UNIVERSITÀ DEGLI STUDI DI PADOVA

Dipartimento di Scienze Statistiche

Corso di Dottorato di Ricerca in Scienze Statistiche

Ciclo XXXIV

BAYESIAN MODELING OF CALCIUM IMAGING DATA

Coordinatore del Corso: Prof. Nicola Sartori

Supervisore: Prof. Antonio Canale

Co-supervisore: Prof. Michele Guindani

Dottoranda: Laura D'Angelo

12 Gennaio 2022

Supervisor: Prof. Antonio Canale
Co-supervisor: Prof. Michele Guindani

ABSTRACT

Recent advancements in miniaturized fluorescence microscopy have made it possible to investigate neuronal responses to external stimuli in awake behaving animals through the analysis of intra-cellular calcium signals. An ongoing challenge is deconvolving the noisy calcium signals to extract the spike trains, and understanding how this activity is affected by external stimuli and conditions. In this thesis, we aim to provide novel approaches to tackle various aspects of the analysis of calcium imaging data within a Bayesian framework.

Following the standard methodology to the analysis of calcium imaging data based on a two-stage approach, we investigate efficient computational methods to link the output of the deconvolved fluorescence traces with the experimental conditions. In particular, we focus on the use of Poisson regression models to relate the number of detected spikes with several covariates. Motivated by this framework, but with a general impact in terms of application to other fields, we develop an efficient Metropolis-Hastings and importance sampling algorithm to simulate from the posterior distribution of the parameters of Poisson log-linear models under conditional Gaussian priors, with superior performance with respect to the state-of-the-art alternatives.

Motivated by the lack of clear uncertainty quantification resulting from the use of a two-stage approach, and the impossibility to borrow information between the two stages, we focus on the analysis of individual neurons, and develop a coherent mixture model that allows for estimation of spiking activity and, simultaneously, reconstructing the distributions of the calcium transient spikes' amplitudes under different experimental conditions. More specifically, our modeling framework leverages two nested layers of random discrete mixture priors to borrow information between experiments and discover similarities in the distributional patterns of the neuronal response to different stimuli.

Finally, we move to the multivariate analysis of populations of neurons. Here the interest is not only to detect and analyze the spiking activity but also to investigate the existence of groups of co-activating neurons. Estimation of such groups is a challenging problem due to the need to deconvolve the calcium traces and then cluster the resulting latent binary time series of activity. We describe a nonparametric mixture model that allows for simultaneous deconvolution and clustering of time series based on common patterns of activity. The model makes use of a latent continuous process for the spike probabilities to identify groups of co-activating cells. Neurons' dependence is taken into account by informing the mixture weights with their spatial location, following the common neuroscience assumption that neighboring neurons often activate together.

SOMMARIO

Grazie alle recenti innovazioni tecnologiche nel campo della microscopia miniaturizzata e, in particolare, allo sviluppo di una speciale tecnica che permette di misurare otticamente il livello intra-cellulare di ioni di calcio, si è resa possibile l'analisi dell'attività neuronale in risposta alla stimolazione esterna in animali svegli e liberi di muoversi. Tuttavia, l'analisi del livello di fluorescenza osservato presenta diverse complessità. Una prima difficoltà deriva dalla necessità di estrarre le serie del segnale (i cosiddetti *spike train*), ovvero le serie di attività neuronale. Dopodiché, il segnale estratto deve essere messo in relazione con con le condizioni sperimentali che l'hanno generato. Con questa tesi si vogliono introdurre degli approcci innovativi per l'analisi di dati di *imaging* del calcio, nell'ambito di un'analisi statistica bayesiana.

L'approccio classico all'analisi di dati di *imaging* del calcio si basa su una procedura in due passi: in una prima fase vengono estratti gli *spike train*; successivamente, queste serie vengono messe in relazione alle condizioni esterne. Muovendoci all'interno di questo contesto, ma con validità più generale in termini di ambiti di applicazione, si introducono dei nuovi metodi computazionali per stimare in modo efficiente la relazione tra il segnale osservato e le condizioni sperimentali. In particolare, si pone l'interesse su modelli di regressione di Poisson, comunemente usati per studiare la dipendenza del numero di attivazioni da un insieme di fattori esterni. Si sviluppano un algoritmo Metropolis-Hastings e un *importance sampler* per simulare dalla distribuzione a posteriori dei coefficienti di tali modelli, sotto l'assunzione di distribuzioni a priori Gaussiane (o condizionatamente Gaussiane) sui parametri di regressione.

Un'analisi in due passi comporta alcuni svantaggi: per esempio, l'impossibilità di ottenere una chiara valutazione dell'incertezza complessiva, oltre all'impossibilità di condividere informazione tra le due fasi. Per questo motivo, ci focalizziamo sull'analisi di singoli neuroni e introduciamo un modello mistura che permette di stimare l'attività neuronale e, allo stesso tempo, di analizzare la distribuzione delle attivazioni in risposta a diverse condizioni sperimentali. In particolare, il modello proposto sfrutta una distribuzione a priori basata su due livelli annidati di misture finite, che permette di condividere l'informazione tra condizioni sperimentali, e indagare similitudini e differenze nella risposta ai diversi stimoli.

Infine, si introduce un'analisi multivariata di popolazioni di neuroni. In questo contesto l'interesse non è volto solo ad analizzare le singole serie di attività, ma anche ad indagare l'esistenza di gruppi di neuroni con modelli di attivazioni simili. In questa tesi si introduce un modello mistura nonparametrico che permette di estrarre le serie delle attivazioni e, allo stesso tempo, di raggruppare i neuroni con un modello di attivazioni simile. Il modello si basa sull'introduzione di un processo latente continuo che descrive, per ogni istante temporale, la probabilità di osservare un'attivazione.

Inoltre, per includere la dipendenza spaziale tra neuroni, i pesi della mistura sono funzione della distanza tra le cellule, come suggerito da diversi studi di neuroscienze.

CONTENTS

INTRODUCTION **1**
Overview . . . 1
Main contributions of the thesis . . . 2

1 BACKGROUND: STATISTICAL MODELING OF CALCIUM IMAGING DATA **5**
1.1 Overview of calcium imaging data . . . 5
1.1.1 Deconvolution methods . . . 5
1.1.2 Spike train data analysis . . . 7
1.2 Data sets . . . 7
1.2.1 Allen Brain Observatory data . . . 7
1.2.2 Hippocampal neurons data . . . 9
1.3 A brief review of some Bayesian nonparametric models . . . 10
1.3.1 Finite mixture models . . . 10
1.3.2 Dirichlet process mixture models . . . 12
1.3.3 Finite mixtures with an unknown number of components . . . 16
1.3.4 Bayesian nonparametric models for nested data . . . 18

2 NOVEL POSTERIOR SAMPLING SCHEMES FOR POISSON ENCODING MODELS **23**
2.1 Efficient posterior sampling strategies . . . 25
2.1.1 Approximate posterior distribution . . . 25
2.1.2 Metropolis-Hastings sampler . . . 26
2.1.3 Adaptive importance sampler . . . 27
2.1.4 Tuning parameters r_i . . . 27
2.2 Numerical illustrations . . . 28
2.2.1 Synthetic data . . . 28
2.2.2 Spike train data . . . 29

3 MODELING SINGLE-NEURON ACTIVATIONS VIA NONPARAMETRIC MIXTURES **33**
3.1 Model and prior specification . . . 34
3.2 Posterior inference . . . 36
3.3 Simulation study . . . 37
3.4 Allen Brain Observatory data analysis . . . 42
3.4.1 Analysis of additional neurons . . . 46

4 CLUSTERING ACTIVATION PATTERNS OF SPATIALLY-REFERENCED NEURONS **51**
4.1 Model and prior specification . . . 52
4.2 Posterior inference . . . 53

4.3 Simulation study 54
4.4 Analysis of hippocampal neurons 56

CONCLUSIONS **65**

BIBLIOGRAPHY **67**

LIST OF FIGURES

1.1 Calcium traces of two neurons from the Allen Brain Observatory data. 8
1.2 Calcium traces of 88 neurons from the hippocampal neurons data. . . . 9

2.1 Comparison between the time per independent sample of the proposed algorithms and of the HMC algorithm. 30
2.2 Comparison between the time per independent sample of the proposed algorithms and of the HMC algorithm, using a horseshoe prior. 31
2.3 Estimated coefficients of the regression on the calcium imaging data set. 32

3.1 Comparison between the misclassification error rate in the simulation study for the four considered methods. 40
3.2 Comparison between the adjusted Rand index of the fCAM and CAM. 41
3.3 Observed fluorescence trace of a neuron from the Allen Brain Observatory data and estimated neuronal activity. 42
3.4 Distribution of the observational cluster parameters for the three experimental conditions of the Allen Brain Observatory data. 44
3.5 Visual representation of the estimated spikes and their amplitudes in the calcium trace from the Allen Brain Observatory data. 45
3.6 Observed fluorescence trace of neuron $\#517398389$ from the Allen Brain Observatory data and estimated neuronal activity. 48
3.7 Distribution of the observational cluster parameters for the three experimental conditions of neuron $\#517398389$. 48
3.8 Visual representation of the estimated spikes and their amplitudes in the calcium trace of neuron $\#517398389$. 48
3.9 Observed fluorescence trace of neuron $\#587435328$ from the Allen Brain Observatory data and estimated neuronal activity. 49
3.10 Distribution of the observational cluster parameters for the three experimental conditions of neuron $\#587435328$. 49
3.11 Visual representation of the estimated spikes and their amplitudes in the calcium trace of neuron $\#587435328$. 49

4.1 Estimated clustering on the simulated data with sample size $n = 20$. . 58
4.2 Estimated clustering on the simulated data with sample size $n = 30$. . 59
4.3 Estimated clustering on the simulated data with sample size $n = 40$. . 60
4.4 Estimated clustering on the hippocampal neurons data using 4 time windows of length 50. 61

4.5 Estimated clustering on the hippocampal neurons data using 2 time windows of length 100. 62
4.6 Distribution of the estimated spike amplitudes on the hippocampal neurons data. 63

LIST OF TABLES

4.1 Misclassification error rate on the simulated data. 55
4.2 Adjusted Rand index of the estimated clusters of activity on the simulated data. 55

INTRODUCTION

Overview

A fundamental but unsolved problem in neuroscience is understanding the functioning of neurons and neuronal networks in processing sensory information, generating locomotion, and mediating learning and memory. The investigation of the structure and function of the nervous system can be dated back to the nineteenth century with the invention of the technique of silver impregnation by Camillo Golgi in 1873, which allowed the visualization of individual neurons (Drouin et al., 2015). The technique initiated the study of the microscopic anatomy of the nervous system, and the investigation of how neurons organize to form the brain. Ever since there has been a significant research effort both to discover the cellular properties of the nervous system, and to characterize behaviors and correlate them with activity imaged in different regions of the brain. However, many scientists recognize that despite the innovative techniques developed to observe and analyze neurons, we are still facing an "explanatory gap" between the understanding of elemental components and the outputs that they produce (Parker, 2006; Parker, 2010; Dudai, 2004). That is, we know a lot about the components of the nervous system, but still we have little insight into how these components work together to enable us to think, remember, or behave. One of the reasons of this gap is the availability of a huge quantity of data, but a lack of tools to integrate these data in order to obtain a coherent picture of the brain functioning (Parker, 2010).

The technological developments of the last few decades have opened fundamentally new opportunities to investigate the nervous system. Large neuronal networks can now be visualized using *in vivo* high-resolution imaging techniques, which permit to record the neuronal activity in freely moving animals over long periods of time. In this thesis, we focus on data resulting from the application of the two-photon calcium imaging technique. Calcium ions generate intracellular signals that determine a large variety of functions in all neurons: when a neuron fires, calcium floods the cell and produces a transient spike in its concentration (Grienberger and Konnerth, 2012). By using genetically encoded calcium indicators, which are fluorescent molecules that react when binding to the calcium ions, it is possible to optically measure the level of calcium by analyzing the observed fluorescence trace. However, extracting these fluorescent calcium traces is just the first step towards the understanding of brain circuits: how to relate the observed pattern of neuronal activity to the external stimuli that triggered it remains an open problem of research.

The first step for analyzing fluorescent calcium traces is to deconvolve them to extract the spike trains, which are the series of recorded firing times and spikes' amplitudes.

From these series it is possible to derive many useful quantities which are commonly used to interpret the neuron's activity: for example, one can compute the number of recorded firing events, neuron-specific or stimulus-specific distributions of the spikes' amplitudes, and other proxies of the intensity of the neuronal response. The set of tools that explicitly try to relate external stimuli with some summary of activity are usually referred to as "encoding models". In this context the stimuli are considered as features, and they are used to predict patterns of neuronal activity. These models allow to investigate how experimental conditions and specific stimulation trigger the neurons' activity, and hence how external information is encoded by individual neurons and neuronal networks.

Main contributions of the thesis

The availability of large quantities of data from calcium imaging studies, and the relative scarcity of tools to analyze them motivated the investigation of new methodologies. In this thesis, we aim to contribute to the development of novel statistical tools to gain new insights into the analysis of calcium imaging data. Herein, we adopted a Bayesian approach: there are several reasons that led to this choice. As it will become clearer from the details of the specific applications, a Bayesian approach had proved necessary in order to deal with models that comprised complex dependence structures, heterogeneous data, and, possibly, the availability of past information from previous studies. Moreover, using a Bayesian approach, it is straightforward to induce some regularization on the model parameters, a feature that is often fundamental in multivariate studies.

Novel posterior sampling scheme for Poisson encoding models

Linear models and generalized linear models are among the most natural classes of encoding models (Paninski et al., 2007). They allow to link the observed output of an experiment with a number of features and experimental conditions in a flexible and interpretable way. In particular, if the variable of interest is the number of spikes, which is a proxy of the intensity of the neuronal response, Poisson regression represents a straightforward choice. However, the dimensionality of the considered data poses a computational challenge and leads to the need for efficient algorithms to obtain a sample from the posterior distribution of parameters. Motivated by the lack of specific and efficient algorithms to sample from the posterior distribution of the parameters of Bayesian log-linear models, in Chapter 2 we develop a novel sampling strategy which exhibits superior performance with respect to the state-of-the-art alternatives.

In particular, we develop an efficient Metropolis-Hastings algorithm and importance sampler to simulate from the posterior distribution of the regression parameters. The key for both algorithms is the introduction of a proposal density based on an approximation of the posterior distribution of parameters under conditional Gaussian priors. With conditional Gaussian prior, we refer to a possibly hierarchical prior with conditional distribution $\beta \sim \mathrm{N}(b, B)$, with b and/or B random. Examples include straightforward Gaussian prior distributions with informative (b, B) fixed using prior

information, and scale mixtures of Gaussian where b is set to zero and the variance has a suitable hierarchical representation, such as the Bayesian lasso prior or the horseshoe prior, among others. Our result leverages the negative binomial convergence to the Poisson likelihood (Casella and Berger, 2002): thanks to this result, we are able to exploit the Pólya-gamma data augmentation of Polson et al. (2013) to derive an efficient sampling scheme.

The performances of the proposed solutions, in terms of mixing and computation time, are comparable or superior to those of the efficient Stan implementation of the Hamiltonian Monte Carlo algorithm in all scenarios considered in an extensive simulation study, and particularly when a hierarchical prior is assumed. The ease of application of the proposed algorithms is further enhanced by their availability via the R package `bpr` (D'Angelo, 2021). Clearly, the impact is broader than the motivating application to calcium imaging data, as Poisson regression is commonly used in several other fields.

Modeling single-neuron activations via nonparametric mixtures

Routine methods to analyze calcium imaging data are based on a two-stage approach: in a first phase, the raw fluorescent trace is deconvolved to extract the spike train, then, in a second phase, some summary statistics is derived and linked to the experimental conditions that generated it. This approach, adopted in the previous contribution, is simple to implement and can be applied in a broad range of applications, however, it has some drawbacks: for example, the impossibility to borrow information between the two stages, and an unclear quantification of the uncertainty propagating from one stage to the next. Only performing the two tasks simultaneously allows to coherently quantify the uncertainty of the results, whereas, in a two-step approach, it is not straightforward to evaluate the overall uncertainty, as it is the result of the contribution of each step.

In Chapter 3, we introduce a nested Bayesian finite mixture model that allows for estimating the spiking activity and, simultaneously, reconstructing the distributions of the calcium transient spikes' amplitudes under different experimental conditions, for example, in response to different types of stimuli. More specifically, our modeling framework estimates and clusters the distributions of the calcium transient spikes' amplitudes via a nested formulation of the generalized mixtures of finite mixtures prior recently proposed by Frühwirth-Schnatter et al. (2021). The proposed model further adopts the use of a common atom specification as in Denti et al. (2021) for estimating the distribution of the spikes' amplitudes under each experimental condition. These two nested layers of random discrete mixture priors allow to borrow information between experiments and discover similarities in the distributional patterns of neuronal responses to different stimuli. Furthermore, the spikes' intensity values are also clustered within and between experimental conditions to determine the existence of common (recurring) response amplitudes.

Clustering activation patterns of spatially-referenced neurons

When analyzing populations of neurons, often the interest is in identifying groups of neurons with a highly correlated pattern of activity. For many areas of the brain, there is a general agreement on the organization of these networks of neurons and their behavior. However, the functional and anatomical organization of hippocampal neurons is still an open research problem.

In Chapter 4, we formulate a nonparametric mixture model that allows for deconvolution of several calcium traces and, simultaneously, detection of groups of co-activating neurons. Specifically, our model clusters the latent binary time series of activity based on similarities of the spiking activity over seconds-long periods of time. The model makes use of a latent continuous process for the spike probabilities to identify groups of co-activating cells. Neurons' dependence is taken into account by informing the mixture weights with their spatial location through the use of a probit stick breaking process (Rodríguez and Dunson, 2011), following the common neuroscience assumption that neighboring neurons often activate together.

1 BACKGROUND: STATISTICAL MODELING OF CALCIUM IMAGING DATA

1.1. Overview of calcium imaging data

Calcium ions generate intracellular signals that control key functions in all types of neurons. At rest, most neurons have an intracellular calcium concentration of about 100 nm; however, during electrical activity, the concentration can rise transiently up to levels around 1000 nm (Berridge et al., 2000). The development of techniques that enable the visualization and quantitative estimation of the intracellular calcium signals have thus greatly enhanced the investigation of neuronal functioning. The development of calcium imaging techniques involved two parallel processes: the development of calcium indicators, which are fluorescent molecules that react when binding to the calcium ions, and the implementation of the appropriate imaging instrumentation, in particular, the introduction of two-photon microscopy (Denk et al., 1990). In recent years, the innovation achieved in these two fields has allowed for real-time observation of biological processes at the single-cell level simultaneously for large groups of neurons (Grienberger and Konnerth, 2012).

The output of two-photon calcium imaging is a movie of time-varying fluorescence intensities, and a first complex pre-processing phase deals with the identification of the spatial location of each neuron in the optical field and source extraction (Mukamel et al., 2009; Dombeck et al., 2010). The resulting processed data consist of a fluorescent calcium trace for each observable neuron in the targeted area which, however, is only a proxy of the underlying neuronal activity. Hence further analyses are needed to deconvolve the fluorescence trace to extract the spike train (i.e. the series of recorded firing times), and to try to explain how these firing events are linked with the experiment that generated that particular pattern of activity.

1.1.1. **Deconvolution methods**

There is currently a rich literature of methods addressing the issue of deconvolving the raw fluorescent trace to extract the spike train. A successful approach is to assume a biophysical model to relate the spiking activity to the calcium dynamics, and to the observed fluorescence. Vogelstein et al. (2010) proposed a simple but effective model that has later been adopted by several authors (Pnevmatikakis et al., 2016; Friedrich and Paninski, 2016; Friedrich et al., 2017; Jewell and Witten, 2018; Jewell et al., 2019). The model considers the observed fluorescence as a linear (and noisy) function of the intracellular calcium concentration; the calcium dynamics is then modeled using an autoregressive process with jumps in correspondence of the neuron's firing events. Denoting with y_t the observed fluorescence trace of a neuron and with Ca_t the

underlying calcium concentration, for time $t = 1, \dots, T$, the model can be written as

$$\begin{aligned} y_t &= b + \mathrm{Ca}_t + \epsilon_t, \qquad \epsilon_t \sim \mathrm{N}(0, \sigma^2), \\ \mathrm{Ca}_t &= \gamma\, \mathrm{Ca}_{t-1} + A_t + w_t, \qquad w_t \sim \mathrm{N}(0, \tau^2), \end{aligned} \tag{1.1}$$

where b models the baseline level of the observed trace and ϵ_t is a Gaussian measurement error. In the absence of neuronal activity, the true calcium concentration Ca_t is considered to be centered around zero. The parameter A_t captures the neuronal activity: in the absence of a spike ($A_t = 0$), the calcium level follows a AR(1) process controlled by the parameter γ; when a spike occurs, the concentration increases instantaneously of a value $A_t > 0$. A challenge remains estimating the neuronal activity A_t in a precise and computationally efficient way.

Vogelstein et al. (2010) assumed that all spikes have a fixed amplitude, and interpreted the parameter A_t as the *number* of spikes at time t. Following this definition, they placed a Poisson prior distribution on A_t; however, the maximum a posteriori estimation of the spike train using a Poisson distribution is computationally intractable. Hence they searched an approximate solution by replacing the Poisson distribution with an exponential distribution of the same mean. This leads to some loss of interpretation of the parameters A_t, as now they are no longer integer values but rather non-negative real numbers, but turns the problem into a convex optimization, which can be solved efficiently. Adopting this approach leads to solving a non-negative lasso problem for estimating the calcium concentration, where the L_1 penalty enforces sparsity of the neural activity. Efficient algorithms to obtain a solution of this problem have also been proposed by Pnevmatikakis et al. (2016), Friedrich and Paninski (2016), and Friedrich et al. (2017).

A different perspective is instead proposed by Jewell and Witten (2018) and Jewell et al. (2019): rather than interpreting A_t in model (1.1) as the number of spikes at the t-th timestep, they interpreted its sign as an indicator for whether or not *at least one* spike occurred, that is, $A_t = 0$ indicates no spikes at time t, and $A_t > 0$ indicates the occurrence of at least one spike. The model so formulated includes an indicator variable, which corresponds to using an L_0 penalization and which makes the optimization problem highly non-convex. In their work, Jewell and Witten (2018) and Jewell et al. (2019) developed fast algorithms to compute the spike trains under these assumptions. Jewell and Witten (2018) asserted that the solutions discussed by Vogelstein et al. (2010), Friedrich and Paninski (2016), and Friedrich et al. (2017) can actually be seen as convex relaxations of this optimization problem, to overcome the computational intractability of the L_0 penalization.

Finally, Pnevmatikakis et al. (2013) proposed a fully Bayesian approach. Although less computationally efficient than optimization methods, it allows to obtain a posterior distribution of all model parameters instead of just a point estimate, hence improving uncertainty quantification. Differently from previous models, they defined the parameter A_t as the *amplitude* of a spike at time t, taking values in the non-negative real numbers. They formulated the presence/absence of a spike and its amplitude by using the product of a Bernoulli random variable (taking value 0 if there is no spike at time t, and 1 otherwise) with a half-Gaussian random variable (modeling the positive amplitudes). However, they did not explicitly assume sparsity of the spikes.

1.1.2. **Spike train data analysis**

Standard methods to analyze calcium imaging data rely on a two-step approach: in a first phase, some deconvolution method, such as those described in the previous section, is applied to identify the spikes; then, a different method is used on the deconvolved output to analyze it and, possibly, to relate it with some covariates. However, while there is a rich literature on deconvolution methods, there is still little research on methods that try to derive inferential results from their output.

The set of models that explicitly try to understand how the brain encodes external variables and stimuli into spike train is usually referred to as "encoding models". Among them, Paninski et al. (2007) focused on the use of generalized linear models (GLMs). GLMs allow spike trains to be regressed against a potentially large number of covariates such as behavioral parameters, experimental conditions, and other relevant factors. Moreover, regression models provide simple and interpretable results of the effect of each covariate on the neuronal response. In particular, Paninski et al. (2007) highlighted the importance of regularization methods and inclusion of prior knowledge to improve estimation of the model parameters, as in many cases the number of covariates is potentially very large, leading to noisy results and a loss in interpretability.

A different approach has been proposed by Wei et al. (2019): instead of focusing the relationship between the experimental conditions and some summary statistic of the resulting spike train, they studied the distribution of the deconvolved output. In particular, this allows to analyze quantities such as the spike probability and the spikes' amplitudes. They proposed a mixture model, with a Dirac mass at zero, representing the absence of neuronal response, and a translated Gamma distribution to model the positive amplitudes.

1.2. Data sets

1.2.1. **Allen Brain Observatory data**

The Allen Brain Observatory (Allen Institute MindScope Program, 2016; de Vries et al., 2020) is a large public data repository for investigating how sensory stimulation is represented by neural activity in the mouse visual cortex in both single cells and populations. The project aims to provide a standardized and systematic survey to measure and analyze visual responses from neurons across cortical areas, using genetically encoded fluorescent calcium indicators, measured by *in vivo* two-photon calcium imaging.

The study is an extended survey of physiological activity in the mouse visual cortex in response to a range of visual stimuli (Allen Brain Observatory, 2017). Each mouse was placed in front of a screen where different types of visual stimuli were shown, while the mouse's neuronal activity was recorded. The stimuli vary from simple synthetic images such as locally sparse noise or static gratings, to complex natural scenes and movies. The goal of the study is to investigate how neurons at different depths and in different areas of the visual cortex respond to stimuli of varying complexity, to understand their functional properties. An important finding from mammalian is

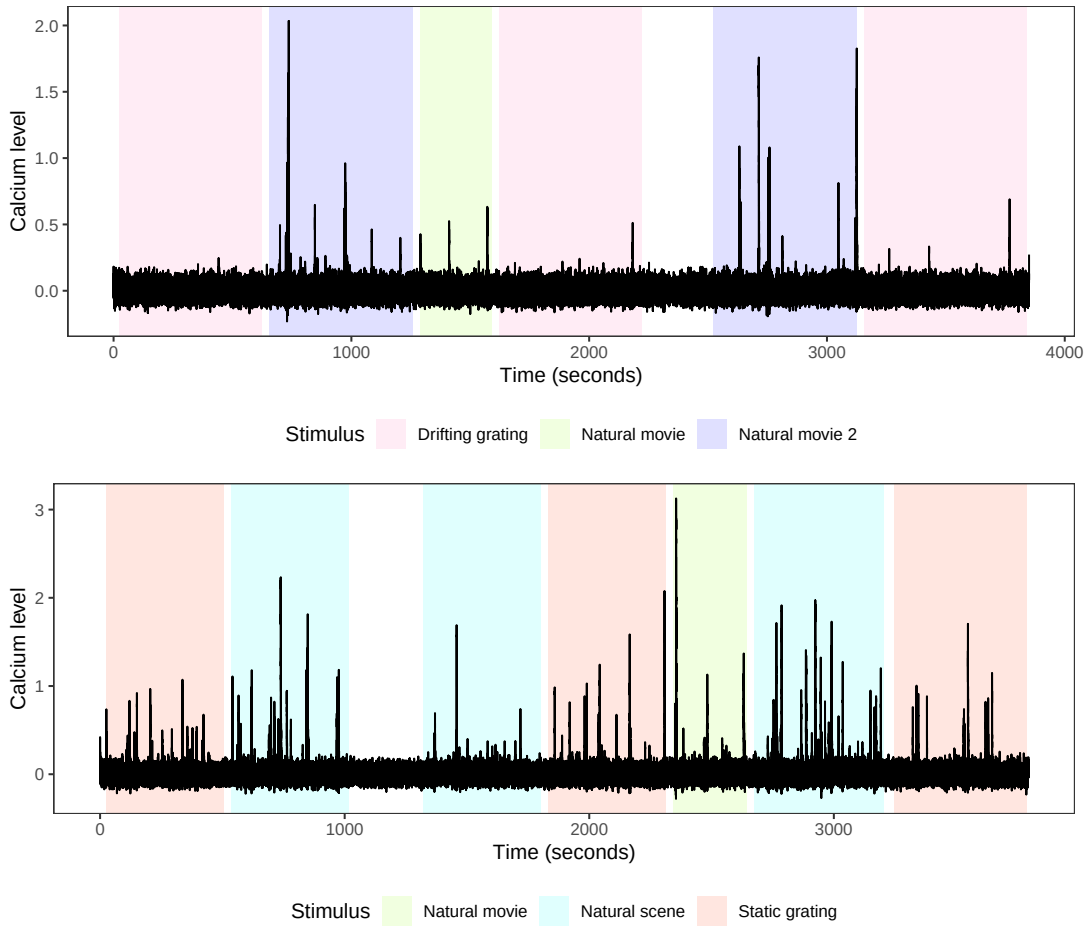

FIGURE 1.1: Allen Brain Observatory data: calcium traces (black line) of two neurons located in the primary visual area during Session A (upper plot) and Session B (bottom plot) of the experiment. The background colors denote the visual stimulus displayed at each time.

that higher visual areas tend to respond to more complex stimuli relative to lower areas. These differences indicate that the different neurons and visual areas have distinct functional properties. Hence, it is of critical interest to devise methods that allow inferring how the neuronal response varies under the different types of visual stimuli. As an example, Figure 1.1 shows the calcium traces of two neurons recorded during two different experiment sessions from the Allen Brain Observatory study. Each experiment comprises three types of visual stimuli, and has a duration of approximately one hour. These plots highlight that the neuronal response is highly variable, both across experimental conditions and between neurons.

The Allen Brain Observatory study comprises records of neuronal activity from over 60000 cells from six visual areas (VISp, VISl, VISal, VISrl, VISam, and VISpm) and different imaging depths (ranging from 175 to 625 microns). The data were collected analyzing the brain activity of several genetically engineered mice, using different transgenic Cre lines. The neuronal response of each mouse was recorded during three experimental sessions: specifically, each session was made up of different types of

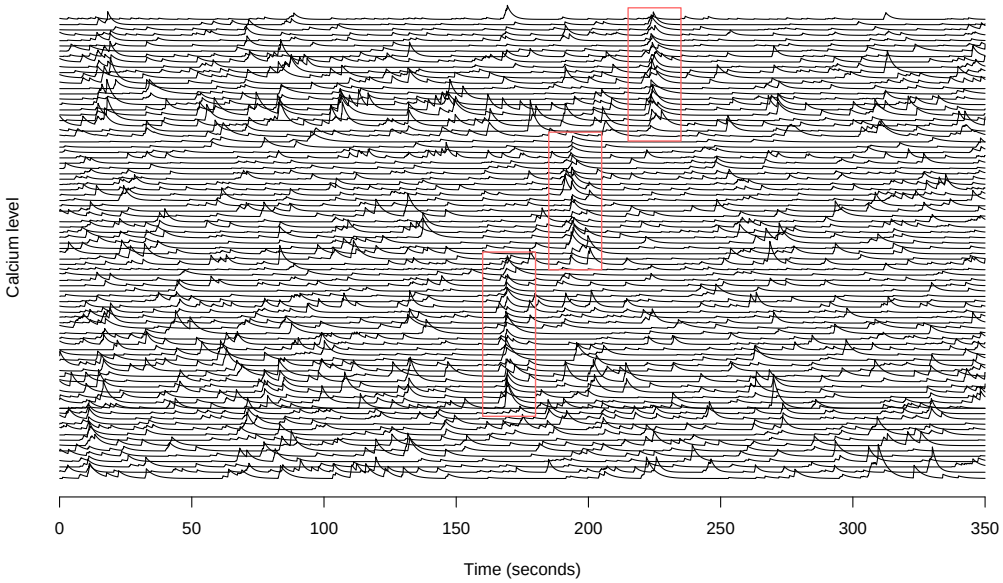

FIGURE 1.2: Hippocampal neurons data: denoised calcium traces (black lines) recorded in the first 350 seconds of the experiment for 86 neurons. The colored rectangles highlight the presence of co-activating neurons.

visual stimuli displayed sequentially. Session "A" comprises two natural movies and a drifting gratings stimulus; session "B" comprises both natural movies and natural scenes, and a static gratings stimulus; finally, session "C" again includes two natural movies, and a locally sparse noise. Moreover, in all sessions a period of absence of stimuli was used to evaluate the baseline response during spontaneous activity. A detailed description of the visual stimuli can be found in a technical report (Allen Brain Observatory, 2017).

1.2.2. **Hippocampal neurons data**

Neurons located in the hippocampus are the focus of many studies investigating the neuronal circuits that underlie cognition, learning and memory. Specifically, experiments involving freely moving mice within particular environments allow to study the location-specific firing of hippocampal neurons during navigation. These particular neurons, that activate when the animal enters a specific place in the environment (the *place field*), are often called *place cells*. Place cells are thought, collectively, to act as cognitive representations of a specific location in space (O'Keefe and Nadel, 1978). However, the anatomical organization of these subpopulations of co-activating cells is still under investigation: it is unclear whether hippocampal neurons with similar place fields are spatially organized within the hippocampus (Dombeck et al., 2010).

The dataset we consider collects the activity of 325 neurons located in the hippocampus of a mouse during navigation in a closed environment for about 12 minutes. The calcium level is recorded with a frequency of 15 frames per second, resulting in 10870 calcium measurements over time for each series. Moreover, for each neuron, it is also

available its location in the region of interest of the hippocampus.

1.3. A brief review of some Bayesian nonparametric models

In this section we review some statistical tools that will be employed in this thesis in the analysis of calcium imaging data. The purpose of this section is not to provide a comprehensive review, but rather to outline the theoretical framework we adopted and fix some notation. The core topic will be the Bayesian methodology, with a focus on Bayesian nonparametric models. Concise and exhaustive reviews of Bayesian nonparametrics can be found in Canale et al. (2016), and in the books by Hjort et al. (2010) and Ghosal and van der Vaart (2017).

1.3.1. **Finite mixture models**

We start our discussion by reviewing finite mixtures. Although they are not part of the Bayesian nonparametric methodology, they provide the starting point for many models that we will review in the following. The content of this brief overview on finite mixtures is largely based on the dedicated chapter in Gelman et al. (2013).

Definition and hierarchical representations

Mixtures are a popular tool to model heterogeneous data, characterized by the presence of subpopulations within the overall population. In many practical problems the data are collected under different conditions – unfortunately, it is not always possible to have information on the subpopulation to which each individual observation belongs. Mixture models can be used in problems of this type, where the population consists of a number of latent subpopulations, and where each subpopulation can be described using a relatively simple model.

Denote the observed data as a vector of n units $\boldsymbol{y} = (y_1, \dots, y_n)$, where each $y_i \in \mathcal{Y}$ is a scalar observation and $\mathcal{Y}$ is a measurable space (extension to the multivariate case is straightforward). Also, assume that the n observations are exchangeable, meaning that the joint probability distribution $p(y_1, \dots, y_n)$ is invariant to permutations of the indices. In the framework of finite mixtures, we assume that the population is made of $K \leq n$ subpopulations, with K known and fixed. We assume that within each of these groups, the distribution of y_i, $i = 1, \dots, n$, can be modeled as $f(y_i \mid \theta_k^*)$, for $k = 1, \dots, K$, with f a density on $\mathcal{Y}$. In this way we assume that a common parametric family is assumed for all these component distributions, which however depend on specific parameter vectors θ_k^*. The last missing piece to construct a mixture model is the parameter describing the proportion of population from each component k: we denote this parameter with π_k, satisfying $\sum_{k=1}^K \pi_k = 1$. Denoting the full vectors of parameters as $\boldsymbol{\theta}^* = (\theta_1^*, \dots, \theta_K^*)$ and $\boldsymbol{\pi} = (\pi_1, \dots, \pi_K)$, the data distribution for observation i can be formulated as

$$p(y_i \mid \boldsymbol{\theta}^*, \boldsymbol{\pi}) = \pi_1 \, f(y_i \mid \theta_1^*) + \cdots + \pi_K \, f(y_i \mid \theta_K^*).$$

In mixture models it is convenient to think of the component indicators as missing data, and to impute them to obtain a much simpler form of the data distribution.

Hence we introduce the cluster indicator variables $c_i \in \{1, \dots, K\}$, with $c_i = k$ if y_i belongs to the k-th mixture component. Conditionally on the mixture proportions, it holds $\Pr(c_i = k \mid \boldsymbol{\pi}) = \pi_k$, for $k = 1, \dots, K$. Moreover, given these allocation variables, the y_i are assumed to be independent and their distribution factorizes as $p(y_i \mid c_i, \boldsymbol{\theta}^*) = \prod_{k=1}^{K} f(y_i \mid \theta_k^*)^{\mathbb{I}(c_i=k)}$, where $\mathbb{I}(A)$ is the indicator variable, taking value 1 if event A occurs and 0 otherwise. The joint density of the observed data and the unobserved allocation variables, conditionally on the model parameters, can now be written as

$$p(\boldsymbol{y}, \boldsymbol{c} \mid \boldsymbol{\theta}^*, \boldsymbol{\pi}) = p(\boldsymbol{y} \mid \boldsymbol{c}, \boldsymbol{\theta}^*)\, p(\boldsymbol{c} \mid \boldsymbol{\pi}) = \prod_{i=1}^{n} \prod_{k=1}^{K} \{\pi_k f(y_i \mid \theta_k^*)\}^{\mathbb{I}(c_i=k)}.$$

Having defined the data distribution, in a Bayesian framework we need to specify adequate prior distributions on the model parameters $\boldsymbol{\pi}$ and $\boldsymbol{\theta}^*$. The prior G_0 on $\boldsymbol{\theta}^*$ is usually chosen depending on the specific application and on the basis of the component distribution f. For the mixture proportions π_k, the conjugate and most natural prior distribution is the Dirichlet distribution, $\boldsymbol{\pi} \sim \text{Dirichlet}_K(\alpha_1, \dots, \alpha_K)$.

Thanks to the data augmentation based on the cluster allocation variables, the model also admits a useful hierarchical representation: for $i = 1, \dots, n$

$$\begin{aligned} y_i \mid c_i, \theta_1^*, \dots, \theta_K^* &\sim f(y_i \mid \theta_{c_i}^*) \\ \Pr(c_i = k \mid \pi_1, \dots, \pi_K) = \pi_k \quad &\text{for } k = 1, \dots, K \\ \theta_1^*, \dots, \theta_K^* &\sim G_0 \\ \pi_1, \dots, \pi_K &\sim \text{Dirichlet}_K(\alpha_1, \dots, \alpha_K). \end{aligned} \tag{1.2}$$

It is possible to rewrite Equation (1.2) in a slightly different way by thinking that each observation y_i is associated with a parameter θ_i, where these parameters are drawn from a discrete distribution G with support on the K locations $\{\theta_1^*, \dots, \theta_K^*\}$. The model then becomes, for $i = 1, \dots, n$

$$\begin{aligned} y_i \mid \theta_i &\sim f(y_i \mid \theta_i) \\ \theta_i \mid \boldsymbol{\theta}^*, \boldsymbol{\pi} &\sim G \\ G = \sum_{k=1}^{K} \pi_k \delta_{\theta_k^*} & \\ \theta_1^*, \dots, \theta_K^* &\sim G_0 \\ \pi_1, \dots, \pi_K &\sim \text{Dirichlet}_K(\alpha_1, \dots, \alpha_K). \end{aligned} \tag{1.3}$$

Posterior inference for finite mixture models

Posterior inference for mixture models is usually performed through Markov Chain Monte Carlo (MCMC) methods and, in particular, the Gibbs sampler, as the full conditional distributions after imputing the cluster indicators $\mathcal{C} = \{c_1, \dots, c_n\}$ are greatly simplified. Moreover, for the distribution of the mixture weights it is possible to exploit the conjugacy of the Dirichlet distribution with the multinomial model. A Gibbs sampler then simply iterates these three steps:

1. Update the cluster-specific parameters θ_k^*, for $k = 1, \ldots, K$, from

$$p(\theta_k^* \mid \mathcal{C}, \boldsymbol{y}) \propto G_0(\theta_k^*) \prod_{i:c_i=k} f(y_i \mid \theta_k^*).$$

2. Update the weights $\pi_1, \ldots, \pi_K$ by sampling from a Dirichlet distribution with updated parameters

$$\pi_1, \ldots, \pi_K \mid \mathcal{C} \sim \text{Dirichlet}_K(\alpha_1 + n_1, \ldots, \alpha_K + n_K)$$

where n_k is the number of observations allocated to cluster k, for $k = 1, \ldots, K$.

3. Update the cluster indicators: for $i = 1, \ldots, n$ and $k = 1, \ldots, K$,

$$\Pr(c_i = k \mid \boldsymbol{\pi}, \boldsymbol{\theta}^*, y_i) \propto \pi_k f(y_i \mid \theta_k^*).$$

1.3.2. **Dirichlet process mixture models**

Nonparametric mixtures extend model (1.3) by placing a nonparametric prior on G. The most common prior on random probability measures is the Dirichlet process (DP), introduced by Ferguson (1973; 1974). Draws from a DP are discrete distributions with probability one, hence they turned out useful as flexible mixing measures in discrete mixtures.

The Dirichlet process

The Dirichlet process is a stochastic process whose realizations are probability distributions with probability one. Stochastic processes are distributions over function spaces, with their realizations being random functions. In the case of the DP, it is a distribution over the space of probability measures, which are real-valued functions with particular properties, which can be interpreted as distributions over some probability space. In this Section we just briefly review some of the main properties of the DP, the reader can refer to, e.g., Müller et al. (2015) and Hjort et al. (2010) for an exhaustive review of the Dirichlet process prior and related models.

Formally, a random distribution G on some probability space Θ is said to follow a DP prior with base measure G_0 and concentration parameter α, denoted $G \sim \text{DP}(\alpha, G_0)$, if for any partition $\{B_1, \ldots, B_H\}$ of Θ

$$(G(B_1), \ldots, G(B_H)) \sim \text{Dirichlet}_H\big(\alpha G_0(B_1), \ldots, \alpha G_0(B_H)\big).$$

That is, the finite-dimensional marginal distributions of a DP are Dirichlet distributions.

The success of the DP mainly arises from two appealing characteristics: its large support, with respect to the space of probability distributions, and tractability of the posterior distribution. Closely related to this last aspect is the conjugacy property of the DP: as the finite dimensional Dirichlet distribution is conjugate to the multinomial likelihood, the DP is conjugate with respect to i.i.d. sampling, that is, with respect to a completely unknown distribution from i.i.d. data. More precisely, if we take $\{\theta_1, \ldots, \theta_n\}$ a sequence of independent draws from $G \sim \text{DP}(\alpha, G_0)$, then the posterior

distribution of G given these observed values is still a DP. Letting again $\{B_1, \ldots, B_H\}$ be a finite measurable partition of Θ, and letting n_h be the number of observed values in B_h, for $h = 1, \ldots, H$, the posterior distribution is given by

$$(G(B_1), \ldots, G(B_H)) \mid \theta_1, \ldots, \theta_n \sim \text{Dirichlet}_H\big(\alpha G_0(B_1) + n_1, \ldots, \alpha G_0(B_H) + n_H\big).$$

In other terms, the posterior distribution is still DP with updated parameters:

$$G \mid \theta_1, \ldots, \theta_n \sim \text{DP}\left(\alpha + n, \frac{\alpha G_0 + \sum_{i=1}^n \delta_{\theta_i}}{\alpha + n}\right)$$

where the posterior base measure is a weighted average between the prior base measure G_0 and the empirical distribution $\sum_{i=1}^n \delta_{\theta_i}/n$. The weight associated with the prior base distribution is proportional to α, while the empirical distribution has weight proportional to the number of observations.

Another useful result, which allows to get a better understanding of the effect of using a DP as mixing measure, is the represented by Blackwell-MacQueen urn scheme (Blackwell and MacQueen, 1973), which describes the predictive distribution of draws from a DP. Consider again a sequence $\{\theta_1, \ldots, \theta_n\}$ of independent draws from $G \sim \text{DP}(\alpha, G_0)$. The predictive distribution of θ_{n+1} conditioned on these values, and with G marginalized out is given by

$$\theta_{n+1} \mid \theta_1, \ldots, \theta_n \sim \frac{1}{\alpha + n}\left(\alpha G_0 + \sum_{i=1}^n \delta_{\theta_i}\right).$$

Therefore the posterior base measure given $\{\theta_1, \ldots, \theta_n\}$ is also the predictive distribution of θ_{n+1}. This distribution highlights the discreteness of draws from a DP, and allows to investigate the clustering structure induced by the DP when is used as mixing measure in mixture models. Since the distribution is discrete, it is possible that some of the values $\{\theta_1, \ldots, \theta_n\}$ will be repeated. In particular, the unique values of $\{\theta_1, \ldots, \theta_n\}$ induce a partition of the set $\{1, \ldots, n\}$ into clusters defined by observations with the same value. Denoting with $\{\theta_1^*, \ldots, \theta_{K^+}^*\}$ the unique values among the θ_i, and letting n_k be the number of θ_i equal to θ_k^*, for $i = 1, \ldots, n$ and $k = 1, \ldots, K^+$, the predictive distribution can be written as

$$\theta_{n+1} \mid \theta_1, \ldots, \theta_n \sim \frac{1}{\alpha + n}\left(\alpha G_0 + \sum_{k=1}^{K^+} n_k \delta_{\theta_k^*}\right).$$

From this equation, it is possible to notice that θ_{n+1} will take a value θ_k^* with a probability proportional to n_k, the number of times it has already been observed. Hence, the larger n_k is, the higher the probability that it will grow. This is a rich-gets-richer phenomenon, where large clusters grow larger faster.

The DP admits several nice representations. An intuitive constructive definition of a DP random probability measure is given by Sethuraman (1994) and is based on the discrete nature of the process, which can be represented as a a weighted sum of point masses. This definition states that if $G \sim \text{DP}(\alpha, G_0)$, then it can be expressed as

follows:

$$\begin{aligned} v_k \sim \text{Beta}(1, \alpha), \quad \theta_k^* \sim G_0 \qquad & \text{for } k \geq 1 \\ \pi_1 = v_1, \quad \pi_k = v_k \prod_{h=1}^{k-1} (1 - v_h) \qquad & \text{for } k \geq 2 \\ G = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k^*}. & \end{aligned} \tag{1.4}$$

The construction of the weights $\{\pi_k\}_{k=1}^{\infty}$ by means of beta random variable is usually called stick-breaking process, also denoted as $\boldsymbol{\pi} \sim \text{GEM}(\alpha)$ after the names of Griffiths, Engen and McCloskey. The name arises from a metaphor for this construction, where a unit stick is broken in infinitely many parts, and each piece is used to define a weight. Because of its simplicity, this representation has motivated extensions of the process as well as new algorithms for posterior inference.

Dirichlet process mixtures

Getting back to the framework of mixture models, from the last representation of the DP it is clear that this process can be conveniently used as mixing measure in mixture models. Following the structure of Equation (1.3) for finite mixtures, DP mixtures (DPMs) can be written as

$$\begin{aligned} y_i \mid \theta_i &\sim f(y_i \mid \theta_i) \\ \theta_i \mid G &\sim G \\ G &\sim \text{DP}(\alpha, G_0). \end{aligned}$$

Alternatively, making use of the set of unique values $\{\theta_k^*\}_{k=1}^{\infty}$ and of Sethuraman's representation, the model can be expressed as

$$p(y_i \mid G) = \sum_{k=1}^{\infty} \pi_k f(y_i \mid \theta_k^*) \tag{1.5}$$

where the weights $\{\pi_k\}_{k=1}^{\infty}$ follow a stick-breaking construction and the $\{\theta_k^*\}_{k=1}^{\infty}$ are i.i.d. samples from the base measure G_0.

Differently from finite mixtures, DPMs are infinite mixture models, as they assume a countably infinite number of components. However, because the π_k's decrease exponentially quickly, only a small number of components will be used to model the data a priori: in the following, we will define *clusters* these non-empty components. In general, in a sample of size n, the prior expected number of clusters $\mathbb{E}(K^+)$ is approximately equal to $\alpha \log(1 + n/\alpha)$. This means that while in finite mixture models the number of clusters has to be fixed in advance, in DPMs the number of clusters is determined by the data and can be inferred.

Posterior inference for DPMs

Applying MCMC techniques to DPMs directly is not feasible, as it would require imputing the infinite-dimensional distribution G. Instead, successful algorithms to

perform posterior inference on DP mixtures have been developed using the particular representations that this process admits. It is common to divide these methods into two classes depending on the strategy they adopt to deal with the infinite-dimensional component of the model: marginal algorithms are based on model representations with the DP is integrated out, while conditional algorithms explicitly represent the measure generated by the process using its stick-breaking construction. Here we will focus on conditional methods. Avoiding marginalization of the random measure allows to perform inference on it directly, moreover, these methods are often computationally more efficient than marginal ones; however, they require to devise good strategies to deal with the infinite dimension of the process.

In the blocked Gibbs sampler of Ishwaran and James (2001) the infinite random measure is truncated to an upper bound K for the number of components. The motivation that justifies this approach is that the weights $\{\pi_k\}_{k=1}^{\infty}$ determined by the stick-breaking construction are stochastically decreasing in k. By choosing a sufficiently large value K one can assume that $\sum_{k=K+1}^{\infty} \pi_k$ has a distribution concentrated near zero. Adopting such truncation leads to a representation of the model as finite mixture, so the sampler described in Section 1.3.1 can be applied with just few changes on the weights distribution. In particular, the weights are now sampled from a stick-breaking process with $v_k \sim \text{Beta}(1 + n_k, \alpha + \sum_{h=k+1}^{K} n_h)$ for $k = 1, \dots, K-1$. Finally, letting $v_K = 1$ ensures that the first K weights sum to one. Adopting this approach leads to a straightforward MCMC algorithm for posterior inference, however, in some applications K needs to be set to very large values in order to obtain an adequate approximation.

The slice sampler (Walker, 2007; Kalli et al., 2011) has been adopted as an alternative "dynamic" truncation method to automatically select the active components of the mixture. The slice sampler relies on the introduction of latent uniform random variables u_i, $i = 1, \dots, n$, such that marginalizing the joint density of (y_i, u_i) still returns the desired density in Eq. (1.5), and such that the conditional density of $y_i \mid u_i$ only involves a finite number of mixture components. Specifically, model (1.5) can be obtained as marginal density with respect to u_i of

$$p(y_i, u_i \mid G) = \sum_{k=1}^{\infty} \mathbb{I}(u_i < \pi_k) f(y_i \mid \theta_k^*).$$

The key of introducing this variable is that, conditionally on u_i, the density can be written as

$$p(y_i \mid u_i, G) = N_u^{-1} \sum_{k \in A_u} f(y_i \mid \theta_k^*)$$

where A_u is the set of indices of active components $A_u = \{k : \pi_k > u\}$ and $N_u = \sum_{k \in A_u} \pi_k$. This model defines a finite mixture with equal weights N_u^{-1}. Introducing the cluster indicators $\mathcal{C} = \{c_1, \dots, c_n\}$ further simplifies the density, similarly to the case of finite mixtures. Finally, Kalli et al. (2011) also introduce a non-stochastic positive sequence $\{\xi_1, \xi_2, \dots\}$ in order to improve mixing (we refer to the original paper for a discussion on the choice of the sequence). With all these elements, the joint density of

$(\boldsymbol{y}, \boldsymbol{u}, \mathcal{C})$ conditioned on $\boldsymbol{\pi}$ and $\boldsymbol{\theta}^*$ becomes

$$p(\boldsymbol{y}, \boldsymbol{u}, \mathcal{C} \mid \boldsymbol{\pi}, \boldsymbol{\theta}^*) = \prod_{i=1}^{n} \mathbb{I}(u_i < \xi_{c_i}) \frac{\pi_{c_i}}{\xi_{c_i}} f(y_i \mid \theta^*_{c_i}). \tag{1.6}$$

The slice sampler algorithm iterates the following steps:

1. Update the cluster-specific parameters θ^*_k, for $k = 1, \dots, K$, where K is defined as the maximum index h such that $\xi_h > u_i$ for all $i = 1, \dots, n$, from

$$p(\theta^* \mid \mathcal{C}) \propto G_0(\theta^*_k) \prod_{i:c_i=k} f(y_i \mid \theta^*_k).$$

2. Update the weights π_k for $k = 1, \dots, K$ using the stick-breaking construction (1.4) with beta random variables $v_k \sim \text{Beta}(1 + n_k, \alpha + \sum_{h=k+1}^{K} n_h)$.

3. Sample the latent variables u_i from a uniform distribution $u_i \mid c_i \sim \text{Unif}(0, \xi_{c_i})$

4. Update the cluster allocations c_i for $i = 1, \dots, n$ from

$$\Pr(c_i = k \mid u_i, y_i, \theta^*_k) \propto \mathbb{I}(k : \xi_k > u_i) \frac{\pi_k}{\xi_k} f(y_i \mid \theta^*_k).$$

1.3.3. **Finite mixtures with an unknown number of components**

To avoid fixing the number of clusters, a different approach to DP mixtures is to consider finite mixtures with a prior on the number of components. Although this may seem as the most natural way to infer the number of groups, application of finite mixtures with an unknown number of components, recently often called mixtures of finite mixtures (MFMs), has long been hindered by the difficulty of performing posterior inference. Several inference methods have been proposed for this class of models (Richardson and Green, 1997; Nobile, 2004; Nobile and Fearnside, 2007; McCullagh and Yang, 2008) often exploiting the reversible jump Markov chain Monte Carlo techniques. However, applying the reversible jump in new situations can be hard, as it requires designing new reversible jump moves.

Because of these difficulties, researchers have also investigated alternative strategies: for example, Stephens (2000) exploited birth-death processes to construct an ergodic Markov chain with the appropriate stationary distribution. More recently, different models have been proposed inspired by nonparametric mixtures (Miller and Harrison, 2018; Argiento and De Iorio, 2019). Miller and Harrison (2018) explicitly derived MFMs counterparts for several properties exhibited by DPMs. However, their approach is restricted to a fixed parameter α for the Dirichlet distribution, regardless of the dimension K, and only relies on marginal methods for performing posterior inference. Argiento and De Iorio (2019) instead introduced a new class of random measures based on the normalization of a point process, and used it as mixing measure in a MFMs. Moreover, they outlined marginal and conditional methods for posterior inference using this class of discrete random measures.

Another approach is discussed by Malsiner-Walli et al. (2016) and Frühwirth-Schnatter and Malsiner-Walli (2019), based on the use of sparse mixtures. Similarly

to DPMs, in this formulation they distinguish between the mixture components and the clusters, which are defined as the components actually used by the data. In their approach, the number of components K is fixed to a large and clearly overfitting value, and a Dirichlet prior with small parameter on the mixture weights ensures that the (random) number of clusters K_+ will take values smaller than K with high probability a priori and also a posteriori.

Generalized mixtures of finite mixtures

A general formulation that includes several models described above as special cases (finite mixtures with a fixed number of clusters, MFMs and overfitted mixtures) has recently been described by Frühwirth-Schnatter et al. (2021): they call this specification generalized mixture of finite mixtures (gMFM). Similarly to overfitted mixtures, a key aspect of this approach is the distinction between the number of components K and the number of clusters K_+; however, here, the number of components is also random. In the following we will review some aspects of these models that will be used later in this thesis. The gMFM model can be defined in a hierarchical form analogous to Equation (1.2) as

$$\begin{gathered} y_i \mid K, c_i = k, \theta_k^* \sim f(y_i \mid \theta_k^*) \\ \Pr(c_i = k \mid K, \pi_1, \dots, \pi_K) = \pi_k \quad \text{for } k = 1, \dots, K \\ \theta_1^*, \dots, \theta_K^* \sim G_0 \\ \pi_1, \dots, \pi_K \mid K, \alpha \sim \text{Dirichlet}_K(\alpha/K, \dots, \alpha/K) \\ K \sim p(K) \end{gathered}$$

Including a prior on K also induces a prior on the number of clusters K_+: here a crucial role is played by the sequence of concentration parameters of the Dirichlet distribution. Considering fixed parameters as in Miller and Harrison (2018), where a $\text{Dirichlet}_K(1, \dots, 1)$ is used regardless of the value of K, leads to a prior expected number of clusters $\mathbb{E}(K_+)$ close to $\mathbb{E}(K)$ for many priors $p(K)$. Having instead concentration parameters that decrease with increasing K induces randomness in the prior distribution of $K_+ \mid K$, allowing for a gap between K_+ and K. In this formulation the parameters decrease linearly with K, and a F prior distribution is used for the hyperparameter α. The specification of a gMFM is completed with a suitable prior $p(K)$ on the number of components. In their work, Frühwirth-Schnatter et al. (2021) discuss different choices and compare the resulting prior on the number of clusters. A desirable prior on K should be weakly informative on the number of clusters, and should lead to a prior on K_+ which is concentrated on moderate number of clusters, with fat tails to ensure that also a high number of clusters may be estimated. They suggest to use a translated prior for K, where $K - 1$ follows a beta-negative-binomial distribution, which is a hierarchical generalization of the Poisson, the geometric and the negative-binomial distribution.

Posterior inference for gMFMs

An important contribution of the work of Frühwirth-Schnatter et al. (2021) is the introduction of a new inference algorithm "telescoping sampler" to obtain the posterior

distribution of all model parameters without resorting to reversible jump MCMC techniques. Their algorithm is a trans-dimensional Gibbs sampler which at each iteration alternates two key steps: first, it updates the partition of the observations $\mathcal{C} = \{c_1, \ldots, c_n\}$ conditionally on the number of components, then, it samples a new value for K on the basis of this partition. Explicitly sampling the number of mixture components greatly simplifies inference as, conditionally on K, the updates of the partition and of the model parameters are brought back to standard steps. Hence the crucial aspect is sampling from the full conditional of the number of components. This is achieved by combining the conditional exchangeable partition probability function $p(\mathcal{C} \mid n, K, \alpha)$, derived in Section 2.2 of Frühwirth-Schnatter et al. (2021), with the prior $p(K)$. The algorithm performs the following steps:

1. Update the partition $\mathcal{C}$:
   (a) Sample c_i, for $i = 1, \ldots, n$ from $\Pr(c_i = k \mid \boldsymbol{\pi}, \boldsymbol{\theta}^*, y) \propto \pi_k f(y_i \mid \theta_k^*)$;
   (b) Determine the number K_+ of non-empty clusters and relabel the components such that the first K_+ clusters are non-empty.
2. Conditional on $\mathcal{C}$, update the parameters of the non-empty components (and eventual hyperparamters):
$$p(\theta_k^* \mid \mathcal{C}, \boldsymbol{y}) \propto G_0(\boldsymbol{\theta}^*) \prod_{i:c_i=k} f(y_i \mid \theta_k^*) \quad \text{for } k = 1, \ldots, K.$$
3. Conditional on $\mathcal{C}$, draw a new value for K from
$$p(K \mid \mathcal{C}, \alpha) \propto p(\mathcal{C} \mid n, K, \alpha) p(K) = p(K) \frac{K!\, \alpha^{K_+}}{(K - K_+)!\, K^{K_+}} \prod_{k=1}^{K_+} \frac{\Gamma(n_k + \alpha/K)}{\Gamma(1 + \alpha/K)}$$
   for $K = K_+, K_+ + 1, \ldots$; where n_k is the number of observations in cluster k.
4. Update α by performing a Metropolis-Hastings step to sample α from its full conditional
$$p(\alpha \mid \mathcal{C}, K) \propto p(\alpha) \frac{\alpha^{K_+} \Gamma(\alpha)}{\Gamma(n + \alpha)} \prod_{k=1}^{K_+} \frac{\Gamma(n_k + \alpha/K)}{\Gamma(1 + \alpha/K)}$$
5. Add $K - K_+$ empty components,
   (a) if $K > K_+$, sample $K - K_+$ new values θ_k^* from the prior, $k = K_+ + 1, \ldots, K$;
   (b) update the weight vector as
$$\pi_1, \ldots, \pi_K \mid K, \alpha, \mathcal{C} \sim \text{Dirichlet}_K(\alpha/K + n_1, \ldots, \alpha/K + n_K).$$

1.3.4. **Bayesian nonparametric models for nested data**

All models described so far assumed that the data were exchangeable, that is, they arise from one unknown distribution. However, there is growing interest in modeling scenarios where the data come from different but related groups, as for example different populations or experiments. In these cases it is desirable to individually

model the distribution of each group, while also borrowing information between them. When data are collected from individuals in multiple groups, the exchangeability assumption is no longer valid: in these cases, observations are said to be partially exchangeable, meaning that they are exchangeable *within* groups.

Suppose that in addition to each $y_i, i = 1, \dots, n$, we also observe a categorical variable g_i with values in $\{1, \dots, J\}$ indicating the population to which y_i belongs, so that when $g_i = j$ means that y_i comes from the j-th group. In the Bayesian nonparametric framework, several approaches have been proposed to deal with these nested data sets.

The hierarchical Dirichlet process (Teh et al., 2006) arises from the desire to flexibly model the distributions of the observations of each group, while also performing clustering to capture latent structures among all individuals. Each group-specific distribution is modeled with a mixture model which uses a random probability measure G_j as mixing measure, where the G_j's are distributed according to a DP: this allows to obtain a partition of individuals within each group. In order to borrow information across groups, they propose a hierarchical formulation where one draw from a Dirichlet process is used as the base measure G_0 of the Dirichlet process generating the individual G_j's. This construction implies that the distributions $\{G_1, \dots, G_J\}$ share the same atoms (the atoms of G_0), thus the model yields a clustering of the individuals also across groups. However, as these G_j's are independent draws, in general, they will have different weights: as a result, there is no clustering of these group-specific distributions. The hierarchical DP hence does not allow to investigate similarities between the distributions, but only to cluster individuals.

To overcome this limitation, Rodríguez et al. (2008) introduced the nested Dirichlet process, which allows to obtain a clustering both of the observations within each group, and of the groups themselves. Consider again a collection of distributions $\{G_1, \dots, G_J\}$ serving as group-specific mixing measures of a nonparametric mixture. The nested DP assumes that $G_j \mid Q \sim Q$ for $j = 1, \dots, J$ and $Q \sim DP(\alpha, DP(\beta, G_0))$. Using the stick-breaking representation of the DP, this model can be expressed as

$$G_1, \dots, G_J \mid Q \sim Q, \qquad Q = \sum_{k=1}^{\infty} \pi_k \delta_{G_k^*} \tag{1.7}$$

$$G_k^* = \sum_{l=1}^{\infty} \omega_{l,k} \delta_{\theta_{l,k}^*}$$

where the sequences of weights $\{\pi_k\}_{k=1}^{\infty}$ and $\{\omega_{l,k}\}_{l=1}^{\infty}$ for $k \geq 1$ follow a stick breaking construction, $\boldsymbol{\pi} \sim \text{GEM}(\alpha)$ and $\boldsymbol{\omega}_k \sim \text{GEM}(\beta)$, and the parameters $\theta_{l,k}^*$ are i.i.d. samples from G_0. A main difference from the hierarchical DP is that this formulation allows to cluster the group-specific distributions: indeed, if two G_j's are assigned to the same G_k^*, then the observations in the two groups have exactly the same generating distribution. When two groups share the same distribution, they are said to belong to the same distributional cluster.

In a recent paper by Camerlenghi et al. (2019), however, they noted a degeneracy property that occurs in the nested DP in case of ties across samples at the observed or latent level. In particular, if two distributions G_1 and G_2 share at least one atom, then their posterior distribution degenerates on $\{G_1 = G_2\}$, forcing homogeneity across the two samples. To overcome this drawback, they introduce a novel class of latent nested

processes. These processes are based on a mixture of two random distributions: an idiosyncratic component and a shared component. The shared random distribution explicitly accounts for the possibility of observing some common atoms, while the idiosyncratic one allows some atoms to be distinct and specific to each subpopulation. However, implementation of this model becomes challenging and computationally burdensome when the number of groups increases.

The common atoms model

Another nested nonparametric prior, more suited to practical applications, is proposed by Denti et al. (2021). They formulate a constrained modification of the nested DP which, however, does not suffer from the degeneracy issue. Moreover, compared to the models introduced by Camerlenghi et al. (2019), it is computationally more efficient and allows to perform posterior inference in a quite straightforward way even when the number of groups is moderate. The first level of their common atoms model (CAM) is analogous to the nested DP (Eq. 1.7), however, the specification of the distributional atoms G_k^* makes use of a common set of atoms for all k. Hence the distributions G_k^* can be seen as realizations of a single-atom dependent DP,

$$G_k^* = \sum_{l=1}^{\infty} \omega_{l,k} \delta_{\theta_l^*}$$

where the sequences of weights $\{\omega_{l,k}\}_{l=1}^{\infty}$ for $k \geq 1$ again follow a stick breaking construction, and the common atoms $\{\theta_l^*\}_{l=1}^{\infty}$ are independent draws from a base measure G_0. Similarly to the nested DP, the first mixture level of the CAM allows to perform a clustering of the group-specific distributions G_j; however, as the set of atoms defining the G_k^* is common across distributions, the CAM also allows to obtain a clustering of individuals both within each group and across them, in a similar fashion to the hierarchical DP.

Convoluting this nested prior with a kernel we obtain a nested infinite mixture model: the density for observations $\boldsymbol{y}_j = \{y_i : g_i = j; i = 1, \dots, n\}$ belonging to group j can be written as

$$p(\boldsymbol{y}_j \mid \boldsymbol{\theta}^*, \boldsymbol{\pi}, \boldsymbol{\omega}) = \sum_{k=1}^{\infty} \pi_k \prod_{i:g_i=j} \sum_{l=1}^{\infty} \omega_{l,k} f(y_i \mid \theta_l^*).$$

Introducing two sequences of latent cluster allocations $\mathcal{C}^D = \{c_{g_i}^D\}$ and $\mathcal{C} = \{c_i\}$ for $i = 1, \dots, n$, corresponding respectively to the distributional cluster allocation of the J groups and to the observational cluster allocation of individuals, the model admits a hierarchical representation as

$$\begin{aligned} y_i \mid c_i, \boldsymbol{\theta}^* &\sim f(y_i \mid \theta_{c_i}^*) \\ c_i \mid c_{g_i}^D = k, \boldsymbol{\omega}_k \sim \sum_{l=1}^{\infty} \omega_{l,k} \delta_l \qquad & \boldsymbol{\omega}_k \sim \text{GEM}(\beta) \quad \text{for } k \geq 1 \\ c_{g_i}^D \mid \boldsymbol{\pi} \sim \sum_{k=1}^{\infty} \pi_k \delta_k \qquad & \boldsymbol{\pi} \sim \text{GEM}(\alpha) \\ \theta_l^* &\sim G_0 \quad \text{for } l \geq 1. \end{aligned}$$

Posterior inference for the CAM

To infer on the posterior distribution of the CAM model, Denti et al. (2021) develop a nested version of the slice sampler of Kalli et al. (2011) (described in the last part of Section 1.3.2). This sampler relies on two sequences of latent uniform variables $\boldsymbol{u}^D = \{u_{g_i}\}_{g_i=1}^J$ (on the distributional level) and $\boldsymbol{u}^O = \{u_i^O\}_{i=1}^n$ (on the observational level) pointing to the active mixture components. Moreover, they also introduce two deterministic sequences $\boldsymbol{\xi}^D = \{\xi_k\}_{k=1}^\infty$ and $\boldsymbol{\xi}^O = \{\xi_{l,k}^O\}_{l=1}^\infty$ for $k \geq 1$. Similarly to Equation 1.6, the model can be written as

$$p(\boldsymbol{y}, \boldsymbol{u}^D, \boldsymbol{u}^O, \mathcal{C}^D, \mathcal{C} \mid \boldsymbol{\theta}^*, \boldsymbol{\pi}, \boldsymbol{\omega}) = \prod_{j=1}^J \mathbb{I}(u_j^D < \xi_{c_j^D}^D) \frac{\pi_{c_j^D}}{\xi_{c_j^D}^D} \prod_{i:g_i=j} \mathbb{I}(u_i^O < \xi_{c_i,c_j^D}^O) \frac{\omega_{c_i,c_j^D}}{\xi_{c_i,c_j^D}^O} f(y_i \mid \theta_{c_i}^*).$$

Then, the algorithm iterates the following steps

1. Sample the latent uniform random variables
   (a) At the distributional level, for $j = 1, \ldots, J$, sample $u_j^D \sim \text{Unif}(0, \xi_{c_j^D}^D)$
   (b) At the observational level, for $i = 1, \ldots, n$, sample $u_i^O \sim \text{Unif}(0, \xi_{c_i,c_{g_i}^D}^O)$
2. Update the weight vectors $\boldsymbol{\pi}$ and $\boldsymbol{\omega}_k$, for $k \geq 1$
   (a) At the distributional level, sample the distributional stick-breaking proportions $v_k \sim \text{Beta}(1+m_k, \alpha+\sum_{h=k+1}^{K^*} m_h)$, where m_k is the number of groups in distributional cluster k.
   (b) At the observational level, for each $k = 1, \ldots, K^*$, sample the stick-breaking proportions $u_{l,k} \sim \text{Beta}(1 + n_l^k, \beta + \sum_{h=l+1}^{L^*} n_h^k)$, where n_l^k is the number of individuals assigned to observational cluster l and distributional cluster k.
3. Update the cluster indicators
   (a) For the distributional clusters, sample the variables c_j^D from

$$\Pr(c_j^D = k \mid \boldsymbol{u}^D, \boldsymbol{\pi}, \boldsymbol{\omega_k}, \mathcal{C}) \propto \mathbb{I}(u_j^D < \xi_{c_j^D}^D) \frac{\pi_{c_j^D}}{\xi_{c_j^D}^D} \prod_{i:g_i=j} \omega_{c_i,k}$$

   (b) For the observational clusters, sample the variables c_i from

$$\Pr(c_i = l \mid y_i, g_i = j, c_j^D, \boldsymbol{u}^O, \omega_{l,c_j^D}) \propto \mathbb{I}(u_i^O < \xi_{l,c_j^D}^O) \frac{\omega_{l,c_j^D}}{\xi_{l,c_j^D}^O} f(y_i \mid \theta_l^*).$$

4. Conditional on the observational clusters, sample the cluster-specific parameters θ_l^* from

$$p(\theta^* \mid \mathcal{C}) \propto G_0(\theta_l^*) \prod_{i:c_i=l} f(y_i \mid \theta_l^*).$$

We refer to the Supplementary Material of the original paper for details about the specific sequences ξ^O and ξ^D and computation of the upper bounds K^* and L^* involved at step 2.

2 NOVEL POSTERIOR SAMPLING SCHEMES FOR POISSON ENCODING MODELS

As we have seen in Section 1.1.2, encoding models are an important tool to study how the deconvolved spike trains vary with the underlying experimental conditions. To this end, GLMs provide a simple and flexible strategy to estimate the impact of each covariate on the mean of the variable of interest. A relevant question that can be addressed using GLMs is how the number of spikes detected during a specific experiment is affected by the experimental conditions and the characteristics of the neurons. A plausible assumption to model the spike counts per time bin is to use a Poisson distribution (Paninski et al., 2007): in the context of GLMs, this setting naturally leads to the use of Poisson regression models.

Poisson regression models are common in statistics and represent one of the most popular choices to model how the distribution of count data varies with predictors. A typical assumption is that, under an independent sample of counts, $y_1, \dots, y_n$, the probability mass function of the generic y_i conditionally on a p-dimensional vector of covariates x_i is

$$f(y_i \mid \lambda_i) = \frac{\lambda_i^{y_i}}{y_i!} e^{-\lambda_i}, \quad \log(\lambda_i) = x_i^T \beta, \tag{2.1}$$

where β is a p-dimensional vector of unknown coefficients. Linking the linear predictor $x_i^T \beta$ and the parameter λ_i with the logarithm represents the most natural choice, as the logarithm is the canonical link for the Poisson family (Nelder and Wedderburn, 1972). Besides encoding models for spike train analyses, this model has broad application in several fields, including medicine and epidemiology (Frome, 1983; Frome and Checkoway, 1985; Hutchinson and Holtman, 2005), manufacturing process control (Lambert, 1992), analysis of accident rates (Joshua and Garber, 1990; Miaou, 1994), and crowd counting (Chan and Vasconcelos, 2009), among others.

Adopting a Bayesian approach can be particularly convenient in the context of calcium imaging studies. As pointed out by Paninski et al. (2007), often some regularization technique is needed to obtain reliable estimates of the effects, as in some experiments the large number of covariates leads to a sensible risk of overfitting. The Bayesian paradigm offers a natural approach to regularized regression: there is a large literature on prior inducing some kind of shrinkage or selection, as, for example, the spike-and-slab prior (Mitchell and Beauchamp, 1988), the Bayesian lasso (Park and Casella, 2008), the horseshoe prior and its extensions (Carvalho et al., 2010; Piironen and Vehtari, 2017).

However, model (2.1) does not enjoy any conjugacy property and, thus, regardless of the prior used, the posterior distribution of β is not available in close form. Consequently, inference is conducted using Markov Chain Monte Carlo (MCMC) methods, which obtain a sample from the posterior distribution of the parameters.

Several approaches have focused on how to easily obtain the posterior distribution of the coefficients of Poisson models without requiring complex tuning strategies or long computation times. In the context of count-valued time series, Frühwirth-Schnatter and Wagner (2006) proposed a formulation of the model based on two levels of data augmentation, to derive an efficient approximate Gibbs sampler. Frühwirth-Schnatter et al. (2009) exploited a data augmentation strategy to simplify the computation of hierarchical models for count and multinomial data. Data augmentation strategies have also been employed in the case of models for multivariate dependent count data (Karlis and Meligkotsidou, 2005; Bradley et al., 2018). However, the simplest Poisson regression in (2.1) still lacks a specific and efficient algorithm to sample from the posterior distribution of the parameters β for any prior choice, making the Metropolis-Hastings (Hastings, 1970) or Hamiltonian Monte Carlo (HMC) (Neal, 2011) algorithms the only available options.

On the other hand, several efficient computational strategies for binary regression models have been proposed in the literature. Using the probit link, Albert and Chib (1993) proposed an efficient data augmentation based on a latent Gaussian variable, while the more recent contribution by Polson et al. (2013) exploited the canonical logit link, introducing an efficient Pólya-gamma data augmentation scheme. Leveraging Polson et al. (2013) approach, we propose a novel approximation of the posterior distribution that can be exploited as proposal distribution of a Metropolis-Hastings algorithm or as importance density of an importance sampling for Poisson log-linear models with conditional Gaussian prior distributions on the regression parameters. With conditional Gaussian prior, we refer to a possibly hierarchical prior with conditional distribution $\beta \sim \mathrm{N}(b, B)$, with b and/or B random. Examples include straightforward Gaussian prior distributions with informative (b, B) fixed using prior information, and scale mixtures of Gaussian where b is set to zero and the variance has a suitable hierarchical representation, such as the Bayesian lasso prior or the horseshoe prior, among others.

More specifically, we introduce an approximation of the posterior density that exploits the negative binomial convergence to the Poisson distribution. Thanks to this result, we are able to leverage the Pólya-gamma data augmentation scheme of Polson et al. (2013) to derive an efficient sampling scheme. In the next section, we introduce and discuss the proposed algorithms, starting from the definition of an *approximate* posterior distribution whose sampling can be performed straightforwardly. Sampling from this approximate posterior is then used as proposal density for the Metropolis-Hastings or importance sampler. The performance of the proposed algorithms in terms of computational efficiency is compared with that of state-of-the-art methods in a simulation study. Finally, we employ the proposed method to estimate a deconvolution model on spike train data from the Allen Brain Observatory.

2.1. Efficient posterior sampling strategies

2.1.1. Approximate posterior distribution

Assume $y_1, \dots, y_n$ is an independent sample of counts from model (2.1). We introduce an approximation of the posterior density which exploits the negative binomial convergence to the Poisson distribution, i.e., we approximate the i-th contribution to the likelihood function $f(y_i \mid \lambda_i)$ with $\tilde{f}_{r_i}(y_i \mid \lambda_i)$ where

$$\tilde{f}_{r_i}(y_i \mid \lambda_i) = \left(\frac{r_i}{r_i + \lambda_i}\right)^{r_i} \left(\frac{\lambda_i}{r_i + \lambda_i}\right)^{y_i}, \tag{2.2}$$

which corresponds to the probability mass function of a negative binomial random variable with parameter r_i, the number of failures until the experiment is stopped, and success probability $\lambda_i/(r_i + \lambda_i)$. As r_i approaches infinity, this quantity converges to a Poisson likelihood.

Following Polson et al. (2013), we rewrite each i-th contribution to the approximate likelihood (2.2) by introducing augmented Pólya-gamma random variables $\omega_i \sim \text{PG}(y_i + r_i, 0)$, i.e

$$\begin{aligned}\tilde{f}_{r_i}(y_i \mid \beta) = &\exp\left\{\frac{(x_i^{\text{T}}\beta - \log r_i)(y_i - r_i)}{2}\right\} 2^{-(y_i + r_i)} \times \\ &\int_0^{+\infty} \exp\left\{-\omega_i \frac{(x_i^{\text{T}}\beta - \log r_i)^2}{2}\right\} f_{\text{PG}}(\omega_i; y_i + r_i, 0)\, d\omega_i,\end{aligned}$$

where $f_{\text{PG}}(\cdot; \xi, \zeta)$ denotes the density of a Pólya-gamma with parameters (ξ, ζ).

In what follows, we assume that prior knowledge about the unknown β parameters is represented by a conditionally Gaussian prior, i.e. $\beta \sim \text{N}(b, B)$, with a possible hierarchical representation for the parameters b and B. Examples include default informative Gaussian with fixed (b, B) or scale mixtures of Gaussian where b is set to zero and the variance has a suitable hierarchical representation (Park and Casella, 2008; Carvalho et al., 2010; Piironen and Vehtari, 2017).

The *approximate* posterior based on the conditionally Gaussian prior $\beta \sim \text{N}(b, B)$ and approximate likelihood $\prod_{i=1}^n \tilde{f}_{r_i}(y_i \mid \beta)$ is consistent with the successful Gibbs sampler of Polson et al. (2013); i.e., sampling from the *approximate* posterior is equivalent to sampling iteratively from the following full conditionals

$$\omega_i | \beta \sim \text{PG}(y_i + r_i, x_i^{\text{T}}\beta - \log r_i), \qquad \beta | y, \omega \sim \text{N}_p(m_\omega, V_\omega), \tag{2.3}$$

where $V_\omega = (X^{\text{T}}\Omega X + B^{-1})$ and $m_\omega = V_\omega(X^{\text{T}}\kappa + B^{-1}b)$, with $\Omega = \text{diag}\{\omega_1, \dots, \omega_n\}$ and $\kappa = (\omega_1 \log r_1 + (y_1 - r_1)/2, \dots, \omega_n \log r_n + (y_n - r_n)/2)$.

The adherence of this approximate posterior to the true posterior highly depends on the values of r_i, with larger values of r_i resulting in better approximations. However, when employing this result in posterior sampling, large values of r_i imply longer computation time due to the computational cost of sampling Pólya-gamma random variables with large parameters. Although the specific choice of r_i remains an open point—discussed later in Section 2.1.4—in the context of MCMC sampling, we propose to reduce the computational burden related to the sampling of n Pólya-gamma random

variables marginalizing the Gaussian distribution in (2.3) with respect to the related Pólya-gamma density conditioned on $\beta^{(t-1)}$, the last available β sampled. Since this marginalization is not in a closed form we introduce a second level of approximation of the true posterior. Specifically, we introduce $q(\beta \mid \beta^{(t-1)})$ a density that depends on $\beta^{(t-1)}$, defined as the first-order Taylor expansion of the marginalized Gaussian distribution, i.e.

$$q(\beta \mid \beta^{(t-1)}) = (2\pi)^{-p/2}\det(V_{\mathbb{E}(\omega)})^{-1/2}\exp\left\{-\frac{1}{2}(\beta - m_{\mathbb{E}(\omega)})^{\mathrm{T}}V_{\mathbb{E}(\omega)}^{-1}(\beta - m_{\mathbb{E}(\omega)})\right\}, \tag{2.4}$$

where $V_{\mathbb{E}(\omega)} = (X^{\mathrm{T}}\Omega X + B^{-1})$, $m_{\mathbb{E}(\omega)} = V_{\mathbb{E}(\omega)}(X^{\mathrm{T}}\kappa + B^{-1}b)$, $\Omega = \mathrm{diag}\{\mathbb{E}(\omega_1), \ldots, \mathbb{E}(\omega_n)\}$, $\kappa = (\mathbb{E}(\omega_1)\log r_i + (y_1 - r_i)/2, \ldots, \mathbb{E}(\omega_n)\log r_i + (y_n - r_i)/2)$, and for each $i = 1, \ldots, n$ the conditional expectation of each ω_i is simply

$$\mathbb{E}\left(\omega_i^*\right) = \frac{r_i + y_i}{2(x_i^{\mathrm{T}}\beta^{(t-1)} - \log r_i)}\left(\frac{e^{x_i^{\mathrm{T}}\beta^{(t-1)}} - r_i}{e^{x_i^{\mathrm{T}}\beta^{(t-1)}} + r_i}\right),$$

or equivalently

$$\beta \mid \beta^{(t-1)} \sim \mathrm{N}(m_{\mathbb{E}(\omega)}, V_{\mathbb{E}(\omega)}). \tag{2.5}$$

The above construction is eventually used as the building block of efficient Metropolis-Hastings and importance sampling algorithms, as described in the following sections.

2.1.2. **Metropolis-Hastings sampler**

We employ the above sampling mechanism as the proposal density in a Metropolis-Hastings algorithm. Consistent with this, at each iteration of the MCMC sampler, an additional step that accepts or rejects the proposed draw is introduced. Specifically, we assume that conditionally on the current state of the chain $\beta^{(t-1)}$, a new value β^* is sampled from (2.5). Then, the acceptance probability

$$\alpha(\beta^*, \beta^{(t-1)}) = \min\left\{1, \frac{\pi(\beta^* \mid y)}{\pi(\beta^{(t-1)} \mid y)}\frac{q(\beta^{(t-1)} \mid \beta^*)}{q(\beta^* \mid \beta^{(t)})}\right\}, \tag{2.6}$$

is evaluated to decide whether to accept or reject the proposed β^*, where $\pi(\beta^* \mid y)$ is the exact posterior distribution.

To compute the acceptance probability in (2.6), the forward and backward transition densities $q(\beta^* \mid \beta^{(t-1)})$ and $q(\beta^{(t-1)} \mid \beta^*)$ must be computed. Consistent with this, approximation (2.4) is particularly useful: without it, it would be necessary to compute the marginal density where the Pólya-gamma random variables are integrated out. However, the marginalization with respect to the Pólya-gamma density does not lead to a closed form expression; thus, the Metropolis-Hastings algorithm cannot be defined.

Clearly, for increasing r_i the proposal density (2.5) is closer to the true full conditional distribution; hence, the related acceptance rate will be higher, and the Metropolis-Hastings algorithm will be similar to a Gibbs sampler. On the other hand, setting this parameter to get a lower acceptance rate can result in smaller autocorrelation, and hence a better mixing (Robert and Casella, 2010). We discuss an approach to choose r_i balancing these two extremes in Section 2.1.4.

2.1.3. **Adaptive importance sampler**

The sampling mechanism (2.5) can also be exploited within the context of importance sampling, where the posterior expectation of a function of the parameter β, $\mathbb{E}\left(h(\beta)\right) = \int h(\beta)\,\pi(\beta \mid y)d\beta$ is evaluated via Monte Carlo integration without direct sampling from $\pi(\beta \mid y)$. To this end, the general approach is to define an importance density $q(\beta)$ that is used to sample values $\beta^{(1)}, \ldots, \beta^{(T)}$, which are eventually averaged to obtain an approximation of $\mathbb{E}\left(h(\beta)\right)$ through

$$\widehat{\mathbb{E}(h(\beta))} = \frac{1}{T}\sum_{t=1}^{T} \tilde{w}(\beta^{(t)})h(\beta^{(t)}),$$

with weights

$$\tilde{w}(\beta^{(t)}) = \frac{\pi(\beta^{(t)} \mid y)}{q(\beta^{(t)})}.$$

The efficiency of this algorithm is determined by the ability of the importance density to sample values relevant to the target density. To improve this aspect, we modify the original algorithm and, at each iteration, we simulate values from (2.5), updating the importance density with (2.4). Thus, the importance density is adaptively updated based on the previously extracted value $\beta^{(t-1)}$ and the weights become

$$\tilde{w}(\beta^{(t)}) = \frac{\pi(\beta^{(t)} \mid y)}{q(\beta^{(t)} \mid \beta^{(t-1)})}.$$

2.1.4. **Tuning parameters** r_i

The values of the parameters r_i, $i = 1, \ldots, n$, have to be tuned to balance the trade-off between acceptance rate and autocorrelation in the Metropolis-Hastings, and to control the mixing of the weights in the importance sampler. However, tuning n parameters is not practical, especially for moderate to large n. The first simple solution sets all parameters equal to a single value r, however, in our experience, this resulted in a low effective sample size for some of the sampled chains.

As an alternative strategy, we choose to tune instead the distance of the proposal density from the target posterior. As the expression of the posterior distribution is unknown, we control the distance between the Poisson and negative binomial likelihood. Based on Teerapabolarn (2012), we consider the upper bound of the relative error between the Poisson and negative binomial cumulative distribution functions. This result is particularly useful owing to its simplicity, which allows to analytically derive adequate parameters to bound the error to a specific value. Specifically, if Y is a Poisson random variable with mean λ_i and V is a negative binomial random variable with parameters r_i and p_i, as defined in Section 2.1.1, we have the following result:

$$\sup_{y_i \geq 0} \left| \frac{\Pr(Y \leq y_i)}{\Pr(V \leq y_i)} - 1 \right| = e^{-\lambda_i} p_i^{-r_i} - 1.$$

Hence, by setting an upper bound d for the distance between the Poisson and negative binomial distribution, all the values of the parameters r_i can be automatically derived

to obtain a proposal density whose distance from the target posterior is constant for every y_i, even for heterogeneous data. Under our notation $p_i = \lambda_i/(r_i + \lambda_i)$, thus $d = e^{-\lambda_i}(1 + r_i/\lambda_i)^{r_i} - 1$, which is solved by

$$r_i = -\lambda \log c \cdot \left\{ \log c + \lambda \cdot \mathrm{W}\left(\frac{-c^{-1/\lambda} \log c}{\lambda} \right) \right\}^{-1}, \tag{2.7}$$

where $c = e^{\lambda}(d^2 + 1)$ and $\mathrm{W}(\cdot)$ is the Lambert-W function (Lambert, 1758), which can be computed numerically using standard libraries. Hence, in the algorithm, at the beginning of each iteration, the values $r_1, \ldots, r_n$ are computed according to (2.7) conditionally on the current value of β.

2.2. Numerical illustrations

2.2.1. **Synthetic data**

We conducted a simulation study under various settings to compare the efficiency of the proposed Metropolis-Hastings and importance sampler with that of state-of-the-art methods. We focused on the Hamiltonian Monte Carlo approach—as implemented in the Stan software (Stan Development Team, 2021)—as the successful Metropolis-Hastings with standard random walk proposal would require, different from the proposed approaches, the tuning of p parameters, which becomes cumbersome for moderate to elevate p. The proposed methods are implemented via the R package `bpr`, which is written in efficient C++ language exploiting the `Rcpp` package (Eddelbuettel and Francois, 2011) and available from the Comprehensive R Archive Network (D'Angelo, 2021) and at `github.com/laura-dangelo/bpr`.

Data were generated from a Poisson log-linear model with sample sizes $n \in \{25, 50, 100, 200\}$ and number of covariates $p \in \{5, 10, 20\}$. Specifically, for each combination of n and p, we consider 50 independent n dimensional vectors of counts where each y_i $(i = 1 \ldots, n)$ is sampled from a Poisson distribution with mean $\lambda_i = e^{x_i^{\mathrm{T}}\beta}$, with common parameter β. The covariates were generated from continuous or discrete/categorical random variables under the constraints that the continuous variables have mean zero and variance one and that $1 \leq \lambda_i \leq 200$.

Two prior distributions for the coefficients β were assumed, namely a vanilla Gaussian prior with independent components $\beta_j \sim \mathrm{N}(0, 2)$, $j = 1, \ldots, p$, and the more complex horseshoe prior (Carvalho et al., 2010) which allows for the following conditionally Gaussian representation

$$\begin{aligned} \beta_j \mid \eta_j^2, \tau^2 &\sim \mathrm{N}(0, \eta^2\tau^2) \\ \eta \sim \mathrm{C}^+(0,1), \quad \tau &\sim \mathrm{C}^+(0,1), \end{aligned}$$

for $j = 1, \ldots, p$, where $\mathrm{C}^+(0, 1)$ is the standard half-Cauchy distribution. To implement the samplers under the horseshoe prior, we used the details of Makalic and Schmidt (2016), and fixed the global shrinkage parameter τ to the "optimal value" $\tau_n(p_n) = (p_n/n)\sqrt{\log(n/p_n)}$, where p_n is the number of non-zero parameters (van der Pas et al., 2017).

Each method introduced in Section 2.1 was run for 10000 iterations with 5000 of them discarded as burn-in. The convergence of each algorithm was assessed by graphical inspection of the trace plots of the resulting chains. The convergence was satisfactory for all simulations and comparable for all algorithms, as no systematic bias was found in the posterior mean of the estimated parameters.

To assess the efficiency of the proposed methods, we used a proxy of the time per independent sample, which is estimated as the total time (in seconds) necessary to simulate the entire chain, over the effective sample size of the resulting chain. For the proposed adaptive importance sampler, an estimate of the effective sample size was obtained using the quantity $\sum_{t=1}^{T} w(\beta^{(t)})^2/(\sum_{t=1}^{T} w(\beta^{(t)}))^2$, which takes values between 1 and n (Robert and Casella, 2010). Notably, the burn-in samples were removed from the chains before computing the effective sample size. Thus, the obtained times per independent sample do not represent exactly the number of seconds necessary to generate one independent sample—they rather represent an overestimate. Nonetheless, this approach provides a robust and fair comparison between the different competing algorithms. The experiment has been run on a Linux machine with 8 GB DDR4 2400 MHz RAM, CPU Intel i7-7700HQ 3.8 GHz, running R 4.1.1.

Figure 2.1 and 2.2 show, for each combination of n and p, the distribution of the median time per independent sample for the three algorithms computed on the 50 replications under a Gaussian and horseshoe priors, respectively. The plots are presented in the logarithmic scale for clarity.

For the Gaussian prior the performances of the proposed algorithms are better than those obtained using the HMC implemented in Stan, for small values of the dimension p. For $p = 20$, instead, the performances of the HMC are quite competitive with respect to the importance sampling and broadly comparable to the proposed efficient Metropolis-Hastings algorithm. Notably, the differences are less evident with increasing sample size.

For the horseshoe prior, the proposed Metropolis-Hastings presents a stable superior performance with respect to the HMC sampler implemented in Stan for each sample size n ad number of covariates p. The performance of the importance sampler remains competitive. As previously observed for the Gaussian prior, the differences are less evident for increasing sample size.

2.2.2. **Spike train data**

Herein, we illustrate the proposed sampling method on data of brain activity in mice in response to visual stimulation. The data set was generated using a small subset of data from the Allen Brain Observatory (Allen Institute MindScope Program, 2016), described in detail in Section 1.2.1. In the original data set, for each neuron the fluorescent calcium traces are recorded, which is a proxy of the neuronal activity, under different experimental conditions. From these traces, it is of interest to detect and analyze the activations of neurons, which correspond to transient spikes of the intracellular calcium level. We applied the method reported by Jewell et al. (2019) as described in de Vries et al. (2020) to extract and count the activations of each neuron, to understand how they are affected by the experimental conditions and the location of the neurons in the brain.

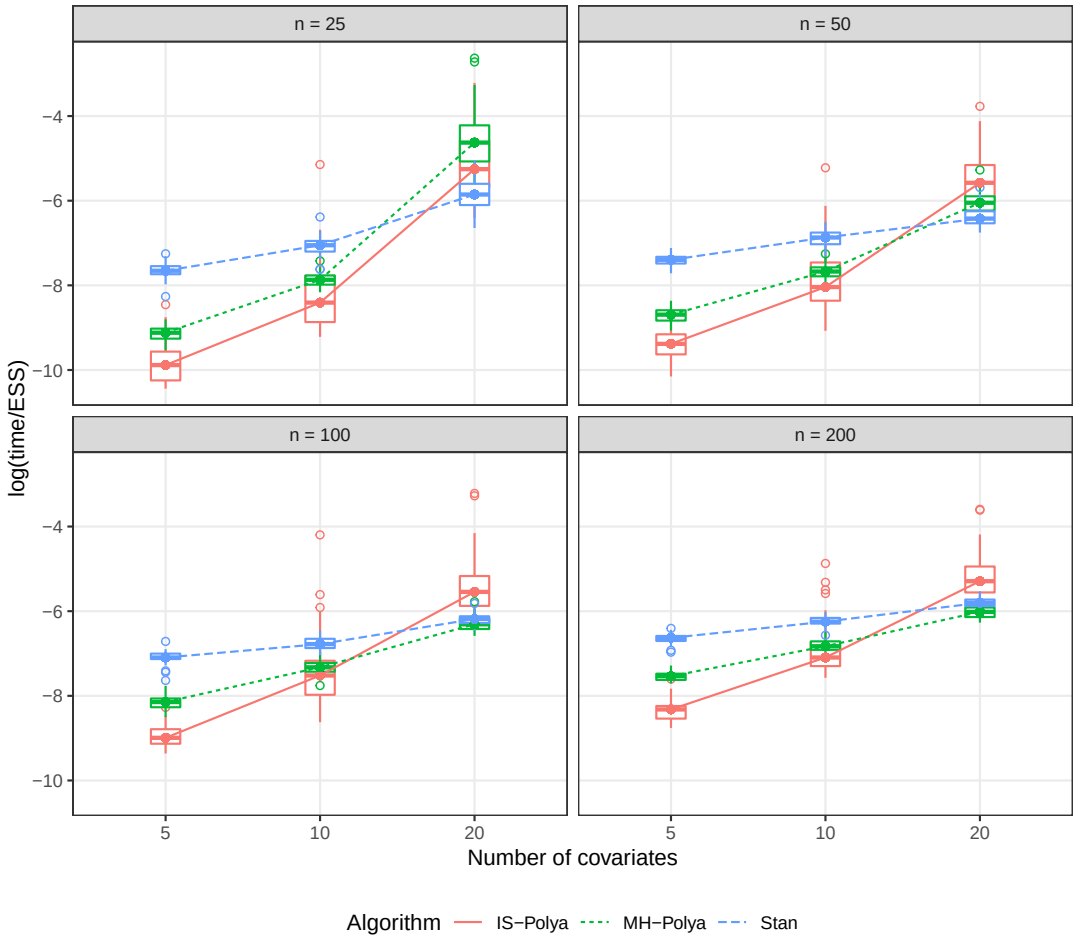

FIGURE 2.1: Time per independent sample (in logarithmic scale) for the three algorithms. For each combination of n and p the boxplots represent the distribution of the (log) time (in seconds) over the effective sample size using a Gaussian prior, over 50 replications.

The covariates that we considered are the depth of the neuron, the area of the visual cortex where the neuron is located (factor with 6 levels), the cre transgenic mouse line (factor with 13 levels), and the type of visual stimulation (factor with 4 levels). The depth of the neurons is discretized to 22 levels, ranging from 175 to 625 microns, thus, we could obtain a data set having a full factorial design with 5 replications for each available covariate combination. Moreover, we included a quadratic term of the depth to improve the fitting. The obtained data set is made of 920 observations on 23 variables.

We ran the proposed Metropolis-Hastings algorithm for 9000 iterations, discarding the first 5000 as burn-in. The computation time was 98 seconds. The posterior estimates of the coefficients of the dummies on three categorical variables are shown in Figure 2.3; and for the numeric covariate depth, the posterior mean and 95% credible intervals were equal to -2.72×10^{-3} $(-2.90 \times 10^{-3}, -2.42 \times 10^{-3})$ for the linear term, and 5.59×10^{-6} $(5.11 \times 10^{-6}, 6.08 \times 10^{-6})$ for the quadratic term. Given these estimates of

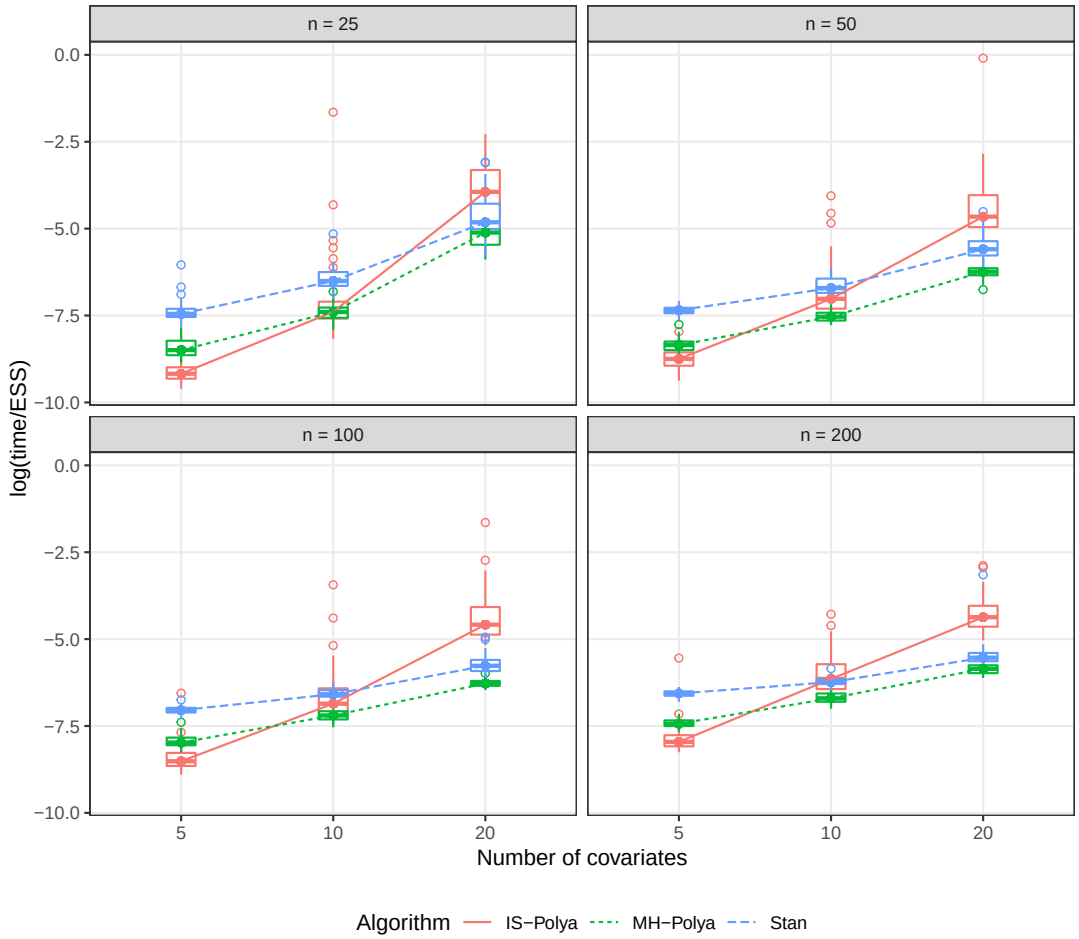

FIGURE 2.2: Time per independent sample (in logarithmic scale) for the three algorithms. For each combination of n and p the boxplots represent the distribution of the (log) time (in seconds) over the effective sample size using the horseshoe prior, over 50 replications.

the coefficients, the number of spikes increased with the largest depths. Moreover, as shown in Figure 2.3, the response of neurons is heterogeneous across the cre-lines and, coherent with the results of de Vries et al. (2020), we obtained that the mean response is lower for the VISam, VISpm and VISrl areas.

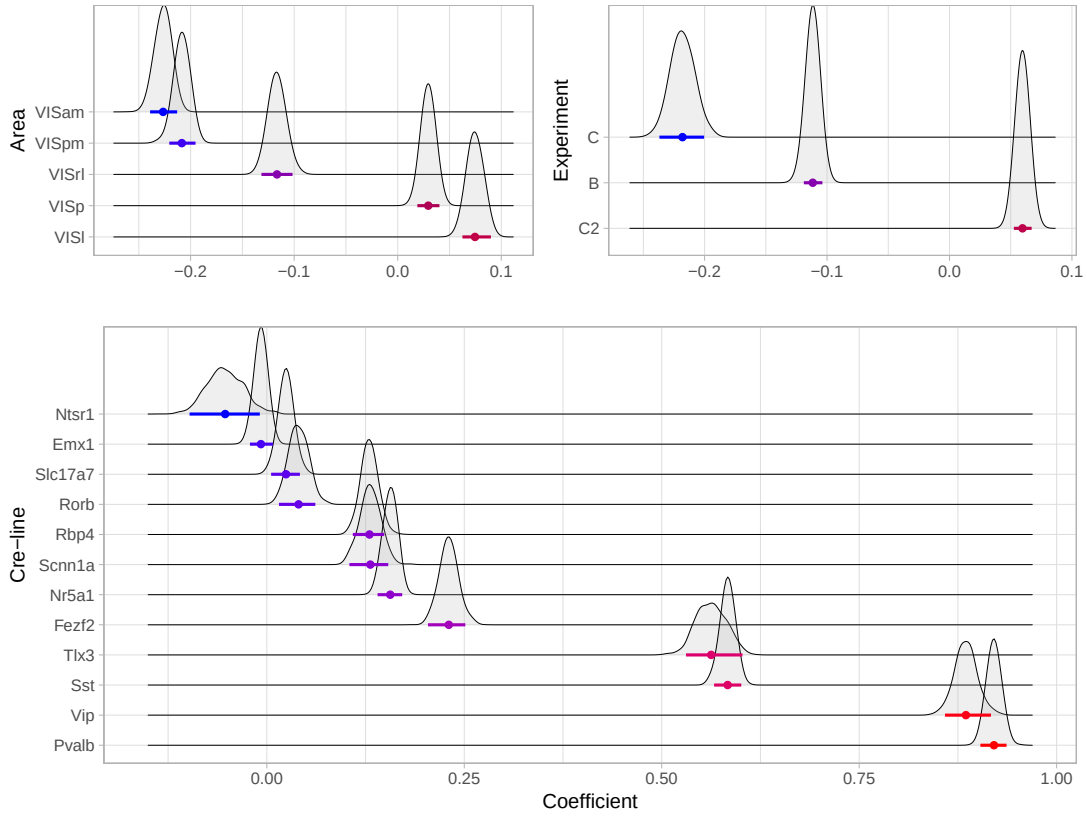

FIGURE 2.3: Estimated coefficients of the regression on the calcium imaging data set: posterior density, with the posterior mean and 95% credible interval (colored dot and segment).

3 MODELING SINGLE-NEURON ACTIVATIONS VIA NONPARAMETRIC MIXTURES

As discussed in Section 1.1.2, routine methods to analyze calcium imaging data are based on a two-step approach. However, it is expected the rate and the distribution of spikes to be stimulus-dependent (Brenner et al., 2002), but none of the previously described approaches allows taking into account explicitly the heterogeneity of spikes' behaviors as a function of the stimulus. As Figure 1.1 clearly shows for the Allen Brain Observatory data, the spikes' intensities vary greatly according to the type of stimulus.

In this chapter, we introduce a coherent nested Bayesian finite mixture model that allows for the estimation of the spiking activity of each neuron – which could be seen as a first step for the analysis of larger brain activity combining multiple neurons in a region. In addition, our model *simultaneously* allows for reconstructing the distributions of spikes under various experimental conditions; for example, in response to different types of visual stimuli in the Allen Brain Observatory data set.

More specifically, our modeling framework estimates and clusters the distributions of the calcium transient spikes' amplitudes via a nested formulation of the generalized mixture of finite mixtures (gMFM) prior recently proposed by Frühwirth-Schnatter et al. (2021). The proposed model further adopts the use of a common atom specification as in Denti et al. (2021) for estimating the distribution of the spikes' amplitudes under each experimental condition. The proposed common atom gMFM has several advantages with respect to typical Bayesian nonparametric models for nested data. With respect to models based on Dirichlet process priors, the gMFM provides increased flexibility to estimate partitions characterized either by many, well-balanced, clusters or by a small set of large clusters. The common atom model allows us to obtain nested inference on densities without incurring in the degeneracy issues pointed out by Camerlenghi et al. (2019) for the widely used nested Dirichlet process of Rodríguez et al. (2008). At the same time, the common atom formulation still leverages two nested layers of random discrete mixture priors to borrow information between experiments and to identify similarities in the distributional patterns of the neuronal responses to different stimuli. In addition, differently than in the nested Dirichlet process, the common atom model also allows clustering the inferred spikes' intensity values both within and between experimental conditions, so to infer common (recurring) response amplitudes. Finally, we allow our model to enforce sparsity of neuron firing over time by assuming a spike-and-slab prior specification on the marginal distribution of the amplitudes.

3.1. Model and prior specification

We consider the biophysical model for the calcium dynamics (1.1) introduced in Section 1.1.1, and the interpretation of the A_t parameters as *amplitude* of a spike at time t, taking value 0 if there is no spike and a positive value otherwise.

We are interested in characterizing the neuronal activity under different experimental conditions. For each time point $t = 1, \dots, T$, let g_t be a discrete categorical variable, taking values in $\{1, \dots, J\}$, where J is the number of distinct experimental settings, so that $g_t = j$ indicates that the neuronal activity at time t is observed under condition j. The experimental conditions are often designed to capture variations in neuronal activity with respect to a baseline process, which may represent a "typical" brain process. For example, in the Allen Observatory data, the interest is to investigate visually-evoked functional responses of neurons in the mouse's visual cortex. Therefore, neurons associated with visual decoding should be expected to activate in all conditions. It is then of interest to study not only *if* but also *how* the neurons differentially respond to the presentation of a variety of visual stimuli.

In this chapter, we propose a hierarchical Bayesian approach to investigate similarities and differences in the distribution of spikes over time and conditions. In order to borrow information across different experimental conditions, one option is to fit a parametric hierarchical random effect model, and obtain a post-MCMC clustering of the estimated spikes A_t by grouping together those spikes with similar magnitudes. This approach has several limitations: on the one hand, the distribution of the random effects is constrained into a specific parametric form; on the other hand, the clustering of, say, the posterior mean estimates of the parameters A_t's does not allow the model to fully describe stimulus-specific distributional differences and to take into account the posterior uncertainty in the spikes.

In order to allow flexible modeling of distributions and to describe the heterogeneity of distributional features, we assume a nested Bayesian finite mixture specification. More specifically, we rewrite (1.1) as

$$y_t \mid b, \gamma, \mathrm{Ca}_{t-1}, A_t, \sigma^2, \tau^2 \sim \mathrm{N}(b + \gamma\, \mathrm{Ca}_{t-1} + A_t, \sigma^2 + \tau^2)$$

and we assume that the spikes A_t are from stimulus-specific distributions, i.e. $(A_t \mid g_t = j,\, G_j) \sim G_j$, $j = 1, \dots, J$, to account for the observed variety of neuronal activity under different experiment settings. We further allow for clustering the distributions across conditions, in order to capture similar patterns of neuronal activity. Indeed, one may typically expect $K < J$ distributional clusters. For example, a neuron may respond to general visual stimulation and not specifically to the type of stimulus considered. More specifically, we assume the following generalized mixture of finite mixtures structure:

$$G_1, \dots, G_J \mid Q \sim Q, \qquad Q = \sum_{k=1}^{K} \pi_k \delta_{G_k^*} \tag{3.1}$$

where $\pi_1, \dots, \pi_K \mid K \sim \mathrm{Dirichlet}_K(\alpha/K, \dots \alpha/K)$, $\alpha > 0$, and $G_1^*, \dots, G_K^*$ are a set of cluster-defining distributions, obtained as realizations of an underlying random probability measure, specified further below. Equation (3.1) implies that the G_j's,

$j = 1, \ldots, J$ have a positive probability of clustering together, thereby giving rise to *distributional clusters*. In practice, the number of mixture components, K, is typically larger than the number of clusters, K_+, and some of the atoms G_k^* are not assigned to any of the G_j's (empty components). The prior on the number of mixture components K is a translated beta-negative-binomial distribution as in Frühwirth-Schnatter et al. (2021). Including a prior $p(K)$ leads to both K_+ and K being random a priori. Finally, the distributional atoms G_k^*, $k = 1, \ldots, K$ are also obtained as a realization from an underlying generalized mixture of finite mixtures,

$$G_k^* = \sum_{l=1}^{L} \omega_{l,k} \delta_{A_l^*} \tag{3.2}$$

with $\omega_{1,k}, \ldots, \omega_{L,k} \mid L \sim \text{Dirichlet}_L(\beta/L)$, for some positive real number $\beta > 0$. The set of atoms A_l^* is common across all distributions $G_1^*, \ldots, G_K^*$ and they are obtained as i.i.d. draws from a centering measure, $A_l^* \sim G_0(A_l^*)$. Therefore, equation (3.2) defines a clustering of the inferred spike intensities both within a given condition (i.e. for fixed G_k^*) and across conditions (i.e. across the G_k^*'s; hence, across the G_j's). In the following, we adopt common terminology in the literature on nested Bayesian nonparametric priors and indicate the clustering induced on the A_t through the proposed two-layers prior as *observational clustering*. The nested gMFM formulation requires the specification of a prior on the number of components that specify the lower-level distributional atoms G_k^*, $L \sim p(L)$. Once again, some of the components may be empty.

We enforce sparsity in the detection of the spikes by modeling the base measure G_0 for the parameters A_l^* with a spike-and-slab specification (Mitchell and Beauchamp, 1988), which is a convex mixture between a Dirac mass at zero – representing the absence of neuronal response – and a diffuse density on the positive real numbers – representing the intensity of the neuronal response. More specifically, we assume

$$G_0 = (1 - p)\, \delta_0 + p\, \text{Gamma}\,(h_{A1}, h_{A2}), \tag{3.3}$$

where the slab is a gamma distribution, $\text{Gamma}(a, b)$ with mean a/b and variance a/b^2. The choice of a gamma distribution in (3.3) is particularly relevant for sparsity-inducing purposes, as the gamma density belongs to the set of moment non-local prior densities, as defined by Johnson and Rossell (2010). Therefore, a negligible probability density is assigned to values in a neighborhood of zero, thus inducing a clear separation between the baseline neuronal activity and the neuronal responses. In particular, the higher the shape parameter h_{A1}, the larger is the separation. We assume a $\text{Beta}(h_{1p}, h_{2p})$ prior for the proportion of spikes p with h_{1p} much smaller than h_{2p} in order to favor sparsity of detections.

The proposed formulation can be seen as a special case of *inner* spike-and-slab nonparametric priors, following a terminology introduced by Canale et al. (2017) and Canale et al. (2022). In the following, we will refer to the proposed specification as a finite common atom model (fCAM).

The Bayesian model elicitation is completed by assuming conjugate priors for the underlying calcium level concentration parameters, i.e. the baseline calcium level b, and the variances σ^2 and τ^2. Specifically, the following conjugate prior distributions

are assumed:

$$\begin{aligned} \mathrm{Ca}_0 \sim \mathrm{N}(0, C_0), \quad b \sim \mathrm{N}(b_0, B_0) \\ 1/\sigma^2 \sim \mathrm{Gamma}(h_{1\sigma}, h_{2\sigma}), \quad 1/\tau^2 \sim \mathrm{Gamma}(h_{1\tau}, h_{2\tau}), \end{aligned} \tag{3.4}$$

Finally, under the assumption that the process is stationary with positive correlation between the calcium level at consecutive times, we constrain $\gamma \in (0, 1)$ and let $\gamma \sim \mathrm{Beta}(h_{1\gamma}, h_{2\gamma})$, a priori.

3.2. Posterior inference

For computational purposes, it is convenient to rewrite the likelihood for an observation y_t under condition $g_t = j$ by introducing two latent cluster allocation variables, $c_j^D = c_{g_t}^D$ and c_t, indicating the distributional cluster for the group j and the observational cluster for y_t, respectively.

Given K and $\{\pi_k\}_{k=1}^K$, the distributional allocation variable $c_j^D \in \{1, \ldots, K\}$, with $\Pr(c_j^D = k) = \pi_k$. Similarly, conditionally on $c_{g_t}^D = k$, and given L and $\{\omega_{l,k}\}_{l=1}^L$, the observational allocation variable $c_t \in \{1, \ldots, L\}$, with $\Pr(c_t = l) = \omega_{l,k}$. Therefore, conditionally on the other model parameters, the joint distribution of the observed data and the latent cluster allocations can be written as

$$f(\boldsymbol{y}, \boldsymbol{c}, \boldsymbol{c}^D \mid \boldsymbol{\pi}, \boldsymbol{\omega}, \boldsymbol{A}^*) = \prod_{j=1}^{J} \pi_{c_j^D} \prod_{t:g_t=j} \omega_{c_t, c_j^D}\, p(y_t \mid A_{c_t}^*),$$

which facilitates posterior inference.

More specifically, posterior inference for the proposed fCAM can be carried out quite straightforwardly by means of Markov chain Monte Carlo (MCMC) techniques. The sampling of the latent calcium level Ca_t uses an iterative approach based on the Kalman filter and on a forward filtering backward sampling algorithm (Prado and West, 2010). Full conditional posteriors for b, p, σ^2 and τ^2 are available in closed form thus leading to straightforward Gibbs sampling steps. For the autoregressive parameter γ, we use a Metropolis-Hastings within the Gibbs step. The sampling of A_t exploits a combination of the nested slice sampler of Denti et al. (2021) and of the telescoping sampler of Frühwirth-Schnatter et al. (2021). A detailed description of the latter step is reported in Algorithm 1 below. Here, we just present a schematic description of the MCMC steps:

1) Sample the calcium level Ca_t, for $t = 0, \ldots, T$, using a forward filtering backward sampling:
   a) Run Kalman filter: set $a_0 = m_0 = 0$, $R_0 = C_0 = \mathrm{var}(\mathrm{Ca}_0)$. For $t = 1, \ldots, T$ let
   $$a_t = \gamma\, m_{t-1} + A_t$$
   $$R_t = \gamma^2\, C_{t-1} + \tau^2.$$
   Compute the filtering distribution's parameters, m_t and C_t, for $t = 1, \ldots, T$, where
   $$m_t = a_t + R_t\, (R_t + \sigma^2)^{-1}\, (y_t - b - a_t)$$

$$C_t = R_t - R_t^2 \, (R_t + \sigma^2)^{-1}.$$

b) Draw $\mathrm{Ca}_T \sim \mathrm{N}(m_T, C_T)$;

c) For $t = T-1, \ldots, 0$, draw $\mathrm{Ca}_t \sim \mathrm{N}(h_t, H_t)$, with

$$h_t = m_t + \gamma \, C_t \, R_{t+1}^{-1} (\mathrm{Ca}_{t+1} - a_{t+1})$$

$$H_t = C_t - \gamma^2 \, C_t^2 \, R_{t+1}^{-1}.$$

2) Sample a new value for the baseline level b:

$$b \sim \mathrm{N}\left(\frac{b_0}{B_0} + \frac{1}{\sigma^2} \sum_{t=1}^{T} (y_t - \mathrm{Ca}_t), \sqrt{\frac{1}{B_0} + \frac{T}{\sigma^2}} \right).$$

3) Sample the variance on the output equation σ^2 and the variance on the state equation τ^2:

$$1/\sigma^2 \sim \mathrm{Gamma}\left(h_{1\sigma} + \frac{T}{2}, \, h_{2\sigma} + \frac{1}{2} \sum_{t=1}^{T} (y_t - \mathrm{Ca}_t - b)^2 \right)$$

$$1/\tau^2 \sim \mathrm{Gamma}\left(h_{1\tau} + \frac{T}{2}, \, h_{2\tau} + \frac{1}{2} \sum_{t=1}^{T} (\mathrm{Ca}_t - \gamma \, \mathrm{Ca}_{t-1} - A_t)^2 \right).$$

4) Update the autoregressive parameter γ using a Metropolis-Hastings step.

5) Update the parameter p of the spike-and-slab base measure from

$$p \sim \mathrm{Beta}(h_{1p} + T - n_0, \, h_{2p} + n_0),$$

where n_0 is the number of y_t assigned to the the spike component.

6) Update the cluster allocations variables $\boldsymbol{c}^D$ and $\boldsymbol{c}$, the number of mixture components K and L, and the cluster parameters $\boldsymbol{A}^*$ using the nested telescoping sampling for the finite common atom model reported in Algorithm 1.

3.3. Simulation study

The performances of the proposed method are assessed through a simulation study. The purpose of this section is twofold, namely to assess both the ability to correctly identify the spike times, and the accuracy of the inferred clustering structure.

We simulated synthetic data exhibiting a baseline level and a number of spikes representing the effect of the response of a neuron to a stimulus, thus mimicking the characteristics of real series of calcium imaging following the structure of model (1.1). Specifically, we first divided the time frame into J hypothetical experimental conditions of equal length, with J varying in the different scenarios described below. Consistent with our motivating assumption that the neuronal response depends on the type of

Algorithm 1 Nested telescoping sampling

Denote with $\mathcal{C}^D$ the current partition on the distributions and with $\mathcal{C}^O$ the partition on the observations.

1: Sample the weights on the distributions:

$$(\pi_1, \dots, \pi_K) \mid K, \alpha, \mathcal{C}^D \sim \text{Dirichlet}(e_1, \dots, e_K);$$

where $e_k = \alpha/K + J_k$, and J_k is the number of groups assigned to distribution k.

2: Sample the weights on the observations: for all $k \in \{1, \dots, K\}$ sample a vector $\boldsymbol{\omega}_k$ from

$$(\omega_{1,k}, \dots, \omega_{L,k}) \mid L, \beta, \mathcal{C}^O, \mathcal{C}^D \sim \text{Dirichlet}(f_{1,k}, \dots, f_{L,k});$$

where $f_{l,k} = \beta/L + N_{l,k}$, and $N_{l,k}$ is the number of observations in the observational cluster l and distributional cluster k.

3: Update the partition on the distributions $\mathcal{C}^D$ by sampling from the posterior distribution of the latent cluster allocation variables $\boldsymbol{c}^D$. For $j = 1, \dots, J$

$$\Pr(c_j^D = k \mid \boldsymbol{\pi}, K, \boldsymbol{A}^*, \boldsymbol{y}, \boldsymbol{g}) \propto \pi_k \prod_{t:g_t=j} \omega_{c_t, c_j^D}\, p(y_t \mid A^*_{c_t}),$$

with $k \in \{1, \dots, K\}$. Determine $J_k = \#\{j : c_j^D = k\}$, for $k = 1 \dots, K$, and the number of non-empty components $K_+ = \sum_{k=1}^K I\{J_k > 0\}$. Relabel the components so that the first K_+ are non-empty.

4: Update the partition on the observations $\mathcal{C}^O$ by sampling from the posterior distribution of the latent cluster allocation variables $\boldsymbol{c}$. For $t = 1, \dots, T$

$$\Pr(c_t = l \mid c_{g_t}^D = k, \boldsymbol{c}, \boldsymbol{\omega}, L, K, \boldsymbol{A}^*, \boldsymbol{y}, \boldsymbol{g}) \propto \omega_{l,k}\, p(y_t \mid A^*_{c_t}),$$

with $l \in \{1, \dots, L\}$, $k \in \{1, \dots, K\}$. Determine $N_l = \#\{t : c_t = l\}$, for $l = 1 \dots, L$, and the number of non-empty components $L_+ = \sum_{l=1}^L I\{N_l > 0\}$. Relabel the components so that the first L_+ are non-empty. Because all the mixtures share the same atoms, the cluster parameters are sorted regardless of the distributional cluster allocation.

5: Sample the cluster parameters for the non-empty components: $p(A_l^* \mid -) \propto p(A_l^*) \prod_{t:c_t=l} p(y_t \mid A_l^*)$.

6: Conditional on $\mathcal{C}^D$, sample the number of components K of the mixture on distributions.

7: Conditional on $\mathcal{C}^O$, sample the number of components L of the mixtures on observations. If $L > L_+$, sample a new parameter A^* for the empty components from the prior distribution.

8: Update the hyperparameter α on the Dirichlet distribution on the mixture weights on distributions.

9: Update the hyperparameter β on the Dirichlet distribution on the mixture weights on observations.

The posterior distributions for steps 6-9 are given in Frühwirth-Schnatter et al. (2021).

stimulus, each experimental condition is assumed to belong to one of the K distributional clusters. Then, for each experimental condition, we generated the neuronal activity: first, we generated the presence or absence of a neuron response uniformly in time, where the spike probability can vary across groups. Then, conditionally on the obtained activations, we generated some additional spikes in a short subsequent interval, so that it is very likely to observe close or even successive spikes. In this way, the data mimic a real calcium imaging time series. Moreover, we are able to conduct a careful assessment of the ability of the model to distinguish the presence of a single high spike versus the convolution of several spikes in consecutive times. Finally, the values A_t, conditionally on their distributional cluster, are generated from one of the finite sets of spike amplitudes described below.

We simulated 50 independent data sets for each of the three scenarios described henceforth. In Scenario 1 we assumed $J = 6$ experimental conditions, generated from $K = 4$ distributional clusters. The spike amplitudes in the distributional clusters are $(0.35, 0.89, 1.15, 1.80, 2.20)$, $(0.65, 0.89, 1.40, 1.80)$, $(0.35, 0.65, 1.15)$, and $(0.35, 0.89, 1.60)$. Scenario 2 assumes $J = 4$ experimental conditions and $K = 3$ distributional clusters with spike amplitudes equal to $(0.3, 0.5, 0.7, 0.9, 1.1, 1.5)$, $(0.3, 0.9, 1.5, 1.8)$, and $(0.5, 0.9, 1.5)$. Finally, Scenario 3 sets $J = 5$ and $K = 3$ with the spike amplitudes in the distributional clusters being $(0.3, 0.5, 0.7, 0.9, 1.1)$, $(0.3, 0.9, 1.1, 1.3)$, and $(0.7, 0.9, 1.3)$. While in Scenario 1 the amplitudes of the spikes are quite large, spaced apart, and with the corresponding distributional clusters well distinct, in Scenario 3 the spike amplitudes are more homogeneous and more clustered in time. Scenario 2 represents an in-between situation. Hence, from the first to the last scenario, we are assuming an increasing degree of complexity. The R script generating these synthetic datasets is available at `github.com/laura-dangelo/fCAM_calcium_imaging`.

The results attained by the proposed fCAM are compared to those obtained exploiting the common atom model (CAM) of Denti et al. (2021) – which provides a benchmark for the clustering of the spikes and the stimulus-specific distributions – and to those obtained with the L_0 penalization method of Jewell et al. (2019) and the L_1 penalization method of Friedrich et al. (2017), which provide a benchmark for the task of spikes' detection. For the latter two methods we have assumed complete knowledge of the autoregressive constant controlling the rate of the calcium decay, since we found that the results were quite sensitive to this estimate. To assess the sensitivity of the proposed fCAM to the prior specification, we repeated the numerical experiment for different values of the hyperparameters h_{A1} and h_{A2} in (3.3). In particular, the shape parameter h_{A1} was supposed to play a key role in the detection of spikes. Keeping fixed the ratio h_{A1}/h_{A2}, the parameters were set equal to 3, 4, 6, and 8: a small value implies, *a priori*, less separation between zero and the distribution of the positive spikes while a large value corresponds to the opposite effect.

Focusing on the classification of each time point as a spike or not, Figure 3.1 summarizes the misclassification rate for all competing methods under the three scenarios. The results of the 50 replications are summarized using boxplots. For our fCAM, we report only the results obtained with $h_{A1} = h_{A2} = 8$ as those obtained for the other choices are essentially equivalent. The rates are small in absolute value and broadly comparable across the different methods, thus confirming that all the competing models are effective in detecting the spikes.

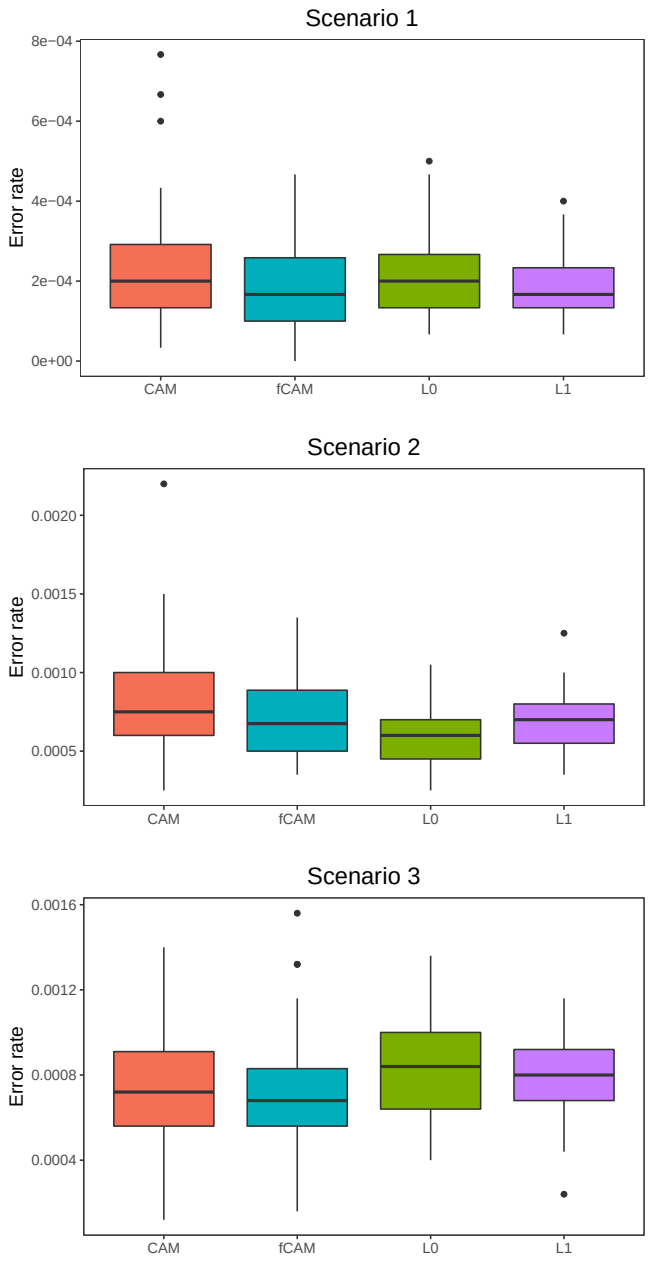

FIGURE 3.1: Distribution of the misclassification error rate in the simulation study for the four considered methods: CAM, fCAM, and the methods of Jewell et al. (2019) "L0" and Friedrich et al. (2017) "L1".

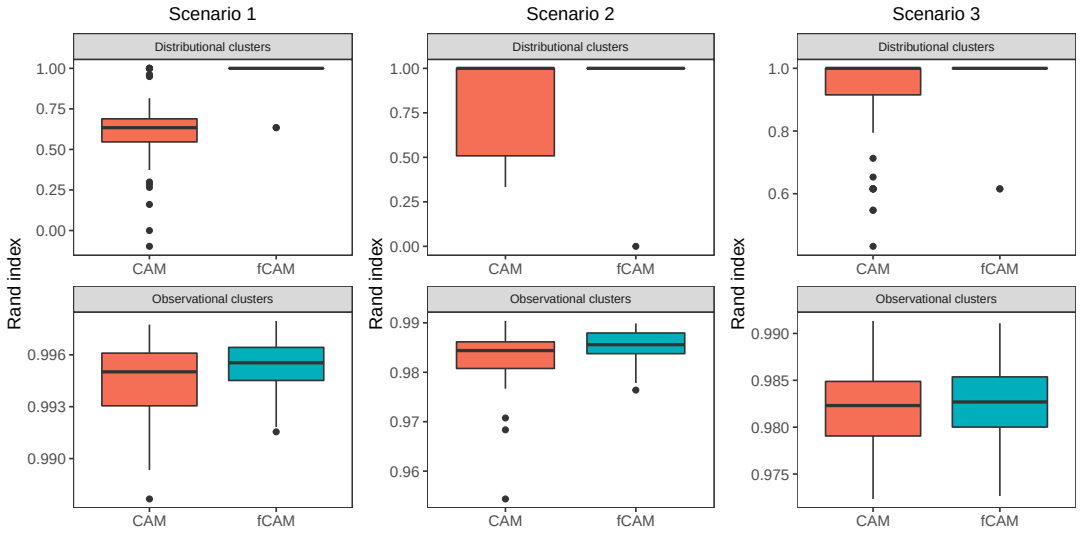

FIGURE 3.2: Distribution of the adjusted Rand index on the distributional and observational clusters of CAM and fCAM, computed on the 50 simulations for the three scenarios of the simulated data.

However, the proposed fCAM not only enables the detection of spikes but also allows us to conduct inference on the clustering structure. Therefore, we report on its ability to identify the clustering structure. Figure 3.2 reports the adjusted Rand index (Rand, 1971; Hubert and Arabie, 1985) computed on both the observational and the distributional clusters for $h_{A1} = h_{A2} = 8$ (results for other settings are similar). Values of the adjusted Rand index close to 1 denote that the identified structure resemble the true clustering. While for the observational clusters the results are broadly comparable, for the distributional clusters, the performance of the proposed fCAM is uniformly superior. In addition, the variability of the results generally appears to be drastically smaller for the fCAM, thus providing evidence of greater efficiency. This is consistent to the results of Frühwirth-Schnatter et al. (2021) where the generalized mixtures of finite mixtures is compared to a standard Dirichlet process mixture model.

From a computational point of view, the proposed algorithm is clearly more demanding than the optimization methods of Jewell et al. (2019) and Friedrich et al. (2017). However, the computing time is comparable to the slice sampler adopted for the CAM, and in general a full run requires just few minutes on a Linux machine with an i7-7700HQ 3.8 GHz Intel processor, 8 GB RAM, running R 4.1.0. For example, for a calcium trace of length 50,000, the computing time of the proposed method is around 2 minutes. Indeed, our experience suggests that the main factor affecting the computing time is the length of the series. In general, in the analysis of spike activity, we expect the number of clusters to be small and – in particular – much smaller than the number of observations.

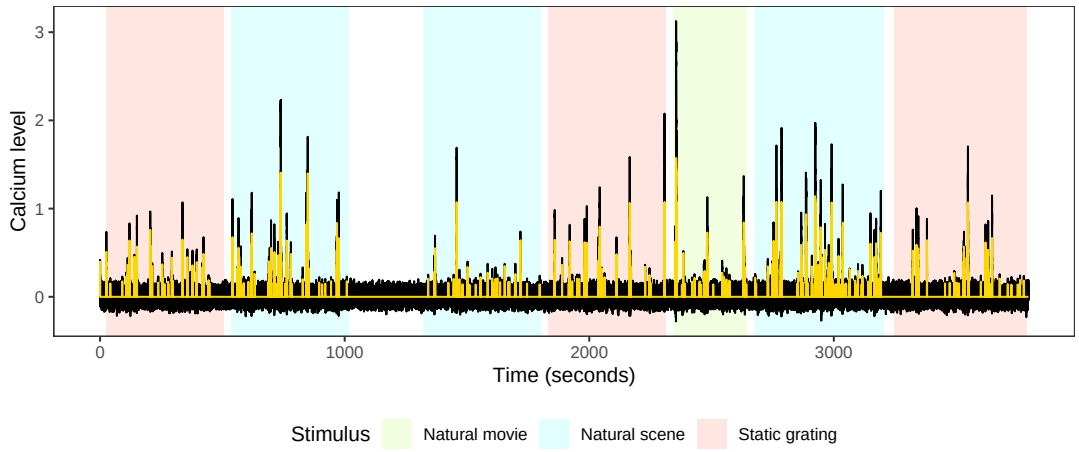

FIGURE 3.3: Observed fluorescence trace y_t of a neuron from the Allen Brain Observatory data (black line), and visual stimulus to which the mouse is exposed (shaded areas). The yellow line represents the estimated neuronal activity.

3.4. ALLEN BRAIN OBSERVATORY DATA ANALYSIS

We now revert to the analysis of the data from the Allen Brain Observatory (Allen Institute MindScope Program, 2016). The data comprise the dF/F-transformed fluorescence trace for a cell during session-B of the experiment (cell id 508596945). This session comprises three types of visual stimuli (static gratings, natural scene and natural movie) in addition to some period of spontaneous activity (absence of visual stimuli). Since the data are recorded at a frequency of 30 Hz, the resulting series consists of 113,865 time points for a total of 63.2 minutes. We focus the first analysis on a neuron located in the primary visual area, at an imaging depth equal to 350 microns. Additional analyses for other neurons are reported later in this section.

The observed fluorescence trace is shown with a continuous black line in Figure 3.3. Different shaded backgrounds indicate the types of visual stimuli. Using the notation introduced in the previous Sections, $J = 4$ with $j = 1, 2, 3$ corresponding to static grating, natural scene, and natural movie, respectively and $j = 4$ indicating no stimulus presence.

We ran the MCMC algorithm of Section 3.2 using the same prior specification of Section 3.1 for 15,000 iterations discarding the first 7,000 iterations as burn-in and keeping one iteration every four to improve mixing. Visual inspection of the traceplots and Geweke diagnostics showed no issues with convergence. The superimposed light line in Figure 3.3 represents the estimated neuronal activity in terms of the inferred amplitude A_t, i.e. removing the measurement errors and the result of the accumulation of calcium from the previous spikes. Here and henceforth, we identified the presence of a spike if the posterior probability of a spike at time t, say PPS_t, estimated by the proportion of non-zero A_t's over all MCMC iterations, was greater than $\kappa = 75.5\%$. This threshold allows us to control the (estimated) Bayesian false discovery rate at the

pre-set value 0.05, that is κ solves the equation

$$\text{FDR}\,(\kappa) = \frac{\sum_{t=1}^{T}(1-\text{PPS}_t)\,I_{(\text{PPS}_t>\kappa)}}{\sum_{t=1}^{T} I_{(\text{PPS}_t>\kappa}\Big)} = 0.05.$$

For more details, we refer to Newton et al. (2004) and Müller et al. (2007). See also Sun et al. (2015) for a discussion with dependent hypotheses.

As already mentioned in the Introduction, in calcium imaging it is of interest studying the distribution of the spikes in response to each experimental stimulus, and identifying similarities and differences in these distributions across stimuli.

We start by investigating the presence of similarities in the neuronal response to different types of visual stimuli. This corresponds to analyzing the clustering of the spike distributions induced by the proposed fCAM. The model clusters together the groups corresponding to the natural scene and natural movie stimuli with high posterior probability, while the static grating stimulus and the absence of stimuli are assigned to two separate distributional clusters. In other terms, the neuron appears to show similar neuronal responses in the natural scene and natural movie stimuli whereas the responses appear distinctly different under the other two conditions.

To understand whether and how the neuronal response depends on the type of stimulus, we estimated the spike amplitude distribution for each of the four types of stimuli. Figure 3.4 shows the histograms of posterior means of the non-zero spike amplitudes for the three types of stimuli. The distribution for the time interval between 1018-1319 sec in Figure 3.3 (absence of stimuli) is not presented because no activity was detected. Despite the apparent similarities of the distributions in Figure 3.4, the second cluster of spike amplitude distributions (natural scene and natural movie) shows a heavier tail. Specifically, the highest observed cluster during the static grating stimulus (top plot) is centered at 1.06, while for the other two stimuli we obtained several higher values, with the largest cluster centered around 1.43.

A qualitative representation of how these spike clusters are distributed within the three groups is given in Figure 3.5. The three plots show a short interval of the observed calcium series, chosen in correspondence of one of the highest observed spikes. Each plot also shows a series of colored vertical lines: the lines are placed at the estimated spike times, and the colors correspond to the estimated spike amplitudes. The represented partition is the posterior point estimate obtained by minimizing the variation of information loss, as proposed in Wade and Ghahramani (n.d.). Conditionally on the obtained partition, for each cluster a representative value for the cluster parameter is obtained as follows: first, for each MCMC iteration, the group-specific average of A_t is computed keeping the partition fixed; then, these values are averaged over all the MCMC iterations. We notice that for all experiments, high values of the observed calcium level are often produced as the result of several consecutive spikes, since, individually, the spikes are characterized by a relatively low amplitude, and the observed calcium level is cumulated due to its autoregressive behavior. The autoregressive parameter γ has a posterior mean equal to 0.493 with a 95% credible interval of $(0.481, 0.505)$. This result corresponds to the understanding that the observed calcium response may be generated by high-frequency firing neurons: due to the low-sampling rate, the non-linear calcium signal essentially captures a super-imposition of multiple spikes (Hoang et al., 2020).

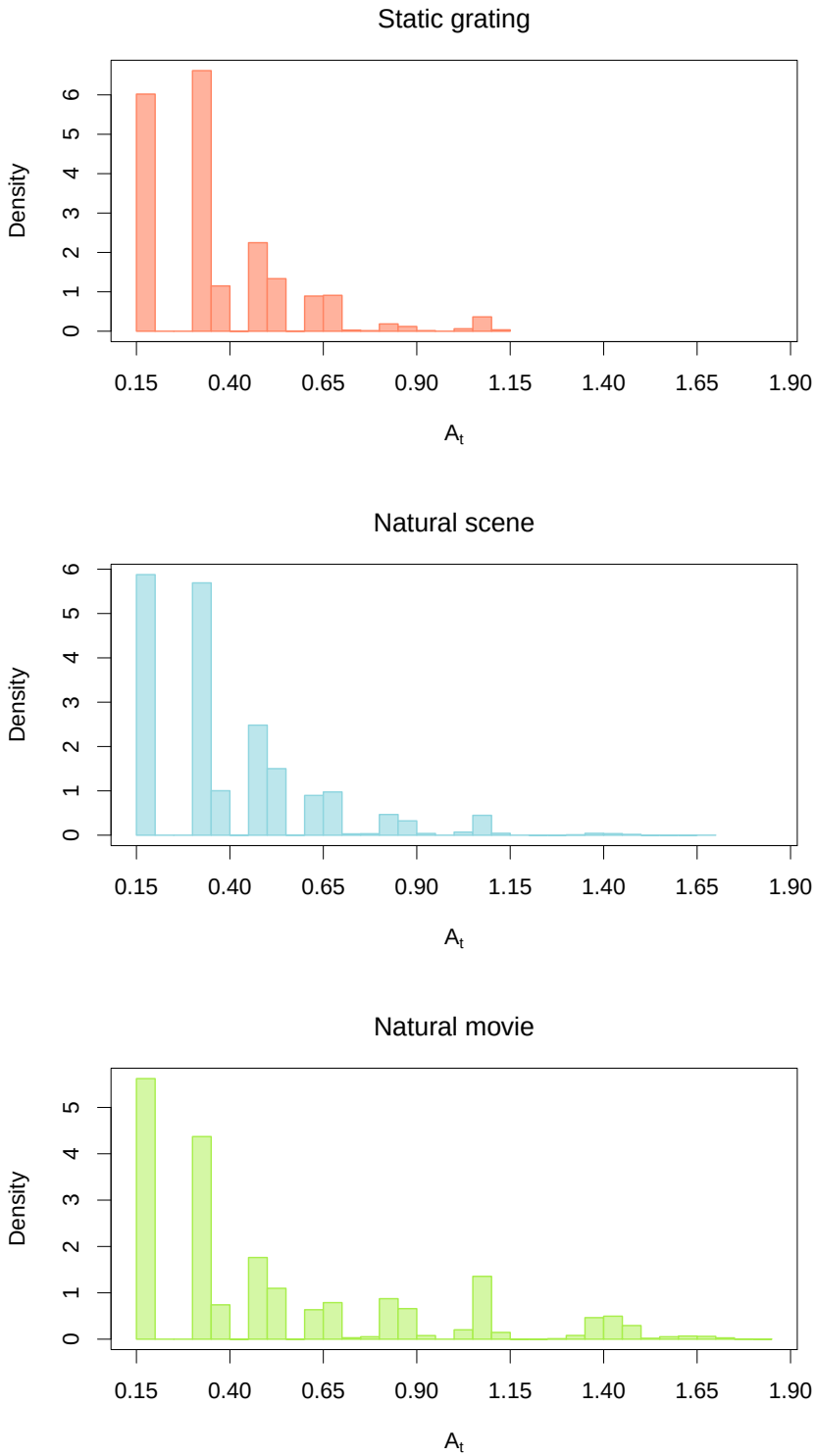

FIGURE 3.4: Empirical distribution of the posterior means of the observational cluster parameters A_t for the three experimental conditions of the Allen Brain Observatory data.

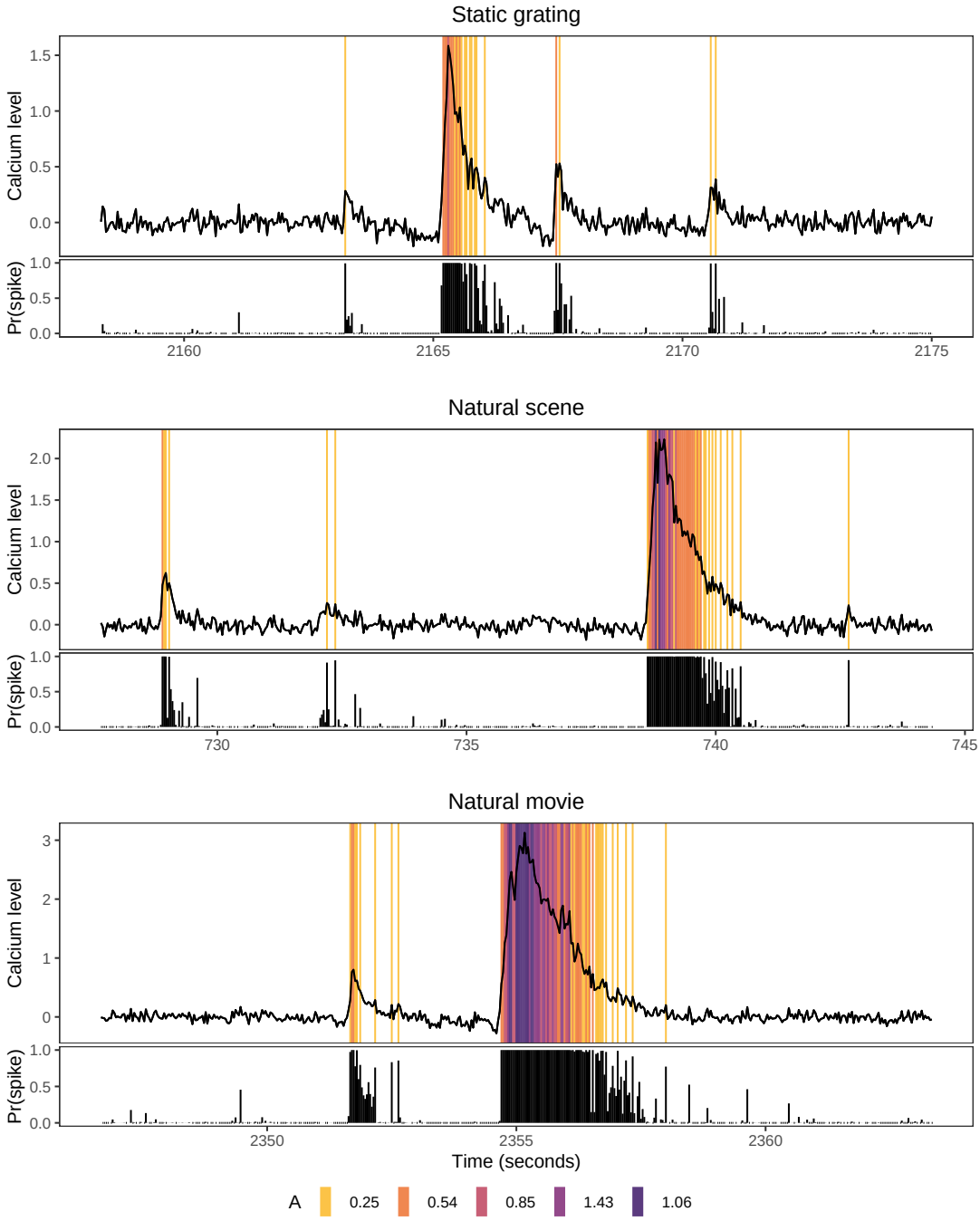

FIGURE 3.5: Short interval of length 500 of the Allen Brain Observatory data in correspondence of a spike, for the three stimuli. The vertical lines indicate the time of a spike and the colors correspond to the observational cluster of its amplitude. The bottom panels show the estimated posterior probability of spike presence, for each time point.

As a matter of fact, another useful quantity we can compute to compare the neuronal activity between stimuli is the firing rate, which provides a measure of how often the neuron has activated during a specific visual stimulus. The rate computes the number of detected spikes per second, to take into account the different duration of the experiments. For the static grating stimulus the posterior mean rate (and related 95% credible interval) is 0.223 $(0.216, 0.229)$, while for the natural scene and natural movie stimuli they are 0.419 $(0.410, 0.428)$ and 0.511 $(0.495, 0.531)$, respectively. These results highlight the role of spike-frequency adaptation, whereby some neurons show an increased activity when exposed to more complex stimuli, thus exhibiting higher firing rates and larger calcium concentration measurements (Peron and Gabbiani, 2009).

3.4.1. **Analysis of additional neurons**

We analyzed two additional neurons: the first was chosen from the same targeted area and depth as the one presented above (primary visual area, 350 microns, cell id 517398389); while the second neuron was chosen from the same area but at a different depth, and, specifically, at a much shallower depth (175 microns, cell id 587435328). Although all neurons are located in the same area, we do not necessarily expect similar results, as neurons do not show spatial dependence in some regions of the brain (Rosenbaum et al., 2017). In the following paragraphs we present the same type of analyses and plots reported in the main analysis.

Figure 3.6 shows the observed fluorescence trace of the second considered neuron ($\#517398389$), together with the estimated activity. Even if the trace appears to be similar to the previously analyzed neuron, we obtained a different clustering of the stimuli. Here the static grating and natural movie are pooled together with high posterior probability, while the natural scene is assigned to a separate cluster. Similarly to the previous analysis, no activity is detected during the absence of stimuli. The firing rate has a posterior mean (and related 95% credible interval) equal to 0.488 $(0.471, 0.506)$ for the static grating stimulus, to 1.192 $(1.166, 1.220)$ for the natural scene, and to 0.304 $(0.286, 0.325)$ for the natural movie. Figure 3.7 shows the histograms of the posterior means of the non-zero spike amplitudes for the three stimuli. Compared to the natural movie, for both the static grating and the natural scene we observe heavier tails and the presence of clusters associated with larger parameters. Notice also that the estimated spikes' amplitudes are quite similar to the ones estimated in the previous analysis. This behavior is consistent with what observe in Figure 3.8, which shows a short interval of the observed calcium series, chosen in correspondence of one of the highest observed spikes.

Moving to the analysis of the last neuron ($\#587435328$), Figure 3.9 shows the observed fluorescence trace, together with the estimated activity. It is evident that this trace is very different from the previously considered ones, as it shows very little activity during the static grating stimulus and especially during the natural movie stimulus. This difference is reflected in the estimated distributional clusters, as now the natural movie is even pooled together with the absence of stimuli with high posterior probability; while the static grating and natural scene are assigned to two separate clusters. The firing rate was estimated equal to 0.067 $(0.062, 0.072)$ spikes per second for the static grating stimulus, to 1.354 $(1.326, 1.383)$ for the natural scene, and to 0.0063

$(0.0033, 0.0066)$ for the natural movie. Figure 3.10 shows the histograms of the distribution of the posterior means of the observational cluster parameters A_t. We also report the histogram associated with the natural movie, as few spikes were detected in correspondence of the end of the stimulus. It is evident how the natural scene is associated with the most intense activity, while the static grating leads a lower, but still present, activity. Finally Figure 3.11, similarly to the previous analyses, shows a short interval of the observed calcium series, together with colored vertical lines at the estimated spike times.

In line with the current literature, our approach is limited to the analysis of the calcium responses observed from single neurons. Inferences from our work could possibly be used to identify patterns across multiple neurons. For example, it is reasonable to assume that neurons exhibiting similar activity patterns may be grouped into homogeneous (spatial) clusters. Therefore, a second stage of the analysis may explicitly cluster across neurons the inferred spikes and the posterior means of the amplitudes within successive time-intervals of calcium traces. Alternatively, one could apply the zero-inflated gamma model recently proposed by Wei et al. (2019) to study the densities of the deconvolved activity estimates and similarly heuristically compare such densities across neurons.

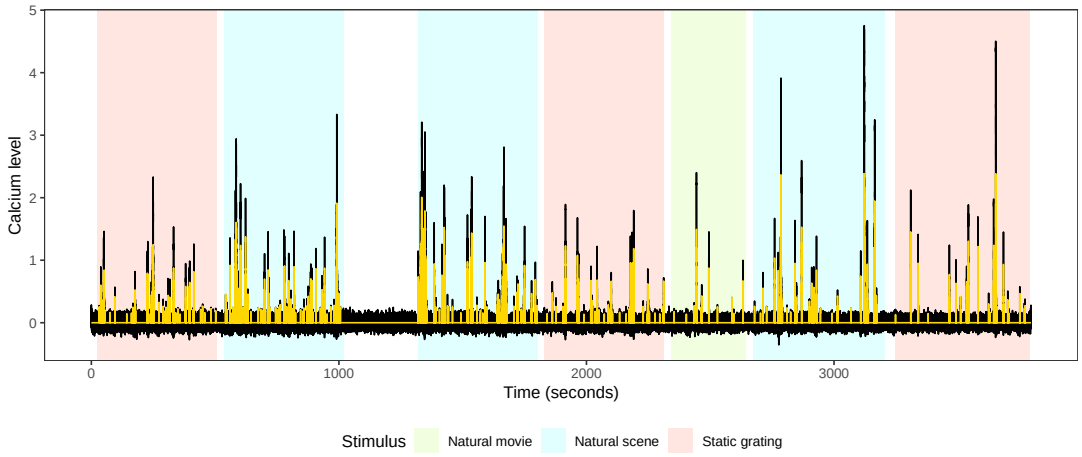

FIGURE 3.6: Observed fluorescence trace for neuron #517398389 at depth 350 microns (black line), and visual stimulus to which the mouse is exposed (shaded areas). The yellow line represents the estimated neuronal activity.

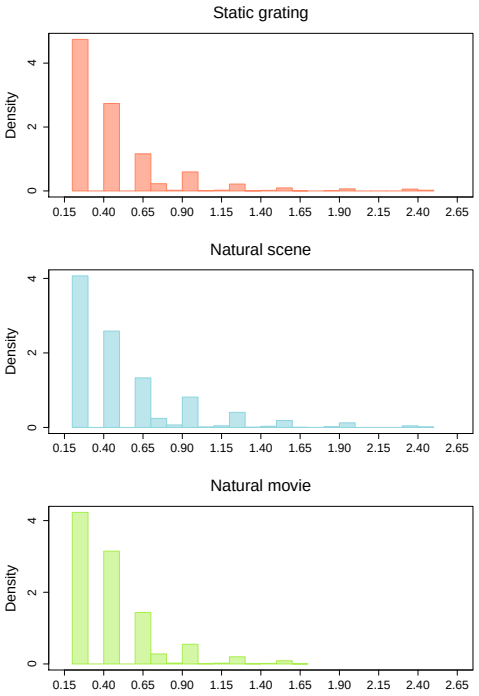

FIGURE 3.7: Empirical distribution of the posterior means of the observational cluster parameters A_t for the three experimental conditions, for neuron #517398389.

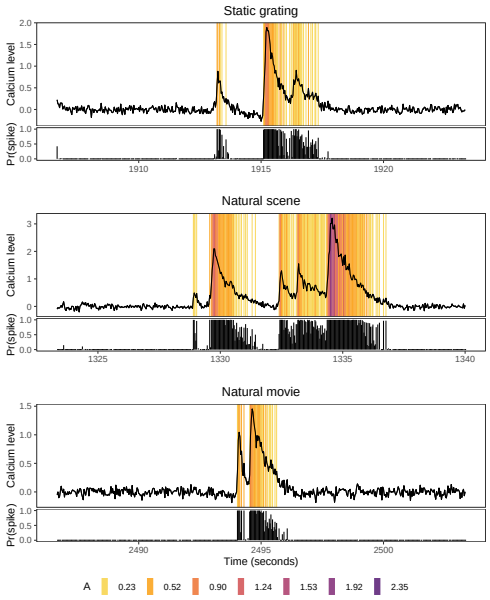

FIGURE 3.8: Short interval of length 500 in correspondence of a spike for the three stimuli, for neuron #517398389. The vertical lines indicate the spikes' times and the colors correspond to the observational cluster of their amplitudes. The bottom panels show the estimated posterior probability of spike presence, for each time point.

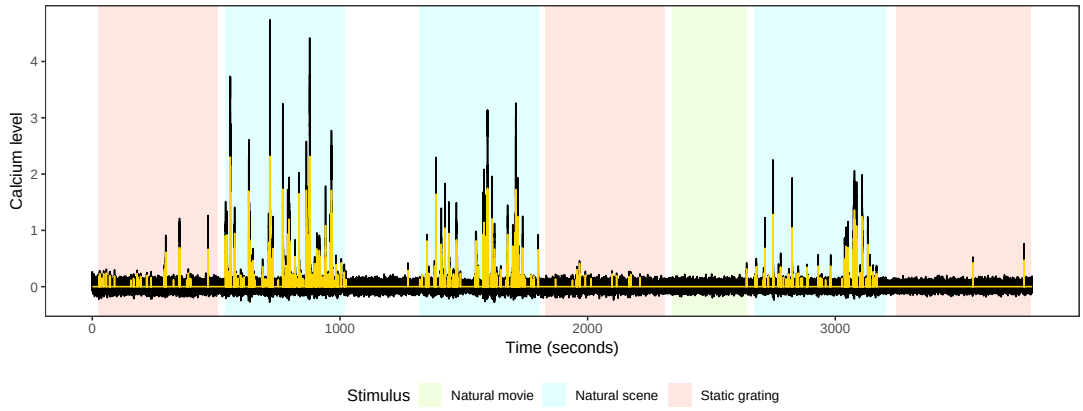

FIGURE 3.9: Observed fluorescence trace for neuron neuron #587435328 at a depth 175 (black line), and visual stimulus to which the mouse is exposed (shaded areas). The yellow line represents the estimated neuronal activity.

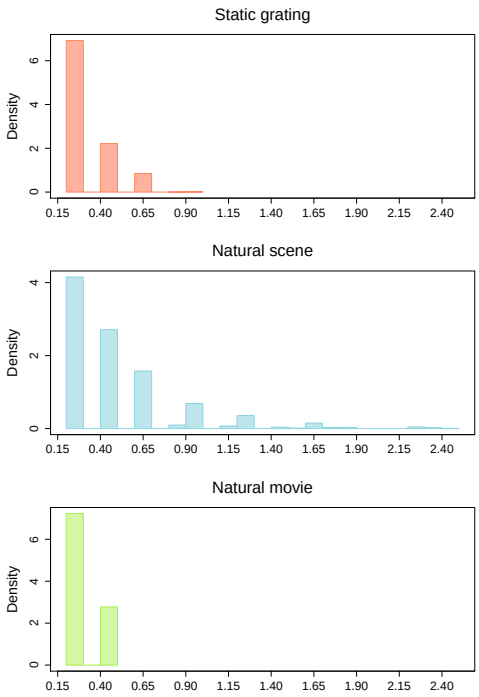

FIGURE 3.10: Empirical distribution of the posterior means of the observational cluster parameters A_t for the three experimental conditions, for neuron #587435328.

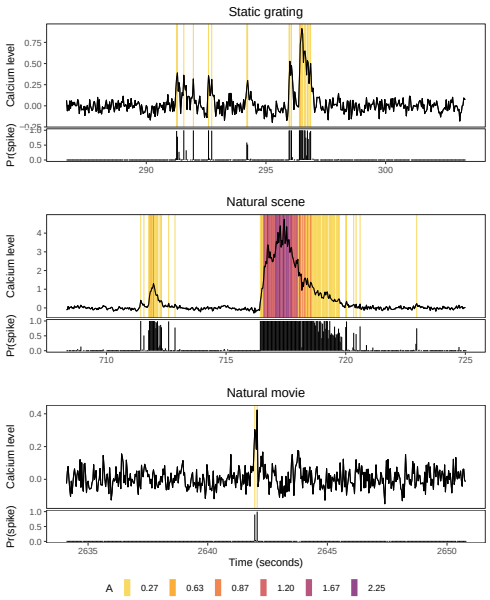

FIGURE 3.11: Short interval of length 500 in correspondence of a spike for the three stimuli, for neuron #587435328. The vertical lines indicate the spikes' times and the colors correspond to the observational cluster of their amplitudes. The bottom panels show the estimated posterior probability of spike presence, for each time point.

4 CLUSTERING ACTIVATION PATTERNS OF SPATIALLY-REFERENCED NEURONS

The previous approach only accounted for the analysis of calcium traces of individual neurons. The analysis of populations of neurons could only be performed as a second phase by combining the results of multiple analyses. Although not ideal, this approach is sometimes the only viable option due to the complexity and the large size of the data as, for example, the length of the individual series in the Allen Brain Observatory data.

In this chapter we develop a model to analyze the activity of groups of neurons and to cluster this activity on the basis of recurring patterns of activation. As motivating application, we consider the hippocampal neurons data described in Section 1.2.2: differently from the Allen Brain Observatory data, here the animal is not subjected to different types of stimuli, and it is freely moving within an environment. This type of experimental setting is often used to investigate hippocampal dynamics and connectivity patterns, which consist of groups of co-activating neurons. Hippocampal neurons underlying spatial navigation are thought to have a distributed organization, not strictly connected to the anatomical structure; however, many studies have shown that neurons with the same place field tuning often tend to neighbor one another (Eichenbaum et al., 1989; Redish et al., 2001). Hence, when investigating clusters of co-activating neurons, it can be relevant to take into account their spatial location and, in particular, their proximity to each other. In particular, the interest is aimed to identifying subpopulations of neurons which share a common spiking activity over seconds-long periods of time (Bittner et al., 2017). In the case of the data that we considered (recorded with a frequency of 15 frames per second), it means estimating clusters of temporal activity patterns with a duration up to a few hundreds of time points.

We develop a nonparametric mixture model that allows for simultaneous deconvolution of calcium traces and identification of groups of neurons with a similar pattern of activity during seconds-long periods. Moreover, the weights of the mixture prior are informed using the spatial proximity between cells. In particular, to identify such clusters, our modeling framework looks for similarities in the deconvolved binary time series describing the active/resting state of the neurons at each time point. A possible difficulty in clustering these time series arises from the presence of isolated or erratic spikes, which make the observed series somehow different, even if the overall patterns match. To overcome this drawback, we perform clustering at a latent level through a process that describes, at each time point, the probability of observing a spike, hence allowing for some degree of discrepancy across series in the same cluster.

4.1. Model and prior specification

Once again, we employ the general model for the calcium dynamics introduced in equation (1.1) of Chapter 1. However, differently from the previous chapter, here we consider n neurons, so we also introduce the index $i = 1, \dots, n$ corresponding to each fluorescence trace. Moreover, we split the parameters A_t of equation (1.1) into two separate components: $s_{i,t} \in \{0, 1\}$ describing the presence/absence of a spike (the *signal*), and $a_{i,t} \in \mathbb{R}^+$ describing the spike amplitude when present. With these modifications, the model can be written, for time $t = 1, \dots, T$, as

$$\begin{aligned} y_{i,t} &= b_i + \mathrm{Ca}_{i,t} + \epsilon_{i,t}, \qquad \epsilon_{i,t} \sim \mathrm{N}(0, \sigma^2), \\ \mathrm{Ca}_{i,t} &= \gamma\, \mathrm{Ca}_{i,t-1} + s_{i,t} \cdot a_{i,t} + w_{i,t}, \qquad w_{i,t} \sim \mathrm{N}(0, \tau^2), \end{aligned} \tag{4.1}$$

where the baseline parameters b_i are now neuron-specific. Moreover, for each observation it is also provided information on the spatial location of the neuron in the region of interest, $\boldsymbol{l}_i \in \mathcal{L} \subseteq \mathbb{R}^2$.

In this context, the interest is in clustering the n neurons according to their pattern of activation, which is described by the binary series $\boldsymbol{s}_i = \{s_{i,1}, \dots, s_{i,T}\}$. However, we would like these clusters to comprise all neurons with a *similar* activation pattern, even if the series differ for some occasional or isolated spikes. Instead of clustering directly the binary time series, we assume that these series are functions of an underlying continuous process that describes the spike probabilities, and we perform clustering at this latent level.

Specifically, we assume that, for each $t = 1, \dots, T$, the latent signal $\boldsymbol{s}_i$ is the realization of independent Bernoulli random variables whose probability depends on an underlying mixture of Gaussian processes through a probit transformation. Denoting with $\tilde{\boldsymbol{s}}_i = \{\tilde{s}_{i,1}, \dots, \tilde{s}_{i,T}\}$ the realization of this underlying process, we write

$$s_{i,t} \sim \mathrm{Bernoulli}(\Phi(\tilde{s}_{i,t})),$$

where $\Phi(\cdot)$ is the cumulative distribution function of a standard Gaussian distribution. Assuming a latent Gaussian process also allows us to easily describe the observed temporal dependence among spikes through the covariance function. As already noticed in the application to the Allen Brain Observatory data in the previous chapter, often the observed longer duration of a transient is the result of the summation of multiple spikes (Dombeck et al., 2010). Hence it is clear that the spikes are not uniformly distributed in time, and that explicitly modeling this behavior might improve detection and interpretation.

To obtain a clustering of neurons, we assume a mixture prior on the underlying Gaussian process that controls the probability of observing a spike at each time point. To include information on the spatial location of each neuron, we make use of the probit stick-breaking process (PSBP) of Rodríguez and Dunson (2011), where the weights are informed using the proximity matrix between neurons $\Sigma(\boldsymbol{l})$. This nonparametric

prior on $\tilde{\boldsymbol{s}}_i$ can be written as

$$\begin{aligned} \tilde{\boldsymbol{s}}_i \mid \boldsymbol{l}_i &\sim G_{\boldsymbol{l}_i} \\ G_{\boldsymbol{l}_i} &= \sum_{k\geq 1} \pi_k(\boldsymbol{l}_i) \cdot \delta_{\tilde{\boldsymbol{s}}_k^*} \\ \pi_k(\boldsymbol{l}_i) &= \Phi\big(\alpha_k(\boldsymbol{l}_i)\big) \prod_{r<k} \big\{1 - \Phi\big(\alpha_r(\boldsymbol{l}_i)\big)\big\} \end{aligned} \tag{4.2}$$

with

$$\begin{bmatrix} \alpha_k(\boldsymbol{l}_1) \\ \alpha_k(\boldsymbol{l}_2) \\ \vdots \\ \alpha_k(\boldsymbol{l}_n) \end{bmatrix} \sim \mathrm{N}_n \left(\mathbf{0},\, \Sigma(\boldsymbol{l}) = \begin{bmatrix} 1 & k(\boldsymbol{l}_1, \boldsymbol{l}_2) & \dots & k(\boldsymbol{l}_1, \boldsymbol{l}_n) \\ k(\boldsymbol{l}_1, \boldsymbol{l}_2) & 1 & \dots & k(\boldsymbol{l}_2, \boldsymbol{l}_n) \\ \vdots & \vdots & \ddots & \vdots \\ k(\boldsymbol{l}_1, \boldsymbol{l}_n) & k(\boldsymbol{l}_2, \boldsymbol{l}_n) & \dots & 1 \end{bmatrix} \right)$$

where $k(\boldsymbol{l}_i, \boldsymbol{l}_{i'})$ is a covariance function. Finally, the atoms of the mixture are independent draws from a Gaussian process (GP), i.e.,

$$\tilde{\boldsymbol{s}}_k^* \sim \mathrm{GP}(\boldsymbol{\mu}, \Omega),$$

where the covariance function $\Omega(t, t')$ describes the temporal dependence, that we model using a squared exponential kernel.

Consistently with the approach described in the previous chapter, we model the positive spike amplitudes using a Gamma prior, $a_{i,t} \sim \mathrm{Gamma}(h_{1a}, h_{2a})$. Moreover, for the remaining parameters, we adopt the same prior specification as in Eq. (3.4).

4.2. Posterior inference

Posterior inference for the proposed model can be carried out using MCMC methods, and, in particular, the Gibbs sampler, as for most parameters the full conditional distributions are available analytically. In line with the previous work, also here it is convenient to introduce the latent cluster allocation variables $c_i \in \{1, 2, \dots\}$, that, in this context, identify the groups of neurons with a similar activation pattern. Conditionally on $\pi_k(\boldsymbol{l}_i)$, we have $\Pr(c_i = k \mid \boldsymbol{l}_i) = \pi_k(\boldsymbol{l}_i)$. Hence the distribution of the signal for neuron i can be expressed, conditionally on the cluster allocation, as

$$p(\boldsymbol{s}_i \mid c_i = k, \tilde{\boldsymbol{s}}_k^*) = \prod_{t=1}^{T} \Phi(s_{k,t}^*)^{s_{i,t}} \big(1 - \Phi(s_{k,t}^*)\big)^{1 - s_{i,t}}. \tag{4.3}$$

Notice that the cluster allocation only affects the latent process controlling the spike probabilities, hence the c_i's are independent of the observed traces, given the series of the estimated signal s_i. Moreover, as the temporal dependence between spikes is expressed only at the latent level, the distribution of each observed series is simply

$$f(\boldsymbol{y}_i \mid b_i, \mathbf{Ca}_i, \gamma, \boldsymbol{s}_i, \boldsymbol{a}_i, \sigma^2, \tau^2) = \prod_{t=1}^{T} \phi(y_{i,t} \mid\; b_i + \gamma\, \mathrm{Ca}_{i,t-1} + s_{i,t} \cdot a_{i,t};\; \sigma^2 + \tau^2), \tag{4.4}$$

where $\phi(\cdot \mid \mu; \varsigma^2)$ is the density function of a normal random variable of mean μ and variance ς^2. Hence to obtain a sample from the posterior distribution of b_i, γ, $\mathbf{Ca}_i$, σ^2 and τ^2 we can adapt the MCMC steps described in Section 3.2 for the multivariate case in a straightforward manner.

Combining the prior distribution of the signal in Eq. (4.3) with the likelihood (4.4), the full conditional distribution of $\boldsymbol{s}_i$ is easily obtained as

$$\Pr(s_{it} = 1 \mid y_{it}, c_i = k, \tilde{s}_{k,t}, -) = \frac{1}{\sqrt{2\pi(\sigma^2+\tau^2)}} e^{-\frac{1}{2(\sigma^2+\tau^2)}(y_{i,t}-b_i-\gamma \mathrm{Ca}_{i,t-1}-a_{i,t})^2} \Phi(\tilde{s}_{k,t})$$

$$\Pr(s_{it} = 0 \mid y_{it}, c_i = k, \tilde{s}_{k,t}, -) = \frac{1}{\sqrt{2\pi(\sigma^2+\tau^2)}} e^{-\frac{1}{2(\sigma^2+\tau^2)}(y_{i,t}-b_i-\gamma \mathrm{Ca}_{i,t-1})^2} \Phi(-\tilde{s}_{k,t}).$$

Notice that the probability of observing a spike at time t also depends on the specific amplitude $a_{i,t}$. Hence at each iteration we need to sample a new value for all amplitude parameters $a_{i,t}$, even if a spike was not detected for that particular neuron and time. Regarding the sampling of the amplitudes, assuming a Gamma prior does not lead to a simple expression of the full conditional, hence we make use of a Metropolis-Hastings step.

The update of the cluster allocation variables c_i using the location-dependent PSBP is performed using the data augmentation strategy outlined in Rodríguez and Dunson (2011). For simplicity, we used a finite PSBP with a large number of components, as it constitutes a fair approximation of the original process based on an infinite number of components (Rodríguez and Dunson, 2011; Ishwaran and James, 2001).

Finally, slightly more demanding and computationally intensive, is the sampling of the realizations of the latent Gaussian process. To this end, we exploit the exponentially decreasing correlation between time points in our definition of Ω to approximate the Gaussian process to a collection of conditionally independent multivariate random variables. Specifically, since after a certain lag p the covariance $\Omega(t, t+p) = \Omega(t, t-p)$ is *virtually* equal to zero, we set all the corresponding elements in the covariance matrix exactly to zero. In this way, we obtain a T-variate Gaussian distribution with a band covariance matrix. This device allows us to write the model in state-space form and to estimate the latent process using the closed-form filter developed by Fasano et al. (2021) for binary time series. In our specific case, the observed level is the set of binary series of signal $\boldsymbol{s}_i$ for all neurons in the same activation cluster, and the state equation can be expressed as depending on a p-variate Gaussian random vector.

4.3. Simulation study

The performances of the proposed model in detecting the spikes and clustering the extracted activation patterns are investigated through a simulation study. Unfortunately, the high computational cost of the proposed algorithm constitutes an obstacle to running a full and thorough simulation study, hence here we only present a preliminary analysis of the results. Additional work will be needed in order to devise computationally efficient strategies to perform posterior inference.

We simulated data according to our model, and we considered a high signal-to-noise ratio, in order to focus more on the clustering performances, rather than the spike detec-

	$n = 20$		$n = 30$		$n = 40$	
	$T = 100$	$T = 200$	$T = 100$	$T = 200$	$T = 100$	$T = 200$
MLGP	0	$5.25 \cdot 10^{-3}$	0	$1.00 \cdot 10^{-3}$	0	$7.50 \cdot 10^{-4}$
L_0	$2.20 \cdot 10^{-2}$	$2.97 \cdot 10^{-2}$	$1.73 \cdot 10^{-2}$	$2.33 \cdot 10^{-2}$	$2.22 \cdot 10^{-2}$	$2.30 \cdot 10^{-2}$

TABLE 4.1: Misclassification error rate on the simulated data for the proposed model (MLGP) and the method of Jewell et al. (2019) (L_0).

	$n = 20$		$n = 30$		$n = 40$	
	$T = 100$	$T = 200$	$T = 100$	$T = 200$	$T = 100$	$T = 200$
MLGP	1	0.915	0.870	0.923	0.912	0.912
L_0 + hierarchical	0.469	0.853	0.820	0.889	0.516	0.648
true + hierarchical	1	0.968	0.820	0.966	1	0.934

TABLE 4.2: Adjusted Rand index of the estimated clusters of activity on the simulated data for the proposed model (MLGP), the two-stage approach (L_0 + hierarchical) and a hierarchical clustering on the true signal (true + hierarchical).

tion task. Specifically, the spike amplitudes were generated from a Gamma distribution with mean and variance equal to 3 and 0.7^2, respectively; while the measurement error variance was set to 0.3^2. Moreover, we set all baseline parameters b_i to zero, and we fixed the decay parameter $\gamma = 0.5$.

We considered scenarios with varying sample size ($n = 20, 30, 40$) and length of the calcium traces ($T = 100, 200$). Here we only present one simulated scenario for each setting to briefly describe the model's behavior and assess its performance. To compare the results of the proposed mixture of latent Gaussian processes (MLGP), we also applied a standard two-stage approach. Specifically, we (1) first deconvolved the simulated calcium traces using the approach introduced by Jewell et al. (2019), then, (2) we clustered the extracted series of the signal using a hierarchical clustering based on the Hamming distance. The penalization parameters in step (1) were selected following the procedure illustrated in de Vries et al. (2020) to minimize the number of estimated spikes smaller than 2 standard deviations of the trace. To assess the influence of spike detection on the estimated clustering in the two-stage approach, we also estimated a hierarchical clustering on the true signal.

We evaluated our model and step (1) of the two-stage approach by comparing the misclassification error rate obtained in the simulated scenarios (Tab. 4.1). In general, the performances of the proposed model are superior to those obtained using the L_0 penalization approach. This is consistent with the findings of the previous chapter, where we assessed that a simultaneous deconvolution and estimation of the spiking activity can improve spike detection.

Estimation of the clustering structure was assessed using the adjusted Rand index (Rand, 1971; Hubert and Arabie, 1985). Table 4.2 compares the results of the proposed model with those attained by the hierarchical clustering (2), both on the estimated signal, after the deconvolution phase (1), and by applying step (2) directly on the true signal, to exclude the impact of spike detection on cluster recognition. The proposed model shows superior performance compared to the two-stage approach in all simulated scenarios. The hierarchical clustering based on the true signal has

overall the best results, however, it is clearly not applicable in a real application, as it assumes perfect identification of the spikes. Moreover, in hierarchical or centroid-based clustering, one has to fix some parameters, whose choice is somehow arbitrary, and that heavily affect the resulting partition (e.g. the number of clusters). Conversely, model-based clustering is relatively free from tuning parameters and subjective choices, leading to more stable and data-driven results.

4.4. Analysis of hippocampal neurons

Finally, we move to the analysis of the hippocampal neurons data described in Sec. 1.2.2. Also in this case we will present only a preliminary analysis, as the computational issues mentioned in the previous section are particularly relevant in this context, because of the length of the series and the large number of neurons.

We considered the first 2000 time points of the experiment, and we selected a subset of 30 neurons that, from a preliminary analysis, showed at least one spike in some pre-specified intervals (to avoid the inclusion of neurons with no activations in the considered time window). On the calcium traces that we considered, we first performed a pre-processing phase by down-sampling the traces, keeping one observation every 10, hence leading to 200 time points for each series. This step was necessary to make the data more conform to other fluorescent traces, as, for example, those from the Allen Brain Observatory analyzed in the previous chapter, and for computational reasons.

Neuroscience research aims to identify groups of neurons with a pattern of activity that coincides over seconds-long periods. In real experiments, the calcium traces are recorded for very long periods of time, hence, the clustering of neurons can not be performed on the whole series. It is, indeed, unlikely that neurons exhibit a similar activity for the entire duration of the experiment, being the dynamics that regulate the neuronal activity very complex. This issue leads to the need to specify the length of the intervals where we seek for clustering structures: neuroscience provides some guidelines on this choice (Bittner et al., 2017), however, there are no clear indications on the exact length of such time windows, as it may depend on the specific scope and setting of the research. Here, we considered both 4 non-overlapping windows of 50 time points (75 seconds) each and 2 windows of 100 time points (150 seconds), to assess influence of this choice on the resulting clustering. Figg. 4.4 and 4.5 compare these two scenarios: the top panel of each figure shows the observed calcium traces, and the colors correspond to the estimated cluster label, which can vary in each time window. The bottom panels show the location of neurons in the hippocampus, and the colors again show the evolution of the estimated clustering. In both scenarios, most neurons are assigned to the same cluster, and this large group remains together for the whole duration of the experiment. From a visual inspection, one would expect a larger number of clusters, as many traces which are now grouped together appear to be quite different. However, the model is able to detect some specific patterns of activity, which are reasonably assigned to separate groups.

Both scenarios show that there is scope for improvement in the clustering procedure, and there are a few directions which are worth investigation. An important factor that affects the results is the process on the amplitudes. Figure 4.6 shows the distribution

of the estimated parameters a_{it}: the spike amplitudes appear to be very heterogeneous and with a large variance. Using a gamma distribution to model the positive spikes could hence be too restrictive, and a more flexible prior could improve spike estimation and, in turn, the clustering structure. For example, adopting a mixture prior, similarly to the previous chapter, could be appropriate to address these issues. Another aspect which can be investigated is the use of overlapping time-windows, which could lead to a more progressive transition between partitions.

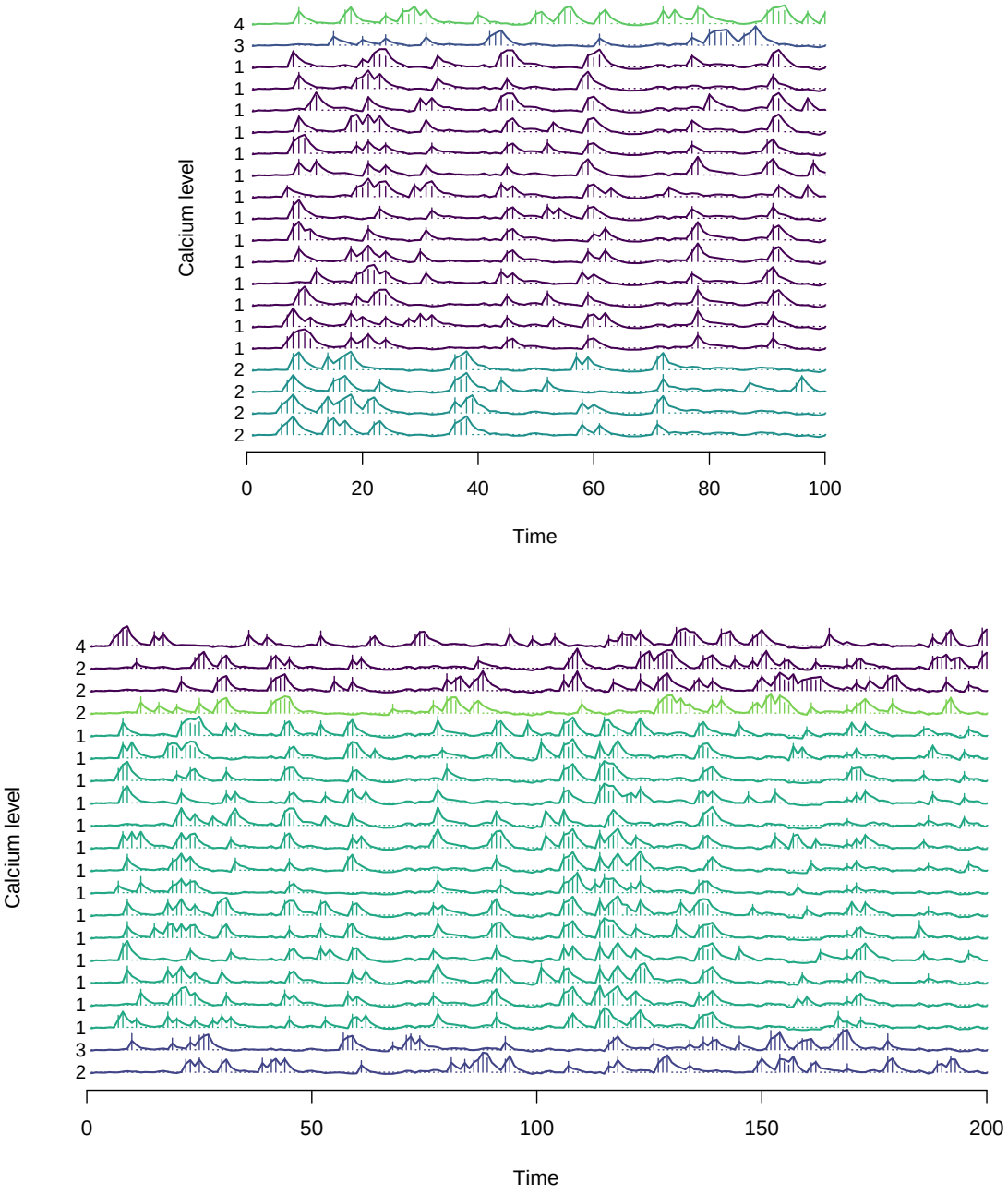

FIGURE 4.1: Estimated clustering on the simulated data with sample size $n = 20$, and series length $T = 100$ (top) and $T = 200$ (bottom). Each series is the simulated calcium trace; the colors correspond to the estimated clusters, while the numbers on the left of each series correspond to the true partition. The vertical segments correspond to the detected spikes.

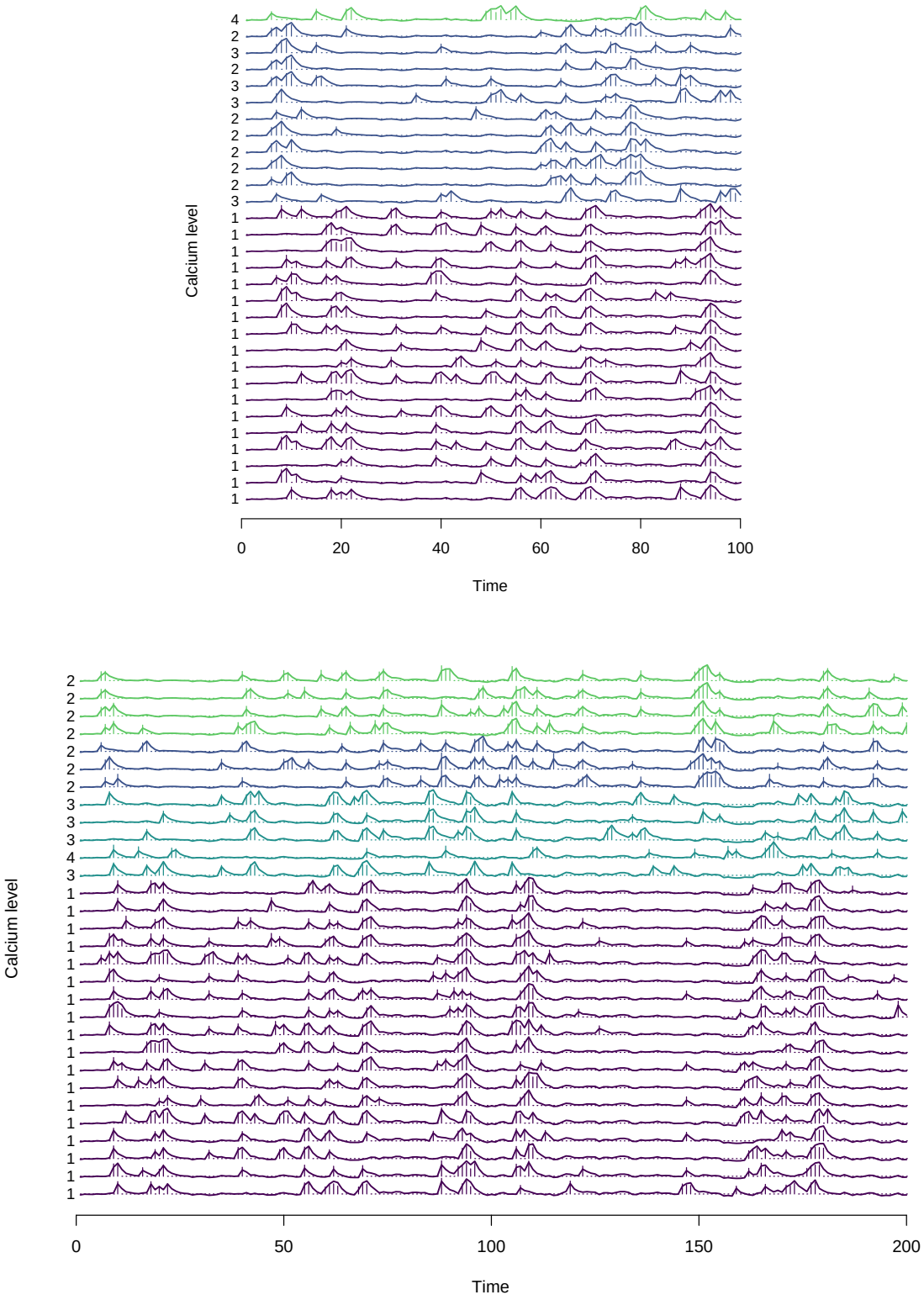

FIGURE 4.2: Estimated clustering on the simulated data with sample size $n = 30$, and series length $T = 100$ (top) and $T = 200$ (bottom). Each series is the simulated calcium trace; the colors correspond to the estimated clusters, while the numbers on the left of each series correspond to the true partition. The vertical segments correspond to the detected spikes.

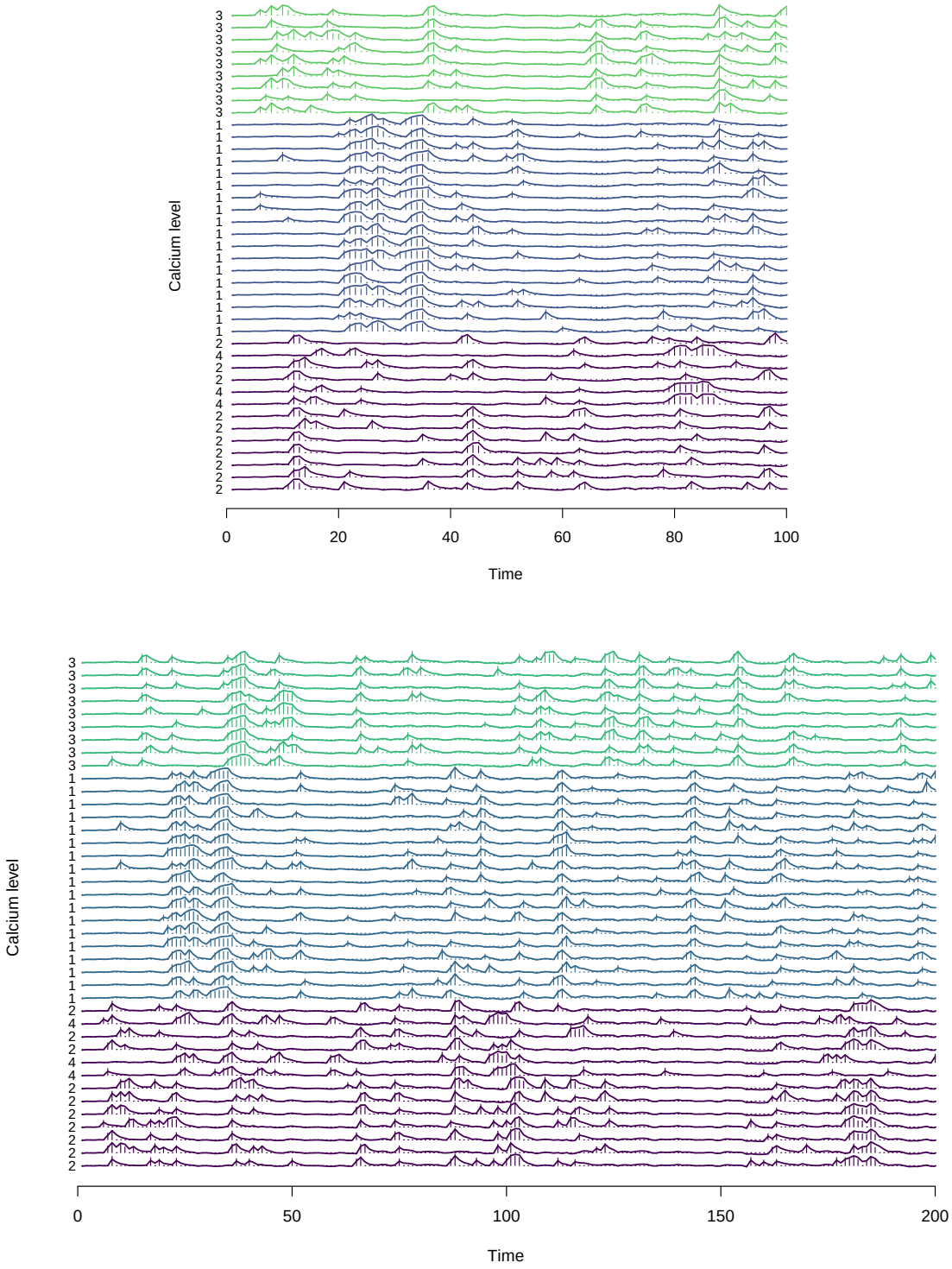

FIGURE 4.3: Estimated clustering on the simulated data with sample size $n = 40$, and series length $T = 100$ (top) and $T = 200$ (bottom). Each series is the simulated calcium trace; the colors correspond to the estimated clusters, while the numbers on the left of each series correspond to the true partition. The vertical segments correspond to the detected spikes.

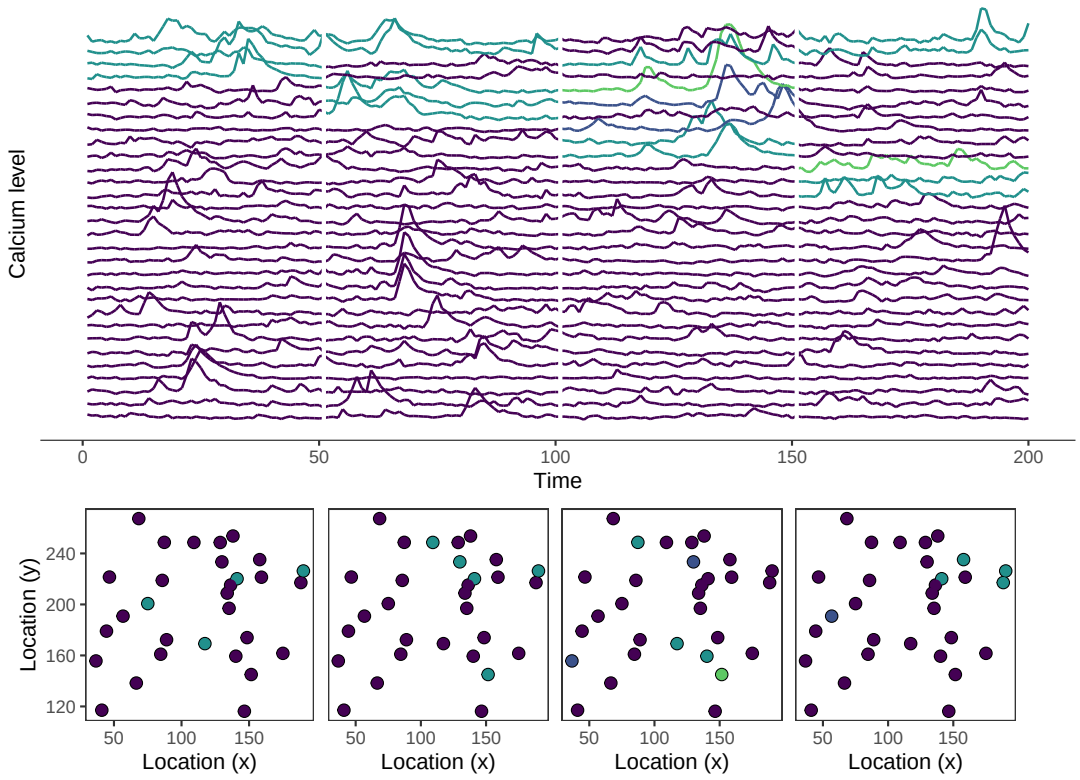

FIGURE 4.4: Hippocampal neurons data: observed calcium traces (top) and locations (bottom) for the considered subset of neurons. The colors correspond to the estimated clustering in 4 non-overlapping time windows of length 50.

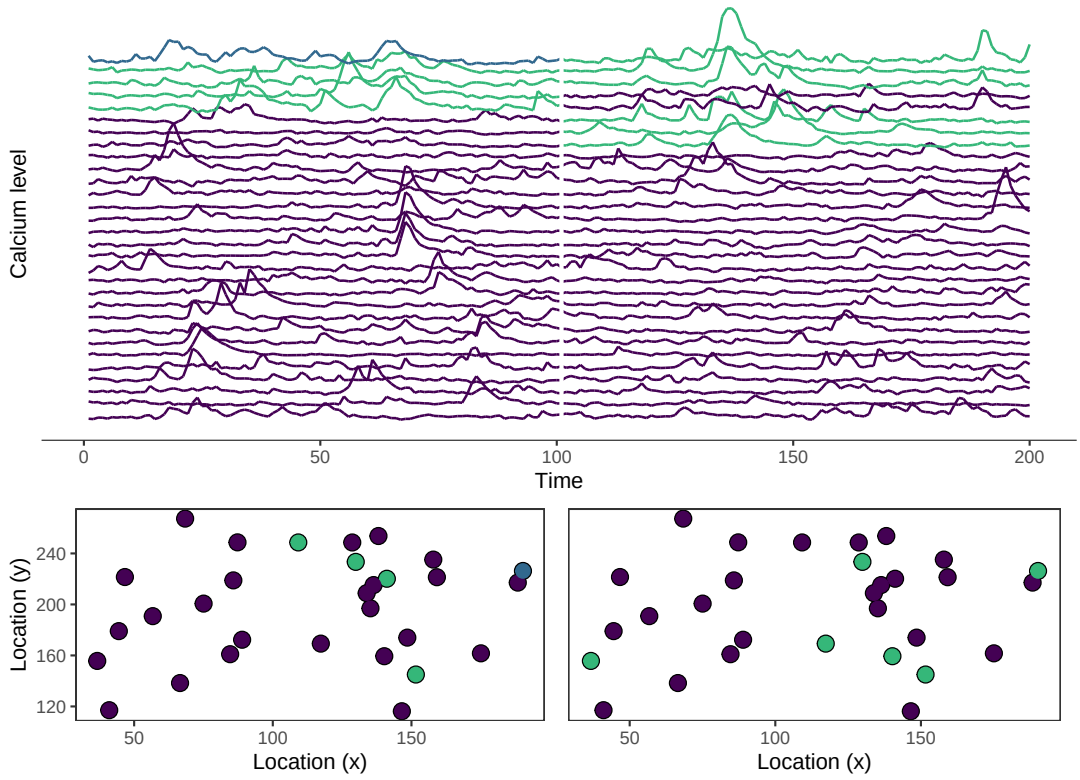

FIGURE 4.5: Hippocampal neurons data: observed calcium traces (top) and locations (bottom) for the considered subset of neurons. The colors correspond to the estimated clustering in 2 non-overlapping time windows of length 100.

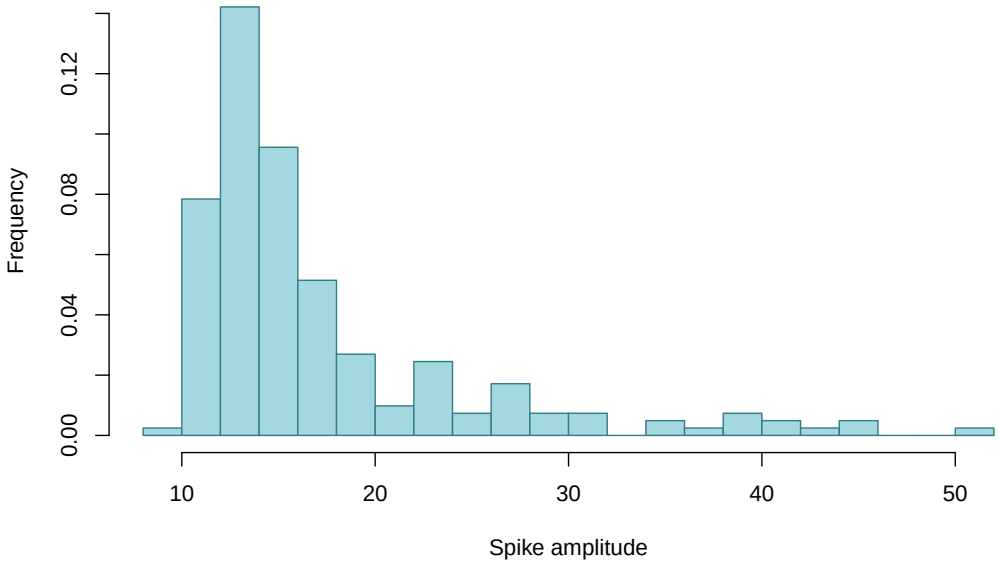

FIGURE 4.6: Distribution of the estimated spike amplitudes on the hippocampal neurons data.

CONCLUSIONS

Discussion

In recent years the technological advances have enabled the collection of increasingly complex data. Calcium imaging data fit perfectly into this context: being high dimensional, often collected in elaborate experimental settings, with spatial and temporal dependence structures, and with a non-homogeneous response between neurons, they present several modeling challenges. Analyzing these data hence fosters investigation of new statistical and computational tools in many directions. In this thesis, we have examined three different, although related, aspects of a Bayesian analysis of these data.

In the first chapter, we have considered a classical two-stage approach, based on a first deconvolution phase and a successive statistical analysis of the output. Specifically, we have examined the use of Poisson regression models to relate the number of detected spikes with several covariates describing the experimental conditions. However, although we focused on this specific application, Poisson log-linear models are routinely used in many contexts, making our work applicable also outside of the scope of calcium imaging studies. We have developed two Markov chain Monte Carlo algorithms to sample from the posterior distribution of the regression parameters under the assumption of conditionally Gaussian prior distributions. The algorithms exploit the introduction of an approximate posterior distribution, which is used as the building block for a Metropolis-Hastings and importance sampling algorithms. The proposed sampling strategies show good performances in terms of efficiency compared to state-of-the-art methods.

In the second chapter, we have developed a nonparametric nested mixture model that allows for simultaneous deconvolution and estimation of the spiking activity, hence overcoming standard two-stage approaches. The model makes use of two nested layers of random discrete mixture priors to borrow information between experiments and discover similarities in the neuronal response to different types of stimuli. The results on simulated data show how simultaneous deconvolution and estimation of the spike amplitudes leads to lower misclassification error, thanks to the borrowing of information between the two phases. Application to a real data set from the Allen Brain Observatory demonstrates the ability to capture characteristic features of neuronal activity.

Finally, in the last chapter, we have moved to the multivariate analysis of populations of neurons. In general, neurons do not exhibit a homogeneous response to stimulation, and a relevant research question in neuroscience is studying groups of co-activating cells. This motivated the investigation of new clustering strategies to identify calcium traces with a similar underlying pattern of activity over seconds-long periods of time.

We have formulated a nonparametric mixture model that deconvolves the fluorescence traces and clusters the latent binary series of activity. The latter task is achieved through the introduction of a latent continuous process that explicitly characterizes the spike probabilities and models their temporal dependence. Moreover, spatial dependence is also taken into account by using location-dependent mixture weights.

Future directions of research

The work described in Chapter 4 presents some aspects that are prone to possible extensions and improvements, some of which have already been described in the dedicated sections. A first issue, already raised in Sec. 4.3, is related to the computational cost of the algorithm used to perform posterior inference. To estimate the realizations of the latent mixture of Gaussian process, we applied the particle filtering algorithm proposed by Fasano et al. (2021). Although the algorithm is a very good strategy to perform inference on binary state space models, it is not the most appropriate method in our context, as it must be run at every iteration of the Gibbs sampler. Hence, it would be useful to devise computationally more efficient ways to sample the realizations of the latent mixture of Gaussian processes.

The process on the amplitudes is very simple and somehow restrictive: as already discussed, it would be useful to replace the gamma prior with a more flexible distribution such as, for example, a mixture prior, similarly to Chapter 3.

Another possible improvement is relative to the time window used in the application to the hippocampal neurons data. As already pointed out, the choice of the time window can heavily affect the resulting clustering and hence the implications of the findings. Moreover, one could choose between segmenting the series or opting for a sliding window approach. All these issues are worth additional research, and, in particular, it would be useful to develop some methods to evaluate the sensitivity of the clustering to the different alternatives, and, possibly, to provide a way to evaluate what choice led to the "best" results.

In the hippocampal neurons data set it is also available an additional covariate that at each time records the spatial coordinates of the mouse into the environment. It would be interesting to conduct further research to extend the model of Chapter 4 to include this information. To model dependence of the spiking activity from this covariate, one possibility could be to divide the environment into quadrants, and adopt the nested mixture of Chapter 3 by considering each quadrant as a different experimental condition. Another possibility could be to represent the movements of the mouse using a network structure. A regression model could then be defined to relate the neurons' activity at each time point with the mouse's location by using a prior distribution on graphs (Cai et al., 2019).

BIBLIOGRAPHY

Albert, J. H. and Chib, S. (1993). "Bayesian analysis of binary and polychotomous response data". *Journal of the American Statistical Association* 88(422), 669–679.

Allen Brain Observatory (2017). *Technical whitepaper: stimulus set and response analyses*. URL: help.brain-map.org/display/observatory/Documentation.

Allen Institute MindScope Program (2016). *Allen Brain Observatory – 2-photon visual coding* [*dataset*]. brain-map.org/explore/circuits.

Argiento, R. and De Iorio, M. (2019). "Is infinity that far? A Bayesian nonparametric perspective of finite mixture models". *arXiv:1904.09733*.

Berridge, M. J., Lipp, P., and Bootman, M. D. (2000). "The versatility and universality of calcium signalling". *Nature Reviews Molecular Cell Biology* 1, 11–21.

Bittner, K. C., Milstein, A. D., Grienberger, C., Romani, S., and Magee, J. C. (2017). "Behavioral time scale synaptic plasticity underlies CA1 place fields". *Science* 357(6355), 1033–1036.

Blackwell, D. and MacQueen, J. B. (1973). "Ferguson distributions via Pólya urn schemes". *The Annals of Statistics* 1(2), 353–355.

Bradley, J. R., Holan, S. H., and Wikle, C. K. (2018). "Computationally efficient multivariate spatio-temporal models for high-dimensional count-valued data (with discussion)". *Bayesian Analysis* 13(1), 253–310.

Brenner, N., Agam, O., Bialek, W., and Ruyter van Steveninck, R. de (2002). "Statistical properties of spike trains: universal and stimulus-dependent aspects". *Physical review. E, Statistical, nonlinear, and soft matter physics* 66 (3), 031907.

Cai, Q., Kang, J., and Yu, T. (2019). "Bayesian network marker selection via the thresholded graph Laplacian Gaussian prior". *Bayesian Analysis* 15(1), 79–102.

Camerlenghi, F., Dunson, D. B., Lijoi, A., Prünster, I., and Rodríguez, A. (2019). "Latent nested nonparametric priors (with discussion)". *Bayesian Analysis* 14(4), 1303–1356.

Canale, A., Lijoi, A., Nipoti, B., and Prünster, I. (2017). "On the Pitman–Yor process with spike and slab base measure". *Biometrika* 104(3), 681–697.

Canale, A., Lijoi, A., Nipoti, B., and Prünster, I. (2022). "Inner spike and slab Bayesian nonparametric models". *Econometrics and Statistics*, in press.

Canale, A., Lijoi, A., and Prünster, I. (2016). "Bayesian Nonparametrics". *Wiley StatsRef: Statistics Reference Online*. John Wiley & Sons, Ltd, 1–11.

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010). "The horseshoe estimator for sparse signals". *Biometrika* 97(2), 465–480.

Casella, G. and Berger, R. (2002). *Statistical Inference*. Duxbury advanced series in statistics and decision sciences. Thomson Learning.

Chan, A. B. and Vasconcelos, N. (2009). "Bayesian Poisson regression for crowd counting". *2009 IEEE 12th International Conference on Computer Vision*, 545–551.

D'Angelo, L. (2021). *R package bpr: Bayesian Poisson regression*. URL: CRAN.R-project.org/package=bpr.

de Vries, S. E. J., Lecoq, J. A., Buice, M. A., Groblewski, P. A., Ocker, G. K., et al. (2020). "A large-scale standardized physiological survey reveals functional organization of the mouse visual cortex". *Nature neuroscience* 23(1), 138–151.

Denk, W., Strickler, J. H., and Webb, W. W. (1990). "Two-photon laser scanning fluorescence microscopy". *Science* 248, 73–76.

Denti, F., Camerlenghi, F., Guindani, M., and Mira, A. (2021). "A common atoms model for the Bayesian nonparametric analysis of nested data". *Journal of the American Statistical Association*, in press.

Dombeck, D. A., Harvey, C. D., Tian, L., Looger, L. L., and Tank, D. W. (2010). "Functional imaging of hippocampal place cells at cellular resolution during virtual navigation". *Nature Neuroscience* 13, 1433–1440.

Drouin, E., Piloquet, P., and Péréon, Y. (2015). "The first illustrations of neurons by Camillo Golgi". *The Lancet Neurology* 14(6), 567.

Dudai, Y. (2004). "The neurosciences: the danger that we will think that we have understood it all". *The new brain sciences: perils and prospects*. Ed. by S. Rose and D. Rees. Cambridge, UK: Cambridge University Press, 167–180.

Eddelbuettel, D. and Francois, R. (2011). "Rcpp: Seamless R and C++ integration". *Journal of Statistical Software, Articles* 40(8), 1–18.

Eichenbaum, H., Wiener, S. I., Shapiro, M., and Cohen, N. J. (1989). "The organization of spatial coding in the hippocampus: a study of neural ensemble activity". *The Journal of neuroscience : the official journal of the Society for Neuroscience*.

Fasano, A., Rebaudo, G., Durante, D., and Petrone, S. (2021). "A closed-form filter for binary time series". *Statistics and Computing* 31.

Ferguson, T. S. (1973). "A Bayesian analysis of some nonparametric problems". *The Annals of Statistics* 1(2), 209–230.

– (1974). "Prior distributions on spaces of probability measures". *The Annals of Statistics* 2(4), 615–629.

Friedrich, J. and Paninski, L. (2016). "Fast active set methods for online spike inference from calcium imaging". *Advances In Neural Information Processing Systems* (NIPS 2016). Ed. by D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett. Barcelona, Spain, 1984–1992.

Friedrich, J., Zhou, P., and Paninski, L. (2017). "Fast online deconvolution of calcium imaging data". *PLOS Computational Biology* 13(3), 1–26.

Frome, E. L. (1983). "The analysis of rates using Poisson regression models". *Biometrics* 39(3), 665–674.

Frome, E. L. and Checkoway, H. (1985). "Use of Poisson regression models in estimating incidence rates and ratios". *American Journal of Epidemiology* 121(2), 309–323.

Frühwirth-Schnatter, S. and Malsiner-Walli, G. (2019). "From here to infinity: sparse finite versus Dirichlet process mixtures in model-based clustering". *Advances in Data Analysis and Classification* 13, 33–64.

Frühwirth-Schnatter, S., Frühwirth, R., Held, L., and Rue, H. (2009). "Improved auxiliary mixture sampling for hierarchical models of non-Gaussian data". *Statistics and Computing* 19(479).

Frühwirth-Schnatter, S., Malsiner-Walli, G., and Grün, B. (2021). "Generalized Mixtures of Finite Mixtures and Telescoping Sampling". *Bayesian Analysis* 16(4), 1279–1307.

Frühwirth-Schnatter, S. and Wagner, H. (2006). "Auxiliary mixture sampling for parameter-driven models of time series of counts with applications to state space modelling". *Biometrika* 93(4), 827–841.

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., et al. (2013). *Bayesian Data Analysis, Third Edition*. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis.

Ghosal, S. and van der Vaart, A. (2017). *Fundamentals of Nonparametric Bayesian Inference*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Grienberger, C. and Konnerth, A. (2012). "Imaging calcium in neurons". *Neuron* 73(5), 862–885.

Hastings, W. K. (1970). "Monte Carlo sampling methods using Markov chains and their applications". *Biometrika* 57(1), 97–109.

Hjort, N., Holmes, C., Müller, P., and Walker, S. (2010). *Bayesian Nonparametrics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Hoang, H., Sato, M.-a., Shinomoto, S., Tsutsumi, S., Hashizume, M., et al. (2020). "Improved hyperacuity estimation of spike timing from calcium imaging". *Scientific Reports* 10(1), 17844.

Hubert, L. and Arabie, P. (1985). "Comparing partitions". *Journal of Classification* 2(336), 193–218.

Hutchinson, M. K. and Holtman, M. C. (2005). "Analysis of count data using Poisson regression". *Research in Nursing & Health* 28(5), 408–418.

Ishwaran, H. and James, L. F. (2001). "Gibbs sampling methods for stick-breaking priors". *Journal of the American Statistical Association* 96(453), 161–173.

Jewell, S. and Witten, D. (2018). "Exact spike train inference via L0 optimization". *The Annals of Applied Statistics* 12(4), 2457–2482.

Jewell, S. W., Hocking, T. D., Fearnhead, P., and Witten, D. M. (2019). "Fast nonconvex deconvolution of calcium imaging data". *Biostatistics* 21(4), 709–726.

Johnson, V. E. and Rossell, D. (2010). "On the use of non-local prior densities in Bayesian hypothesis tests". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 72(2), 143–170.

Joshua, S. C. and Garber, N. J. (1990). "Estimating truck accident rate and involvements using linear and Poisson regression models". *Transportation Planning and Technology* 15(1), 41–58.

Kalli, M., Griffin, J., and Walker, S. (2011). "Slice sampling mixture models". *Statistics and Computing* 21, 93–105.

Karlis, D. and Meligkotsidou, L. (2005). "Multivariate Poisson regression with covariance structure". *Statistics and Computing* 15, 255–265.

Lambert, D. (1992). "Zero-inflated Poisson regression, with an application to defects in manufacturing". *Technometrics* 34(1), 1–14.

Lambert, J. H. (1758). "Observations variae in mathesin puram". *Acta Helvitica, physico-mathematico-anatomico-botanico-medica* 3, 128–168.

Makalic, E. and Schmidt, D. F. (2016). "A simple sampler for the horseshoe estimator". *IEEE Signal Processing Letters* 23(1), 179–182.

Malsiner-Walli, G., Frühwirth-Schnatter, S., and Grün, B. (2016). "Model-based clustering based on sparse finite Gaussian mixtures". *Statistics and Computing* 26, 303–324.

McCullagh, P. and Yang, J. (2008). "How many clusters?" *Bayesian Analysis* 3(1), 101–120.

Miaou, S.-P. (1994). "The relationship between truck accidents and geometric design of road sections: Poisson versus negative binomial regressions". *Accident Analysis & Prevention* 26(4), 471–482.

Miller, J. W. and Harrison, M. T. (2018). "Mixture models with a prior on the number of components". *Journal of the American Statistical Association* 113(521), 340–356.

Mitchell, T. J. and Beauchamp, J. J. (1988). "Bayesian variable selection in linear regression". *Journal of the American Statistical Association* 83(404), 1023–1032.

Mukamel, E. A., Nimmerjahn, A., and Schnitzer, M. J. (2009). "Automated analysis of cellular signals from large-scale calcium imaging data". *Neuron* 63(6), 747–760.

Müller, P., Parmigiani, G., and Rice, K. (2007). "FDR and Bayesian multiple comparisons rules". *Bayesian Statistics 8*. Ed. by J. Bernardo, M. Bayarri, J. Berger, A. Dawid, D. Heckerman, et al. Oxford, UK: Oxford University Press.

Müller, P., Quintana, F. A., Jara, A., and Hanson, T. (2015). *Bayesian nonparametric data analysis*. Springer.

Neal, R. M. (2011). "MCMC using Hamiltonian dynamics". *Handbook of Markov chain Monte Carlo* 2(11), 2.

Nelder, J. A. and Wedderburn, R. W. M. (1972). "Generalized linear models". *Journal of the Royal Statistical Society. Series A (General)* 135(3), 370–384.

Newton, M. A., Noueiry, A., Sarkar, D., and Ahlquist, P. (2004). "Detecting differential gene expression with a semiparametric hierarchical mixture method". *Biostatistics* 5(2).

Nobile, A. (2004). "On the posterior distribution of the number of components in a finite mixture". *The Annals of Statistics* 32(5), 2044–2073.

Nobile, A. and Fearnside, A. (2007). "Bayesian finite mixtures with an unknown number of components: the allocation sampler". *Statistics and Computing* 17, 147–162.

O'Keefe, J. and Nadel, L. (1978). *The hippocampus as a cognitive map*. Oxford university press.

Paninski, L., Pillow, J., and Lewi, J. (2007). "Statistical models for neural encoding, decoding, and optimal stimulus design". *Computational Neuroscience: Theoretical Insights into Brain Function*. Ed. by P. Cisek, T. Drew, and J. F. Kalaska. Vol. 165. Progress in Brain Research. Elsevier, 493–507.

Park, T. and Casella, G. (2008). "The Bayesian lasso". *Journal of the American Statistical Association* 103(482), 681–686.

Parker, D. (2006). "Complexities and uncertainties of neuronal network function". *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* 361(1465), 81–99.

– (2010). "Neuronal network analyses: premises, promises and uncertainties". *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* 365(1551), 2315–2328.

Peron, S. P. and Gabbiani, F. (2009). "Role of spike-frequency adaptation in shaping neuronal response to dynamic stimuli". *Biological cybernetics* 100(6), 505–520.

Piironen, J. and Vehtari, A. (2017). "Sparsity information and regularization in the horseshoe and other shrinkage priors". *Electronic Journal of Statistics* 11(2), 5018–5051.

Pnevmatikakis, E. A., Soudry, D., Gao, Y., Machado, T. A., Merel, J., et al. (2016). "Simultaneous denoising, deconvolution, and demixing of calcium imaging data". *Neuron* 89(2), 285–299.

Pnevmatikakis, E., Merel, J., Pakman, A., and Paninski, L. (2013). "Bayesian spike inference from calcium imaging data". *In Signals, Systems and Computers* (2013 Asilomar Conference). Pacific Grove, CA, 349–353.

Polson, N. G., Scott, J. G., and Windle, J. (2013). "Bayesian inference for logistic models using Pólya-gamma latent variables". *Journal of the American Statistical Association* 108(504), 1339–1349.

Prado, R. and West, M. (2010). *Time Series: modeling, computation, and inference*. 1st. Chapman and Hall.

Rand, W. M. (1971). "Objective criteria for the evaluation of clustering methods". *Journal of the American Statistical Association* 66(336), 846–850.

Redish, A. D., Battaglia, F. P., Chawla, M. K., Ekstrom, A. D., Gerrard, J. L., et al. (2001). "Independence of firing correlates of anatomically proximate hippocampal pyramidal cells". *Journal of Neuroscience* 21(5).

Richardson, S. and Green, P. J. (1997). "On Bayesian analysis of mixtures with an unknown number of components (with discussion)". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 59(4), 731–792.

Robert, C. and Casella, G. (2010). *Introducing Monte Carlo methods with R*. Springer.

Rodríguez, A. and Dunson, D. (2011). "Nonparametric Bayesian models through probit stick-breaking processes". *Bayesian Analysis* 6, 145–178.

Rodríguez, A., Dunson, D. B., and Gelfand, A. E. (2008). "The nested Dirichlet process". *Journal of the American Statistical Association* 103(483), 1131–1154.

Rosenbaum, R., Smith, M. A., Kohn, A., Rubin, J. E., and Doiron, B. (2017). "The spatial structure of correlated neuronal variability". *Nature neuroscience* 13(1), 107–114.

Sethuraman, J. (1994). "A constructive definition of Dirichlet priors". *Statistica Sinica* 4(2), 639–650.

Stan Development Team (2021). *Stan modeling language users guide and reference manual*. URL: `mc-stan.org/`.

Stephens, M. (2000). "Bayesian analysis of mixture models with an unknown number of components—an alternative to reversible jump methods". *The Annals of Statistics* 28(1), 40–74.

Sun, W., Reich, B. J., Cai, T. T., Guindani, M., and Schwartzman, A. (2015). "False discovery control in large-scale spatial multiple testing". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 77(1), 59–83.

Teerapabolarn, K. (2012). "The least upper bound on the Poisson-negative binomial relative error". *Communications in Statistics - Theory and Methods* 41(10), 1833–1838.

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2006). "Hierarchical Dirichlet processes". *Journal of the American Statistical Association* 101(476), 1566–1581.

van der Pas, S., Szabó, B., and Vaart, A. van der (2017). "Adaptive posterior contraction rates for the horseshoe". *Electronic Journal of Statistics* 11(2), 3196–3225.

Vogelstein, J. T., Packer, A. M., Machado, T. A., Sippy, T., Babadi, B., et al. (2010). "Fast nonnegative deconvolution for spike train inference from population calcium imaging". *Journal of Neurophysiology* 104(6), 3691–3704.

Wade, S. and Ghahramani, Z. (n.d.). "Bayesian cluster analysis: point estimation and credible balls (with discussion)". *Bayesian Analysis* (2) (), 559–626.

Walker, S. G. (2007). "Sampling the Dirichlet mixture model with slices". *Communications in Statistics - Simulation and Computation* 36(1), 45–54.

Wei, X.-X., Zhou, D., Grosmark, A., Ajabi, Z., Sparks, F., et al. (2019). "A zero-inflated gamma model for post-deconvolved calcium imaging traces". *bioRxiv: 637652*.

Laura D'Angelo

CURRICULUM VITAE

Contact Information

Department of Economics, Management and Statistics (DEMS)
University of Milano-Bicocca
Piazza dell'Ateneo Nuovo 1,
20126 Milan, Italy.

e-mail: laura.dangelo@unimib.it

Current Position

from December 2021
Research fellow in Statistics
Department of Economics, Management and Statistics (DEMS)
University of Milano-Bicocca

Research interests

- Statistical modeling
- Bayesian statistics
- Bayesian nonparametrics
- Computational statistics

Education

October 2018 - November 2021
PhD in Statistical Science, University of Padova.
Thesis title: Bayesian modeling of calcium imaging data
Supervisor: Prof. Antonio Canale
Co-supervisor: Prof. Michele Guindani.

October 2015 – November 2017
Master's degree (*laurea specialistica/magistrale***) in Statistical Science**.
University of Padova, Department of Statistics
Title of dissertation: "Modelli Bayesiani nonparametrici: applicazioni al settore assicurativo "
Supervisor: Prof. Antonio Canale
Final mark: 110/110 cum laude.

October 2012 – September 2015
Bachelor's degree (*laurea triennale***) in Statistics, Economics and Finance**.
University of Padova, Department of Statistics
Title of dissertation: "L'area sotto la curva ROC specifica per caratteristiche"
Supervisor: Prof. Gianfranco Adimari
Final mark: 110/110 cum laude.

Visiting periods

January 2020 – November 2020
University of California, Irvine;
Irvine, CA (USA).
Supervisor: Prof. Michele Guindani.

Work experience

January 2018 – September 2018
BIP S.p.A..
Consultant, data analyst.

Awards and Scholarship

July 2021 - ISBA 2021 World Meeting.
ISBA 2021 Best Student/Postdoc Contributed Paper Award.

Computer skills

- R (advanced)
- C++, GitHub, Python (working level)

Language skills

Italian: native;
English: fluent.

Publications

Articles in journals
D'Angelo L., Canale A., Yu Z. and Guindani M. (2021) Bayesian nonparametric analysis for the detection of spikes in noisy calcium imaging data. *Biometrics* (in press).

D'Angelo L. and Canale A. (2021) Contributed Discussion on: "Centered partition processes: informative priors for clustering", in *Bayesian Analysis*, **16**(1), 356–358.

D'Angelo L., Canale A., Yu Z. and Guindani M. (2021) Detection of neural activity in calcium imaging data via Bayesian mixture models, in *Book of Short Papers SIS 2021* (Editors: Perna C., Salvati N., Schirripa Spagnolo F.), ISBN: 9788891927361.

D'Angelo L. (2019) Model based clustering in group life insurance via Bayesian nonparametric mixtures, in *Book of Short Papers SIS 2019* (Editors: Arbia, G., Peluso, S., Pini, A. and Rivellini, G.), ISBN: 978889191510.

Working papers
D'Angelo L. and Canale A. (2021) Efficient posterior sampling for Bayesian Poisson regression. *arXiv preprint arXiv:2109.09520*

Conference presentations

D'Angelo L., Canale A., Yu Z. and Guindani M. (2021). Bayesian nonparametric analysis for the detection of spikes in noisy calcium imaging data. (contributed talk) *JSM 2021*, August 8 - 12, 2021.

D'Angelo L., Canale A., Yu Z. and Guindani M. (2021). Bayesian nonparametric analysis for the detection of spikes in noisy calcium imaging data. (contributed talk) *ISBA 2021 World Meeting*, June 23 - July 2, 2021. Pre-recorded video available at https://youtu.be/SLLSJVuFnMs .

D'Angelo L., Canale A., Yu Z. and Guindani M. (2021). Detection of neural activity in calcium imaging data via Bayesian mixture models. (contributed talk) *SIS 2021 Intermediate meeting*, Pisa, Italy, June 21 - 25, 2021.

D'Angelo L. (2019). Model based clustering in group life insurance via Bayesian nonparametric mixtures. *SIS 2019 Intermediate meeting*, Milan, Italy, June 12-14, 2019.

Teaching experience

April 2021
Tirocinio formativo
use of Latex for scientific writing, 2.5 hours
University of Padova

Other Interests

Member of LIPU (Lega Italiana Protezione Uccelli) since 2019.
Volunteer for the project Lipu LIFE Choo-na in 2019.

References

Prof. Antonio Canale
University of Padova
via Cesare Battisti, 241-243;
35121 Padova. Italy.
Phone: +39 049 827 4168
e-mail: canale@stat.unipd.it

Prof. Michele Guindani
University of California, Irvine
Donald Bren School of Information and Computer Sciences; Irvine, CA 92697-1250
Phone: +1 949 824 3276
e-mail: michele.guindani@UCI.edu