Tallinn University of Technology
Mathematics and Natural Science Department
Institute of Cybernetics

THE BUILDING BLOCKS OF THE NEURAL CODE-- STATISTICAL MODELING OF CO-OCCURRENCE PATTERNS IN NEURAL SPIKE TRAINS

Master's Thesis

Maria Kesa

Supervisor: Margus Pihlak, Mathematics department, professor
Co-supervisors: Raul Vicente Zafra, Tartu University, professor
Michael J. Berry, Princeton Institute of Neuroscience, Associate professor

Engineering Physics

2017

TALLINNA TEHNIKAÜLIKOOL

Matemaatika-loodusteaduskond

Küberneetika instituut

STATISTILINE MUDEL NEURONITE NÄRVIIMPULSSIDE KAASESINEMISE MUSTRITE MODELLEERIMISEKS

Magistritöö

Maria Kesa

Juhendaja: Margus Pihlak, Matemaatika õppetool, dotsent
Kaasjuhendaja: Raul Vicente Zafra, Tartu Ülikool, Andmeteaduse professor
Michael J. Berry, Princetoni Neuroteaduste Instituut, profesor

Tehniline Füüsika
2017

Deklareerin, et käesolev lõputöö on minu iseseisva töö tulemus ning kinnitan, et esitatud materjalide põhjal ei ole varem akadeemilist kraadi taotletud.

Kinnitan, et antud töö koostamisel olen kõikide teiste autorite seisukohtadele, probleemipüstitustele, kogutud arvandmetele jmt viidanud.

Maria Kesa

Juhendaja: *Margus Pihlak*

Töö vastabmagistritööle esitatavatele nõuetele.

Kaitsmiskomisjoni esimees:

Lubatud kaitsmisele

……

(nimi, allkiri, kuupäev)

Table of Contents

Introduction

The meaning of the activity of a single neuron depends on the circuit context in which it occurs (Kruskal et al, 2013, Carrillo-Reid, 2015). Cell assemblies are a distinct feature of information processing in the brain (Harris, 2005). Cell assemblies are groups of cells that co-activate in time. They form ensembles whose properties cannot be explained by studying the firing of individual neurons (Miller et al, 2014). Individual cells can participate in multiple assemblies, which enriches the coding capacity of the circuit (Miller, 2014). Furthermore, whilst coding by individual neurons is unreliable, multidimensional codes of population level firing patterns, are more robust (Montijn et al, 2016). In this thesis we use a statistical machine learning model called Latent Dirichlet Allocation (LDA) to automatically identify cell assemblies from data. The model can identify a mixture of assemblies active during a response. We apply this analysis to publicly available calcium imaging data from the Allen Brain Observatory (Allen Brain Observatory, 2017). Our modeling approach provides a principled way to quantify cell assembly activations in neural data sets.

Allen Brain Observatory data is an example of a striking new development in neuroscience. The developments in the field of neuroinformatics coalesce towards making large-scale neural recordings publicly available, so that scientists can mine the data to extract pieces of information that are relevant to their theories and

research. With the introduction of large-scale neural datasets comes the need for algorithms that can extract useful information and thus make browsing the data faster and more insightful. Crucially this involves summarizing the data in a meaningful way. This thesis is concerned with applying a model that has been successfully used to model the statistics of large text corpora and construct interfaces for data navigation and searching, to neuroscience data.

Latent factor models have been extensively used in neuroscience to model the shared low-dimensional structure that generates the variability in high-dimensional data (Buesing et al, 2012). These models have been very successful in decoding applications for the Brain-Machine Interface. Recently, hidden Markov models have also been successfully used to model the responses in retinal population data (Prentice et al, 2016). These methods have a temporal dimension, they use latent variables that influence the evolution of the system in time. LDA is not a temporal model, although variants have been developed to extend it to the temporal domain (Wang and McCallum, 2006). We argue that LDA can still be useful in the neuroscience domain, because there are applications where time as a dimension is not important. For example, one might wish to construct simple summaries of experiments without having to follow the temporal ordering of the stimuli that were applied. Discarding the temporal dimension may also be beneficial for analysis if the data contains a lot of inter-trial variability, leading to low correlations between the recorded time series'. In those situations one may simply want to know which cell assemblies activated during stimulus presentations, not necessarily which order they were activated in. Averaging over time is also intrinsic to sensory discrimination tasks, where pooling responses over time increases the signal-to-noise ratio (Berry, communications). LDA is a useful model to have in your toolbox for the analysis and construction of summaries that are handy for certain data mining tasks on large volumes of experimental data.

Novelty

While our ability to record from a large number of neurons keeps improving, the algorithmic question of how to extract knowledge from this data is open to investigation. Here we use a statistical learning approach, that has been successful at compressing a large amount of text data into interpretable distributions, on neural calcium imaging data. This model has never been applied to large neural datasets and exploring its utility in this context is completely novel.

1. Background

1.1. Calcium Imaging

Calcium is an extremely important intracellular messenger in mammalian neurons (for a review see Grienberger, 2012). Calcium influx and efflux into the cell and its release by internal stores determine its cellular concentration. Calcium influx happens mainly through voltage-gated calcium channels, ionotropic glutamate receptors, nicotinic acetylcholine receptors (nAChR), and transient receptor potential type C (TRPC) channels. Calcium is pumped out of the cell by the plasma membrane calcium ATPase (PMCA) and the sodium-calcium exchanger (NCX). Calcium release from internal stores is mediated by inositol trisphosphate receptors and ryanodine receptors. In this thesis the important events that lead to calcium level elevations in the cell are the openings of voltage gated channels during synaptic activity and action potential firing. This influx of calcium is a signal that can be detected and we next explain how.

Intracellular calcium levels can be detected by calcium indicators. These are molecules that change their fluorescence properties (e.g. emit light in a narrow band) when they bind calcium. There are two main kinds: chemical indicators and genetically encoded calcium indicators. The experiments that this thesis is based on used a genetically encoded calcium indicators. Mice can be genetically engineered to express a fluorescent protein selectively in certain cells. During the experiment,

when calcium enters the cell and binds to the expressed fluorescent protein, it changes its fluorescence properties and thus the wavelength at which it emits light and this can be detected with a microscope. The fluorescence changes in the cell can be recorded over time.

Calcium elevation in cells is necessary for vesicle release in action potential firing. The movies of fluorescence changes during the experiment have to be transformed into a time series of fluorescence changes (called dF/F, typically defined as $\frac{(F - F_b)}{F_b}$ where F_b is the baseline of the fluorescence). Spikes cause an elevation in the dF/F signal above a baseline. They can be detected either by simple thresholding (the strategy used in this thesis), deconvolution algorithms or supervised machine learning algorithms, trained on ground-truth data with intracellular recordings (Vogelstein et al, 2010).

In summary, action potentials in large populations of cells (hundreds, even thousands) can be detected from changes in fluorescence through calcium imaging, because action potential firing depends on an influx of calcium into the cell. When calcium binds to a fluorescent indicator it emits a light signal that can be capture with a microscope and a video camera. Calcium imaging thus permits to image patterns of neural firing in populations across time.

1.2. Latent Dirichlet Allocation

1.2.1. Developing domain analogies for applying Latent Dirichlet Allocation to neural data

Latent Dirichlet Allocation (LDA) was initially invented to model the statistical structure of large text collections. It is a solution to the question, 'How can we use co-occurrence patterns of words in documents to summarize large amounts of text and detect clusters in the data?'. However, since its inception, this algorithm has been also used in other fields where there is a need for finding a probabilistic description of high-dimensional data sets, for example genomics (Liu et al, 2010, Shivashankar et al, 2011). Here we apply Latent Dirichlet Allocation to neural data.

So what is LDA? LDA is an unsupervised Bayesian generative model that employs latent variables (variables that influence the observed data, but are not measured) to capture the statistical patterns in observed data. We first describe the model in the context in which it was created, e.g. modeling text corpora and then describe a mapping from the original application domain to the domain of neural data sets.

LDA extracts topics from data. Topics describe what a document is about and, in the framework of LDA, they are probability distributions over words. For example, a topic that can be called 'Statistics', will place high probability on words such as 'Gaussian', 'Multivariate', 'p-value'. Each document is described by a distribution over latent topics. Generative models specify the probabilistic process that is assumed to have generated the data and in LDA this process is: for each word slot in

the document, sample a topic from the topic distribution describing a document and then sample a word from the topic.

In this thesis we define a cell assembly to be a probability distribution over the cells imaged in the experiment. This probability describes how likely a cell is to spike. Different cell assemblies correspond to different groups of neurons that are likely to spike together. During each stimulus presentation, different cell assemblies can co-activate and thus we must learn the distribution over cell assemblies for each recording. In table 1 we establish analogies between the model applied to text and applied to spiking data.

Text domain	**Neuronal spiking domain**
a document	the recording during a given stimulus presentation
a word in a document	a cell spiking
topic	a cell assembly
a distribution over topics	which cell assemblies are co-active during a particular stimulus presentation

Table 1. Domain transfer from text to spiking data.

The power of probabilistic models lies in their generality. The same mathematical model or framework can often be used to understand different phenomena. We thus exploit a model that has been developed in one context to explore the statistical structure in a different data modality.

1.2.2. Mathematical description of Latent Dirichlet Allocation

We first describe mathematically the assumptions for how the data was generated, e.g. the generative process, and then describe how to invert the model to go from data to inferring the model parameters.

To generate each response r, we first describe which cell assemblies are active in that response. The distribution describing what cell assemblies are active in a response π_k^r, with $k = 1, .., K$ form a multinomial distribution. They have a Dirichlet prior with concentration hyperparameters α_k, which is a conjugate prior of the multinomial distribution. Conjugate prior for the likelihood function means that the posterior is in the same family of distributions as the prior, which simplifies inference. The Dirichlet distribution is given by:

$$Dirichlet(\pi|\alpha) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_0)...\Gamma(\alpha_K)} \prod_{k=1}^{K} \pi_k^{\alpha_k - 1},$$

where

$$\alpha_0 = \sum_{i=1} \alpha_i \alpha_i$$

$$\Gamma(x) = \int_0^\infty u^{x-1} e^{-u} du.$$

The generative process begins by sampling a π_k^r from the prior, which describes what cell assemblies are active in the response. This induces a multinomial distribution over cell assemblies, from which we then sample a particular cell assembly, c_i^r. Finally, we sample the occurrence of a spike of a particular neuron, s_i^r from the distribution over cells d conditioned on the cell assembly. To do this we use a row corresponding to the cell assembly c_i^r in a $K \times M$ matrix d, whose entries represent the spiking probabilities of M neurons in K cell assemblies. α and β are

parameters of the Dirichlet distribution. The generative process thus proceeds as follows :

$$\pi^r \sim Dirichlet_K(\alpha)$$

$$c_i^r \sim Multinomial_K(\pi^r)$$

$$d \sim Dirichlet_{K \times M}(\beta)$$

$$s_i^r \sim Multinomial_M(d|c_i^r)$$

The following procedure generates one spike s_i from a particular cell. To generate the entire response of the cells to the experimental condition (e.g. the presentation of a visual stimulus), we repeat this procedure N times for the number of spikes in the response.

The likelihood of a particular response can be decomposed as follows (Blei et al, 2003):

$$p(r) = \int_\pi \left(\prod_{n=1}^{N} \sum_{c_n=1}^{k} p(s_n|c_n;\beta) p(c_n|\pi) \right) p(\pi|\alpha) d\pi$$

In summary, to generate a response, we sequentially sample from a distribution over cell assemblies and then conditioned on the cell assembly, we sample a spike from the distribution over cells. This is a generative process that models how a particular response was generated. The next step is to invert the model and perform inference, e.g. to learn the distributions over cell assemblies and the distributions over cells from data.

1.2.3. Inference in Latent Dirichlet Allocation

The aim of inference is to obtain a posterior distribution over the latent variables and parameters given the data.

$$p(\pi, c|s, \alpha, \beta) = \frac{p(\pi, c, s|\alpha, \beta)}{p(s|\alpha, \beta)}$$

This posterior is intractable, because it assumes a summation over all possible settings of latent variables and parameters. There is a variety of algorithms that can be applied to probabilistic graphical models with latent variables. Broadly, they fall into two classes-- Markov Chain Monte Carlo (MCMC) sampling and variational Inference. The MCMC algorithms construct a Markov chain that has the target posterior distribution as its stationary distribution. Variational inference approximates the posterior by searching for a simpler distribution in a parametrized family that best explains the data.

Here we briefly outline the variational inference algorithm that is implemented in the library that we used for fitting the model in this thesis (for a more extensive coverage, see Blei et al, 2003).

We first define a family of variational distributions, q.

$$q(\pi, c|\gamma, \phi) = q(\pi|\gamma) \prod_{n=1}^{N} q(c_n|\phi_n)$$

γ is a Dirichlet parameter and ϕ is a multinomial parameter and they are the free variational parameters.

The key idea of variational inference algorithms is to use Jensen's inequality to obtain a tractable bound on the log-likelihood and to use this bound to find the optimal free variational parameters, e.g. the ones that make the bound as tight as possible. The bound is constructed as follows:

$$
\begin{aligned}
&\log p(s|\alpha,\beta) = \log \int \sum_c p(\pi, c, s|\alpha, \beta) d\pi \\
&= \log \int \sum_c \frac{p(\pi, c, s|\alpha, \beta) q(\pi, c|\gamma, \phi)}{q(\pi, c|\gamma, \phi)} d\pi \\
&\geq \int \sum_c \log p(\pi, c, s|\alpha, \beta) d\pi - \int \sum_c q(\pi, c|\gamma, \phi) \log q(\pi, c|\gamma, \phi) d\pi \\
&= E_q[\log p(\pi, c, s|\alpha, \beta)] - E_q[\log q(\pi, c|\gamma, \phi)]
\end{aligned}
$$

This allows us to formulate an optimization problem to find the parameters γ and ϕ that minimize this bound:

$(\gamma*, \phi*) = argmin_{\gamma,\phi} KL(q(\pi, c|\gamma, \phi)||p(\pi, c|s, \alpha, \beta))$

This objective minimizes the Kullback-Leibler divergence between the posterior and the variational distribution q. This minimization is done via an iterative fixed point method.

In summary, in order to learn the approximate posterior distribution of the topic proportions and the topic distributions (cell assembly proportions and cell assemblies) given the training data, we construct a simpler variational distribution with its own parameters and then find the parameters that minimize a lower bound on the log-likelihood.

2. Methods

2.1. Data from Allen Brain Observatory

The data in this study is obtained from a public data set on primary visual cortex (V1) from the Allen Brain Observatory. In this data set calcium recordings were made in different transgenic mouse lines and in different cortical depths in response to visual stimuli such as gratings, natural images and movies.

In the Allen experiments data is available for 6 different transgenic mouse lines. Here we restrict our analysis to a particular transgenic mouse line Emx1-IRES-Cre with the GCaMP6f genetic calcium indicator and Cam2ka-tta tetracycline controlled trans-activation protein (the calcium indicator expression depends both on the activity of Cre recombinase and the tetracycline controlled transactivator protein (ttA)). This mouse line expresses Cre in the excitatory neurons in the neocortex, permitting fluorescent imaging of neural populations therein.

The depths that were recorded from are 175, 250, 275 and 375 microns. We analyze the data from depth 275 microns. For the particular mouse and depth that we analyze, there are 222 neurons in the imaged population.

The mice were presented with random noise patterns, gratings, moving gratings, natural images and natural movies. The data is essentially a long dF/F signal for every cell, which describes the change of fluorescence over time and a table that

records which images are shown at particular time points. The order of the images during the presentation is randomized. Each image is shown for 250 ms at the rate 32 Hz (8 timepoints for each image presentation). There are 50 repetitions of each image, with blank screen shown approximately after every 25 images. Here we limit our analysis to the data to natural images. The natural images consisted of 118 black and white images with pictures of landscapes, animals and natural patterns.

In summary, while data from Allen Brain Observatory is quite extensive, we limit our analysis to a particular mouse cell line, imaged depth and class of stimuli. However, our analysis strategy can be equally well applied to other subsets of the data.

2.2. Data Pre-Processing

Allen Brain Observatory data contains dF/F traces of calcium signals for the experiments. We use a simple strategy for determining whether the cell fired in a particular sampling time bin. We thresholded the signal at two values, 0.1 and 0.3 (dF/F is dimensionless) to obtain whether a spike was fired in the bin, thereby binarizing the data. The threshold 0.1 approximately corresponds to the emission of an action potential for the fluorescent protein, GCaMP6f (Berry, communication). However, the data is noisy (see the section "Variability in Data") and we therefore also use a higher threshold to get a signal with higher confidence.

Because the data is sampled at 32 Hz (e.g. each image presentation is 8 timepoints within 250 ms) and one action potential lasts for 1 ms, it is possible that there are multiple spikes in one time point. We don't take this into account as we are

interested in whether the cell spiked in a time window resolved by the setup of the experiment, but not its rate, e.g. how many times it spiked.

We chose the method of simple thresholding over deconvolution algorithms and supervised machine learning algorithms due to its simplicity. Complex preprocessing steps introduce artefacts that are difficult to intuit and may also require setting of parameters by hand (for example extracting spikes from a deconvolved trace also requires the setting of a threshold by hand). We prefer to keep the analysis as close to the data as possible, because the modeling is then more interpretable.

For this thesis we filtered out 30 (out of 222) most active cells for the topic model analysis. This permitted us to carry out certain calculations without the need for parallelization, e.g. on a desktop computer. In particular, we could fit a large number of models in acceptable time for selecting the optimal number of topics for describing the data (see the next section). Subsetting the data also permits more compact visualizations more appropriate for a thesis. Web-based automated interactive visualizations (for example, using tools such as d3js) can of course permit clear visualization for a much larger number of cells. Also, the most active neurons have highest signal-to-noise ratio (Koch et al, 2004) and therefore give more robust results. With more computational resources and time, larger models involving more cells can be computed. Topic models trained on text corpora often have a vocabulary size of tens of thousands of words. Future research will reveal whether LDA is robust enough for modeling larger populations of cells.

2.3. Latent Dirichlet Allocation applied to Allen Brain Observatory data

In order to apply LDA, we first pool all of the data together. Each document or response is whether a spike occurred in a time point in the experiment (there were 8 time points for every image presentation) among the 30 most active cells. We obtain a data matrix of dimensionality (47200, 30). We have two of these data matrices, one corresponding to threshold 0.1 and the other to the threshold 0.3. After filtering out response vectors that are all zero, we obtain dimensionality (47184, 30) for threshold 0.1 and (46260, 30) for threshold 0.3.

The main modeling question in applying LDA is how many topics should you use to model the data. We used a metric from (Arun et al, 2010), implemented in the topic modeling library TomLib. The method fits a range of models (here from 5-30 topics) and gives each a score based on a symmetric Kullback-Leibler divergence measure. The method uses the fact that LDA can be seen as a matrix factorization of the data matrix into two matrices, the document-topic matrix, which has the dimensions $T \times W$ and the document-topic matrix, which has the dimensions $D \times T$, where T is the number of topics (cell assemblies), W is the number of words (cells) and D is the number of documents (responses). Both matrices can be used to proportion of topics assigned to the corpus and the authors use this fact to derive that the distributions of the singular values of the matrix topic-word matrix and the row L1 norm of the topic-document matrix, should be similar at the number of topics that best describe the corpus (in the sense of the symmetric Kullback-Leibler divergence). When applied to real data, the measure grows initially

and then shows a strong dip and then increases again. The location of the dip is at the number of topics that best describe the corpus.

We trained this method on held-out set of 10000 response vectors. The results from this method is given in figures 1 and 2. The first graph (corresponding to threshold 0.1) shows a dip at 11 topics. The second graph (corresponding to threshold 0.3) displays a dip at 6 topics. These are sensible number of topics for a small vocabulary of 30 cells. We used these numbers to subsequently train a topic model with that number of topics on the rest of the data.

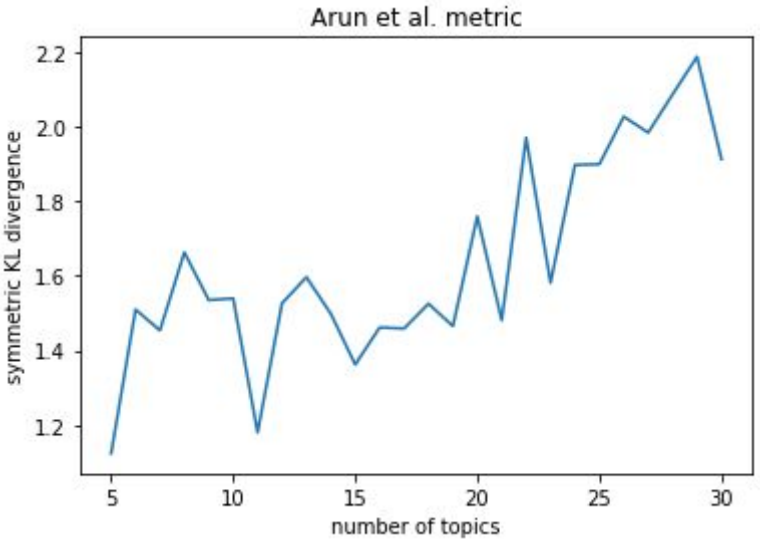

Figure 1. A metric for determining the appropriate number of topics, data was thresholded at 0.1.

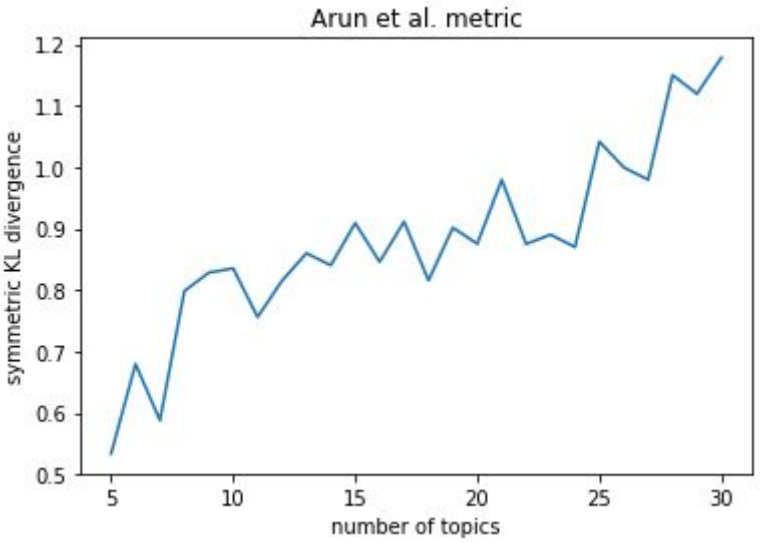

Figure 2. A metric for determining the appropriate number of topics, data was thresholded at 0.3

After choosing an appropriate number of topics, we used Scikit-Learn library in Python to fit the LDA model.

3. Results

3.1. Variability in the data

Neural responses in the cortex are variable (Gerstner, 2014). Figure shows the dF/F signal for 10 trials in response to the same image in four different cells. Figure 3 illustrates that the responses are variable.

Binarizing the fluorescence signal and construct a peristimulus time histogram (PSTH) (figures 4 and 5) makes it apparent that the cells have no strong temporal preference when the stimulus is shown and fire the same number of spikes on average.

Across trials, cell firing is highly variable (figures 6 and 7) even for the same stimulus presentation. This indicates that there is a lot of irreproducibility in the data set, which makes modeling more difficult.

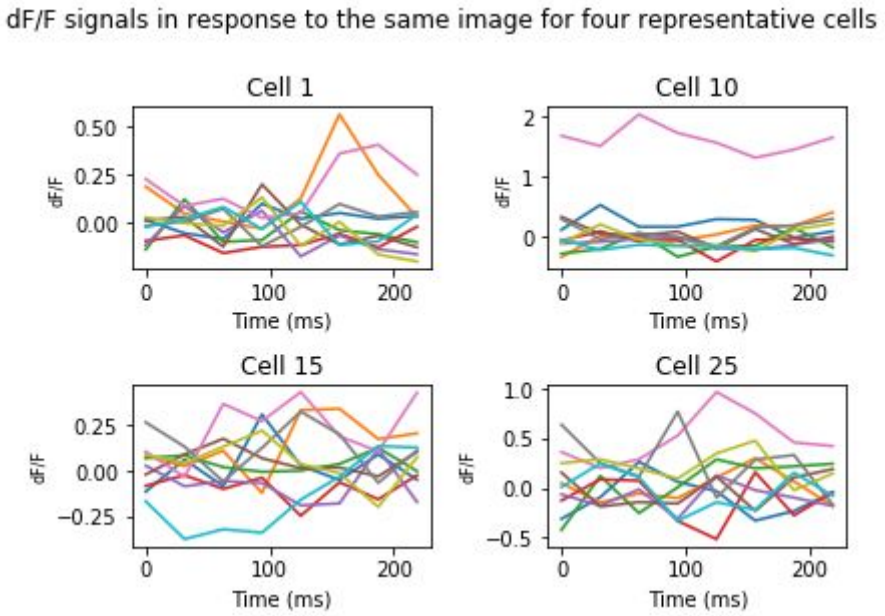

Figure 3. dF/F signal for representative cells

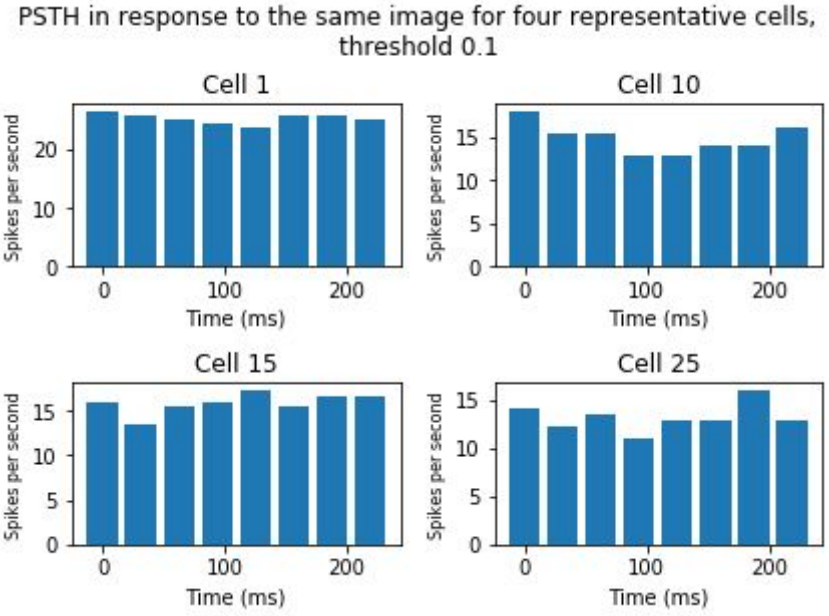

Figure 4. Peristimulus time histograms for four representative cells, threshold 0.1

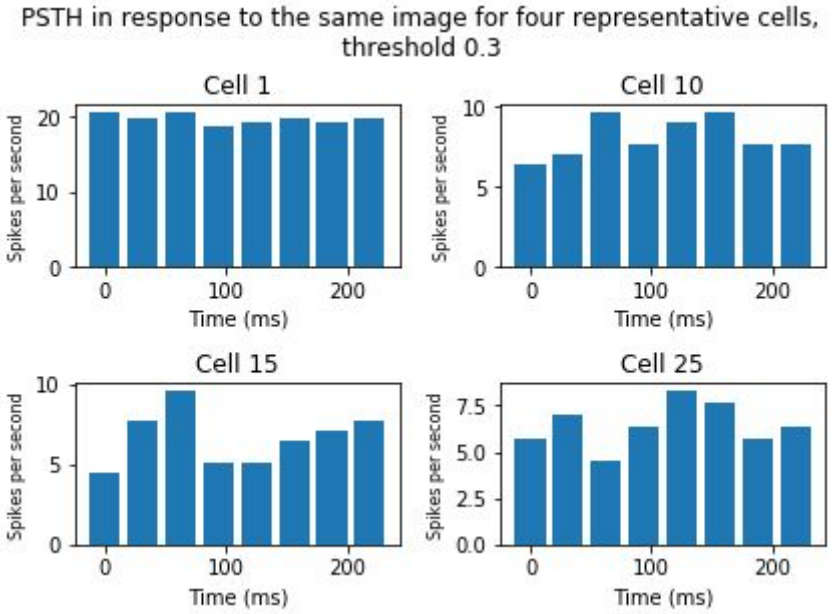

Figure 5. Peristimulus time histogram for four representative cells, threshold 0.3

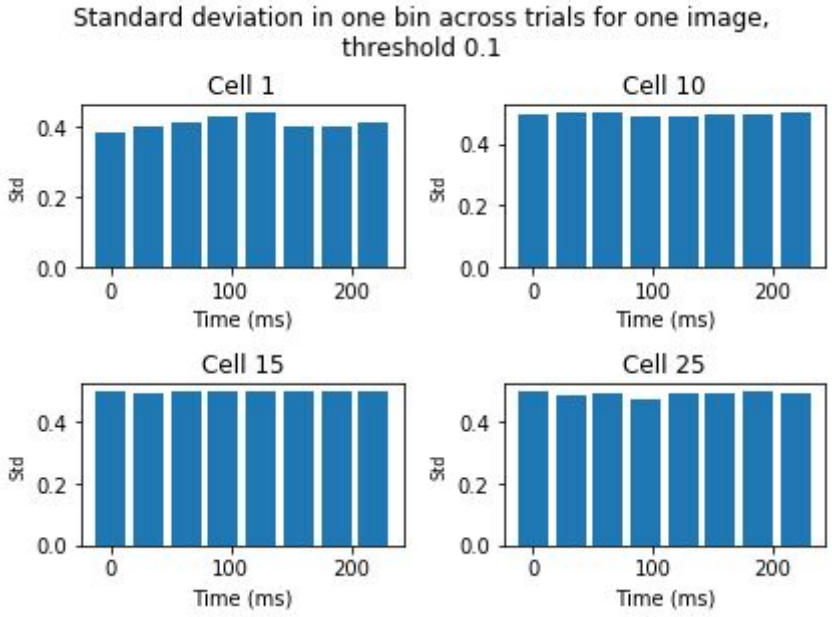

Figure 6. Standard deviation of thresholded cell signals across time points in the stimulus for one stimulus, threshold 0.1.

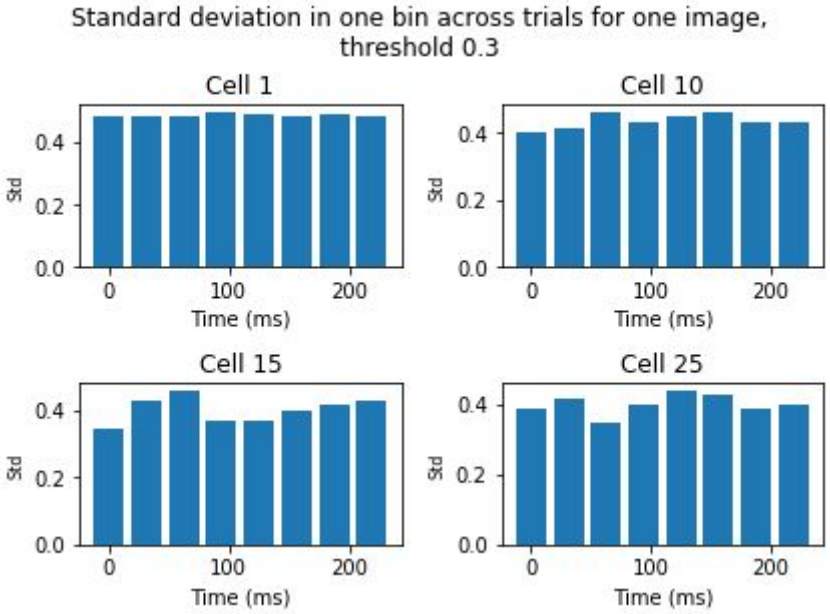

Figure 7. Standard deviation of thresholded cell signals across time points in the stimulus for one stimulus, threshold 0.3.

3.2. Applying Latent Dirichlet Allocation to calcium imaging data

We fit LDA to data thresholded at 0.1 and 0.3 and plot the resulting topics (cell assemblies) in figure 8 and 9. The topics in 0.3 binarized data look much more useful, e.g. they aren't as broad as some of the topics in 0.1 binarized data and are more sparse. This means that the data, where there are less spikes, resolves into more meaningful topics, e.g. there is less noise and irreproducibility.

We also plot the distribution over cell assemblies (topics) for a sample of responses (threshold 0.3) in figure 10. These distributions are strongly peaked, indicating that at each time frame particular cell assemblies are activated.

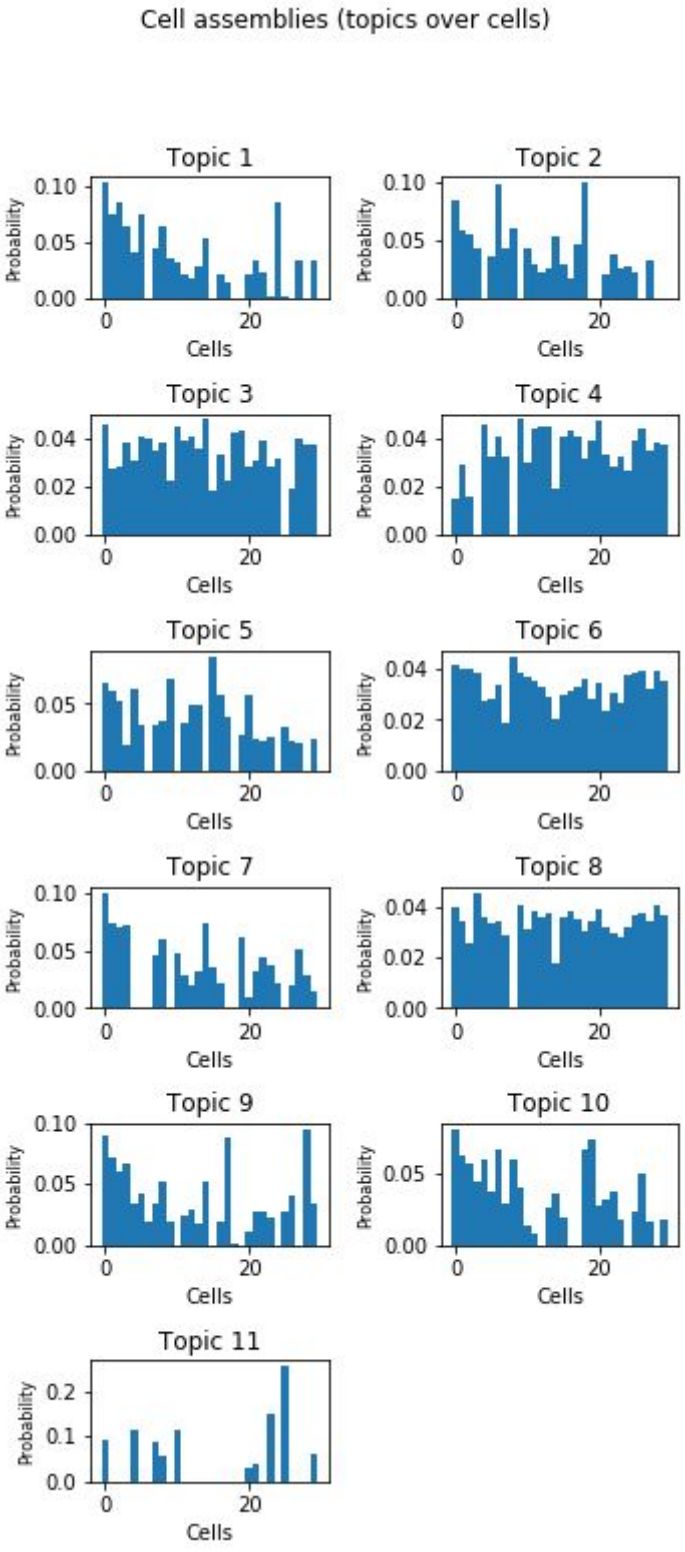

Figure 8. Topics (cell assembly distributions) for data thresholded at 0.1.

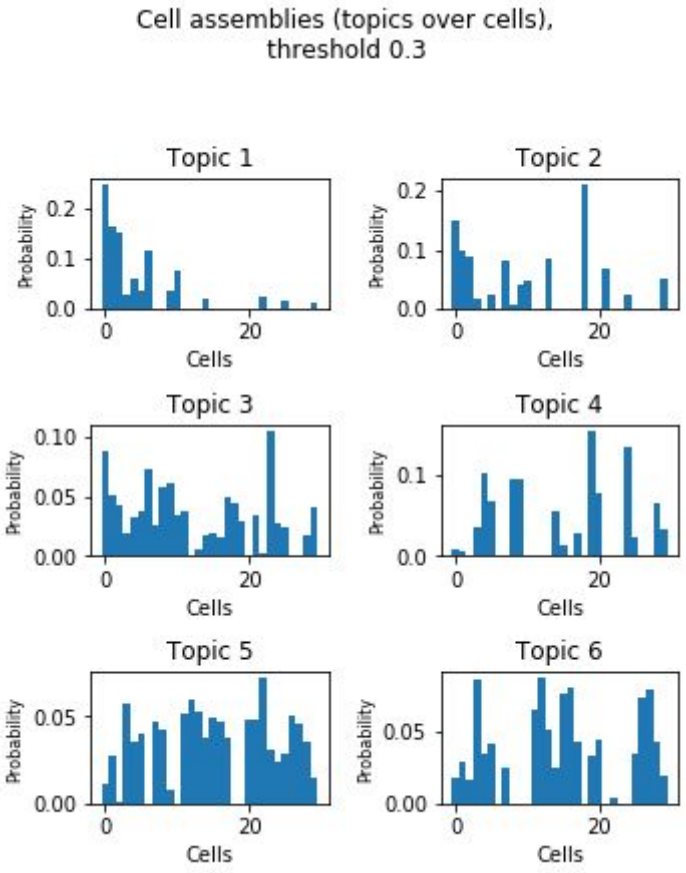

Figure 9. Topics (cell assembly distributions) for data thresholded at 0.3.

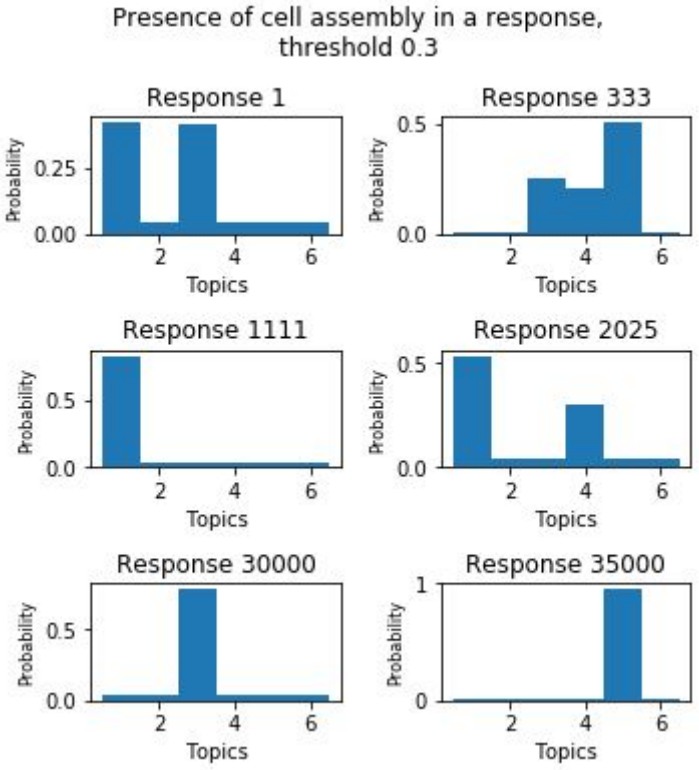

Figure 10. Cell assemblies (topics) distributions for representative cell responses.

Summary

We applied a statistical modeling framework, LDA, to calcium imaging data in a particular layer of the mouse V1 neocortex in response to natural image stimuli. We used a method based on spectral analysis to determine a good number of topics to use in the modeling (Arun, 2010). We found that cell assemblies, e.g. groups of co-activating cells can be extracted from the data and give more or less reasonable topic distributions depending on the threshold chosen to binarize the data. Finding the optimal parameter for thresholding is thus an important step in data modelling and future work will focus on how to best determine this parameter.

This method could be very useful for compressing experimental data into interpretable distributions for exploratory analysis of the data. Each time point in the experiment can be summarized by the cell assemblies that were most strongly active. These distributions derived from LDA compactly capture the modules of the neural code, e.g. groups of co-firing neurons.

Limitations and future work

On the technical side, our analysis shows sensitivity to the threshold that we used to binarize the data. A more principled way to choose the threshold for binarization would thus be a valuable addition to the modeling methodology. For example, one could make the threshold adaptive by taking into account the individual statistics of

each of the cell's activity. Inventing novel metrics for assessing which threshold parameter best describes the data is also left to future work. One idea would be to use log-likelihood to see which parameters place the highest probability on the data given the model, but this does not assess the quality of the topics, e.g. how sparse and interpretable they are. One measure of the goodness of the topic could be its Kullback-Leibler divergence from a uniform distribution, however this measure is maximized when the topic peaks at one cell. Designing an appropriate measure of the usefulness of the topic distribution is thus left to future work.

Alternatively, one could also extract the rates of firing from the data directly (Ganmor, 2016). The topic model also works when there are multiple occurrences of the same word in the document, e.g. a cell spikes several times. We expect that extracting rates from the data could present a viable alternative to simple binarization.

On the conceptual side, topic models as summaries of texts have clear semantics. People are often able to give a name to a topic after viewing which words it places a high probability on. What knowledge, insight and meaning can be gleaned from knowing the cell assemblies extracted from data, however? Answering this question is left to future work, where applications of this modeling methodology will be explored.

Here we used simple visualizations of the topics and their distributions in responses. The framework of LDA permits constructing interactive visualizations that enable an enhanced browsing experience of the data (PyLDAvis, 2017). Applying these more advanced visualization tools to Allen Brain Observatory data is future work.

Kokkuvõte

Käesolevas töös rakendati statistilist mudelit Latentne Dirichlet Allokatsioon (Latent Dirichlet Allocation) suurele hulgale kaltsiumi pildistamise tehnika andmetele Alleni Aju Observatooriumist (Allen Brain Observatory, 2017), millest tuletati närviimpulsside esinemist. Andmed tulenesid eksperimentidest, milles mõõdeti esmase visuaalse ajukoore rakkude reaktsiooni looduslikele piltidele.

Statistiline mudel grupeeris edukalt kokku rakud, millel olid sarnased närviimpulside kaasesinemise mustrid. Statistilise mudeldamise tulemused sõltusid tugevalt andmestikule rakendatud lävest, mida kasutati andemete diskreetimiseks (närviimpulside tuletamiseks).

Tuleviku ülesanneteks (mis jäävad väljapoole käesolevast tööst) on leida kvantitatiivne mõõt, mis iseloomustab kui kasulik ja hea on närviimpulsside kaasesinemise jaotus, et kindlaks teha milline lävi on parim diskreetimiseks.

Resümee

Käesolevas töös rakendati statistilist mudelit-- Latentne Dirichlet Allokatsioon (Latent Dirichlet Allocation), et kvantitatiivselt uurida neuronite närviimpulsside kaasesinemise mustreid. Andmestikus leiduvad statistilised mustrid võimaldasid edukalt tihendada andmeid, esitades need jaotustena, mis näitasid millised rakud tõenäoliselt koos aktiveeruvad. Iga rakkude reaktsioon stiimuli ühele ajapunktile oli siis esitatav jaotusena üle rakkude närviimpulsi kaasesinemise jaotuste, sest modelleerimise metoodikaks oli hierarhiline Bayes'i mudel. Andmete tihendus ehk iseloomustus läbi mudeli prisma võimaldas esitada suurt andmekogu kompaktselt. Antud statistiline meetod võib aidata teadlastel teha uurimuslike analüüse andmetest (exploratory analysis), et kvantifitseerida statistilised seaduspärasused, mis andmestikus esinevad.

Citations

Allen Brain Observatory, http://observatory.brain-map.org/visualcoding/, Last accessed on 31 May, 2017

Arun, R., Suresh, V., Veni Madhavan, C.E., Narasimha Murthy, M.N. (2010). On finding the natural number of topics with Latent Dirichlet Allocation: Some observations. In Advances in Knowledge Discovery and Data Mining. PAKDD 2010. Lecture notes in Computer Science, vol 6118. Springer, Berlin, Heidelberg.

Berry, M.J. (2017). Spoken communication.

Blei, D.M., Ng, A.Y., Jordan, M.I. (2003). Latent Dirichlet Allocation. Journal of Machine Learning Research 3, 993-1022.

Buesing, L., Macke, J., Sahani, M. (2012). Learning stable, regularised latent models of neural population dynamics. Network: Computation in Neural Systems, 1-2.

Carillo-Reid, L., Miller, J.-K., Hamm, J.P., Jackson, J., Yuste, R. (2015). Endogenous sequential cortical activity evoked by visual stimuli. The Journal of Neuroscience 35-23, 8813– 8828

Ganmor, E., Krumin, M., Rossi, L.F., Carandini, M., Simoncelli, E.P. (2016). Direct estimation of firing rates from calcium imaging data. ArXiv pre-print, q-bio.

Gerstner, W., Kistler, W., Naud, R., Paninski, L. (2014). Neuronal dynamics: from single neurons to networks and models of cognition. Cambridge University Press.

Grienberger, C., Konnerth, A. (2012). Imaging calcium in neurons. Neuron 73, 862-885.

Harris, K.D. (2005). Neural signatures of cell assembly organization. Nature Reviews Neuroscience 6, 399-407.

Koch, K., McLean, J., Berry, M., Sterling, P., Balasubramanian, V., Freed,M. (2004). Efficiency of information transmission by retinal ganglion cells. Current Biology 14, 1-20.

Kruskal, P.B., Li, L., MacLean, J.N. (2013). Circuit reactivation dynamically regulates synaptic plasticity in neocortex. Nature Communications 4, 2574.

Liu, B., Liu, L., Tsykin, A., Goodall, G., Green, J., Zhu, M., Kim, C.H., Li, J. (2010). Identifying functional miRNA-mRNA regulatory modules with correspondence latent dirichlet allocation. Bioinformatics 26-24, 3105-3111.

Miller, J.-K., Ayzenshtat, I., Carillo-Reid, L., Yuste, R. (2014). Visual stimuli recruit intrinsically generated cortical ensembles. Proceedings of National Academy of Science, early edition, 1-9.

Montijn, J.S., Meijer, G.T., Lansink, C.S., Pennarzt, C.M.A. (2016). Population-level neural codes are robust single-neuron variability from a multidimensional coding perspective. Cell Reports 16, 2486–2498.

Prentice, J.S., Marre, O., Ioffe, M.L., Loback, A.R., Tkacic, G., Berry, M.J.B. (2016). Error-robust modes of the retinal population code. PLOS Computational Biology.

PyLDAvis, https://github.com/bmabey/pyLDAvis, last accessed 28 May, 2017

Shivashankar, S., Srivathsan, S., Ravindran, S., Tendulkar, A. (2011). Multi-view methods for protein structure comparison using latent dirichlet allocation. Bioinformatics 27-13, 161-168.

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

Wang, X., McCallum, A. (2006). Topics over time: a non-markov continuous-time model of topical trends. In: Conference in Knowledge Discovery and Data Mining, New York, NY, USA, ACM Press, 424–433