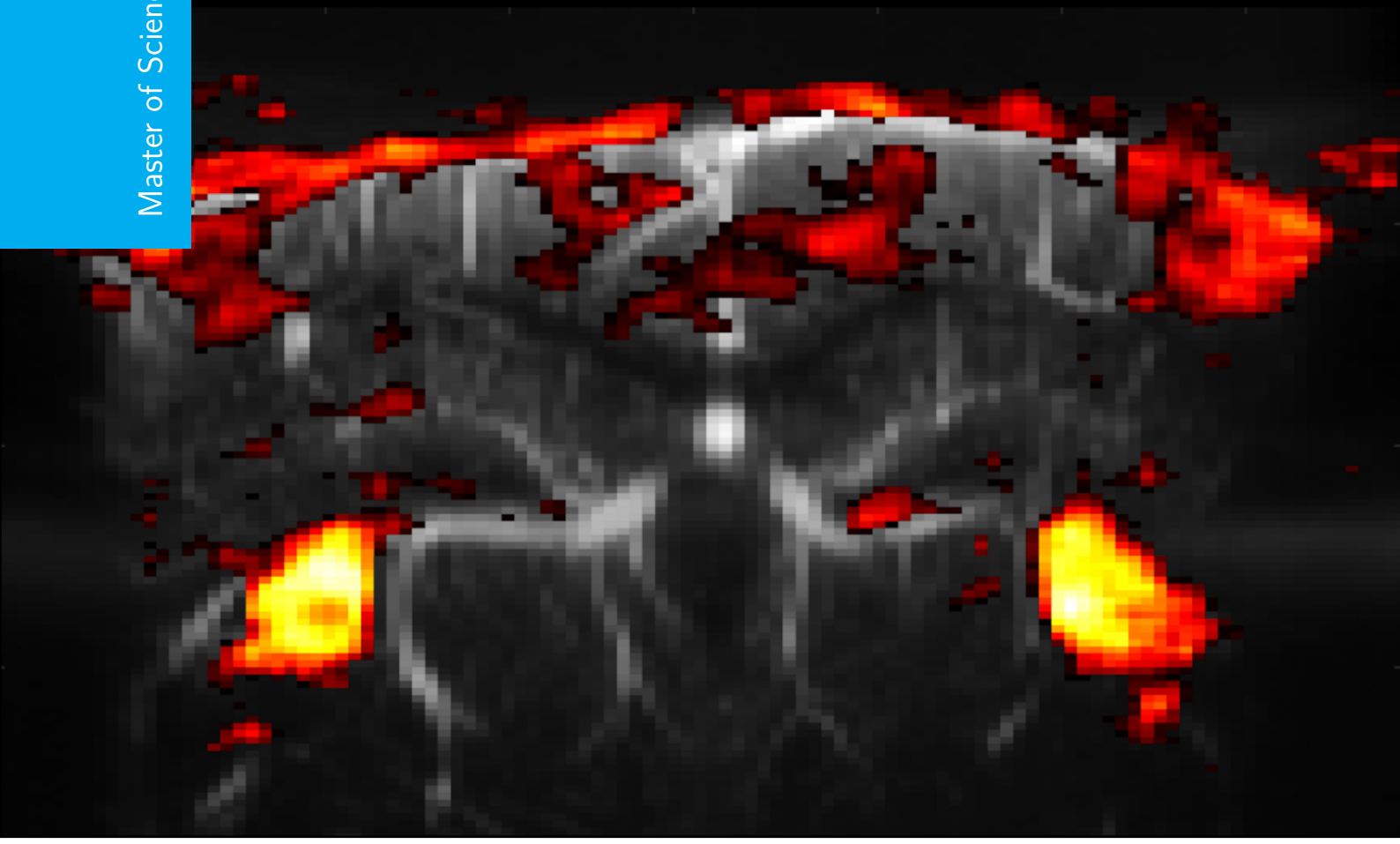
Joint Estimation of Neural and Hemodynamic Responses in Functional Ultrasound Using Variational Inference

T. Fujioka

Master of Science Thesis



TU Delft Delft University of Technology

Delft Center for Systems and Control

Joint Estimation of Neural and Hemodynamic Responses in Functional Ultrasound Using Variational Inference

Master of Science Thesis

For the degree of Master of Science in Systems and Control at Delft University of Technology

T. Fujioka

July 4, 2025

Faculty of Mechanical Engineering (ME) · Delft University of Technology

Abstract

Functional Ultrasound (fUS) is an emerging neuroimaging technique capable of capturing brain activity similarly to functional magnetic resonance imaging (fMRI) but with higher spatiotemporal resolution and lower operational cost. This thesis investigates the extension of the joint detection-estimation (JDE) framework to jointly estimate both the hemodynamic response function (HRF) and neural response function (NRF) from fUS imaging data. In the proposed model, the two response functions are represented as a cascade linear time-invariant (LTI) convolution systems, enabling indirect estimation of neural activity signals from fUS measurements. Since direct recordings of neural activity are often unavailable, this Bayesian approach offers a data-driven means of probing the brain's functional organization.

Inference is performed within a coordinate ascent variational inference (CAVI) framework. The proposed algorithm was applied to fUS datasets and validated against simultaneous recordings of neural firing rates. Results demonstrate that the model successfully captures neural activity, achieving a Pearson correlation coefficient (PCC) of approximately 0.24-0.30, and provides a modest improvement over the conventional boxcar input stimulus model. Additionally, the jointly estimated HRFs were consistent with existing literature, and regional HRF estimates revealed differences in response dynamics between the visual cortex and hippocampus, highlighting region-specific hemodynamic properties.

Table of Contents

Chapter 1

Introduction

Understanding the neural activity underlying functional brain imaging signals is a fundamental step toward decoding how the brain processes information and how different regions interact. Hemodynamic imaging techniques such as functional magnetic resonance imaging (fMRI) and, more recently, functional Ultrasound (fUS), provide indirect measurements of brain activity by capturing local changes in deoxygenated hemoglobin level or blood volume. While these signals contain valuable information, the true neural activity and the mechanisms linking it to hemodynamic responses remain hidden in the data.

Compared to fMRI, fUS offers higher spatio-temporal resolution with a significantly smaller form factor of its measurement equipment, making it a promising tool for studying fast and localized neural processes in small animals. One prior study [15] proposed a blind deconvolution approach to jointly estimate neural sources and hemodynamic responses from fUS recordings using a constrained tensor decomposition framework. However, to date, no existing work has applied a fully probabilistic model to jointly infer both neural dynamics and hemodynamic responses in fUS. In this thesis, we address this gap by designing a Bayesian model within the variational inference framework to uncover these hidden variables in fUS data.

In this chapter, we introduce key background concepts from the neuroscience field that are essential for the formulation of the thesis goals. We then provide a brief overview of the current functional neuroimaging data analysis methods. Finally, we outline the the existing gaps in literature and the goals of this thesis.

1-1 Background

This section covers preliminary knowledge from the neuroscience domain, which includes a basis of the dynamics we are trying to identify, as well as the working principle of fUS.

1-1-1 Neurovascular Coupling

Both fMRI and fUS capture the neural activity in the brain indirectly via the neurovascular coupling. The neurovascular coupling is a phenomenon derived from a complex chain of metabolic processes which connects the dynamic changes in the blood and neural activity. Despite extensive research efforts [42, 21, 22], the exact link between hemodynamics and neural activity remains unclear. However, prevailing hypotheses suggest the following cascading process; first, neural activity is triggered in response to a stimulus, which generates energetic demand to restore the neurotransmitter systems involved in the activation [46]. Subsequently, the energetic demand is met by the metabolic process involving the delivery of glucose, lactate and oxygen in neighboring arterioles and capillaries [13]. This in turn, causes a local increase of blood flow, an oversupply of oxygen, and red blood cells (RBCs) velocity [8, 28]. fMRI measures changes in blood volume and deoxygenated hemoglobin level, while fUS measures changes in blood volume.

1-1-2 functional Ultrasound (fUS)

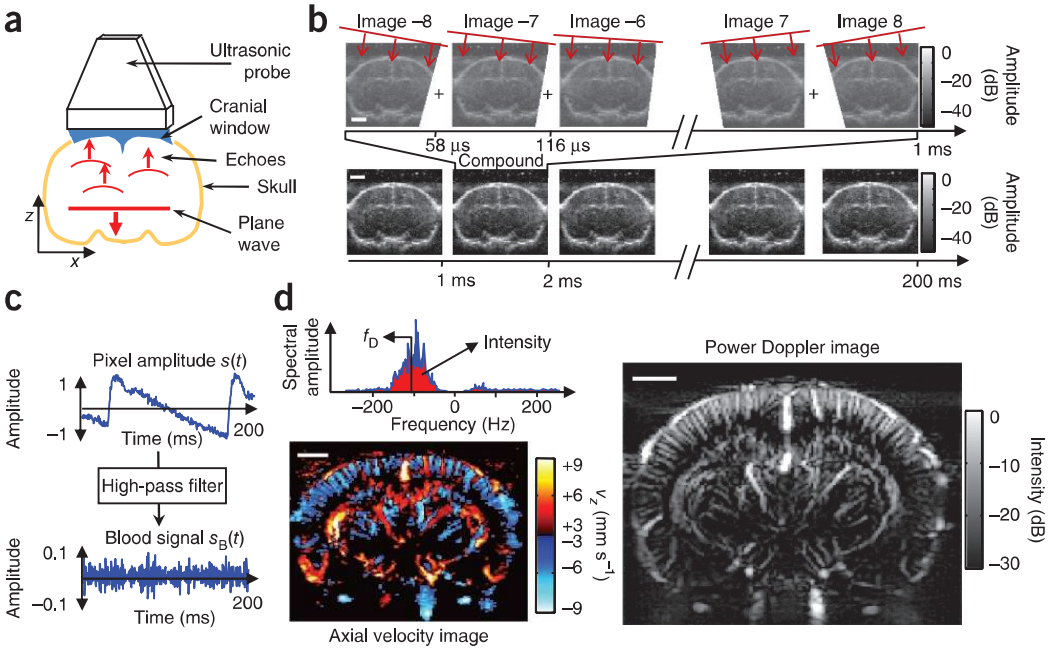


Figure 1-1: Schematic of fUS and its measured data [29]

fUS measures local changes in blood volume through the Doppler shift that ultrasonic pulses undergo. As illustrated in Figure 1-1a, an ultrasonic probe mounted on the cranial window emits plane waves at ultrasonic frequency. As the ultrasonic waves propagate through the brain, some of their energy is backscattered by red blood cells (RBCs) and slow moving tissues. The backscattered echoes are measured by the ultrasonic probe. The frequency of the backscattered echoes is shifted depending on the direction and velocities of the RBCs and the slow moving tissues due to the Doppler effect [30]. For example, given an object

moving towards the ultrasonic probe at velocity v, the central frequency f_D in the echo that is backscattered from the object is

$$f_D = -\frac{2vf_0}{c}, \tag{1-1}$$

where c is the speed of sound in the medium and f_0 is the central frequency of the ultrasonic pulses emitted from the probe.

Since the Doppler frequency shift does not occur in the echoes backscattered from the RBCs that are moving in the direction orthogonal to the propagation of the ultrasonic plane wave, the ultrasonic probe emits the plane pulse wave and measures the echoes at different angles (see Figure 1-1b) to detect moving RBCs in lateral direction. This is also to improve image quality. The measurements at different angles are summed through coherent compounding [36, 35], resulting in a beam-formed image known as a *compound image*.

The output of the ultrasonic probe is pixel amplitude $s(h, w, t_i)$ of the compound image at pixel (h, w) and time t_i where $i = 1, 2, \dots N$ indexes the successive emission of the ultrasonic pulses. The pixel amplitude fluctuates due to the RBCs and tissue motions caused by cardiovascular pulses and respiration. Because the fluctuation in the pixel amplitude from the tissue motions does not reflect the functional activity of the brain, it is considered an (unwanted) clutter signal. As the tissue motions are slower than the RBCs, the echoes from the tissue contribute to a low frequency signal in the pixel amplitude, which can be rejected by a high-pass filter [29] (see Figure 1-1 c). Besides the high pass filtering, the recent study in [12] proposed the clutter (tissue motion signal) rejection filter using the singular value decomposition (SVD), which became the most prevalent pre-processing method in practice.

After the clutter rejection, the remaining signal is called the *blood signal* and contains signals from the RBCs and the noise from electronic components in the probe. By computing mean signal intensity of the blood signal as

$$I(h, w) = \frac{1}{N}\sum_{i=1}^{N} |s(h, w, t_i)|^2, \tag{1-2}$$

a power Doppler image (PDI) (see Figure 1-1d (right)) can be obtained, which is an intensity measure for the volume of moving blood.

The acquisition of the PDI image is repeated for a given number of windows of length N. For the case in Figure 1-1, 200 compound images acquired at sampling rate of 1kHz forms one PDI, resulting in sampling rate of 5Hz for the PDI.

1-2 Existing Functional Neuroimaging Data Analysis

In functional neuroimaging data analysis, the objective can be distinguished in three different aspects: functional connectivity, resting-state activity and event-evoked activity. Functional connectivity analysis aims to describe the interactions between activities of different brain regions in resting state or during execution of experimental tasks and stimuli. In resting-state activity analysis, spontaneous brain activity without tasks is studied. In connection to the functional connectivity, one of the objectives of this analysis is to capture and understand intrinsic organization such as the default mode network [41], which is found to be altered in pathological conditions [44].

In contrast to resting-state activity analysis, task-evoked activity reflects brain responses to stimuli or tasks, such as sensory stimuli and motor control or cognitive tasks. Investigating how different brain regions are activated under these conditions could help researchers identify biomarkers that enable cross-subject and cross-condition comparisons of the brain.

In this thesis, we focus on the event-evoked activity analysis of the brain.

The existing methods in literature for the event-evoked activity follow two separate approaches: hypothesis-driven and data-driven approaches.

The hypothesis-driven approach begins with constructing a model to incorporate prior knowledge about the used experimental stimuli or tasks. Typically, a linear time-invariant (LTI) is used for a hypothesis-driven approach, which often models the measured neuroimaging data as a convolution between neural activity and hemodynamic response function (HRF), representing the neurovascular coupling. The detailed description on the LTI model is provided in Section 2-1. For fMRI domain, the implementation of such LTI model can range from constrained blind deconvolution to Bayesian estimation [10, 43, 52, 7].

The data-driven approach, on the other hand, does not involve explicit models like the hypothesis-driven approach. Instead, the data-driven approach employs decomposition of the observed data with structural constraints, such as principle component analysis (PCA), independent component analysis (ICA), and tensor decompositions [50, 6, 9] to separate source activities of the brain for given experimental condition.

In this thesis, the event-evoked activity analysis is tackled in the hypothesis-driven approach.

1-2-1 fUS Data Analysis in Literature

A recent study by [38] demonstrated a strong correlation between fUS signals and local neuronal firing rates. The same study also showed that a temporal smoothing function mapping firing rates to fUS signals closely resembled the HRFs reported in the literature. Similarly, another independent study [2] reported a significant link between fUS signals and local neuronal activity, showing that fUS signals can be predicted from measured calcium (Ca^{2+}) concentrations. These findings suggest that the fUS signal has information of neural activity but hidden in the data.

1-2-2 Gaps in Current Data Analysis in fUS

To date, the challenge of hemodynamic estimation in event-evoked analysis for fMRI has been addressed using approaches such as the data-driven methods such as ICA, PCA, and tensor-based decompositions or the hypothesis-driven method in constrained optimization framework and probabilistic framework. Despite the recent emergence of fUS technique, methods like ICA and tensor decompositions in both hypothesis- and data-driven approaches have also been applied to fUS [48, 23, 14]. However, the Bayesian approach like [7, 1] remains largely unexplored. This is important because Bayesian methods enable flexible integration of physiological priors, enabling more informed and interpretable inference.

Another limitation is that the majority of a current hypothesis-driven method uses a binary stimulus time series as a surrogate for the neural activity signal. While this assumption

simplifies the computational complexity and identifiability of the brain's underlying hemodynamics, it imposes two key drawbacks: (a) it restricts the flexibility of neural activation by assuming it to be binary, and (b) it alters the original interpretation of the neurovascular coupling by replacing the neural input to the hemodynamic system with the external stimulus signal.

1-3 Thesis Goals and Outline

The primary goal of this thesis is to develop methods for jointly estimating the hemodynamic response and underlying neural activity from fUS data, aiming to provide neuroscientists with an improved tool for further understanding of mysterious organization and functionality of the brain.

As mentioned in Section 1-2-1, recent studies have demonstrated that fUS signals strongly correlate with local neuronal activity, such as firing rates and calcium concentrations, suggesting that information about neural responses is hidden in the fUS data. However, current hypothesis-driven methods in fUS primarily rely on models that use a fixed stimulus time series as the input to the hemodynamic system, limiting the flexibility in capturing dynamics of neural activity. Moreover, Bayesian modeling approaches, widely recognized for their ability to incorporate uncertainty, remain unexplored in fUS analysis.

Our aim is to extract this hidden information, specifically, neural response function (NRF), which describes how neural activity responds to external stimuli, and hemodynamic response function (HRF), which characterizes the vascular response to the neural activity. We model these as a cascade linear time-invariant (LTI) convolution system.

Thus, the general research goals and the existing gap in the literature lead to the following research question:

Can we jointly estimate hemodynamic response function and neural response function in a probabilistic framework using fUS data?

To address this question, we first design a cascade convolution model in which the HRF and NRF are represented as two finite impulse response (FIR) filters connected in series. We then cast this model into a Bayesian framework, where prior physiological knowledge about the brain is incorporated in the form of prior distributions. Subsequently, we derive a computational algorithm for the extended model using a Variational expectation-maximization (VEM) method within the coordinate ascent variational inference (CAVI) framework. Finally, we validate the proposed model through numerical simulations, followed by its application to real fUS data to estimate the hemodynamic and neural responses.

This work begins with an introduction to the basic hemodynamic LTI system, and variational inference, a powerful class of probabilistic methods, in Chapter 2, including a review of the joint detection-estimation (JDE) framework previously applied to fMRI. Thereafter, Chapter 3 presents our proposed model, which extends the existing JDE approach by simultaneously estimation both hemodynamic and neural responses from fUS data. The results of our method, applied to both synthetic and real experimental data, are presented in Chapters 4

and 5, respectively. Finally, Chapter 6 concludes the thesis and discusses potential directions for future work.

Chapter 2

Theoretical Background

This section summarizes the existing literature related to the system identification of the hemodynamic model in functional brain imaging data. First we describe a commonly used hemodynamic model in LTI. Then, we present the probabilistic framework, called variational inference, which serves as a basis of algorithms presented in this thesis. At the end of this section, we present an algorithm developed in the past, namely the joint detection-estimation (JDE) [7, 32], which will be extended in our proposed algorithm.

2-1 Hemodynamic LTI Model

In analysis of hemodynamics of the brain, a commonly used model is LTI convolution model:

$$y(t) = h(t) * x(t) + \epsilon(t). \tag{2-1}$$

Here, $y(t)$ is observed time-series of the blood signal for a single pixel/voxel, $\epsilon(t)$ is measurement noise, and the operator $*$ represents convolution. The neurovascular coupling comprised of complex metabolic process is abstracted into an impulse response function $h(t)$ that maps neural activity $x(t)$ in the brain to the blood signal $\boldsymbol{y}(t)$. This impulse function is commonly referred to as the HRF.

In practice, the continuous-time model in Equation (2-1) is discretized, as the observed signals are sampled at fixed intervals. The discrete version of the model is commonly written in matrix form as:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{h} + \boldsymbol{\epsilon}, \tag{2-2}$$

where $\mathbf{y}$ is the observed signal vector, and $\boldsymbol{\epsilon}$ is the measurement noise vector. $\mathbf{h}$ is the finite impulse response of HRF, and $\mathbf{X}$ is a Toeplitz matrix constructed from neural activity $x(t)$. Together, $\boldsymbol{X}\boldsymbol{h}$ represents a discrete convolution of neural activity signal and hemodynamic response function, and thus the change in blood volume caused by the neural activity.

The HRF in the LTI model is parametrized differently for fMRI and fUS data. For fMRI, the HRF is commonly modeled using two gamma distribution functions [26]:

$$h(t) = A\left(\frac{t^{\alpha_1-1}\beta_1^{\alpha_1}e^{-\beta_1 t}}{\Gamma(\alpha_1)} - c\frac{t^{\alpha_2-1}\beta_2^{\alpha_2}e^{-\beta_2 t}}{\Gamma(\alpha_2)}\right) \tag{2-3}$$

where $\Gamma(\cdot)$ denotes the gamma function. Here, the HRF is parametrized by six parameters: A controls the overall amplitude of the HRF, while α_1 and β_1 control the timing and dispersion of the main response. Similarly, α_2 and β_2 controls undershooting behavior commonly observed in fMRI data, and c adjusts the peak of the undershoot.

For fUS, a single gamma distribution function

$$h(t) = A\frac{t^{\alpha_1-1}\beta_1^{\alpha_1}e^{-\beta_1 t}}{\Gamma(\alpha_1)} \tag{2-4}$$

was introduced recently in [2] and adopted by [14]. The negative gamma distribution function as in Equation 2-3 is excluded for fUS since fUS data lacks the post-stimulus undershoot observed in fMRI [2].

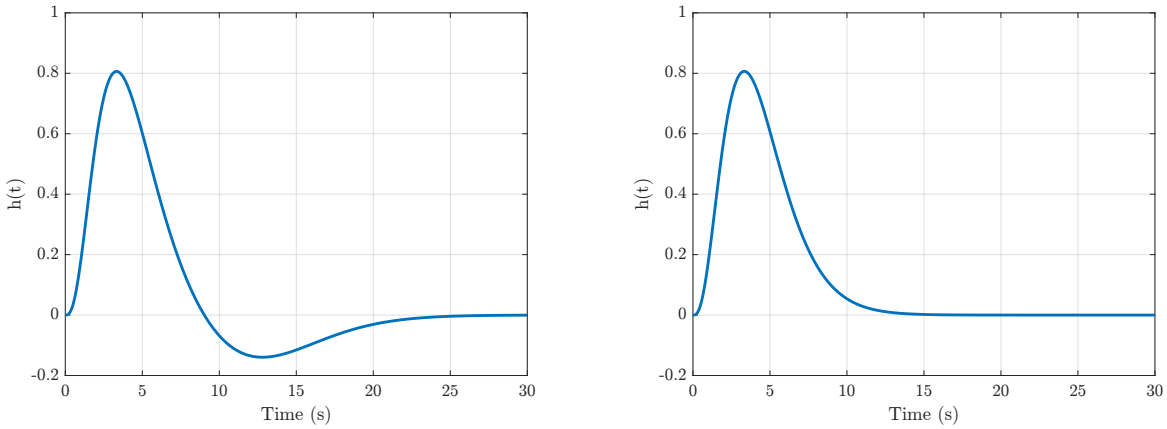


Figure 2-1: (a): Example of commonly used HRF in fMRI. Post-stimulus peak is followed by post-peak undershoot. (b) Example of commonly used HRF in fUS. The HRF has no post-stimulus undershoot.

In practice, the neural activity signal $x(t)$ is often unavailable. Therefore, the timings of external stimulus or experimental paradigm are used for $x(t)$ as a surrogate of the neural activity. Common choices for the surrogate signal $x(t)$ are a binary boxcar signal or a series of Dirac-delta signals [20, 18]. The boxcar signal indicates the onset and duration of the external stimulus or experimental events, while the Dirac-delta signal is used to indicate their onset timings (see Figure 2-2).

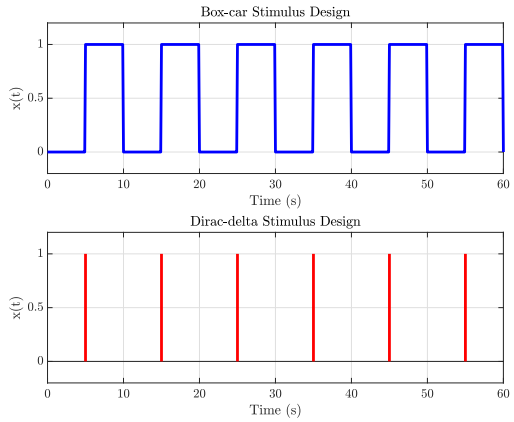


Figure 2-2: Examples of commonly used surrogate signals $x(t)$.

2-2 Variational Inference as an Optimization

Variation inference is a Bayesian inference method to estimate latent (unobserved) variables hidden in observed data. While this goal is the same as any other estimation methods, variational inference distinguishes itself by approximating the posterior distribution of the latent variables, rather than producing a single point estimate.

In the context of this thesis, the variational inference serves as the foundational framework for the estimation procedures in our proposed JDE model, introduced later in Section 3. Therefore, we begin by presenting the core concepts of variational inference.

The general problem of variational inference starts with a set of observed variables $\boldsymbol{y} = \{y_n\}_{n=1,2,...N}$ and a set of latent variables $\boldsymbol{x} = \{x_m\}_{m=1,2,...M}$, including unknown model parameters. The goal of variational inference is to estimate the posterior distributions $p(\boldsymbol{x}|\boldsymbol{y})$ of the latent variables derived from Bayes rule,

$$p(\boldsymbol{x}|\boldsymbol{y}) = \frac{p(\boldsymbol{y},\boldsymbol{x})}{p(\boldsymbol{y})} = \frac{p(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x})}{\int p(\boldsymbol{y},\boldsymbol{x})d\boldsymbol{x}}, \tag{2-5}$$

where prior distribution $p(\boldsymbol{x})$ encodes our prior beliefs or knowledge about latent variables $\boldsymbol{x}$, and the likelihood $p(\boldsymbol{y}|\boldsymbol{x})$ reflects the statistical fitness of our parametrized model to the observed data. In the denominator, $p(\boldsymbol{y})$ is known as evidence or marginal likelihood. As $p(\boldsymbol{y})$ is obtained by integrating the joint probability $p(\boldsymbol{y}|\boldsymbol{x})$ over all possible values of $\boldsymbol{x}$, it represents the preference of the observed data to our model.

Although the ultimate goal in Bayesian inference frameworks is to find the posterior $p(\boldsymbol{x}|\boldsymbol{y})$, the evidence $p(\boldsymbol{y})$ in the denominator is often intractable for complex Bayesian models due to the lack of analytical solution for the integral.

Variational inference circumvents this intractability issue by approximating the true posterior with a variational posterior $q(\boldsymbol{x}|\boldsymbol{y};\boldsymbol{\nu})$ with variational parameter $\boldsymbol{\nu}$ [3]. Here, the variational parameter $\boldsymbol{\nu}$ is a parameter that characterizes the variational posterior distribution. For instance, if we choose to approximate the true posterior $p(\boldsymbol{x}|\boldsymbol{y})$ by a multivariate Gaussian distribution, the variational parameter is $\boldsymbol{\nu} = [\boldsymbol{\mu}, \boldsymbol{\Sigma}]$. The approximation with the variational

posterior aims to minimize Kullback-Leiber (KL) divergence between $q(\boldsymbol{x}|\boldsymbol{y};\boldsymbol{\nu})$ and $p(\boldsymbol{x}|\boldsymbol{y})$, which is generally thought of as the measure of "distance" between the two distributions despite non-symmetry of the KL divergence. To avoid cluttered symbols, we omit conditional notation of $\boldsymbol{y}$ in the variational posterior throughout this report.

The (backward-) KL divergence is defined as

$$\mathrm{KL}\left(q(\boldsymbol{x};\boldsymbol{\nu})||p(\boldsymbol{x}|\boldsymbol{y})\right) = \int q(\boldsymbol{x};\boldsymbol{\nu}) \ln \frac{q(\boldsymbol{x};\boldsymbol{\nu})}{p(\boldsymbol{x} \mid \boldsymbol{y})} d\boldsymbol{y} \geq 0. \tag{2-6}$$

Since the equality holds when $q(\boldsymbol{x};\boldsymbol{\nu}) = p(\boldsymbol{x}|\boldsymbol{y})$, minimizing the KL divergence is thought of finding the variational posterior $q(\boldsymbol{x};\boldsymbol{\nu})$ as close to the true posterior $p(\boldsymbol{x}|\boldsymbol{y})$ as possible. However, the KL divergence cannot be minimized directly since we do not know the true posterior $p(\boldsymbol{x}|\boldsymbol{y})$. Instead, variational inference maximizes the evidence lower bound (ELBO) $\mathcal{L}(q(\boldsymbol{x};\boldsymbol{\nu}))$, defined as

$$\begin{aligned}
\mathcal{L}(q(\boldsymbol{x};\boldsymbol{\nu})) &= \int q(\boldsymbol{x};\boldsymbol{\nu}) \ln \frac{p(\boldsymbol{x},\boldsymbol{y})}{q(\boldsymbol{x};\boldsymbol{\nu})} d\boldsymbol{x} && \text{(2-7)}\\
&= \int q(\boldsymbol{x};\boldsymbol{\nu}) \ln p(\boldsymbol{x},\boldsymbol{y}) d\boldsymbol{x} - \int q(\boldsymbol{x};\boldsymbol{\nu}) \ln q(\boldsymbol{x};\boldsymbol{\nu}) d\boldsymbol{x} && \text{(2-8)}\\
&= \mathbb{E}_{x\sim q(\boldsymbol{x};\boldsymbol{\nu})}[\ln p(\boldsymbol{x},\boldsymbol{y})] - \mathbb{E}_{x\sim q(\boldsymbol{x};\boldsymbol{\nu})}[\ln q(\boldsymbol{x};\boldsymbol{\nu})] && \text{(2-9)}
\end{aligned}$$

Further decomposition of $\mathcal{L}(q(\boldsymbol{x};\boldsymbol{\nu}))$ yields the relation

$$\begin{aligned}
\mathcal{L}(q(\boldsymbol{x};\boldsymbol{\nu})) &= \mathbb{E}_{x\sim q(\boldsymbol{x};\boldsymbol{\nu})}[\ln p(\boldsymbol{x},\boldsymbol{y})] - \mathbb{E}_{x\sim q(\boldsymbol{x};\boldsymbol{\nu})}[\ln q(\boldsymbol{x};\boldsymbol{\nu})] && \text{(2-10)}\\
&= \mathbb{E}_{x\sim q(\boldsymbol{x};\boldsymbol{\nu})}\left[\ln p(\boldsymbol{x}|\boldsymbol{y}) + \ln p(\boldsymbol{y})\right] - \mathbb{E}_{x\sim q(\boldsymbol{x};\boldsymbol{\nu})}[\ln q(\boldsymbol{x};\boldsymbol{\nu})] && \text{(2-11)}\\
&= \mathbb{E}_{x\sim q(\boldsymbol{x};\boldsymbol{\nu})}[\ln p(\boldsymbol{x}|\boldsymbol{y})] + \ln p(\boldsymbol{y}) - \mathbb{E}_{x\sim q(\boldsymbol{x};\boldsymbol{\nu})}[\ln q(\boldsymbol{x};\boldsymbol{\nu})] && \text{(2-12)}\\
&= \mathbb{E}_{x\sim q(\boldsymbol{x};\boldsymbol{\nu})}\left[\ln \frac{p(\boldsymbol{x}|\boldsymbol{y})}{q(\boldsymbol{x};\boldsymbol{\nu})}\right] + \ln p(\boldsymbol{y}) && \text{(2-13)}\\
&= -\mathrm{KL}\left(q(\boldsymbol{x};\boldsymbol{\nu})||p(\boldsymbol{x}|\boldsymbol{y})\right) + \ln p(\boldsymbol{y}). && \text{(2-14)}
\end{aligned}$$

Equation 2-14 leads to the relation

$$\ln p(\boldsymbol{y}) - \mathcal{L}(q(\boldsymbol{x};\boldsymbol{\nu})) = \mathrm{KL}\left(q(\boldsymbol{x};\boldsymbol{\nu})||p(\boldsymbol{x}|\boldsymbol{x})\right) \geq 0. \tag{2-15}$$

Since the evidence $\ln p(\boldsymbol{y})$ is constant with respect to the variational function $q(\boldsymbol{x};\boldsymbol{\nu})$, the minimization of the KL divergence with respect to $q(\boldsymbol{x};\boldsymbol{\nu})$ is equivalent to maximization of $\mathcal{L}(q(\boldsymbol{x};\boldsymbol{\nu}))$ with constraint $\ln p(\boldsymbol{y}) \geq \mathcal{L}(q(\boldsymbol{x};\boldsymbol{\nu}))$, in which log of the evidence is bounded below by $\mathcal{L}(q(\boldsymbol{x};\boldsymbol{\nu}))$ and hence the name ELBO.

The optimization problem in variational inference is thus formulated as

$$\begin{aligned}
&\min_{q(\boldsymbol{x};\boldsymbol{\nu})} \mathrm{KL}\left(q(\boldsymbol{x};\boldsymbol{\nu})||p(\boldsymbol{x}|\boldsymbol{y})\right) && \text{(2-16)}\\
&= \max_{q(\boldsymbol{x};\boldsymbol{\nu})} \mathcal{L}(q(\boldsymbol{x};\boldsymbol{\nu})) && \text{(2-17)}\\
&\text{s.t. } \ln p(\boldsymbol{y}) \geq \mathcal{L}(q(\boldsymbol{x};\boldsymbol{\nu})). && \text{(2-18)}
\end{aligned}$$

2-3 Coordinate Ascent Variational Inference

In this section, we present one of the most commonly used variational inference optimization algorithm, called coordinate ascent variational inference (CAVI) [3]. In CAVI, an additional constraint is imposed to the optimization problem in Equation (2-17), which fully factorizes the variational posterior. Then, CAVI iteratively solves the optimization problem for each factor of $q(\boldsymbol{x}|\boldsymbol{y};\boldsymbol{\nu})$. As the ELBO in the optimization problem is maximized iteratively, the style of the optimization can be analogous to that of the alternating minimization algorithm.

The full factorization of $q(\boldsymbol{x}|\boldsymbol{y};\boldsymbol{\nu})$, commonly known as mean-field approximation, is

$$q(\boldsymbol{x};\boldsymbol{\nu}) = \prod_{m=1}^{M} q(\boldsymbol{x}_m;\boldsymbol{\nu}_m), \tag{2-19}$$

where the variational posterior is factorized into separate factors for each latent variable $\boldsymbol{x}_m$. As the factorization eliminates the dependencies between $\boldsymbol{x}$, it allows the CAVI to separately solve the optimization problem in Equation 2-17 for each variational parameter $\boldsymbol{\nu}_m$. On one hand, this reduces the complexity of the optimization as the optimizer $\boldsymbol{\nu}$ is reduced into the lower dimensional vector of $\boldsymbol{\nu}_m$. On the other hand, the factorized variational posterior fails to capture the correlations between the latent variables in the underlying true posterior. This leads to the under representation of the true posterior, especially, the second-order statistics. The visualization of this under representation is shown in Figure 2-3, where a two-dimensional Gaussian distribution as a true posterior is approximated by product of two univariate Gaussian distributions, i.e., $p(x_1, x_2|\boldsymbol{y}) \approx q(x_1;\boldsymbol{\nu}_1)q(x_2;\boldsymbol{\nu}_2)$ with $\boldsymbol{\nu}_m = [\mu_m, \sigma_m^2]$. Thus, the mean field approximation is simplifies the optimization problem at the expense of reduced accuracy in the posterior estimation.

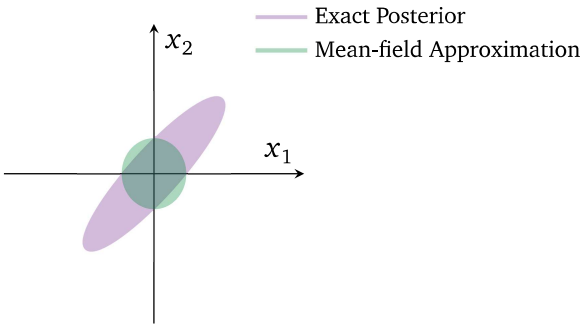


Figure 2-3: Visualization of approximating a two-dimensional Gaussian posterior by mean field approximation [4]

Applying the mean-field approximation from Equation 2-19 to the ELBO in Equation 2-17, the optimal update solution for $q(x_m;\nu_m)$ is given by [3]

$$q(x_m;\nu_m^*) \propto \mathbb{E}_{i\neq m}[\ln p(\boldsymbol{x},\boldsymbol{y})], \tag{2-20}$$

where the operator $\mathbb{E}_{i\neq m}[\cdot]$ denotes taking the expectation with respect to all factors $q(x_i;\nu_i)$ except $q(x_m;\nu_m)$.

The derivation of Equation 2-20 from [3] is as follows. By applying the mean field approximation in Equation 2-19, the ELBO in Equation 2-9 becomes

$$\mathcal{L}(q(\boldsymbol{x};\boldsymbol{\nu})) = \int \prod_{m=1}^{M} q(\boldsymbol{x}_m;\boldsymbol{\nu}_m) \ln \frac{p(\boldsymbol{x},\boldsymbol{y})}{\prod_{m=1}^{M} q(\boldsymbol{x}_m;\boldsymbol{\nu}_m)} d\boldsymbol{x}_m \tag{2-21}$$

$$= \int \prod_{m=1}^{M} q(\boldsymbol{x}_m;\boldsymbol{\nu}_m) \left\{ \ln p(\boldsymbol{x},\boldsymbol{y}) - \sum_{m=1}^{M} \ln q(\boldsymbol{x}_m;\boldsymbol{\nu}_m) \right\} d\boldsymbol{x}_m \tag{2-22}$$

$$= \int q(\boldsymbol{x}_m;\boldsymbol{\nu}_m) \left\{ \ln p(\boldsymbol{x},\boldsymbol{y}) \prod_{i\neq m} q(\boldsymbol{x}_i;\boldsymbol{\nu}_i) d\boldsymbol{x}_i \right\} d\boldsymbol{x}_m - \int \sum_{m=1}^{M} \ln q(\boldsymbol{x}_m;\boldsymbol{\nu}_m) \prod_{m=1}^{M} q(\boldsymbol{x}_m;\boldsymbol{\nu}_m) d\boldsymbol{x}_m \tag{2-23}$$

$$= \int q(\boldsymbol{x}_m;\boldsymbol{\nu}_m) \ln \tilde{p}(\boldsymbol{x}_m,\boldsymbol{y}) d\boldsymbol{x}_m - \int q(\boldsymbol{x}_m;\boldsymbol{\nu}_m) \ln q(\boldsymbol{x}_m;\boldsymbol{\nu}_m) d\boldsymbol{x}_m + \text{const} \tag{2-24}$$

where the substitution in the integrand of Equation 2-24 is defined as

$$\ln \tilde{p}(\boldsymbol{x}_m,\boldsymbol{y}) = \int \ln p(\boldsymbol{x},\boldsymbol{y}) \prod_{i\neq m} q(\boldsymbol{x}_i;\boldsymbol{\nu}_i) d\boldsymbol{x}_i = \mathbb{E}_{i\neq m}[\ln p(\boldsymbol{x},\boldsymbol{y})] + \text{const.} \tag{2-25}$$

In Equation 2-22, the fraction in the logarithm is separated into the numerator and the denominator terms. In Equation 2-23, the integral is separated for each term and the order of the first integral is rearranged. Note that the rearrangement of the order is valid due to the independence between the variational posteriors. From Equation 2-23 to Equation 2-24, all terms that are not related to the latent variable $\boldsymbol{x}_m$ and its variational parameter $\boldsymbol{\nu_m}$ are absorbed into the constant term. By inspecting Equation 2-24, one can notice that the first two terms form the negative KL divergence of the variational posterior $q(\boldsymbol{x}_m,\boldsymbol{\nu}_m)$ from the expected log-likelihood $\ln \tilde{p}(\boldsymbol{x}_m,\boldsymbol{y})$. Therefore, the maximization of the ELBO with respect to $q(\boldsymbol{x}_m;\boldsymbol{\nu}_m)$ is

$$\max_{q(\boldsymbol{x}_m,\boldsymbol{\nu}_m)} \mathcal{L}(q(\boldsymbol{x}_m;\boldsymbol{\nu}_m)) = \max_{q(\boldsymbol{x}_m,\boldsymbol{\nu}_m)} -\text{KL}(q(\boldsymbol{x}_j;\nu_j)||\tilde{p}(\boldsymbol{x}_j,\boldsymbol{y})) \tag{2-26}$$

$$= \min_{q(\boldsymbol{x}_m,\boldsymbol{\nu}_m)} \text{KL}(q(\boldsymbol{x}_j;\nu_j)||\tilde{p}(\boldsymbol{x}_j,\boldsymbol{y})). \tag{2-27}$$

Since the KL divergence is minimized when $q(\boldsymbol{x}_j;\nu_j) = \tilde{p}(\boldsymbol{x}_j,\boldsymbol{y})$, the optimal solution for $q(\boldsymbol{x}_j;\nu_j)$ is

$$\ln q_m(\boldsymbol{x}_m;\boldsymbol{\nu}_m^*) = \mathbb{E}_{i\neq m}[\ln p(\boldsymbol{x},\boldsymbol{y})] + \text{const}, \tag{2-28}$$

where the constant term accounts for the normalization constant of the variational posterior. Equivalently, this can be also expressed as

$$\ln q_j(\boldsymbol{x}_m;\boldsymbol{\nu}_m^*) = \mathbb{E}_{i\neq m}[\ln p(\boldsymbol{x}_m|\boldsymbol{y},\boldsymbol{x}_{i\neq m})] + \text{const}, \tag{2-29}$$

where $p(\boldsymbol{x}_m|\boldsymbol{y},\boldsymbol{x}_{i\neq m})$ is the probability of $\boldsymbol{x}_m$ conditioned on the observed data and all other latent variables, which is also known as complete conditional.

So far, we have not made specific assumptions on the form of the variational posterior $q(\boldsymbol{x}_m;\boldsymbol{\nu}_m)$ for which we are optimizing. As can be seen in the optimal solution of CAVI in Equation 2-20 or 2-28, we have a solution up to multiplicative (or additive constant in the log form) and hence we still need to compute the normalization constant.

In CAVI, the computation of the normalization constant for the continuous distribution is avoided by choosing specific form of prior distributions for given likelihood functions. Specifically, prior distributions for given likelihood functions are chosen such that the resulting posterior distributions will be in the same form as the priors. These prior distributions are called conjugate priors and will be described in the following section.

2-4 Conjugate prior

In Bayesian statistics, a likelihood function can be combined with a prior distribution over the parameter of the likelihood, as seen in the numerator of Equation 2-5. When the resulting posterior distribution is in the same functional form as the prior distribution, the prior is referred to as a conjugate prior [3]. The conjugacy of the prior distribution to the likelihood function is an important concept that simplifies the computation of the posterior distribution. In the simplest case, consider the linear observation model:

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b} \tag{2-30}$$

where $\boldsymbol{y}$ is a vector of observed data, $\boldsymbol{b}$ is a zero-mean Gaussian noise vector ($\boldsymbol{b} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$), $\boldsymbol{x}$ is an unobserved vector distributed as $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Gamma})$, and matrix $\boldsymbol{A}$ is assumed to be known. In this model, the likelihood is

$$p(\boldsymbol{y} \mid \boldsymbol{x}) = \mathcal{N}(\boldsymbol{A}\boldsymbol{x}, \boldsymbol{\Sigma}) \tag{2-31}$$

with unknown mean $\boldsymbol{A}\boldsymbol{x}$ and the prior is

$$p(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Gamma}). \tag{2-32}$$

Because the Gaussian distribution is a conjugate prior to the likelihood with an unknown mean, we know that the posterior distribution is in the form of Gaussian distribution with unknown mean and covariance. This can be shown as follows:

$$\begin{aligned}
p(\boldsymbol{x} \mid \boldsymbol{y}) &= \frac{p(\boldsymbol{y} \mid \boldsymbol{x})p(\boldsymbol{x})}{p(\boldsymbol{y})} && \text{(2-33)}\\
&\propto p(\boldsymbol{y} \mid \boldsymbol{x})p(\boldsymbol{x}) && \text{(2-34)}\\
&= \exp\left(-\frac{1}{2}(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}) - \frac{1}{2}\boldsymbol{x}^T\boldsymbol{\Gamma}^{-1}\boldsymbol{x}\right) && \text{(2-35)}\\
&= \exp\left(-\frac{1}{2}\left(\boldsymbol{x}^T(\boldsymbol{A}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{A} + \boldsymbol{\Gamma}^{-1})\boldsymbol{x} - 2\boldsymbol{x}^T\boldsymbol{A}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{y} + \boldsymbol{y}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{y}\right)\right) && \text{(2-36)}\\
&= \exp\left((\boldsymbol{x} - \boldsymbol{\mu})^T\boldsymbol{P}(\boldsymbol{x} - \boldsymbol{\mu}) - \boldsymbol{\mu}^T\boldsymbol{P}\boldsymbol{\mu}\right) && \text{(2-37)}\\
&\propto \exp\left((\boldsymbol{x} - \boldsymbol{\mu})^T\boldsymbol{P}(\boldsymbol{x} - \boldsymbol{\mu})\right) && \text{(2-38)}\\
&\text{with} \\
&\boldsymbol{P} = \boldsymbol{A}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{A} + \boldsymbol{\Gamma}^{-1} \text{ and } \boldsymbol{\mu} = \boldsymbol{P}^{-1}\boldsymbol{A}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{y}. && \text{(2-39)}
\end{aligned}$$

In the derivations above, Equation 2-36 is rearranged into Equation 2-37 by completing the square of the exponents, and the second term in Equation 2-37 is omitted in Equation 2-38 as the term is independent of $\boldsymbol{x}$. Since the posterior in Equation 2-38 is in the Gaussian form with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{P}^{-1}$, the result shows that the posterior can be computed

without needing to evaluate the normalization constant. However, if non-conjugate prior is used, the posterior distribution will not belong to a known family of distributions. In such cases, numerical approximation methods are required to compute the normalization constant to fully characterize the posterior. Since inference becomes computationally prohibitive in more complex hierarchical models, the complex models are usually constructed from the conjugate pairs of prior and likelihood [4].

2-5 Variational Expectation-Maximization

Variational expectation-maximization (VEM) extends the variational inference by iteratively updating not only the variational posterior $q(\boldsymbol{x};\boldsymbol{\nu})$ but also hyperparameters $\boldsymbol{\theta}$ of the hierarchical Bayesian model using the maximum-likelihood [45]. Specifically, given the hyperparameters $\boldsymbol{\theta}$ (i.e., tuning parameters of the hyperpriors), Equation 2-15 is extended to:

$$\ln p(\boldsymbol{y};\boldsymbol{\theta}) - \mathcal{L}(q(\boldsymbol{x};\boldsymbol{\nu}),\boldsymbol{\theta}) = \mathrm{KL}\left(q(\boldsymbol{x};\boldsymbol{\nu})||p(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta})\right) \geq 0, \tag{2-40}$$

and the ELBO is

$$\mathcal{L}(q(\boldsymbol{x};\boldsymbol{\nu}),\boldsymbol{\theta}) = \int q(\boldsymbol{x};\boldsymbol{\nu}) \ln \frac{p(\boldsymbol{x},\boldsymbol{y};\boldsymbol{\theta})}{q(\boldsymbol{x};\boldsymbol{\nu})} d\boldsymbol{x} \tag{2-41}$$

$$= \mathbb{E}_{z\sim q(\boldsymbol{x};\boldsymbol{\nu})}[\ln p(\boldsymbol{x},\boldsymbol{y};\boldsymbol{\theta})] - \mathbb{E}_{z\sim q(\boldsymbol{x};\boldsymbol{\nu})}[\ln q(\boldsymbol{x};\boldsymbol{\nu})]. \tag{2-42}$$

Assuming the ELBO is not jointly concave in $q(\boldsymbol{x};\boldsymbol{\nu})$ and $\boldsymbol{\theta}$, it is maximized alternatively with respect to $q(\boldsymbol{x};\boldsymbol{\nu})$ and $\boldsymbol{\theta}$. Applying the mean-field approximation and following the coordinate ascent method, the alternating maximization of the ELBO is defined in two steps:

- E-step: update $q(\boldsymbol{x}_j,\nu_j)$ with

$$\ln q_j(\boldsymbol{x}_j;\nu_j^*) = \mathbb{E}_{i\neq j}[\ln p(\boldsymbol{x},\boldsymbol{y};\boldsymbol{\theta}^{(t)})] + \text{const}, \tag{2-43}$$

  where $\boldsymbol{\theta}^{(t)}$ is the current estimate of the hyperparameter $\boldsymbol{\theta}$

- M-step: update the estimate of $\boldsymbol{\theta}$ by

$$\boldsymbol{\theta}^{t+1} = \arg\max_{\boldsymbol{\theta}} \mathbb{E}_{\boldsymbol{x}\sim\prod_i q(\boldsymbol{x}_i;\nu_i)}[\ln p(\boldsymbol{x},\boldsymbol{y};\boldsymbol{\theta})]. \tag{2-44}$$

In the E-step, the maximization of the ELBO with respect to $q(\boldsymbol{x};\boldsymbol{\nu})$, is the same as in Equation 2-20, except that the expectation of the complete log-likelihood is conditioned on the current estimate $\boldsymbol{\theta}^{(t)}$ of the hyperparameters (i.e., $\mathbb{E}_{i\neq j}[\ln p(\boldsymbol{x},\boldsymbol{y};\boldsymbol{\theta}^{(t)})]$). In the M-step, the maximization of the ELBO with respect to the hyperparameters is performed by maximizing the expected complete data log-likelihood, where the expectation is taken using the most recent estimations of the posteriors. The entropy term in the ELBO (Equation 2-42) is constant with respect to $\boldsymbol{\theta}$ and therefore does not appear in the M-step.

Alternatively, the VEM can be seen as the ordinary expectation-maximization (EM) with modification on the E-step where the intractable true posteriors are replaced with variational posteriors[45].

2-6 Connection to MAP

Maximum a posteriori (MAP) estimation is a point estimation method used for inference on unobserved variables of hierarchical Bayesian models, defined as

$$\hat{\boldsymbol{x}} = \arg\max_{\boldsymbol{x}} \ln p(\boldsymbol{y}, \boldsymbol{x}). \tag{2-45}$$

While the coordinate ascent variational inference generally requires careful construction of iterative update equations, solving Equation 2-45 directly using nonlinear optimization methods might be appealing due to its simplicity. However, as MAP is a point estimation method, the direct optimization of the MAP estimation is prone to overfitting, particularly when the true posterior distribution is multimodal [25]. To better understand this tendency, we first revisit the interpretation of the terms in the ELBO. Recall that the ELBO is expressed as

$$\mathcal{L}(q(\boldsymbol{x};\boldsymbol{\nu})) = \int q(\boldsymbol{x};\boldsymbol{\nu}) \ln p(\boldsymbol{x},\boldsymbol{y}) d\boldsymbol{x} - \int q(\boldsymbol{x};\boldsymbol{\nu}) \ln q(\boldsymbol{x};\boldsymbol{\nu}) d\boldsymbol{x} \tag{2-46}$$

$$= \mathbb{E}_{z\sim q(\boldsymbol{x};\boldsymbol{\nu})}[\ln p(\boldsymbol{x},\boldsymbol{y})] - \mathbb{E}_{z\sim q(\boldsymbol{x};\boldsymbol{\nu})}[\ln q(\boldsymbol{x};\boldsymbol{\nu})]. \tag{2-47}$$

The first term of the ELBO represents the fitness of the model to the observed data and encourages the variational posterior $q(\boldsymbol{x},\boldsymbol{\nu})$ to place its probability mass around the point estimate of the MAP since matching the mass of $q(\boldsymbol{x},\boldsymbol{\nu})$ to the mass of $p(\boldsymbol{x},\boldsymbol{y})$ maximizes the first term. The second term, the entropy of $q(\boldsymbol{x},\boldsymbol{\nu})$ promotes more dispersed distributions since entropy of a distribution increases as its probability mass diffuses [4]. In other words, the second term prefers $q(\boldsymbol{x};\boldsymbol{\nu})$ to be as general as possible.

In contrast, MAP can be interpreted as maximization of the ELBO using a dirac-delta function as a variational posterior, i.e., $q(\boldsymbol{x};\boldsymbol{\nu}) = \delta(\boldsymbol{x} - \boldsymbol{x}_{\mathrm{MAP}})$ with unknown MAP estimate $\boldsymbol{x}_{\mathrm{MAP}}$. Following the derivations in [25], substituting into the ELBO leads

$$\mathcal{L}(\delta(\boldsymbol{x} - \boldsymbol{x}_{\mathrm{MAP}})) = \int \delta(\boldsymbol{x} - \boldsymbol{x}_{\mathrm{MAP}}) \ln p(\boldsymbol{x},\boldsymbol{y}) d\boldsymbol{x} - \int \delta(\boldsymbol{x} - \boldsymbol{x}_{\mathrm{MAP}}) \ln \delta(\boldsymbol{x} - \boldsymbol{x}_{\mathrm{MAP}}) d\boldsymbol{x} \tag{2-48}$$

$$= \ln p(\boldsymbol{x}_{\mathrm{MAP}}, \boldsymbol{y}) + \text{const.} \tag{2-49}$$

Here, the second term becomes $-\infty$ but independent of $\boldsymbol{x}_{\mathrm{MAP}}$. Consequently, the MAP can be seen as the variational inference without the regulation term introduced by the entropy term.

Thus, while MAP estimation offers computational simplicity, the maximization of the ELBO in the variational inference incorporates the balance between overfitting and underfitting of the variational posterior distribution, making the variational inference more robust than MAP estimation when the true posterior distribution is multimodal.

2-7 Related work: Joint Detection-Estimation (JDE)

In this section, prior work on the joint detection-estimation (JDE) developed for fMRI in literature is presented, based on which our proposed model will be designed.

The JDE was first introduced by Makni et. al. [31] in 2005 as a Bayesian method to simultaneously perform two tasks that were traditionally handled in separate steps: (1) a detection

step to localize brain regions activated by a given stimulus, and (2) an estimation step to capture the hemodynamics of the brain.

To achieve this, the JDE employs variational inference framework and jointly estimates three key components: the neural activation class labels (e.g., "active" or "non-active") for each pixel, the neural response level (NRL) indicating how strongly a region is activated, and the last component is the HRF.

Figure 2-4 provides a high-level overview of the components estimated by the JDE framework.

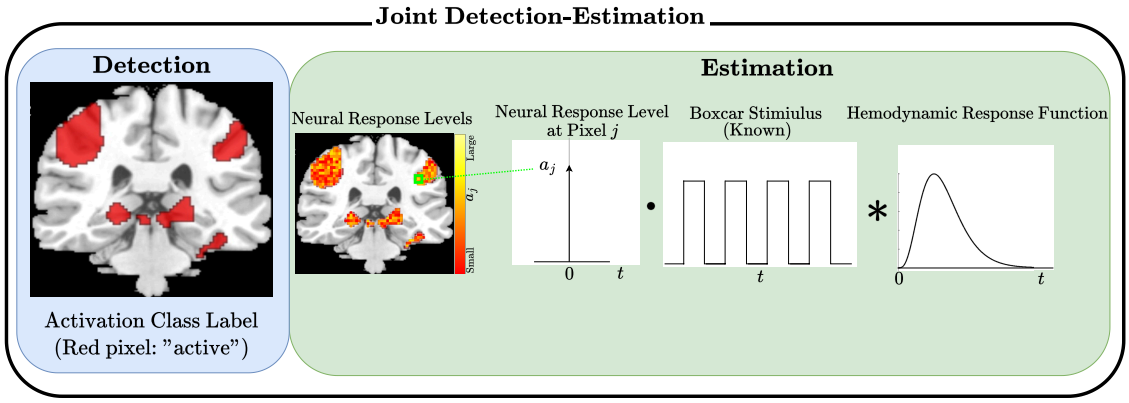


Figure 2-4: Schematic of the JDE framework. Detection of activation class and estimation of pixel-level NRLs and a single common HRF are performed jointly. Activation image under "Detection" is adapted from Zeithamova (2008) [53], © 2008 Society for Neuroscience. Used under educational fair use.

Within the variational inference framework, the computation of the JDE has been implemented in two different ways in literature: (a) with a sampling-based method known as Markov Chain Monte Carlo (MCMC), and (b) with a deterministic method like CAVI described in Section 2-3. Most of the prior works before 2012 such as [31, 32, 49] use MCMC which is typically computationally more intensive than CAVI.

The adaptation to CAVI was pioneered in 2012 by Chaari et al. [7] and most of the prior works after 2012 implement the JDE using the deterministic method. The most recent extension of the JDE method is joint parcellation-detection-estimation (JPDE) [1] where the parcellation of the brain is also estimated simultaneously with the spatial activity map and the hemodynamic response.

In this thesis, we extend the JDE to capture not only the hemodynamic response but also the neural response of the brain using the CAVI implementation. Thus, in this section we limit our scope to the JDE in [7].

2-7-1 Observation Model

The observation model in the JDE involves the convolution between the timings of the experimental conditions and the HRF. Measurement noise and physiological artifacts are added to the convolution signal. As most hemodynamic models used in literature, the timings of

experimental conditions are used as a surrogate of neural activity signals, disregarding the variability of neural activation level in time and space. In the JDE, the notion of the variability in neural activity is reflected on NRLs, which are modeled as unknown pixel-specific scaling factors of the convolution for each experimental conditions.

Let $\boldsymbol{y}_j \in \mathbb{R}^{N\times 1}$ denote the blood oxygenation level dependent (BOLD) signal at pixel j in region of interest (ROI) of the brain, where N is the temporal signal length, and $\boldsymbol{x}_m \in \mathbb{R}^{N\times 1}$ denotes the time series of the experimental condition. The HRF is assumed to be shared across all pixels in the ROI and is modeled as an FIR filter, of length L_h, denoted by $\boldsymbol{h} \in \mathbb{R}^{L_h\times 1}$. The NRL for the external condition m at pixel j is a_j^m, and a_j^m is assumed to be independent of other experimental conditions and pixels.

Assume M experimental conditions, the observed BOLD signal at pixel j is modeled as

$$\boldsymbol{y}_j = \sum_{m=1}^{M} a_j^m \boldsymbol{X}_m \boldsymbol{h} + \boldsymbol{P}\boldsymbol{\ell}_j + \boldsymbol{\epsilon}_j. \tag{2-50}$$

Here:

- $\boldsymbol{X}_m \in \mathbb{R}^{N\times L_h}$ is a Toeplitz matrix of $\boldsymbol{x_m}$.
- $a_j^m \boldsymbol{X}_m \boldsymbol{h}$ represents the convolution of the HRF and the surrogate mth input signal $\boldsymbol{x}_m$ which is weighted by NRL a_j^m, reflecting the neural activation at pixel j.
- $\boldsymbol{\epsilon}_j \in \mathbb{R}^{N\times 1}$ is the measurement noise at pixel j.
- $\boldsymbol{P}\boldsymbol{\ell}_j \in \mathbb{R}^{N\times 1}$ represents low frequency artifacts in the BOLD at pixel j.
- $\boldsymbol{P} \in \mathbb{R}^{N\times U}$ is a discrete cosine transform (DCT) matrix containing one DC offset vector and $U-1$ low frequency cosine vectors of length N.
- $\boldsymbol{\ell_j}$ is a coefficient vector for $\boldsymbol{P}$.

As for the number U of basis vectors in the matrix $\boldsymbol{P}$, U is determined as [11]:

$$U = \lceil 2N f_{\min} \rceil + 1, \tag{2-51}$$

where $f_{\min}$ is the lowest frequency component that is attributable to the low frequency artifact in the BOLD signal. Alternatively, one may think of $f_{\min}$ as a cut-off frequency of a low-pass filter typically used in preprocessing.

In Equation 2-50, the BOLD signal $\boldsymbol{y}_j$ and the time course of the experimental condition $\boldsymbol{x}_m$ (and thus $\boldsymbol{X}_m$) are the only variables that we observe. In the context of variational inference, we treat HRF $\boldsymbol{h}$ and NRL a_j^m as random variables for all j and m, which we refer to as latent variables. Since we treat these variables as random variables, they are distributed according to some prior distributions. Their prior distributions are chosen to reflect our prior physiological knowledge on the latent variables, which we will discuss in details in the following section. On the other hand, we refer to other non-random quantities $\boldsymbol{P}$, $\boldsymbol{\ell_j}$ as model parameters.

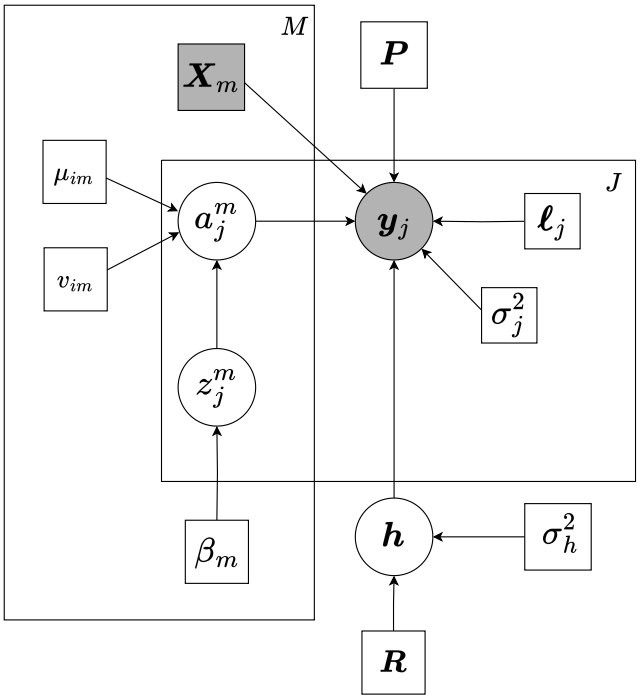


Figure 2-5: Graphical model for hierarchical bayesian model of JDE [7]: circular nodes represent latent random variables and square nodes represent model parameters. Nodes for observed variable and parameter are shaded. The rectangular plate with number (M or J) indicates the number of identical copies of the plate. The direction of edges connecting nodes indicate the causal dependence between the nodes.

2-7-2 Hierarchical Bayesian Model

Apart from the latent variables $\boldsymbol{h}$ and a_j^m from Equation 2-50, we define additional latent variable, z_j^m, the neural activation class label. Before we discuss these latent variables and their priors, we shall first present the graphical Bayesian model in Figure 2-5 to provide an overview of the hierarchical Bayesian model.

The graphical model represents the hierarchical Bayesian structure in which the causal relations between variables are indicated by directed edges. For example, the node $\boldsymbol{y}_j$ is directly connected to nodes $\boldsymbol{X}_m$, a_j^m, $\boldsymbol{h}$, σ_j, $\boldsymbol{P}$ and $\boldsymbol{\ell}_j$, through the arrows directed to $\boldsymbol{y}_j$. This represents that the observed measurement $\boldsymbol{y}_j$ is generated by the variables and parameters that are directly connected to it. Assuming σ_j^2 is the parameter related to the noise vector $\boldsymbol{\epsilon}_j$, this agrees with the observation equations in Equation 2-50, where the generation of the observation is $\boldsymbol{y}_j$ directly dependent on those parameters.

The graphical representation helps preserve the clarity of the model's structure as we introduce additional equations and symbolic notation in the following sections.

To complete the hierarchical Bayesian model, the likelihood from the observation equation and a prior distribution of each latent variable are specified in the following.

2-7-3 Likelihood

The likelihood of the observed data reflects the fitness of the chosen model for the data and its distribution is characterized by the noise model. In the JDE framework [32, 7], a first-order auto-regressive (AR(1)) model is used to model colored noise in the BOLD signal due to the physiological noise. Specifically, the noise $\boldsymbol{\epsilon}_j$ is modeled as

$$\boldsymbol{\epsilon}_j \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Gamma}_j^{-1}) \tag{2-52}$$

with

$$\boldsymbol{\Gamma}_j = \sigma_j^{-2} \boldsymbol{\Lambda}_j, \tag{2-53}$$

where $\boldsymbol{\Lambda}_j$ is

$$\boldsymbol{\Lambda}_j = \begin{bmatrix} 1 & -\rho_j & 0 & \cdots & 0 \\ -\rho_j & 1+\rho^2 & -\rho_j & \cdots & 0 \\ & \ddots & \ddots & \ddots & \\ 0 & \cdots & -\rho_j & 1+\rho^2 & -\rho_j \\ 0 & \cdots & 0 & -\rho_j & 1 \end{bmatrix} \in \mathbb{R}^{N \times N} \tag{2-54}$$

and ρ_j is the AR(1) coefficient for pixel j. $\boldsymbol{\Gamma}_j$ represents the precision matrix of the AR(1) noise and $\sigma_j^2 = \frac{\sigma_0^2}{1-\rho_j^2}$, with σ_0^2 (the variance) of the white noise in the AR(1).

From the noise model in Equation 2-52 and the observation model in Equation 2-50, the likelihood for $\boldsymbol{y}_j$ is

$$p(\boldsymbol{y}_j | \boldsymbol{h}, \boldsymbol{a}_j \, ; \boldsymbol{X}, \boldsymbol{P}, \boldsymbol{\ell}_j, \boldsymbol{\Gamma}_j) = \boldsymbol{\mathcal{N}} \left(\sum_{m=1}^{M} a_j^m \boldsymbol{X}_m \boldsymbol{h} + \boldsymbol{P} \boldsymbol{\ell}_j \, , \boldsymbol{\Gamma}_j^{-1} \right) \tag{2-55}$$

$$= (2\pi)^{-\frac{N}{2}} \det(\boldsymbol{\Gamma}_j)^{\frac{1}{2}} \exp \left(-\frac{1}{2} \left(\boldsymbol{y}_j - \sum_{m=1}^{M} a_j^m \boldsymbol{X}_m \boldsymbol{h} - \boldsymbol{P} \boldsymbol{\ell}_j \right)^{\top} \boldsymbol{\Gamma}_j \left(\boldsymbol{y}_j - \sum_{m=1}^{M} a_j^m \boldsymbol{X}_m \boldsymbol{h} - \boldsymbol{P} \boldsymbol{\ell}_j \right) \right), \tag{2-56}$$

where $\boldsymbol{a}_j$ is represents the set $\boldsymbol{a}_j = \{a_j^m\}_{m=1:M}$. We assume that the noise is independent between pixels and define sets here to avoid cluttering: $\boldsymbol{y} = \{\boldsymbol{y}_j\}_{j=1:J}$, $\boldsymbol{a} = \{\boldsymbol{a}_j^m\}_{j=1:J}^{m=1:M}$, $\boldsymbol{\ell} = \{\boldsymbol{\ell}_j\}_{j=1:J}$, $\boldsymbol{X} = \{\boldsymbol{X}_m\}_{m=1:M}$ and $\boldsymbol{\Gamma} = \{\boldsymbol{\Gamma}_j\}_{j=1:J}$. The likelihood for all observation $\boldsymbol{y}$ is then expressed as:

$$p(\boldsymbol{y} | \boldsymbol{h}, \boldsymbol{a} \, ; \boldsymbol{X}, \boldsymbol{P}, \boldsymbol{\ell}, \boldsymbol{\Gamma}) = \prod_{j=1}^{J} p(\boldsymbol{y}_j | \boldsymbol{h}, \boldsymbol{a}_j \, ; \boldsymbol{X}, \boldsymbol{P}, \boldsymbol{\ell}_j, \boldsymbol{\Gamma}_j) \tag{2-57}$$

$$= \prod_{j=1}^{J} \boldsymbol{\mathcal{N}} \left(\sum_{m=1}^{M} a_j^m \boldsymbol{X}_m \boldsymbol{h} + \boldsymbol{P} \boldsymbol{\ell}_j \, , \boldsymbol{\Gamma}_j^{-1} \right). \tag{2-58}$$

Throughout this report, the symbol with omitted subscript and superscript denotes a set of such symbol over the omitted subscripts and superscripts for compactness.

2-7-4 Prior on Hemodynamic Response Function

For the model prior of the HRF, Chaari et al. [7] follows the prior models proposed in [5, 16], in which two physiological assumptions are incorporated: (a) the HRF is causal and decays to a baseline; (b) variations of the HRF are smooth. For (a), the first and the last coefficients of the HRF are fixed to zero, i.e., $\boldsymbol{h}(1) = \boldsymbol{h}(L_h) = 0$. For (b), a zero-mean multivariate Gaussian prior with a covariance matrix based on the second-order difference matrix $\boldsymbol{D}$ is used:

$$\boldsymbol{h}(2 : L_h - 1) \sim \mathcal{N}\left(\boldsymbol{0}, \sigma_h^2 \boldsymbol{R}\right) \tag{2-59}$$

with

$$\boldsymbol{D} = \begin{bmatrix} -2 & 1 & 0 & 0 & \cdots & 0 \\ 1 & -2 & 1 & 0 & \cdots & 0 \\ 0 & 1 & -2 & 1 & \cdots & 0 \\ 0 & 0 & 1 & -2 & \cdots & 0 \\ & \vdots & & \ddots & \ddots & 1 \\ 0 & 0 & \cdots & 0 & 1 & -2 \end{bmatrix} \in \mathbb{R}^{(L_h-2)\times(L_h-2)} \tag{2-60}$$

and

$$\boldsymbol{R} = (\Delta t)^4 (\boldsymbol{D}^\top \boldsymbol{D})^{-1}, \tag{2-61}$$

where $\boldsymbol{h}(2 : L_h - 1)$ contains the FIR coefficients of the HRF from 2 and $L_h - 1$ time points and Δt is the sampling period of the observation. σ_h^2 is the scaling factor of the covariance matrix, which represents our uncertainty about the shape of HRF and controls the amplitude of $\boldsymbol{h}$ and how much the observation $\boldsymbol{y}$ is reflected on the solution of $\boldsymbol{h}$. For instance, a small σ_h^2 means we are confident that the second derivative of HRF is small, which in turn encourages the solution of $\boldsymbol{h}$ to be flat. On the other hand, a large σ_h^2 represents that we are uncertain about our smooth fluctuation assumption, allowing both $\boldsymbol{h}$ to have large second derivative and the observation data to have large influence on the shape of $\boldsymbol{h}$.

2-7-5 Prior on Neural Response Level

The NRL a_j^m reflects the amplitude of the neural activity and is conditioned on the activation class z_j^m. The distribution of a_j^m is modeled as Gaussian:

$$p(a_j^m | z_j^m = k;\ \mu_{im}, v_{im}) \sim \mathcal{N}(\mu_{im}, v_{im}) \text{ for } k \in \{0, 1\}, \tag{2-62}$$

where μ_{im} and v_{im} are the mean and variance of the neural activation level at pixel j for the activation class $z_j^m = k$. Here, $z_j^m = 1$ represents that the neural population at pixel j is activated for stimulus m. $z_j^m = 0$ represents that pixel j is not activated for stimulus m. With the assumption that the non-activated pixel does not contribute to the observed signal, the mean NRL for non-activated pixel is set to zero (i.e., $\mu_{0m} = 0$) while all other parameters are unknown and are to be estimated. Additionally, the JDE assumes that the NRL for all experimental conditions are independent and the NRL at each pixel is independent of other pixels. Thus, the prior distribution for a set of NRLs $\boldsymbol{a}$ is modeled as:

$$p(\boldsymbol{a}|\boldsymbol{z}\ ; \boldsymbol{\mu}_{\{1:M\}}, \boldsymbol{v}_{\{1:M\}}) = \prod_{j=1}^{J} \prod_{m=1}^{M} p(a_j^m | z_j^m; \boldsymbol{\mu}_m, \boldsymbol{v}_m) \tag{2-63}$$

$$\tag{2-64}$$

where $\boldsymbol{\mu}_m = \{\mu_{0m}, \mu_{1m}\}$ and $\boldsymbol{v}_m = \{v_{0m}, v_{1m}\}$.

2-7-6 Prior on Neural Activation Class

The physiological knowledge used for activation class label z_j^m is the regional homogeneity in the brain. The regional homogeneity is the hypothesis that the neural populations in a given region are likely to share the same activation state [27, 24, 47]. For instance, when neural population at pixel j is active, we expect neighboring pixels to be also active. Assume there are two activation classes: 1 for "active" and 0 for non-active. The neural activation class for a set $\boldsymbol{z}^m = \{z_j^m\}_{j=1:J}$ is modeled using Ising model[34]:

$$p(\boldsymbol{z}^m; \beta_m) = \frac{1}{Z(\beta_m)} \exp\left(\beta_m \sum_{j=1}^{J} \sum_{i \in \delta_j} I(z_i^m = z_j^m)\right) \tag{2-65}$$

where β_m is the spatial regularization parameter and $I(\cdot)$ is the indicator function: $I(z_i^m = z_j^m) = 1$ if $z_i^m = z_j^m$ and 0 otherwise. δ_j denotes all neighboring pixels of pixel j. The indicator function in the exponents accounts for the number of neighboring pixels in the same activation class. $Z(\beta_m) = \sum_{\boldsymbol{z}^m} e^{\beta_m \sum_{j \sim i} I(z_i^m = z_j^m)}$ is the normalization constant, which is the summation over all possible configurations of $\boldsymbol{z}^m$.

The Ising model promotes the regional homogeneity of the activation between pixels. For instance, the exponent becomes larger when more neighboring pixels share the same activation class, resulting in larger probability for configuration $\boldsymbol{z^m}$ with a high regional homogeneity. Similarly, the degree of the regional homogeneity is controlled by β_m. When β_m is small, the regional homogeneity becomes weak as the exponent becomes small.

Chapter 3

Proposed Model

In this chapter, we present a cascade LTI model to capture the dynamics of the neurovascular coupling of the brain. Specifically, we model the neurovascular system in two FIR filters: the first one is neural response function (NRF) which models the dynamics of neural activity and the second one is HRF. We then extend this LTI system to a hierarchical Bayesian model, in which we incorporate some known physiological phenomena. We proceed with derivation of the optimal solutions in CAVI framework and present the iterative update solutions to the designed model.

3-1 Neurovascular Model

As stated in Chapter 1, our objective is to model and estimate dynamics of the complex neurovascular system in the brain. Figure 3-1 illustrates the high-level structure of our cascade LTI model. Given an external stimulus presented to a subject, the neurons in some brain regions react and process the incoming stimulus and generate neural activity. The NRF is modeled as an LTI system, where the stimulus onset serves as the input and the resulting neural activity as the output. This output reflects the collective response of the neuronal population within a given pixel. The resulting neural activity causes contraction and dilation of capillaries surrounding the activated neurons to meet the metabolic demand, which in turn causes the change in blood volume through HRF.

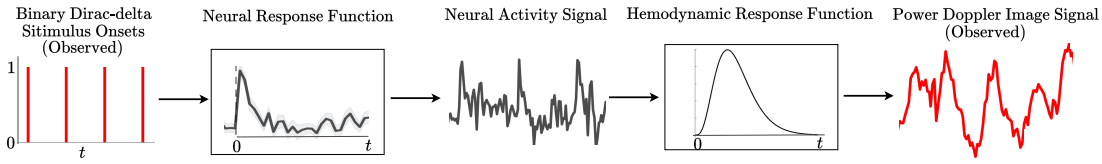


Figure 3-1: High-level illustration of our neurovascular model. The convolution of the binary stimulus onset input signal and the neural response function yields the neural activity signal. The neural activity signal is then mapped to the measured power Doppler image signal via convolution with the hemodynamic response function. The traces of the neural response function, neural activity, and power Doppler image signals in this figure are adapted from Nunez-Elizalde et al., 2022, [38], © 2025 Elsevier B.V. Used under fair use for academic purposes.

The PDI signal in fUS is commonly in three dimensional format, two spatial dimensions for a 2D slice of the brain and one dimension for the time evolution of the signal. Throughout this paper, we consider ROI of PDI data: a subset of pixel js in the 2D slice and N samples in time and we refer pixel to the index as j. Let $\boldsymbol{y}_j \in \mathbb{R}^{N\times 1}$ denote the PDI signal at pixel j, and $\boldsymbol{x}_m \in \mathbb{R}^{N\times 1}$ denote the time series of the onset of external stimulus m. We consider that M different stimuli or experimental conditions are used.

The neurovascular system is represented as a two-stage convolution system:

$$\boldsymbol{y}_j = \sum_{m=1}^{M} \boldsymbol{X}_m \boldsymbol{R}_j^m \boldsymbol{h} \tag{3-1}$$

or

$$\boldsymbol{y}_j = \sum_{m=1}^{M} \boldsymbol{X}_m \boldsymbol{H} \boldsymbol{r}_j^m. \tag{3-2}$$

Here:

- $\boldsymbol{r}_j^m \in \mathbb{R}^{L_r\times 1}$ denotes the NRF at pixel j for stimulus m, modeled as a FIR of length L_r,
- $\boldsymbol{h} \in \mathbb{R}^{L_h\times 1}$ denotes the HRF, modeled as a FIR of length L_h,
- $\boldsymbol{X}_m \in \mathbb{R}^{N\times(L_h+L_r+1)}$ is a Toeplitz matrix of $\boldsymbol{x_m}$,
- $\boldsymbol{R}_j^m \in \mathbb{R}^{(L_h+L_r-1)\times L_h}$ is a Toeplitz matrix of $\boldsymbol{r}_j^m$,
- $\boldsymbol{H} \in \mathbb{R}^{(L_h+L_r-1)\times L_r}$ is a Toeplitz matrix of $\boldsymbol{h}$.

In the model above, we assume that:

- the NRF is specific to each pixel and depends on the stimulus type.
- The HRF is shared across all pixels in the ROI.

As noted in Chapter 1, PDI signals contain not only the blood signal, but also measurement noise from the electronic equipments. Furthermore, PDI signals often contain low frequency

artifacts that are not rejected in the clutter filtering, which originate from cardiovascular pulsatility and breathing.

Assuming that the noise and the artifacts are additive to the blood signal in the PDI signal, and following [7], the model equation becomes

$$\boldsymbol{y}_j = \sum_{m=1}^{M} \boldsymbol{X}_m \boldsymbol{R}_j^m \boldsymbol{h} + \boldsymbol{P}\boldsymbol{\ell}_j + \boldsymbol{\epsilon}_j \tag{3-3}$$

$$\boldsymbol{y}_j = \sum_{m=1}^{M} \boldsymbol{X}_m \boldsymbol{H} \boldsymbol{r}_j^m + \boldsymbol{P}\boldsymbol{\ell}_j + \boldsymbol{\epsilon}_j. \tag{3-4}$$

where $\boldsymbol{\epsilon}_j \in \mathbb{R}^{N\times 1}$ is the measurement noise, and $\boldsymbol{P}\boldsymbol{\ell}_j$ are the artifacts. Here, $\boldsymbol{P} \in \mathbb{R}^{N\times U}$ contains a set of low frequency sinusoids and DC offset, and $\boldsymbol{\ell}_j \in \mathbb{R}^{U\times 1}$ is the coefficient vector for $\boldsymbol{P}$.

3-2 Bayesian model

In this section, we extend our observation model defined in Equations 3-3 and 3-4 by adopting a Bayesian framework. Specifically, we treat the unknown variables $\boldsymbol{h}$ and $\boldsymbol{r}_j^m$ as latent variables with specific distributions to incorporate our prior physiological knowledge.

In addition to these latent variables, we introduce two more latent variables: z_j^m and β_m. z_j^m represents the activation class of neural population at pixel j under stimulus m and β_m is related spatial correlation of z_j^m. Compared to the JDE presented in Section 2-7-2, the roles of z_j^m and β_m remain unchanged. However, in our model, β_m is explicitly treated as a latent variable with a prior distribution.

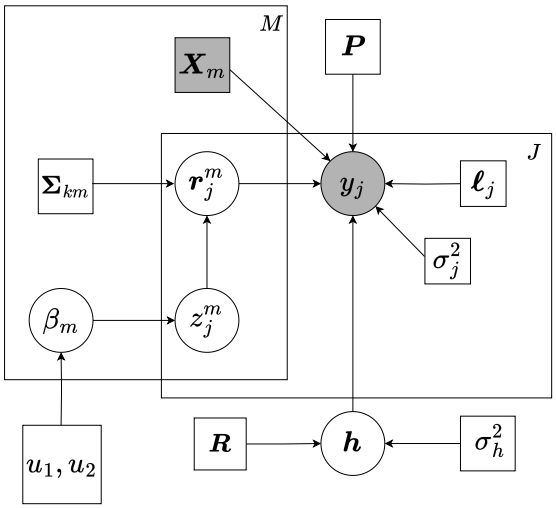


Figure 3-2: Graphical model of the proposed JDE. In comparison to the graphical model in Figure 2-5, the NRL a_j^m node is replaced by NRF $\boldsymbol{r}_j^m$ in our model. In addition, the spatial regularization parameter β_m is treated as a latent variable.

Figure 3-2 shows the graphical representation of our proposed Bayesian model. As described in Section 2-7-2, directed edges in graphical Bayesian models denote causal dependencies between variables. Although the overall structure resembles the earlier model, our proposed model differs in $\boldsymbol{r}_j^m$ node and β_m node.

In what follows, we define the likelihood from the observation equations and a prior distribution of each latent variable.

3-2-1 Likelihood

We model the measurement noise $\boldsymbol{\epsilon}_j$ from Equations 3-3 and 3-4 as independently and identically distributed (iid) additive zero-mean Gaussian noise with noise variance of σ_j^2 for all pixels and time steps. While an $AR(1)$ noise model is often used in fMRI data, since fMRI is known to have correlated noise, the characteristics of measurement noise in fUS data have not been investigated, and thus we use the simplest noise model here.

Since $\boldsymbol{\epsilon}_j$ is a random variable, the observed signal $\boldsymbol{y}_j$ follows a Gaussian distribution, the likelihood of the observed signal $\boldsymbol{y}_j$ as a function of latent variables and parameters is:

$$p(\boldsymbol{y}_j|\boldsymbol{r}_j,\boldsymbol{h},\boldsymbol{\ell}_j;\sigma_j^2) = \mathcal{N}\left(\sum_{m=1}^{M}\boldsymbol{X}_m\boldsymbol{R}_j^m\boldsymbol{h}+\boldsymbol{P}\boldsymbol{\ell}_j,\sigma_j^2\boldsymbol{I}_N\right) \tag{3-5}$$

$$= \frac{1}{(2\pi\sigma_j^2)^{\frac{N}{2}}}\exp\left(-\frac{1}{2\sigma_j^2}\left(\boldsymbol{y}_j-\sum_{m=1}^{M}\boldsymbol{X}_m\boldsymbol{R}_j^m\boldsymbol{h}-\boldsymbol{P}\boldsymbol{\ell}_j\right)^\top\left(\boldsymbol{y}_j-\sum_{m=1}^{M}\boldsymbol{X}_m\boldsymbol{R}_j^m\boldsymbol{h}-\boldsymbol{P}\boldsymbol{\ell}_j\right)\right), \tag{3-6}$$

or alternatively,

$$p(\boldsymbol{y}_j|\boldsymbol{r}_j,\boldsymbol{h},\boldsymbol{\ell}_j;\sigma_j^2) = \mathcal{N}\left(\sum_{m=1}^{M}\boldsymbol{X}_m\boldsymbol{H}\boldsymbol{r}_j^m+\boldsymbol{P}\boldsymbol{\ell}_j,\sigma_j^2\boldsymbol{I}_N\right) \tag{3-7}$$

$$= \frac{1}{(2\pi\sigma_j^2)^{\frac{N}{2}}}\exp\left(-\frac{1}{2\sigma_j^2}\left(\boldsymbol{y}_j-\sum_{m=1}^{M}\boldsymbol{X}_m\boldsymbol{H}\boldsymbol{r}_j^m-\boldsymbol{P}\boldsymbol{\ell}_j\right)^\top\left(\boldsymbol{y}_j-\sum_{m=1}^{M}\boldsymbol{X}_m\boldsymbol{H}\boldsymbol{r}_j^m-\boldsymbol{P}\boldsymbol{\ell}_j\right)\right), \tag{3-8}$$

where $\boldsymbol{I}_N$ is the $N\times N$ identity matrix. For compactness, we denote the set of NRFs involved at pixel j as $\boldsymbol{r}_j=\{\boldsymbol{r}_j^m\}_{m=1:M}$. Note that similar to the two observations equations, we have Equations 3-5 and 3-7 that are equivalent but with different ordering of the two convolutions, due to the commutativity of convolution.

Since we assume the noise is iid for all pixels, the set of PDI signals for all pixels, denoted by $\boldsymbol{y}=\{\boldsymbol{y}_j\}_{j=1:J}$, has likelihood function

$$p(\boldsymbol{y}|\boldsymbol{r},\boldsymbol{h},\boldsymbol{\ell};\sigma^2) = \prod_{j=1}^{J}p(\boldsymbol{y}_j|\boldsymbol{r}_j,\boldsymbol{h},\boldsymbol{\ell}_j;\sigma_j^2). \tag{3-9}$$

Again, the set of drift coefficients involved with pixel j is denoted by $\boldsymbol{\ell}=\{\boldsymbol{\ell}_j\}_{j=1:J}$ for compactness.

3-2-2 Prior on HRF

To construct the prior model of the HRF, we use two physiological assumptions on the HRF as in [5, 7] (also from Section 2-7):

- HRF has smooth variation,
- HRF is causal and decays to a baseline.

The first assumption stems from the fact that the HRF maps from the neural activity to the PDI signal which is related to the change in the blood volume. Since the blood volume in the brain cannot change rapidly due to physiological limitations, the HRF should not contain rapid fluctuation but rather has smoothly varying shape.

The second assumption incorporates two key features: first, the amplitude of HRF at onset should be zero or near zero to ensure the causality of the hemodynamic response; second, the HRF is modeled as a stable FIR, reflecting the empirical observations that the change in blood volume due to neurovascular coupling returns to the baseline within a finite time.

All JDE methods in literature [31, 32, 49, 7, 1] incorporate the first assumption using the second-order finite difference matrix and the second assumption by fixing the first and last coefficients of $\boldsymbol{h}$ to zero (See Section 2-7-4). While the effect of setting the two coefficients after every posterior update on $\boldsymbol{h}$ has not been reported in literature, this causes inconsistency in the CAVI algorithm, especially when the posterior variance of $\boldsymbol{h}$ is used in the update solutions of other variables. Therefore, we modify the covariance matrix in Equation 2-60 to encourage incorporating the second assumption without fixing the first and last coefficient of $\boldsymbol{h}$.

Recall that the prior distribution of $\boldsymbol{h}$ is a multivariate Gaussian. The proposed covariance matrix in our design is

$$\boldsymbol{R} = (\boldsymbol{A}^T\boldsymbol{A})^{-1} \in \mathbb{R}^{L_h \times L_h}, \tag{3-10}$$

where

$$\boldsymbol{A} = \frac{1}{\Delta t^2}(\boldsymbol{D}_2 + \boldsymbol{B}),\ \boldsymbol{B} = \begin{bmatrix} -a & 0 & \cdots & 0 \\ 0 & 0 & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & -a \end{bmatrix} \in \mathbb{R}^{L_h \times L_h}. \tag{3-11}$$

where the sampling period of PDI Δt and the second-order difference matrix $\boldsymbol{D}_2$ are kept the same as in Equation 2-59. Here, the difference from the earlier design is that the additional matrix $\boldsymbol{B}$ reduces the variance of the first and last coefficients of $\boldsymbol{h}$ by a.

From a Ridge regression perspective, the addition of $\boldsymbol{B}$ is effectively equivalent to the increasing penalty on the first and the last coefficients of $\boldsymbol{h}$ and their related cross-terms in a quadratic cost function. Figure 3-3 visualizes the effect of the modification on the covariance matrix, where the first and the last coefficients in randomly generated samples of $\boldsymbol{h}$ are close to zero.

In addition to this modification, we include a nonnegative prior mean $\boldsymbol{\mu}_{h0}$, as opposed to the earlier design where the prior mean is set to zero. This helps guide the estimation of $\boldsymbol{h}$

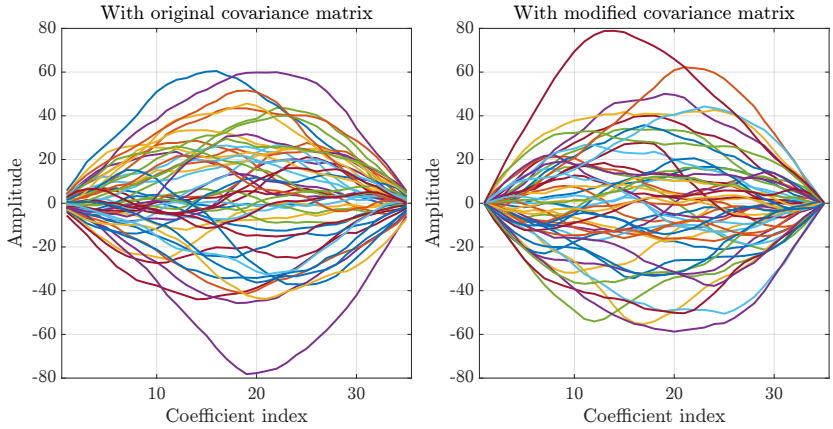


Figure 3-3: 50 randomly generated samples of $\boldsymbol{h}$ of length $L_h = 35$ from zero-mean multivariate Gaussian distribution with the original covariance matrix (Left) and with the modified covariance matrix (Right). $\Delta t = \sigma_h^2 = 1$ is used. In both plots, the random vectors generated from Gaussian distribution exhibit smooth shape, which helps the estimation of $\boldsymbol{h}$ avoid overfitting to the noise. Comparing to the left and the right plots, it can be observed that the modification of the covariance matrix drives the first and the last coefficients of the random vectors towards zero.

towards positive response function, which is important because our neurovascular model is more ill-posed due to the the presence of two convolution stages.

The resulting prior distribution for $\boldsymbol{h}$ is:

$$\boldsymbol{h} \sim \mathcal{N}\left(\mu_{h0}, \sigma_h^2 \boldsymbol{R}\right). \tag{3-12}$$

where σ_h^2 is a scaling factor of the covariance matrix, which represents our uncertainty about the shape of HRF and controls the amplitude of $\boldsymbol{h}$, as touched upon in Section 2-7-4.

To sum up, we model the distribution of $\boldsymbol{h}$ as a multivariate Gaussian distribution whose random realizations have smooth variation, causal onset and decay. By choosing this prior, we effectively confine the solution space of $\boldsymbol{h}$ to the space where $\boldsymbol{h}$ resembles the HRF in Figure 2-1b.

3-2-3 Prior on NRF

The prior work of the JDE in Section 2-7-5 captures the notion of neural activity as a scalar (a_j^m), representing the amplitude of the convolution between the HRF and the stimulus signal $\boldsymbol{x}_m$. In our model, we extend the notion of the neural activity to a vector $\boldsymbol{r}_j^m$ to capture its temporal dynamics in response to the onsets of stimulus. Similar to the prior model of HRF in the previous section, we use a multivariate Gaussian distribution as a prior distribution of the NRF. However, unlike the HRF, an empirical shape of NRF has not been specified in literature. Therefore, we design an expected shape of the NRF based on the neural response reported in [2, 17, 38, 40].

Our model assumptions on NRF are:

- Change in neural activity signal is instantaneous to the onset of stimulus,

- The NRF decays exponentially to a baseline after the onset of stimulus.

As regards the first assumption, while we recognize the causality of the neural activity in real life, the sampling rate of the NRF in our cascade LTI needs to be aligned with the sampling rate of the HRF, which is limited to the sampling rate of the PDI signal (typically up to 10Hz). Therefore, we assume that the onset of the neural activity can be instantaneous to the onset of stimulus when the sampling period of the neural activity is adjusted to the PDI signal.

As for the second assumption, although an actual neural response is generally considered to be nonlinear and depends on various factors such as types, intensity and durations of the external stimulus, we ignore the effect of the intensity and duration of stimulus in our model by approximating the neural response as a simple linear FIR that decays rapidly after the onset of the stimulus.

To impose the prior model assumptions on the NRF, we use a zero-mean multivariate Gaussian distribution conditioned on the neural activation class $z_j^m \in \{0, 1\}$. We assume $z_j^m = 1$ represents that the neural population in pixel j is active for stimulus m. We model the prior distribution of $\boldsymbol{r}_j^m$ for $z_j^m = 1$ as:

$$p(\boldsymbol{r}_j^m | z_j^m = 1) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{1m}) . \tag{3-13}$$

where covariance matrix $\boldsymbol{\Sigma}_{1m}$ is based on the second-order stable spline kernel [39], whose a-th row and b-th column entry is defined as:

$$\boldsymbol{\Sigma}_{1m}(a, b) = \lambda_r \left(\frac{\alpha_r^{a+b+\max(a,b)}}{2} - \frac{\alpha_r^{3\max(a,b)}}{6} \right) \tag{3-14}$$

$$\lambda_r \geq 0, \ 0 \leq \alpha_r < 1 . \tag{3-15}$$

Here, α_r is a decay factor which accounts for the exponential decay of the NRF and λ_r is a scaling factor which represents the degree of our prior belief on the shape of NRF and controls the amplitude of $\boldsymbol{r}_j^m$, analogous to σ_h^2 in the prior of HRF.

Figure 3-4 illustrates random vectors generated using the stable spline, incorporated with the assumptions of the NRF. Assuming that the zero NRF corresponds to the baseline neural activity, the negative NRF can be interpreted as a suppression of the baseline neural activity.

For the case where the neural population is not activated by stimulus m ($z_j^m = 0$), we model the prior distribution of $\boldsymbol{r}_j^m$ as:

$$p(\boldsymbol{r}_j^m | z_j^m) = \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{0m}) \tag{3-16}$$

with a diagonal covariance matrix

$$\boldsymbol{\Sigma}_{0m} = \sigma_0^2 \boldsymbol{I}_{L_r}, \tag{3-17}$$

where $\boldsymbol{I}_{L_r}$ denotes the $L_r \times L_r$ identity matrix.

The prior distribution of $\boldsymbol{r}_j^m$ conditioned on the class assignment z_j^m from Equations 3-13 and 3-16 can be summarized as:

$$p(\boldsymbol{r}_j^m | z_j^m) = \begin{cases} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{1m}) & \text{for } z_j^m = 1. \\ \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{0m}) & \text{for } z_j^m = 0. \end{cases} \tag{3-18}$$

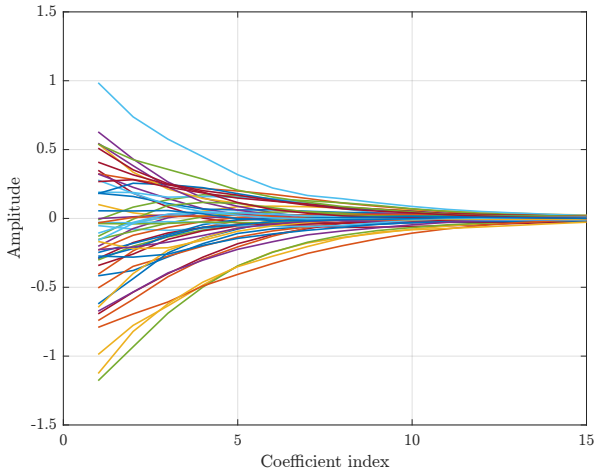


Figure 3-4: 50 random realizations of $\boldsymbol{r}_j^m$ ($L_r = 15$) from a zero-mean multivariate Gaussian distribution with the second-order stable spline kernel.($\lambda_r = 1$, $\alpha_r = 0.84$). The random vectors follow the exponential decay. This encourages the estimation of the NRF towards decaying functions, which helps the estimation algorithm avoid overfitting to the noise.

As z_j^m is a hard mixture assignment taking a value of either 0 or 1, $p(\boldsymbol{r}_j^m|z_j^m)$ can be interpreted as a hard Gaussian mixture of two, given by:

$$p(\boldsymbol{r}_j^m|z_j^m) = \mathcal{N}\left(\boldsymbol{0}, \boldsymbol{\Sigma}_{1m}\right)^{I(z_j^m=1)} + \mathcal{N}\left(\boldsymbol{0}, \boldsymbol{\Sigma}_{0m}\right)^{I(z_j^m=0)} \tag{3-19}$$

where $I(\cdot)$ is the indicator function.

So far, we have made assumptions and constructed the prior model only on the NRF at pixel j and stimulus m. As a final step of the Bayesian model design on NRF, we consider independence of NRF between all pixels and stimulus types. The complete prior distribution of NRF for the set of all NRF $\boldsymbol{r} = \{\boldsymbol{r}_j^m\}_{j=1:J,\ m=1:M}$ conditioned on $\boldsymbol{z} = \{z_j^m\}_{j=1:J,\ m=1:M}$ is thus given as:

$$p(\boldsymbol{r}|\boldsymbol{z}) = \prod_{j=1}^{J}\prod_{m=1}^{M} p(\boldsymbol{r}_j^m|z_j^m). \tag{3-20}$$

3-2-4 Prior on Activation Class

To detect the neural activations in the brain, we use the same model as used in [7] (see also Section 2-7-6), namely, Ising model. Although the prior model remains the same, we shall provide brief summary of the Ising model here for completeness.

We consider two activation classes of neural populations in the brain: "active" and "non-active". Specifically, suppose $z_j^m \in \{0, 1\}$ denotes the activation class label at pixel j for stimulus m. When the neural population at pixel j is activated in response to stimulus m, $z_j^m = 1$. On the other hand, $z_j^m = 0$ denotes that the neural population is not activated for stimulus m.

The physiological knowledge used for the activation class z_j^m is the regional homogeneity hypothesis, in which the neural populations in a given region are likely to share the same

activation state [27, 24, 47]. The neural activation class for a set $\boldsymbol{z}^m = \{z_j^m\}_{j=1:J}$ is modeled using the Ising model as:

$$p(\boldsymbol{z}^m; \beta_m) = \frac{1}{Z(\beta_m)} \exp\left(\beta_m \sum_{j=1}^{J} \sum_{i \in \delta_j} I(z_i^m = z_j^m)\right) \tag{3-21}$$

where β_m is the spatial regularization parameter and $I(\cdot)$ is the indicator function. δ_j denotes the set of all neighboring pixels of pixel j and thus, $\sum_{i \in \delta_j}$ denotes summation over all neighboring pixels of pixel j. In the Ising model, we consider 4-neighbor interactions in 2D lattice.

The indicator function in the exponents accounts for the number of neighboring pixels in the same activation class. Because of this, when more neighboring pixels share the same activation class, the probability of such configuration becomes higher, promoting regional homogeneity.

The strength of the regional homogeneity is controlled by β_m. For instance, a large β_m encourages the neighboring pixels to be in the same activation class. In general, the value $\beta_m \approx 0.88$ is known as a critical value of the Ising model, at which the activation pattern undergoes from disordered pattern to coherent pattern [34].

The normalization constant $Z(\beta_m) = \sum_{\boldsymbol{z}^m} \exp\left(\beta_m \sum_{j=1}^{J} \sum_{i \in \delta_j} I(z_i^m = z_j^m)\right)$ is the summation of probabilities over all possible configurations of $\boldsymbol{z}^m$. In practice, this normalization constant is intractable even for relatively small 2D image as the number of possible configuration grows exponentially, i.e., 2^J, with the number of pixels J.

As a final remark, the following expression of the Ising model is also commonly used in literature:

$$p(\boldsymbol{z}^m; \bar{\beta}_m) = \frac{1}{\bar{Z}(\bar{\beta}_m)} e^{\bar{\beta}_m \sum_{j \sim i} I(z_i^m = z_j^m)} \tag{3-22}$$

where $\sum_{j \sim i}$ denotes the summation over all possible neighboring pairs and critical value in this case is $\bar{\beta}_m \approx \frac{1}{2}\beta_m = 0.44$. We present both expressions in Equations 3-21 and 3-22 to avoid confusion for readers who may be more familiar with the alternative expression shown above. Both formulations are equivalent since:

$$p(\boldsymbol{z}^m; \beta_m) = \frac{1}{\bar{Z}(\bar{\beta}_m)} \exp\left(\bar{\beta}_m \sum_{j \sim i} I(z_i^m = z_j^m)\right) \tag{3-23}$$

$$= \frac{1}{\bar{Z}(\bar{\beta}_m)} \exp\left(\frac{1}{2}\bar{\beta}_m \sum_{j=1}^{J} \sum_{i \in \delta_j} I(z_i^m = z_j^m)\right) \tag{3-24}$$

$$= \frac{1}{\bar{Z}(2\beta_m)} \exp\left(\frac{1}{2} \cdot 2\beta_m \sum_{j=1}^{J} \sum_{i \in \delta_j} I(z_i^m = z_j^m)\right) \tag{3-25}$$

$$= \frac{1}{Z(\beta_m)} \exp\left(\beta_m \sum_{j=1}^{J} \sum_{i \in \delta_j} I(z_i^m = z_j^m)\right), \tag{3-26}$$

where $\frac{1}{2}$ is due to compensate for the double counting of edges between neighboring pixels.

3-2-5 Prior on Spatial Regularization Parameter

The earlier work on JDE in [7] assumes that the spatial regularization parameter β_m of the Ising model is a fixed model parameter. In our JDE model, we consider β_m as a latent variable. Based on the study of the Ising model in [34], we assume the independence of β_m between stimulus types and impose a uniform prior distribution on β_m for all $m = 1, 2, \dots M$. The resulting prior distribution for the set $\boldsymbol{\beta} = \{\beta_m\}_{m=1:M}$ is given by:

$$p(\boldsymbol{\beta}) = \prod_{m=1}^{M} p(\beta_m) = \prod_{m=1}^{M} \mathcal{U}(u_1, u_2) \tag{3-27}$$

where $\mathcal{U}(\cdot, u_1, u_2)$ denotes a uniform distribution with lower bound u_1 and upper bound u_2. As regards the values of the bounds, $u_1 = 0.8$ and $u_2 = 2$ are considered in this study. These specific values of the lower and upper bounds are chosen to enforce the spatial regularization to stay around the critical value ≈ 0.88 [34].

3-2-6 Summary

Our hierarchical Bayesian model is summarized as follows:

$$\beta_m; u_1, u_2 \sim \mathcal{U}(u_1, u_2) \qquad \text{for } m = 1, 2, \dots M \tag{3-28}$$

$$\boldsymbol{z}^m | \beta_m \sim \frac{1}{Z(\beta_m)} e^{\beta_m \sum_{j \sim i} I(z_i^m = z_j^m)} \qquad \text{for } m = 1, 2, \dots M \tag{3-29}$$

$$\boldsymbol{r}_j^m | z_j^m; \boldsymbol{\Sigma_{m1}}, \boldsymbol{\Sigma_{m0}} \sim \mathcal{N}(\mathbf{0}, \Sigma_{m1})^{I(z_j^m = 1)} + \mathcal{N}(\mathbf{0}, \Sigma_{m0})^{I(z_j^m = 0)} \qquad \text{for } m = 1, 2, \dots M \text{ and } j = 1, 2, \dots J \tag{3-30}$$

$$\boldsymbol{h}; \sigma_h^2, \boldsymbol{R} \sim \mathcal{N}(\boldsymbol{\mu}_{h0}, \sigma_h^2 \boldsymbol{R}) \tag{3-31}$$

$$\boldsymbol{y}_j | \boldsymbol{h}, \boldsymbol{r}_j; \boldsymbol{X}, \boldsymbol{P}, \boldsymbol{\ell_j}, \sigma_j^2 \sim \mathcal{N}\left(\sum_{m=1}^{M} \boldsymbol{X}_m \boldsymbol{H} \boldsymbol{r}_j^m + \boldsymbol{P} \boldsymbol{\ell}_j, \sigma_j^2 \boldsymbol{I}_N \right) \tag{3-32}$$

$$\sim \mathcal{N}\left(\sum_{m=1}^{M} \boldsymbol{X}_m \boldsymbol{R}_j^m \boldsymbol{h} + \boldsymbol{P} \boldsymbol{\ell}_j, \sigma_j^2 \boldsymbol{I}_N \right) \qquad \text{for } j = 1, 2, \dots J \tag{3-33}$$

The corresponding complete data likelihood is:

$$p(\boldsymbol{y}, \boldsymbol{h}, \boldsymbol{r}, \boldsymbol{z}, \boldsymbol{\beta}; \boldsymbol{\Theta}) = p(\boldsymbol{y}|\boldsymbol{h}, \boldsymbol{r}) p(\boldsymbol{h}) p(\boldsymbol{r}|\boldsymbol{z}) p(\boldsymbol{z}|\boldsymbol{\beta}) p(\boldsymbol{\beta}). \tag{3-34}$$

Here, it should be noted that the model parameters are omitted for compactness.

3-3 Variational Expectation-Maximization Estimation

In this section, we present the iterative update rules for our Bayesian model under the VEM framework. As described in Section 2-5, the latent variables and some model parameters are estimated by maximizing the evidence lower bound (ELBO) of the Bayesian model. The estimates include the sets of latent variables $\boldsymbol{r}$, $\boldsymbol{h}$, $\boldsymbol{z}$, and $\boldsymbol{\beta}$, as well as the sets of model parameters $\boldsymbol{\ell}$, σ_h^2, and σ^2.

3-3-1 Mean Field Approximation

We begin by specifying the form of the variational posterior distribution over the latent variables. Using the mean-field approximation, the variational posterior is factorized as:

$$q(\boldsymbol{h}, \boldsymbol{r}, \boldsymbol{z}, \boldsymbol{\beta}) = q(\boldsymbol{h})q(\boldsymbol{r})q(\boldsymbol{z})q(\boldsymbol{\beta}) \tag{3-35}$$

$$= \prod_{j=1}^{J} q(\boldsymbol{r}_j) \prod_{m=1}^{M} q(z_j^m)q(\beta_m)q(\boldsymbol{h}) \tag{3-36}$$

Equation 3-35 assumes independence among the sets of latent variables. In Equation 3-36, we further assume independence among pixels and stimulus types. For the posterior activation class $q(z_j^m)$, we assume that the label at pixel j and stimulus m is independent of other labels. Similarly, for $q(\beta_m)$, we assume that the spatial regularization parameter is independent across stimulus types.

However, for $q(\boldsymbol{r})$, we only assume independence across pixels and we allow dependence across stimulus types within the same pixel, similar to the JDE model in [7]. This allows our model to capture correlations in the NRF across stimuli.

As described in 2-3, the mean-field approximation is used to avoid intractable computation during estimation via optimization, at the expense of reduced accuracy in the posterior estimation. In this case, breaking independence between shapes NRF and HRF is more physiologically sound since HRF depends on the structure of the surrounding vascular systems while NRF depends on neuronal types. However, removing dependency among pixels tends to underrepresent the posterior distribution. This is because closely spaced non-neighboring neuronal populations are likely to be correlated.

3-3-2 ELBO and Update Strategy

Let $\boldsymbol{\theta} = \{\boldsymbol{\ell}, \sigma_h^2, \sigma^2\}$ denote the set of model parameters to be estimated. The ELBO under the factorized variational posterior is:

$$\mathcal{L}\left(q(\boldsymbol{h})q(\boldsymbol{r})q(\boldsymbol{z})q(\boldsymbol{\beta}), \boldsymbol{\theta}\right) = \int q(\boldsymbol{h})q(\boldsymbol{r})q(\boldsymbol{z})q(\boldsymbol{\beta}) \ln \frac{p(\boldsymbol{y}, \boldsymbol{h}, \boldsymbol{r}, \boldsymbol{z}, \boldsymbol{\beta}; \boldsymbol{\theta})}{q(\boldsymbol{h})q(\boldsymbol{r})q(\boldsymbol{z})q(\boldsymbol{\beta})} d\boldsymbol{h}\ d\boldsymbol{r}\ d\boldsymbol{z}\ d\boldsymbol{\beta} \tag{3-37}$$

$$= \mathbb{E}_{q(\boldsymbol{h})q(\boldsymbol{r})q(\boldsymbol{z})q(\boldsymbol{\beta})}[\ln p(\boldsymbol{y}, \boldsymbol{h}, \boldsymbol{r}, \boldsymbol{z}, \boldsymbol{\beta}; \boldsymbol{\theta})] - \mathbb{E}_{q(\boldsymbol{h})q(\boldsymbol{r})q(\boldsymbol{z})q(\boldsymbol{\beta})}[\ln q(\boldsymbol{h})q(\boldsymbol{r})q(\boldsymbol{z})q(\boldsymbol{\beta})], \tag{3-38}$$

where $\mathbb{E}_{q(\cdot)}$ denotes taking the expectation of a function with respect to the variational posterior.

Let $\boldsymbol{\theta}^{(t)}$ and $q(\cdot)^{(t)}$ denote estimates at iteration t. The VEM algorithm alternates between optimizing the ELBO with respect to the variational distributions (E-step) and the model parameters (M-step), as outlined below:

Algorithm 1 Variational EM Algorithm

1: **Initialize:** $\boldsymbol{\theta}^{(0)}, q^{(0)}(\boldsymbol{h}), q^{(0)}(\boldsymbol{r}), q^{(0)}(\boldsymbol{z}), q^{(0)}(\boldsymbol{\beta})$
2: **for** $t = 1, 2 \dots$ until ELBO converges **do**
3: **E-step:**

- Update-$\boldsymbol{r}$:

$$\ln q^{(t)}(\boldsymbol{r}) \stackrel{c}{=} \mathbb{E}_{q^{(t-1)}(\boldsymbol{h})q^{(t-1)}(\boldsymbol{z})q^{(t-1)}(\boldsymbol{\beta})} \left[\ln p(\boldsymbol{y}, \boldsymbol{h}, \boldsymbol{r}, \boldsymbol{\beta}; \boldsymbol{\theta}^{(t-1)})\right] \tag{3-39}$$

- Update-$\boldsymbol{h}$:

$$\ln q^{(t)}(\boldsymbol{h}) \stackrel{c}{=} \mathbb{E}_{q^{(t)}(\boldsymbol{r})q^{(t-1)}(\boldsymbol{z})q^{(t-1)}(\boldsymbol{\beta})} \left[\ln p(\boldsymbol{y}, \boldsymbol{h}, \boldsymbol{r}, \boldsymbol{\beta}; \boldsymbol{\theta}^{(t-1)})\right] \tag{3-40}$$

- Update-$\boldsymbol{z}$:

$$\ln q^{(t)}(\boldsymbol{z}) \stackrel{c}{=} \mathbb{E}_{q^{(t)}(\boldsymbol{h})q^{(t)}(\boldsymbol{r})q^{(t-1)}(\boldsymbol{\beta})} \left[\ln p(\boldsymbol{y}, \boldsymbol{h}, \boldsymbol{r}, \boldsymbol{\beta}; \boldsymbol{\theta}^{(t-1)})\right] \tag{3-41}$$

- Update-$\boldsymbol{\beta}$:

$$\ln q^{(t)}(\boldsymbol{\beta}) \stackrel{c}{=} \mathbb{E}_{q^{(t)}(\boldsymbol{h})q^{(t)}(\boldsymbol{r})q^{(t)}(\boldsymbol{z})} \left[\ln p(\boldsymbol{y}, \boldsymbol{h}, \boldsymbol{r}, \boldsymbol{\beta}; \boldsymbol{\theta}^{(t-1)})\right] \tag{3-42}$$

4: **M-step:**

$$\boldsymbol{\theta}^{(t)} = \arg\max_{\boldsymbol{\theta}} \mathbb{E}_{q^{(t)}(\boldsymbol{h})q^{(t)}(\boldsymbol{r})q^{(t)}(\boldsymbol{z})q^{(t)}(\boldsymbol{\beta})} \left[\ln p(\boldsymbol{y}, \boldsymbol{h}, \boldsymbol{r}, \boldsymbol{\beta}; \boldsymbol{\theta})\right] \tag{3-43}$$

5: **end for**

Each E-step update maximizes the ELBO with respect to one variational distribution while keeping the others fixed. The M-step maximizes the ELBO with respect to the model parameters $\boldsymbol{\theta}$ using the current variational distributions.

In the following, we present the update equations by evaluating the expectations in Equations 3-39–3-42 and 3-43, which yields closed-form solutions for each variational distribution and model parameter. Especially, we will observe that each one of update equations for $q(\boldsymbol{h})$ and $q(\boldsymbol{r})$ reduces to a closed-form solution given by variational parameters in the form of Gaussian distributions. This is simply because we use Gaussian priors for $\boldsymbol{h}$ and $\boldsymbol{r}$, which are conjugate prior to the Gaussian likelihood with an unknown mean. Full derivations can be found in Appendix A. Regarding the ELBO evaluated at each iteration in Algorithm 1, its computation can be found in Appendix B. The source code for the implementation of this algorithm is available at Github. For the sake of brevity, we omit the superscript (t) and the irrelevant model parameters.

Update of $\boldsymbol{r}_j$: Each latent response function $\boldsymbol{r}_j$ follows a Gaussian distribution:

$$q(\boldsymbol{r}_j) = \mathcal{N}(\boldsymbol{\mu}_{r_j}, \boldsymbol{\Sigma}_{r_j}) \tag{3-44}$$

with the posterior mean $\boldsymbol{\mu}_{r_j}$ and covariance $\boldsymbol{\Sigma}_{r_j}$

$$\boldsymbol{\Sigma}_{r_j} = \left(\frac{1}{\sigma_j^2} \mathbb{E}_{q(\boldsymbol{h})} \left[\boldsymbol{G}^\top \boldsymbol{G} \right] + \sum_{k=1}^{K} \boldsymbol{\Phi}_{jk} \boldsymbol{\Sigma}_k^{-1} \right)^{-1} \in \mathbb{R}^{ML_r \times ML_r} \tag{3-45}$$

and

$$\boldsymbol{\mu}_{r_j} = \boldsymbol{\Sigma}_{r_j} \frac{1}{\sigma_j^2} \mathbb{E}_{q(\boldsymbol{h})}[\boldsymbol{G}]^\top \left(\boldsymbol{y}_j - \boldsymbol{P}\boldsymbol{\ell}_j \right) \in \mathbb{R}^{ML_r}. \tag{3-46}$$

Here, $\boldsymbol{r}_j$ is a vertical stacking of $\boldsymbol{r}_j^m$, defined as

$$\boldsymbol{r}_j = \begin{bmatrix} r_j^1 \\ r_j^2 \\ \vdots \\ r_j^M \end{bmatrix} \in \mathbb{R}^{ML_r}, \tag{3-47}$$

and $\boldsymbol{G}$ is defined as

$$\boldsymbol{G} = \begin{bmatrix} \boldsymbol{X}_1 \boldsymbol{H} & \boldsymbol{X}_2 \boldsymbol{H} & \cdots & \boldsymbol{X}_M \boldsymbol{H} \end{bmatrix} \in \mathbb{R}^{N \times (ML_r)} \tag{3-48}$$

for shorthand notation. $\phi_{jk}^m = q(z_j^m = k)$ represents posterior probability of pixel j belonging to activation class z_j^m for stimulus m. The other symbols are defined as:

$$\boldsymbol{\Sigma}_k^{-1} = \begin{bmatrix} \boldsymbol{\Sigma}_{k1}^{-1} & & & \\ & \boldsymbol{\Sigma}_{k2}^{-1} & & \\ & & \ddots & \\ & & & \boldsymbol{\Sigma}_{kM}^{-1} \end{bmatrix} \in \mathbb{R}^{ML_r \times ML_r}, \tag{3-49}$$

and

$$\boldsymbol{\Phi}_{jk} = \begin{bmatrix} \phi_{jk}^1 \boldsymbol{I}_{L_r} & & & \\ & \phi_{jk}^2 \boldsymbol{I}_{L_r} & & \\ & & \ddots & \\ & & & \phi_{jk}^M \boldsymbol{I}_{L_r} \end{bmatrix} \in \mathbb{R}^{ML_r \times ML_r}, \tag{3-50}$$

where $\boldsymbol{I}_{(a)}$ is the identity matrix of the size $a \times a$.

Update of $\boldsymbol{h}$: The posterior of HRF also follows a Gaussian variational posterior:

$$q(\boldsymbol{h}) = \mathcal{N}(\boldsymbol{\mu}_h, \boldsymbol{\Sigma}_h) \tag{3-51}$$

with closed-form expressions for $\boldsymbol{\mu}_h$ and $\boldsymbol{\Sigma}_h$ given by:

$$\boldsymbol{\Sigma}_h = \left(\sum_{j=1}^{J} \frac{1}{\sigma_j^2} \mathbb{E}_{q(\boldsymbol{r}_j^m)} \left[\left(\sum_{m=1}^{M} \boldsymbol{X}_m \boldsymbol{R}_j^m \right)^\top \left(\sum_{m=1}^{M} \boldsymbol{X}_m \boldsymbol{R}_j^m \right) \right] + \frac{1}{\sigma_h^2} \boldsymbol{R}^{-1} \right)^{-1} \tag{3-52}$$

and

$$\boldsymbol{\mu}_h = \boldsymbol{\Sigma}_h \left(\sum_{j=1}^{J} \frac{1}{\sigma_j^2} \left(\sum_{m=1}^{M} \boldsymbol{X}_m \mathbb{E}_{q(\boldsymbol{r}_j^m)} \left[\boldsymbol{R}_j^m \right] \right)^\top \left(\boldsymbol{y}_j - \boldsymbol{P}\boldsymbol{\ell}_j \right) + \frac{1}{\sigma_h^2} \boldsymbol{R}^{-1} \boldsymbol{\mu}_{h0} \right). \tag{3-53}$$

Update of z_j^m: The posterior of the activation class label z_j^m at each pixel-stimulus pair is given by a categorical distribution:

$$q(z_j^m = k) \propto \boldsymbol{\mathcal{N}}(\boldsymbol{\mu}_{r_j^m}; \mathbf{0}, \boldsymbol{\Sigma}_{km}) \exp\left(-\mathrm{tr}(\boldsymbol{\Sigma}_{km}^{-1}\boldsymbol{\Sigma}_{r_j^m}) + \sum_{i\in\delta_j} \phi_{ik}^m \mu_{\beta_m}\right). \tag{3-54}$$

Note that this is an unnormalized distribution. The normalization constant can be computed by simply summing over all k.

Update of β_m: Each spatial regularization parameter β_m is updated through

$$q(\beta_m) \propto \prod_{j=1}^{J} \frac{\exp\left(\beta_m \sum_{j=1}^{J} \sum_{i\in\delta_j} \sum_{k=1}^{K} \phi_{jk}^m \phi_{ik}^m\right)}{\sum_{k=1}^{K} \exp\left(\beta_m \sum_{i\in\delta_j} \phi_{ik}^m\right)} \mathcal{U}(\beta_m, u_1, u_2) \tag{3-55}$$

Note that this is not in the closed-form. To compute mean μ_{β_m} and normalization constant, we use numerical integration in the interval (u_1, u_2).

Update of model parameters $\boldsymbol{\theta} = \{\boldsymbol{\ell}, \sigma_h^2, \sigma^2\}$: In the M-step, the parameters are estimated by maximizing the expected complete-data log-likelihood with respect to the current variational distributions:

$$\boldsymbol{\ell} = \arg\max_{\boldsymbol{\ell}} \ \mathbb{E}_{q(\boldsymbol{h})q(\boldsymbol{r})q(\boldsymbol{z})q(\boldsymbol{\beta})}[\ln p(\boldsymbol{y}, \boldsymbol{h}, \boldsymbol{r}, \boldsymbol{z}, \boldsymbol{\beta}; \boldsymbol{\theta})] \tag{3-56}$$

$$\sigma_h^2 = \arg\max_{\sigma_h^2} \ \mathbb{E}_{q(\boldsymbol{h})q(\boldsymbol{r})q(\boldsymbol{z})q(\boldsymbol{\beta})}[\ln p(\boldsymbol{y}, \boldsymbol{h}, \boldsymbol{r}, \boldsymbol{z}, \boldsymbol{\beta}; \boldsymbol{\theta})] \tag{3-57}$$

$$\sigma^2 = \arg\max_{\sigma^2} \ \mathbb{E}_{q(\boldsymbol{h})q(\boldsymbol{r})q(\boldsymbol{z})q(\boldsymbol{\beta})}[\ln p(\boldsymbol{y}, \boldsymbol{h}, \boldsymbol{r}, \boldsymbol{z}, \boldsymbol{\beta}; \boldsymbol{\theta})] \tag{3-58}$$

By evaluating expectation, each of these leads to a closed-form update solution:

$$\boldsymbol{\ell}_j = \boldsymbol{W}^{-1}\boldsymbol{b} = \left(\boldsymbol{P}^\top \boldsymbol{P}\right)^{-1} \boldsymbol{P}^\top \left(\boldsymbol{y}_j - \mathbb{E}_{q(\boldsymbol{h})}\left[\boldsymbol{G}\right] \boldsymbol{\mu}_{r_j}\right), \tag{3-59}$$

$$\sigma_j^2 = \frac{(\boldsymbol{y}_j - 2\mathbb{E}_{q(\boldsymbol{h})}[\boldsymbol{G}]\boldsymbol{\mu}_{r_j} - \boldsymbol{P}\boldsymbol{\ell}_j)^\top (\boldsymbol{y}_j - \boldsymbol{P}\boldsymbol{\ell}_j) + \boldsymbol{\mu}_{r_j}^\top \mathbb{E}_{q(\boldsymbol{h})}[\boldsymbol{G}^\top \boldsymbol{G}]\boldsymbol{\mu}_{r_j} + \mathrm{tr}\left(\mathbb{E}_{q(\boldsymbol{h})}[\boldsymbol{G}^\top \boldsymbol{G}]\boldsymbol{\Sigma}_{r_j}\right)}{N}, \tag{3-60}$$

$$\sigma_h^2 = \frac{-L_h + \sqrt{8\lambda(\boldsymbol{\mu}_h - \boldsymbol{\mu}_{h0})^\top \boldsymbol{R}^{-1}(\boldsymbol{\mu}_h - \boldsymbol{\mu}_{h0}) + 8\lambda \ \mathrm{tr}\left(\boldsymbol{R}^{-1}\boldsymbol{\Sigma}_h\right) + (L_h)^2}}{4\lambda}. \tag{3-61}$$

3-4 Practicalities

As a final note in this chapter, we briefly address an important practical consideration regarding the choice of the scaling parameters in the Gaussian priors of $\boldsymbol{h}$ and $\boldsymbol{r}_j^m$, particularly when applying the proposed JDE algorithm (Algorithm 1) to real fUS data.

As discussed in Sections 3-2-2 and 3-2-3, the scaling parameters σ_h^2 from the HRF and λ_r from the NRF determine the strength of our prior belief relative to the data. In particular, they control the influence of the prior means and the overall amplitudes of $\boldsymbol{h}$ and $\boldsymbol{r}_j^m$. In practice, their values must be chosen with the noise variance σ_j^2 in mind.

When σ_j^2 is large and the scaling parameters are small (i.e., $\sigma_h^2 \ll \sigma_j^2$), the posterior updates for $\boldsymbol{h}$ and $\boldsymbol{r}_j^m$ are dominated by the prior. As a consequence, the posterior estimates of $\boldsymbol{h}$ and $\boldsymbol{r}j^m$ tend to collapse towards the prior mean.

To demonstrate, recall the posterior update equations for the posterior covariance of $\boldsymbol{h}$:

$$\boldsymbol{\Sigma}_h = \left(\sum_{j=1}^{J} \frac{1}{\sigma_j^2} \mathbb{E}_{q(\boldsymbol{r}_j^m)} \left[\left(\sum_{m=1}^{M} \boldsymbol{X}_m \boldsymbol{R}_j^m \right)^\top \left(\sum_{m=1}^{M} \boldsymbol{X}_m \boldsymbol{R}_j^m \right) \right] + \frac{1}{\sigma_h^2} \boldsymbol{R}^{-1} \right)^{-1} \tag{3-62}$$

and the posterior mean

$$\boldsymbol{\mu}_h = \boldsymbol{\Sigma}_h \left(\sum_{j=1}^{J} \frac{1}{\sigma_j^2} \left(\sum_{m=1}^{M} \boldsymbol{X}_m \mathbb{E}_{q(\boldsymbol{r}_j^m)} \left[\boldsymbol{R}_j^m \right] \right)^\top \left(\boldsymbol{y}_j - \boldsymbol{P}\boldsymbol{\ell}_j \right) + \frac{1}{\sigma_h^2} \boldsymbol{R}^{-1} \boldsymbol{\mu}_{h0} \right). \tag{3-63}$$

When $\sigma_h^2 \ll \sigma_j^2$, the data-dependent terms become negligible. As a result, we have $\boldsymbol{\Sigma}_h \approx \sigma_h^2 \boldsymbol{R}$ and $\boldsymbol{\mu}_h \approx \boldsymbol{\mu}_{h0}$, meaning that the posterior is driven mostly by the prior. If the prior mean $\boldsymbol{\mu}_{h0}$ is set to zero, the posterior HRF will also be close to zero regardless of the data. Furthermore, since the amplitude of $\boldsymbol{r}_j^m$ directly influences the data-dependent term in both equations, the collapse of one of these variables reinforces the collapse of the other.

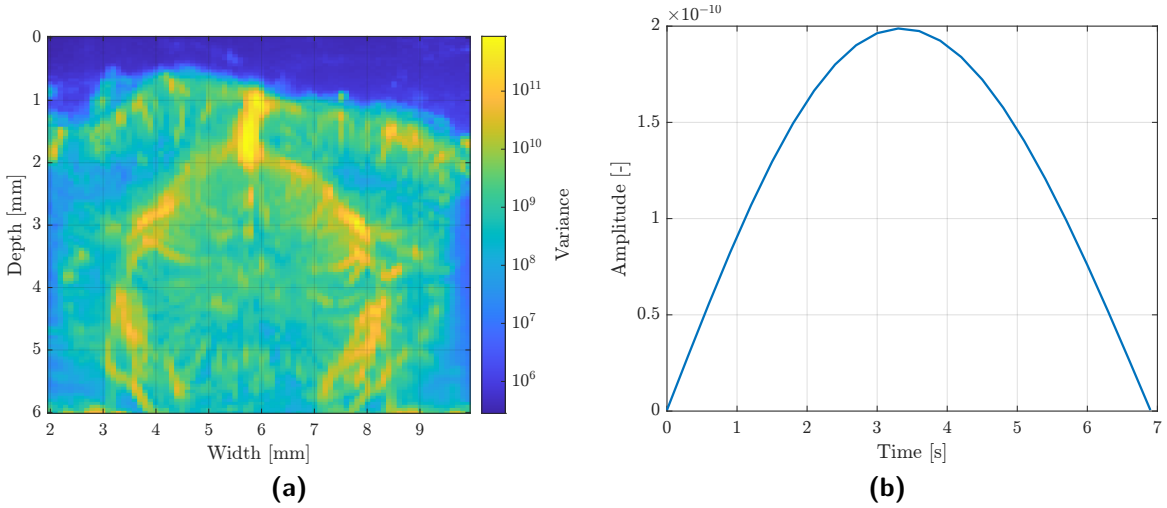


Figure 3-5: (a): Variance of PDI signals used in this thesis. The average variance over 2D slices is approximately 7.7×10^9. (b): Estimated HRF with $\sigma_h^2 = 10$, $\lambda_r = 1$ and a zero-mean Gaussian prior. This demonstrates the posterior collapse.

This issue is especially relevant for the real fUS dataset used in this thesis as the variances range from 10^5 to over 10^{11}, as shown in Figure 3-5a. Figure 3-5b illustrates the consequence of choosing $\sigma_h^2 = 10$ and $\lambda_r = 1$, with the zero-mean priors and an initial half sine waveform for the HRF. As can be seen, the posterior HRF estimate collapses to near-zero.

To address this, we choose larger values for the scaling parameters in our real data experiments (see Chapter 5), specifically $\sigma_h^2 = 100$ and $\lambda_r = 100$ to ensure that the model remains sensitive to the observed data.

In summary, when dealing with high noise variance, it is essential to scale the prior variances accordingly to avoid the collapse of the posterior estimates. This also ensures that the model remains data-driven. Moreover, as the $\boldsymbol{h}$ and $\boldsymbol{r}_j^m$ influence each other through the update equations, careful tuning of both σ_j^2 and λ_r is crucial.

Chapter 4

Numerical Experiments

This chapter presents and discusses the results from the proposed cascade convolution model of Chapter 3 using synthetic data. The goal is to validate the model's ability to recover latent neural and hemodynamic response functions and detect activated pixels. We begin by assessing the convergence behavior and estimation quality across different datasets in Section 4-2. We then examine the model's robustness to varying noise levels in Section 4-3, followed by an analysis of how prior scaling parameters affect estimation performance in Section 4-4.

4-1 Experimental Setup

We consider a spatial grid of size 20×20 ($J = 400$ pixels) and a temporal length $N = 240$. For the external stimuli, we use two stimuli ($M = 2$), each modeled as a sequence of Dirac-delta functions representing stimulus onset times.

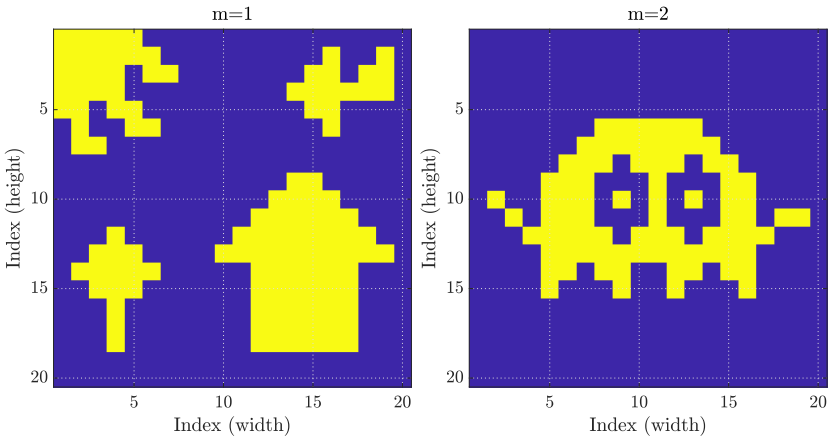
The lengths of the HRF and NRF are set to $L_h = 35$ and $L_r = 15$, respectively.

For the synthetic data, we generated three distinct data sets. Figure 4-1a shows two activation patterns obtained from the PyHRF Python library, which are used consistently across all three data sets. In these activation maps, yellow pixels represent active regions for the corresponding stimulus, while blue pixels indicate inactive regions. Note that some pixels may be active for both stimuli.

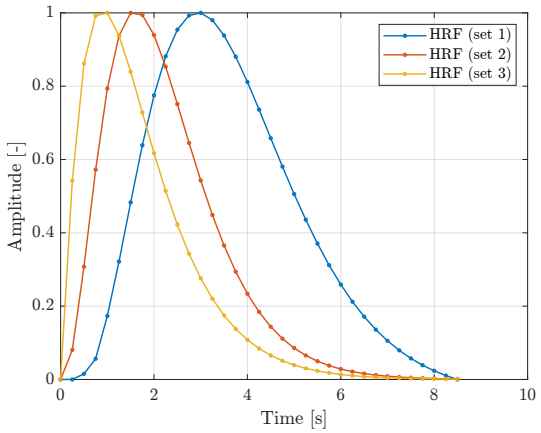
Figure 4-1b shows the unique HRF used in each of the three synthetic data sets.

For the NRF, we first define a reference NRF trace for each stimulus, shown as solid lines in Figure 4-1c. Then, for each active pixel, we generate its specific NRF by adding iid Gaussian noise with a standard deviation of 0.015 to the corresponding reference trace (shown as transparent lines in Figure 4-1c). Although the same reference traces are used across all data sets, the pixel-specific NRFs are independently generated for each set.

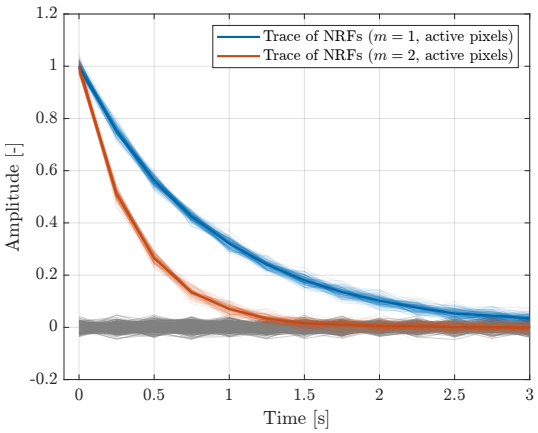
Stimulus onset times are randomly generated for each data set. Each stimulus sequence contains approximately 24 onsets.



(a) Ground truth activation patterns for two stimuli: activation patterns for stimulus 1 ($m = 1$)(Left) and stimulus 2 ($m = 2$) (Right). Active pixels ($z_j^m = 1$) for each stimulus are colored yellow and non-active pixels ($z_j^m = 0$) are colored blue.



(b) Ground truth HRFs used for three different synthetic data sets.



(c) Ground truth NRFs of all 400 pixels from Set 1. Two solid lines are used to generate $\boldsymbol{r}_j^m$ for pixels that are active for each stimulus by adding iid zero-mean Gaussian noise with standard deviation of 0.015. $\boldsymbol{r}_j^m$ for non-active pixels are generated from the same Gaussian distribution. The gray traces represent the NRFs of non-active pixels.

Figure 4-1: Ground truth of latent variables

In all numerical experiments, we ignore slow drift terms by setting $\boldsymbol{P}\boldsymbol{\ell}_j = \boldsymbol{0}$ for all j. For model initialization, the posterior probability of activation assignment is set uniformly: $q(z_j^m = 1) = q(z_j^m = 0) = 0.5$ for all j and m, meaning that the probability of a pixel being active or inactive under each stimulus is initially equal for all pixels. The HRF is initialized as a half sine wave spanning the entire HRF length. The spatial regularization parameter $\boldsymbol{\beta}_m$ is initialized at 0.88.

The full set of model parameters used in this experiment is listed in Tables 4-1 and 4-2.

Table 4-1: Model parameters and initializations

Model parameter	Value	Description
Δt	0.25s	Sampling period of synthetic data (= 4Hz)
σ_j	$2\sigma_{\text{ground truth}}$	Initial noise variance
λ_r	1	Scaling factor of second-order stable-spline kernel
α_r	0.84	Decay factor of second-order state-spline kernel
σ_0^2	1.3×10^{-4}	Scaling parameter of the prior NRF for non-active pixels.
σ_h	1	Initial scaling factor for HRF prior
$\boldsymbol{\mu}_{h0}$	$\mathbf{0}$	Mean of HRF prior
u_1	0.6	Lower bound of uniform distribution on $\beta_m \; \forall \; m$
u_2	2	Upper bound of uniform distribution on $\beta_m \; \forall \; m$

Table 4-2: Variational parameters and initializations

Variational parameter	Value	Description
$\boldsymbol{\mu}_h$	$\mathbf{0}$	Initial posterior mean of HRF
$\boldsymbol{\Sigma}_h$	$\boldsymbol{I}_{L_h}$	Initial posterior variance of HRF
$q(z_j^m)$	$0.5 \; \forall \; m, \; j$	Initial posterior activation class assignment weight
β_m	$0.88 \; \forall \; m$	Initial spatial regularization parameter

4-2 Convergence

In this section, we demonstrate the convergence of estimations obtained by our proposed model. Measurement noise is added to the synthetic data as zero-mean Gaussian noise distributed as $\boldsymbol{\mathcal{N}}(\mathbf{0}, \sigma_j^2 \boldsymbol{I}_N)$. In this experiment, we assume a signal-to-noise ratio (SNR) of 10dB. Since the HRF in each synthetic data set has a different response gain (i.e., area under the curve), the noise variances are set to 2.76, 1.36, and 1.04 for Sets 1, 2, and 3, respectively. The noise variance is constant across all pixels within each data set.

The posterior probability map (PPM) for both stimuli ($m = 1, 2$) across all data sets is shown in the left two columns of Figure 4-2. Here, the PPM represents the estimated posterior probability of each pixel being in the "active" state, i.e., $q(z_j^m = 1)$.

In Set 3, the estimated PPM perfectly matches the ground truth activation patterns. In Set 2, there is only a single false detection at location (height, width) = $(13, 4)$. In contrast, Set 1 shows a significantly higher number of false detections, totaling 85 misclassified pixels. Since these activation class assignments condition the prior distribution of the NRF, a higher misclassification rate can degrade the quality of the NRF estimates. Figure 4-2 displays the posterior means of the NRFs for the detected active pixels (transparent lines), along with the average NRF across these means for each stimulus (solid blue line for stimulus 1 and solid red line for stimulus 2). In line with the detection estimate, the average NRFs from Set 2 and 3 closely resemble the ground truth traces. For Set 1, on the other hand, they deviate from the ground truth due to the large number of false detections. In fact, the NRFs from the misclassified pixels of Set 1 correspond to those that are around zero in the plot.

One potential cause for these misclassifications lies in the factor $\boldsymbol{\mathcal{N}}(\boldsymbol{\mu}_{r_j^m} | \mathbf{0}, \boldsymbol{\Sigma}_{m \; z_j^m})$ in the update of $q(z_j^m)$ (Equation 3-54). Because NRF priors for both active and non-active classes are zero-mean, their probability mass overlap substantially. As a result, the spatial coherence

term can dominate the update step, causing non-active pixels to be misclassified as active. Moreover, as the NRFs from these misclassified pixels remain close to zero, the contribution from observed data may not be strong enough to avoid direction of this local maximum induced by spatial coherence.

In contrast to the activation class and NRF estimates, the estimated HRFs in Figure 4-2 closely match the ground truth across all sets.

As a remark, it must be noted that the peak amplitudes of HRFs in the plots have been rescaled to 1 and the reciprocal of the same scaling has been applied to the corresponding NRFs. This normalization reflects the inherent scaling ambiguity between the estimated HRF and NRF in blind deconvolution problems.

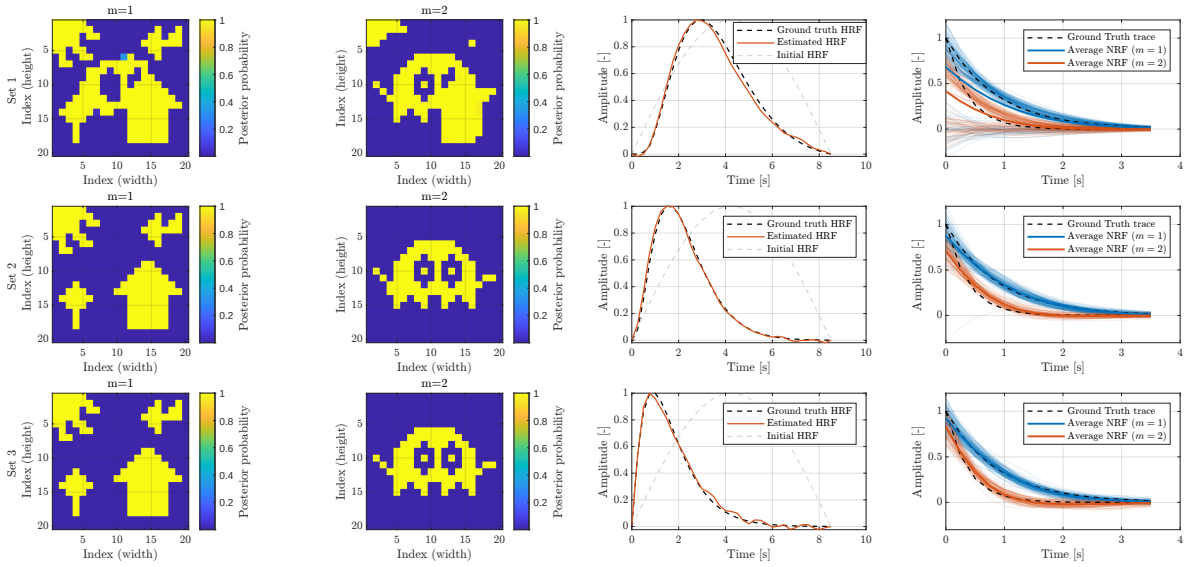


Figure 4-2: Estimation Results for SNR of 10dB

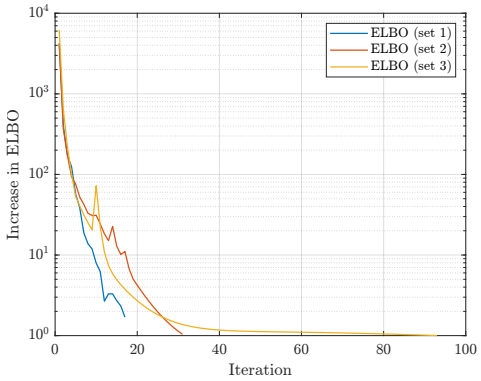


Figure 4-3: Increase in ELBO at each iteration

Finally, Figure 4-3 illustrates the iterative increase of ELBO for each data set. The JDE is terminated when the improvement of ELBO falls below a threshold of 1. Interestingly, Set 1, which yielded the worst estimation quality converged after only iterations of 19, whereas Set 2 and 3 required 34 and 96 iterations, respectively. This early convergence in Set 1 suggests that

the estimation may have been trapped in a bad local maximum, leading to poor estimation results. Similar convergence patters were observed across 10 different realizations, where the average number of iterations to convergence was 19.5 for Set 1, 34.1 for Set 2, and 106.1 for Set 3, confirming the observed trends.

4-3 Robustness to Noise

In this experiment, the generated data from the previous section were used with a range of noise variances in order to investigate the robustness of the proposed method to variations in noise levels.

The noise variances were selected to correspond to SNR values of $20, 15, 10, 5, 0, -5$dB across the datasets. For each dataset, 10 realizations of the noise were added. To isolate the effect of noise from the effect of initialization, the same model parameters and initializations from Tables 4-1 and 4-2 were used for all conditions.

We evaluated the performance of the proposed algorithm by computing the average relative error for the HRF, NRFs, and reconstructed synthetic data, as well as the average number of misclassifications in activation detection.

The relative error between an estimate $\bar{\boldsymbol{A}}$ and the ground truth $\boldsymbol{A}$ is defined as:

$$\text{Relative error } [\%] = \frac{\|\bar{\boldsymbol{A}} - \boldsymbol{A}\|_F}{\|\boldsymbol{A}\|_F} \times 100 \tag{4-1}$$

where $\|\cdot\|_F$ is the Frobenius norm. Regarding the computation of the relative error for NRFs and the reconstructed data, the error is computed by summing the Frobenius norms of residuals over the active pixels.

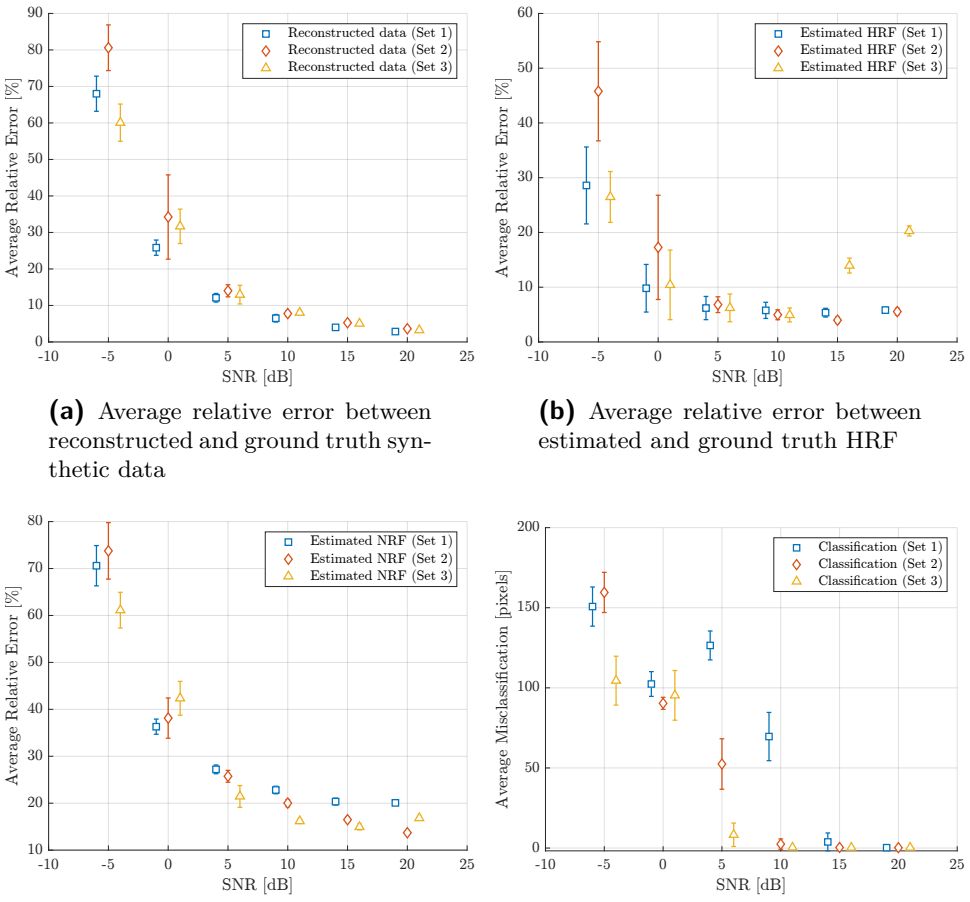
Figure 4-4a illustrates the relative errors of the reconstructed data averaged across noise realizations for each dataset, alongside 2 standard deviation interval. The relative reconstruction error behaves similarly across datasets in the SNR range of 20 to 5dB. At lower SNR levels, the reconstruction error increases rapidly, especially for Set 2. Although we do not have a reference model for direct comparison of noise robustness, assuming a practical SNR range between 5dB and 15dB, the relative reconstruction error remains below 20%, suggesting that our JDE exhibits sufficient robustness to noise for practical datasets.

Figure 4-4b shows the average relative error between the estimated and ground truth HRFs for each dataset. The relative errors for Sets 1 and 2 behave similarly to the reconstruction errors shown in Figure 4-4a. However, Set 3 shows notably larger relative errors at SNR levels of 15dB and 20dB. Figure 4-5 displays the HRF and NRFs estimated at 15dB for Set 3. In this figure, the slope of the estimated HRF after the peak is delayed and shifted to the right compared to the ground truth. The reason of the relatively large relative error is attributed to the bias from the prior distribution of HRF. As seen during the model design of Figure 3-3, peak values of random vectors generated from the HRF prior tend to be concentrated around the middle of a vector. This tendency causes the prior to bias the peak location toward the center. Despite this bias, the reconstruction errors at 15dB and 20dB remain low because the average traces of the NRFs shift slightly to the left, compensating for the delay in the HRF.

As shown in Figure 4-4c, the average relative errors for the NRFs display a similar trend across datasets, resembling the behavior of the reconstruction errors. Notably, the NRFs from Set 1 exhibit higher relative errors in the $5 - 20$dB range.

When comparing Figure 4-4d to Figures 4-4a and 4-4c, it can be seen that the average relative errors for both reconstruction and NRFs remain relatively unaffected by the high number of misclassifications in Sets 1 and 2 at 5dB. This indicates that the misclassified pixels mostly consists of false positive detections.

To sum up, our JDE model demonstrates moderate robustness to noise in capturing the underlying hemodynamic signals over a practical SNR range, as evidenced by the relative reconstruction error plots. The model also estimates the HRF accurately within this range, with average relative errors below 10%, except for Set 3. The NRFs are similarly well estimated. However, the model struggles with distinguishing the activation class of non-active pixels, particularly when the estimated NRF is smooth and close to zero at 5dB and 10dB. Thus, the activation detection capability of the JDE model remains sensitive to noise levels.



(a) Average relative error between reconstructed and ground truth synthetic data

(b) Average relative error between estimated and ground truth HRF

(c) Average relative error between estimated and ground truth NRFs

(d) Average number of misclassified pixels

Figure 4-4: Estimation results for SNR[dB]$= \{20, 15, 10, 5, 0, -5\}$. The bar at each point of the plot represents 2 standard deviation across 10 noise realization of each condition.

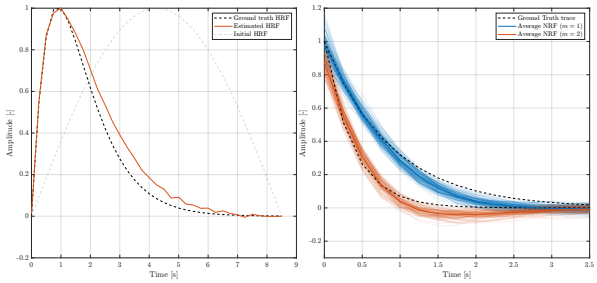


Figure 4-5: Estimated HRF and NRFs from Set 3 at 15dB

4-4 Effect of Scaling Parameters

In this section, we investigate how the scaling parameters λ_r(from the NRF prior) and σ_h^2(from the HRF prior) influence the estimation performance. In specific, we repeated the same experiment again for Set 1 using a common scaling value $c \in \{0.1,\ 1\ 10\}$ for both parameters, i.e., $\lambda_r = \sigma_h^2 = c$. The case $c = 1$ corresponds to the value used in the previous section and serves as a baseline for comparison. Again, 10 realizations of noise were added to the ground truth data for SNR levels $\{20, 15, 10, 5, 0, -5\}$dB and all other model parameters and initializations were kept the same as in Tables 4-1 and 4-2. Performance was evaluated using the average relative error for HRF, NRFs, and reconstructed data, as well as the average number of misclassified pixels in activation detection. Across most SNR levels, relative errors in the reconstructed data ,estimated HRF and NRFs were largely unaffected by the choice of c, except for a modest increase in reconstruction error at $c = 10$ in the 5–15dB range. Thus, the plots for these results are omitted for brevity.

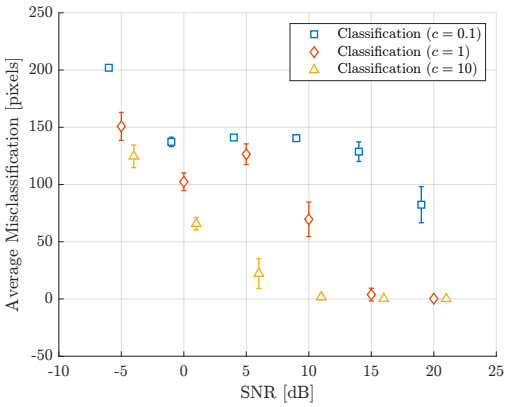


Figure 4-6: Average number of misclassified pixels for various range of scaling parameters $c \in \{0.1,\ 1\ 10\}$ ($\lambda_r = \sigma_h^2 = c$) over SNR[dB]= $\{20, 15, 10, 5, 0, -5\}$

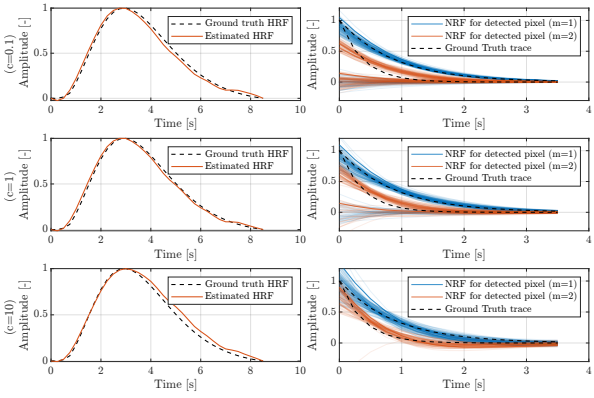


Figure 4-7: The estimated HRFs and NRFs at 10dB for $c = 0.1$ (Top row), $c = 1$ (Middle row) and $c = 10$ (Bottom row). NRFs at detected pixels are shown in blue for $m = 1$ and red for $m = 2$.

Figure 4-6 shows the average number of misclassified pixels. As can be seen, increasing the scaling parameter consistently reduces misclassifications, while decreasing it made detection more sensitive to noise. Although we could not directly explain this trend through analytical expressions, visual inspection of the estimated HRF and NRFs empirically suggests that

scaling influences classification indirectly via shifting effect on HRF and NRF estimation as observed in the previous section. Figure 4-7 illustrates these effects at 10dB. For $c = 10$, the HRF estimate is skewed to the right while for $c = 0.1$, it is biased to the left. These shifts occur with systematic changes in the decay rate and amplitude of the corresponding NRFs. Specifically, a right-biased HRF corresponds to NRFs with higher amplitudes and faster decays, while a left-biased HRF is associated with lower amplitudes and slower decays. When the NRF estimates lean toward sharp and high amplitude patterns, misclassified non-active pixels are penalized more severely in the likelihood, increasing residuals and making them easier to reject. This leads to fewer false positives.

These observations suggest that scaling the prior covariance introduces a trade-off between bias in the estimated HRF and NRF shapes and robustness in activation detection.

Chapter 5

Real Experiments

In this chapter, we apply our JDE model to the real fUS imaging data of the brain. We begin our estimations with the initialization that contains information about possible underlying structures of the data in Section 5-3. The validation of the estimated components are assessed by comparing the measured PDI signals and neural recordings from the dataset in [38].

Subsequently, the estimation is repeated with non-informative initialization and the resulting estimates are validated in Section 5-4.

5-1 Data Acquisition

We used a dataset of simultaneous fUS and neural recordings from Nunez-Elizalde et al [38]. The general data acquisition procedure and preprocessing steps conducted in their study are summarized below. For full details, the reader is referred to [38].

Data were collected from awake, head-fixed mice (C57/BL6) exposed to a 2Hz flashing checkerboard stimulus at three spatial locations: left, center, and right, totaling number of stimulus conditions $M = 3$. Each recording session lasted approximately 8 minutes. Doppler signals were recorded at 500Hz from an ultrasound transducer over a coronal slice covering the primary visual cortex and hippocampus. The Doppler signals were segmented into 400ms batches with 50ms overlap. Each batch was high-pass filtered with a cut-off of 15Hz and the first 15 principal components were removed to eliminate artifacts due to slow-moving tissues [12]. PDIs were then computed by measuring the power of the Doppler signals in the cental non-overlapping 300ms window, resulting in PDI signals with a temporal resolution of 3.33Hz and a spatial resolution of $48.3\mu m \times 100\mu m$.

Neural activity was recorded using Neuropixels probes [19] inserted along parasagittal trajectories to intersect the PDI imaging plane at the visual cortex. Low field potential (LFP) signals were recorded at each electrode site of Neuropixels. The recorded LFP signals were spike sorted and manually curated to extract single-unit and multi-unit activities. These were summed across electrode sites overlapping the imaging plane to yield a single firing rate trace

for the visual cortex. The traces were binned at 300ms intervals to match the temporal resolution of PDI signals. Figure 5-1 illustrates the data acquisition setup using a single probe. Note that we used the data with two Neuropixels probes in this thesis, meaning there is an additional Neuropixels probe on the right hemisphere of the brain.

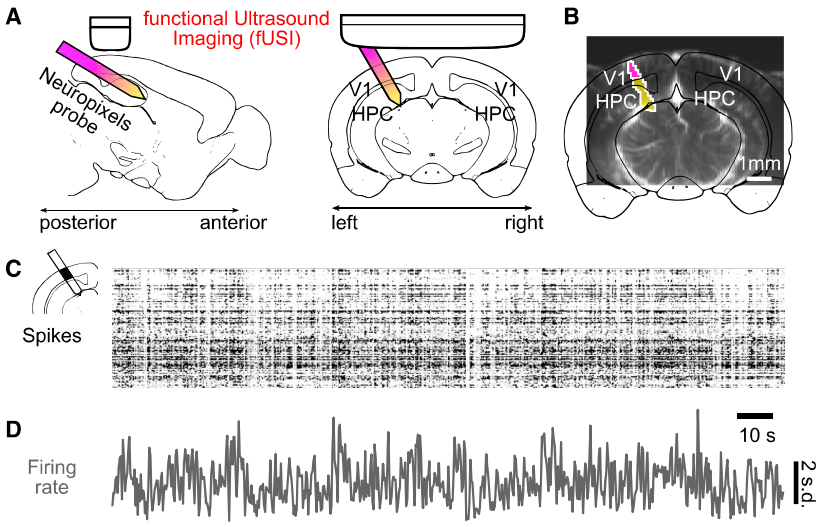


Figure 5-1: Example of data acquisition setup with a single Neuropixels probe [38]: (A) Schematic of concurrent fUS and electrophysiological recordings in the primary visual cortex (V1) and hippocampus (HPC). (B) Coronal PDI image showing the Neuropixels probe trajectory intersecting the imaging plane (purple) and extending anterior to it (yellow). (C) Example of spike activity over time and depth recorded in (V1). (D) Averaged firing rates derived from the spike data.

5-2 Experimental Setup

For this thesis, we used event-evoked recordings on a mouse labeled CR017 in the dataset of Nunez-Elizalde et al [38]. The mouse was shown 40 trials of checkerboard stimuli at three screen positions (left, center, right), repeated four times. Only one screen flashed at a time and stimuli were not overlapping in time. We denote these stimuli as $m = 1$, $m = 2$, and $m = 3$ for the left, center, and right screens of flashing checkerboards, respectively. The illustrations of the presented visual stimuli and the Pearson correlation coefficient (PCC) between the PDI signals and the boxcar stimulus sequence (delayed by 2.1s for maximum correlation) are shown in Figure 5-2.

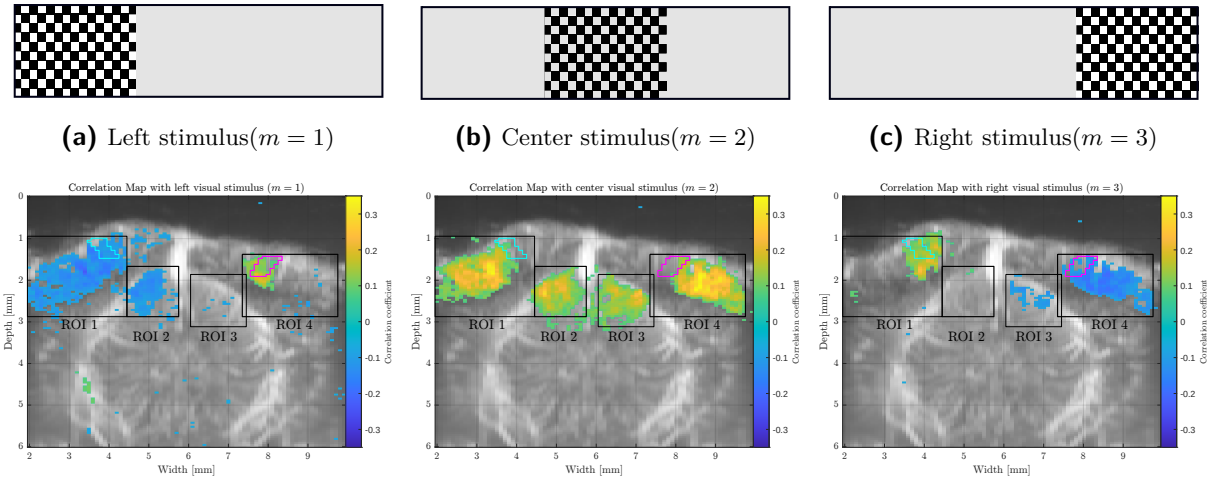
The original PDI data had dimensions 125×80 in 2D space and 1593 in time. We focus on four ROIs depicted in Figure 5-2. We focused on four region of interest (ROI)s (see Figure 5-2) selected based on pixels showing magnitude of a Pearson correlation coefficient (PCC) above 0.08 with the delayed (2.1s) boxcar stimulus signals. These ROIs correspond roughly to the visual cortex (ROI 1 and 4) and hippocampus (ROI 2 and 3). To remove edge artifacts, we truncated the first 9 and the last 10 samples of the original fUS data, resulting in a temporal length of 1574 samples. The spatio-temporal dimension for each ROI is as follows:

- ROI 1: $40 \times 25 \times 1574$ ($J = 40 \times 25 = 1000$ and $N = 1574$)

- ROI 2: $24 \times 13 \times 1574$ ($J = 24 \times 13 = 312$ and $N = 1574$)
- ROI 3: $26 \times 14 \times 1574$ ($J = 26 \times 14 = 364$ and $N = 1574$)
- ROI 4: $31 \times 24 \times 1574$ ($J = 31 \times 24 = 744$ and $N = 1574$)

Neural signals were taken from spike-sorted and curated firing rates, binned at 300 ms. Final traces were averaged over 52 electrodes (left probe) and 266 electrodes (right probe), and also truncated to 1574 samples for alignment.

As a final remark, we did not apply the normalization to PDI signals as that was used in [38] during the estimation of the JDE. Furthermore, consistent with prior JDE design principles in [7], we performed no low-pass filtering or de-trending on the PDI signals before applying our method.



(a) Left stimulus($m = 1$)

(b) Center stimulus($m = 2$)

(c) Right stimulus($m = 3$)

(d) Correlation with left stimulus

(e) Correlation with center stimulus

(f) Correlation with right stimulus

Figure 5-2: (a)-(c): illustrations of the flashing checker board stimuli [38]. (d)-(f): correlation maps between the PDI signals and the boxcar stimulus sequences (delayed by = 2.1s). Thresholded at $|\text{PCC}| > 0.08$ and overlaid on average PDI signals. Cyan and magenta lines indicate intersections of Neuropixels probes.

5-3 Estimation with Informative Initialization

In this section, we present the results of our JDE with informative initializations on the ROIs. "Informative" refers to initializations that contain information extracted from the real data set. In specific, we initialized activation class of a pixel to "active" ($z_j^m = 1$) if its absolute correlation with the delayed boxcar stimulus exceeded 0.08; otherwise, it was initialized to "non-active" ($z_j^m = 1$). For HRF, we initialized the mean of HRF with a single gamma canonical HRF with a peak amplitude of 1, a time-to-peak (TTP) of 2.4s and a full width at half maximum (FWHM) of 2.7s, based on the reported HRF in [38]. The values of the model parameters and other latent variables used for the estimation are compiled in Tables 5-1 and 5-2.

Table 5-1: Model parameters and initializations

Model parameter	Value	Description
Δt	0.3s	Sampling period of real data (= 3.3Hz)
σ_j	$\frac{\sum_{n=1}^{N}(\boldsymbol{y}_j(n)-\bar{\boldsymbol{y}}_j)^2}{(N-1)}$	Initialized at sample variance of $\boldsymbol{y}_j$ ($\bar{\boldsymbol{y}}$ denotes sample mean)
λ_r	100	Scaling factor of second-order stable-spline kernel
α_r	0.84	Decay factor of second-order state-spline kernel
σ_0^2	0.157	Set to match determinants ($\det \boldsymbol{\Sigma}_{0m} = \det \boldsymbol{\Sigma}_{1m}$)
σ_h	100	Initial scaling factor for HRF prior
$\boldsymbol{\mu}_{h0}$	$\boldsymbol{\mu}_h$	Set to initial $\boldsymbol{\mu}_h$ (TTP= 2.4s, FWHM= 2.7s, amplitude 1)
$\boldsymbol{\ell}_j$	-	Estimated initially with $\boldsymbol{\mu}_h = \mathbf{0}$, $\boldsymbol{\mu}_{r_j} = \mathbf{0}$
u_1	0.8	Lower bound of uniform distribution on $\beta_m \ \forall \ m$
u_2	2	Upper bound of uniform distribution on $\beta_m \ \forall \ m$

Table 5-2: Variational parameters and initializations

Variational parameter	Value	Description
$\boldsymbol{\mu}_h$	-	TTP= 2.4s, FWHM= 2.7s and peak amplitude of 1
$\boldsymbol{\Sigma}_h$	$\boldsymbol{I}_{L_h}$	Initial posterior variance of HRF
$q(z_j^m)$	-	Initialized at $q(z_j^m = 1) = 1$ if PCC > 0.08 (Zero otherwise)
β_m	0.88 $\forall \ m$	Initial spatial regularization parameter

5-3-1 Estimated HRFs

Figure 5-3 displays the estimated HRF for each ROI alongside the initial gamma HRF. The initial HRF amplitude was rescaled for comparison.

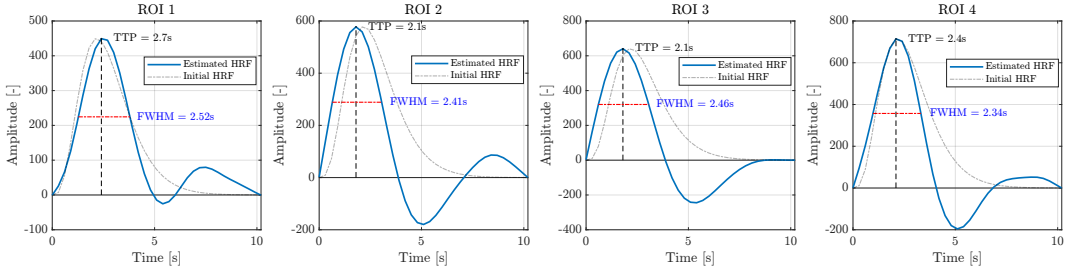


Figure 5-3: Initial HRF and estimated HRFs for four ROIs. Note that the initial HRF with peak amplitude of 1 is rescaled to the peak amplitude of the estimated HRF for visualization purpose.

Comparing to the initial HRF, we observe that the peaks of all estimated HRFs are followed by the undershoots, diverging from the initial single-gamma HRF. Contradicting to the reported HRF in [2], our results suggest that the underlying HRFs are likely to follow two- or higher-gamma HRF model than the single-gamma.

In terms of response speed, estimated HRFs from ROI 2 and 3 have faster response as both TTPs and FWHM are shorter and their overall shapes are shifted to the left. On the other hand, the HRFs from ROI 1 and 4 have relatively similar response time to the initial HRF. ROI 2 and 3 roughly correspond to the hippocampus in the left and right hemisphere and ROI 1 and 4 roughly correspond to the left and right visual cortices. The observations of

the response time indicate that the hippocampus has faster response than the visual cortex. Assuming structural symmetry in the brain, the estimated HRFs can be seen as a supportive indication for the regional dependency hypothesis on HRF.

Overall, despite the observed differences between the estimated HRF and the initial HRF, the estimated HRFs align with multi-subject values of TTP of 2.1 ± 0.3s (median $\pm$ median absolute deviation) and FWHM of 2.9 ± 0.6s in [38], confirming the physiological plausibility of our HRF estimates.

5-3-2 Detected Activation Maps

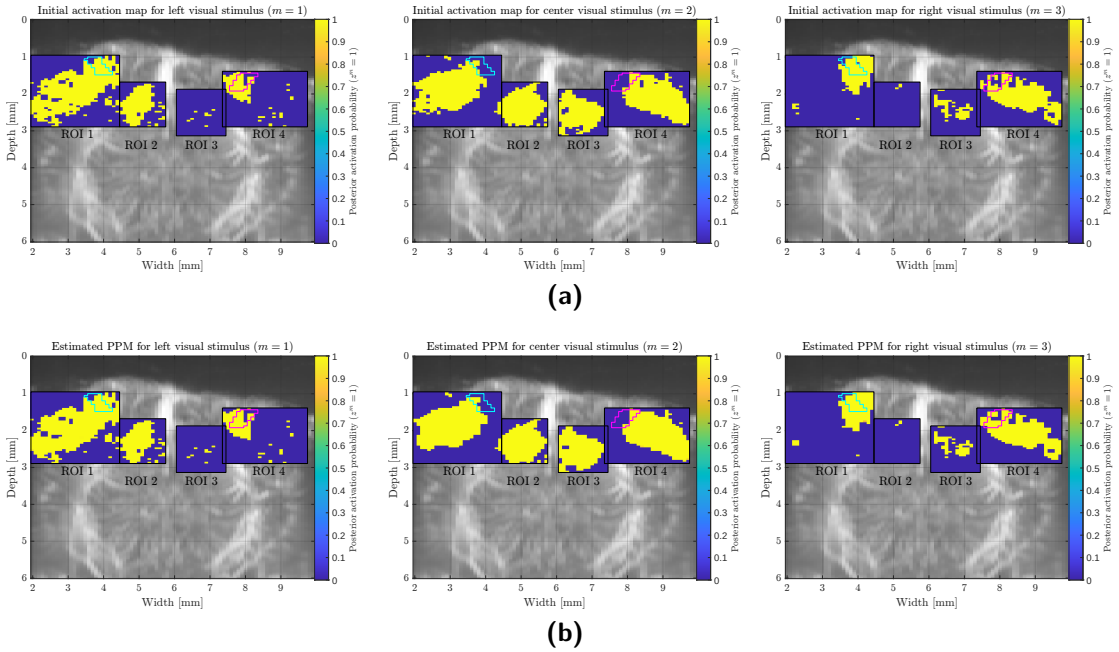


Figure 5-4: (a): Initial activation maps. (b): Estimated activation maps for left ($m = 1$), center ($m = 2$), and right ($m = 3$) stimuli

Figure 5-4a displays the initial activation maps and Figure 5-4b shows the estimated activation maps. The estimated activation maps are nearly identical to the initial ones, as expected given that the initial activation maps already contain information about pixels with input-output correlation. The estimated activation maps have slightly higher spatial coherence, particularly in ROI 1 and 4 as they have fewer non-active pixels within clusters. Although the estimated activation maps do not yield more information than the initializations, the near-equivalence confirms that our JDE solution remained close the initial sub-optimal point.

Regarding both the initial and the estimated activation maps, it can be observed that the activation patterns for the left and right visual stimuli have asymmetric structures and opposite to each other while the center visual stimulus yields a symmetric activation map. These structures are consistent with lateralized visual processing [51], where each hemisphere primarily handles information from the opposite side of the visual field, and commitment from

the opposite hemisphere could be modulated. This will become clearer as we present the estimated neural responses in the following section.

5-3-3 Estimated NRFs

Traces of the estimated NRFs from the active pixels in the activation maps are shown in 3×4 panels in Figure 5-5. In this figure, a column of the panel corresponds to a ROI and a row corresponds to a stimulus type. For instance, three plots of NRF in the left most column show the estimated NRFs in ROI 1 for left ($m = 1$), center ($m = 2$), right ($m = 3$) visual stimuli.

ROI 1 and 2 (the left visual cortex and hippocampus) reveal negative neural responses to the left stimulus and positive responses to the center and right stimuli. On the other hand, ROI 3 and 4 (right hemisphere) have the opposite relation between the signs of neural response and the stimulus types. These sign differences align with the signs of coefficients in the correlation map in Figure 5-2.

Knowing that the estimated HRFs for all ROIs have large positive response (see Figure 5-3), the negative NRFs imply a decrease in blood volume in the left visual cortex and hippocampus during the left visual stimulus decreases and similarly in the right visual cortex and hippocampus during the right stimulus. Although the underlying mechanism for these negative responses remains uncertain, based on the lateralized visual processing pattern observed in the previous section, the estimated neural responses are consistent with neural inhibition hypothesis and attentional modulation processes [37, 33]. These frameworks propose that neuronal activity in task-irrelevant areas is suppressed to facilitate information processing in task-relevant areas.

In summary, the stimulus-dependent polarity of the estimated neural responses across hemispheres supports the hypothesis of functionally lateralized processing. Furthermore, the alignment of neural activity patterns with estimated activation maps and correlation structures (Figure 5-2) reinforces the physiological plausibility and interpretation of the estimated NRFs.

5-3-4 Validity of Estimates

To assess the reliability of the estimated components, we evaluated how well the reconstructed PDI signals matched the measured data. Specifically, we computed the PCC between the reconstructed and the low-pass filtered PDI signals (cut-off of 0.3Hz), shown in Figure 5-6a. The correlation map shows moderate to high PCC values in each ROI. The average PCC in ROI 1, 2, 3 and 4 are found to be 0.347, 0.403, 0.386 and 0.393, respectively. Although these values might be generally considered low, we found that the maximum PCC between the low-pass filtered PDI signals and the delayed boxcar stimuli is only 0.326. This suggests that the dataset contains low input-output correlation information. Given this, the obtained average PCC values indicate that our JDE model can moderately capture the relationship between the stimulus and the measured PDI signals.

Next, we examined the validity of the estimated neurovascular dynamics within the probe regions. We first computed the PCC between the average of the reconstructed and low-pass

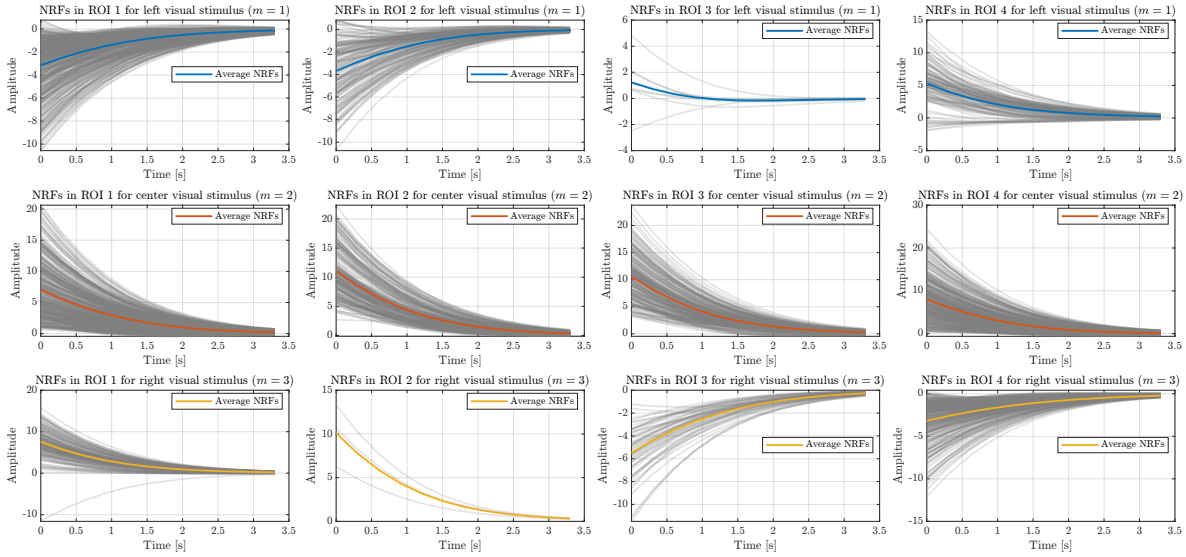
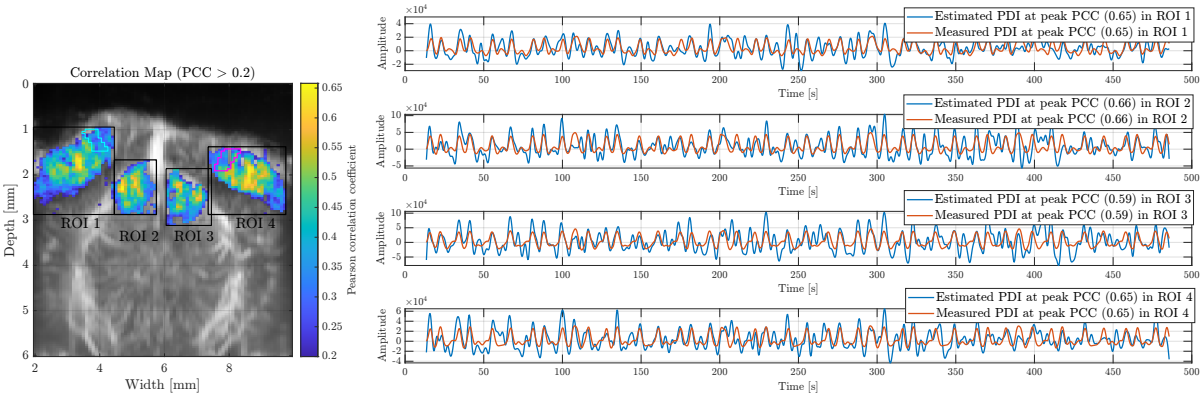


Figure 5-5: Estimated NRFs of four ROIs with informative initializations. The estimated NRFs. Columns: ROI $1-4$; Rows: stimulus types $m = 1$ (top), $m = 2$ (middle), $m = 3$ (bottom).



(a) **(b)**

Figure 5-6: (a): Correlation between low-pass filtered measured PDI and reconstructed PDI. (b): Low-pass filtered measured and reconstructed PDI signals at peak PCC pixel for each ROI

filtered measured PDI signals within the probe areas, which are found to be 0.41 and 0.44. These results further indicate moderate accuracy in capturing the stimulus-blood relation.

We then reconstructed the aggregated neural activity signal within each Neuropixels probe area to evaluate how well the neural activity estimated from the PDI signals reflects the actual neural activity. Figure 5-7b shows the bandpass-filtered average firing rate (pass band $0.08 - 0.8$Hz) and the reconstructed neural activity on a normalized scale.

In this figure, the two signals for each probe appear noticeably different. The aggregate of reconstructed neural activities showed low correlations of 0.24 and 0.30 for the left and right probes, respectively. As a baseline for comparison, we also evaluated the correlations between the measured firing rate and the sum of (delayed) boxcar stimuli with adjusted polarity, which yielded correlations of 0.20 and 0.27, respectively.

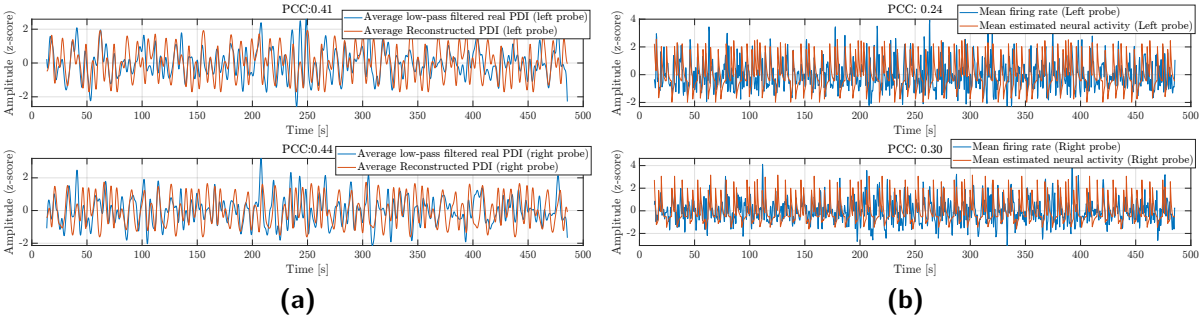


Figure 5-7: (a): Average reconstructed PDI signal within probes. Left probe (Top) and right probe (Bottom). (b): Average reconstructed neural activities in left probe (Top) and right probe (Bottom)

Here, "adjusted polarity" refers to using a weighted combination of stimulus boxcar functions, specifically $\sum_{m=1}^{M} c_m \boldsymbol{x}_m$, where $\boldsymbol{x}_m$ is the boxcar stimulus for stimulus m and $c_m \in -1, 1$ is a sign factor chosen from the polarities of the average NRF for each stimulus. For instance, $c_1 = -1$, $c_2 = c_3 = 1$ were used when computing correlation to the neural firing rate from left Neuropixels probe, according to the polarity of the average NRFs in ROI 1.

This implies that despite their low correlation values, the reconstructed neural activity represent the measured neural activity better than the commonly assumed surrogate signal models.

Moreover, without these polarity adjustments (i.e., $c_m = 1$ for all m), the boxcar stimulus model yielded significantly low correlation coefficients < 0.08. This strongly indicates that the neural firing rates decrease in one side of the hemisphere when the direction of incoming visual stimulus is on the same side and thus, further supporting the neural inhibition hypothesis mentioned before.

Taken together with the moderate PCC values for the average PDI signals in the probe areas, these observations suggest that our JDE model can moderately capture not only the direct input-output dynamics between stimulus onsets and measured PDI signals but also the intermediate dynamics between the stimulus and the underlying neural activity inferred from PDI.

Although we can directly validate the estimated neural activity only in the probe regions due to the limited spatial coverage, the similarity between the PCC values in the ROIs and those in the probe areas supports the validity of the estimated components in other regions as well.

To sum up, the correlations between the reconstructed and measured PDI signals were relatively high in both the ROIs and the Neuropixels probes areas, despite the inherently low input-output correlation in the measured data. While the reconstructed neural activities showed only moderate correlations with the measured neural activity ($\approx$ 0.24–0.30), they nonetheless outperformed commonly used surrogate input models. These results indicate that our JDE model can moderately capture and represent the neurovascular dynamics in the fUS data.

5-4 Verification with Non-informative Initialization

In Section 5-3, we presented estimation results using informative initialization, where activation patterns and the mean HRF were initialized based on correlation maps and the reported HRF in [38]. To verify the consistency of these results, we repeated the estimation on the same experimental data using a non-informative initialization. Specifically, the activation patterns were initialized uniformly ($q(z_j^m) = 0.5$ for all pixels j and stimulus types m), and the mean HRF was initialized as a single gamma function with a TTP of 1.5s and FWHM of 2.03, which differs notably from the prior-informed shape in [38]. We initially attempted to keep the same values used as in the previous estimation. However, ROI 3 quickly converged to a poor local minimum within 10 iterations, resulting in zero activation detections and negative HRF, likely due to low detection sensitivity. To mitigate this, we gradually increased the scaling parameter of the NRF prior until convergence improved, resulting in $\lambda_r = 300$.

Following this adjustment, the estimated HRFs from both initializations matched closely (PCC > 0.999). This indicates that the informative initialization does not introduce significant bias into the HRF estimation and thus, reinforcing the validity of the HRF estimates and the possibility of region-specific HRF hypothesis observed earlier.

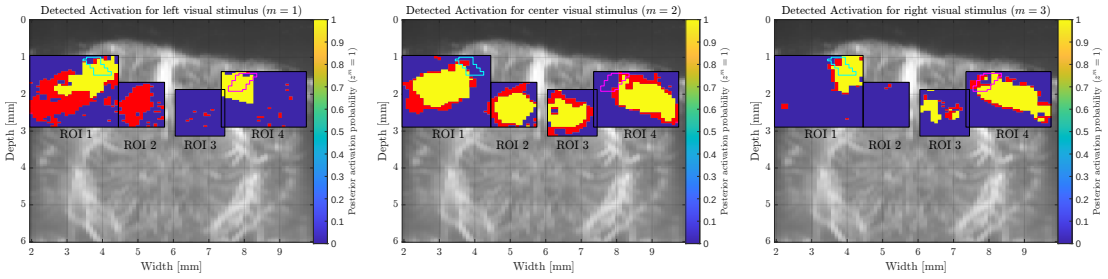


Figure 5-8: Estimated activation maps with non-informative initialization for left ($m = 1$), center ($m = 2$), and right ($m = 3$) stimuli. Red pixels denote activations detected with informative initialization.

Figure 5-8 shows the estimated activation maps under the non-informative initialization on top of the informative ones. Compared to the informative results, notable differences can be observed in ROIs 2 and 3 for the left stimulus, where no activations were detected. especially ROI 2 and 3 where there is no detected active pixel, whereas the rest of the activations patterns are similar. ROI 1 also showed reduced detection for left stimulus (from 504 to 203 active pixels), amounting to a 40% drop. Together with the necessary parameter adjustment mentioned before, these differences highlight moderate dependence of our JDE on the initialization. Nonetheless, the activation patterns still preserved key spatial features such as lateralized visual processing structures found earlier, suggesting that the model retains biologically plausible organization even with uninformative starting points.

To examine the variability in estimated neural responses, Figure 5-9 displays NRFs from overlapping active regions. Although the average traces are similar across initializations, substantial variation can be seen between individual NRFs. The average relative root mean square error with respect to the informative ones was approximately 10% with a standard deviation of 13. Consequently, the reconstructed PDI signals under non-informative initialization differ by 20% in relative Frobenius norm from those with the informative initialization.

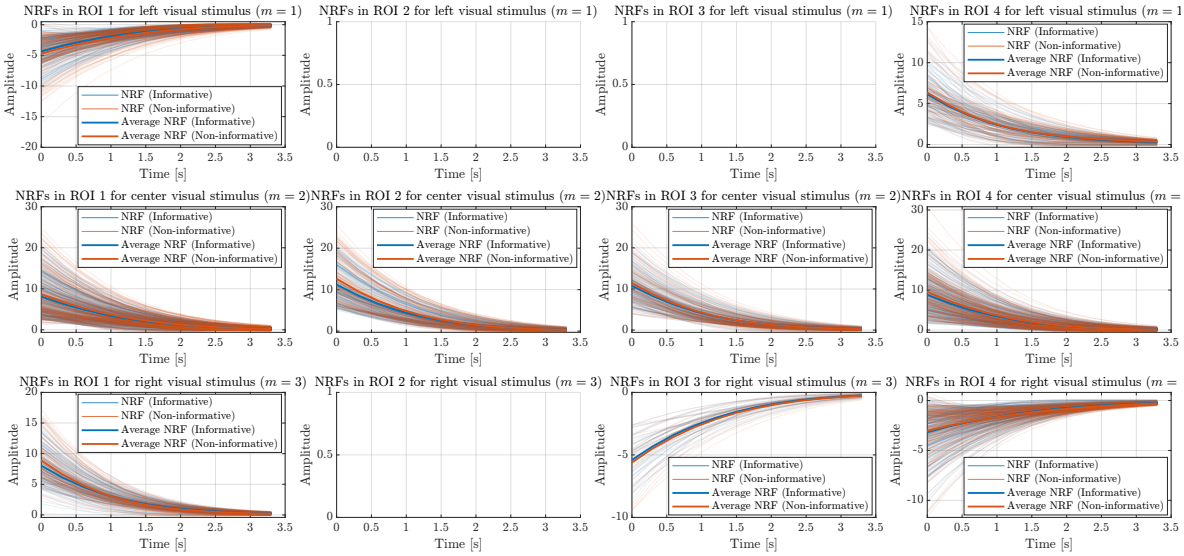


Figure 5-9: Estimated NRFs of four ROIs for co-detected pixels. Columns: ROI $1-4$; Rows: stimulus types $m=1$ (top), $m=2$ (middle), $m=3$ (bottom). NRFs in each ROI are rescaled using HRF amplitude between initializations to eliminate scaling ambiguity.

These findings suggest that while average dynamics remain stable, pixel-wise neural responses from our JDE are more sensitive to the choice of initialization on the activation pattern.

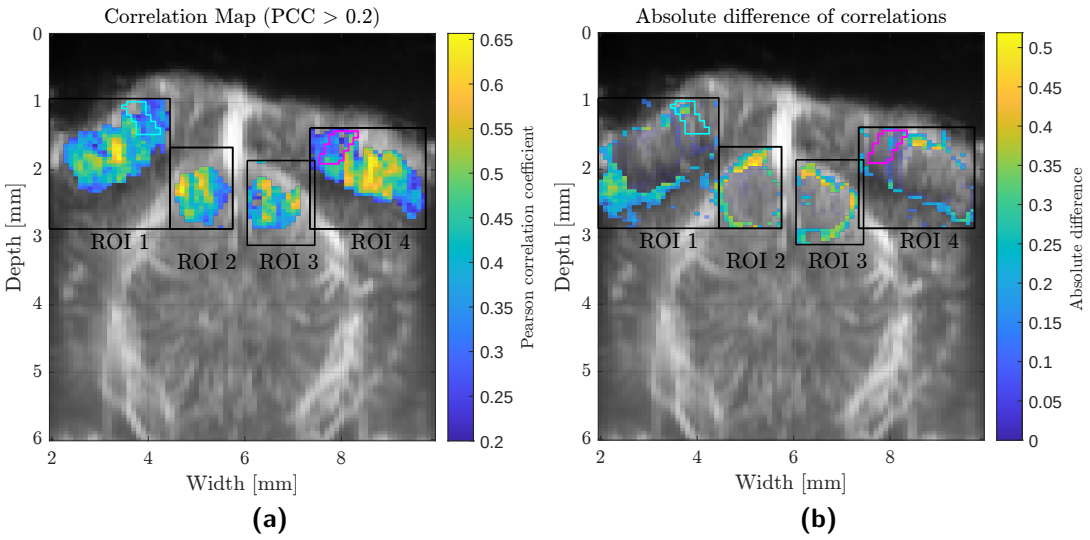


Figure 5-10: (a): Correlation between low-pass filtered measured and reconstructed PDI with non-informative initialization. (b): Absolute difference in correlation coefficients between the informative and non-informative initializations.

Figure 5-10a shows the correlation between the measured and reconstructed PDI signals. Across all ROIs, we found slightly higher average PCCs ($0.402-0.442$) than those under informative initialization $0.347-0.403$. Figure 5-10b displays the absolute difference in correlation maps between the two initializations, highlighting that both reconstructions fit the

data comparably well. Reconstructed neural activity for the left probe showed the same correlation (PCC = 0.24) across both initializations, while the right probe yielded a slightly lower correlation (0.27 vs. 0.30). Despite the substantial differences observed earlier in the activation maps and NRFs, the correlations to the measured data remained similar between the two initializations. This suggests that the JDE can still represent the hidden neural activity signals from the PDI signals with non-informative initialization.

To summarize, non-informative initialization led to reduced activation detection in ROI 1 (by 40%) and complete suppression in ROI 2, along with relative errors of approximately 10% in NRFs and 20% in reconstructed PDI signals compared to the informative ones. Despite these differences, the HRFs were nearly identical and the global signal reconstructions remained comparable, indicating that the informative initialization does not bias the overall estimation. However, the variability in detection and neural response quality suggests that the JDE model remains sensitive to initialization, particularly in pixel-wise neural response estimation.

Chapter 6

Conclusions and Future Works

This chapter presents summary of this thesis including the key results found during the experiments, and suggestions for future work based on these results.

In this thesis, we developed an extension of joint detection-estimation (JDE) algorithm to estimate not only hemodynamics, but also dynamics of neural activity by considering cascade convolution model. Such a design is considered with respect to a potential utility for neuroscientists in understanding mysterious neurovascular coupling mechanism.

In Chapter 4, we validated our JDE on synthetic data to assess its performance in estimating hemodynamic response function (HRF), neural response function (NRF), and detecting activations. Convergence analysis showed that the model accurately recovered HRFs and NRFs in some setup when activation detection was reliable while it partially struggled with misclassified activations and degraded NRF estimates in other set. Robustness tests demonstrated that the model maintained low reconstruction and HRF errors for SNRs above 5dB whereas activation detection became less reliable in noisy settings. Finally, we found that scaling parameters significantly impacted performance, revealing sensitivity to scaling parameters for reliable estimates.

In Chapter 5, the proposed JDE model was applied to real fUS imaging data on a mouse with visual stimulus experiments. We found the estimated region-specific HRFs diverge from the canonical single-gamma shape and suggest a multi-gamma form. The HRFs also revealed distinct temporal trends between the hippocampus and visual cortex, supporting the regional dependency hypothesis on HRF. The detected activations aligned well with stimulus based correlations and exhibited lateralized patterns which were consistent with known neurophysiological visual processing. Estimated neural response functions revealed stimulus-dependent polarity difference, indicative of neural inhibition and attentional modulation mechanisms. Despite inherently low input-output correlation in the data, the model achieved moderate correlation between reconstructed and measured PDI signals, as well as improved representation of neural activity compared to commonly used boxcar surrogate models. However, the results revealed a moderate sensitivity of activation and neural response estimations to initialization conditions, underscoring some limitations in robustness at the pixel level. Overall,

these findings demonstrate the capability of our model to capture both hemodynamics and hidden dynamics of neural activity from input stimulus onset and real fUS data without noise filtering or detrending on the beam-formed PDI signals.

Returning to our research question
"Can we jointly estimate hemodynamic response function and neural response function in a probabilistic framework using fUS data?",
we conclude that our JDE model can successfully estimate hemodynamic response function, and it can estimate neural response function well in collective sense, but not in pixel level.

6-1 Limitations and Future Works

The JDE design developed in this thesis is associated with limitations and some perspectives to be considered in future work. The following list is ordered by relevance and urgency.

1. **Increasing robustness to initialization**
   The model demonstrated moderate sensitivity to initialization, particularly for activation maps and neural responses, occasionally converging to poor local minima without structured guidance. This motivates future work on improving the robustness of activation detection with respect to initialization. A possible approach is to use structured mean-field approximations for the activation class labels to better capture dependencies between pixels. For instance, limiting the factorization of $q(\boldsymbol{z}^m)$ to a 3×3 patch. Although this involves the evaluation of 2^9 configurations per patch, it could increase robustness to initialization by spreading the posterior probability mass and avoiding poor local minima.

2. **Conducting a study for systematic parameter tuning**
   Although the region-averaged NRF traces remained stable, the pixel-level estimates exhibited high variability. Consequently, interpreting pixel-wise dynamics remains speculative. While this limitation arises partially due to the issue stated in the previous point, the choice of model parameters also significantly affects the resulting estimates. This limitation greatly hinders further extensions of this model, and thus, future work should be directed towards investigating systematic model parameter selection methods or automatic scaling parameter inference.

3. **Incorporation of tensor decomposition with 1D Ising models**
   The integration of tensor decomposition techniques, combined with spatial regularization via the 1D Ising model on activation patterns, could potentially improve spatial coherence and detection reliability.

4. **Switch to stochastic optimization**
   The prior distributions used in this thesis were limited to conjugate priors, as the closed-form update rules in our JDE were derived within a CAVI framework. Moving beyond conjugate priors and closed-form updates, adopting stochastic or black-box variational inference would enable the use of non-conjugate priors and nonlinear NRF models. This would expand model flexibility and reduce reliance on tedious derivations while potentially improving robustness to noise and initialization. Moreover, even if the same

Bayesian model is used, stochastic optimization could potentially escape poor local minima thanks to its inherent randomness. However, it should be noted that stochastic variational inference may also suffer from high variance in the estimates. Nonetheless, as stochastic variational inference has not yet been tested on fUS data in the literature, it warrants exploration in future work.

5. **Nonlinear extension to a delay-normalized model for NRF**
   In line with the limitation of the CAVI framework mentioned in the previous point, our observation model is restricted to a linear convolution model. By switching to stochastic variational inference, we could implement a nonlinear NRF model, such as a delay-normalization model. The cascade model, consisting of a nonlinear block followed by a linear convolution block, is known as a Hammerstein model. Although the optimization algorithm does not necessarily have to be Bayesian for this system, it would be interesting to investigate whether this improves the estimation of neural activity, as neural responses are commonly known to exhibit nonlinear characteristics.

Appendix A

Derivations of Update Equations

A-1 Update for $q(\boldsymbol{h})$

For the sake of convenience, the CAVI update rule and the complete data likelihood in Equations 3-40 and 3-34 are recalled below.

CAVI update rule:

$$\ln q(\boldsymbol{h})^{(t)} \stackrel{c}{=} \mathbb{E}_{q(\boldsymbol{r})^{(t)} q(\boldsymbol{z})^{(t-1)} q(\boldsymbol{\beta})^{(t-1)}} \left[\ln p\left(\boldsymbol{y}, \boldsymbol{h}, \boldsymbol{r}, \boldsymbol{\beta}; \boldsymbol{\theta}^{(t-1)} \right) \right] \tag{A-1}$$

Complete data likelihood:

$$p(\boldsymbol{y}, \boldsymbol{h}, \boldsymbol{r}, \boldsymbol{z}, \boldsymbol{\beta}; \boldsymbol{\Theta}) = p(\boldsymbol{y}|\boldsymbol{h}, \boldsymbol{r}) p(\boldsymbol{h}) p(\boldsymbol{r}|\boldsymbol{z}) p(\boldsymbol{z}|\boldsymbol{\beta}) p(\boldsymbol{\beta}) \tag{A-2}$$

Furthermore, for the sake of brevity, we omit in the derivations the superscript (t) and the model parameters that are not estimated.

Starting from Equation A-1 and substituting the complete-data likelihood (Equation A-2):

$$\ln q(\boldsymbol{h}) \stackrel{c}{=} \mathbb{E}_{q(\boldsymbol{r}) q(\boldsymbol{z}) q(\boldsymbol{\beta})} \left[\ln p(\boldsymbol{y}, \boldsymbol{h}, \boldsymbol{r}, \boldsymbol{z}, \boldsymbol{\beta}; \boldsymbol{\theta}) \right] \tag{A-3}$$

$$= \mathbb{E}_{q(\boldsymbol{r}) q(\boldsymbol{z}) q(\boldsymbol{\beta})} \left[\ln p(\boldsymbol{y}|\boldsymbol{h}, \boldsymbol{r}; \boldsymbol{\ell}, \sigma^2) + \ln p(\boldsymbol{h}; \sigma_h^2) + \ln p(\boldsymbol{r}|\boldsymbol{z}) + \ln p(\boldsymbol{z}|\boldsymbol{\beta}) + \ln p(\boldsymbol{\beta}) \right]. \tag{A-4}$$

Since expectation is a linear operator, we can split this as:

$$\begin{aligned} \ln q(\boldsymbol{h}) = {} & \mathbb{E}_{q(\boldsymbol{r}) q(\boldsymbol{z}) q(\boldsymbol{\beta})} \left[\ln p(\boldsymbol{y}|\boldsymbol{h}, \boldsymbol{r}; \boldsymbol{\ell}, \sigma^2) \right] + \mathbb{E}_{q(\boldsymbol{r}) q(\boldsymbol{z}) q(\boldsymbol{\beta})} \left[\ln p(\boldsymbol{h}; \sigma_h^2) \right] \\ & + \mathbb{E}_{q(\boldsymbol{r}) q(\boldsymbol{z}) q(\boldsymbol{\beta})} \left[\ln p(\boldsymbol{r}|\boldsymbol{z}) \right] + \mathbb{E}_{q(\boldsymbol{r}) q(\boldsymbol{z}) q(\boldsymbol{\beta})} \left[\ln p(\boldsymbol{z}|\boldsymbol{\beta}) \right] + \mathbb{E}_{q(\boldsymbol{r}) q(\boldsymbol{z}) q(\boldsymbol{\beta})} \left[\ln p(\boldsymbol{\beta}) \right]. \end{aligned} \tag{A-5}$$

Now, observe that the last three terms in Equation A-5 are constant with respect to $\boldsymbol{h}$. Furthermore, the function inside the first expectation depends only on $\boldsymbol{h}$ and $\boldsymbol{r}$, while the second term depends solely on $\boldsymbol{h}$. Thus, we can simplify:

$$\ln q(\boldsymbol{h}) \stackrel{c}{=} \mathbb{E}_{q(\boldsymbol{r})} \left[\ln p(\boldsymbol{y}|\boldsymbol{h}, \boldsymbol{r}; \boldsymbol{\ell}, \sigma^2) \right] + \ln p(\boldsymbol{h}; \sigma_h^2). \tag{A-6}$$

In what follows, we will carry out similar simplifications without explicitly remarking on constant terms with respect to the current variational factor.

The prior term is a standard Gaussian log-density:

$$\ln p(\boldsymbol{h};\sigma_h^2) \stackrel{c}{=} -\frac{1}{2\sigma_h^2}\left(\boldsymbol{h}^\top \boldsymbol{R}^{-1}\boldsymbol{h} - 2\boldsymbol{\mu}_{h0}^\top \boldsymbol{R}^{-1}\boldsymbol{h}\right). \tag{A-7}$$

Next, we evaluate the expectation of log-likelihood in Equation A-6. Substituting Equations 3-9 and 3-6 into the log-expectation gives

$$\mathbb{E}_{q(\boldsymbol{r})}\left[\ln p(\boldsymbol{y}|\boldsymbol{h},\boldsymbol{r};\boldsymbol{\ell},\sigma^2)\right] = \mathbb{E}_{q(\boldsymbol{r})}\left[\ln \prod_{j=1}^{J} p(\boldsymbol{y}_j|\boldsymbol{h},\boldsymbol{r}_j;\boldsymbol{\ell}_j,\sigma_j^2)\right] \tag{A-8}$$

$$\stackrel{c}{=} \mathbb{E}_{q(\boldsymbol{r})}\left[\sum_{j=1}^{J} -\frac{1}{2\sigma_j^2}\left(\boldsymbol{y}_j - \sum_{m=1}^{M}\boldsymbol{X}_m\boldsymbol{R}_j^m\boldsymbol{h} - \boldsymbol{P}\boldsymbol{\ell}_j\right)^\top\left(\boldsymbol{y}_j - \sum_{m=1}^{M}\boldsymbol{X}_m\boldsymbol{R}_j^m\boldsymbol{h} - \boldsymbol{P}\boldsymbol{\ell}_j\right)\right] \tag{A-9}$$

$$\stackrel{c}{=} \sum_{j=1}^{J} -\frac{1}{2\sigma_j^2}\left(-2\left(\boldsymbol{y}_j - \boldsymbol{P}\boldsymbol{\ell}_j\right)^\top\left(\sum_{m=1}^{M}\boldsymbol{X}_m\mathbb{E}_{q(\boldsymbol{r}_j^m)}\left[\boldsymbol{R}_j^m\right]\right)\boldsymbol{h}\right.$$
$$\left. -\boldsymbol{h}^\top\mathbb{E}_{q(\boldsymbol{r}_j^m)}\left[\left(\sum_{m=1}^{M}\boldsymbol{X}_m\boldsymbol{R}_j^m\right)^\top\left(\sum_{m=1}^{M}\boldsymbol{X}_m\boldsymbol{R}_j^m\right)\right]\boldsymbol{h}\right). \tag{A-10}$$

In the step above, the expectations over $\boldsymbol{r}$ are factorized via mean-field: $q(\boldsymbol{r}) = \prod_j \prod_m q(\boldsymbol{r}_j^m)$.

Grouping quadratic terms in $\boldsymbol{h}$ form Equations A-7 and A-10 as:

$$\boldsymbol{W}_h = \sum_{j=1}^{J}\frac{1}{\sigma_j^2}\mathbb{E}_{q(\boldsymbol{r}_j^m)}\left[\left(\sum_{m=1}^{M}\boldsymbol{X}_m\boldsymbol{R}_j^m\right)^\top\left(\sum_{m=1}^{M}\boldsymbol{X}_m\boldsymbol{R}_j^m\right)\right] + \frac{1}{\sigma_h^2}\boldsymbol{R}^{-1}, \tag{A-11}$$

and other terms:

$$\boldsymbol{b}_h^\top = \sum_{j=1}^{J}\frac{1}{\sigma_j^2}\left(\boldsymbol{y}_j - \boldsymbol{P}\boldsymbol{\ell}_j\right)^\top\left(\sum_{m=1}^{M}\boldsymbol{X}_m\mathbb{E}_{q(\boldsymbol{r}_j^m)}\left[\boldsymbol{R}_j^m\right]\right) + \frac{1}{\sigma_h^2}\boldsymbol{\mu}_{h0}^\top\boldsymbol{R}^{-1} \tag{A-12}$$

Completing the square, Equation A-6 can be rewritten in the quadratic form:

$$\ln q(\boldsymbol{h}) = -\frac{1}{2}(\boldsymbol{h}^\top\boldsymbol{W}_h\boldsymbol{h} - 2\boldsymbol{b}_h^\top\boldsymbol{h}) \tag{A-13}$$

$$= -\frac{1}{2}(\boldsymbol{h} - \boldsymbol{W}_h^{-1}\boldsymbol{b}_h)^\top\boldsymbol{W}_h(\boldsymbol{h} - \boldsymbol{W}_h^{-1}\boldsymbol{b}_h) - \boldsymbol{b}_h^\top\boldsymbol{W}_h^{-1}\boldsymbol{b}_h \tag{A-14}$$

$$\stackrel{c}{=} -\frac{1}{2}(\boldsymbol{h} - \boldsymbol{W}_h^{-1}\boldsymbol{b}_h)^\top\boldsymbol{W}_h(\boldsymbol{h} - \boldsymbol{W}_h^{-1}\boldsymbol{b}_h) \tag{A-15}$$

Notice that the posterior $\ln q(\boldsymbol{h})$ is in a standard Gaussian log-density. Therefore, the posterior $q(\boldsymbol{h})$ is

$$q(\boldsymbol{h}) = \boldsymbol{\mathcal{N}}(\boldsymbol{\mu}_h, \boldsymbol{\Sigma}_h), \tag{A-16}$$

and the posterior variational parameters are updated by

$$\boldsymbol{\Sigma}_h = \boldsymbol{W}_h^{-1} \tag{A-17}$$

$$= \left(\sum_{j=1}^{J} \frac{1}{\sigma_j^2} \mathbb{E}_{q(\boldsymbol{r}_j^m)} \left[\left(\sum_{m=1}^{M} \boldsymbol{X}_m \boldsymbol{R}_j^m \right)^\top \left(\sum_{m=1}^{M} \boldsymbol{X}_m \boldsymbol{R}_j^m \right) \right] + \frac{1}{\sigma_h^2} \boldsymbol{R}^{-1} \right)^{-1} \tag{A-18}$$

and

$$\boldsymbol{\mu}_h = \boldsymbol{W}_h^{-1} \boldsymbol{b}_h$$

$$= \boldsymbol{\Sigma}_h \left(\sum_{j=1}^{J} \frac{1}{\sigma_j^2} \left(\sum_{m=1}^{M} \boldsymbol{X}_m \mathbb{E}_{q(\boldsymbol{r}_j^m)} \left[\boldsymbol{R}_j^m \right] \right)^\top \left(\boldsymbol{y}_j - \boldsymbol{P} \boldsymbol{\ell}_j \right) + \frac{1}{\sigma_h^2} \boldsymbol{R}^{-1} \boldsymbol{\mu}_{h0} \right). \tag{A-19}$$

As regards $\mathbb{E}_{q(\boldsymbol{r}_j^m)} \left[\boldsymbol{R}_j^m \right]$, this is equivalent to constructing the Toeplitz matrix using posterior mean $\boldsymbol{\mu}_{r_j^m}$ of $\boldsymbol{r}_j^m$, which will be obtained in later derivations. For the quadratic expectation in $\boldsymbol{W}_h$, we first expand the inner product as:

$$\left(\sum_{m=1}^{M} \boldsymbol{X}_m \boldsymbol{R}_j^m \right)^\top \left(\sum_{m=1}^{M} \boldsymbol{X}_m \boldsymbol{R}_j^m \right) = \sum_{m=1}^{M} \boldsymbol{R}_j^{m\top} \boldsymbol{X}_m^\top \boldsymbol{X}_m \boldsymbol{R}_j^m \tag{A-20}$$

$$+ \sum_{m=1}^{M} \sum_{n=1}^{m-1} \left(\boldsymbol{R}_j^{m\top} \boldsymbol{X}_m^\top \boldsymbol{X}_n \boldsymbol{R}_j^n + \boldsymbol{R}_j^{n\top} \boldsymbol{X}_n^\top \boldsymbol{X}_m \boldsymbol{R}_j^m \right). \tag{A-21}$$

Denoting the Toeplitz matrix $\boldsymbol{R}_j^m$ as

$$\boldsymbol{R}_j^m = \begin{bmatrix} \boldsymbol{r}_{\boldsymbol{j(1)}}^{\boldsymbol{m}} & \boldsymbol{r}_{\boldsymbol{j(2)}}^{\boldsymbol{m}} & \cdots & \boldsymbol{r}_{\boldsymbol{j(L_h)}}^{\boldsymbol{m}} \end{bmatrix} \tag{A-22}$$

where $\boldsymbol{r}_{j(i)}^m$ represents ith column and $\boldsymbol{r}_{j(i)}^m$ is given by

$$\boldsymbol{r}_{j(i)}^m = \begin{bmatrix} \boldsymbol{0}_{i-1} \\ \boldsymbol{r}_j^m \\ \boldsymbol{0}_{L_h - i} \end{bmatrix} \in \mathbb{R}^{L_h + L_r - 1} \tag{A-23}$$

Here $\boldsymbol{0}_a$ represents a zero column vector of length a. Using these notations, the expectation of $\boldsymbol{R}_j^{m\top} \boldsymbol{X}_m^\top \boldsymbol{X}_n \boldsymbol{R}_j^n$ is

$$\mathbb{E}_{q(\boldsymbol{r}_j^m)} \left[\boldsymbol{R}_j^{m\top} \boldsymbol{X}_m^\top \boldsymbol{X}_n \boldsymbol{R}_j^n \right] = \mathbb{E}_{q(\boldsymbol{r}_j^m)} \left[\begin{bmatrix} \boldsymbol{r}_{j(1)}^{m\top} \\ \boldsymbol{r}_{j(2)}^{m\top} \\ \vdots \\ \boldsymbol{r}_{j(L_h)}^{m\top} \end{bmatrix} \boldsymbol{X}_m^\top \boldsymbol{X}_n \begin{bmatrix} \boldsymbol{r}_{j(1)}^m & \boldsymbol{r}_{j(2)}^m & \cdots & \boldsymbol{r}_{j(L_h)}^m \end{bmatrix} \right] \tag{A-24}$$

Finally, using expectation in quadratic form: $\mathbb{E}[\boldsymbol{x}^\top \boldsymbol{A} \boldsymbol{x}] = \boldsymbol{\mu}^\top A \boldsymbol{\mu} + \mathrm{tr}(\boldsymbol{A}\boldsymbol{\Sigma})$ with trace operation tr, (a, b) entry of the expectation in Equation A-24 is computed as

$$\mathbb{E} \left[\left(\boldsymbol{R}_j^{m\top} \boldsymbol{X}_m^\top \boldsymbol{X}_n \boldsymbol{R}_j^n \right)_{(a,b)} \right] = \mathbb{E} \left[(\boldsymbol{r}_{j_{(a)}}^m) \right]^\top \boldsymbol{X}_m^\top \boldsymbol{X}_n \mathbb{E} \left[\boldsymbol{r}_{j_{(b)}}^n \right] \tag{A-25}$$

$$+ \mathrm{tr} \left(\boldsymbol{X}_m^\top \boldsymbol{X}_n \, \mathrm{Circ} \left(\begin{bmatrix} \Sigma_{r_j^n r_j^m} & \boldsymbol{0}_{L_n \times (L_m - 1)} \\ \boldsymbol{0}_{(L_n - 1) \times L_m} & \boldsymbol{0}_{(L_n - 1) \times (L_m - 1)} \end{bmatrix}, b - 1, a - 1 \right) \right) \tag{A-26}$$

where $\Sigma_{r_j^n r_j^m}$ is posterior covariance between $\boldsymbol{r}_j^n$ and $\boldsymbol{r}_j^m$ and $\mathrm{Circ}(\boldsymbol{A}, i, j)$ is a circular shifting operator that shifts down matrix $\boldsymbol{A}$ in row direction by i and in column direction by j.

A-2 Update for $q(\boldsymbol{r})$

CAVI update rule:

$$\ln q(\boldsymbol{r})^{(t)} \stackrel{c}{=} \mathbb{E}_{q(\boldsymbol{h})^{(t-1)} q(\boldsymbol{z})^{(t-1)} q(\boldsymbol{\beta})^{(t-1)}} \left[\ln p\left(\boldsymbol{y}, \boldsymbol{h}, \boldsymbol{r}, \boldsymbol{\beta}; \boldsymbol{\theta}^{(t-1)} \right) \right] \tag{A-27}$$

Similar to the derivation of $\ln q(\boldsymbol{h})$ omitting constant terms yields

$$\ln q(\boldsymbol{r}) \stackrel{c}{=} \mathbb{E}_{q(\boldsymbol{h})} \left[\ln p(\boldsymbol{y}|\boldsymbol{h}, \boldsymbol{r}; \boldsymbol{\ell}, \sigma^2) \right] + \mathbb{E}_{q(\boldsymbol{z})} \left[\ln p(\boldsymbol{r}|\boldsymbol{z}) \right] \tag{A-28}$$

Using the factorized posterior: $q(\boldsymbol{z}) = \prod_j \prod_m q(\boldsymbol{z}_j^m)$, the log-expectation of the prior is:

$$\mathbb{E}_{q(\boldsymbol{z})} \left[\ln p(\boldsymbol{r}|\boldsymbol{z}) \right] \stackrel{c}{=} \mathbb{E}_{q(\boldsymbol{z})} \left[\sum_{j=1}^{J} \sum_{m=1}^{M} \sum_{k=1}^{K} \boldsymbol{I}(z_j^m = k) \left(-\frac{1}{2} \boldsymbol{r}_j^{m\top} \boldsymbol{\Sigma}_{km}^{-1} \boldsymbol{r}_j^m \right) \right] \tag{A-29}$$

$$= \sum_{j=1}^{J} \sum_{m=1}^{M} \sum_{k=1}^{K} \mathbb{E}_{q(z_j^m)} \left[\boldsymbol{I}(z_j^m = k) \right] \left(-\frac{1}{2} \boldsymbol{r}_j^{m\top} \boldsymbol{\Sigma}_{km}^{-1} \boldsymbol{r}_j^m \right) \tag{A-30}$$

where $\mathbb{E}_{q(z_j^m)} \left[\boldsymbol{I}(z_j^m = k) \right] = q(z_j^m = k)$ is a posterior probability of pixel j belonging to activation class k under mth stimulus since

$$\mathbb{E}_{q(z_j^m)} \left[\boldsymbol{I}(z_j^m = k) \right] = \sum_{z_j^m \in \{1 \ldots K\}} \boldsymbol{I}(z_j^m = k) q(z_j^m) \tag{A-31}$$

$$= q(z_j^m = k). \tag{A-32}$$

For compactness, let $\phi_{jk}^m = q(z_j^m = k)$,

$$\boldsymbol{\Sigma}_k^{-1} = \begin{bmatrix} \boldsymbol{\Sigma}_{k1}^{-1} & & & \\ & \boldsymbol{\Sigma}_{k2}^{-1} & & \\ & & \ddots & \\ & & & \boldsymbol{\Sigma}_{kM}^{-1} \end{bmatrix} \in \mathbb{R}^{ML_r \times ML_r}, \qquad \boldsymbol{r}_j = \begin{bmatrix} \boldsymbol{r}_j^1 \\ \boldsymbol{r}_j^2 \\ \vdots \\ \boldsymbol{r}_j^M \end{bmatrix} \in \mathbb{R}^{ML_r}, \tag{A-33}$$

and

$$\boldsymbol{\Phi}_{jk} = \begin{bmatrix} \phi_{jk}^1 \boldsymbol{I}_{L_r} & & & \\ & \phi_{jk}^2 \boldsymbol{I}_{L_r} & & \\ & & \ddots & \\ & & & \phi_{jk}^M \boldsymbol{I}_{L_r} \end{bmatrix} \in \mathbb{R}^{ML_r \times ML_r}, \tag{A-34}$$

where $\boldsymbol{I}_{(a)}$ is the identity matrix of the size $a \times a$. Using these notations, the log-expectation of the prior is from Equation A-30:

$$\mathbb{E}_{q(\boldsymbol{z})} \left[\ln p(\boldsymbol{r}|\boldsymbol{z}) \right] = \sum_{j=1}^{J} \sum_{m=1}^{M} \sum_{k=1}^{K} \phi_{jk}^m \left(-\frac{1}{2} \boldsymbol{r}_j^{m\top} \boldsymbol{\Sigma}_{km}^{-1} \boldsymbol{r}_j^m \right) \tag{A-35}$$

$$= -\frac{1}{2} \sum_{j=1}^{J} \sum_{k=1}^{K} \boldsymbol{r}_j^\top \boldsymbol{\Phi}_{jk} \boldsymbol{\Sigma}_k^{-1} \boldsymbol{r}_j \tag{A-36}$$

Define

$$\boldsymbol{G} = \begin{bmatrix} \boldsymbol{X}_1\boldsymbol{H} & \cdots \boldsymbol{X}_M\boldsymbol{H} \end{bmatrix} \in \mathbb{R}^{N \times ML_r}. \tag{A-37}$$

Using these notations, the log-likelihood term in Equation A-28 is:

$$\mathbb{E}_{q(\boldsymbol{h})}\left[\ln p(\boldsymbol{y} \mid \boldsymbol{r}, \boldsymbol{h}; \sigma^2)\right] \overset{c}{=} -\mathbb{E}_{q(\boldsymbol{h})}\left[\sum_{j=1}^{J} \frac{1}{2\sigma_j^2}\left(\boldsymbol{y}_j - \boldsymbol{G}\boldsymbol{r}_j - \boldsymbol{P}\boldsymbol{\ell}_j\right)^\top \left(\boldsymbol{y}_j - \boldsymbol{G}\boldsymbol{r}_j - \boldsymbol{P}\boldsymbol{\ell}_j\right)\right] \tag{A-38}$$

$$\overset{c}{=} -\sum_{j=1}^{J} \frac{1}{2\sigma_j^2}\left(-2\left(\boldsymbol{y}_j - \boldsymbol{P}\boldsymbol{\ell}_j\right)^\top \mathbb{E}_{q(\boldsymbol{h})}[\boldsymbol{G}]\boldsymbol{r}_j + \boldsymbol{r}_j^\top \mathbb{E}_{q(\boldsymbol{h})}\left[\boldsymbol{G}^\top\boldsymbol{G}\right]\boldsymbol{r}_j\right) \tag{A-39}$$

Grouping quadratic terms w.r.t. $\boldsymbol{r}_j$ in the log-likelihood and log-priors (Equations A-36 and A-39) into

$$\boldsymbol{W}_r = \frac{1}{\sigma_j^2}\mathbb{E}_{q(\boldsymbol{h})}\left[\boldsymbol{G}^\top\boldsymbol{G}\right] + \sum_{k=1}^{K} \boldsymbol{\Phi}_{jk}\boldsymbol{\Sigma}_k^{-1}, \tag{A-40}$$

and other terms

$$\boldsymbol{b}_r^T = \frac{1}{\sigma_j^2}\left(\boldsymbol{y}_j - \boldsymbol{P}\boldsymbol{\ell}_j\right)^\top \mathbb{E}_{q(\boldsymbol{h})}[\boldsymbol{G}]. \tag{A-41}$$

Using the grouped terms and the mean-field factorization ($q(\boldsymbol{r}) = \prod_j q(\boldsymbol{r}_j)$) yields a factor of the log-posterior (Equation A-28) in the quadratic form:

$$\ln q(\boldsymbol{r}_j) \overset{c}{=} -\frac{1}{2}(\boldsymbol{h} - \boldsymbol{W}_h^{-1}\boldsymbol{b}_h)^\top \boldsymbol{W}_h(\boldsymbol{h} - \boldsymbol{W}_h^{-1}\boldsymbol{b}_h). \tag{A-42}$$

As in the previous section, this factor is in Gaussian log-density form and thus, the posterior distribution of $\boldsymbol{r}_j$ is

$$q(\boldsymbol{r}_j) = \boldsymbol{\mathcal{N}}(\boldsymbol{\mu}_{r_j}, \boldsymbol{\Sigma}_{r_j}) \tag{A-43}$$

and its variational parameters are updated through

$$\boldsymbol{\Sigma}_{r_j} = \boldsymbol{W}_r^{-1} \tag{A-44}$$

$$= \left(\frac{1}{\sigma_j^2}\mathbb{E}_{q(\boldsymbol{h})}\left[\boldsymbol{G}^\top\boldsymbol{G}\right] + \sum_{k=1}^{K} \boldsymbol{\Phi}_{jk}\boldsymbol{\Sigma}_k^{-1}\right)^{-1} \in \mathbb{R}^{ML_r \times ML_r} \tag{A-45}$$

and

$$\boldsymbol{\mu}_{r_j} = \boldsymbol{W}_r^{-1}\boldsymbol{b}_r \tag{A-46}$$

$$= \boldsymbol{\Sigma}_{r_j}\frac{1}{\sigma_j^2}\mathbb{E}_{q(\boldsymbol{h})}[\boldsymbol{G}]^\top\left(\boldsymbol{y}_j - \boldsymbol{P}\boldsymbol{\ell}_j\right) \in \mathbb{R}^{ML_r} \tag{A-47}$$

The expectation $\mathbb{E}_{q(\boldsymbol{h})}[\boldsymbol{G}]$ by replacing $\boldsymbol{h}$ in the Toeplitz matrix by μ_h derived in Equation

A-19. For the computation of the quadratic expectation in $\boldsymbol{\Sigma}_{r_j}$, we first expand it as:

$$\mathbb{E}_{q(\boldsymbol{h})}\left[\boldsymbol{G}^\top \boldsymbol{G}\right] = \mathbb{E}_{q(\boldsymbol{h})}\left[\begin{bmatrix}\boldsymbol{H}^\top \boldsymbol{X}_1^\top \\ \vdots \\ \boldsymbol{H}^\top \boldsymbol{X}_M^\top\end{bmatrix}\begin{bmatrix}\boldsymbol{X}_1\boldsymbol{H} & \cdots & \boldsymbol{X}_M\boldsymbol{H}\end{bmatrix}\right] \tag{A-48}$$

$$\tag{A-49}$$

$$= \begin{bmatrix}\mathbb{E}_{q(\boldsymbol{h})}\left[\boldsymbol{H}^\top \boldsymbol{X}_1^\top \boldsymbol{X}_1 \boldsymbol{H}\right] & \cdots & \mathbb{E}_{q(\boldsymbol{h})}\left[\boldsymbol{H}^\top \boldsymbol{X}_1^\top \boldsymbol{X}_M \boldsymbol{H}\right] \\ \vdots & \ddots & \vdots \\ \mathbb{E}_{q(\boldsymbol{h})}\left[\boldsymbol{H}^\top \boldsymbol{X}_M^\top \boldsymbol{X}_1 \boldsymbol{H}\right] & \cdots & \mathbb{E}_{q(\boldsymbol{h})}\left[\boldsymbol{H}^\top \boldsymbol{X}_M^\top \boldsymbol{X}_M \boldsymbol{H}\right]\end{bmatrix}. \tag{A-50}$$

Define

$$\boldsymbol{H} = \begin{bmatrix}\boldsymbol{h_{(1)}} & \boldsymbol{h_{(2)}} & \ldots & \boldsymbol{h_{(L_r)}}\end{bmatrix} \tag{A-51}$$

where $\boldsymbol{h}(i)$ is ith column of Toeplitz matrix $\boldsymbol{H}$, given by

$$\boldsymbol{h}_{(i)} = \begin{bmatrix}\boldsymbol{0}_{i-1} \\ \boldsymbol{h} \\ \boldsymbol{0}_{L_r - i}\end{bmatrix} \in \mathbb{R}^{L_h + L_r - 1}. \tag{A-52}$$

A $L_r \times L_r$ block of $\mathbb{E}_{q(\boldsymbol{h})}\left[\boldsymbol{G}^\top \boldsymbol{G}\right]$ is then computed as:

$$\mathbb{E}_{q(\boldsymbol{h})}\left[\boldsymbol{H}\boldsymbol{X}_m^\top \boldsymbol{X}_n \boldsymbol{H}\right] = \mathbb{E}_{q(\boldsymbol{h})}\left[\begin{bmatrix}\boldsymbol{h}_{(1)}^\top \\ \vdots \\ \boldsymbol{h}_{(L-1)}^\top\end{bmatrix}\boldsymbol{X}_m^\top \boldsymbol{X}_n \begin{bmatrix}\boldsymbol{h}_{(1)} & \cdots & \boldsymbol{h}_{(L-1)}\end{bmatrix}\right] \tag{A-53}$$

$$= \mathbb{E}_{q(\boldsymbol{h})}\left[\begin{bmatrix}\boldsymbol{h}_{(1)}^\top \boldsymbol{X}_m^\top \boldsymbol{X}_n \boldsymbol{h}_{(1)} & \cdots & \boldsymbol{h}_{(1)}^\top \boldsymbol{X}_m^\top \boldsymbol{X}_n \boldsymbol{h}_{(L-1)} \\ \vdots & \ddots & \vdots \\ \boldsymbol{h}_{(L-1)}^\top \boldsymbol{X}_m^\top \boldsymbol{X}_n \boldsymbol{h}_{(1)} & \cdots & \boldsymbol{h}_{(L-1)}^\top \boldsymbol{X}_m^\top \boldsymbol{X}_n \boldsymbol{h}_{(L-1)}\end{bmatrix}\right] \tag{A-54}$$

$$= \begin{bmatrix}\mathbb{E}_{q(\boldsymbol{h})}\left[\boldsymbol{h}_{(1)}^\top \boldsymbol{X}_m^\top \boldsymbol{X}_n \boldsymbol{h}_{(1)}\right] & \cdots & \mathbb{E}_{q(\boldsymbol{h})}\left[\boldsymbol{h}_{(1)}^\top \boldsymbol{X}_m^\top \boldsymbol{X}_n \boldsymbol{h}_{(L-1)}\right] \\ \vdots & \ddots & \vdots \\ \mathbb{E}_{q(\boldsymbol{h})}\left[\boldsymbol{h}_{(L-1)}^\top \boldsymbol{X}_m^\top \boldsymbol{X}_n \boldsymbol{h}_{(1)}\right] & \cdots & \mathbb{E}_{q(\boldsymbol{h})}\left[\boldsymbol{h}_{(L-1)}^\top \boldsymbol{X}_m^\top \boldsymbol{X}_n \boldsymbol{h}_{(L-1)}\right]\end{bmatrix} \tag{A-55}$$

Finally, (a, b) entry of such block is computed by

$$\begin{aligned}\mathbb{E}_{q(\boldsymbol{h})}\left[\boldsymbol{h}_{(a)}^\top \boldsymbol{X}_m^\top \boldsymbol{X}_n \boldsymbol{h}_{(b)}\right] &= \mathbb{E}_{q(\boldsymbol{h})}\left[\boldsymbol{h}_{(a)}\right]\boldsymbol{X}_m^\top \boldsymbol{X}_n \mathbb{E}_{q(\boldsymbol{h})}\left[\boldsymbol{h}_{(b)}\right] \\ &\quad + \operatorname{tr}\left(\boldsymbol{X}_m^\top \boldsymbol{X}_n \operatorname{Circ}\left(\begin{bmatrix}\boldsymbol{\Sigma}_h & \boldsymbol{0}_{L_h \times (L_r-1)} \\ \boldsymbol{0}_{(L_r-1)\times L_h} & \boldsymbol{0}_{(L_r-1)\times(L_r-1)}\end{bmatrix}, b-1, a-1\right)\right)\end{aligned} \tag{A-56}$$

A-3 Update for $q(\boldsymbol{z})$

CAVI update rule

$$\ln q(\boldsymbol{z})^{(t)} \stackrel{c}{=} \mathbb{E}_{q(\boldsymbol{h})^{(t)}q(\boldsymbol{r})^{(t)}q(\boldsymbol{\beta})^{(t-1)}}\left[\ln p\left(\boldsymbol{y},\boldsymbol{h},\boldsymbol{r},\boldsymbol{\beta};\boldsymbol{\theta}^{(t-1)}\right)\right] \tag{A-57}$$

The Ising prior is a random Markov field model with dependence on classes of other pixels. Together with the mean-field factorization $q(\boldsymbol{z}) = \prod_j \prod_m q(\boldsymbol{z}_j^m)$, this leads the CAVI update rule of a factor log-posterior to be:

$$\ln q(z_j^m) \stackrel{c}{=} \mathbb{E}_{q(\boldsymbol{h})q(\boldsymbol{r})q(\boldsymbol{\beta})q(z_{\setminus j}^m)}\left[\ln p\left(\boldsymbol{y},\boldsymbol{h},\boldsymbol{r},\boldsymbol{\beta};\boldsymbol{\theta}\right)\right] \tag{A-58}$$

where $z_{\setminus j}^m$ represents all other pixels except for z_j^m, and thus the expectation of the complete data likelihood is also taken over posterior distribution of other pixels.

Omitting constant terms in the log-posterior gives:

$$\ln q(\boldsymbol{z}_j^m) \stackrel{c}{=} \mathbb{E}_{q(z_{\setminus j}^m)q(\boldsymbol{r})}\left[\ln p(\boldsymbol{r}|\boldsymbol{z})\right] + \mathbb{E}_{q(z_{\setminus j}^m)q(\beta_m)}\left[\ln p(\boldsymbol{z}|\beta_m)\right] \tag{A-59}$$

$$\stackrel{c}{=} \mathbb{E}_{q(\boldsymbol{r}_j^m)}\left[\ln p(\boldsymbol{r}_j^m|z_j^m)\right] + \mathbb{E}_{q(z_{\setminus j}^m)q(\beta_m)}\left[\ln p(z_j^m|\beta_m)\right] \tag{A-60}$$

Note that given neighboring pixels, z_j^m is conditionally independent of all other pixels.

The second expectation term is evaluated as

$$\mathbb{E}_{q(z_{\setminus j}^m)q(\beta_m)}\left[\ln p(\boldsymbol{z}_j^m|\boldsymbol{\beta})\right] = \mathbb{E}_{q(z_{\setminus j}^m)q(\beta_m)}\left[-\ln Z(\beta_m) + \beta_m \sum_{j=1}^{J}\sum_{i\in\delta_j} I(z_i^m = z_j^m)\right] \tag{A-61}$$

$$\stackrel{c}{=} \mathbb{E}_{q(z_{\setminus j}^m)q(\beta_m)}\left[\beta_m \sum_{j=1}^{J}\sum_{i\in\delta_j} I(z_i^m = z_j^m)\right] \tag{A-62}$$

$$= \mathbb{E}_{q(z_{\setminus j}^m)}\left[\sum_{j=1}^{J}\sum_{i\in\delta_j} I(z_i^m = z_j^m)\right]\mu_{\beta_m} \tag{A-63}$$

$$= \mathbb{E}_{q(z_{\setminus j}^m)}\left[\sum_{j=1}^{J}\sum_{i\in\delta_j}\sum_{k=1}^{K} I(z_j^m = k)I(z_i^m = k)\right]\mu_{\beta_m} \tag{A-64}$$

$$\stackrel{c}{=} \sum_{k=1}^{K} I(z_j^m = k)\mathbb{E}_{q(z_{\setminus j}^m)}\left[\sum_{i\in\delta_j} I(z_i^m = k)\right]\mu_{\beta_m} \tag{A-65}$$

$$\stackrel{c}{=} \sum_{k=1}^{K} I(z_j^m = k)\sum_{i\in\delta_j}\phi_{ik}^m\mu_{\beta_m}, \tag{A-66}$$

where δ_j represents a set of all neighboring pixels of j and $\phi_{jk}^m = q(z_j^m = k)$ is the posterior probability defined in the derivation of $q(\boldsymbol{r})$.

Knowing the independence in the NRF prior $p(\boldsymbol{r}|\boldsymbol{z}) = \prod_j \prod_m \prod_k q(\boldsymbol{r}_j^m|z_j^m = k)$, the first

expectation term in Equation A-60 is

$$\begin{aligned}
&\mathbb{E}_{q(\boldsymbol{r}_j^m)}\left[\ln p(\boldsymbol{r}_j^m|z_j^m=k)\right] \\
&\quad= \mathbb{E}_{q(\boldsymbol{r}_j^m)}\left[\sum_{m=1}^{M}\sum_{k=1}^{K}\boldsymbol{I}(z_j^m=k)\left(-\frac{L_r}{2}\ln(2\pi)-\frac{1}{2}\ln(\det\boldsymbol{\Sigma}_{km})-\frac{1}{2}\boldsymbol{r}_j^{m\top}\boldsymbol{\Sigma}_{km}^{-1}\boldsymbol{r}_j^m\right)\right] \quad \text{(A-67)}\\
&\quad= \sum_{m=1}^{M}\sum_{k=1}^{K}\boldsymbol{I}(z_j^m=k)\left(-\frac{L_r}{2}\ln(2\pi)-\frac{1}{2}\ln(\det\boldsymbol{\Sigma}_{km})-\frac{1}{2}\mathbb{E}_{q(\boldsymbol{r}_j^m)}\left[\boldsymbol{r}_j^{m\top}\boldsymbol{\Sigma}_{km}^{-1}\boldsymbol{r}_j^m\right]\right), \quad \text{(A-68)}
\end{aligned}$$

where the quadratic expectation w.r.t. $\boldsymbol{r}_j^m$ is evaluated as:

$$\mathbb{E}_{q(\boldsymbol{r}_j^m)}\left[\boldsymbol{r}_j^{m\top}\boldsymbol{\Sigma}_{km}^{-1}\boldsymbol{r}_j^m\right]=\boldsymbol{\mu}_{r_j^m}{}^{\top}\boldsymbol{\Sigma}_{km}^{-1}\boldsymbol{\mu}_{r_j^m}+\mathrm{tr}(\boldsymbol{\Sigma}_{km}^{-1}\boldsymbol{\Sigma}_{r_j^m}). \quad \text{(A-69)}$$

Substituting Equations A-66 and A-68 into the log-posterior (Equation A-60), we obtain unnormalized posterior probability:

$$q(z_j^m=k)\propto\boldsymbol{\mathcal{N}}(\boldsymbol{\mu}_{r_j^m};\mathbf{0},\boldsymbol{\Sigma}_{km})\exp\left(-\mathrm{tr}(\boldsymbol{\Sigma}_{km}^{-1}\boldsymbol{\Sigma}_{r_j^m})+\sum_{i\in\delta_j}\phi_{ik}^m\mu_{\beta_m}\right) \quad \text{(A-70)}$$

A-4 Update for $q(\boldsymbol{\beta})$

CAVI update rule:

$$\ln q(\boldsymbol{\beta})^{(t)}\overset{c}{=}\mathbb{E}_{q(\boldsymbol{h})^{(t)}q(\boldsymbol{r})^{(t)}q(\boldsymbol{z})^{(t)}}\left[\ln p\left(\boldsymbol{y},\boldsymbol{h},\boldsymbol{r},\boldsymbol{\beta};\boldsymbol{\theta}^{(t-1)}\right)\right] \quad \text{(A-71)}$$

Starting with omitting constant terms in Equation A-71, we have:

$$\begin{aligned}
\ln q(\beta_m)&\overset{c}{=}\mathbb{E}_{q(\boldsymbol{h})q(\boldsymbol{r})q(\boldsymbol{z})}\left[\ln p\left(\boldsymbol{y},\boldsymbol{h},\boldsymbol{r},\boldsymbol{\beta};\boldsymbol{\theta}\right)\right] &\text{(A-72)}\\
&\overset{c}{=}\mathbb{E}_{q(\boldsymbol{z}^m)}\left[\ln p(\boldsymbol{z}^m|\beta_m)\right]+\ln p(\beta_m) &\text{(A-73)}
\end{aligned}$$

Since the prior distribution on $\boldsymbol{\beta}$ is defined as

$$p(\boldsymbol{\beta})=\prod_{m=1}^{M}p(\beta_m)=\prod_{m=1}^{M}\mathcal{U}(\beta_m,u_1,u_2), \quad \text{(A-74)}$$

the second term in Equation A-73 is

$$\ln q(\beta_m)=-\ln(u_2-u_1). \quad \text{(A-75)}$$

The log-expectation of the Ising model is evaluated as:

$$\begin{aligned}
\mathbb{E}_{q(\boldsymbol{z}^m)}\left[\ln p(\boldsymbol{z}^m|\beta_m)\right]&=\mathbb{E}_{q(\boldsymbol{z}^m)}\left[-\ln Z(\beta_m)+\beta_m\sum_{j=1}^{J}\sum_{i\in\delta_j}\sum_{k=1}^{K}I(z_j^m=k)I(z_i^m=k)\right] &\text{(A-76)}\\
&=-\ln Z(\beta_m)+\beta_m\sum_{j=1}^{J}\sum_{i\in\delta_j}\sum_{k=1}^{K}\phi_{jk}^m\phi_{ik}^m\ . &\text{(A-77)}
\end{aligned}$$

However, the normalization constant $\ln Z(\beta_m)$ from the Ising model is intractable.

To overcome this intractability, we first replace $p(\boldsymbol{z}^m|\beta_m)$ by pseudo-likelihood:

$$p^{PL}(\boldsymbol{z}^m|\beta_m) = \prod_{j=1}^{J} p(z_j^m|\beta_m, z_{\setminus j}^m) = \prod_{j=1}^{J} p(z_j^m|\beta_m, z_{\delta_j}^m), \tag{A-78}$$

where the conditional distribution is

$$p(z_j^m|\beta_m, z_{\delta_j}^m) = \frac{\exp\left(\beta_m \sum_{i\in\delta_j} \sum_{k=1}^{K} I(z_j^m = k) I(z_i^m = k)\right)}{\sum_{z_j^m \in \{1,2,\dots K\}} \exp\left(\beta_m \sum_{i\in\delta_j} \sum_{k=1}^{K} I(z_j^m = k) I(z_i^m = k)\right)} \tag{A-79}$$

$$= \frac{\exp\left(\beta_m \sum_{i\in\delta_j} \sum_{k=1}^{K} I(z_j^m = k) I(z_i^m = k)\right)}{\sum_{k=1}^{K} \exp\left(\beta_m \sum_{i\in\delta_j} I(z_i^m = k)\right)} \tag{A-80}$$

Using this pseudo-likelihood, the log-expectation of the Ising model yields is approximated as:

$$\mathbb{E}_{q(\boldsymbol{z}^m)}\left[\ln p(\boldsymbol{z}^m|\beta_m)\right] \approx \mathbb{E}_{q(\boldsymbol{z}^m)}\left[\ln p^{PL}(\boldsymbol{z}^m|\beta_m)\right] \tag{A-81}$$

$$= -\sum_{j=1}^{J} \mathbb{E}_{q(\boldsymbol{z}^m)}\left[\ln\left(\sum_{k=1}^{K} \exp\left(\beta_m \sum_{i\in\delta_j} I(z_i^m = k)\right)\right)\right] + \sum_{j=1}^{J}\sum_{i\in\delta_j}\sum_{k=1}^{K} \phi_{jk}^m \phi_{ik}^m \tag{A-82}$$

Assuming that one pixel has four neighboring pixels in 2D grid, the evaluation of the expectation in Equation A-82 requires computation of JK^4 terms, which is computationally expensive.

Following [34], this term is approximated by applying Jensen's inequality twice, resulting in

$$\mathbb{E}_{q(\boldsymbol{z}^m)}\left[\ln\left(\sum_{k=1}^{K} \exp\left(\beta_m \sum_{i\in\delta_j} I(z_i^m = k)\right)\right)\right] \approx \ln\left(\sum_{k=1}^{K} \exp\left(\beta_m \sum_{i\in\delta_j} \phi_{ik}^m\right)\right). \tag{A-83}$$

Using this approximation, we finally have unnormalized posterior density:

$$q(\beta_m) \propto \prod_{j=1}^{J} \frac{\exp\left(\beta_m \sum_{j=1}^{J} \sum_{i\in\delta_j} \sum_{k=1}^{K} \phi_{jk}^m \phi_{ik}^m\right)}{\sum_{k=1}^{K} \exp\left(\beta_m \sum_{i\in\delta_j} \phi_{ik}^m\right)} \mathcal{U}(\beta_m, u_1, u_2) \tag{A-84}$$

We use numerical quadrature to compute normalization constant and mean μ_β.

A-5 Update for $\boldsymbol{\ell}$

The closed-form update equation for $\boldsymbol{\ell}$ in the M-step is obtained by maximizing the ELBO with respect to $\boldsymbol{\ell}$ as follows.

$$\arg\max_{\boldsymbol{\ell}} \ \mathbb{E}_{q(\boldsymbol{h})q(\boldsymbol{r})q(\boldsymbol{z})q(\boldsymbol{\beta})}\left[\ln p(\boldsymbol{y},\boldsymbol{h},\boldsymbol{r},\boldsymbol{\beta}\mid\boldsymbol{\theta})\right]$$

$$= \arg\max_{\boldsymbol{\ell}} \mathbb{E}^{q(\boldsymbol{h})q(\boldsymbol{r})}\left[\sum_{j=1}^{J}\ln p(\boldsymbol{y}_j\mid\boldsymbol{h},\boldsymbol{r}_j,\boldsymbol{\theta})\right] \tag{A-85}$$

$$= \arg\max_{\boldsymbol{\ell}} \mathbb{E}^{q(\boldsymbol{h})q(\boldsymbol{r})}\left[\sum_{j=1}^{J}-\frac{1}{2}\left(\boldsymbol{y}_j-\sum_{m=1}^{M}\boldsymbol{X}_m\boldsymbol{H}\boldsymbol{r}_j-\boldsymbol{P}\boldsymbol{\ell}_j\right)^{\top}\left(\boldsymbol{y}_j-\sum_{m=1}^{M}\boldsymbol{X}_m\boldsymbol{H}\boldsymbol{r}_j-\boldsymbol{P}\boldsymbol{\ell}_j\right)\right] \tag{A-86}$$

$$= \arg\min_{\boldsymbol{\ell}} \sum_{j=1}^{J}\mathbb{E}_{q(\boldsymbol{h})q(\boldsymbol{r})}\left[\left(\boldsymbol{y}_j-\boldsymbol{P}\boldsymbol{\ell}_j\right)^{\top}\left(\boldsymbol{y}_j-\boldsymbol{P}\boldsymbol{\ell}_j\right)-2\left(\sum_{m=1}^{M}\boldsymbol{X}_m\boldsymbol{H}\boldsymbol{r}_j^m\right)^{\top}\left(\boldsymbol{y}_j-\boldsymbol{P}\boldsymbol{\ell}_j\right)\right] \tag{A-87}$$

Recall the following notations:

$$\boldsymbol{G}=\begin{bmatrix}\boldsymbol{X}_1\boldsymbol{H} & \boldsymbol{X}_2\boldsymbol{H} & \cdots & \boldsymbol{X}_M\boldsymbol{H}\end{bmatrix}\in\mathbb{R}^{N\times(ML_r)},\qquad \boldsymbol{r}_j=\begin{bmatrix}r_j^1\\ r_j^2\\ \vdots\\ r_j^M\end{bmatrix}\in\mathbb{R}^{(ML_r)}. \tag{A-88}$$

Using these, the last step becomes:

$$\arg\min_{\boldsymbol{\ell}}\sum_{j=1}^{J}-\frac{1}{2}\left[-2\boldsymbol{y}_j^{\top}\boldsymbol{P}\boldsymbol{\ell}_j+\boldsymbol{\ell}_j^{\top}\boldsymbol{P}^{\top}\boldsymbol{P}\boldsymbol{\ell}_j-2\left(\mathbb{E}_{q(\boldsymbol{h})}\left[\boldsymbol{G}\right]\boldsymbol{\mu}_{r_j}\right)^{\top}\boldsymbol{P}\boldsymbol{\ell}_j\right] \tag{A-89}$$

$$= \arg\min_{\boldsymbol{\ell}}\sum_{j=1}^{J}-\frac{1}{2}\left[-2\left(\boldsymbol{y}_j-\mathbb{E}_{q(\boldsymbol{h})}\left[\boldsymbol{G}\right]\boldsymbol{\mu}_{r_j}\right)^{\top}\boldsymbol{P}\boldsymbol{\ell}_j+\boldsymbol{\ell}_j^{\top}\boldsymbol{P}^{\top}\boldsymbol{P}\boldsymbol{\ell}_j\right]. \tag{A-90}$$

Since $\boldsymbol{\ell}$ is separable for each j, we have:

$$\Rightarrow \arg\min_{\boldsymbol{\ell}_j}-2\left(\boldsymbol{y}_j-\mathbb{E}_{q(\boldsymbol{h})}\left[\boldsymbol{G}\right]\boldsymbol{\mu}_{r_j}\right)^{\top}\boldsymbol{P}\boldsymbol{\ell}_j+\boldsymbol{\ell}_j^{\top}\boldsymbol{P}^{\top}\boldsymbol{P}\boldsymbol{\ell}_j \tag{A-91}$$

Let

$$\boldsymbol{b}^{\top}=\left(\boldsymbol{y}_j-\mathbb{E}_{q(\boldsymbol{h})}\left[\boldsymbol{G}\right]\boldsymbol{\mu}_{r_j}\right)^{\top}\boldsymbol{P},\quad \boldsymbol{W}=\boldsymbol{P}^{\top}\boldsymbol{P} \tag{A-92}$$

The minimization is in quadratic form as:

$$\arg\min_{\boldsymbol{\ell}_j}\left(\boldsymbol{\ell}_j-\boldsymbol{W}^{-1}\boldsymbol{b}\right)^{\top}\boldsymbol{W}\left(\boldsymbol{\ell}_j-\boldsymbol{W}^{-1}\boldsymbol{b}\right)-\boldsymbol{b}^{\top}\boldsymbol{W}^{-1}\boldsymbol{b} \tag{A-93}$$

Thus, the update solution is:

$$\boldsymbol{\ell}_j=\boldsymbol{W}^{-1}\boldsymbol{b}=\left(\boldsymbol{P}^{\top}\boldsymbol{P}\right)^{-1}\boldsymbol{P}^{\top}\left(\boldsymbol{y}_j-\mathbb{E}_{q(\boldsymbol{h})}\left[\boldsymbol{G}\right]\boldsymbol{\mu}_{r_j}\right) \tag{A-94}$$

A-6 Update for σ^2

Similarly to the previous derivation, the closed-form update equation for σ^2 in the M-step is obtained by maximizing the ELBO.

$$\begin{aligned}
&\arg\max_{\sigma^2} \ \mathbb{E}_{q(\boldsymbol{h})q(\boldsymbol{r})q(\boldsymbol{z})q(\boldsymbol{\beta})}\left[\ln p(\boldsymbol{y},\boldsymbol{h},\boldsymbol{r},\boldsymbol{\beta}\mid\boldsymbol{\theta})\right] \\
&= \arg\max_{\sigma^2} \mathbb{E}_{q(\boldsymbol{h})q(\boldsymbol{r})}\left[\sum_{j=1}^{J}\ln p(\boldsymbol{y}_j\mid\boldsymbol{h},\boldsymbol{r}_j,\boldsymbol{\theta})\right] \qquad \text{(A-95)} \\
&= \arg\max_{\sigma_j^2}\sum_{j=1}^{J} -\frac{N}{2}\ln(\sigma_j^2) - \frac{1}{2\sigma_j^2}\mathbb{E}_{q(\boldsymbol{h})q(\boldsymbol{r})}\left[\left(\sum_{m=1}^{M}\boldsymbol{X}_m\boldsymbol{H}\boldsymbol{r}_j^m\right)^\top\left(\sum_{m=1}^{M}\boldsymbol{X}_m\boldsymbol{H}\boldsymbol{r}_j^m\right)\right] \\
&\qquad - \frac{1}{2\sigma_j^2}\left(\boldsymbol{y}_j - 2\sum_{m=1}^{M}\mathbb{E}_{q(\boldsymbol{h})q(\boldsymbol{r})}\left[\boldsymbol{X}_m\boldsymbol{H}\boldsymbol{r}_j^m\right] - \boldsymbol{P}\boldsymbol{\ell}_j\right)^\top\left(\boldsymbol{y}_j - \boldsymbol{P}\boldsymbol{\ell}_j\right) \qquad \text{(A-96)}
\end{aligned}$$

Again, using the notations in Equation A-88, the last step becomes:

$$\begin{aligned}
&\arg\max_{\sigma_j^2}\sum_{j=1}^{J} -\frac{N}{2}\ln(\sigma_j^2) - \frac{1}{2\sigma_j^2}\left(\boldsymbol{\mu}_{r_j}^\top\mathbb{E}_{q(\boldsymbol{h})}\left[\boldsymbol{G}^\top\boldsymbol{G}\right]\boldsymbol{\mu}_{r_j} + \operatorname{tr}\left(\mathbb{E}_{q(\boldsymbol{h})}[\boldsymbol{G}^\top\boldsymbol{G}]\boldsymbol{\Sigma}_{r_j}\right)\right) \qquad \text{(A-97)} \\
&\qquad - \frac{1}{2\sigma_j^2}(\boldsymbol{y}_j - 2\mathbb{E}_{q(\boldsymbol{h})}[\boldsymbol{G}]\boldsymbol{\mu}_{r_j} - \boldsymbol{P}\boldsymbol{\ell}_j)^\top(\boldsymbol{y}_j - \boldsymbol{P}\boldsymbol{\ell}_j) \qquad \text{(A-98)}
\end{aligned}$$

Setting derivative w.r.t. σ_j^2 to zero, we have:

$$\sigma_j^2 = \frac{(\boldsymbol{y}_j - 2\mathbb{E}_{q(\boldsymbol{h})}[\boldsymbol{G}]\boldsymbol{\mu}_{r_j} - \boldsymbol{P}\boldsymbol{\ell}_j)^\top(\boldsymbol{y}_j - \boldsymbol{P}\boldsymbol{\ell}_j) + \boldsymbol{\mu}_{r_j}^\top\mathbb{E}_{q(\boldsymbol{h})}[\boldsymbol{G}^\top\boldsymbol{G}]\boldsymbol{\mu}_{r_j} + \operatorname{tr}\left(\mathbb{E}_{q(\boldsymbol{h})}[\boldsymbol{G}^\top\boldsymbol{G}]\boldsymbol{\Sigma}_{r_j}\right)}{N} \qquad \text{(A-99)}$$

A-7 Update for σ_h^2

Again, the ELBO is maximized w.r.t. σ_h^2 to obtain the closed-form update equation as follows.

$$\begin{aligned}
&\arg\max_{\sigma_h^2} \ \mathbb{E}_{q(\boldsymbol{h})q(\boldsymbol{r})q(\boldsymbol{z})q(\boldsymbol{\beta})}\left[\ln p(\boldsymbol{y},\boldsymbol{h},\boldsymbol{r},\boldsymbol{\beta}\mid\boldsymbol{\theta})\right] \\
&= \arg\max_{\sigma_h^2} \mathbb{E}_{q(\boldsymbol{h})}\left[\ln p(\boldsymbol{h}\mid\sigma_h^2)\right] \qquad \text{(A-100)} \\
&= \arg\max_{\sigma_h^2} \ -\frac{L_h}{2}\ln(2\pi) - \frac{1}{2}\ln\det(\sigma_h^2\boldsymbol{R}) - \frac{1}{2\sigma_h^2}\mathbb{E}_{q(\boldsymbol{h})}\left[(\boldsymbol{h}-\boldsymbol{\mu}_{\boldsymbol{h}\boldsymbol{0}})^\top\boldsymbol{R}^{-1}(\boldsymbol{h}-\boldsymbol{\mu}_{\boldsymbol{h}\boldsymbol{0}})\right] \qquad \text{(A-101)} \\
&= \arg\max_{\sigma_h^2} \ -\frac{L_h}{2}\ln\sigma_h^2 - \frac{1}{2\sigma_h^2}\left((\boldsymbol{\mu}_h-\boldsymbol{\mu}_{h0})^\top\boldsymbol{R}^{-1}(\boldsymbol{\mu}_h-\boldsymbol{\mu}_{h0}) + \operatorname{tr}\left(\boldsymbol{R}^{-1}\boldsymbol{\Sigma}_h\right)\right) \qquad \text{(A-102)}
\end{aligned}$$

Again, setting derivative w.r.t. σ_h^2 to zero as:

$$-\frac{L_h}{\sigma_h} + \frac{1}{\sigma_h^3}\left((\boldsymbol{\mu}_h - \boldsymbol{\mu}_{h0})^\top \boldsymbol{R}^{-1}(\boldsymbol{\mu}_h - \boldsymbol{\mu}_{h0}) + \mathrm{tr}\left(\boldsymbol{R}^{-1}\boldsymbol{\Sigma}_h\right)\right) = 0, \tag{A-103}$$

we have

$$\sigma_h^2 = \frac{(\boldsymbol{\mu}_h - \boldsymbol{\mu}_{h0})^\top \boldsymbol{R}^{-1}(\boldsymbol{\mu}_h - \boldsymbol{\mu}_{h0}) + \mathrm{tr}\left(\boldsymbol{R}^{-1}\boldsymbol{\Sigma}_h\right)}{L_h} \tag{A-104}$$

As was done in [7], we can also impose an exponential prior on σ_h^2, i.e., $p(\sigma_h^2) = \lambda \exp(-\lambda\sigma_h^2)$ with scaling parameter λ. The maximization becomes:

$$\arg\max_{\sigma_h^2} \mathbb{E}^{q(\boldsymbol{h})}\left[\ln p(\boldsymbol{h} \mid \sigma_h^2)\right] + \ln\left(\lambda \exp(-\lambda\sigma_h^2)\right)$$

Following the same step, we have the solution:

$$\sigma_h^2 = \frac{-L_h + \sqrt{8\lambda(\boldsymbol{\mu}_h - \boldsymbol{\mu}_{h0})^\top \boldsymbol{R}^{-1}(\boldsymbol{\mu}_h - \boldsymbol{\mu}_{h0}) + 8\lambda\ \mathrm{tr}\left(\boldsymbol{R}^{-1}\boldsymbol{\Sigma}_h\right) + (L_h)^2}}{4\lambda} \tag{A-105}$$

Appendix B

Computation of ELBO

Recall the ELBO in Equation 3-38:

$$\mathcal{L}\left(q(\boldsymbol{h})q(\boldsymbol{r})q(\boldsymbol{z})q(\boldsymbol{\beta}),\boldsymbol{\theta}\right)$$

$$= \mathbb{E}_{q(\boldsymbol{h})q(\boldsymbol{r})q(\boldsymbol{z})q(\boldsymbol{\beta})}[\ln p(\boldsymbol{y},\boldsymbol{h},\boldsymbol{r},\boldsymbol{z},\boldsymbol{\beta};\boldsymbol{\theta})] - \mathbb{E}_{q(\boldsymbol{h})q(\boldsymbol{r})q(\boldsymbol{z})q(\boldsymbol{\beta})}[\ln q(\boldsymbol{h})q(\boldsymbol{r})q(\boldsymbol{z})q(\boldsymbol{\beta})]. \tag{B-1}$$

We begin the evaluation of the ELBO in Equation 3-38 by expanding the entropy term as:

$$-\mathbb{E}_{q(\boldsymbol{h})q(\boldsymbol{r})q(\boldsymbol{z})q(\boldsymbol{\beta})}[\ln q(\boldsymbol{h})q(\boldsymbol{r})q(\boldsymbol{z})q(\boldsymbol{\beta})]$$

$$= -\mathbb{E}_{q(\boldsymbol{h})}[\ln q(\boldsymbol{h})] - \mathbb{E}_{q(\boldsymbol{r})}[\ln q(\boldsymbol{r})] - \mathbb{E}_{q(\boldsymbol{z})}[\ln q(\boldsymbol{z})] - \mathbb{E}_{q(\boldsymbol{\beta})}[\ln q(\boldsymbol{\beta})] \tag{B-2}$$

Because the posteriors of $\boldsymbol{h}$ and $\boldsymbol{r}$ are both Gaussian, the first entropy terms are given by:

$$-\mathbb{E}_{q(\boldsymbol{h})}[\ln q(\boldsymbol{h})] = \frac{(L_h+1)}{2}\ln(2\pi) + \frac{1}{2}\ln\det(\boldsymbol{\Sigma}_h), \tag{B-3}$$

and

$$-\mathbb{E}_{q(\boldsymbol{r})}[\ln q(\boldsymbol{r})] = -\sum_{j=1}^{J}\mathbb{E}_{q(\boldsymbol{r}_j)}[\ln q(\boldsymbol{r}_j)] \tag{B-4}$$

$$= \frac{(ML_r+1)}{2}\ln(2\pi) + \frac{1}{2}\ln\det\left(\boldsymbol{\Sigma}_{r_j}\right) \tag{B-5}$$

where $\boldsymbol{\Sigma}_h$ and $\boldsymbol{\Sigma}_{r_j}$ are the updated posterior covariances of $\boldsymbol{h}$ (Equation A-18) and $\boldsymbol{r}_j$ (Equation A-45), respectively.

For the entropy of $q(\boldsymbol{z})$, each factor of $q(\boldsymbol{z})$ is distributed as the categorical distribution. We first express $q(z_j^m)$ as:

$$q(z_j^m) = \prod_{k=1}^{K}\left(\phi_{jk}^m\right)^{\boldsymbol{I}\left(z_j^m=k\right)}, \tag{B-6}$$

where, recall that ϕ_{jk}^m denotes the posterior probability of pixel j belonging to activation class k for stimulus m. Using this expression, the entropy of the posterior distribution for $\boldsymbol{z}$ is computed as:

$$-\mathbb{E}_{q(\boldsymbol{z})}[\ln q(\boldsymbol{z})] = -\sum_{j=1}^{J}\sum_{m=1}^{M}\mathbb{E}_{q(z_j^m)}\left[\ln q(z_j^m)\right] \tag{B-7}$$

$$= -\sum_{j=1}^{J}\sum_{m=1}^{M}\mathbb{E}_{q(\boldsymbol{z}_j^m)}\left[\sum_{k=1}^{K}\boldsymbol{I}(z_j^m = k)\ln\phi_{jk}^m\right] \tag{B-8}$$

$$= -\sum_{j=1}^{J}\sum_{m=1}^{M}\sum_{k=1}^{K}\phi_{jk}^m \ln\phi_{jk}^m \tag{B-9}$$

For the last entropy $q(\boldsymbol{\beta})$, we use numerical integration again to evaluate for each m as:

$$-\mathbb{E}_{q(\beta_m)}[\ln q(\beta_m)] = \sum_{s=u_1}^{u_2}\Delta s\ q(\beta_m = s)\ln q(\beta_m = s) \tag{B-10}$$

where Δs is the grid size of the numerical integration, and recall that u_1 and u_2 denote lower and upper bounds of the uniform prior we used.

Next, we expand the the expected complete data log-likelihood term as:

$$\begin{aligned}
&\mathbb{E}_{q(\boldsymbol{h})q(\boldsymbol{r})q(\boldsymbol{z})q(\boldsymbol{\beta})}[\ln p(\boldsymbol{y},\boldsymbol{h},\boldsymbol{r},\boldsymbol{z},\boldsymbol{\beta};\boldsymbol{\theta})] \\
&= \mathbb{E}_{q(\boldsymbol{h})q(\boldsymbol{r})}[\ln p(\boldsymbol{y}|\boldsymbol{h},\boldsymbol{r};\boldsymbol{\theta})] + \mathbb{E}_{q(\boldsymbol{h})}[\ln p(\boldsymbol{h};\boldsymbol{\theta})] + \mathbb{E}_{q(\boldsymbol{r})q(\boldsymbol{z})}[\ln p(\boldsymbol{r}|\boldsymbol{z};\boldsymbol{\theta})] \\
&\qquad + \mathbb{E}_{q(\boldsymbol{z})q(\boldsymbol{\beta})}[\ln p(\boldsymbol{z}|\boldsymbol{\beta};\boldsymbol{\theta})] + \mathbb{E}_{q(\boldsymbol{\beta})}[\ln p(\boldsymbol{\beta};\boldsymbol{\theta})]
\end{aligned} \tag{B-11}$$

The first term is given by

$$\begin{aligned}
&\mathbb{E}_{q(\boldsymbol{h})q(\boldsymbol{r})}\left[\ln p(\boldsymbol{y}|\boldsymbol{h},\boldsymbol{r};\boldsymbol{\theta})\right] \\
&= \sum_{j=1}^{J}\Bigg(-\frac{N}{2}\ln(2\pi) - \frac{N}{2}\ln(\sigma_j^2) - \frac{1}{2\sigma_j^2}(\boldsymbol{y}_j - \boldsymbol{P}\boldsymbol{\ell}_j)^\top(\boldsymbol{y}_j - \mathbb{E}[\boldsymbol{G}]\boldsymbol{\mu}_{r_j} - \boldsymbol{P}\boldsymbol{\ell}_j) \\
&\qquad -\frac{1}{2\sigma_j^2}\left(\boldsymbol{\mu}_{r_j}^\top\mathbb{E}_{q(\boldsymbol{h})}\left[\boldsymbol{G}^\top\boldsymbol{G}\right]\boldsymbol{\mu}_{r_j} + \mathrm{tr}\left(\mathbb{E}_{q(\boldsymbol{h})}\left[\boldsymbol{G}^\top\boldsymbol{G}\right]\boldsymbol{\Sigma}_{r_j}\right)\right)\Bigg)
\end{aligned} \tag{B-12}$$

The second term is the cross-entropy between the posterior and prior distributions of $\boldsymbol{h}$, which is:

$$\mathbb{E}_{q(\boldsymbol{h})}\left[\ln p(\boldsymbol{h};\boldsymbol{\theta})\right] = -\frac{1}{2}\left(L_h\ln(2\pi) + \ln\det(\boldsymbol{\sigma}_h^2\boldsymbol{R}) + \boldsymbol{\mu}_h^\top\boldsymbol{R}\boldsymbol{\mu}_h + \mathrm{tr}\left(\boldsymbol{R}^{-1}\boldsymbol{\Sigma}_h\right)\right). \tag{B-13}$$

The third term is given by

$$\mathbb{E}_{q(\boldsymbol{r})q(\boldsymbol{z})}\left[\ln p(\boldsymbol{r} \mid \boldsymbol{z}, \boldsymbol{\theta})\right]$$

$$= -\sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{m=1}^{M}\frac{\phi_{jk}^{m}}{2}\left(L_r \ln(2\pi) + \ln(\det \boldsymbol{\Sigma}_{km}) + \boldsymbol{\mu}_{r_j}^{m\top}\boldsymbol{\Sigma}_{km}^{-1}\boldsymbol{\mu}_{r_j}^{m} + \mathrm{tr}(\boldsymbol{\Sigma}_{km}^{-1}\boldsymbol{\Sigma}_{r_j^m})\right) \tag{B-14}$$

The fourth term is the cross-entropy in the ELBO between posterior and conditional prior distributions of $\boldsymbol{z}$ (with pseudo-likelihood) and is evaluated by:

$$\mathbb{E}_{q(\boldsymbol{z})q(\boldsymbol{\beta})}\left[\ln p(\boldsymbol{z}|\boldsymbol{\beta})\right] \approx \sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{m=1}^{M}\sum_{i\in\delta_j}\phi_{jk}^{m}\phi_{ik}^{m} - \sum_{j=1}^{J}\sum_{m=1}^{M}\ln\left(\sum_{k=1}^{K}\exp\left(\boldsymbol{\mu}_{\beta_m}\sum_{i\in\delta_j}\phi_{ik}^{m}\right)\right) \tag{B-15}$$

The fifth term is the cross-entropy for β_m, which is simply the entropy of the uniform prior, given by:

$$\mathbb{E}_{q(\beta_m)}\left[\ln p(\beta_m : \boldsymbol{\theta})\right] = \mathbb{E}_{q(\boldsymbol{\beta}_m)}\left[\ln\frac{1}{u_2 - u_1}\right] = -\ln(u_2 - u_1) \tag{B-16}$$

Finally, by substituting all evaluated terms into Equation B-1, we can compute the ELBO at each iteration of the VEM algorithm.

Bibliography

[1] Mohanad Albughdadi. *Bayesian joint detection-estimation in functional MRI with automatic parcellation and functional constraints*. PhD thesis, Institut National Polytechnique de Toulouse-INPT, 2016.

[2] Ali-Kemal Aydin, William D Haselden, Yannick Goulam Houssen, Christophe Pouzat, Ravi L Rungta, Charlie Demené, Mickael Tanter, Patrick J Drew, Serge Charpak, and Davide Boido. Transfer functions linking neural calcium to single voxel functional ultrasound signal. *Nature communications*, 11(1):2954, 2020.

[3] Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*. Springer, 2006.

[4] David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017.

[5] Richard B Buxton and Lawrence R Frank. A model for the coupling between cerebral blood flow and oxygen metabolism during neural stimulation. *Journal of cerebral blood flow & metabolism*, 17(1):64–72, 1997.

[6] Vince D Calhoun and Tülay Adali. Unmixing fmri with independent component analysis. *IEEE Engineering in Medicine and Biology Magazine*, 25(2):79–90, 2006.

[7] Lotfi Chaari, Thomas Vincent, Florence Forbes, Michel Dojat, and Philippe Ciuciu. Fast joint detection-estimation of evoked brain activity in event-related fmri using a variational approach. *IEEE transactions on Medical Imaging*, 32(5):821–837, 2012.

[8] Emmanuelle Chaigneau, Martin Oheim, Etienne Audinat, and Serge Charpak. Two-photon imaging of capillary blood flow in olfactory bulb glomeruli. *Proceedings of the National Academy of Sciences*, 100(22):13081–13086, 2003.

[9] Christos Chatzichristos, Eleftherios Kofidis, Manuel Morante, and Sergios Theodoridis. Blind fmri source unmixing via higher-order tensor decompositions. *Journal of neuroscience methods*, 315:17–47, 2019.

[10] Hamza Cherkaoui, Thomas Moreau, Abderrahim Halimi, Claire Leroy, and Philippe Ciuciu. Multivariate semi-blind deconvolution of fmri time series. *NeuroImage*, 241:118418, 2021.

[11] Philippe Ciuciu, J-B Poline, Guillaume Marrelec, Jérôme Idier, Christophe Pallier, and Habib Benali. Unsupervised robust nonparametric estimation of the hemodynamic response function for any fmri experiment. *IEEE Transactions on medical imaging*, 22(10):1235–1251, 2003.

[12] Charlie Demené, Thomas Deffieux, Mathieu Pernot, Bruno-Félix Osmanski, Valérie Biran, Jean-Luc Gennisson, Lim-Anna Sieu, Antoine Bergel, Stephanie Franqui, Jean-Michel Correas, et al. Spatiotemporal clutter filtering of ultrafast ultrasound data highly increases doppler and fultrasound sensitivity. *IEEE transactions on medical imaging*, 34(11):2271–2285, 2015.

[13] Gerald A Dienel. Brain glucose metabolism: integration of energetics with function. *Physiological reviews*, 99(1):949–1045, 2019.

[14] Aybüke Erol, Chagajeg Soloukey, Bastian Generowicz, Nikki Van Dorp, Sebastiaan Koekkoek, Pieter Kruizinga, and Borbála Hunyadi. Deconvolution of the functional ultrasound response in the mouse visual pathway using block-term decomposition. *Neuroinformatics*, 21(2):247–265, 2023.

[15] Aybüke Erol, Simon Van Eyndhoven, Sebastiaan Koekkoek, Pieter Kruizinga, and Borbala Hunyadi. Joint estimation of hemodynamic response and stimulus function in functional ultrasound using convolutive mixtures. In *2020 54th Asilomar Conference on Signals, Systems, and Computers*, pages 246–250. IEEE, 2020.

[16] Cyril Goutte, Finn Aarup Nielsen, and KH Hansen. Modeling the hemodynamic response in fmri using smooth fir filters. *IEEE transactions on medical imaging*, 19(12):1188–1201, 2000.

[17] Iris IA Groen, Giovanni Piantoni, Stephanie Montenegro, Adeen Flinker, Sasha Devore, Orrin Devinsky, Werner Doyle, Patricia Dugan, Daniel Friedman, Nick F Ramsey, et al. Temporal dynamics of neural responses in human visual cortex. *Journal of Neuroscience*, 42(40):7562–7580, 2022.

[18] Rik Henson and Karl Friston. Convolution models for fmri. *Statistical parametric mapping: The analysis of functional brain images*, pages 178–192, 2007.

[19] James J Jun, Nicholas A Steinmetz, Joshua H Siegle, Daniel J Denman, Marius Bauza, Brian Barbarits, Albert K Lee, Costas A Anastassiou, Alexandru Andrei, Çağatay Aydın, et al. Fully integrated silicon probes for high-density recording of neural activity. *Nature*, 551(7679):232–236, 2017.

[20] Ming-Hung Kao, M'hamed Temkit, and Weng Kee Wong. Recent developments in optimal experimental designs for functional magnetic resonance imaging. *World journal of radiology*, 6(7):437, 2014.

[21] Hiroki Kitaura, Naonori Uozumi, Manavu Tohmi, Maya Yamazaki, Kenji Sakimura, Masaharu Kudoh, Takao Shimizu, and Katsuei Shibuki. Roles of nitric oxide as a vasodilator in neurovascular coupling of mouse somatosensory cortex. *Neuroscience research*, 59(2):160–171, 2007.

[22] Raymond C Koehler, Debebe Gebremedhin, and David R Harder. Role of astrocytes in cerebrovascular regulation. *Journal of applied physiology*, 100(1):307–317, 2006.

[23] Sofia-Eirini Kotti and Borbála Hunvadi. Extracting hemodynamic activity with low-rank spatial signatures in functional ultrasound using tensor decompositions. In *2024 32nd European Signal Processing Conference (EUSIPCO)*, pages 1347–1351. IEEE, 2024.

[24] Jean-Philippe Lachaux, Eugenio Rodriguez, Jacques Martinerie, and Francisco J Varela. Measuring phase synchrony in brain signals. *Human brain mapping*, 8(4):194–208, 1999.

[25] Harri Lappalainen and James W Miskin. Ensemble learning. In *Advances in Independent Component Analysis*, pages 75–92. Springer, 2000.

[26] Martin A Lindquist, Ji Meng Loh, Lauren Y Atlas, and Tor D Wager. Modeling the hemodynamic response function in fmri: efficiency, bias and mis-modeling. *Neuroimage*, 45(1):S187–S198, 2009.

[27] Yong Liu, Kun Wang, YU Chunshui, Yong He, Yuan Zhou, Meng Liang, Liang Wang, and Tianzi Jiang. Regional homogeneity, functional connectivity and imaging markers of alzheimer's disease: a review of resting-state fmri studies. *Neuropsychologia*, 46(6):1648–1656, 2008.

[28] Nikos K Logothetis and Brian A Wandell. Interpreting the bold signal. *Annu. Rev. Physiol.*, 66:735–769, 2004.

[29] Emilie Macé, Gabriel Montaldo, Ivan Cohen, Michel Baulac, Mathias Fink, and Mickael Tanter. Functional ultrasound imaging of the brain. *Nature methods*, 8(8):662–664, 2011.

[30] Emilie Mace, Gabriel Montaldo, Bruno-Felix Osmanski, Ivan Cohen, Mathias Fink, and Mickael Tanter. Functional ultrasound imaging of the brain: theory and basic principles. *IEEE transactions on ultrasonics, ferroelectrics, and frequency control*, 60(3):492–506, 2013.

[31] Salima Makni, Philippe Ciuciu, Jérôme Idier, and J-B Poline. Joint detection-estimation of brain activity in functional mri: a multichannel deconvolution solution. *IEEE Transactions on Signal Processing*, 53(9):3488–3502, 2005.

[32] Salima Makni, Jérôme Idier, Thomas Vincent, Bertrand Thirion, Ghislaine Dehaene-Lambertz, and Philippe Ciuciu. A fully bayesian approach to the parcel-based detection-estimation of brain activity in fmri. *Neuroimage*, 41(3):941–969, 2008.

[33] John HR Maunsell and Erik P Cook. The role of attention in visual processing. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 357(1424):1063–1072, 2002.

[34] Clare A McGrory, D Michael Titterington, Robert Reeves, and Anthony N Pettitt. Variational bayes for estimating the parameters of a hidden potts model. *Statistics and Computing*, 19:329–340, 2009.

[35] Gabriel Montaldo, Emilie Macé, Ivan Cohen, Jeremy Berckoff, Mickael Tanter, and Mathias Fink. Ultrafast compound doppler imaging: A new approach of doppler flow analysis. In *2010 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pages 324–327, 2010.

[36] Gabriel Montaldo, Mickaël Tanter, Jérémy Bercoff, Nicolas Benech, and Mathias Fink. Coherent plane-wave compounding for very high frame rate ultrasonography and transient elastography. *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, 56(3):489–506, 2009.

[37] Notger G Müller and Andreas Kleinschmidt. The attentional 'spotlight's' penumbra: center-surround modulation in striate cortex. *Neuroreport*, 15(6):977–980, 2004.

[38] Anwar O Nunez-Elizalde, Michael Krumin, Charu Bai Reddy, Gabriel Montaldo, Alan Urban, Kenneth D Harris, and Matteo Carandini. Neural correlates of blood flow measured by ultrasound. *Neuron*, 110(10):1631–1640, 2022.

[39] Gianluigi Pillonetto, Francesco Dinuzzo, Tianshi Chen, Giuseppe De Nicolao, and Lennart Ljung. Kernel methods in system identification, machine learning and function estimation: A survey. *Automatica*, 50(3):657–682, 2014.

[40] M Andrea Pisauro, Neel T Dhruv, Matteo Carandini, and Andrea Benucci. Fast hemodynamic responses in the visual cortex of the awake mouse. *Journal of Neuroscience*, 33(46):18343–18351, 2013.

[41] Marcus E. Raichle, Ann Mary MacLeod, Abraham Z. Snyder, William J. Powers, Debra A. Gusnard, and Gordon L. Shulman. A default mode of brain function. *Proceedings of the National Academy of Sciences*, 98(2):676–682, 2001.

[42] Arien J Smith, Hal Blumenfeld, Kevin L Behar, Douglas L Rothman, Robert G Shulman, and Fahmeed Hyder. Cerebral energetics and spiking frequency: the neurophysiological basis of fmri. *Proceedings of the National Academy of Sciences*, 99(16):10765–10770, 2002.

[43] Karthik Ramakrishnan Sreenivasan, Martin Havlicek, and Gopikrishna Deshpande. Nonparametric hemodynamic deconvolution of fmri using homomorphic filtering. *IEEE transactions on medical imaging*, 34(5):1155–1163, 2014.

[44] Anil Man Tuladhar, Liselore Snaphaan, Elena Shumskaya, Mark Rijpkema, Guillen Fernandez, David G Norris, and Frank-Erik de Leeuw. Default mode network connectivity in stroke patients. *PloS one*, 8(6):e66556, 2013.

[45] Richard Eric Turner and Maneesh Sahani. *Two problems with variational expectation maximisation for time series models*, page 104–124. Cambridge University Press, 2011.

[46] Christopher W Tyler and Lora T Likova. The neurometabolic underpinnings of fmri bold dynamics. *Advanced Brain Neuroimaging Topics in Health and Disease—Methods and Applications*, pages 21–47, 2014.

[47] Peter Uhlhaas, Gordon Pipa, Bruss Lima, Lucia Melloni, Sergio Neuenschwander, Danko Nikolić, and Wolf Singer. Neural synchrony in cortical networks: history, concept and current status. *Frontiers in integrative neuroscience*, 3:543, 2009.

[48] Benjamin Vidal, Marine Droguerre, Ludovic Venet, Luc Zimmer, Marco Valdebenito, Franck Mouthon, and Mathieu Charvériat. Functional ultrasound imaging to study brain dynamics: application of pharmaco-fus to atomoxetine. *Neuropharmacology*, 179:108273, 2020.

[49] Thomas Vincent, Philippe Ciuciu, and Jerome Idier. Application and validation of spatial mixture modelling for the joint detection-estimation of brain activity in fmri. In *2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 5218–5222, 2007.

[50] Roberto Viviani, Georg Grön, and Manfred Spitzer. Functional principal component analysis of fmri data. *Human brain mapping*, 24(2):109–129, 2005.

[51] Lupeng Wang and Richard J Krauzlis. Involvement of striatal direct pathway in visual spatial attention in mice. *Current Biology*, 30(23):4739–4744, 2020.

[52] Guo-Rong Wu, Wei Liao, Sebastiano Stramaglia, Ju-Rong Ding, Huafu Chen, and Daniele Marinazzo. A blind deconvolution approach to recover effective connectivity brain networks from resting state fmri data. *Medical image analysis*, 17(3):365–374, 2013.

[53] Dagmar Zeithamova, W Todd Maddox, and David M Schnyer. Dissociable prototype learning systems: evidence from brain imaging and behavior. *Journal of Neuroscience*, 28(49):13194–13201, 2008.

Glossary

List of Acronyms

fUS	functional Ultrasound
PDI	Power Doppler Image
RBCs	red blood cells
HRF	hemodynamic response function
LTI	linear time-invariant
fMRI	functional magnetic resonance imaging
SVD	singular value decomposition
PDI	power Doppler image
JDE	joint detection-estimation
KL	Kullback-Leiber
ELBO	evidence lower bound
CAVI	coordinate ascent variational inference
VEM	Variational expectation-maximization
NRF	neural response function
FIR	finite impulse response
SNR	signal-to-noise ratio
ICA	independent component analysis
PCA	principle component analysis
MCMC	Markov Chain Monte Carlo
JPDE	joint parcellation-detection-estimation
ROI	region of interest
BOLD	blood oxygenation level dependent
DCT	discrete cosine transform
NRL	neural response level

TTP	time-to-peak
FWHM	full width at half maximum
PPM	posterior probability map
iid	independently and identically distributed
PCC	Pearson correlation coefficient
LFP	low field potential
MAP	maximum a posteriori