Neural Topic Modelling

Pamela Hathway (p.hathway16@imperial.ac.uk)
Department of Electrical and Electronic Engineering, Imperial College London
South Kensington Campus, London SW7 2AZ, UK

Dan F. M. Goodman (d.goodman@imperial.ac.uk)
Department of Electrical and Electronic Engineering, Imperial College London
South Kensington Campus, London SW7 2AZ, UK

Abstract

We introduce neural topic modelling - an unsupervised, scalable and interpretable neural data analysis tool which can be applied across different spatial and temporal scales. The aim is an approach that can handle the ever-increasing number of neurons recorded by high channel count multi-electrode arrays. Neural topic modelling is based on latent Dirichlet allocation, a method routinely used in text mining to find latent topics in texts. The spike trains are converted into "neural words" - the presence or absence of discrete events (e.g. neuron 1 has a higher firing rate than usual). Neural topic modelling results in a number of topics (probability distributions over words) which best explain the given co-occurrences of neural words over time. Applied to an electrophysiological dataset of mouse visual cortex, hippocampus and thalamus neurons, neural topic modelling groups neural words into topics which exhibit common attributes such as overlapping receptive fields or proximity on the recording electrode. It recovers these relationships despite receiving no knowledge about the cortex topography or about the spatial structure of the stimuli. Choosing neural activity patterns as neural words that are relevant to the brain makes the topics interpretable by both the brain and researchers, setting neural topic modelling apart from other machine learning approaches.

Keywords: multi-electrode recordings; electrophysiology; machine leaning; topic modelling

Introduction

Recent advances in neuronal recording techniques have allowed researchers access to large datasets of neuronal activity. In addition, electrophysiological recordings are often made in freely moving animals carrying out complex behaviours during multiple experimental paradigms (Stringer et al., 2018). Therefore researchers are confronted with the task of combining neural data on different temporal and spatial scales with experimental and behavioural variables. How to go about the analysis of such rich datasets is a huge challenges for neuroscientists and current analysis methodologies.

There are several main challenges for any new methodology of neural data analysis. 1) New analyses need to be scalable for larger numbers of neurons, preferably for hundreds or thousands of neurons. This is where e.g. correlation-based analysis methods will soon reach their limits (the number of correlations to consider increases quadratically for pair-wise correlations and even more so with higher correlations). 2) Recordings are often done across different temporal and spatial scales (e.g. a combination of electrophysiological recording, Ca2+ imaging, local field potentials, etc.) and new analyses should be able to handle all of them and potentially even allow combining their analyses. 3) The application of advanced machine learning techniques to neural data have provided researchers with fascinating results, but at the same time it is unclear whether the structure found in the data using these results corresponds to something that could actually be used by the brain itself, and it is therefore unclear how the results should be interpreted. Creating analysis methods that are interpretable (both by the brain and by researchers) would therefore be helpful to the field and further our understanding of how the brain encodes information.

We propose a new approach to these challenges: neural topic modelling, a neural data analysis tool based on latent Dirichlet allocation (LDA) (Blei, Ng, & Jordan, 2003). LDA is routinely used in text mining to find latent topics hidden in texts (e.g. in newspaper articles or tweets). In neural topic modelling, neural data is converted into the presence or absence of discrete events (e.g. neuron 1 has a higher firing rate than usual), which we call "neural words". This creates a universal framework for neural data on different temporal and spatial scales since any type of data can be translated into this common format. Basing the neural words on neural activity patterns that are detectable by downstream neurons ensures that the input to neural topic modelling is relevant to the brain.

Neural topic modelling provides an ideal starting point for the analysis of large-scale neural datasets. As an unsupervised tool, it uncovers relationships between neurons that we expect to find (visual receptive fields in visual cortex neurons), but also relationships that are more surprising (neurons in the visual cortex and thalamus exhibiting the same receptive field). It is therefore a promising method to inform researchers of possible interesting analysis directions.

Methods

To demonstrate the validity of neural topic modelling we analysed an electrophysiological dataset of 367 neurons recorded with a Neuropixel electrode (Lopez et al., 2016) from a head-fixed mouse viewing spatial noise (sparse black and white squares on grey background). The dataset includes neurons located in the visual cortex, the hippocampus and the thala-

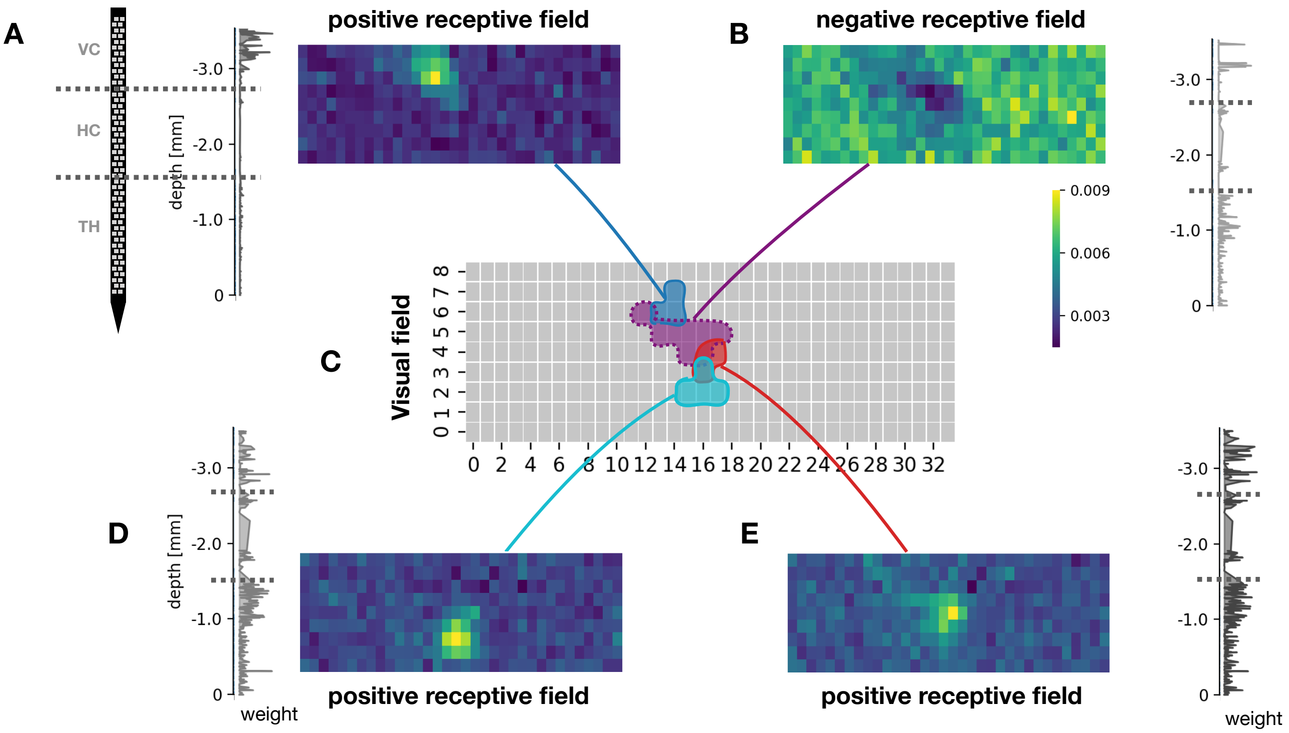

Figure 1: **Example topic receptive fields**. Examples of positive receptive fields (weighted mean probability) (**A**, **D**, **E**) and a negative receptive field **C**. Positive receptive field in **A** only includes neural words from neurons in the visual cortex on the top part of the electrode (left), whereas the other receptive field show no such preference towards a particular brain region. The colour indicated the probability of a word happening given the stimulus location. **B**) An example of a negative receptive field. The receptive fields were masked at 0.8 of maximum value of weighted mean probability maps and overlaid on the visual field where black and white squares were shown during the experiment.

mus.

LDA receives as input a number of documents and the words present in those documents. To create the "neural documents" the recording was divided into time windows during which at least one square was present. The spikes were translated into two simple neural word types: 1) increased firing rate in neuron i, 2) decreased firing rate in neuron i. The creating of neural words in this manner means that each neuron in the recording can give rise to many neural words.

LDA assumes that the words are distributed across documents based on a set of distinct topics - a topic being a probability distribution over words (e.g. word A has a 0.01 probability to be in topic 0 and a 0.3 probability to be in topic 1). Therefore LDA gives an estimation of a set of topics which explain the given occurrences of words in the documents.

Since the LDA algorithm can is prone to finding local minima, we ran 100 iterations with different random seeds. The resulting topics from the 100 iterations were then clustered using K-Means, resulting in the topic clusters that are each comprised of very similar topics (similar probability distributions over neural words). For each topic cluster we calculated the weighted mean probability distribution over words.

Results

Neural topic modelling is a neural data analysis tool which results in groups of neural words based on common characteristics such as preferred stimulus appearance location or location within the same brain region.

The results from applying LDA 100 times to the dataset were clustered to form topic clusters and their homogeneity was confirmed visually (data not shown). Each topic cluster was made up of topics found during separate LDA runs. Topic clusters were comprised of between 16 and 100 topics, where a topic cluster of size 100 means that that particular topic (probability distribution over neural words) was found in every single LDA run. For each topic cluster we related the neural words to the locations of their respective neurons on the electrode and calculated the weighted mean probability distribution over the ten neural words with the highest weights. To visualise the receptive fields of a topic cluster, we investigated the relationship between the topic clusters and stimulus location.

About 1/3 of topic clusters exhibited a concentration of neural words from a spatially limited region on the electrode (0.4

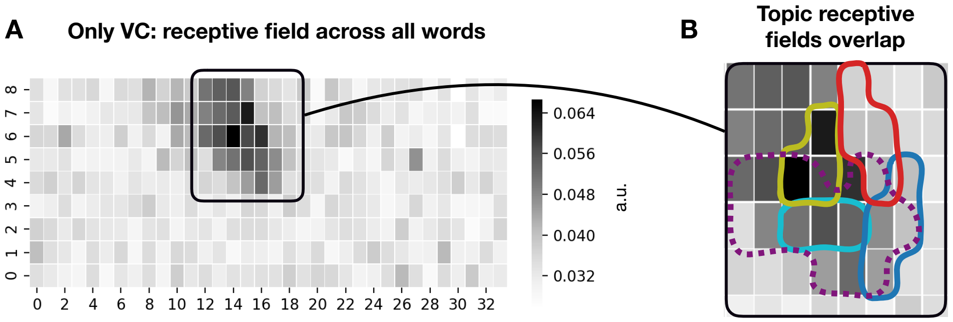

Figure 2: **Topic receptive fields**. **A**) The probability of a word happening given the stimulus location across all words. **B**) The weighted mean probability for four topics with positive receptive fields (solid lines) and one topic with a negative receptive field (dashed line). Overlap of positive and negative receptive fields from different topics masked at 0.8 of max value of weighted mean probability maps.

- 1 mm) (for an example see Figure 1A). The borders of these regions correspond well with the boundaries between visual cortex and hippocampus, and hippocampus and thalamus. Some topic clusters are even further limited to only a subregion of the visual cortex or the thalamus , possibly reflecting topic clusters being restricted to neural words from neurons of a particular cortical layer or thalamic nucleus.

Several topic clusters exhibit a clear pattern of either a small region to which the words in the topic respond preferentially (positive receptive field, see Figure 1A, D, E) or a small region to which none of the neurons respond to preferentially (negative receptive field, see Figure 1B). Visual receptive fields can be found also in about 1/3 of all topic clusters, and some of them also exhibit a limitation to a particular brain region as detailed above (Figure 1A). We show four mostly non-overlapping receptive fields, three with positive receptive fields and one with a negative receptive field.

Running the same analysis for only the visual cortex neurons results in topic clusters that are only sensitive to the region shown in Figure 1A. Interestingly, when concentrating the analysis on this smaller subset of neurons, the topic clusters correspond to separate, mostly non-overlapping sub-regions of a larger receptive field (see Figure 2B). Additionally, some topic clusters were brightness-sensitive and reacted exclusively to black squares but not white squares (data not shown), something that was not seen in other topics and not found when using the whole dataset.

We verified that the weights within a topic were not solely driven by number of times a word occurred or the word order in the input (data not shown).

Discussion

Neural topic modelling discovers distinct topics - groups of neural words (neural activity patterns) - in which the neural words are similar in their preferred stimulus location and/or spatial proximity on the recording electrode without having to search for these characteristics explicitly. In a dataset including visual cortex, hippocampus and thalamus we found topics (groupings of neural words) with visual receptive fields in each of the three brain regions.

Some of the topics found were largely expected (visual receptive fields in visual cortex) but other less so (visual receptive fields in hippocampus, thalamus, and spanning visual cortex and thalamus). Other topics reflected the local connectivity of the brain regions, since neurons located close to each other were grouped into topics together. Without more detailed information on the exact location of the electrode it is not possible to confirm whether similar neural activity is based on synaptic connections or functional connectivity without a direct connection. Further knowledge of the placement of the electrode among the layers and nuclei will only enhance the quality of the interpretation of the results.

The visual receptive fields found in this data set are close to each other if not slightly overlapping in relation to the visual field. This can be explained by the design of the Neuropixel electrode with which the recording was made. The electrode is long, but very thin, and therefore only records from few neurons in each layer of the brain and very probably along a visual cortex column. It is therefore not surprising that the receptive fields of the neurons in the visual cortex are in a similar location of the visual field. It is nevertheless interesting that the neuronal words from the deeper regions show a slightly shifted preferred region.

One focus of neural topic modelling is the potential interpretability of the topics by both the brain and the researchers. If the neural words in a given analysis are chosen so that they can be detected by downstream neurons, then each topic as a combination of neural words will be detectable by the brain as well. In addition, the resulting topics are interpretable by researchers in terms of their combinations of neural words (e.g. a mixture of similar or dissimilar neural words) and in terms of the physical origins of those neural words (neural words from neurons in the same or different brain regions). Importantly, the methodology circumvents the combinatorial explosion of possible neuron combinations by converting neural activity into relevant events, resulting in a scalable method for very large datasets. The general approach to the data also means that many different types of neural data recordings can be analysed without the need to have all datasets on the same spatial or temporal scale.

In addition to the straightforwardly interpretable topics found by neural topic modelling outlined above, there remain several topics that are not connected to the two characteristics we were searching for (visual receptive field and spatial proximity on electrode). This does not mean that these topics are meaningless; on the contrary, it is possible that we have not looked for the correct experimental variable for that topic. We expect to be able to find topics that are sensitive to visual stimulus orientation and movement, other sensory experiences or the animals' behaviour. It will be interesting to apply the model to more complex datasets e.g. in behaving mice with more advanced experimental paradigms, or to datasets where the neural representation of the stimulus structure is less clear e.g. for olfactory experiments.

Conclusion

Neural topic modelling is an unsupervised analysis tool that receives no knowledge about the cortex topography nor about the spatial structure of the stimuli, but is able to recover these relationships anyway. The method provides an ideal starting point for the analysis of large-scale neural datasets by highlighting possible interesting directions for further analysis. The combination of scalability, applicability across temporal and spatial scales and the biological interpretability of neural topic modelling sets this approach apart from other machine learning approaches to neural data analysis. We will make neural topic modelling available to all researchers in the form of a Python software package.

Acknowledgments

This work was partly supported by the Engineering and Physical Sciences Research Council and The Royal Society (grant RG170298).

References

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

Lopez, C. M., Mitra, S., Putzeys, J., Raducanu, B., Ballini, M., Andrei, A., . . . Yazicioglu, R. F. (2016, jan). Paper 22.7 A 966-electrode neural probe with 384 configurable channels in 0.13um SOI CMOS. In *Digest of technical papers - ieee international solid-state circuits conference.* IEEE.

Stringer, C., Pachitariu, M., Steinmetz, N., Reddy, C. B., Carandini, M., & Harris, K. D. (2018). Spontaneous behaviors drive multidimensional, brain-wide population activity. *bioRxiv*.