UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Sede Amministrativa: Università degli Studi di Padova

Dipartimento di Scienze Statistiche
Corso di Dottorato di Ricerca in Scienze Statistiche
Ciclo XXXIV

Thesis Title

Coordinatore del Corso: Prof. Nicola Sartori

Supervisore: Prof. Antonio Canale

Co-supervisore: Prof. Michele Guindani

Dottoranda: Laura D'Angelo

Day Month Year of submission

Abstract

Abstract content.

Sommario

Contenuto del sommario.

Contents

List of Figures

List of Tables

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, 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 with its output remains an open problem of research.

Main contributions of the thesis

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

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

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 c_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) assume that all spikes have a fixed amplitude, and interpret the parameter A_t as the *number* of spikes at time t. Following this definition, they place 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 search 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 were also 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 interpret 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) develop fast algorithms to compute the spike trains under these assumptions. Jewell and Witten (2018) assert 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) propose 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 define the parameter A_t as the *amplitude* of a spike at time t, taking values in the non-negative real numbers. They formulate 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 do not explicitly assume sparsity of the spikes.

1.2 Data sets

Una frase introduttiva? Parlo dei dati dell'Allen Brain Observatory e poi ci sarebbe da mettere i nuovi dati se riesco a fare qualcosa del progetto 3...

1.2.1 Allen Brain Observatory data

The Allen Brain Observatory (Allen Institute for Brain Science, 2016) is a public large 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 and layers, utilizing transgenic Cre lines to drive expression of genetically encoded fluorescent calcium indicators, and 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 is placed in front of a screen where different types of visual stimuli are shown, while the mouse's neuronal activity is 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 in the visual areas respond to stimuli of different complexity. Specifically, each neuron in the visual cortex can be characterized by their *receptive field*, i.e. the features of the visual stimulus that trigger the signaling of that neuron. Hence, it is of critical interest to devise methods that allow inferring how the neuronal response varies under the different types of visual stimuli.

1.2.2 Altri dati?

Paragrafo qui.

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.

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 within each of them a relatively simple model can be applied.

Denote the observed data as a vector of n units $\boldsymbol{y} = (y_1, \dots, y_n)$; also, assume that the n observations are exchangeable, meaning that the joint probability density $p(y_1, \dots, y_n)$ 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 $k = 1, \dots, K$. Usually 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 indicator S_{ik} of component k for observation i, with

$$S_{ik} = \begin{cases} 1 & \text{if } y_i \text{ is drawn from component } k \\ 0 & \text{otherwise.} \end{cases}$$

Given $\boldsymbol{\pi}$, the distribution of $\boldsymbol{S}_i = (S_{i1}, \ldots, S_{iK})$ is $\text{Multinomial}(1; \pi_1, \ldots, \pi_K)$. Conditionally on $\boldsymbol{S}_i$, the data distribution of y_i is simply $p(y_i \mid \boldsymbol{S}_i, \boldsymbol{\theta}^*) = \prod_{k=1}^{K} f(y_i \mid \theta_k^*)^{S_{ik}}$; moreover, given $\boldsymbol{S} = (\boldsymbol{S}_1, \ldots, \boldsymbol{S}_n)$, the y_i are assumed to be independent. The joint density of the observed data and the unobserved indicators, conditionally on the model parameters, can now be written as

$$p(\boldsymbol{y}, \boldsymbol{S} \mid \boldsymbol{\theta}^*, \boldsymbol{\pi}) = p(\boldsymbol{y} \mid \boldsymbol{S}, \boldsymbol{\theta}^*)\, p(\boldsymbol{S} \mid \boldsymbol{\pi}) = \prod_{i=1}^{n} \prod_{k=1}^{K} \{\pi_k f(y_i \mid \theta_k^*)\}^{S_{ik}}.$$

Having defined the data distribution, 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, $\boldsymbol{\pi} \sim \text{Dir}_K(\alpha_1, \ldots, \alpha_K)$.

This model also admits a useful hierarchical representation. Rewriting the latent allocation variables using the cluster indicators $c_i \in \{1, \ldots, K\}$, with $c_i = k$ if y_i belongs to the k-th mixture component (i.e. $S_{ik} = 1$), the model is, for $i = 1, \ldots, n$

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

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 $\{\theta_1^*, \ldots, \theta_K^*\}$. The model then becomes, for $i = 1, \ldots, n$

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

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 conditionals after imputing the cluster indicators $\mathcal{C} = \{c_1, \ldots, c_n\}$ 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

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

   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. For this short review of some of the main properties of the DP we followed the unpublished report by Yee Whye Teh, which illustrates in a clear and simple way the fundamental properties of this process.

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 $G \sim \mathrm{DP}(\alpha, G_0)$, if for any

partition $\{B_1, \ldots, B_H\}$ of Θ

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

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 $G \sim \mathrm{DP}(\alpha, G_0)$, 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

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

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

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

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 $G \sim \mathrm{DP}(\alpha, G_0)$. 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 $G \sim \mathrm{DP}(\alpha, G_0)$, then it can be expressed as follows:

$$\begin{aligned} \beta_k \sim \mathrm{Beta}(1, \alpha), \quad \theta_k^* \sim G_0 \qquad & \text{for } k \geq 1 \\ \pi_1 = \beta_1, \quad \pi_k = \beta_k \prod_{l=1}^{k-1} (1 - \beta_l) \qquad & \text{for } k \geq 2 \\ G = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k^*}. & \end{aligned} \tag{1.4}$$

The construction of the weights $\{\pi_k\}_{k=1}^{\infty}$ by means of Beta random variable is usually called stick-breaking process. 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 inference procedures, as for example the algorithms described in the following to sample from the posterior distribution of Dirichlet process mixtures.

Dirichlet process mixtures

Getting back to the framework of mixture models, from the representation of the process as discrete measure it is straightforward to extend the finite mixtures described in the previous section to nonparametric mixtures: following the structure of Equation (1.3), DP mixtures (DPMs) can be written as

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

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, the prior expected number of clusters in a sample of size n 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 clusters. 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 $\beta_k \sim \text{Beta}(1 + n_k, \alpha + \sum_{h=k+1}^{K} n_h)$ for $k = 1, \ldots, K-1$. Finally, letting $\beta_K = 1$ 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

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

where $\mathbb{I}(A)$ is the indicator variable, taking value 1 if condition A is satisfied and 0 otherwise. 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. With all these elements, the joint density of (y_i, u_i, c_i) conditioned on $\boldsymbol{\pi}$ and $\boldsymbol{\theta}^*$ becomes

$$p(y_i, u_i, c_i \mid \boldsymbol{\pi}, \boldsymbol{\theta}^*) = \mathbb{I}(u_i < \xi_{c_i}) \frac{\pi_{c_i}}{\xi_{c_i}} f(y_i \mid \theta_{c_i}^*).$$

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 (1.3.2) with beta random variables $\beta_k \sim \text{Beta}(1 + n_k, \alpha + \sum_{l=k+1}^{K} n_l)$.

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 Mixtures of finite mixtures

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 mixtures of finite mixtures (MFMs) has long been hindered by the difficulty of performing posterior inference. Several inference methods have been proposed for this type of models (McCullagh and Yang, 2008; Nobile, 2004; Nobile and Fearnside, 2007; Richardson and Green, 1997) 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.

More recently, different models have been proposed inspired by nonparametric mixtures. Miller and Harrison (2018) explicitly derived MFMs counterparts for several properties exhibited by DPMs. They considered a finite mixture model similar to the one defined in section 1.3.1, where however the number of clusters is random and is assigned a prior distribution. A main limitation of their approach is that it assumes a fixed parameter α for the Dirichlet distribution, regardless of the dimension K, which can lead to undesired

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 all three 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.* (2020): 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

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

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 $\mathrm{Dir}_K(1, \dots, 1)$ is used regardless of the value of K, leads to a prior expected number of clusters $E(K_+)$ close to $E(K)$ 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.* (2020) 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.* (2020) 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 $\mathcal{C} = \{c_1, \dots, c_n\}$ 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.* (2020), 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

$$\pi_1, \ldots, \pi_K \mid K, \alpha, \mathcal{C} \sim \mathrm{Dir}_K(\alpha/K + n_1, \ldots, \alpha/K + n_K).$$

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, $i = 1, \ldots, n$, we also observe a categorical variable g_i with values in $\{1, \ldots, J\}$ 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

$$G_1, \dots, G_J \mid Q \sim Q, \qquad Q = \sum_{k=1}^{\infty} \pi_k \delta_{G_k^*} \tag{1.6}$$
$$G_k^* = \sum_{l=1}^{\infty} \omega_{l,k} \delta_{\theta_{l,k}^*}$$

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 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 atom 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. 1.6), 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.

Posterior inference for the CAM

Chapter 2

Efficient posterior sampling for Bayesian log-linear models

2.1 Algorithm

Chapter 3

Bayesian nonparametric analysis for partially exchangeable single-neuron data

3.1 Model specification

3.2 Posterior computation

Bibliography

Allen Brain Observatory (2017) Technical whitepaper: stimulus set and response analyses.

Allen Institute for Brain Science (2016) Allen brain observatory. Available at: http://observatory.brain-map.org/visualcoding.

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.

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

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.

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* pp. 1–12.

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. In *The new brain sciences: perils and prospects*, eds S. Rose and D. Rees, pp. 167 – 180. Cambridge University Press.

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

Ferguson, T. S. (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. In *Advances In Neural Information Processing Systems*, eds D. Lee, M. Sugiyama, U. Luxburg, I. Guyon and R. Garnett, pp. 1984 – 1992.

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

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., Malsiner-Walli, G. and Grün, B. (2020) Generalized mixtures of finite mixtures and telescoping sampling.

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

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

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.

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

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.

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.

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.

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.

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.

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

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

Pnevmatikakis, E. A., Soudry, D., Gao, Y., Machado, T. A., Merel, J., Pfau, D., Reardon, T., Mu, Y., Lacefield, C., Yang, W., Ahrens, M., Bruno, R., Jessell, T. M., Peterka, D. S., Yuste, R. and Paninski, L. (2016) Simultaneous denoising, deconvolution, and demixing of calcium imaging data. *Neuron* **89**(2), 285 – 299.

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.

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

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

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.

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

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