A LATENT VARIABLE MODELING APPROACH FOR COGNITIVE EEG DATA: AN EXAMPLE FROM NEUROLINGUISTICS

Davide Turco[*], **Conor Houghton**
Faculty of Engineering
University of Bristol
Bristol, BS8 1UB, UK
{davide.turco, conor.houghton}@bristol.ac.uk

ABSTRACT

Electroencephalography (EEG) provides high temporal resolution data that are valuable for analyzing cognitive processes, but the high noise and dimensionality make analysis difficult. Traditional event-related potential studies lose single-epoch information through epoch averaging and restricting analysis to specific landmarks. To address this, we apply a latent variable model (LVM), LFADS, to encode EEG epochs and infer lower-dimensional dynamical factors reflecting cognitive processes. We first validate LFADS on synthetic EEG data, proving it recovers latent dynamics and external inputs. We then apply LFADS to real EEG data from a reading experiment and find it can reconstruct epochs' signal and distinguish responses to words with different syntactic roles. Moreover, we decode two word features from the inferred factors, with performance comparable to decoding using components obtained from traditional dimensionality-reduction techniques. Our results illustrate the potential of dynamical LVMs as an alternative approach for EEG dimensionality reduction, preserving interpretable factors encoding cognitive information. Applying such models to clinical EEG may uncover temporal biomarkers of cognitive processes.

1 INTRODUCTION

Non-invasive neuroimaging techniques are critical for analyzing brain activity related to cognitive processing and developing neuroengineering solutions for patients with neurodegenerative conditions (Tayebi et al., 2023). EEG, with its millisecond-level temporal resolution, is invaluable for investigating neural processing in which temporal structure is important. However, its high dimensionality and noise make decoding difficult.

Traditional event-related potential (ERP) studies attempt to overcome these issues by averaging EEG activity evoked by a stimulus across epochs and subjects, to improve the signal-to-noise ratio. Analysis is often restricted to the ERP components, i.e. the signal amplitude at fixed times after the stimulus onset, traditionally associated with specific cognitive functions; for instance, the N400 is considered a marker of semantic processing (Kutas & Federmeier, 2000). However, the use of ERP has limitations: the ERP waveform is the average of a large number of EEG epochs, leading to the loss of single-epoch and time-course information. Moreover, ERP components may be associated to more than one underlying cause (Kutas & Federmeier, 2011), making their interpretation intricate (Turco & Houghton, 2022).

Seminal electrophysiology work has shown that neural dynamics can be well described by lower-dimensional factors (Churchland et al., 2014), motivating the development of LVMs that infer latent dynamics from single-trial recordings. Early work employed Gaussian Process Factor Analysis (Yu et al., 2009) or linear dynamical systems (Gao & Archer, 2015). The state-of-the-art Latent Factor Analysis via Dynamical Systems (LFADS) model (Pandarinath et al., 2018) consists of a sequential variational autoencoder (VAE) (Kingma & Welling, 2013) that models neural dynamics by learning

[*] Corresponding author

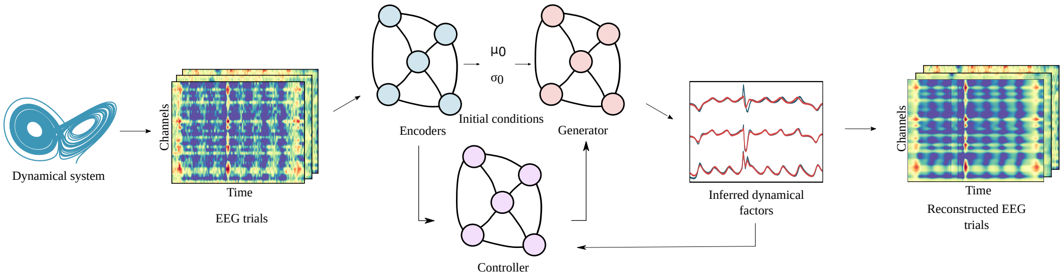


Figure 1: Overview of the model, here shown for synthetic EEG data. GRU encoders summarize the EEG epochs into initial conditions for a GRU generator, which in turn infers lower-dimensional latent factors. These factors are then used to reconstruct the original epochs. The factors can be fed back into a GRU controller, which provides an input to the generator at each time-step.

trial-specific initial conditions from the input data. The inferred dynamical factors have been shown to be highly predictive of the subject's behavior in motor tasks. The model has been applied to other sources of data, such as calcium imaging (Zhu et al., 2021) and EMG (Wimalasena et al., 2022).

Dimensionality reduction on EEG, on the other hand, is usually performed using simple non-dynamical methods: principal component analysis (PCA), independent component analysis (ICA), factor analysis (FA) or traditional autoenconders (Zhang et al., 2020; Tăuţan et al., 2021). Decoding of linguistic features from EEG usually relies on analysis of ERPs or full data (Ling et al., 2019; Murphy et al., 2022). Extending recent progress in non-linear dimensional reduction will mean that EEG becomes a better tool for research and diagnosis for cognitive disorders, and could be power future brain-machine interfaces.

Inspired by recent advances in the analysis of multidimensional electrophysiological data, here, we consider the application of a dynamical LVM to neurolinguistic EEG data. Language processing presents an excellent test case due to its fine temporal dynamics. Moreover, many neurodegenerative conditions have linguistic deficits. We first validate the model on synthetic EEG data, for which we know the latent dynamics. We then apply the model to publicly available EEG data and test whether the inferred lower-dimensional latent factors are informative of the cognitive process underlying language processing. This work represents, to our knowledge, the first application of a dynamical LVM to neurolinguistic EEG data, and is a novel contribution to the recent literature on LVMs or single-trial methods for cognitive EEG data (Ghaderi-Kangavari et al., 2022; Vo et al., 2024).

2 Methodology

2.1 Data

Synthetic EEG data were generated from the Lorenz dynamical system in its chaotic regime. The three dynamical factors were mapped to 32 dimensions via a randomly generated linear transformation, to simulate 32-channel EEG data. To mimic the effect of external stimuli, we perturbed the epochs with a delta pulse at a random time between 250 and 750 ms. Full details of the synthetic data generation process are included in Appendix A.1.

We then analyzed data from previously published EEG recordings from 24 subjects reading 205 natural sentences, comprising 1931 words (Frank et al., 2015); our analysis is limited to a subset of 12 subjects. The signal was recorded at a sampling frequency of 500 Hz, then downsampled to 250 Hz, using a 32-electrode cap, and re-referenced to the average of the mastoids. Filtering and artifact removal were already performed in the original study. We epoched the data from -100 to 700 ms relative to the word onset, and we baseline-normalized the epochs with respect to the average signal in the 100 ms pre-onset segment. We then annotated the stimulus with part-of-speech (POS) tags and grouped them into content (e.g nouns, verbs) and function (e.g. articles, prepositions) words. We also considered annotations already included in the dataset, such as log-transformed word frequency.

2.2 Model

The latent variable modeling analysis used LFADS, which was adapted for EEG data by employing a Gaussian likelihood model. The architecture consists of a sequential VAE, as can be seen in Fig. 1. The dynamical factors f_t are inferred by a generator, which is initialized with a set of epoch-specific initial conditions $g_0 = (\mu_0, \sigma_0)$ that are determined by the encoded representations of the original data x_t. All the networks are gated recurrent units (GRU). Mappings to lower and higher dimensional spaces are obtained with linear weight matrices W. In brief, the generative process is:

$$\begin{aligned} g_0 &\sim \mathcal{N}(\mu_0, \sigma_0^2) \\ g_t &= \text{GRU}(g_{t-1}) \\ f_t &= W_f(g_t) \\ \hat{x}_t &\sim \mathcal{N}(x_t; W_x(f_t), \sigma^2). \end{aligned} \tag{1}$$

Additionally, the architecture can include a controller to model external stimuli or perturbations, c_t. The controller is an additional GRU that, at each time-step, receives the dynamical factors from the previous time-step and provides an input to the generator, $c_t = \text{GRU}(c_{t-1}, [e_t, f_{t-1}])$, where e_t is a state variable sampled from the encoder latent space.

The training objective is to maximize the log-likelihood of the data given the latent variables or, equivalently, maximizing the evidence lower bound which, for tractability, is approximated using variational methods. When a controller is used, there is an additional regularization term for the controller latent space, with an autoregressive prior. The model is optimized using Adam (Kingma & Ba, 2015) with L2 regularization and annealing the Kullback–Leibler divergence terms to avoid exploding gradients.

2.3 Analysis

To analyze the generative components of the model, we draw 50 samples from the distribution of latent variables and average them to obtain posterior probabilities. The inferred factors are assessed as a reduced-dimensional representation of the original data, considering the proportion of variance in the actual data explained and the effectiveness of the factors in describing linguistic features.

We decode two word properties from the factors: log-transformed frequency and POS tags. For the former, we convert frequency values into two classes based on the median value in the training set. For POS decoding, we consider the six tags with most samples in the training set: `NOUN`, `VERB`, `PRON`, `DET`, `ADP`, `ADV`. Our classifier consists of a GRU, followed by a fully-connected layer and a final activation function. Due to the high class imbalance in the dataset, we give the loss function additional rescaling weights for each class. We compare the decoding from LFADS factors with two non-dynamical dimensionality-reduction techniques, ICA and FA, with the same number of components/factors. We also evaluate the performance of the classifier on randomly generated factors, which act as baseline.

3 Results

3.1 Synthetic Data

The ability of the model to recover the true dynamics was assessed using the coefficient of determination, yielding scores of $(0.80 \pm 0.14, 0.77 \pm 0.12, 0.89 \pm 0.09)$ for x, y, and z respectively. Alignment of inferred factors with true factors was achieved through a simple linear transformation. To validate the encoder's ability to summarize initial conditions in the latent space, we visualized a 2D representation using UMAP (McInnes et al., 2018). Fig. 2A illustrates distinct clusters representing epoch-specific initial conditions. The controller's performance was evaluated by estimating the time

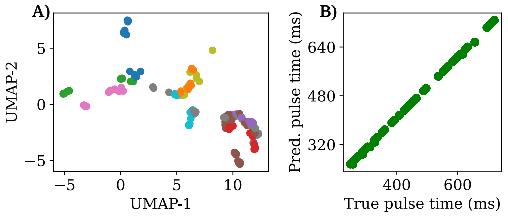


Figure 2: Main results for synthetic EEG data. A) 2D UMAP view of the latent space shows that initial conditions are properly encoded (here shown for ten states, represented by ten different colors). B) The pulse times inferred by the controller strongly correlate to the ground truth.

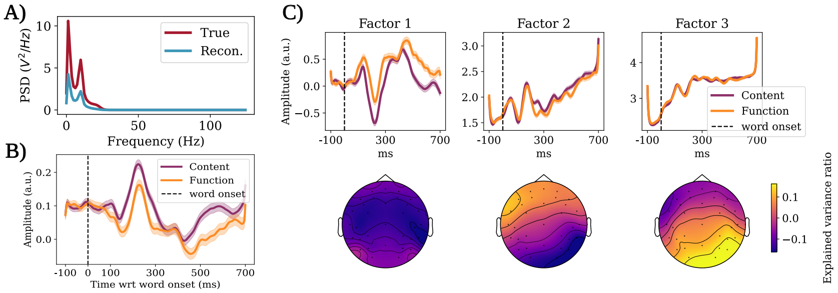


Figure 3: Model output for held-out epochs. A) PSD plot for true and reconstructed EEG epochs, showing similar peaks in delta and alpha bands, but less power in the reconstructed data. B) The averaged factor (±s.e.m.) corresponding to function words displays a smaller amplitude than that corresponding to content words ($p \ll 10^{-4}$, one-tailed paired t-test). C) 3D PCA visualization of the inferred factors (mean ± s.e.m.). Topographic plots showing the explained variance ratio of each factor, highlighting higher scores in left frontotemporal and occipital areas.

Table 1: Decoding on held-out epochs from LFADS factors significantly outperforms chance ($p \leq 10^{-4}$, independent t-test), and is comparable to ICA and FA for both features (mean across 10 seeds).

Feature	**Model**			
	LFADS	***ICA***	***FA***	***Random***
POS tag	0.189 ± 0.007	0.195 ± 0.007	0.192 ± 0.006	0.132 ± 0.006
Log freq.	0.574 ± 0.012	0.590 ± 0.013	0.582 ± 0.011	0.543 ± 0.014

of the delta pulse. Following Pandarinath et al. (2018), we determined the pulse time as the point at which the input signal reaches maximum absolute value. Fig. 2B demonstrates strong correlation between inferred and ground truth pulse times.

While ICA and FA can also recover the true Lorenz dynamics, they do not have any mechanism to account for external perturbation, as a controller, or indeed any time-dependent component.

3.2 Neurolinguistic EEG Data

We evaluated the reconstruction ability of the model by comparing the power spectral density (PSD) of held-out epochs with the PSD of the reconstructed data. Fig. 3A shows that reconstructed epochs preserve prominent power peaks in the delta and alpha frequency bands, but have lower signal power overall. This indicates that while reconstructing the noisy EEG signal, there has been a loss of information, most likely during the dimensionality reduction/expansion stages.

After hyperparameter tuning, the number of factors was fixed at eight. These factors were categorized into two groups representing content and function word epochs, then averaged across epochs. Fig. 3B shows that the average factor for function word epochs has lower amplitude than content words (one-tailed paired t-test), demonstrating that the inferred factors can distinguish brain responses to different word classes. We further reduced the dimensionality of the factors to three using PCA and depicted the components in Fig. 3C. By displaying the ratio of variance in the original EEG data explained by each factor on a topographic plot, we observed that factor 2 is associated with the left frontotemporal area, traditionally linked to language processing, while factor 3 corresponds to the visual cortex, implicated in reading experiments.

We report the average results for the decoding analysis on held-out epochs across ten random seeds in Table 1. Due to high class imbalances, we chose F-1 score as metrics for these classification tasks, as it is more robust than accuracy. Classification from LFADS-inferred dynamical factors is significantly better than chance for both the linguistic features of interest (one-tailed independent t-test). Moreover, decoding from LFADS factors leads to F-1 scores comparable to those obtained decoding from ICA or FA components, proving that LFADS can be safely used as an alternative method for reducing the dimensionality of EEG data and get similar decoding performances as traditional methods.

4 Discussion

In this study, we investigated the potential of the dynamical LVM LFADS in inferring latent factors which are informative of cognitive processes, using data from a language experiment as an example. We first demonstrated that the model can recover the latent dynamics from synthetic EEG data, for which the true factors are known. We then showed that the model allows for reconstruction of the original EEG epochs and it infers factors that can distinguish the response to different word classes and from which two word features can be decoded with scores comparable to those obtained using traditional dimensionality-reduction techniques. Our preliminary work illustrates the potential of such models as an alternative method for EEG dimensionality reduction and motivates further applications of LVMs to clinical EEG data.

In future work, we plan to better optimize the model by performing a large-scale hyperparameter search. We will also consider the temporal decoding performance, that would allow a more natural comparison with traditional ERPs. Finally, the model will be compared to other dynamical LVMs.

Acknowledgments

We thank the Frank lab for the data. DT is funded by an UKRI Centre of Doctoral Training grant (EP/S022937/1). This work was carried out using the HPC facilities of the ACRC, University of Bristol. We are grateful to Dr Stewart whose philanthropy supported the purchase of GPU nodes.

References

Michael M Churchland, John P Cunningham, Matthew T Kaufman, and Krishna V Shenoy. Neural population dynamics during reaching. *Nature neuroscience*, 17(10):1570–1578, 2014.

Stefan L. Frank, Leun J. Otten, Giulia Galli, and Gabriella Vigliocco. The ERP response to the amount of information conveyed by words in sentences. *Brain and Language*, 140:1–11, January 2015.

Jianfeng Gao and Ben Archer. Linear dynamical neural population models through nonlinear embeddings. In *Advances in Neural Information Processing Systems*, pp. 1857–1865, 2015.

Amin Ghaderi-Kangavari, Jamal Amani Rad, Kourosh Parand, and Michael D. Nunez. Neurocognitive models of single-trial EEG measures describe latent effects of spatial attention during perceptual decision making. *Journal of Mathematical Psychology*, 111:102725, 2022. ISSN 0022-2496.

Olaf Hauk, Claire Coutout, Andrew Holden, and Yi Chen. The time-course of single-word reading: Evidence from fast behavioral and brain responses. *NeuroImage*, 60(2):1462–1477, 2012.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

M. Kutas and K. D. Federmeier. The N400: A neural system for semantic integration. *Trends in cognitive sciences*, 15(10):449–456, 2011.

Marta Kutas and Kara D. Federmeier. Electrophysiology reveals semantic memory use in language comprehension. *Trends in Cognitive Sciences*, 4(9):388–393, 2000.

Shouyu Ling, Andy C. H. Lee, Blair C. Armstrong, and Adrian Nestor. How are visual words represented? insights from eeg-based visual word decoding, feature derivation and image reconstruction. *Human Brain Mapping*, 40(17):5056–5068, 2019.

Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. Umap: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29):861, 2018.

Alex Murphy, Bernd Bohnet, Ryan McDonald, and Uta Noppeney. Decoding part-of-speech from human EEG signals. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2022.

Chethan Pandarinath, Daniel J. O'Shea, Jasmine Collins, Rafal Jozefowicz, Sergey D. Stavisky, et al. Inferring single-trial neural population dynamics using sequential auto-encoders. *Nature Methods*, 15(10):805–815, September 2018.

H. Tayebi, S. Azadnajafabad, S.F. Maroufi, M.H. Jamali, A. Ghaderi, et al. Applications of brain-computer interfaces in neurodegenerative diseases. *Neurosurgical Review*, 46(1):131, 2023.

Davide Turco and Conor Houghton. Bayesian modeling of language-evoked event-related potentials. *Conference on Cognitive Computational Neuroscience (CCN)*, 2022.

Alexandra-Maria Tăuţan, Alessandro C. Rossi, Ruben de Francisco, and Bogdan Ionescu. Dimensionality reduction for EEG-based sleep stage detection: comparison of autoencoders, principal component analysis and factor analysis. *Biomedical Engineering*, 66(2):125–136, 2021.

Khuong Vo, Qinhua Jenny Sun, Michael D. Nunez, Joachim Vandekerckhove, and Ramesh Srinivasan. Deep latent variable joint cognitive modeling of neural signals and human behavior. *NeuroImage*, pp. 120559, 2024. ISSN 1053-8119.

Lahiru N Wimalasena, Jonas F Braun, Mohammad Reza Keshtkaran, David Hofmann, Juan Álvaro Gallego, et al. Estimating muscle activation from EMG using deep learning-based dynamical systems models. *Journal of Neural Engineering*, 19(3):036013, May 2022.

Byron M Yu, John P Cunningham, Gopal Santhanam, Stephen I Ryu, Krishna V Shenoy, et al. Gaussian-process factor analysis for low-dimensional single-trial analysis of neural population activity. *Journal of Neurophysiology*, 102(1):614–635, 2009.

Yu Zhang, Wei Zhang, Xiao Liu, Rui Wang, and Yu Wang. Latent factor decoding of multi-channel EEG for emotion recognition through autoencoder-like neural networks. *Front. Neurosci.*, 14:87, 2020.

Feng Zhu, Harrison A. Grier, Raghav Tandon, Changjia Cai, Anjali Agarwal, Andrea Giovannucci, et al. A deep learning framework for inference of single-trial neural population dynamics from calcium imaging with sub-frame temporal resolution. *bioRxiv*, November 2021.

A Appendix

A.1 Lorenz-system Synthetic Data

The dynamics of the Lorenz system is described by the following system of differential equations:

$$\begin{aligned}\frac{dx}{dt} &= \sigma(y-x) \\ \frac{dy}{dt} &= x(\rho-z)-y \\ \frac{dz}{dt} &= xy-\beta z\end{aligned} \tag{2}$$

Synthetic data was generated by running the system with parameter values $\beta = \frac{8}{3}, \rho = 28, \sigma = 10$. We solved the system using the Isoda method with a time-step of 0.006. The system was initialized with 65 states and run for 1 s. The three dynamical factors were then mapped to 32 dimensions via a randomly generated linear transformation. 20 epochs per condition were extracted by applying Gaussian noise to the synthetic EEG data. To mimic the effect of external stimuli, we perturbed the epochs with a delta pulse at a random time between 250 and 750 ms. The pulse was applied per-condition while solving the system.

Table 2: LFADS hyperparameters chosen for models trained on Lorenz synthetic EEG data and real neurolinguistic EEG data.

Hyperparameter	**Lorenz**	**Real**
Encoder dim.	64	64
Controller dim.	64	64
Latent dim.	64	16
Generator dim.	64	100
Factor dim.	3	8
Input dim.	1	4
Initial lr.	0.004	0.004
Dropout	0.3	0.3
Batch size	128	256
LFADS seed	100	0

Table 3: Hyperparameters chosen for the GRU-based classifier used in the decoding analysis.

Hyperparameter	**Value**
Hidden dim.	64
Num. of layers	1
Initial lr.	0.003
Dropout	0.3
Batch size	512

A.2 Loss Function

The objective of the model is to maximize the evidence lower bound (ELBO) of the observed data x, marginalizing over latent variables. The full loss function has the same form as the original LFADS loss function (Pandarinath et al., 2018), adapted to account for the normality of EEG data, and is defined as:

$$\begin{aligned}\text{ELBO}(x_t) = &- D_{\text{KL}}\left[\mathcal{N}(g_0|\mu^{g_0}, \sigma^{g_0})||P^{g_0}(g_0)\right] \\ &- \sum_{t=1}^{T} D_{\text{KL}}\left[\mathcal{N}(u_t|\mu_t^u, \sigma_t^u)||P^u(u_t|u_{t-1})\right] \\ &+ \sum_{t=1}^{T} \log(\mathcal{N}(x_t; W(f_t), \sigma))\end{aligned} \tag{3}$$

where T is the length of the trial, g_0 are the initial conditions, f_t the inferred factors, u_t the input inferred by the controller at each time-step and W is a linear mapping. KL indicates the Kullback–Leibler divergence. For simplicity, we have not included the marginalization over latent variables in the above equation.

The prior for the initial condition latent space was set to $(\mu = 0, \sigma^2 = 0.1)$ for the model trained on real EEG data, and $(0, 1)$ for that trained on synthetic data. In both cases, the prior on the controller latent space was autoregressive, with a process autocorrelation of 10 and a process variance of 0.1.

A.3 Model Hyperparameters

We report here the hyperparameters chosen for the LFADS models trained on synthetic and real data (Table 2), and the hyperparameters used when training the classifiers in the decoding analysis (Table 3). We did not run an exhaustive hyperparameter search, especially for the LFADS model; we used the default values for all the parameters not included in the tables below.

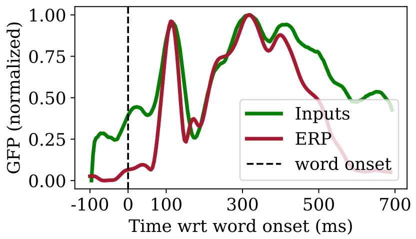
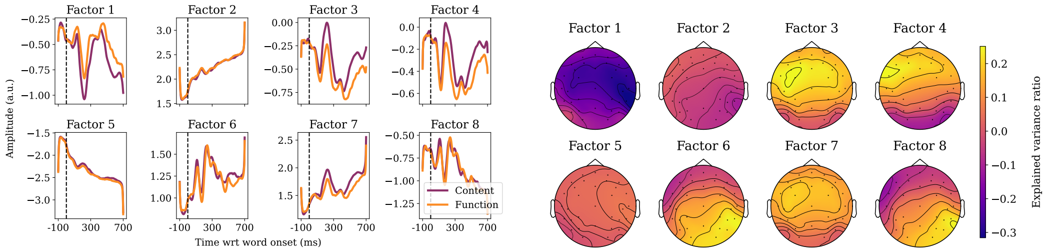


Figure 4: The normalized GFP for inferred inputs closely aligns with that for ERPs, and shows a sharp decline at ~ 200 ms.

A.4 Additional Results

Inferred Inputs The controller was allowed four inputs to model stimulus response. We considered the average variability between inferred inputs across time, the global field power (GFP), and compared it to the GFP of the ERPs estimated from the data. Fig. 4B shows the the GFP of the inferred inputs strongly matches that of the ERPs from raw data, indicating that the controller may perform ERP-like averaging of epochs. Both show a prominent drop in variability at around 200 ms, which is a time-lag usually associated with the earliest brain response to word reading (Hauk et al., 2012), suggesting that the inferred inputs may encode the response to word reading.



(a)

(b)

Figure 5: Individual plots of the eight inferred factors. (a) For most factors, the response to function words is inferred to be weaker than for content words. (b) Topographic plots showing the explained variance ratio of each factors, highlighting higher scores in left frontotemporal and occipital areas.

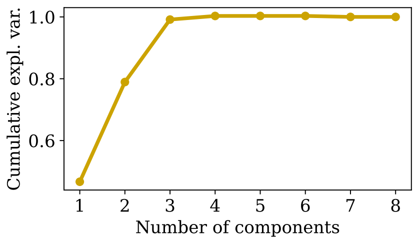


Figure 6: Cumulative proportion of variance explained by PCA models with different number of components fitted on the eight LFADS factors.

Inferred factors We include here the plots of all the eight factors inferred by the model, both as traditional graphs displaying the difference between average factors corresponding to content and function words (Fig. 5a) and as topographic maps showing the explained variance ratio for different EEG channels (Fig. 5b). The presence of pairs of factors with the same score topography (see factors 3 and 4, or factors 6 and 8) motivated a further dimensionality reduction using PCA. The optimal number of PCA components was chosen considering the cumulative explained variance of the principal axes with varying number of components. As Fig. 6 shows, three components can explain nearly all the variance of the original factors.

A.5 Implementation Details

The ICA and FA methods were implemented using `scikit-learn 1.3.0`. All the deep-learning based models, including the classifiers used in the decoding analysis, were implemented in `PyTorch 2.1.0` with `cuda 12.1`. The implementation of the LFADS model is based on the PyTorch version of LFADS[1].

[1] https://github.com/arsedler9/lfads-torch