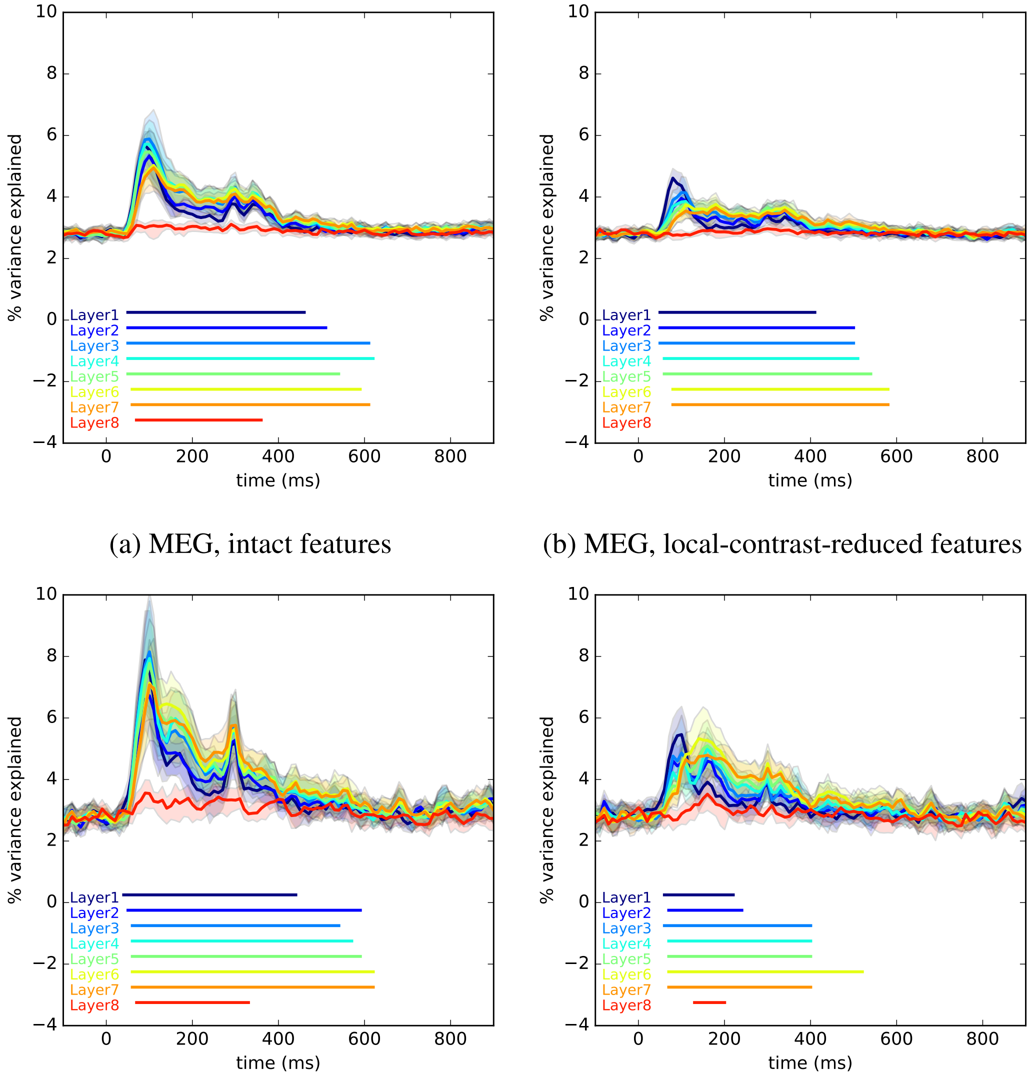
[10] The choice of 10 was arbitrary; we expect similar results as long as the number of components is not too small (e.g., 2 or 3) nor too large such that the regression models overfit.

[11] In the case of intact features, the PCA was run only on the features corresponding to the 362 stimulus images. In the case of *local-contrast-reduced features*, the PCA was run on the features across both the stimulus images and the extra images, but the first 10 principal components corresponding only to the 362 stimulus images were used as regressors.

indicate that the neural responses recorded by MEG/EEG in these time windows were correlated with the `AlexNet` features. In addition, all subplots appeared to show a pattern of early-to-late, lower-level-to-higher-level shift, where lower-level layers explained a larger proportion of variance before 150 ms, and higher-level layers, especially Layer 6 and 7, explained a larger proportion of variance in later time windows (150 to at least 400 ms). This pattern is consistent with the feedforward direction of information flow in the hypothesis of hierarchical organization. Layer 8 generally explained lower variance than the other seven layers in all of the subplots; no significant time window was detected in Figure 5.7(b). This may be because the object categories represented in Layer 8 did not align with the categories that human brains naturally use, or because the neural activity coding object labels is generally weaker than the neural activity coding visual features. Interestingly, although the proportion of variance explained by the principal components from the intact features appeared higher than that by the principal components from the *local-contrast-reduced features*, the early-to-late, lower-level-to-higher-level temporal patterns were much more clear in the latter case. Such results suggest that using the *local-contrast-reduced features* may help us to better separate the neural correlations with low-level and high-level features. Therefore, from now on, we only present results with the *local-contrast-reduced features*; without specific notes, "features" will refer to the *local-contrast-reduced features*.

The results above only showed us the temporal patterns. Next, we visualize spatially which sensors demonstrated strong correlations with the `AlexNet` features, by showing the regression results on a topological map of the sensor layout. For each sensor, and for each time point from 50 ms to 700 ms with a 50 ms step-size, we averaged the R-squared first over a 70-ms window centered at the time point and then across all participants. Note that the same sensor could map to different locations for different participants, due to individual variations in head sizes and the head locations in the MEG helmet or the placement of the EEG cap on the scalp. Hence the results here is mainly for visualizing the overall pattern. We defer more rigorous statistical comparisons of the spatio-temporal correlation profiles to Section 5.3.3 on source-space analysis.

Figure 5.8a includes the topological heat maps of the averaged R-squared (proportion of variance explained) for each MEG sensor by the first 10 principal components of the features in Layer 1, 3, 5 and 7. Similarly, Figure 5.8b shows the heat maps from EEG data. In both modalities, the strongest regression effects were in the posterior sensors, which were close to the visual cortex. From Layer 1 to Layer 7, we can see a shift of the correlation effects from early windows to late windows, especially within 50 to 250 ms. For Layer 1 (and also Layer 3), it is interesting that after the first transient peak at 100 ms, there appears to be a second peak at 300 to 350 ms, which is more apparent in the MEG plots. This second peak was likely to be caused by the neural responses to the disappearance of the stimuli at 200 ms—there was roughly a 80 to 100 ms delay from the stimulus onset to the first peak, so there can be a similar delay in the responses to the disappearance of images. Disappearance of an image causes changes in the visual input on the screen (e.g., an image patch of green grass changes into a gray patch in the fixation screen), and neurons, especially those in the low-level visual cortex, can respond to these changes of visual input. In this sense, these neural responses are specific to the images and thus they can show correlations with the



(a) MEG, intact features

(b) MEG, local-contrast-reduced features

(c) EEG, intact features

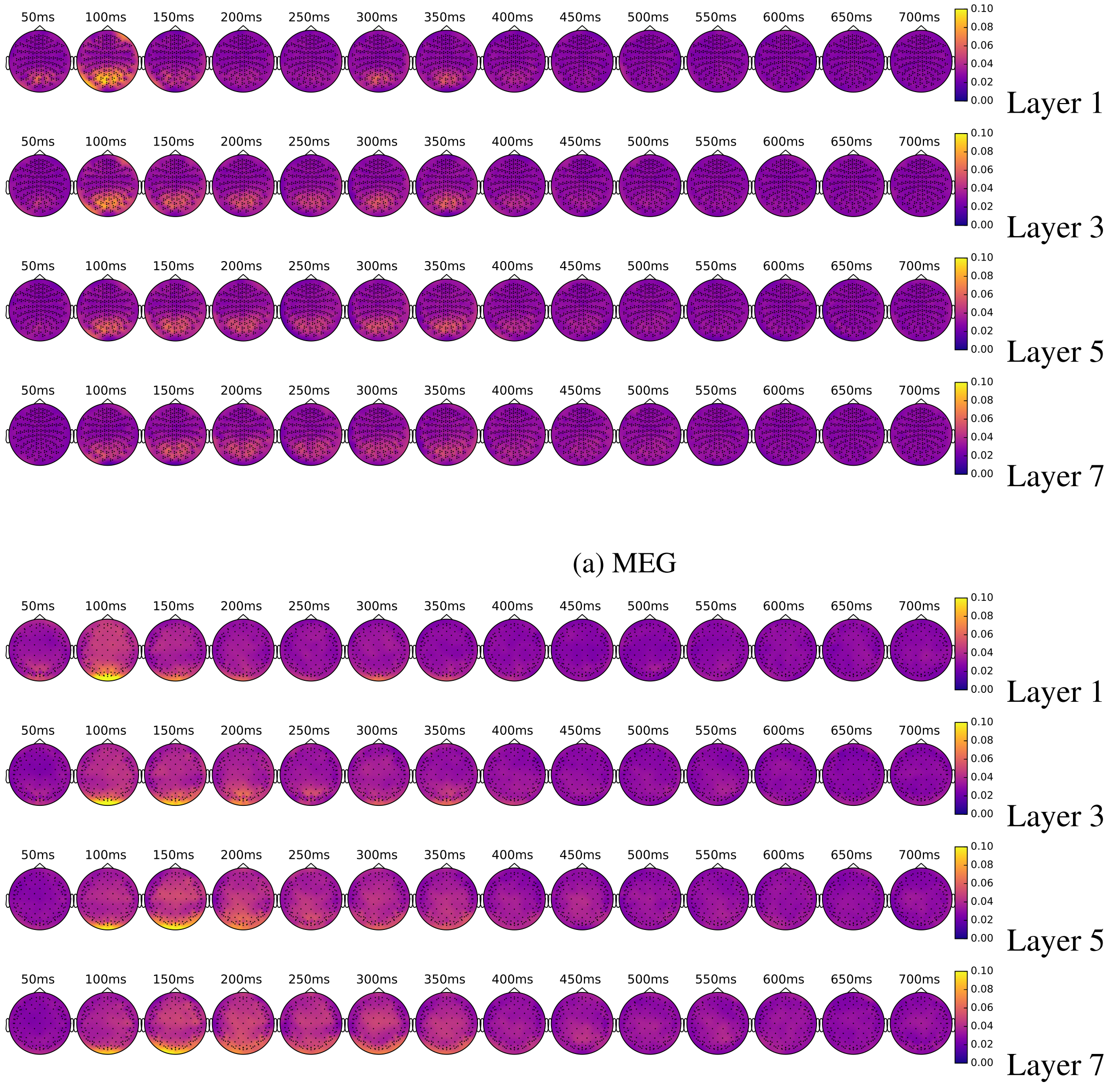
(d) EEG, local-contrast-reduced features

Figure 5.7: Proportion of variance explained by the first 10 principle components of the `AlexNet` features in each layer, averaged across sensors.

features of the images [12].

In addition, we also obtained the topological heat maps for the three groups of decomposed features—the *common components*, the *residual Layer 1* and the *residual Layer 7*, which were obtained from the canonical correlation analysis of the *local-contrast-reduced features* ($p_c = 6$). We note that the *common components* represent the low-level features that are correlated with the high-level features in Layer 7. Since the features in Layer 7 are highly relevant to the outputs of

[12] We collected some preliminary data from three extra participants in EEG, where the stimuli were presented for 500 ms instead of 200 ms. In these preliminary data (not shown here), the second peak appeared to shift to around 600 ms, which validated our speculation to some degree.



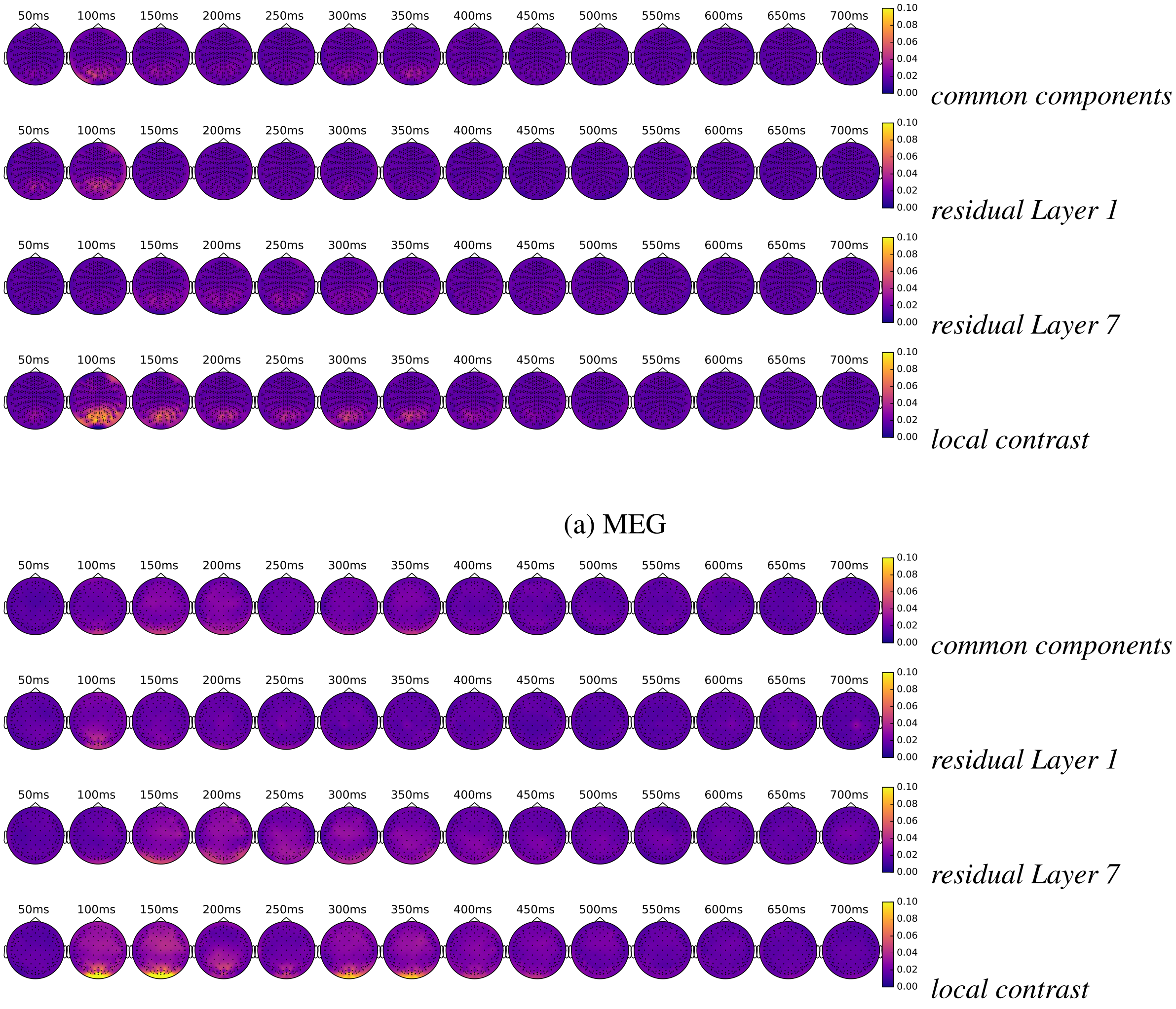
(a) MEG

(b) EEG

Figure 5.8: Topology maps of the proportion of variance explained by the features from Layer 1, 3, 5 and 7.

AlexNet (i.e., the object-category labels), we term the *common components* the *object-category-relevant* low-level features. In contrast, the *residual Layer 1* represents low-level features that are roughly orthogonal to the features in Layer 7 (i.e., roughly orthogonal to *object-category-relevant* features). The *residual Layer 7* represents high-level features that are roughly orthogonal to low-level features; in other words, the components in the *residual Layer 7* are object-category-relevant, and they provide unique high-level information besides the low-level features. For comparison, we also obtained the topological maps for the first 6 principal components of the *local contrast features*. Figure 5.9a shows the corresponding MEG plots and Figure 5.9b shows the corresponding EEG plots. In both the MEG and EEG results, we saw similar patterns. The *residual Layer 1* had an early peak near 100 ms, and the *residual Layer 7* had correlation effects that were later, lasting from 150 ms to 400 ms. The *common components* appeared to have two peaks, one centered at 100 ms, and a weaker one near 300 to 350 ms. Again, the latter one was possibly due to the neural responses to the disappearance of stimuli. If we compare the two low-level groups (*common components* and *residual Layer 1*) with the high-level group (*residual Layer 7*), we can again observe the early-to-late, lower-level-to-higher-level shift, which is consistent with feedforward information flow. The local contrast features had large correlation effects spanning from 100 to 400 ms, with one early peak at 100 to 150 ms and a later peak near 300 to 350 ms, which was possibly due to neural responses to the disappearance of stimuli.

As a short summary of the regression results in the sensor space, we observed an early-to-late, lower-level-to-higher-level shift of the temporal correlation patterns, which supports feedfoward information flow. When we decomposed the *local-contrast-reduced features* from Layer 1 and 7 into three roughly orthogonal groups of features—the *common components*, the *residual Layer 1* and *residual Layer 7*, we observed an apparent temporal separation of the low-level and high-level features (e.g., when comparing the *residual Layer 1* and the *residual Layer 7* or comparing the *common components* and the *residual Layer 7*). In addition, the *local contrast features* explained a larger proportion of variance in the neural data than the three groups derived from the *local-contrast-reduced features*; this result suggests that if the local contrast features were not partialled out from the AlexNet features, most of the correlation effects we observe could be due to the local contrast features.



(a) MEG

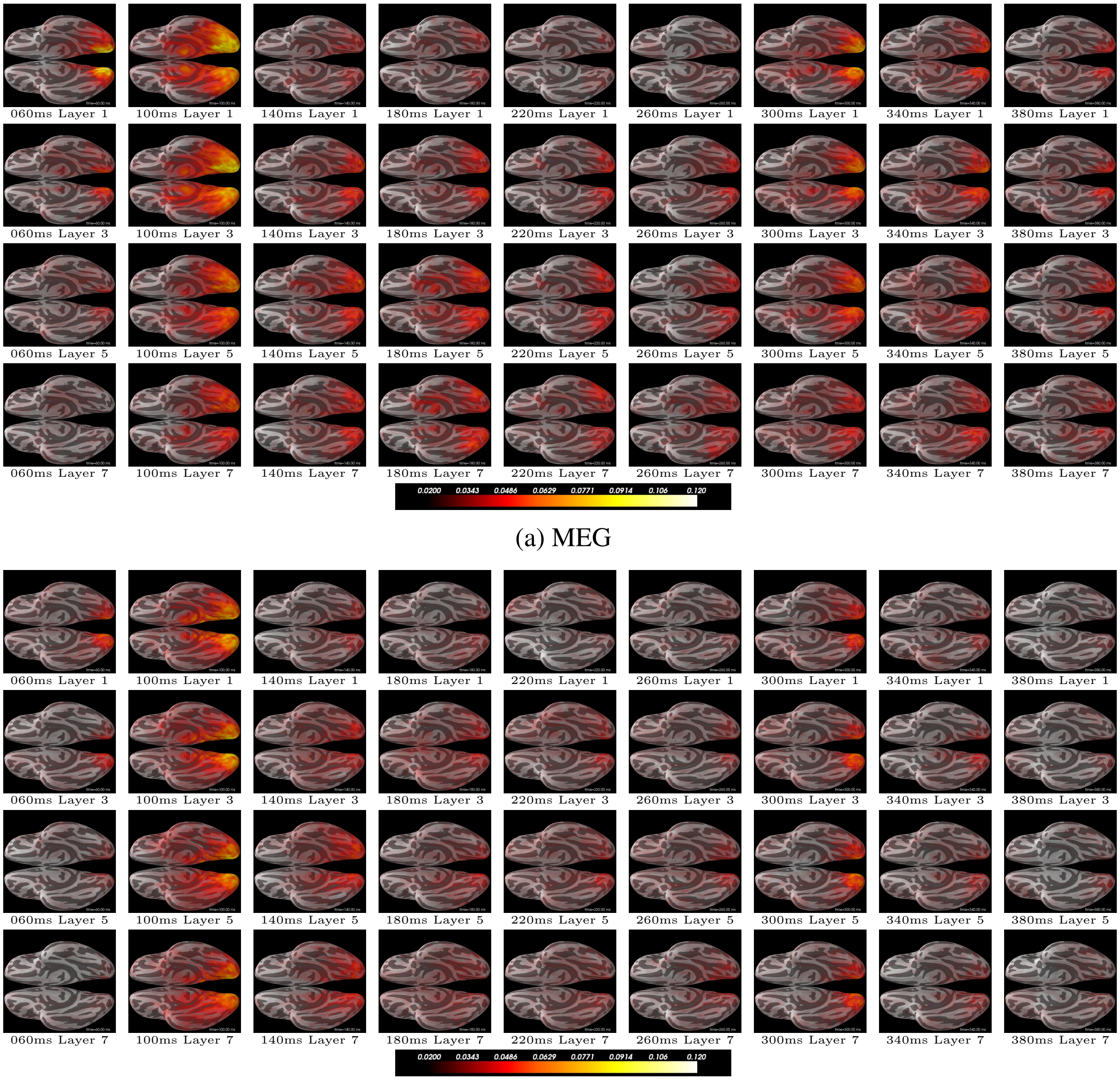
(b) EEG

Figure 5.9: Topology maps of the proportion of variance explained by the *common components*, the *residual Layer 1*, *residual Layer 7* and the *local contrast features*.

5.3.3 Source-space regression

The regression results in the sensor space above only present temporal correlation profiles between the neural activities and the features of the stimuli. In this subsection, we move into the source space to examine the spatio-temporal correlation profiles. We applied two different methods in the source-space regression analyses, the conventional two-step approach (source localization using dSPM and then regression), and our one-step short-time Fourier transform regression (STFT-R) model introduced in Chapter 3. Both methods were applied to the data from individual participants, yielding the regression coefficients for each dimension of the features (regressors) at each time point and each source point.

Based on these results, we can obtain a whole-brain map of the correlation effects or compare the correlation effects in regions along the hierarchy of the visual cortex. In either case, it is reasonable to aggregate the results across participants, for visualization or statistical tests on the group level. In the first case—visualizing the whole-brain correlation effects across all participants on the cortical surfaces, we mapped the source space of each participant onto a default template in `Freesurfer`. It is worth noting that the spatio-temporal map of the regression coefficients given by the STFT-R was spatially sparse; morphing and mapping such sparse patterns onto the template may result in spurious patterns. Therefore, for visualizing the whole-brain maps, we only used the two-step approach—we obtained the dSPM source solutions for each participant, morphed the solutions into the template source space, computed the regression coefficients for each source point, and finally averaged the R-squared statistics across the participants. In this way, we obtained the whole-brain maps of the averaged R-squared statistics for the first 10 principal components for each layer in `AlexNet`. Figure 5.10 shows the visualization at different time points from a ventral view of the cortical surfaces, for Layer 1, 3, 5 and 7. It is worth noting that adjacent source points usually contribute to the sensor recordings in a similar way (i.e., they correspond to correlated columns in the forward matrix); as a result, L_2-norm regularized methods like dSPM generally have a spatial blurring effect, where an underlying single large current dipole can be reconstructed as distributed current dipoles covering a large cortical area. Hence when we see strong effects (i.e., large averaged R-squared) in large areas, it could be due to either local large effects or genuinely distributed effects. Having that in mind, now we look at the results in Figure 5.10 ((a) from MEG and (b) from EEG). In the MEG plots, at 60 ms, the correlation effects localized in the posterior end near the early visual cortex were strongest for Layer 1, and as we move to Layer 3, 5 and 7, the regression effects gradually decreased; at 100 ms, the regression effects had large magnitudes and spread over large areas of the ventral visual cortex; at 140 to 180 ms, the correlation effects were stronger for Layer 3, 5 and 7 than for Layer 1, and the effects spread to more anterior regions for Layer 5 and 7. This pattern is consistent with the feedforward information flow along the hierarchy from posterior to anterior parts. In addition, near 300 to 350 ms, there was a second peak of the correlation effects for most layers in the posterior parts near the low level visual cortex in the hierarchy. As discussed in Section 5.3.2 on the sensor-space regression, this later peak had a delay of roughly 100 ms from the disappearance of the stimuli at 200 ms, which was similar to the delay of the first peak from the stimulus onset; in this sense, the later peak could be due to the correlation between the neural responses to the disappearance of the images and the features, as the disappearance of the images caused image-specific changes in the visual input and neurons in the visual cortex were likely to respond to such changes. The EEG results were similar to those in MEG, yet the pattern differences between layers appear less obvious.



(a) MEG

(b) EEG

Figure 5.10: Proportion of variance explained by different layers, computed from the dSPM source solutions that were morphed into a common template and averaged across participants.

In addition to using the features from each layer in `AleNet`, we also ran source-space regression analyses against the three groups of features, the *common components*, the *residual Layer 1* and the *residual Layer 7*. We aim to run rigorous statistical tests to compare the correlation profiles among these three groups, which represented the object-category-relevant low-level features, the low-level features roughly orthogonal to high-level features, and the high-level features roughly orthogonal to low-level features, respectively. For this purpose, instead of doing a whole-brain analysis, where we needed to control for multiple comparisons at thousands of source points, we focused on several representative ROIs along the hierarchy, including the pericalcarine areas that covered the early visual cortex (EVC), the three partitions of the non-EVC ventral visual cortex along the posterior to anterior axis (Vent 1, 2 and 3) and the object/scene-selective regions (LOC, PPA, RSC and TOS). The corresponding regions in the left and right hemispheres were merged. Note that the object/scene-selective regions might overlap with Vent 1 and 2. By aggregating the correlation effects within each of these ROIs, we have fewer multiple comparisons and therefore the statistical power can be boosted. Moreover, since these regions were defined individually for each participant, we did not have to morph the individual source spaces onto the template. Instead, we were able to apply both the two-step method and the STFT-R model for each individual participant, to obtain summarizing statistics that described the correlation effects for each ROI in each participant, and finally to test these statistics on the group level.

In the two-step approach, after obtaining the dSPM source estimates, we ran a linear regression for each source point at each time point, using each of the three groups (the *common components*, the *residual Layer 1* and the *residual Layer 7*) as the regressors to obtain the R-squared statistics. These time series of R-squared could have different magnitudes in different source points and for different participants, so we further normalized these time series. For each source point, we divided the R-squared values at each time point by the sum across all time points and all three groups. Then we averaged these normalized values across the source points within each ROI. Note that these normalized values reflect the relative correlation effects between the localized neural activities and the three groups of regressors in the ROI, and thus they are the spatio-temporal profiles of the correlation effects we will focus on hereafter. In the first column of Figure 5.11a, we show the average of the normalized values across the participants for each group of regressors in two representative ROIs, obtained from the MEG session. The transparent bands show 95% confidence intervals, bootstrapped at the participant level and corrected for multiple comparisons in all the time points and for the three groups of regressors using the Bonferroni criterion.

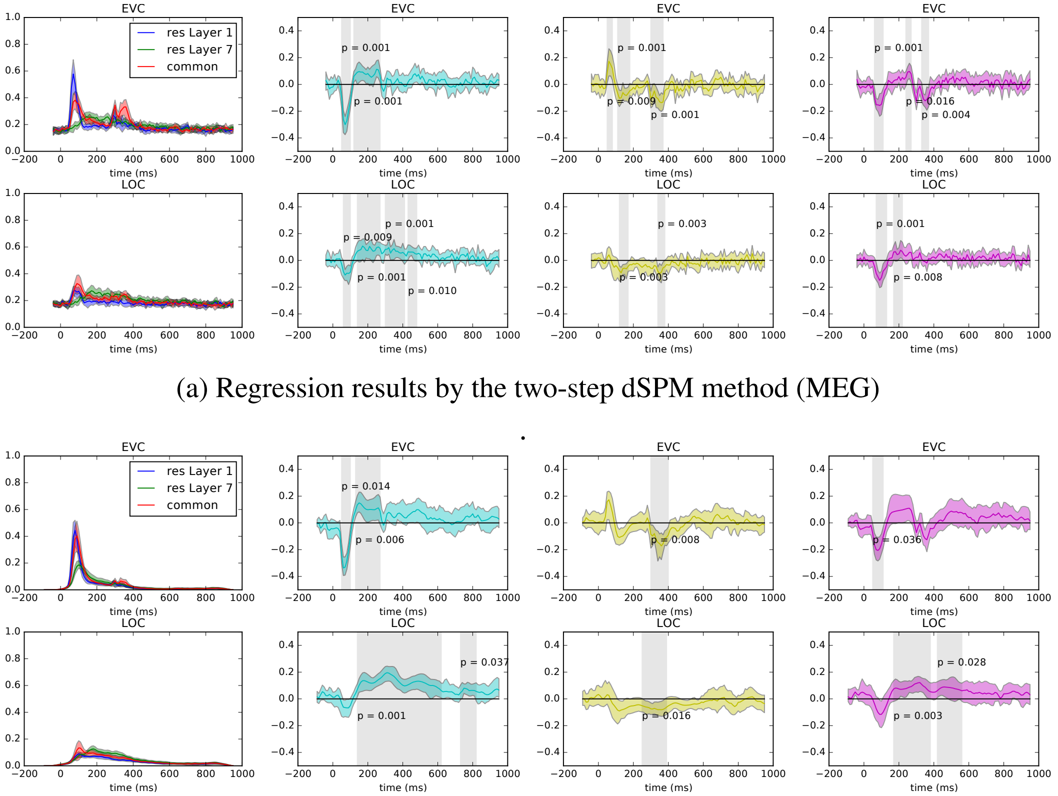
We describe the profiles of the correlation effects in the first column qualitatively, and then we present rigorous pairwise comparisons between the three groups (the *residual Layer 1*, the *residual Layer 7* and the *common components*) in the last three columns. Besides contrasting the patterns between the three groups within a region, we will also focus on the temporal changes of the correlation effects. In addition, we qualitatively compare the profiles in the EVC at the low level of the hierarchy and the profiles in the regions at higher levels than the EVC (e.g., LOC).

In the EVC (the first row), we can see that the *residual Layer 1* and the *common components* (the blue and red curves)—corresponding to the low-level features roughly orthogonal to high-level features and the low-level features that were object-category-relevant—had early transient correlation effects within 60 to 120 ms, peaking near 100 ms. The correlation effects of the *residual Layer 7* (the green curve)—corresponding to the unique high-level features that were roughly orthogonal to low-level features—increased later, starting roughly at 100 ms, peaking near 140 ms, and lasting until at least 400 ms. In addition, the *residual Layer 1* and the *common components* both appeared to have a second peak near 300 ms [13], which was likely due to the neural responses to the disappearance of the stimuli at 200 ms. However, interestingly, this second peak appeared smaller and more transient for the *residual Layer 1*; the correlation effect lasted longer (until at least 380 ms) for the *common components* (the red curve), which represented the object-category-relevant low-level features, than for the *residual Layer 1* (the blue curve), which represented the low-level features roughly orthogonal to high-level features. This result indicates that the early visual cortex, in this time window that was likely to include neural responses to the disappearance of the stimuli, differentiated between the two groups of low-level features, showing a higher and more temporally extended correlation effect with the object-category-relevant low-level features. In contrast, in the LOC (second row), an object-selective region at a higher-level than the EVC in the hierarchy, the early transient peaks of the regression effects for the low-level features (the *residual Layer 1* and the *common components*) were much smaller, whereas the magnitude of the later peak for the high-level features (*residual Layer 7*) was similar to that in the EVC. Moreover, we did not observe second peaks for the *residual Layer 1* and the *common components* that were as prominent as those in the EVC, which suggests that the neural responses corresponding to the disappearance of the images were much smaller in the LOC.

As mentioned above, we did pairwise comparisons between the correlation effects of the *residual Layer 1*, the *common components* and the *residual Layer 7*. In these comparisons, for each source point, each time point and each participant, we computed the ratio of the R-squared value for each regressor group to the sum of the R-squared values across the three groups. In this way, each time point had comparable statistics that described the relative strength of the correlation effects for each of the three regressor groups. Again, for each participant, we averaged these ratios across the source points within each ROI. Then we took pairwise differences (*residual Layer 1-common components*, *residual Layer 7-residual Layer 1* and *residual Layer 7-common components*), and examined whether the averaged differences across participants were significantly different from zero in each ROI. The last three columns in Figure 5.11a show the mean differences across the participants in the two ROIs. The transparent bands show the bootstrapped 95% confidence intervals, corrected for the three comparisons at all the time points using the Bonferroni criterion. The *permutation-excursion t-tests* were used to identify time windows where the two-sided p-values were smaller than 0.05. Note that in this case, we only corrected for multiple comparisons in all the time points. To further correct for the comparisons for different pairs and the multiple ROIs [14], we controlled the false discovery rate (FDR) at 0.05 using the Benjamini-Hochberg procedure.

[13] In the electronic version of this thesis, the reader can enlarge Figure 5.11a for a clear view of the transient peaks.

[14] Besides the EVC and LOC, we also analyzed the results in PPA, RSC and TOS, as well as Vent 1, 2 and 3. See



(a) Regression results by the two-step dSPM method (MEG)

(b) Regression results by STFT-R (MEG)

Figure 5.11: Regression in two ROIs in the source space. First column: averaged statistics of the regression effects across participants (color code: blue: *residual Layer 1*; green: *residual Layer 7*; red: *common components*). Second to fourth columns, pairwise differences among the three feature groups (color code: cyan: *residual Layer 7-residual Layer 1*; yellow: *residual Layer 1- common components*; magenta, *residual Layer 7-common components*). The transparent bands indicates bootstrapped confidence interval at the participant level. Gray areas indicate time windows where the differences were significantly non-zero.

The gray boxes indicate the time windows that survived the correction, and the p-values of the *permutation-excursion t-tests* (before the FDR correction) are marked. There appeared to be some time windows that did not survive the correction, but in which we could visually observe some possibly non-zero differences based on the confidence intervals (which were not corrected for multiple ROIs). However, we were not able to claim that these windows had significantly non-zero difference in the comparisons.

Next we review the comparisons between all pairs among the three groups of features (i.e., the *common components*, the *residual Layer 1* and the *residual Layer 7*). In the second column of Figure 5.11a, the cyan plots show the differences between the correlation effects of the *residual Layer 7* (the unique high-level features) and the *residual Layer 1* (the low-level features roughly orthogonal to high-level features). In the EVC, we can see a negative peak roughly within 60 to

Figure 5.15 and Figure 5.16 in the Appendix (Section 5.5).

100 ms, indicating that the correlation effect of the *residual Layer 7* was smaller than that of the *residual Layer 1*. After this time window, the difference became positive roughly from 120 ms to 260 ms, indicating that the correlation effect of the *residual Layer 7* was greater than that of the *residual Layer 1*. These results are consistent with the patterns of the blue and green curves in the first column, which demonstrated an early-to-late shift of the correlation effects from low-level to high-level features. In the LOC, the difference showed a similar pattern to that in the EVC, but the early negative difference was smaller and the later positive difference lasted longer, corresponding to the profiles of the correlation effects in the first column. Our observations in this comparison are generally consistent with feedforward information flow along the hierarchy. The reader might have expected that the LOC, which is at a higher-level within the hierarchy than the EVC, would show smaller correlation effects with low-level features than the effects observed in our results. We speculate that because the columns in the forward matrix inherently have strong spatial correlations, the reconstructed source solutions can be spatially blurred, and the correlation effects with low-level features can "leak" from lower-level visual areas into the LOC. Moreover, even without spatial blurring, the hierarchy can be gradual such that we may only observe small relative differences between the correlation profiles in the EVC and the LOC.

In the third column, the yellow plots show the differences between the *residual Layer 1* (the low-level features roughly orthogonal to high-level features) and the *common components* (the object-category-relevant low-level features). In the EVC, we observed a positive peak roughly from 60 to 80 ms, followed by a negative peak roughly from 120 to 180 ms; these patterns indicate an early-to-late shift from the *residual Layer 1* to the *common components*, which is reasonably consistent with the pattern in the second column (*residual Layer 7-residual Layer 1*), given that the *common components* were correlated with high-level features. Interestingly, we also observed another negative time window near 300 to 380 ms, which were close to the second peaks of the correlation effects of the *residual Layer 1* and the *common components* starting near 300 ms in the first column. These late peaks of the correlation effects were likely due to the neural responses to the disappearance of the stimuli, and the negative difference in this time window verifies our observation in the first column that the EVC showed higher and longer correlation effects with the *common components* than with the *residual Layer 1*. Now let us consider what mechanisms can lead to such results. When the images disappeared, the changes of the visual input might cause neural responses. Because these changes of visual inputs were image-specific, the corresponding neural responses were likely to be correlated with features of the images. Therefore it was not surprising that the activity in the EVC near 300 ms was correlated with the low-level features of the images, including the *common components* and the *residual Layer 1*. However, the disappearance of the stimuli did not add new information about image contents; thus the neural responses in the EVC were like a "replay" of the stimuli. We observed that the EVC showed higher and longer correlation effects with the object-category-relevant low-level features than with the low-level features that were roughly orthogonal to high-level features; this result is unlikely to be generated by pure feedforward information flow, because in that case, the correlation effect would have been similar for the two groups of low-level features. Indeed, our result suggests that some non-feedforward process (e.g., top-down feedback within the hierarchy) may help the EVC to separate the two groups of low-level features, and mainly "replay" the object-category-relevant low-level features. In the LOC, there were an early and a late negative windows, indicating that the correlation effect of the *common components* was larger than that of the *residual Layer 1*. These results are consistent with the hypothesis of the hierarchical organization, in that the LOC, being at a higher level in the hierarchy, showed higher correlation effects with the object-category-relevant low-level features than with the low-level features that were roughly orthogonal to high-level features.

In the fourth column, the magenta plots show the differences between the *residual Layer 7* (the unique high-level features) and the *common components* (the object-category-relevant low-level features). In the EVC, we observed a negative window roughly from 60 to 100 ms, followed by a positive window near 260 ms; such a pattern again showed an early-to-late, lower-level-to-higher-level shift. There was a later negative window centered near 350 ms, corresponding to the late peak at 300 to 380 ms of the red curve in the first column (i.e., the correlation effects of the *common components* after the disappearance of the stimuli). In the LOC, we observed a negative window centered around 100 ms and a positive window centered around 200 ms, but no later negative window near 350 ms as the one in the EVC. The early-to-late shift was similar to that in the second column; patterns in both the second and the fourth columns showed the differences between the unique high-level features (the *residual Layer 7*) and low-level features (the *common components* or the *residual Layer 1*), supporting feedforward information flow in the hierarchy.

Besides the EVC and the LOC, we also included results in the other ROIs (Vent 1, 2 and 3 and the scene-selective ROIs); see Figure 5.15 (MEG) and Figure 5.16 (EEG) in the Appendix (Section 5.5). In the MEG results in Figure 5.15, the correlation profiles in the Vent 1 were similar to those in the EVC, possibly because the Vent 1 and the EVC were spatially close. However, the relative differences in magnitude between the correlation effects of the unique high-level features (*residual Layer 7*) and the two groups of low-level features (i.e., the *common components* and the *residual Layer 1*) appeared smaller than those in the EVC. Vent 2 had smaller correlation effects in general, and the pattern was more similar to that in the LOC. The correlation effects can be barely seen in Vent 3, probably because the signal strength of visual responses in the anterior temporal lobe was low. The patterns in the PPA, TOS and RSC were generally similar to those in the LOC, Vent 1 and Vent 2. In the EEG results in Figure 5.16, we observed generally similar patterns to those in MEG. However, the regression profiles were more noisy and the separation of profiles between the three groups of regressors was less obvious. Nevertheless, in the EVC, the significant negative difference between the *residual Layer 1* and the *common components* in the third column (yellow) was still observed within the time window of 300 to 350 ms, which was likely to include neural responses to the disappearance of the stimuli.

In addition to the two-step approach of source-space regression analysis, we also applied our one-step STFT-R model. In this method, instead of fitting a model for each group of regressors separately, we concatenated the three groups of 6-dimensional regressors together into 18-dimensional regressors. In this way, we were able to apply the same penalization parameters for each group of regressors. The regression coefficients in the time-frequency domain were transformed back to the time domain, resulting in a time series of regression coefficients for each source point and each dimension of the regressors. Note that all columns in the regressors had zero mean and the same L_2 norm, such that after fitting the regression coefficients, the regressors that corresponded to the regression coefficients with larger magnitudes explained the neural data better. We also note that both positive and negative coefficients were meaningful, so we computed the sum of squares of the fitted regression coefficients for each of the three feature groups, at each time point and each source point. These sums of squares, which indicated the relative strength of the correlation effects in the three groups, played a similar role as the R-squared values in the two-step approach. The remaining normalization and statistical tests were the same as in the two-step approach mentioned above, except that we replaced the R-squared values with the sums of squared coefficients. It is also worth noting that because the STFT-R induced spatial sparsity, some source points within the ROIs had zero regression coefficients; these source points were excluded when we computed the average within each ROI.

Figure 5.11b shows the MEG results for the EVC and the LOC, where the color code was the same as that in Figure 5.11a. We can see that the results were similar to those given by the two-step approach, yet the STFT-R yielded much smoother profiles, and as a result, the detected windows were relatively longer. Note that in the third column in the EVC, the negative difference between the *residual Layer 1* and the *common components* within 300 to 400 ms was robustly detected, indicating that the EVC was able to differentiate the two types of low-level features in this time window, which was likely to include neural responses to the disappearance of the stimuli. Results in the other ROIs are shown in the supplementary figures in the appendix (Section 5.5, Figure 5.17a (MEG) and Figure 5.17b (EEG)). Note that some ROIs only had non-zero source points for a few participants (e.g., PPA and RSC in MEG, and the majority of ROIs in EEG); therefore, we did not show results from those ROIs. In EEG, the profiles of the three groups were very similar, and we were not able to detect any significant time windows in pairwise comparisons. The reason could be that the learned coefficients were spatially very sparse; therefore we did not have enough participants, for whom the ROI had non-zero source points, to obtain reliable comparison results.

As a summary, using the regression analyses in the source space, we obtained spatio-temporal correlation profiles between neural activities and the three groups of features (the *common components*, the *residual Layer 1* and the *residual Layer 7*). By analyzing these profiles, we observed progressive shifts from early to late time windows, from lower-level to higher-level features, and from low-level regions to higher-level regions along the hierarchy. These results strongly support feedforward information flow along the hypothesized hierarchy. More interestingly, we also observed that in a time window that was likely to include neural responses to the disappearance of the stimuli, the early visual cortex showed a higher correlation with the object-category-relevant low-level features (the *common components*) than with the low-level features that were roughly orthogonal to high-level features (the *residual Layer 1*). This result suggests that some non-feedforward process (e.g., top-down influences) might help the EVC to distinguish between the two types of low-level features and mainly represent the object-category-relevant low-level features.

5.3.4 Source-space connectivity analysis of the MEG data

In this section, we address Question (2)—how neural activities in different areas interact with each other dynamically. When an image is presented on the screen, the neurons in the visual cortex are hypothesized to extract information about the image and pass it on within the hierarchy. If one brain region (ROI 1) passes information about the image to another brain region (ROI 2), then the activity in ROI 1 is likely to show some statistical dependence with, or be able to predict the later activity in ROI 2. Here we can use our one-step state-space model in Section 4.3 and a corresponding two-step model that uses the dSPM method for source localization, to examine whether the neural responses to the stimuli in one region were able to predict later neural responses in other regions in our experiments. Such analysis can be viewed as an indirect way of inferring possible information flow between regions. Although the dependence effect identified in this way does not sufficiently mean direct information flow between the two regions, such analysis can provide insights to form hypotheses of cross-region interactions, which can be tested in animal studies where researchers perturb the hypothesized interactions. Due to the limited number of observations (in our case, we only have neural responses to 362 images) and for tractability, we constrain ourselves to linear autoregressive models of order 1 and focus on the autoregressive coefficients for several pre-defined regions of interest.

Because our MEG data had more sensors and more reliable sensor locations in relation to the scalp than the EEG data, the source localization in our MEG data was better; therefore, we only ran the analysis on the MEG data. We focused on six regions of interest, including the early visual cortex (EVC), the object/scene-selective regions at a higher-level than the EVC in the hypothesized hierarchy (LOC, PPA, RSC and TOS), and the medial orbitofrontal cortex (mOFC), which was roughly in or near the location of the mPFC in the "contextual network" introduced by Bar and Aminoff [2003]. Note that during preprocessing, the mean response across all images was already removed; thus we directly applied the connectivity models, which assumed zero mean across observations, on the data. For each participant, we learned the time-varying autoregressive coefficients ($\{\boldsymbol{A}_t\}_{t=1}^T$) and also bootstrapped the 362 observations 15 times to obtain standard deviations of entries in $\{\boldsymbol{A}_t\}_{t=1}^T$. Then we computed a Z-score for each entry, defined as the fitted value divided by the bootstrapped standard deviation. Afterward, we obtained a p-value from a two-sided Wald test against the standard normal null distribution for each entry using its Z-score. For visualization, we first took $-\log_{10}$ of the p-values for each participant at each time point, and then took the average across the participants. Some of the p-values were extremely small; to avoid getting infinitely large logarithms, we truncated the p-values to be at least 10^{-4}.

Figure 5.12 shows the plots of the averaged $-\log_{10}(p\text{-values})$ for each pair of ROIs. The results for each individual participant by the two-step method and our one-step state-space method are shown in the Appendix (Section 5.5, Figure 5.18 and Figure 5.19). Each diagonal subplot corresponds to the autoregressive (AR) coefficients that described the linear dependence of the ROI's current activity on its activity 10 ms ago. Each off-diagonal subplot corresponds to the AR coefficients that described the lagged linear dependence between a pair of ROIs—for example, the subplot in the first row and the second column describes the linear dependence of the activity in the EVC on the activity in the PPA 10 ms ago. In Figure 5.12, we observed strong self-dependence in the diagonal subplots, especially after the stimulus onset (0 ms). This result indicates that the activity in each ROI has strong temporal dependence. For lagged dependence between different ROIs (in the off-diagonal subplots), the one-step method yielded stronger effects than the two-step method for some of the ROI pairs, which can also be seen in the individual subplots (Figure 5.18 and Figure 5.19).

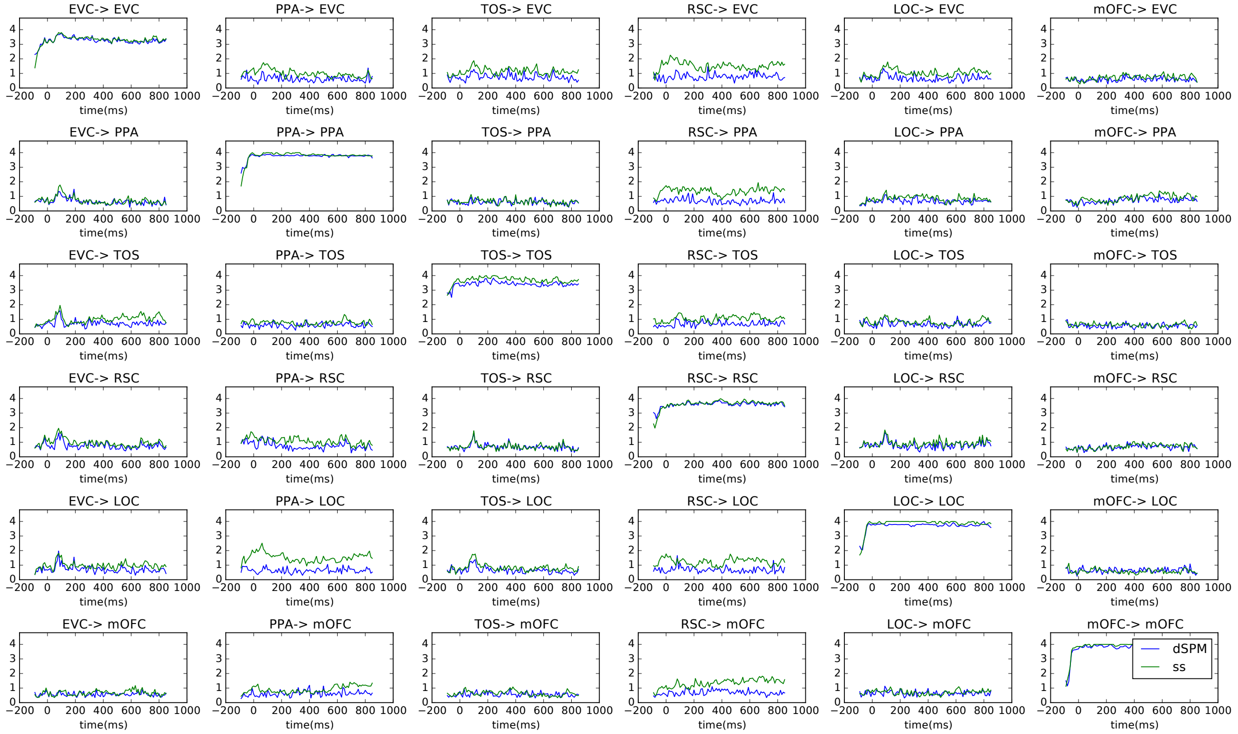


Figure 5.12: Mean $-\log_{10}(p\text{-value})$ across 18 participants in the source-space connectivity analyses, for each pair of ROIs, blue (dSPM): the two-step method; green (ss): the one-step state-space method.

To identify in which time windows the dependence effects were significant, we ran binomial tests [Darlington and Hayes, 2000] to combine the p-values across participants at each corresponding time point and for each corresponding pair. Figure 5.13 shows the $-log_{10}$ of the p-values from the binomial tests, where the threshold at a level of 0.05 after the Bonferroni correction for all the time points and all 6×6 pairs is indicated by the black lines. We are mainly interested in the off-diagonal subplots, which show dependence effects across different ROIs.

The results by the two-step and the one-step methods were similar in the diagonal subplots and some off-diagonal subplots. In other off-diagonal subplots, the one-step method yielded more significant time points than the two-step method. In some of these subplots (e.g., PPA $\rightarrow$ EVC, TOS $\rightarrow$ EVC, LOC $\rightarrow$ EVC), we can see the dependence effects by the one-step model were low in the baseline time window before the stimulus onset and increased afterward; however, in other subplots (e.g, RSC $\rightarrow$ EVC, PPA $\rightarrow$ RSC, RSC $\rightarrow$ LOC), the dependence effects detected by the

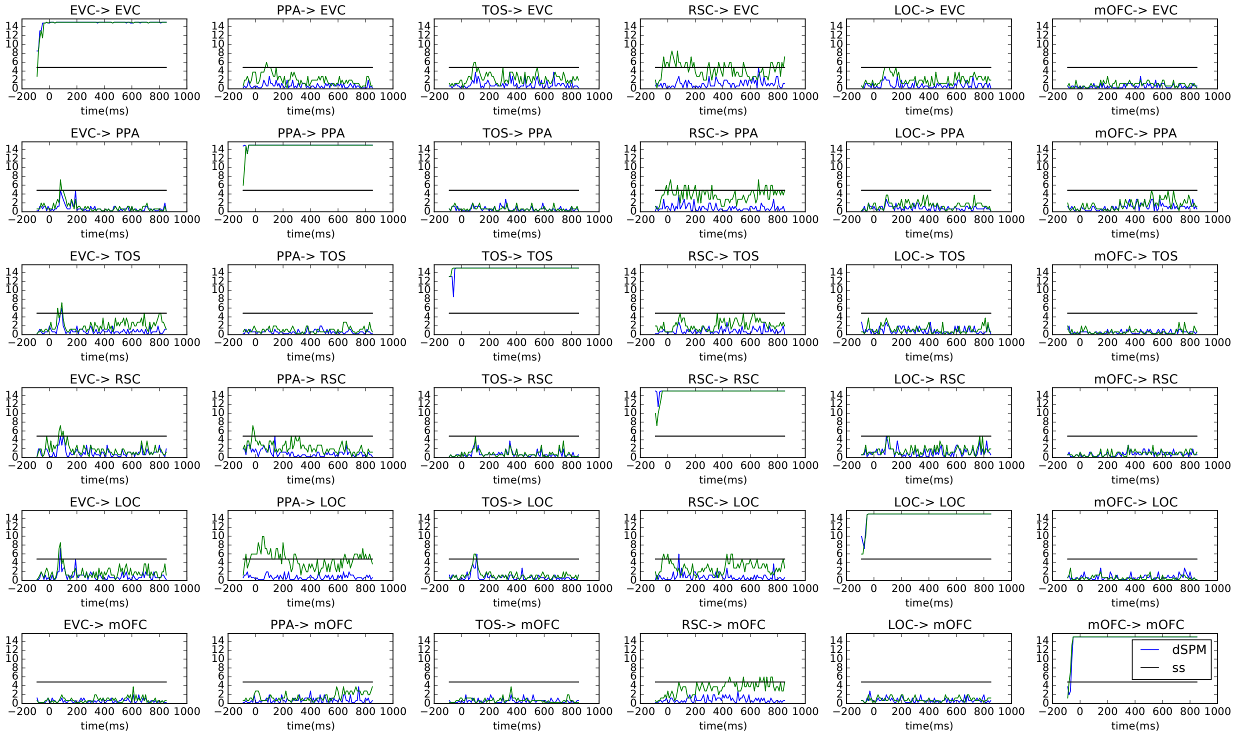


Figure 5.13: The $-\log_{10}(p\text{-value})$ values of the binomial tests, for each pair of ROIs. blue (dSPM): the two-step method; green (ss): the one-step state-space method. Black lines indicate the Bonferroni threshold.

one-step method were high in the baseline time window before the appearance of the images, which could be "false positives" in the sense that they were irrelevant to the image-specific information. As discussed in Chapter 4, the one-step state-space model had specific assumptions about the ROIs involved and about the distribution of source point activity given the ROI activity. Some of the assumptions might not be perfectly met in the data here. Especially, we only included several ROIs for tractability, but activities in missed regions could be incorrectly attenuated or assigned to other ROIs. In contrast, the two-step method is more robust to ROI specification, yet the source estimates by the dSPM could be incorrect due to ignoring the cross-region dependence. The underdetermined nature of source localization and the spatial correlations in the forward matrix bring challenges to both the one-step and the two-step methods, making it difficult to correctly separate activities in ROIs that are spatially close (e.g., EVC and RSC, PPA and RSC, PPA and LOC). Given these challenges, neither the one-step method nor the two-step method is universally better than the other. The significant dependence effects detected only by one-step method might be a mix of false discoveries and genuinely improved estimates. Based on these reasons, we view our results here only as preliminary observations, and hope to further validate them with other neuroimaging modalities (e.g., intra-cranial EEG). Below, we will mainly focus on the subplots where there were relatively small (and insignificant) dependence effects in the baseline time window before the stimulus onset.

Consistently in the results by both methods, we observed significant dependence effects above the

Bonferroni threshold in the first column, (EVC → PPA, TOS, RSC and LOC), starting near 60 to 70 ms and peaking near 80 ms. This time window also corresponded to the initial activation of the early visual cortex that was correlated with low-level features of the images (see the correlation profiles in the EVC in Figure 5.15). The effects in this window could be consistent with feedforward information flow from the early visual cortex to other higher-level regions.

In the symmetric positions, the first row, we also observed significant lagged dependence between the object/scene selective regions and the EVC (PPA, TOS and LOC → EVC), generally starting at 70 to 80 ms and lasting to at least about 120 ms. Such dependence was identified mainly by the one-step method, but the results by the two-step method also show insignificant trend in some subplots (e.g., LOC → EVC). Some of the effects detected in the above pairs by the one-step method also appeared in later time windows (e.g., LOC → EVC at 170 ms). Together, these effects in the first row of off-diagonal subplots may reflect top-down interactions from the scene/object-selective regions to the EVC, after the information is first passed in the feedforward direction.

In the other off-diagonal plots, we can see dependence effects peaking near 100 ms for TOS → RSC, TOS → LOC, RSC → TOS, and LOC → RSC. These regions are all at a relative higher-level than the EVC within the hierarchy, yet the order within these regions has not been clearly established. The significant dependence effects near 100 ms may reflect either interactions among these regions or common influences from other regions not modeled here. Interestingly, we also observed some relatively late dependence at RSC → mOFC and nearly significant dependence at mOFC → PPA by the one-step model; such dependence could indicate some communications between the mOFC and the scene-selective regions, which are possibly involved in contextual processing and episodic memory.

5.4 Discussion

In this chapter, we aimed to address two questions on the information flow in the visual cortex: (1) what kind of information is extracted in different temporal stages and different areas, and (2) how neural activities in different areas interact with each other dynamically. To answer Question (1), we regressed the neural responses on the features from different layers in `AlexNet`, and also on three groups of features derived from Layer 1 and Layer 7. Specifically, we compared the spatio-temporal correlation profiles of these three groups, which represented the object-category-relevant low-level features (the *common components*), the low-level features roughly orthogonal to high-level features (the *residual Layer 1*), and the unique high-level features that were roughly orthogonal to low-level features (the *residual Layer 7*). Our results indicated an early-to-late shift from lower-level features to higher-level features, consistent with feedforward information flow. Moreover, by contrasting the correlation effects of the *common components* and the *residual Layer 1*, we found that in a time window that was likely to include neural responses to the disappearance of the stimuli, the early visual cortex showed a higher and longer correlation effect with the *com-*

mon components (the object-category-relevant low-level features) than with the *residual Layer 1* (the low-level features roughly orthogonal to high-level features). This result indicates that some non-feedforward processes, such as top-down influences, are likely to help the early visual cortex to represent the two types of low-level features differently.

Note that some previous work in MEG also tried relating neural activities to features derived from sophisticated computer vision models [Clarke et al., 2014; Cichy et al., 2016c,a]. However, the majority of the previous work focused on the temporal patterns of correlations, and very few explored the joint spatio-temporal patterns. The temporal patterns in these previous studies showed an early-to-late shift from lower-level to higher-level features, which is consistent with what we found. Among the few studies that briefly looked at spatio-temporal patterns, Clarke et al. [2014] showed correlations of neural activity with visual and semantic features in the source space using a two-step approach (although the authors did not show the full time course in each region of interest), which supported feedforward information flow as well. Additionally, Cichy et al. [2016d] "fused" fMRI and MEG recordings, by comparing the representational similarity of visual objects in the two imaging modalities without using external features; they observed an early-to-late shift along the feedforward direction of the hierarchy in the visual cortex as well. The similar early-to-late, lower-level-to-higher-level shift in our results is consistent with these previous findings. Moreover, our study here provides more comprehensive spatio-temporal profiles to answer Question (1), by using features derived from a more sophisticated computer vision model (`AlexNet`) than the one in [Clarke et al., 2014] (see a comparison in [Yamins et al., 2014]), using a relatively large number of naturalistic scene images (rather than single objects on a blank background in [Clarke et al., 2014] and [Cichy et al., 2016d]), and using our novel decomposition of the features. In addition, the consistency between the results given by the conventional two-step approach and our novel one-step STFT-R method also reassures that our findings were robust to the source localization methods, which was another novel point in our work here. Finally, we also found some novel evidence of non-feedforward processing in the early visual cortex, which is useful in building better computational models of the visual cortex by incorporating recurrent structures.

To address Question (2)—how neural activities in different areas interact with each other dynamically, we examined the time-lagged dependence across several ROIs along the hierarchy, using a time-varying autoregressive model of order 1 in both the two-step framework and our novel one-step framework. In the preliminary results of these dynamic connectivity analyses, we observed leading and lagged linear dependence between the early visual cortex (EVC) and the object/scene-selective regions at a higher level than the EVC. Such a dependence structure during the processing of naturalistic scenes has not been described in previous literature to our knowledge. The dependence can be viewed as indirect evidence of both feedforward and feedback interactions. Limited by the observational nature of our experiments and the challenges of source localization, we could not conclude about causal interactions. However, our novel results may provide some insights on the dynamic connectivity in the visual cortex and help us build detailed hypotheses in terms of timing and directions of cross-region interactions for further tests.

In our analyses, we rigorously sought for evidence of feedforward and feedback information flow along the hypothesized hierarchy of the visual cortex. However, we do admit that the nature of our experiment is exploratory to a large extent. Below we discuss some issues in our experiments and analyses and point out relevant future directions.

Neural responses to the disappearance of stimuli

In our experimental design, after 200 ms of presentation, the stimuli disappeared and the screen switched back to a "+" on a gray background. This disappearance caused changes in the visual input, which might drive neural responses in the early visual cortex. In fact, Liang et al. [2008] characterized the magnitudes of responses to the disappearance of stimuli in cat V1, which were comparable to the magnitudes of responses to the appearance of stimuli. In our analyses, although the mean responses to the appearance and disappearance of all stimuli were subtracted from the data during preprocessing, the image-specific responses (reflected in deviations from the mean) could still be correlated with features of the images. This explained why we observed a late peak of the correlation effects of the low-level features (the *common components* and the *residual Layer 1*) in the EVC near 300 ms (considering a 80 to 100 ms delay for the changes of visual input to get to the EVC). However, the disappearance of stimuli should not be treated as equivalent to the appearance of stimuli, because no additional semantic information is provided in these changes of visual input. Consistently, the unique high-level features (the *residual Layer 7*) did not appear to show an increase of correlation effects in response to the stimulus disappearance in our results. To some extent, the responses to the disappearance of stimuli can be viewed as a "replay" of the images. In our data, the early visual cortex showed a stronger and longer correlation effect with the *common components* (the object-category-relevant low-level features) than with the *residual Layer 1* (the low-level features that were roughly orthogonal to high-level features), as if the early visual cortex was guided to "replay" more of the low-level features that were related to semantic information. If we assume the visual system only has feedforward processing, we expect to see similar correlation effects for both groups of low-level features (as seen in the early correlation effects near 100 ms in the first EVC plot of Figure 5.11a) instead of what we observed. Indeed, some non-feedforward processes, such as feedback from the higher-level cortex to the EVC, or lateral recurrent interactions of the neurons in the early visual cortex, explain our results better.

It is worth noting that the "replay" may evoke visual aftereffect perceived by the participants. Hence our results also provide some insight to further test our speculation via behavioral experiments. We hypothesize that if the participant could perceive some aftereffect after the stimulus disappears, then the aftereffect should be mainly driven by the object-category-relevant low-level features. We could design stimuli to manipulate the such features, and verify this hypothesis in behavioral experiments.

We chose a short presentation duration (200 ms) to reduce artifacts of saccades, and did not use any masks (e.g., white noise patterns) after the stimuli. Our findings related to the responses to the disappearance of stimuli may be viewed as a result of design limitation, where the disappearance of the images interfered with the intact dynamics of visual processing. However, due to this arguable limitation, our experiments also provided some novel observations about the visual system in an unexpected way, which has not been discussed much in the literature as far as we know. In future experiments, we can examine how the differential coding of the different types of low-level features (*common components* and *residual Layer 1*) change if we vary the duration of stimulus presentation.

Local contrast

In the canonical correlation analysis of the intact features from `AlexNet`, we observed that the CCA components between Layer 1 and 7 appeared to show high correlations with local contrast features. In fact, in our first attempt of data analysis, we did not partial out the local contrast features when obtaining the three groups (the *residual Layer 1*, the *common components* and the *residual Layer 7*). In that case, we found the correlation effects of the *common components* were much higher than those of the *residual Layer 1* and the *residual Layer 7*, due to the large the neural correlation with the local contrast features. In this sense, the local contrast is a confounding factor. Nevertheless, this confounding factor could be inherently included in the statistical regularities of naturalistic images. Intuitively, local receptive fields with high contrasts often contain informative features about shapes and boundaries and are consequently related to semantic information. Further tests of this hypothesis, which can be implemented with a large set of images, may help us better understand the nature of naturalistic images. Moreover, in future experiments, to alleviate the influences of this confounding factor, we can design new stimuli by adding high contrast features that are less relevant to semantic information, for example, irregular shadow contours. It would be interesting to study whether these high-contrast features are coded differently from the genuinely informative high-contrast features (e.g., true physical edges or contours) in the human visual cortex, as well as in a CNN trained on regular naturalistic photos. Moreover, we can also use the generative neural networks [Goodfellow et al., 2014] to create stimuli that share some similarity in local contrasts and other low-level features with naturalistic images but do not contain objects.

Confounding factors in data-driven experiments

As mentioned above, when analyzing the correlation effects of the three group of features (the *residual Layer 1*, the *common components* and the *residual Layer 7*), we discovered that local contrast was a confounding factor. We reduced this confounding influence by regressing out from the features the principal components that explained 90% of the variance in the local contrast. However, there might be other similar factors inherently in the image statistical regularities, which have strong a correlation with neural responses but are distributed unevenly across the three groups. This is a limitation of using naturalistic visual stimuli, where the distribution of features could not be designed. However, since the visual inputs in the world typically have a large dimensional and complicated feature space, it is difficult to form good hypotheses that capture the space well from scratch. We view our data-driven exploration here as an important initial step, which helps to make new predictions for future hypothesis-driven experiments.

Choice of using `AlexNet`

We used an 8-layer convolutional neural network (AlexNet), which was trained to classify images into 1000 object categories. However, the stimuli presented to the participants were naturalistic images of scenes, and the participants were likely to do scene recognition and scene understanding during the experiment, although we did not explicitly instruct them to do so. In some preliminary data analysis, which is not presented in this thesis, we used features derived from a network of the same architecture as AlexNet, but trained to classify 250 scene categories [Zhou et al., 2014b], many of which overlapped with the categories in our stimuli. In that case, the neural correlation effects we observed were not significantly higher than those with features from AlexNet, which was trained on object recognition. The reason may be that the features learned in AlexNet have good transferability to other tasks, as suggested by Yosinski et al. [2014]; Huh et al. [2016], possibly because these features well characterize the statistical regularities of the visual world. Moreover, a scene contains many object elements, and thus scene understanding may benefit from good mid-level and high-level representations of objects. Indeed, Zhou et al. [2014a] showed that object detectors emerged in a convolutional neural network that was trained to do scene classification. Therefore it is not surprising that object-category relevant features can explain the neural data during scene processing.

Since the publication of AlexNet, deeper and more sophisticated convolutional neural networks have been developed for more diverse vision tasks. Although it will be interesting to use features from these new networks to explain neural data, we do not speculate the results will change qualitatively. On the one hand, as discussed above, the features in AlexNet are already very rich and expressive. On the other hand, empirically, the number of images we can present in neuroimaging studies are typically fewer than $\sim 10^3$; limited by such numbers, it may be hard to see any large advantage of more sophisticated networks. Given the evidence we found of non-feedforward processing in the visual cortex, in future work, we think it is more intriguing to develop and use non-feedforward neural networks, with built-in feedback according to the connectivity structure in the brain. In this way, we can compare the dynamics of the network with the spatio-temporal neural activity in the brain to better understand the information flow.

Limited number of observations

Here we only presented 362 scene images. Compared with the high-dimensional features given by AlexNet, we did not have enough number of observations to fit large regression models that included all dimensions in the features. In future work, it is important to vastly increase the number of observations, by increasing data collection time, and maybe also slightly reduce the number of image repetitions.

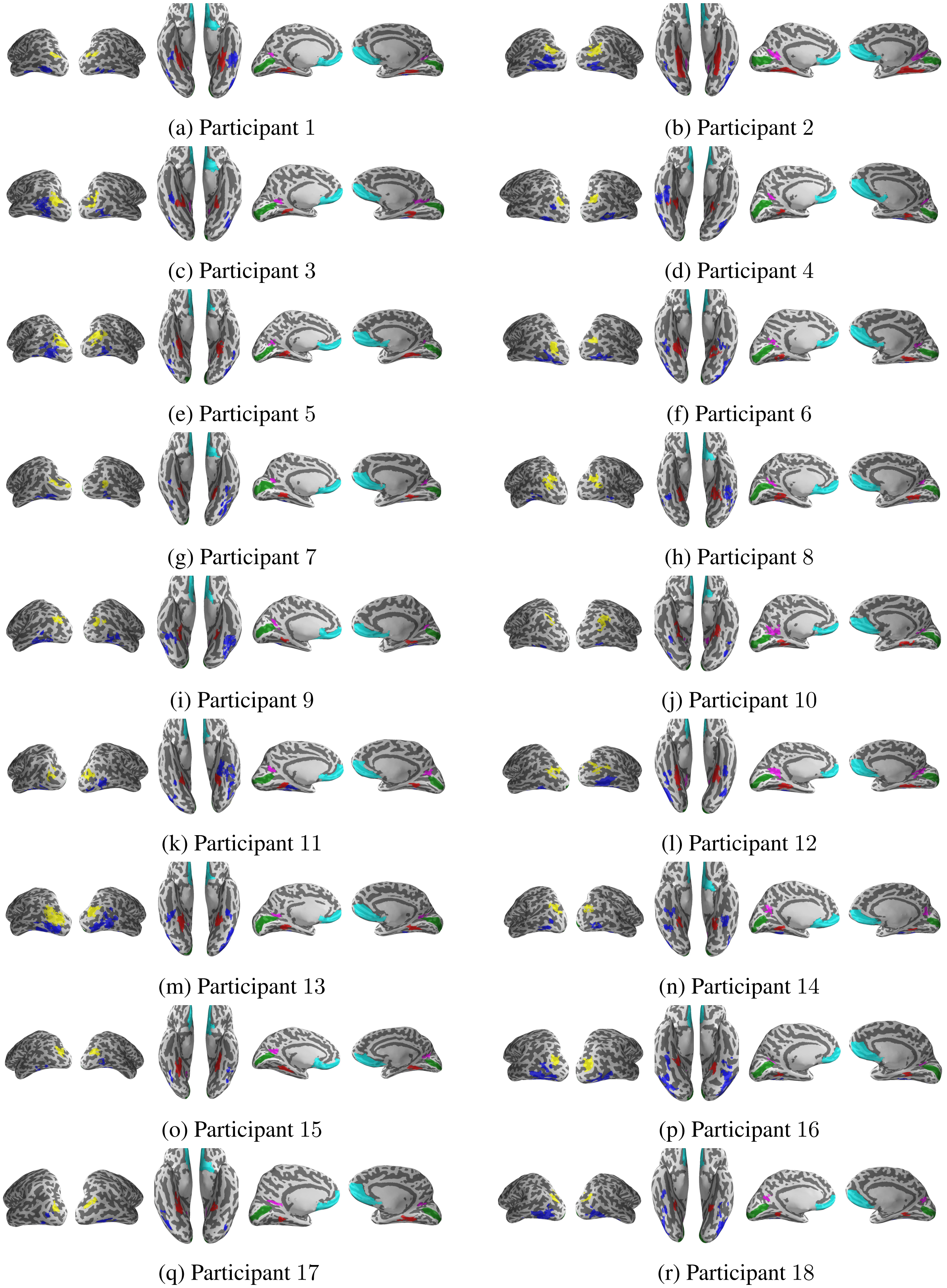
Limitations of the spatial resolution of MEG and EEG

It is worth noting that MEG and EEG are inherently limited by the underdetermined nature of the source localization problem. We have a limited number of sensors, and moreover, the spatial correlations in the columns of the forward matrix can cause spatial blurring effect in the reconstructed source solutions. Both these factors add uncertainty in the localization. Although we have used various methods in our analyses, both traditional two-step methods and our novel one-step models, there is likely a fundamental limit that affects both of them. Moreover, because of the "underdeterminedness", we had to exploit various assumptions in both the two-step and one-step models, and the true neural activity might violate these assumptions. Therefore, although we have tried our best to make reasonable assumptions while keeping the models tractable, we do admit there is some possibility that the source-space results may deviate from the underlying truth. See Chapter 7 for further discussion. We hope the future development of theories on source localization and experimental work with intracranial recordings in patients and animals can further validate our findings.

Comparison between the MEG and EEG results
MEG and EEG differ in their theoretical sensitivity (i.e., lead fields) and the empirical recording systems. In the sensor-space regression, our results in MEG and EEG were similar, yet in the regression analysis in the source space, the EEG results appeared more noisy, and the difference in the correlation profiles of different feature groups were less obvious. Although the difference in the theoretical sensitivity of MEG and EEG may contribute to the difference in the results, the main reason is likely to be that the MEG system had more sensors or higher rank of the data, which contributed to better source localization results. Additionally, in our EEG session, the locations of the sensors were not very accurate due to the digitization system; such errors might also contribute to the difference in results between MEG and EEG. In future work, it would be useful to quantify the impact of such errors in a generic way.

As a summary, in this chapter, by applying source-space regression analysis and connectivity analysis, we explored the spatio-temporal neural dynamics in the human visual cortex while participants processed naturalistic images of scenes. Our results indicate both feedforward and non-feedforward information flow and provide useful insights for further understanding the computation in the visual cortex.

5.5 Appendix: supplementary figures



(a) Participant 1

(b) Participant 2

(c) Participant 3

(d) Participant 4

(e) Participant 5

(f) Participant 6

(g) Participant 7

(h) Participant 8

(i) Participant 9

(j) Participant 10

(k) Participant 11

(l) Participant 12

(m) Participant 13

(n) Participant 14

(o) Participant 15

(p) Participant 16

(q) Participant 17

(r) Participant 18

Figure 5.14: Regions of interest in individual participants, visualized on inflated cortical surfaces. Each subfigure, (from left to right) includes the left and right lateral view, ventral view and left and right medial view of one participant. Color code: parahippocampal place area (PPA): red; transverse occipital sulcus (TOS): yellow; retrosplenial cortex (RSC): magenta; lateral occipital complex (LOC): blue; early visual cortex (EVC): green; medial orbitofrontal cortex (mOFC): cyan.

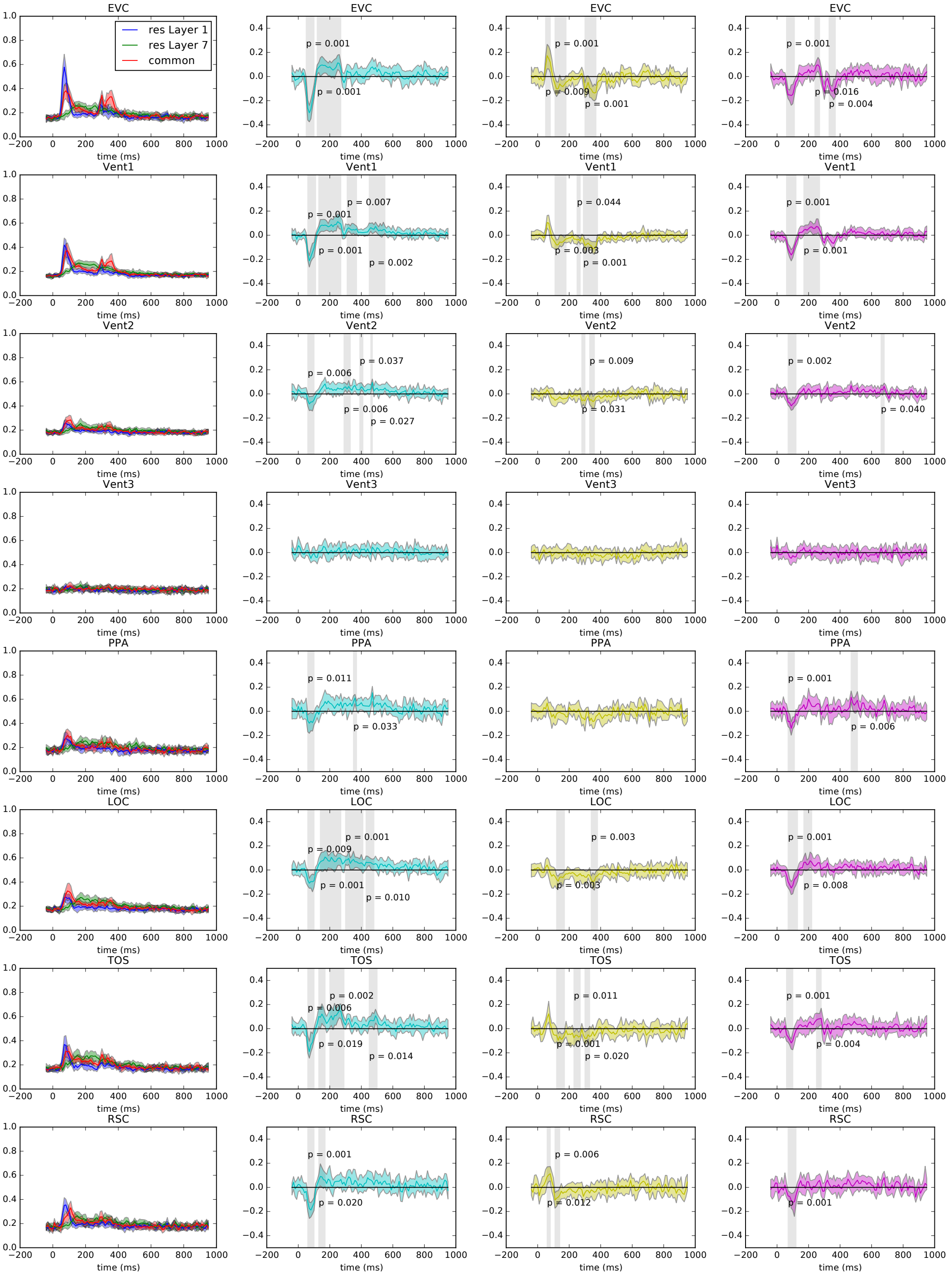


Figure 5.15: The ROI regression results by the two-step dSPM method (MEG). First column: averaged statistics of the correlation effects across participants (color code: blue: *residual Layer 1*; green: *residual Layer 7*; red: *common components*). Second to fourth columns, pairwise differences among the three feature groups. (color code: cyan: *residual Layer 7-residual Layer 1*; yellow: *residual Layer 1- common components*; magenta, *residual Layer 7-common components*). The transparent bands indicates bootstrapped confidence interval at the participant level. Gray areas indicate time windows where the difference was significantly non-zero.

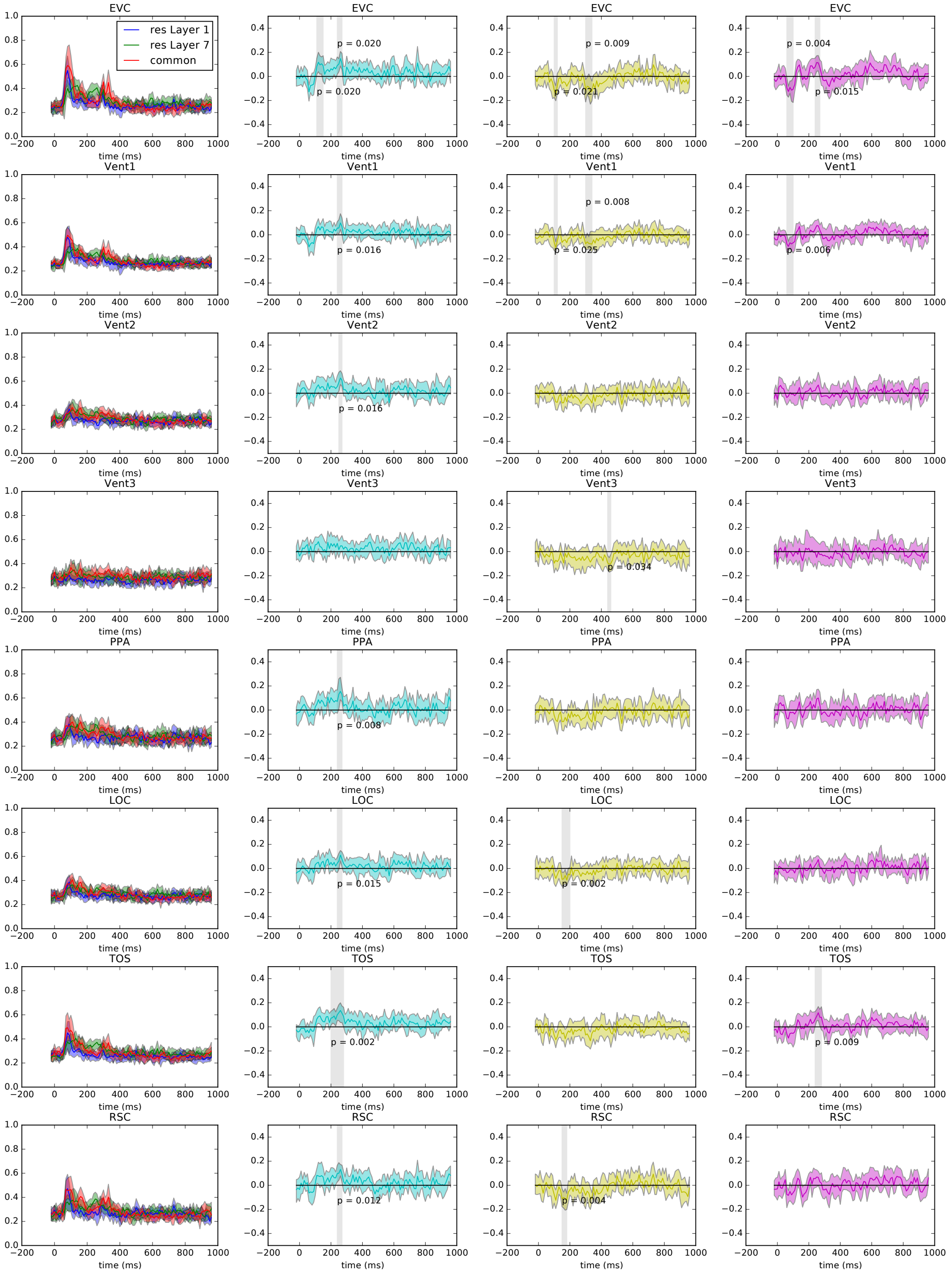
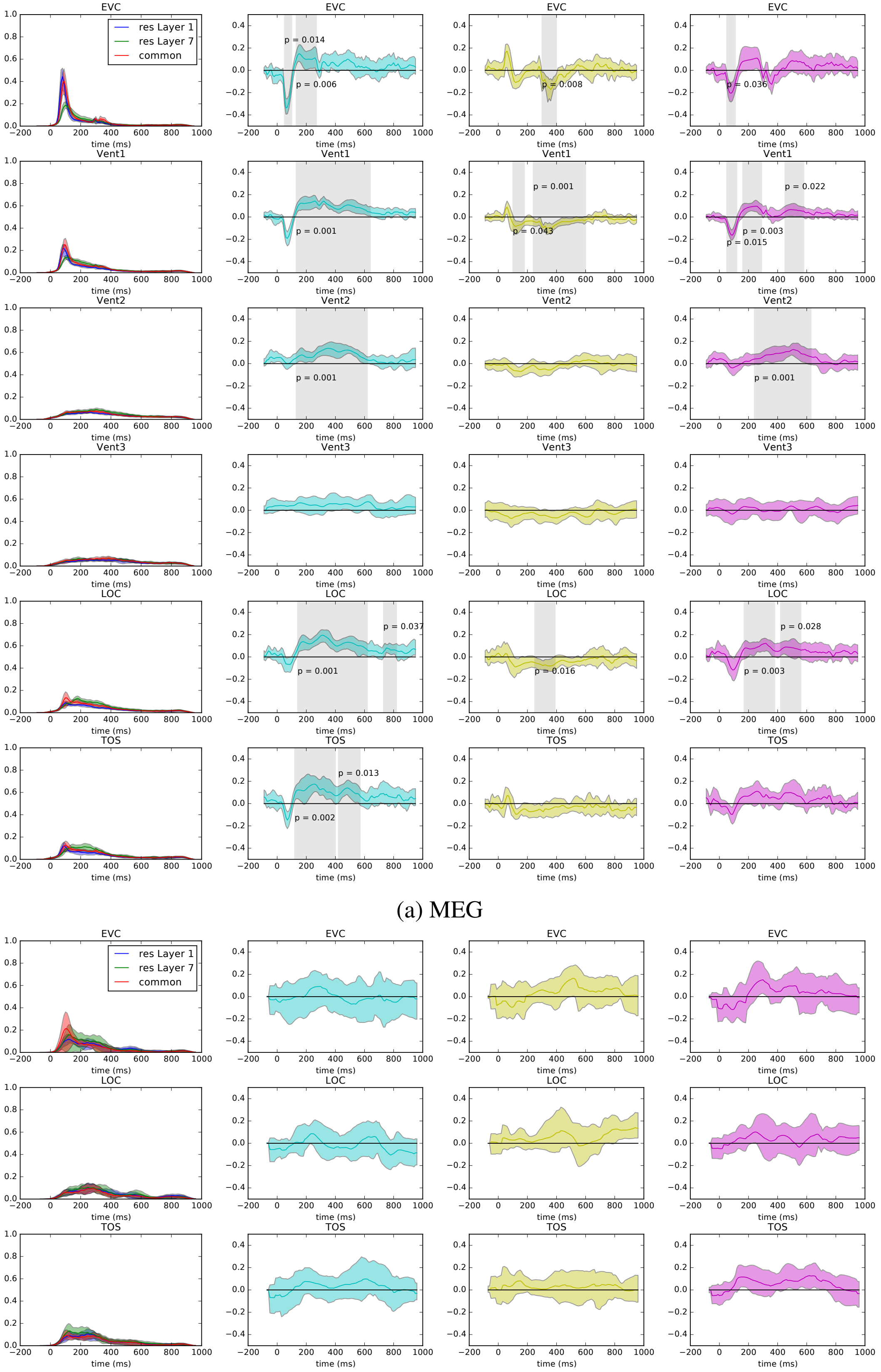


Figure 5.16: The ROI regression results by the two-step dSPM method (EEG). First column: averaged statistics of the correlation effects across participants (color code: blue: *residual Layer 1*; green: *residual Layer 7*; red: *common components*). Second to fourth columns, pairwise differences among the three feature groups. (color code: cyan: *residual Layer 7-residual Layer 1*; yellow: *residual Layer 1- common components*; magenta, *residual Layer 7-common components*). The transparent bands indicates bootstrapped confidence interval at the participant level. Gray areas indicate time windows where the difference was significantly non-zero.



(a) MEG

(b) EEG

Figure 5.17: The ROI regression results by the STFT-R. See the caption of Figure5.15 for detailed descriptions.

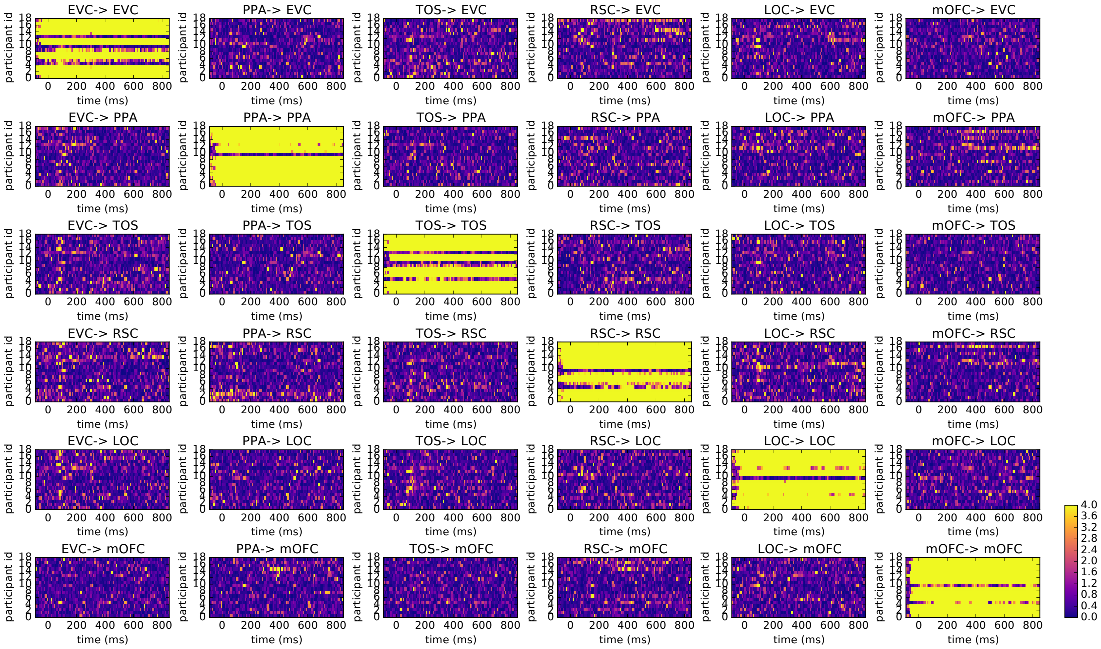


Figure 5.18: Individual $-\log_{10}(p\text{-values})$ from 18 participants by the two-step connectivity analysis.

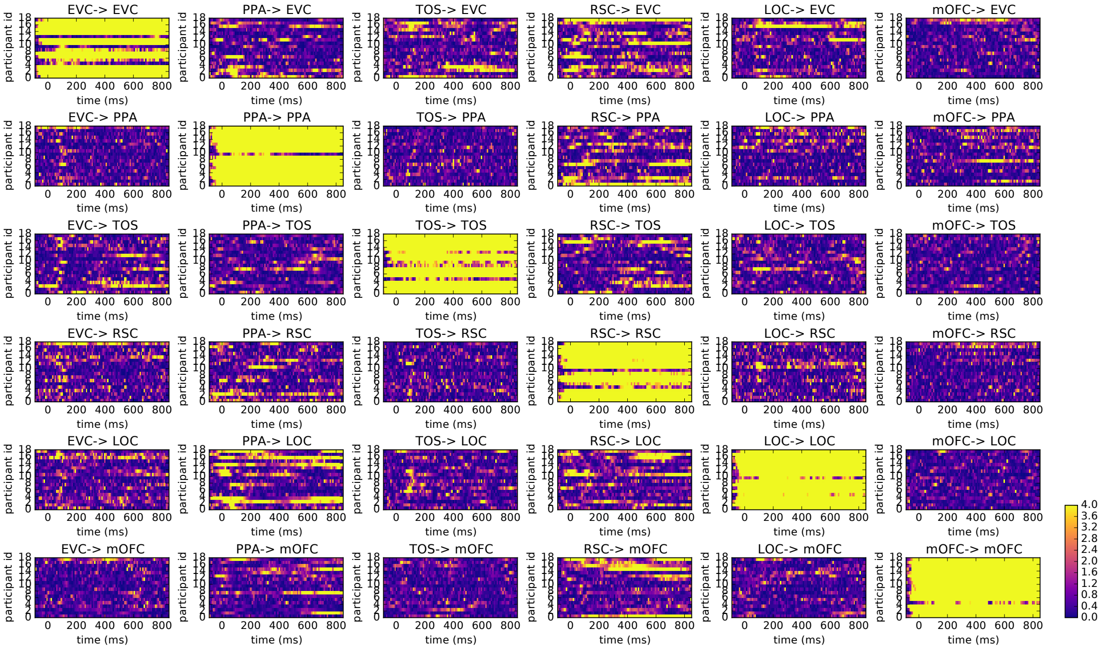


Figure 5.19: Individual $-\log_{10}(p\text{-values})$ from 18 participants by the one-step state-space connectivity analysis.

Chapter 6

Empirical comparison between MEG and EEG

[1]

6.1 Motivation

Although MEG and EEG both measure signals that are induced by neural activity, the two modalities differ in their sensitivity, due to the theoretical properties of their physical mechanisms. In a simplistic case, if we assume the head is a spherical volume conductor, MEG does not detect electric currents along the radial directions, while EEG is able to detect both radial and tangential currents. In more realistic models of the head, there are also similar differences in the sensitivity. Additionally, the magnitudes of the lead fields, which describe the contribution of electric currents in the source space to the sensor recordings, falls off with increasing distance between sensors and sources; such a fall-off effect is larger in MEG than in EEG. As a result, MEG is less sensitive to deeper brain activity than EEG [Hamalainen et al., 1993; Malmivuo, 2012], yet for the shallower currents, the MEG lead fields appear to provide a better spatial resolution [Mosher et al., 1993]. It is worth noting that the spatial resolution of MEG and EEG also depends on the accuracy of forward models. Errors in the conductivity values of head tissues (e.g., scalps and skulls) may result in large errors in the EEG lead fields; in contrast, the MEG lead fields are more robust to such errors [Hamalainen et al., 1993].

In practice, there are more issues that cause differences in spatial resolution and signal quality between MEG and EEG. The first issue is the number of sensors. A typical MEG device is equipped with as many as 306 sensors, including magnetometers, which measure magnetic field strength,

[1] The "stitching" section in this chapter was in collaboration with Dr. William Bishop, who introduced the idea and provided MATLAB implementation of the factor analysis.

and gradiometers, which measure spatial gradients of magnetic fields and thus provide additional information. These sensors give high-dimensional recordings, which cover complementary aspects of the source activities. In contrast, a typical EEG systems has fewer sensors (usually 64 to 128, in some cases 256). Moreover, the scalp voltage measured by adjacent sensors are often highly correlated. Therefore, the rank of EEG recordings is smaller than that in MEG. In this practical perspective, the spatial resolution of source localization is likely to be worse in a 128-sensor EEG system than in a 306-sensor MEG system, simply because the "underdeterminedness" of the source localization problem is even worse with the lower-rank EEG data. The second issue is the signal quality (or noise level) of individual sensors. EEG electrodes are often affected by low-frequency drifts, or noise in the reference, whereas MEG sensors, especially the gradiometers, are less affected by such issues, providing better signal quality.

However, MEG requires expensive superconducting quantum interference devices (SQUIDs) to pick up the weak magnetic field changes induced by neural activity, whereas EEG devices are generally much cheaper. For example, our experimental cost per hour was $\$500$ for MEG and $\leq \$50$ for EEG. In practice, researchers may have relatively tight budgets. In such cases, it is worth evaluating how much extra sensitivity MEG provides. If we are not particularly interested in source-space analysis, but mainly interested in whether the sensor-space data carry certain information in certain time windows (e.g., visual information in scene processing), can we use EEG to achieve comparable sensitivity to that in MEG?

A few previous studies have compared the spatial resolution of MEG and EEG, in which the number of sensors in the two modalities were often equated. These studies gave mixed results depending on whether the study was based on simulations or empirical data, what forward models were used and what source localization methods were used [Lopes da Silva et al., 1991; Liu et al., 2002; Klamer et al., 2015]. The empirical data used in these cases were mostly from relatively simple and well-understood behavioral tasks. Yet there has not been much empirical work on comparing MEG and EEG data acquired with typical devices, in the context of studying complex perceptual or cognitive tasks. In this chapter, we present an empirical comparison of MEG and EEG in our scene-processing experiment in Chapeter 5. Particularly, we focus on neural responses to individual scene images recorded in MEG and EEG. We first compare the signal quality, and explore the similarity between the recordings in the two modalities. Secondly, we compare the sensitivity of MEG and EEG in the sensor space, in testing whether the neural responses at different time points show statistical dependence on visual or semantic features of the images. We note that the empirical spatial resolution of the two modalities depends on the number of sensors (or the rank of data); therefore, we expect the 306-sensor MEG data to yield better spatial resolution than the 128-sensor EEG data in our case. In addition, as mentioned in Chapter 5, the measurement of sensor locations in EEG was significantly worse than in MEG; this issue is also likely to undermine the source-space analysis of EEG data as shown in Chapter 5. Therefore in this chapter, we restrict our comparison in the sensor space.

Due to the theoretical properties of the lead fields of MEG and EEG, the two modalities are complementary, in the sense that if combined, MEG and EEG can provide more complete information about the neural activity than used separately. Later in this chapter, we also explore this direction in our scene-processing experiments. We combine the MEG/EEG recordings in a factor analysis framework and explore the possibility of using both modalities but assigning overlapped subsets of stimuli to each modality. Such a framework may help us obtain good experimental results from the combined data with a smaller cost than using MEG alone.

6.2 Materials and methods

6.2.1 Experimental procedure and data

Details of the experimental procedure, data acquisition and preprocessing can be found in Chapter 5, Section 5.2.5. Data from fifteen participants who finished both the MEG and EEG sessions were used. The preprocessing of data in this chapter was the same as that in Chapter 5 for each participant, with only one difference where trials corresponding to the same image presentation were averaged. The number of repetitions for each image was not always equated between the MEG and EEG sessions for each participant; for some images, there were q_1 repetitions in MEG and q_2 repetitions in EEG, but $q_1 \neq q_2$. In this case, unlike in Chapter 5, where all q_1 (or q_2) repetitions were used in the average in the MEG (or EEG) session, only the first $q_{min} = min(q_1, q_2)$ repetitions were used to compute the averages for the MEG and EEG sessions separately, and the extra ones were discarded. In this sense, the number of repetitions was matched between MEG and EEG for each image.

6.2.2 Visual and semantic features of scene images

We examined whether the neural responses showed statistical dependence with two different sets of features of the stimuli. The first set is a visual feature set, obtained from Layer 5 in AlexNet [2]. We used the *local-contrast-reduced features* of Layer 5 as described in Chapter 5—where the nuisance covariates relevant to image heights, widths, areas and aspect ratios, as well as the principal components explaining 90% of the variance of the local contrast features, were regressed out. Then the first 10 principal components were defined as the visual feature set here [3].

[2] We note that from the results in Chapter 5, the first seven layers all explained a significant amount of variance in the neural responses in the linear regression analysis (Figure 5.7 (b) and (d)); similarly, all these layers would show significant dependence effects on the neural data in the tests in this chapter as well. Nevertheless, our main goal was to examine the sensitivity of the tests here, not to compare the dependence effects between the layers. Therefore we arbitrarily chose Layer 5 here. We expect to get similar results if other layers are used.

[3] The principal component analysis was run on the *local-contrast-reduced features* of the stimulus images and the extra images, but only the principal components corresponding to the 362 stimulus images were included here.

The second feature set included 0/1 semantic features that were used to organize scene categories in the widely used `SUN` database of scene images [Xiao et al., 2010]. Among these features, there were three coarse ones, "indoor', "outdoor natural" and "outdoor man-made"; for each one, there were 4 to 6 finer-grained features, such as "shopping and dining", "workplace" and so on. In total, there were eighteen such features, and we manually assigned 0/1 values corresponding to these eighteen dimensions for each of the 362 stimulus images.

6.2.3 Non-parametric tests of dependence between multivariate variables

A commonly used approach to relating multivariate neural responses to external multivariate variables is the *representational similarity analysis* [Kriegeskorte et al., 2008]. In this approach (see Figure 6.1), consider q multivariate neural responses to q stimuli, and also q external multivariate variables corresponding to the same stimuli (e.g., features of the stimuli or neural responses recorded in a different modality). A $q \times q$ *representational similarity matrix* (RSM) is created for the neural responses (or the external variables), where each (i, j) entry in the RSM is defined as the similarity between the multivariate neural responses (or external variables) corresponding to the ith and jth stimuli. The similarity can be defined using various metrics; here, we use Pearson correlation. After the RSMs for the neural responses and the external variables are obtained, the lower triangular entries of the RSMs (excluding the diagonal), which denote the pairwise similarity between the representations of the stimuli, are concatenated into two large vectors for the neural responses and the external variables respectively. The Pearson correlation of the two concatenated vectors is used as the testing statistic. Under the null hypothesis that the neural responses and the external variables are independent, this statistic should follow a distribution with zero mean. Permutations, where the correspondence of the neural responses to the stimuli is scrambled, can be used to obtain an empirical null distribution and the p-value.

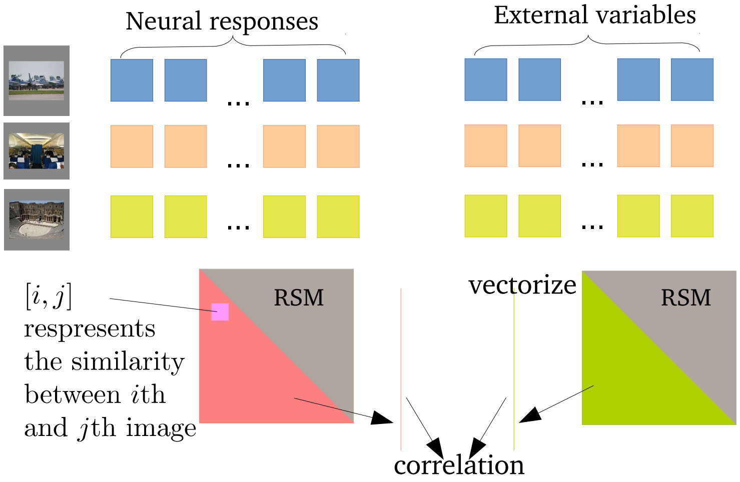

Neural responses
External variables
...
$[i, j]$ respresents the similarity between ith and jth image
RSM
vectorize
RSM
correlation

Figure 6.1: Illustration of the representational similarity analysis.

A similar test, which uses the Hilbert-Schmidt independence criterion (HSIC) [Gretton et al., 2007], can be applied as well. In this test, the pairwise similarity (entries in the RSM) is defined by a kernel function, and the similarity between the two kernel matrices is defined in a similar form to the Pearson correlation in the representational similarity analysis. This HSIC test has good asymptotic properties; the null distribution can be either approximated with analytical forms or obtained by permutations. We used the implementation by Gretton et al. [2007] [4], which exploited a radial-basis Gaussian kernel, where the kernel width was empirically selected based on distances between a small sub-sample of the pairs in the multivariate data.

6.2.4 Quantifying the sensitivity of independence tests

After running the non-parametric independence tests (e.g., the HSIC tests) on the neural responses and the visual or semantic features of the presented images, we defined the *sharp ratio* as the observed testing statistic divided by the critical value at a level of 0.05 according to the empirical null distribution obtained from permutations[5]. In the HSIC tests, the testing statistics were non-negative, so the critical value was obtained in a one-sided manner.

To ideally quantify how sensitive the recordings in MEG and EEG were in detecting the dependence between neural activities and external features, we could compute the statistical power of the tests. However, since we do not know the true alternative distribution, it is not easy to do power analysis. Instead, we use the sharp ratio as an empirical index of the sensitivity. The more observations (images) we had in the test, the higher the sharp ratio would be. To quantify how the sharp ratios vary with increasing numbers of observations, we sub-sampled 10%, 30%, 50%, 70% and 90% of our 362 observations without replacement for 5 times (i.e., the number of sub-samples were 36, 108, 181, 253 and 325). Then we computed the sharp ratio of each sub-sample set at each time point, and took the average across the 5 independent sampling procedures. This procedure was done for each participant in MEG and EEG respectively.

6.2.5 Combining MEG and EEG data using factor analysis

We also explored the possibility of combining (or "stitching") the MEG and EEG recordings using factor analysis. In this analysis, we still focused on examining the statistical dependence between neural activities and visual or semantic features of the scene images. We considered two cases: a case where the neural responses to all stimuli were observed in both MEG and EEG (the full-observation case) and a case where neural responses to some stimuli were missing in MEG (the

[4] http://www.gatsby.ucl.ac.uk/~gretton/indepTestFiles/indep.htm

[5] We note that if there was only a single test at a single time point, a sharp ratio greater than 1 means we could reject the null hypothesis, concluding that the neural responses and the features had a significantly larger dependence than chance. However, here we had multiple time points, so we could not make such conclusions without corrections for multiple comparisons.

missing-observation case). The latter case simulated an experimental design where a subset of stimuli were presented in the MEG session and the full set of stimuli were presented in the EEG session. We learned low-dimensional factors to combine the observations in MEG and EEG corresponding to the overlapped stimuli, and filled in the missing part in MEG using other EEG observations. If for the same number of stimuli, the MEG data had higher sensitivity than the EEG data, then we expected that the filled-in factors corresponding to all stimuli could yield higher sensitivity than the EEG data alone, because they benefited from the MEG data corresponding to the overlapped stimuli besides from the EEG data.

We used the factor analysis model implemented in [Bishop, 2015]. We had $n_{MEG} = 306$ MEG sensors and $n_{EEG} = 128$ EEG sensors; together, we had $n_{both} = n_{MEG} + n_{EEG} = 434$ dimensions of observations at each time point. The mean across the $q = 362$ observations was subtracted from the data. Because the units and scales of the recordings varied across different groups of sensors (including two groups of 102 MEG gradiometers measuring orthogonal gradients, one group of 102 MEG magnetometers, and one group of 128 EEG electrodes), we rescaled the data such that the sum of variances within each of the sensor groups was the same. The observations at different time points were treated as i.i.d. observations in computing the variances. The factor analysis model learned $n_{latent} = 70$ latent factors, from the rescaled n_{both}-dimensional recordings, where data at different time points were treated as i.i.d observations. In the full-observation case, the responses at all time points to all 362 images were used, whereas in the missing-observation case, the MEG responses at all time points were missing for 122 images (from 61 categories randomly sampled from the 181 categories). In both cases, the factor analysis models were estimated using the expectation-maximization algorithm (see Bishop [2015] for details). The number of latent factors $n_{latents} = 70$ was selected because it was an "elbow point", that is, the reconstruction errors in MEG using increasing numbers of latent factors dropped sharply at this value of $n_{latents}$.

6.3 Results

6.3.1 Comparison of signal strength in MEG and EEG

We first focus on the relative strength of signals compared with noise in the MEG and EEG recordings. By "signals", we mean the underlying neural responses to the stimulus images, and by "noise", we refer to the noise from different sources, including inherent sensor noise, physiological artifacts such as eye movements, and the neural activities that are not directly involved in processing the images. It is non-trivial to extract the pure "signals" in this context; as an approximation, we define the signals as the common responses to the stimulus images across participants, up to some linear transformation. In other words, we use the cross-participant reliability, measured in a leave-one-participant-out regression paradigm, as a measurement of signal strength.

For each participant, at each time point and for each sensor, we ran a regression analysis to predict the responses to the 362 images, using the corresponding responses in all the sensors at the same time point from all the other participants. In such an analysis, the regression effect was further quantified in a nested ten-fold cross-validation. For each fold, we used principal component analysis to reduce the high-dimensional regressors (from all the sensors in all the other participants) to a few principal components, for the training and testing data together. We then trained an ordinary least square linear regression model using the training data and computed the regression score—the proportion of variance explained in the testing data—as one minus the mean squared prediction error divided by the variance of the true observations. The regression score was averaged across the ten folds for each sensor, each time point, and each participant.

In terms of how many principal components to use, there was a trade-off—too few components would limit the linear model's ability to explain the variance in the testing data, but too many components would overfit. We were mainly interested in quantifying the highest proportion of variance that can be explained by the other participants data; therefore, we took the maximum of the regression score among those obtained by using various numbers of principal components (20, 40, 60 and 80), for each sensor, each time point, and each participant. Afterwards, the maximum scores were averaged across participants.

The recordings in different sensors had different signal strength; for example, the posterior sensors close to the early visual cortex had higher signal strength than other sensors (similar to what we saw in Figure 5.8 in the regression on the `AlexNet` features). To visualize the overall signal strength, we took the highest regression score (i.e., the maximum scores averaged across participants, as described above) across sensors at each time point. Figure 6.2 shows the results at each time point for MEG and EEG respectively. There seems to be a small time lag between the two modalities, possibly because the different delays in image presentation in MEG and EEG were not perfectly corrected (see Section 5.2.5). Nevertheless, the highest proportion of variance explained was about 28% in EEG and 32% in MEG. Although MEG had slightly higher proportions of variance explained at the majority of time points, the differences were small. In other words, the signal strength in EEG was slightly smaller than, but comparable with that in MEG in this experiment.

6.3.2 Representational similarity between MEG and EEG recordings

Secondly, we tested whether the responses to the 362 stimulus images recorded in MEG and in EEG at corresponding time points showed statistical dependence between each other, using the representational similarity analysis. Figure 6.3 shows the Pearson correlation between the lower triangular entries in the RSMs of the two modalities at each corresponding time point, averaged across 15 participants. We also permuted the image indices 40 times in one modality, using the same permutation sequences across participants. The transparent blue bands show the interval that covered 95% of the permuted values at each time point; these bands are plotted mainly for visualization as no correction for multiple comparisons was applied. We also used a more rigorous two-sided excursion permutation test (see Section 5.2.11) to examine whether the mean Pearson

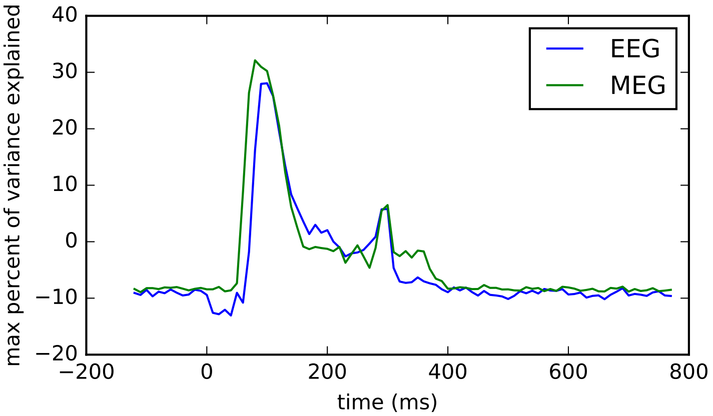


Figure 6.2: The highest proportion of variance explained across sensors in the leave-one-participant-out regression analysis at each time point.

correlations across participants were significantly greater than zero, where the multiple comparisons at all the time points were accounted for; we observed a significant time window from 60 to 450 ms after the stimulus onset (p-value < 0.025), where the MEG and EEG recordings showed significant statistical dependence (or representational similarity) between each other. Note that this window overlapped with the window identified in Chapter 5 (Figure 5.7), where the variance of the MEG and EEG recordings were significantly explained by the `AlexNet` features.

6.3.3 Sensitivity of MEG and EEG in detecting dependence between neural responses and external features

Thirdly, we compare the sensitivity of the MEG and EEG data in addressing one of our main scientific questions—whether the neural activities code information about the visual or semantic features of the stimulus images. Statistically, one can either use regression analysis (as we did in Chapter 5), or use nonparametric dependence tests (e.g., the HSIC test or the representational similarity analysis) to examine the statistical dependence between the neural activities and the features. Here we used the HSIC test because it was computationally fast to implement. The visual features we used here were derived from the features in Layer 5 of `AlexNet`; the semantic features we used here were 0/1 features based on semantic labels of the scene categories (e.g., “indoor”, “shopping and dining”, “outdoor-natural”, etc). There are other visual or semantic features, which can be of interest; however, since our main goal was to compare the sensitivity of MEG and EEG, we used these two particular feature sets as examples and expect the results to be similar with other feature sets.

We aim to compare the sensitivity of MEG and EEG in detecting positive dependence between neural activities and the visual or semantic features using the HSIC tests. Ideally, we could examine the statistical power of the tests, which is determined by the true alternative distributions and

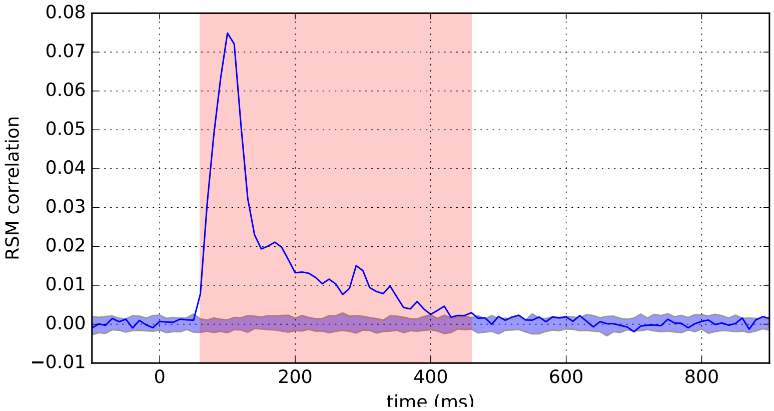


Figure 6.3: Representational similarity analysis between MEG and EEG. The figure shows the Pearson correlation of the entries in the RSMs from MEG and EEG recordings at each time point, averaged across 15 participants. The transparent blue bands show the 95% intervals from 40 permutations without correction for multiple comparisons. The transparent pink band shows the time interval where the mean Pearson correlations across participants were significantly greater than zero (p-value < 0.025).

the noise level in each modality. However, while the null distributions can be obtained by permutations, it is challenging to estimate the alternative distributions of the HSIC statistics, which are needed in power analysis. In this context, we define the *sharp ratio*—the testing statistic divided by the critical value in the null hypothesis at a level of 0.05—as an approximate measurement of the sensitivity. The sharp ratio increases as the number of observations increases; here we computed the sharp ratios for different numbers of observations (10%, 30%, 50%, 70% and 90%, or 36, 109, 181, 253 and 325 observations) at each time point in MEG and EEG separately (see Section 6.2.4 for more details).

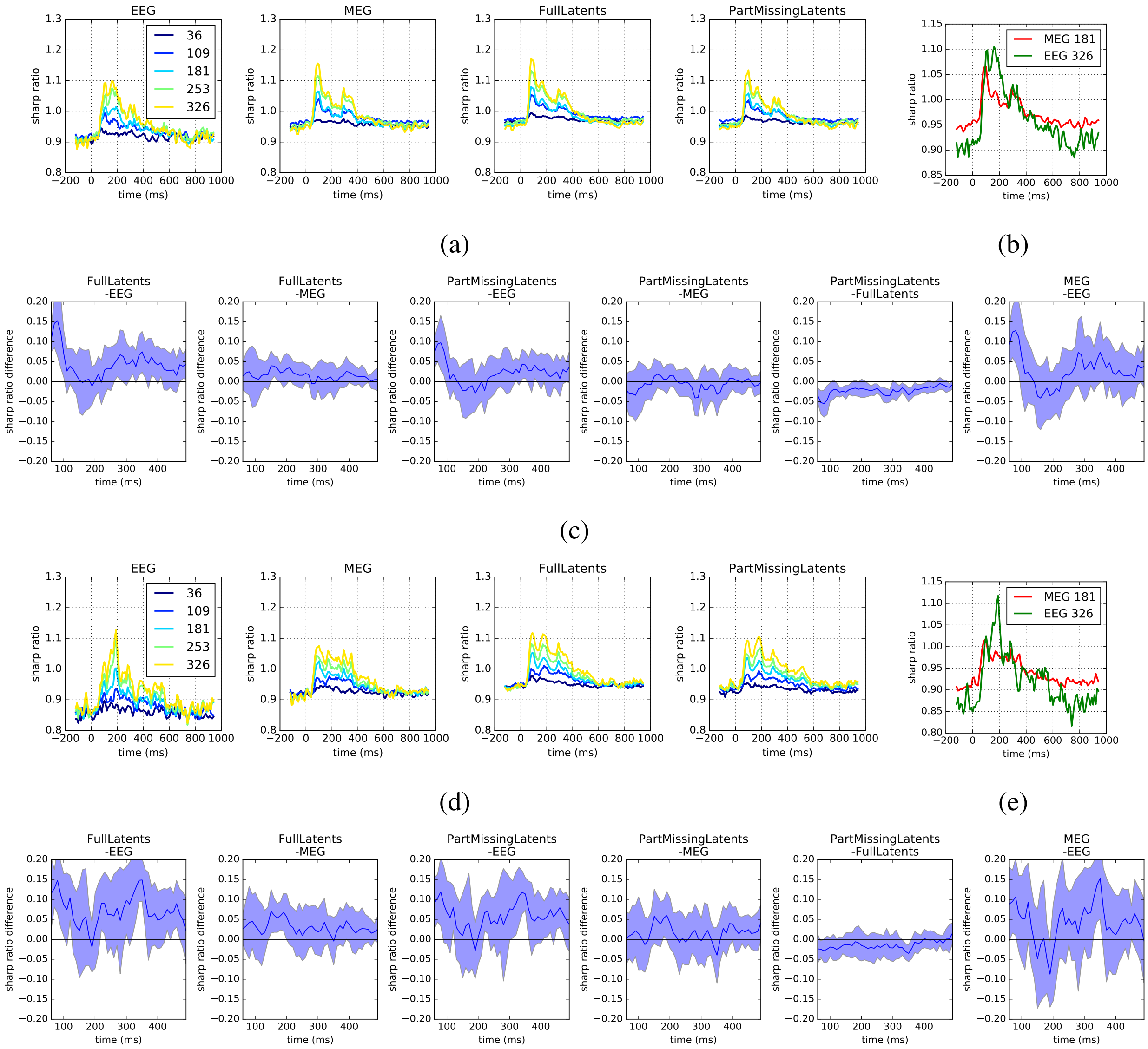
The averaged sharp ratios across participants are shown in the first two plots of Figure 6.4a and 6.4d. In Figure 6.4a, where we tested the dependence between the neural activities and the visual features, we observed that the sharp ratios were higher than the baseline window (i.e., before 0 ms) from around 80 ms to near 500 ms, in both MEG and EEG. As the number of observations increased, the sharp ratios became larger. If we visually compare the MEG and EEG plots in Figure 6.4a for the same number of observations, we see that MEG appeared to have larger sharp ratios than EEG. For more rigorous comparisons, we computed the difference of sharp ratios between MEG and EEG with 325 observations, for each participant respectively, at each time point from 80 to 500 ms, where the sharp ratios were larger than the baseline. The last plot in Figure 6.4c shows the averaged differences between MEG and EEG across participants. The standard error of the averaged difference at each time point was also computed and used to plot the $(2.5\%/43, 97.5\%/43)$ confidence intervals (blue bands), assuming the averaged difference had a normal distribution; note that the confidence intervals were based on the Bonferroni criterion in the correction for multiple comparisons at 43 time points. In this plot, the sharp ratios in MEG were significantly higher than those in EEG before 120 ms, and marginally higher at some time points near 350 ms. Nevertheless, the differences in the sharp ratios between MEG and EEG were not large. If the number observations in EEG was large enough (e.g., 325) compared with that in MEG (e.g., 181) as in Figure 6.4b, the sharp ratios in EEG were comparable with those in MEG at the majority of the time points during 80 to 500 ms (at some time points, the sharp ratios in EEG were even larger). For the test of dependence between the neural activities and the semantic features, we saw similar results (Figure 6.4d, 6.4e and 6.4f); interestingly, the sharp ratios in EEG appeared even marginally larger than those in MEG with 325 observations near 200 ms in Figure 6.4f.

Together, these results indicate that in detecting statistical dependence between neural representations of images and the visual (or semantic) features of the images, EEG appeared to have slightly smaller sensitivity than MEG, when there were the same the number of observations; however, with a larger number of observations, EEG can achieve comparable sensitivity to that in MEG.

6.3.4 "Stitching" data between MEG and EEG in a factor analysis framework

According to the results above, MEG and EEG recordings of neural activities share some representational similarity; in addition, in detecting statistical dependence between neural responses and external features (e.g., the visual or semantic features above), MEG appeared to have slightly higher sensitivity. Next, we consider a setting where we are given a relatively large stimulus set, a tight budget and access to both MEG and EEG. In this setting, we aim to study the statistical dependence between the neural responses and the external features of the stimuli. A possible option is to utilize both MEG and EEG and combine them in a sophisticated way. For example, we can present a subset of stimuli in a relative short MEG session, and present all stimuli in a longer EEG session. The total cost of the experiments can still be low because only a short MEG session is included. Afterward, we can "stitch" the two modalities by learning common latent factors from both modalities, even with some data missing in MEG; then we can run the statistical tests (e.g., HSIC tests) to examine the dependence between the latent factors and the external features. This idea, borrowed from the "neural stitching" concept in [Bishop and Yu, 2014; Bishop, 2015], may lead to higher sensitivity than using EEG alone, with a lower cost than using MEG alone.

Here we explore this possibility by simulating the setting. Consider a missing-observation case, where the MEG responses to 122 of the 362 images were not observed, while the EEG responses to all the images were observed. As a control, we also consider a full-observation case, where both the MEG and EEG responses to all the images were observed. We learned $n_{latent} = 70$ latent factors for each case and computed the sharp ratios of the HSIC tests, sub-sampling different numbers of observations from the latent factors in the same procedure as in Section 6.2.4. The results are shown in the third ("FullLatents" for the full-observation case) and fourth ("PartMissingLatents" for the missing-observation case) plots in Figure 6.4a and Figure 6.4d, for the visual features and the semantic features respectively. In Figure 6.4c and 6.4f, we plotted the pairwise differences between MEG, EEG, "FullLatents" and "PartMissingLatents", averaged across participants when



(a) (b)

(c)

(d) (e)

(f)

Figure 6.4: Sharp ratios in different modalities and with various numbers of observations, in the HSIC tests with the visual features ((a), (b) and (c)), and the semantic features ((d), (e) and (f)). In (a) and (d), besides MEG and EEG, the sharp ratios of the latent factors are also shown. “FullLatents” denote the latent factors that were learned when all the MEG and EEG data were available in the full-observation case, whereas “PartMissingLatents” denote the latent factors that were learned when some observations in MEG were missing (i.e., the missing-observation case). In (b) and (e), the plots show the sharp ratios when there were 181 observations in MEG and 325 observations in EEG. In (c) and (f), the plots show the pairwise differences of sharp ratios between MEG, EEG, “FullLatents” and “PartMissingLatents”, from 80 to 500 ms, with 325 observations. The differences were averaged across participants, and the blue bands show the confidence intervals based on the standard errors of the averaged differences across participants, corrected for multiple comparisons.

there were 325 observations. The blue bands show confidence intervals after the Bonferroni correction. Although the sharp ratios of "PartMissingLatents" were no higher than those in 'FullLatents", they appeared to be higher than those in EEG at least in some time windows, and they were also comparable with (if not always higher than) those in MEG.

6.4 Discussion

In this chapter, we empirically compared the recordings in MEG and EEG in our scene processing experiment. Although MEG appeared to give slightly better signal strength and higher sensitivity in studying the dependence between the neural responses to images and external visual or semantic features, EEG was not much worse. Moreover, in the dependence tests with external features, one can possibly increase the sensitivity of EEG by acquiring more observations, with still a low cost. For example, if we assume MEG sessions cost $\$500$ per hour and EEG sessions cost $\$50$, as in our experiments, then doubling or tripling the EEG session length (thus the number of observations) still costs less. Our results here may serve as informative resources when researchers design experiments that are related to visual processing, and considerate choosing between the MEG and EEG.

Theoretically, MEG and EEG have different sensitivity to source currents in different directions and at different depths, so they each can provide some unique information about the source currents. However, the MEG and EEG recordings have statistical dependence between each other as well, as confirmed by our results here and in Chapter 5. This is not surprising because the lead fields of the two modalities may not be fully orthogonal. More importantly, the neural activities during complex perceptual and cognitive processes can be distributed over large areas on the cortical surfaces, covering different local normal directions and depth (due to the folding of the cortical surfaces), and the distributed activities that are close on the cortical surfaces can be correlated to some degree. The EEG and MEG recordings of such neural activities can be highly dependent on each other.

Here we also explored the possibility of combining MEG and EEG recordings in the statistical analysis of interest, even in cases where the MEG data were partly missing. Our results point in a promising direction where one can assign stimuli to MEG and EEG sessions in a sophisticated way, and "stitch" the data from the two modalities to obtain reasonably good sensitivity with a small cost. The results given here are still preliminary and limited to the specific analysis. In future work, we plan to provide a mathematically formulated "stitching" strategy, taking into account the budget constraints and the sensitivity in relation to the number of stimuli (i.e., the number of observations), such that researchers as users can easily apply it in their daily practice.

It is worth noting that our results here were empirical, in the context of studying visual processing of scene images. Depending on the research questions, the advantages and weaknesses of using MEG and EEG may vary, and further exploration on other research topics is needed. Interestingly, some work on decoding hand movement using MEG and EEG showed similar decoding accuracy rates between the two modalities [Waldert et al., 2008], again indicating that EEG can be almost as good as MEG in this context. Our results are also limited in the sense that we only used the HSIC tests to examine whether there was dependence between the neural responses and the external features. However, we expect to see similar results by other types of analyses, such as predicting images features or image identities using the MEG/EEG recordings.

Another issue worth mentioning is that our analyses here were purely in the sensor space, where the results can be due to the contribution of a few sensors; our results might not reflect the advantage that MEG had more sensors or higher rank data, which is important in source localization or in detecting subtle changes in neural activities at small spatial scales. In future work, we will explore more diverse analyses (e.g., in the source space) that can be affected by the rank of data (or the number of sensors) in MEG and EEG.

Chapter 7

Discussion and future directions

This thesis has included three components: a methodological contribution about one-step source-space analysis, a scientific contribution about the spatial-temporal dynamics in the visual cortex, and an empirical comparison between MEG and EEG. Below, we summarize the major conclusions and discuss related future directions for each component. Because there are a variety of generic issues worth discussing in the first component, we allocate a larger space for it than for the other two.

In the methodological contribution, we presented novel one-step models for source-space analyses in MEG and EEG, which in our simulations outperformed the commonly used two-step methods that apply L_2-norm regularization in the source localization step (i.e., MNE). Our one-step models focused on two types of analyses, (i) regression analyses, which estimated dependence between neural activity and external covariates (e.g., stimulus features or behavioral metrics), and (ii) functional connectivity analysis, which estimated statistical dependence among brain regions of interest while participants performed experimental tasks.

In the one-step regression model (the STFT-R in Chapter 3), we represented the source neural activities as time-frequency components, which were linear combinations of external covariates (regressors). In this way, we directly related the regression coefficients in the time-frequency domain in the source space to the sensor recordings. Moreover, we applied penalties to encourage sparsity in the source space and the time-frequency domain. As a result of the time-frequency sparsity, when the regression coefficients were transformed back to the time domain, the high-frequency components that did not explain the sensor data well (i.e., noise) were attenuated. In addition, during the periods in which we would expect the regression coefficients to be zero (e.g., the baseline time window before the stimulus onset), the method typically recovered zero coefficients. The resulting waveforms were generally smooth, concentrated in the time windows of interest, and thus "clean" and easy to interpret. In contrast, it is difficult to impose such sparsity-inducing constraints on the regression coefficients if we use a two-step approach.

In the one-step connectivity models in Chapter 4, we focused on functional connectivity among pre-defined regions of interest (ROIs). In a stationary setting, the functional connectivity was described through the covariance matrix of the ROI activities; in a dynamic setting, the functional connectivity was described through the time-varying autoregressive coefficients of the ROI activities. We assumed that the activity at each source point within an ROI had an independent normal distribution conditioned on the latent ROI activity at each time point; more specifically, the activity at each source point was centered around a mean, which was the latent ROI activity (or some linearly weighted version of the latent ROI activity in Section 4.2), and the variance of each source point was shared within that ROI. Similarly, we assumed that the activity at each source point outside any ROIs followed an i.i.d normal distribution, with a zero mean and a common variance. By eliminating the source activities, we again built a direct relationship between the latent ROI activities and the sensor data, so that we could directly estimate the covariance matrix or the autoregressive coefficients from the sensor data. In simulations where the above ROI-to-source model was correct, our one-step approaches gave more accurate connectivity estimates than the two-step methods using MNE.

In sum, our results in the methodological contribution demonstrated some advantages of the one-step approach. In future work, we can exploit the one-step framework in more sophisticated analyses. For example, in the one-step framework, we can combine the regression and connectivity analyses to jointly estimate both the correlation with external covariates and the connectivity in the residuals, and we can also decode brain states using distributed neural activity in the source space. In addition, we can incorporate prior knowledge from other imaging modalities (e.g., fMRI) in the one-step framework.

Since the source localization problem is underdetermined, all models rely on some prior assumptions or constraints for tractability. It is worth noting that in our one-step approaches, additional model assumptions were imposed as well (for example, the sparsity assumption in STFT-R, the autoregressive assumption in the state-space connectivity model, and the mapping from the latent ROI activities to the activities at all source points in the connectivity models). We do not think that one-step approaches are always better than two-step approaches. If the model assumptions in the one-step approach are severely wrong, then the one-step model may yield larger errors than a two-step model. Indeed, the main advantage of the one-step framework is the flexibility of jointly constraining all steps in the analysis, instead of using limited constraints for the source localization step, which may sometimes undermine further connectivity or regression analysis.

Nevertheless, in practical uses, all model assumptions may be inaccurate to some degree. Some violations of assumptions may be crucial, yielding less accurate scientific conclusions, while others may be minor—they may reduce the power of the statistical tests, but strong effects may still be detected, especially when the control or null condition also suffers from a similar model mismatch problem. In this context, it is important to consider in what cases the model mismatch is likely to cause inaccurate conclusions and in what cases its effects on the conclusions are minor. Below, we discuss in detail the possible implications of several assumptions we considered and point out related future directions.

- *The sparsity assumption* In our STFT-R model, we assumed sparsity in different domains, including the sparsity of active source points in the spatial domain and the sparsity of neural activities and their correlation with external covariates in the time-frequency domain. The assumption of spatial sparsity has been widely used, not only in the more recent probabilistic models of source localization, but also in earlier methods such as equivalent dipole fitting. However, to some degree, the spatial sparsity is more of a convenient approximation than a justifiable assumption in the biological sense. For example, although the representation of a particular visual object or a scene can be sparse on the neuronal level in a local cortical area (e.g., within a millimeter), on a more coarse level (e.g., several to dozens of millimeters, larger than the highest spatial resolution of fMRI/MEG/EEG), object or scene processing recruits many large brain areas (see the Introduction in Section 5.1). In fact, distributed processing ("connectionism") has been one of the main theories in cognitive science [Fodor and Pylyshyn, 1988]. In this sense, the sparsity assumption in the spatial domain is inaccurate for many sophisticated cognitive processes.

  However, practically, I think violations of the assumption of spatial sparsity only have minor effects on the overall conclusions. For instance, a sparse activation pattern may approximate a denser pattern by only representing the larger activations. Since there are tuning parameters that control the level of sparsity and we can select them via cross-validation, we can empirically observe whether the activation is really spatially sparse. Even if we end up with a pattern more sparse than the truth, it is likely that the non-zero source points are among the true source points involved in the cognitive processes. With that being said, a concerning issue is that the forward matrix $\boldsymbol{G}$ may lack the properties that guarantee the consistency of source-point selection, such as the "irrepresentable condition" or the "restricted isometry property" [Zhao and Yu, 2006], which essentially requires the possible correlations between columns of the forward matrix to be small. In fact, the columns in $\boldsymbol{G}$ that correspond to spatially close source points are correlated; such correlations can make it difficult to separate source points that are spatially close. Moreover, if the true activation pattern is dense in a large area, the detected sparse source points may not be located in the peaks of the true pattern, but a few millimeters off in nearby locations. For future work, various alternative constraints can be used to remedy this issue. One possible direction is to use low-dimensional latent factors to explain the general spatial correlation, but also use sparse patterns to capture strong activations.

  The sparsity in the time-frequency domain appears more justifiable in that the dynamic neural responses to stimuli are often concentrated at certain time windows and certain frequency bands. While the sparsity assumption may be violated, similar arguments to those given above can be applied in this domain as well. Hence the model mismatch problem is likely to have a minor effect on the scientific conclusions in practice.

- *Temporal independence of the sensor noise and source-space noise* Throughout this thesis, we have assumed that the sensor noise $\boldsymbol{e}_t$ in Equation (2.2) is temporally i.i.d. and the covariance matrix can be pre-computed from a baseline time window or from empty-room recordings. In the connectivity analysis, we assumed that within each ROI, the activity in each source point was determined by the latent ROI activity plus a noise term. This source-space noise was also assumed i.i.d. for different source points and across time. These assumptions, especially the assumption of temporal independence can be violated in real data.

In our STFT-R model, the temporal dependence may result in some frequency components across all time windows. However, if the sensor noise is stationary, this issue only has a minor effect on the results for two reasons. First, the phases of the frequency components that are due to the temporal dependence in the sensor noise are not likely to align with the stimulus onset, and these components are not likely to be correlated with the regressors either. Thus they are likely to cancel out. Secondly, in the statistical tests, we either compare the regression coefficients in time windows after the stimulus onset with those in the baseline time window (i.e., before the stimulus onset), or compare the regression coefficients with permuted versions in the permutation tests. In these cases, the temporal dependence of noise also affects the regression coefficients in the baseline window and the permutations, and thus we should still be able to make correct conclusions from the statistical tests. In the time-lagged connectivity analysis, the temporal dependence may have a larger effect (for example, in over-estimating the autoregressive coefficients). Further empirical simulations are needed to evaluate the effect, but again, the effect can be attenuated if we test the difference between the estimates after the stimulus onset and those in the baseline window.

One important future direction is to model the spatio-temporal covariance structure of the sensor noise and the source-space noise. Some previous work proposed to use a Kronecker structure [Bijma et al., 2005], upon which more sophisticated models can be built. Another issue is that the covariance matrix of the sensor noise is pre-computed and fixed. The error in this pre-estimation is not accounted for in source localization or in one-step models. However, even in the typical source localization step (without doing regression or connectivity analysis), simultaneously estimating the source activity of interest, the source-space noise and the sensor noise is challenging, in terms of identifiability and tractability. Some intuitive assumptions about the properties of the noise as well as a Bayesian formulation incorporating the pre-computed noise covariance may help in this case.

- *Short-range temporal dependence in connectivity analysis* In the one-step state-space model of dynamic connectivity in Chapter 4, we only assumed an autoregressive model of order 1. This assumption was mainly chosen for tractability, but longer-term dependence may exist in real data. In a simple case of two ROIs, if the true autoregressive model has a higher order but the influence has only one direction (i.e., only ROI 1 predicts later activity in ROI 2, not vice versa), then we should still be able to infer the influence in the correct direction. However, further analyses are needed for more complicated cases. In future work, we can use a higher-order autoregressive model and impose further regularization for tractability.

- *pre-defined ROIs in connectivity analyses* In our one-step connectivity models in Chapter 4, we only considered a pre-defined set of ROIs and assumed that source points outside any ROIs had zero mean. The connectivity results may depend crucially on the correctness of this assumption. If we missed an important ROI, then we would have shrunken its activity to zero and excluded it from the network. Because of the correlations among columns in the forward matrix $\boldsymbol{G}$, this might inaccurately assign the activity in this ROI to other ROIs, resulting in inaccurate estimates of the connectivity. A remedy is to partition the entire source space into many ROIs and model the connectivity among all of them, with further regularization. Although this will increase the challenge of model fitting, it is an important direction to explore in future work.

Besides what we mentioned above, another very important future direction is theoretical analyses of the source localization problem, such as the fundamental limits or minimax rates based on the biophysical properties of the forward matrix $\boldsymbol{G}$. Advances in this line will provide valuable insights in building both two-step and one-step models.

Finally, the computation of our one-step models was time-consuming, not only because the optimization problems were harder than in the two-step methods, but also because we needed to do model selection and bootstrapping (permutations) of the entire analysis pipeline. Therefore, in future work, we need to scale and speed up the one-step models, especially by exploiting parallel computing.

In our scientific contribution (Chapter 5), we applied source-space regression and connectivity analyses (including both the two-step methods and our novel one-step models) to study the spatio-temporal neural dynamics in human vision. Using MEG/EEG, we recorded neural responses to images of naturalistic scenes. Additionally, we extracted low-level and high-level features of the same images from a pre-trained feedforward convolutional neural network (`AlexNet`). By regressing the source activities on features at different levels, we were able to describe what level of features were coded at different temporal stages and spatial locations in the visual cortex. In this analysis, we observed a strong early-to-late, lower-to-higher-level pattern, which is consistent with feedforward information flow along the hierarchy of the ventral visual cortex. Moreover, we compared the neural correlation with two groups of low-level features, one describing the components that were highly correlated with high-level features (i.e., “object-category-relevant” low-level components), and the other describing the components that were roughly orthogonal to high-level features. We found that in a time window where the neural responses were likely corresponding to the disappearance of the images, the early visual cortex showed a higher and longer correlation with the “object-category-relevant” low-level components than with the the low-level components that were roughly orthogonal to high-level features. This difference suggests there may be non-feedforward influences (e.g., top-down feedback) that help the early visual cortex to separate those two groups of low-level features. The results by the two-step method and our one-step STFT-R method were consistent, and the STFT-R again yielded "cleaner" time courses of regression effects, due to the sparsity in the time-frequency domain. We also used time-varying connectivity analysis to quantify time-lagged dependence between ROIs along the hierarchy; our results, from both the two-step method and our novel one-step state-space model, suggested both leading and lagged dependence between early visual cortex and higher-level scene/object-selective regions. Although such statistical dependence does not necessarily mean causal interactions, the results can help us build specific hypotheses about the timing of feedforward and feedback information flow for future experimental tests. In sum, these explorations, which have not been done in previous literature to our knowledge, provide very intriguing initial results that may help us better understand the vision mechanisms in the brain. In future work, we can explicitly incorporate feedback connections in computational models of the visual cortex and use similar experiments to validate and improve these models.

In our empirical contribution (Chapter 6), using data from the experiments above, we also compared the sensitivity of MEG and EEG in detecting dependence between neural responses and stimulus features in the sensor space. Although MEG showed slightly higher sensitivity, the less costly EEG was able to achieve comparable sensitivity with MEG when the number of observations was doubled. Moreover, our preliminary work showed some advantages of "stitching" the two modalities together. We learned a factor analysis model to combine the EEG responses to a full set of stimuli and the MEG responses to a subset of these stimuli; the "stitched" latent factors demonstrated higher sensitivity compared to the EEG responses alone. These empirical results are helpful for researchers with limited budgets to make the best usage of the two modalities. In future work, we plan to make similar comparisons and apply the "stitching" idea on other experimental analyses for other cognitive tasks.

Finally, another interesting future direction in MEG and EEG is active experimental design. For example, in the context of learning dependence between neural activities and high-dimensional features of stimuli (e.g., visual features of images), researchers often use randomly sampled stimuli, similar to what we did in Chapter 5. However, such random design may require a large number of stimuli, resulting in high acquisition costs especially for MEG. In contrast, one can define an objective metric of the dependence, and exploit an active learning paradigm that maximizes the objective metric by selecting the next stimulus in the feature space based on previous observations [Settles, 2010]. Moreover, in cases where the feature space is too complex, a similar paradigm can be implemented where the candidate stimuli are organized on a graph that describes the between-stimuli similarities [Ma et al., 2013]. In either way, we may be able to present fewer stimuli than in a random design but obtain scientific results that are similarly reliable. We have started some initial work along this line and I will continue further exploration after graduation.

Bibliography

Ahlfors, S. P. and Simpson, G. V. (2004). Geometrical interpretation of fmri-guided MEG/EEG inverse estimates. *NeuroImage*, 22(1):323–332.

Aminoff, E. M. and Tarr, M. J. (2015). Associative processing is inherent in scene perception. *PloS one*, 10(6):e0128840.

Aminoff, E. M., Toneva, M., Shrivastava, A., Chen, X., Misra, I., Gupta, A., and Tarr, M. J. (2015). Applying artificial vision models to human scene understanding. *Frontiers in computational neuroscience*, 9.

Babadi, B., Obregon-Henao, G., Lamus, C., Hamalainen, M. S., Brown, E. N., and Purdon, P. L. (2014). A subspace pursuit-based iterative greedy hierarchical solution to the neuromagnetic inverse problem. *NeuroImage*, 87(0):427 – 443.

Bach, F., Jenatton, R., Mairal, J., and Obozinski, G. (2011). Optimization with sparsity-inducing penalties. *CoRR*, abs/1108.0775.

Baillet, S. and Garnero, L. (1997). A bayesian approach to introducing anatomo-functional priors in the EEG/MEG inverse problem. *Biomedical Engineering, IEEE Transactions on*, 44(5):374–385.

Bar, M. and Aminoff, E. (2003). Cortical analysis of visual context. *Neuron*, 38(2):347–358.

Beck, A. and Teboulle, M. (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, 57:289–300.

Bijma, F., De Munck, J. C., and Heethaar, R. M. (2005). The spatiotemporal meg covariance matrix modeled as a sum of kronecker products. *NeuroImage*, 27(2):402–415.

Bishop, W. E. (2015). Combining neural population recordings: Theory and application. *PhD thesis, Carnegie Mellon University*.

Bishop, W. E. and Yu, B. M. (2014). Deterministic symmetric positive semidefinite matrix completion. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., and Weinberger, K., editors, *Advances in Neural Information Processing Systems 27*, pages 2762–2770. Curran Associates, Inc.

Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press, New York, NY, USA.

Buzsáki, G., Anastassiou, C. A., and Koch, C. (2012). The origin of extracellular fields and currentseeg, ecog, lfp and spikes. *Nature reviews neuroscience*, 13(6):407–420.

Carhart-Harris, R. L., Leech, R., Hellyer, P. J., Shanahan, M., Feilding, A., Tagliazucchi, E., Chialvo, D. R., and Nutt, D. (2015). The entropic brain: a theory of conscious states informed by neuroimaging research with psychedelic drugs. *Psychoanalytical neuroscience: Exploring psychoanalytic concepts with neuroscientific methods*, page 140.

Chen, X., Shrivastava, A., and Gupta, A. (2013). Neil: Extracting visual knowledge from web data. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1409–1416.

Cichy, R. M., Khosla, A., Pantazis, D., and Oliva, A. (2016a). Dynamics of scene representations in the human brain revealed by magnetoencephalography and deep neural networks. *NeuroImage*.

Cichy, R. M., Khosla, A., Pantazis, D., Torralba, A., and Oliva, A. (2016b). Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific reports*, 6.

Cichy, R. M., Khosla, A., Pantazis, D., Torralba, A., and Oliva, A. (2016c). Deep neural networks predict hierarchical spatio-temporal cortical dynamics of human visual object recognition. *arXiv preprint arXiv:1601.02970*.

Cichy, R. M., Pantazis, D., and Oliva, A. (2016d). Similarity-based fusion of meg and fmri reveals spatio-temporal dynamics in human cortex during visual object recognition. *Cerebral Cortex*, page bhw135.

Clarke, A., Devereux, B. J., Randall, B., and Tyler, L. K. (2014). Predicting the time course of individual objects with meg. *Cerebral Cortex*, page bhu203.

Cribben, I., Haraldsdottir, R., Atlas, L. Y., Wager, T. D., and Lindquist, M. A. (2012). Dynamic connectivity regression: determining state-related changes in brain connectivity. *Neuroimage*, 61(4):907–920.

Dalal, S. S., Guggisberg, A. G., Edwards, E., Sekihara, K., Findlay, A. M., Canolty, R. T., Berger, M. S., Knight, R. T., Barbaro, N. M., Kirsch, H. E., et al. (2008). Five-dimensional neuroimaging: localization of the time–frequency dynamics of cortical activity. *Neuroimage*, 40(4):1686–1700.

Dale, A. M., Liu, A. K., Fischl, B. R., Buckner, R. L., Belliveau, J. W., Lewine, J. D., and Halgren, E. (2000). Dynamic statistical parametric mapping: combining fmri and meg for high-resolution imaging of cortical activity. *Neuron*, 26(1):55–67.

Darlington, R. B. and Hayes, A. F. (2000). Combining independent p values: Extensions of the stouffer and binomial methods. *Psychological Methods*, 5(4):496.

David, O., Kiebel, S. J., Harrison, L. M., Mattout, J., Kilner, J. M., and Friston, K. J. (2006). Dynamic causal modeling of evoked responses in EEG and MEG. *NeuroImage*, 30(4):1255–

1272.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE.

DiCarlo, J. J. and Cox, D. D. (2007). Untangling invariant object recognition. *Trends in cognitive sciences*, 11(8):333–341.

Engel, S. A., Glover, G. H., and Wandell, B. A. (1997). Retinotopic organization in human visual cortex and the spatial precision of functional mri. *Cerebral cortex*, 7(2):181–192.

Epstein, R., Harris, A., Stanley, D., and Kanwisher, N. (1999). The parahippocampal place area: Recognition, navigation, or encoding? *Neuron*, 23(1):115–125.

Epstein, R. A. (2008). Parahippocampal and retrosplenial contributions to human spatial navigation. *Trends in cognitive sciences*, 12(10):388–396.

Felleman, D. J. and Van Essen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cerebral cortex*, 1(1):1–47.

Fischl, B., Salat, D. H., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., Van Der Kouwe, A., Killiany, R., Kennedy, D., Klaveness, S., et al. (2002). Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. *Neuron*, 33(3):341–355.

Fisher, R. A. (1925). *Statistical methods for research workers*. Genesis Publishing Pvt Ltd.

Fodor, J. A. and Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28(1-2):3–71.

Freeman, J., Ziemba, C. M., Heeger, D. J., Simoncelli, E. P., and Movshon, J. A. (2013). A functional and perceptual signature of the second visual area in primates. *Nature neuroscience*, 16(7):974–981.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). A note on the group lasso and a sparse group lasso. *arXiv preprint arXiv:1001.0736*.

Friston, K. J. (2011). Functional and effective connectivity: a review. *Brain connectivity*, 1(1):13–36.

Fukushima, M., Yamashita, O., Knösche, T. R., and Sato, M.-a. (2015). Meg source reconstruction based on identification of directed source interactions on whole-brain anatomical networks. *NeuroImage*, 105:408–427.

Galka, A., Ozaki, O. Y. T., Biscay, R., and Valdes-Sosa, P. (2004). A solution to the dynamical inverse problem of eeg generation using spatiotemporal kalman filtering. *NeuroImage*, 23:435–453.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680.

Gow, D. W., Segawa, J. A., Ahlfors, S. P., and Lin, F.-H. (2008). Lexical influences on speech perception: a granger causality analysis of meg and eeg source estimates. *Neuroimage*, 43(3):614–623.

Gramfort, A., Kowalski, M., and Hamaleinen, M. (2012). Mixed-norm estimates for the m/eeg inverse problem using accelerated gradient methods. *Physics in Medicine and Biology*, 57:1937–1961.

Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., Parkkonen, L., and Hmlinen, M. S. (2014). Mne software for processing meg and eeg data. *NeuroImage*, 86(0):446 – 460.

Gramfort, A., Strohmeier, D., Haueisen, J., Hamalainen, M., and Kowalski, M. (2013). Time-frequency mixed-norm estimates: Sparse m/eeg imaging with non-stationary source activations. *NeuroImage*, 70(0):410 – 422.

Granger, C. W. (1988). Some recent development in a concept of causality. *Journal of econometrics*, 39(1):199–211.

Gretton, A., Fukumizu, K., Teo, C. H., Song, L., Schölkopf, B., and Smola, A. J. (2007). A kernel statistical test of independence. In *Advances in neural information processing systems*, pages 585–592.

Grill-Spector, K., Kushnir, T., Edelman, S., Avidan, G., Itzchak, Y., and Malach, R. (1999). Differential processing of objects under various viewing conditions in the human lateral occipital complex. *Neuron*, 24(1):187–203.

Hamalainen, M., Hari, R., Ilmoniemi, R. J., Knuutila, J., and Lounasmaa, O. V. (1993). Magnetoencephalography–theory, instrumentation, to noninvasive studies of the working human brain. *Reviews of Modern Physics*, 65:414–487.

Hamalainen, M. and Ilmoniemi, R. (1994). Interpreting magnetic fields of the brain: minimum norm estimates. *Med. Biol. Eng. Comput.*, 32:35–42.

Henson, R. N., Wakeman, D. G., Litvak, V., and Friston, K. J. (2011). A parametric empirical bayesian framework for the eeg/meg inverse problem: generative models for multi-subject and multi-modal integration. *Frontiers in human neuroscience*, 5.

Hubel, D. H. and Wiesel, T. N. (1968). Receptive fields and functional architecture of monkey striate cortex. *The Journal of physiology*, 195(1):215–243.

Huh, M., Agrawal, P., and Efros, A. A. (2016). What makes imagenet good for transfer learning? *arXiv preprint arXiv:1608.08614*.

Ishai, A. (2008). Lets face it: Its a cortical network. *NeuroImage*, 40(2):415 – 419.

Ito, M. and Komatsu, H. (2004). Representation of angles embedded within contour stimuli in area v2 of macaque monkeys. *The Journal of neuroscience*, 24(13):3313–3324.

Jenatton, R., Mairal, J., Obozinski, G., and Bach, F. (2011). Proximal methods for hierarchical space coding. *J. Mach. Learn. Res*, 12:2297–2334.

Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., and

Darrell, T. (2014). Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 675–678. ACM.

Khaligh-Razavi, S.-M. and Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain it cortical representation. *PLoS Comput Biol*, 10(11):e1003915.

Klamer, S., Elshahabi, A., Lerche, H., Braun, C., Erb, M., Scheffler, K., and Focke, N. K. (2015). Differences between meg and high-density eeg source localizations using a distributed source model in comparison to fmri. *Brain Topography*, 28(1):87–94.

Kriegeskorte, N., Mur, M., and Bandettini, P. A. (2008). Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 2:4.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105.

Kveraga, K., Ghuman, A. S., Kassam, K. S., Aminoff, E. A., Hämäläinen, M. S., Chaumon, M., and Bar, M. (2011). Early onset of neural synchronization in the contextual associations network. *Proceedings of the National Academy of Sciences*, 108(8):3389–3394.

Lachaux, J.-P., Rodriguez, E., Martinerie, J., Varela, F. J., et al. (1999). Measuring phase synchrony in brain signals. *Human brain mapping*, 8(4):194–208.

Lamus, C., Hamalainen, M. S., Temereanca, S., Brown, E. N., and Purdon, P. L. (2012). A spatiotemporal dynamic distributed solution to the meg inverse problem. *NeuroImage*, 63:894–909.

Liang, Z., Shen, W., Sun, C., and Shou, T. (2008). Comparative study on the offset responses of simple cells and complex cells in the primary visual cortex of the cat. *Neuroscience*, 156(2):365–373.

Liu, A. K., Dale, A. M., and Belliveau, J. W. (2002). Monte carlo simulation studies of eeg and meg localization accuracy. *Human brain mapping*, 16(1):47–62.

Liu, Z. and He, B. (2008). fmri–eeg integrated cortical source imaging by use of time-variant spatial constraints. *NeuroImage*, 39(3):1198–1214.

Long, C. J., Purdon, P. L., Temereanca, S., Desai, N. U., Hämäläinen, M. S., and Brown, E. N. (2011). State-space solutions to the dynamic magnetoencephalography inverse problem using high performance computing. *The annals of applied statistics*, 5(2B):1207.

Lopes da Silva, F. H., Wieringa, H. J., and Peters, M. J. (1991). Source localization of EEG versus MEG: Empirical comparison using visually evoked responses and theoretical considerations. *Brain Topography*, 4(2):133–142.

Ma, Y., Garnett, R., and Schneider, J. (2013). σ-optimality for active learning on gaussian random fields. In *Advances in Neural Information Processing Systems*, pages 2751–2759.

Malmivuo, J. (2012). Comparison of the properties of eeg and meg in detecting the electric activity of the brain. *Brain topography*, 25(1):1–19.

Maris, E. and Oostenveld, R. (2007). Nonparametric statistical testing of EEG-and MEG-data.

Journal of neuroscience methods, 164(1):177–190.

Matsuura, K. and Okabe, Y. (1995). Selective minimum-norm solution of the biomagnetic inverse problem. *IEEE Transactions on Biomedical Engineering*, 42(6):608–615.

Mattout, J., Phillips, C., Penny, W. D., Rugg, M. D., and Friston, K. J. (2006). Meg source localization under multiple constraints: an extended bayesian framework. *NeuroImage*, 30(3):753–767.

Mishkin, M., Ungerleider, L. G., and Macko, K. A. (1983). Object vision and spatial vision: two cortical pathways. *Trends in neurosciences*, 6:414–417.

Mosher, J. C., Leahy, R. M., and Lewis, P. S. (1999). EEG and MEG: forward solutions for inverse methods. *IEEE Transactions on Biomedical Engineering*, 46(3):245–259.

Mosher, J. C., Spencer, M. E., Leahy, R. M., and Lewis, P. S. (1993). Error bounds for eeg and meg dipole source localization. *Electroencephalography and clinical Neurophysiology*, 86(5):303–321.

Nestor, A., Plaut, D. C., and Behrmann, M. (2011). Unraveling the distributed neural code of facial identity through spatiotemporal pattern analysis. *Proceedings of the National Academy of Sciences*, 108(24):9998–10003.

Nestor, A., Vettel, J. M., and Tarr, M. J. (2008). Task-specific codes for face recognition: how they shape the neural representation of features for detection and individuation. *PloS one*, 3(12):e3978.

Ou, W., Nummenmaa, A., Ahveninen, J., Belliveau, J. W., Hämäläinen, M. S., and Golland, P. (2010). Multimodal functional imaging using fmri-informed regional EEG/MEG source estimation. *Neuroimage*, 52(1):97–108.

Park, S., Konkle, T., and Oliva, A. (2015). Parametric coding of the size and clutter of natural scenes in the human brain. *Cerebral Cortex*, 25(7):1792–1805.

Pascual-Marqui, R. (2002). Standardized low resolution brain electromagnetic tomography (sloreta): technical details. *Methods Find. Exp. Clin. Pharmacol.*, 24:5–12.

Pascual-Marqui, R. D., Michel, C. M., and Lehmann, D. (1994). Low resolution electromagnetic tomography: a new method for localizing electrical activity in the brain. *International Journal of psychophysiology*, 18(1):49–65.

Petersen, K. B. and Pedersen, M. S. (2008). The matrix cookbook. *Technical University of Denmark*.

Pyles, J. A., Verstynen, T. D., Schneider, W., and Tarr, M. J. (2013). Explicating the face perception network with white matter connectivity. *PloS one*, 8(4):e61611.

Quraan, M. A., Moses, S. N., Hung, Y., Mills, T., and Taylor, M. J. (2011). Detection and localization of hippocampal activity using beamformers with meg: a detailed investigation using simulations and empirical data. *Human brain mapping*, 32(5):812–827.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. (2015). Imagenet large scale visual recognition challenge.

International Journal of Computer Vision, 115(3):211–252.

Scherg, M. (1990). Fundamentals of dipole source potential analysis. *Auditory evoked magnetic fields and electric potentials. Advances in audiology*, 6:40–69.

Ségonne, F., Dale, A., Busa, E., Glessner, M., Salat, D., Hahn, H., and Fischl, B. (2004). A hybrid approach to the skull stripping problem in mri. *Neuroimage*, 22(3):1060–1075.

Settles, B. (2010). Active learning literature survey. *University of Wisconsin, Madison*, 52(55-66):11.

Shumway, R. H. and Stoffer, D. S. (1982). An approach to time series smoothing and forecasting using the em algorithm. *Journal of time series analysis*, 3(4):253–264.

Stegle, O., Lippert, C., Mooij, J. M., Lawrence, N. D., and Borgwardt, K. M. (2011). Efficient inference in matrix-variate gaussian models with\ iid observation noise. In *Advances in neural information processing systems*, pages 630–638.

Stine, R. A. (1985). Bootstrp prediction intervals for regression. *Journal of the American Statistical Association*, 80:1026–1031.

Tanaka, K. (1996). Inferotemporal cortex and object vision. *Annual review of neuroscience*, 19(1):109–139.

Tian, L., Jiang, T., Wang, Y., Zang, Y., He, Y., Liang, M., Sui, M., Cao, Q., Hu, S., Peng, M., and Zhuo, Y. (2006). Altered resting-state functional connectivity patterns of anterior cingulate cortex in adolescents with attention deficit hyperactivity disorder. *Neuroscience letters*, 400(1):39–43.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288.

Van Veen, B. D., Van Drongelen, W., Yuchtman, M., and Suzuki, A. (1997). Localization of brain electrical activity via linearly constrained minimum variance spatial filtering. *Biomedical Engineering, IEEE Transactions on*, 44(9):867–880.

Waldert, S., Preissl, H., Demandt, E., Braun, C., Birbaumer, N., Aertsen, A., and Mehring, C. (2008). Hand movement direction decoded from meg and eeg. *The Journal of neuroscience*, 28(4):1000–1008.

Wandell, B. A., Brewer, A. A., and Dougherty, R. F. (2005). Visual field map clusters in human cortex. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360(1456):693–707.

Wasserman, L. (2010). *All of Statistics: A Concise Course in Statistical Inference*. Springer Publishing Company, Incorporated.

Wipf, D. and Nagarajan, S. (2009). A unified bayesian framework for MEG/EEG source imaging. *Neuroimage*, 44(3):947–966.

Wipf, D. P. and Nagarajan, S. S. (2008). A new view of automatic relevance determination. In *Advances in neural information processing systems*, pages 1625–1632.

Xiao, J., Hays, J., Ehinger, K. A., Oliva, A., and Torralba, A. (2010). Sun database: Large-scale scene recognition from abbey to zoo. In *Computer vision and pattern recognition (CVPR), 2010 IEEE conference on*, pages 3485–3492. IEEE.

Xu, Y. (2013). Cortical spatiotemporal plasticity in visual category learning. *PhD thesis, Cargenie Mellon University*.

Xu, Y., Sudre, G. P., Wang, W., Weber, D. J., and Kass, R. E. (2011). Characterizing global statistical significance of spatiotemporal hot spots in magnetoencephalography/electroencephalography source space via excursion algorithms. *Statistics in medicine*, 30(23):2854–2866.

Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., and DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23):8619–8624.

Yang, Y., Tarr, M. J., and Kass, R. E. (2014). Estimating learning effects: A short-time fourier transform regression model for MEG source localization". In *Lecture Notes on Artificial Intelligence: MLINI 2014: Machine learning and interpretation in neuroimaging*, Montreal, Canada. Springer.

Yosinski, J., Clune, J., Bengio, Y., and Lipson, H. (2014). How transferable are features in deep neural networks? In *Advances in neural information processing systems*, pages 3320–3328.

Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. *The Journal of Machine Learning Research*, 7:2541–2563.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., and Torralba, A. (2014a). Object detectors emerge in deep scene cnns. *arXiv preprint arXiv:1412.6856*.

Zhou, B., Lapedriza, A., Xiao, J., Torralba, A., and Oliva, A. (2014b). Learning deep features for scene recognition using places database. In *Advances in neural information processing systems*, pages 487–495.