Cross-session stability analysis and invariant feature extraction of MEA data

Mario Surlemont*

*surlemontm95@univie.ac.at

universität wien

Introduction

In the field of neuroscience research, the accurate recording and analysis of neural populations is pivotal. Session-to-session variability presents a significant challenge in this domain, often manifesting in recordings through several key causes. Primarily, the continuity of data can be disrupted by the loss of neurons initially present in the recording array. This is compounded by the potential replacement of these neurons by previously unrecorded ones, which introduces new variables into the dataset.
Furthermore, mechanical shifts in the probe array can lead to systematic changes in neuron positions, affecting the consistency of recorded signals. Such shifts can drastically alter the topography of neural recordings, thereby necessitating adjustments in data interpretation.
The stabilization and standardization of neural recordings are crucial for the advancement of our understanding of neural dynamics. Through this poster, we explore methodologies and analytical frameworks for cross-session stability analysis and invariant feature extraction from multi-electrode array (MEA) data. Our goal is to pave the way for more reliable long-term studies of neural populations.

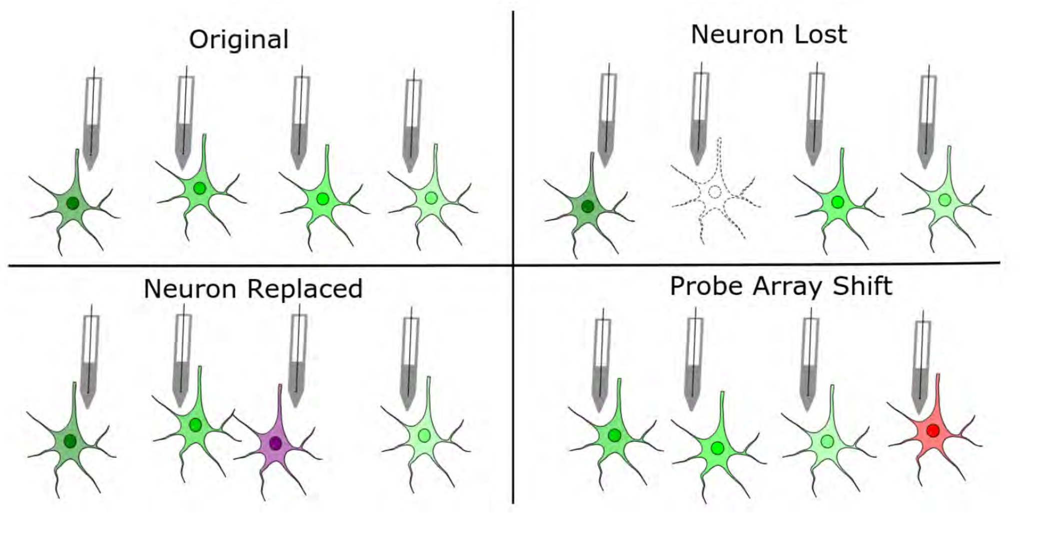
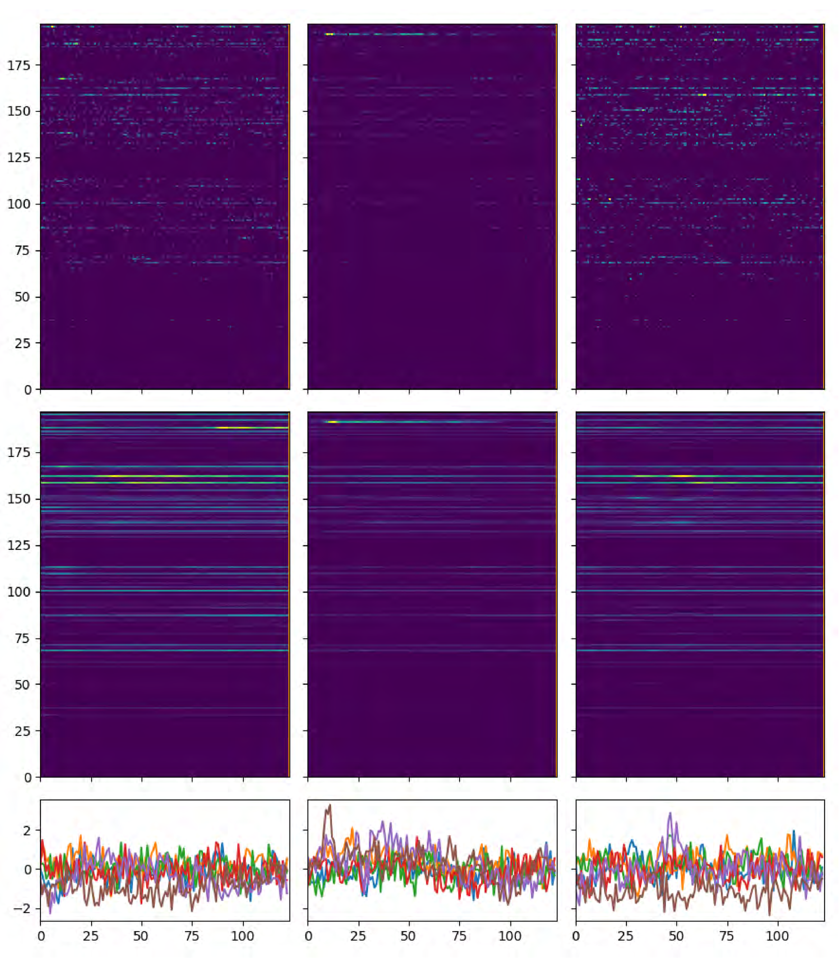


Figure 1 of Jude et al. (2022)

This project involves data from a patient who suffered a stroke and now experiences word-finding difficulties. To aid in her recovery:

- This patient had a Multi-Electrode Array (MEA) implanted, which can capture electrical impulses from nerve cells.
- She undergoes training sessions where she is shown images. Those images can be associated with one of two types of words (nouns or verbs) and one of three semantic categories (body, house, or animal)
- Data are recorded through the MEA during these sessions.

Reconstruction LFADS



The top row displays three randomly selected trials from a session. The heat map represents individual electrodes of the MEA on the Y-axis and time on the X-axis (each point represents 20ms). Each cell of the heat map indicates the frequency at which an electrical activity threshold was exceeded on the respective electrode within the time bin in that time interval. The second row shows the reconstructed signals, while the third row presents the firing rates as line charts, with each electrode represented by one line. It is evident that the reconstructions are significantly closer to a trial-specific average value than the original data. Nevertheless, certain spike patterns can still be recognized in the reconstructions.

Methods

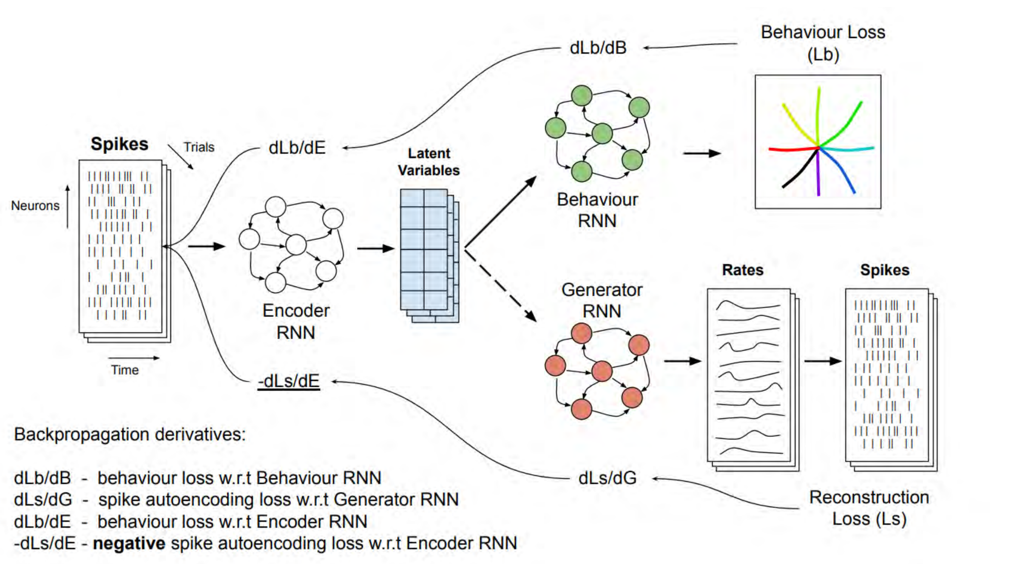


Figure 2 of Jude et al. (2022)

Our investigation into the robustness of neural data recordings across multiple sessions hinges on the utilization of computational models that aim to extract stable and invariant features. Central to our project are two models: the Latent Factor Analysis via Dynamical Systems (LFADS) Pandarinath et al. (2018) and the SABLE Jude et al. (2022) model.
LFADS (Latent Factor Analysis via Dynamical Systems) is a deep learning model that infers the underlying dynamics from high-dimensional and noisy datasets, notably neural spiking data. It utilizes a variational autoencoder to compress observed data into a lower-dimensional latent space that captures essential dynamical features. At its core, an RNN models the temporal evolution of these latent dynamics. The encoder maps high-dimensional input data to initial conditions for the dynamical system, while the decoder reconstructs the observed data from the latent states. LFADS separates different sources of variability in the data, distinguishing noise from meaningful dynamical variations.
The SABLE model is designed for aligning neural activity across different recording sessions without requiring recalibration for behavior decoding. It utilizes unsupervised domain adaptation and a sequential variational autoencoder framework. In its essence it is very similar to the LFADS model and extends it with a particular domain adaption technique.
Initially, we tried to evaluate and implement solutions building on the LFADS base model incorporating different domain adaption techniques. Our first approach was similar to Hurwitz et al. (2021). As this one did not yield promising results, we attempted to adapt the SABLE model to derive a cross-session invariant representation of our data.
However, while we could achieve a good training accuracy, the model failed to generalise. Suprisingly, when doing a sanity check on the original dataset of the model publication we also failed to reproduce the results.
Due to the lack of promising results from domain adaptation technologies, we opted to explore whether the original LFADS model could correctly align the data in a shared space to prepare it for classification. Additionally, we investigated the reasons behind the models' failures, questioning if the probabilistic assumptions of the models, including the VAE posterios, were unsuitable for extracting meaningful information from our data and reconstructing it.

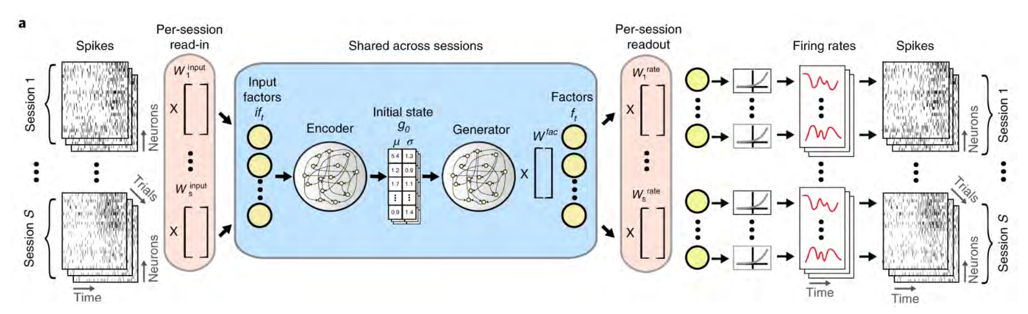
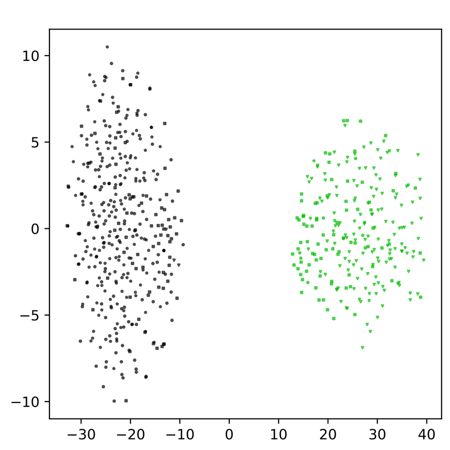
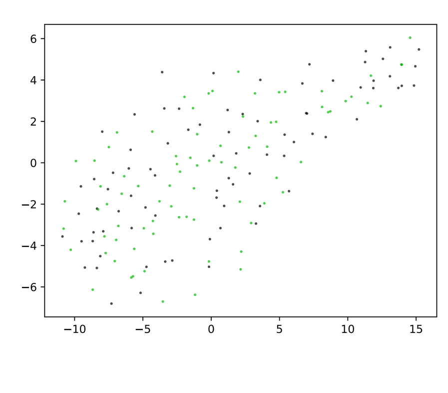


Figure 4 of Pandarinath et al. (2018)

We utilized the PyTorch implementation of the LFADS model as described by Sedler and Pandarinath (2023), which was recently made available. The data preprocessing followed the methodology outlined by Pandarinath et al. (2018), initializing the input matrices through Principal Component Regression (PCR) coefficients. This approach projected the mean of the data for each session onto the principal components derived from session-wise data means, organized by class. This technique was tailor-made for aligning data across multiple sessions within the framework of the LFADS model.
The training protocol integrated all sessions into the LFADS framework, enabling the model to derive a generalized representation of neural activity. The extracted LFADS factors – condensed representations of neural dynamics – served as the foundation for subsequent analyses.
Building on the LFADS factors, we developed a classifier specifically designed to identify the training condition associated with each neural pattern.
We then assessed the decoding performance resulting from this approach against the performance achieved without the PCA initialization.

SABLE Latent Space



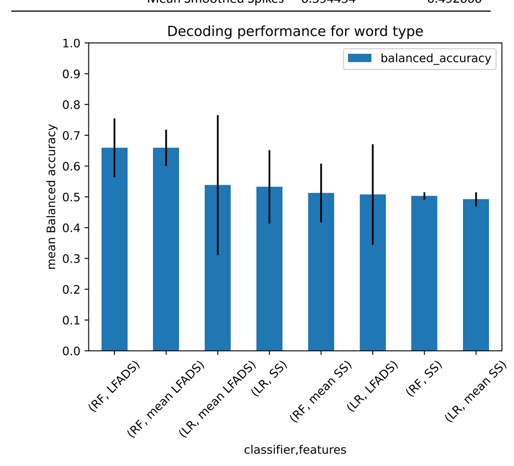


The left plot presents a TSNE visualization of the latent space representation of the initial conditions of the SABLE model applied to the training points. The black dots represent one of the word types, while the green dots represent the other. The model was trained using all available data except for one held-out session. A clear separation between the word types is observed, alongside a mixing of the sessions. The right plot displays the latent space representation of the held-out session. It shows a lack of separation between the word types, indicating a form of overfitting.

Decoding performance

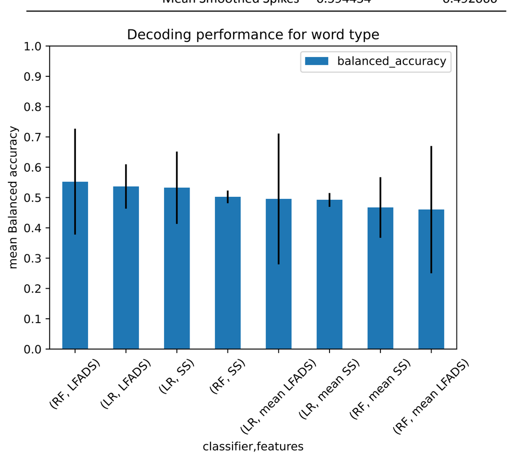
PCR initialised read in matrices

Classifier	Features	Accuracy	
		Accuracy	Balanced Accuracy
Random Forest	LFADS Factors	0.601537	0.685155
	Mean LFADS Factors	0.452356	0.603696
	Smoothed Spikes	0.728840	0.505991
	Mean Smoothed Spikes	0.641693	0.462352
Logistic Regression	Mean LFADS Factors	0.505436	0.538533
	Smoothed Spikes	0.510589	0.532353
	LFADS Factors	0.576918	0.507457
	Mean Smoothed Spikes	0.594454	0.492066



Randomly initialised read in matrices

Classifier	Features	Accuracy	
		Accuracy	Balanced Accuracy
Random Forest	LFADS Factors	0.493565	0.547623
	Smoothed Spikes	0.696789	0.544118
	Mean LFADS Factors	0.545098	0.464518
	Mean Smoothed Spikes	0.565581	0.450099
Logistic Regression	LFADS Factors	0.456607	0.536487
	Smoothed Spikes	0.510589	0.532353
	Mean LFADS Factors	0.497553	0.495313
	Mean Smoothed Spikes	0.594454	0.492066



The models were trained on all available sessions to find a shared space, while classifiers were cross-validated with one held-out session. The reported performance represents the average, with vertical black lines in the plots indicating the standard deviation. In the PCR-initialized approach, the decoding performance reached up to 68% accuracy for the Wordtype. With random initialization, the performance generally decreased compared to the PCR-initialized approach. The analysis suggests that non-linear relationships exist between LFADS factors and the target. Linear classifiers failed to achieve significant performance beyond random guessing. As baseline features, smoothed spike data and their trial mean were used. The PCR initialization notably enhanced and stabilized decoding performance on LFADS factors compared to random initialization. By comparing the two approaches, it is evident that PCR initialization is necessary when utilizing LFADS factors as features.

Discussion

In this project, we faced significant challenges, notably with the SABLE model, whose results we could not reproduce.
Our work with the LFADS model, however, yielded some promising directions. The model's performance was notably improved when incorporating all available sessions and initializing with PCR coefficients that were split by conditions.
Moreover, the implementation of LFADS we used allows for the selection of different posteriors. Our experiments suggests that alternatives, such as a Gaussian posterior, might offer better results in terms of data reconstruction. This insight into posterior choice could guide future efforts in optimizing model performance for neural data analysis.
Despite the setbacks with domain adaptation methods, including our initial unsuccessful attempts and the challenges with SABLE, the LFADS model presents a viable pathway for cross-session alignment. Given these experiences, we recommend further investigation into LFADS-like methods for neural data alignment, considering the nuanced successes and limitations observed.

References

References

Hurwitz, C., Srivastava, A., Xu, K., Jude, J., Perich, M. G., Miller, L. E., and Hennig, M. H. (2021). Targeted neural dynamical modeling.

Jude, J., Perich, M. G., Miller, L. E., and Hennig, M. H. (2022). Robust alignment of cross-session recordings of neural population activity by behaviour via unsupervised domain adaptation.

Pandarinath, C., O'Shea, D. J., Collins, J., Jozefowicz, R., Stavisky, S. D., Kao, J. C., Trautmann, E. M., Kaufman, M. T., Ryu, S. I., Hochberg, L. R., Henderson, J. M., Shenoy, K. V., Abbott, L. F., and Sussillo, D. (2018). Inferring single-trial neural population dynamics using sequential auto-encoders. *Nature Methods*, 15(10):805–815.

Sedler, A. R. and Pandarinath, C. (2023). lfads-torch: A modular and extensible implementation of latent factor analysis via dynamical systems.