Modelling variability in dynamic functional brain networks using embeddings

Rukuang Huang, Chetan Gohil, Mark Woolrich

January 27, 2024

Abstract

Neuroimaging techniques offer unprecedented insights into the dynamic neural processes underlying cognitive functions. With recent studies, data driven models like the Hidden Markov Model (HMM) are getting more attention due to their ability to infer fast temporal dynamics in functional networks in an unsupervised manner. However, these dynamic network models are typically trained at the group level. Whilst it is possible to post-hoc estimate the session-specific networks with the so-called dual estimation, this does not allow the model to discover and benefit from subpopulation structure in the group. We propose an extension to the HMM model that incorporates embedding vectors (c.f. word embedding in Natural Language Processing) to explicitly model individual sessions while training on the entire group. This effectively infers a "fingerprint" for each individual session, which can group together those with similar spatio-temporal patterns. With simulated data, we show that the model can recover the underlying subpopulation structure, achieve higher accuracy than dual estimation on session-specific quantities and can make use of increasing number of sessions to benefit the inference of individual sessions. Applying this model to resting-state and task MEG data, we show the learnt embedding vectors capture meaningful sources of variation across a population. This includes subpopulations related to demographics and systematic differences, such as scanner types or measurement sites. The proposed model offers a powerful new technique for modelling individual sessions while leveraging information from an entire group.

1 Introduction

A number of studies have shown that the brain recruits functional networks during task ([16, 34, 30, 11]), and at rest ([3, 7, 14]). An increasingly important perspective is how networks dynamically activate over time. For example, the Hidden Markov model (HMM, [2, 41]) is a popular model for inferring dynamic functional networks in an unsupervised manner. This model has been successfully applied to both functional magnetic resonance imaging (fMRI) and Magnetoencephalography (MEG) data. For example, [42] has shown with resting-state fMRI data that transitions between networks are nonrandom and are hierarchically organised. It is also shown with MEG resting-state data in [23] that replay bursts coincides with activation of specific resting state networks. Additionally, with task MEG data, [34] uses HMM to identify networks whose activations are associated with task events. However, HMM is limited by only giving a group-level description of the brain regions and spectral content in each brain network.

Accurate estimation of subject-specific or even session-specific networks is important for characterising differences between individuals and for downstream prediction tasks. For instance, HMM has been used to predict behavioural traits ([40]). It is possible to get session-specific estimates with post-hoc analyses, in particular, a method called dual estimation, which has the similar rationale behind dual regression in Independent Component Analysis (ICA, [32], [4]). However, this approach does not allow the inference to benefit from the subpopulation structure within the group. One alternative approach is to use hierarchical models that capture how model parameters vary across subjects. For example, the PROFUMO model has been used successfully to capture aspects of subject variability in functional modes using fMRI data ([22, 21]). However, variability within the population is typically modelled using very simple parametric distributions (e.g. univariate Gaussian), and will fail to capture the rich, multi-channel variability expected in the population. In this paper, we introduce **HIVE** (HMM with Integrated Variability Estimation), an extension to the current HMM model that utilises Bayesian hierarchical modelling ([28]) and incorporates embedding vectors (c.f. word embeddings in Natural Language processing [31]) into the generative model in order to model session-specific variability within a dataset. Embedding vectors have been previously used to deal with between-subject variability in the context of supervised learning ([12, 9]). Here we incorporate it into an unsupervised, generative model of dynamic network activity in the form of the HMM.

In this paper, with simulated and real MEG datasets, we show that our proposed model can infer the underlying population structure in an unsupervised manner, which not only allows us to uncover structure in the population, but also helps to give a more accurate description of the session-specific estimates than traditional method based on the group-level model. Source code for the HIVE model is available in the *osl-dynamics* toolbox ([18]) and scripts to reproduce results on this paper are available at: *github.com/OHBA-analysis/Huang2024_ModelVariabilityWithEmbeddings*.

2 Methods

HMM partitions time series data into a finite number of states and each state has a specific spatial-temporal pattern. The generative model of HMM describes how the time series data is generated given the HMM model parameters. Through training, we find parameters that best generate the observed data.

2.1 Generative model of HMM

For $n \in \mathbb{N}$, let $[n]$ be the set $\{1, \ldots, n\}$. Let $\boldsymbol{x}_t$ be the brain data we observe at time $t \in [T]$. Here T is the total number of samples/time points and $\boldsymbol{x}_t$ is a vector of length N_c - the number of channels. In HMM, it is assumed that the data is generated by

$$\boldsymbol{x}_t | (s_t = j) \sim \mathcal{N}(\boldsymbol{\mu}_j, \boldsymbol{D}_j) \quad \forall t \in [T], \tag{1}$$

where $\boldsymbol{\mu}_j, \boldsymbol{D}_j$ are state-specific means and covariances that describe the spatial-spectral pattern for each state $j \in [J]$, and J is the number of states. In HMM, the process $(s_t)_{t=1}^T$ follows a discrete homogeneous Markov process ([17]). i.e.

$$p(s_{1:T}) = p(s_1) \prod_{t=2}^{T} p(s_t | s_{t-1}), \tag{2}$$

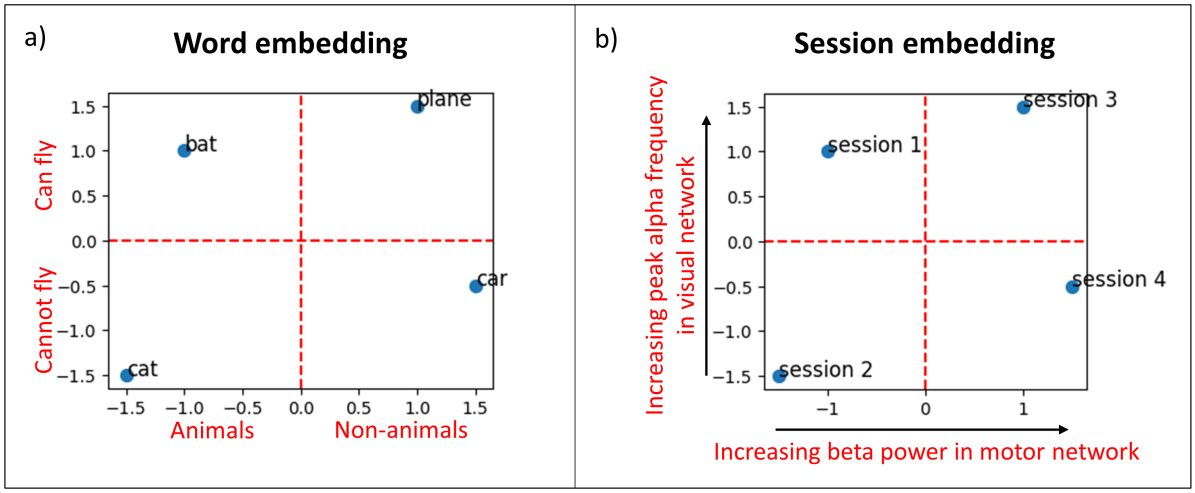

Figure 1: **Examples of embedding vectors**. a) Example of word embedding vectors. The x-axis encodes whether a word is for animals or non-animals and the y-axis encodes whether the object can fly or not. b) Illustration of possible variability captured by embedding vectors. The x-axis encodes the direction of increasing beta power in motor network whereas the y-axis encodes the direction of increasing peak alpha frequency in visual network.

where $p(s_{1:T})$ is a shorthand for the joint distribution $p(s_1, \ldots, s_T)$, $p(s_t = j | s_{t-1} = j') = A_{j'j}$ does not depend on time and is the (j', j)-th entry to the transition probibility matrix $\boldsymbol{A}$.

2.2 Dual estimation in the traditional HMM approach

The state means and covariances $\boldsymbol{\mu}_j, \boldsymbol{D}_j$ are group-level estimates. Similar to dual regression in ICA, dual estimation ([40]) is a way to compute session-specific means and covariances in HMM. For each session, we fix the state probabilities from the group-level HMM and re-estimating the observation model parameters with the session's data (See A.1 for details). In this paper we use **HMM-DE** (HMM with dual estimation) to refer to the pipeline of training a group-level HMM followed by dual estimation for estimating session-specific parameters.

2.3 The use of embedding vectors

The use of word embeddings is a common practice in Natural language processing (NLP, [31]). A word embedding is a real-valued vector that encodes the meaning of the word such that words closer in the vector space have similar meanings. In Figure 1a, we show an example of word embeddings. Here the x-axis encodes whether a word is for animals or non-animals and the y-axis encodes whether the object can fly or not. Similarly, we can assign a vector for each recording session and hope that sessions with similar properties are close to each other in the vector space. In Figure 1b, we show an illustration of possible variability captured by embedding vectors. Here the x-axis encodes the direction of increasing beta power in motor network whereas the y-axis encodes the direction of increasing peak alpha frequency in visual network.

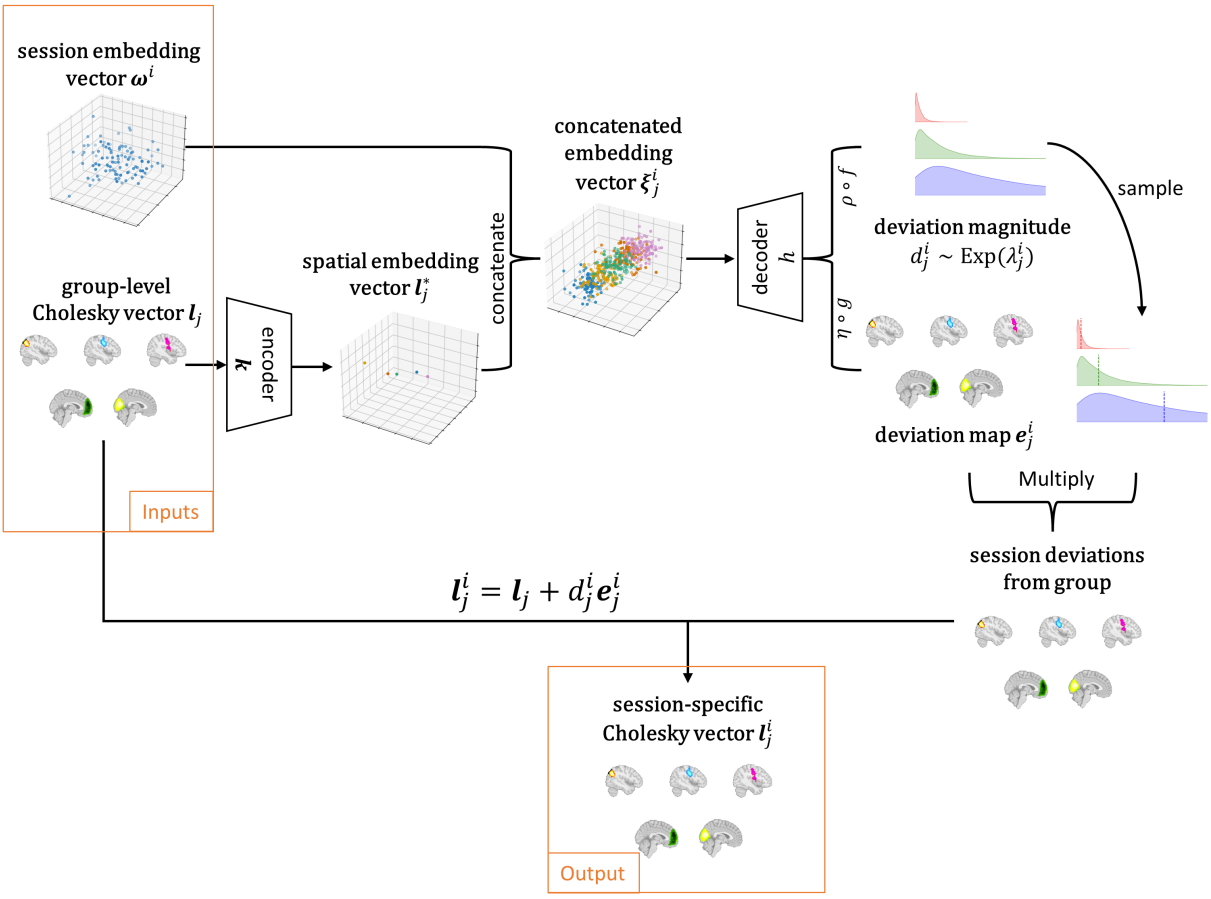

Figure 2: **Variability encoding block**. Session-specific information from the session embedding vectors $\boldsymbol{\omega}^i$ and state-specific information from the group-level Cholesky vectors $\boldsymbol{l}_j$ are combined to generate session and state-specific covariances $\boldsymbol{l}_j^i$. Here the **encoder** k is an affine transformation that condenses state specific information to a lower dimensional space. The **decoder** h is a multi-layer perceptron (MLP) that decodes state and session-specific deviations from the concatenated embedding vectors ξ_j^i.

2.4 Generating session-specific covariances.

In this section we outline the process of modelling session-specific covariances. For simplicity, we will assume that the state-specific means are fixed to zero (i.e. $\mu_j = 0, \forall j \in [J]$), but similar modelling can be easily applied to the means. We still assume that our data is generated from a multivariate Gaussian distribution with

$$\boldsymbol{x}_t^i | (s_t^i = j) \sim \mathcal{N}\left(0, \boldsymbol{D}_j^i\right), \quad \forall t \in [T], i \in [N], \tag{3}$$

where i is the session index and N is the number of sessions. We further assume that the transition probability matrices for all sessions are the same, i.e. $\boldsymbol{A}^i = \boldsymbol{A} \quad \forall i \in [N]$. Importantly, we assume each session i has their own set of covariances $\{\boldsymbol{D}_j^i\}_{j=1}^J$ and we want the covariances for different sessions to respect/be regularised by the relationships between sessions in the population. Furthermore, we want the session-specific covariances to vary around the group-level covariances. This is done by our novel variability encoding block described in Figure 2.

Given a session i, we want to generate the session-specific covariances $\boldsymbol{D}_j^i$. Let $\boldsymbol{\omega}^i$ be the session embedding vector of session i, which is a vector of pre-defined length n_ω that

provides a "fingerprint" for this session. Next, let $\boldsymbol{D}_j$ be the group-level covariance matrix for state j and to ensure legitimacy of generated session-specific covariances, we work in the Cholesky space. Let $\boldsymbol{l}_j, \boldsymbol{l}_j^i$ be the flattened vectors of the lower triangular entries of the Cholesky factors $\boldsymbol{L}_j, \boldsymbol{L}_j^i$ of $\boldsymbol{D}_j, \boldsymbol{D}_j^i$ respectively, then the session-specific Cholesky vector $\boldsymbol{l}_j^i$ is generated by adding a deviation to the group-level Cholesky vector $\boldsymbol{l}_j$ with

$$\boldsymbol{l}_j^i = \boldsymbol{l}_j + d_j^i \boldsymbol{e}_j^i, \tag{4}$$

where d_j^i is a scalar representing the spatial-global magnitude of the deviation and $\boldsymbol{e}_j^i$ is a vector representing the standardised pattern of the spatial deviation in the sense that it has unit sample standard deviation. We would expect the deviation magnitude d_j^i to be small so that the session-specific covariances vary around the group-level covariances, while enforcing positivity of this quantity. Hence we put an exponential prior on d_j^i:

$$d_j^i \sim \mathrm{Exp}(\lambda_j^i). \tag{5}$$

The rate parameter λ_j^i and deviation pattern $\boldsymbol{e}_j^i$ are generated through

$$\begin{aligned} \lambda_j^i &= \rho(f(h(\boldsymbol{\xi}_j^i))), \\ \boldsymbol{e}_j^i &= \eta(g(h(\boldsymbol{\xi}_j^i))), \end{aligned} \tag{6}$$

where ρ is the softplus function to ensure positivity of λ_j^i and η is a Layer Normalisation layer ([1]) with non-trainable scale parameter of 1 to ensure unit standard deviation of $\boldsymbol{e}_j^i$. Affine functions f, g are learnable transformations that extract different information from the hidden state given by the decoder h, a multi-layer perceptron (MLP, [33]). $\boldsymbol{\xi}_j^i$ is the key object called a **concatenated embedding vector** that encodes both session and state-specific information. It is formed by concatenating the session embedding vector $\boldsymbol{\omega}_i$ and spatial embedding vector $\boldsymbol{l}_j^*$, which is a lower dimensional representation of the Cholesky vectors:

$$\boldsymbol{\xi}_j^i = \boldsymbol{\omega}_i +\!\!+ \boldsymbol{l}_j^*, \quad \boldsymbol{l}_j^* = k(\boldsymbol{l}_j), \tag{7}$$

where k is an affine transformation that serves as an encoder to encode spatial information and $+\!\!+$ is a binary operator that concatenates two vectors. The complete generative process is summarised in Figure 3.

2.5 Inference

The parameters of HIVE include the transition probability matrix $\boldsymbol{A}$, the embedding vectors $\boldsymbol{\omega} = \{\boldsymbol{\omega}^i\}_{i=1}^N$, the group-level Cholesky vector $\boldsymbol{l} = \{\boldsymbol{l}_j\}_{j=1}^J$, as well as the weights of the learnable transformations h, f, g, k. To perform inference on these parameters, we employ the variational expectation–maximization (EM) algorithm ([15]), in particular, a variant of the Baum-Welch algorithm ([45]). During the "E-step", we update the state probabilities (the posterior distribution of the states). During the "M-step", given the state probabilities, we update the transition probability matrix $\boldsymbol{A}$ with the stochastic update technique used in [39] and all other parameters by minimising the objective function (variational free energy given the state probabilities, [5])

$$\mathbf{F} = -\mathbf{LL} + \mathbf{KL}. \tag{8}$$

The first term is the negative log-likelihood term. Let T_i be the number of samples for session i, $\boldsymbol{s} = \{s_t^i\}_{i,t}, \boldsymbol{d} = \{d_j^i\}_{i,j}, \boldsymbol{x} = \{x_t^i\}_{i,t}$, and $q(\boldsymbol{s}), q(\boldsymbol{d})$ be the posterior distribution

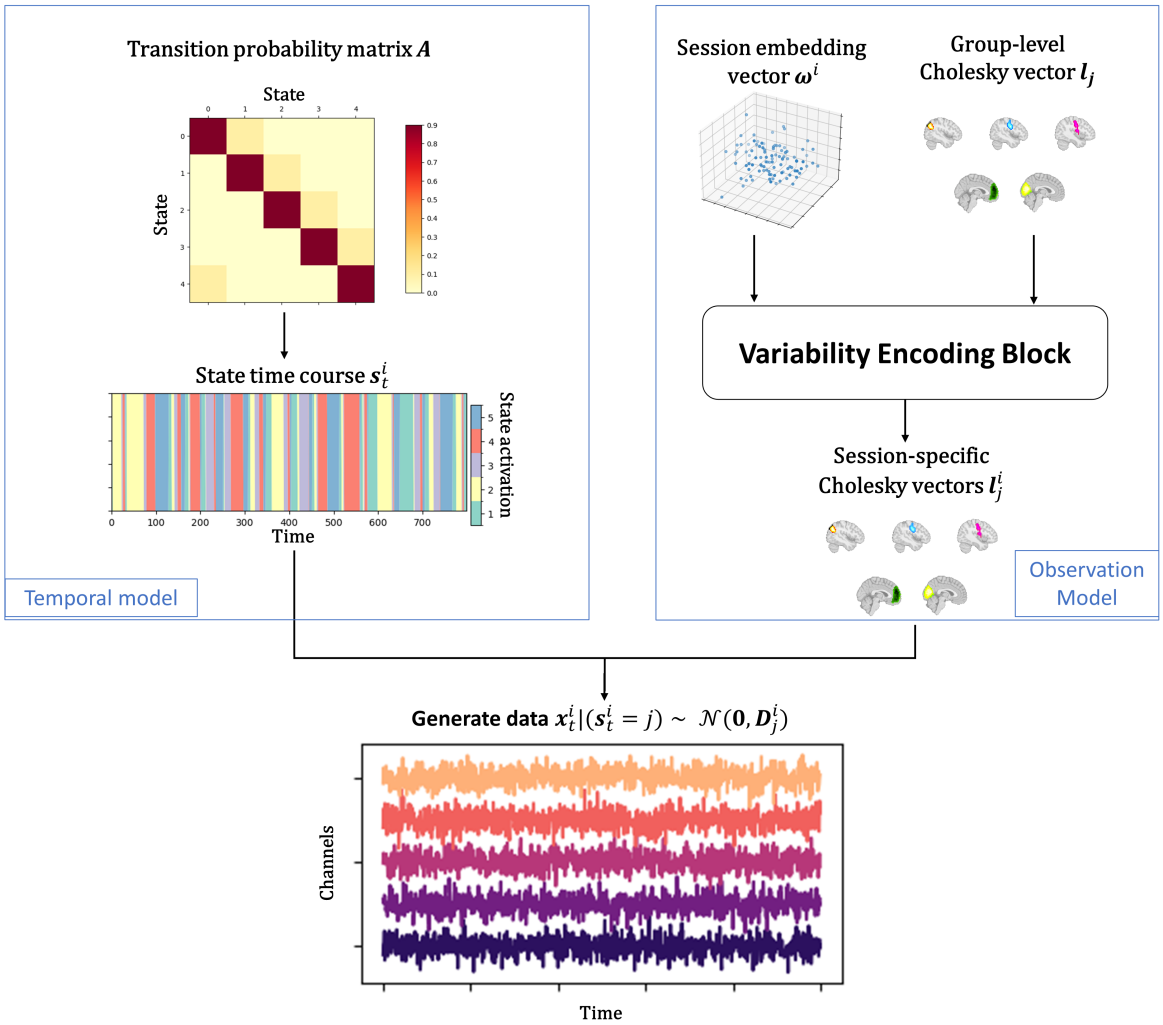

Figure 3: **Generative model of HIVE**. State time courses are generated from a transition probability matrix $\boldsymbol{A}$ through a Markov process. At the same time, session embedding vectors $\boldsymbol{\omega}^i$ and group-level Cholesky vectors $\boldsymbol{l}_j$ are passed to the variability encoding block to give session-specific Cholesky vectors $\boldsymbol{l}_j^i$. Given the state s_t^i at each time, the observed data is generated through the multivariate Gaussian likelihood.

on $\boldsymbol{s}$ and $\boldsymbol{d}$ respectively. Denote $\mathbb{P}(s_t^i = j|\boldsymbol{x})$ the posterior probability that for session i, the state at time t is j, then we have

$$\begin{aligned}\mathbf{LL} &= \mathbb{E}_{q(\boldsymbol{s})}\mathbb{E}_{q(\boldsymbol{d})}\left[\log p(\boldsymbol{x}|\boldsymbol{s},\boldsymbol{d})\right] \\ &= \sum_{i=1}^{N}\sum_{t=1}^{T_i}\mathbb{E}_{q(s_t^i)}\mathbb{E}_{q(\boldsymbol{d})}\left[\log \mathcal{N}(\boldsymbol{x}_t^i|0,\boldsymbol{D}_{s_t^i}^i)\right] \\ &= \sum_{i=1}^{N}\sum_{t=1}^{T_i}\sum_{j=1}^{J}\mathbb{P}(s_t^i = j|\boldsymbol{x})\mathbb{E}_{q(\boldsymbol{d})}\left[\log \mathcal{N}(\boldsymbol{x}_t^i|0,\boldsymbol{D}_j^i)\right].\end{aligned} \tag{9}$$

Notice we omit the dependence of $\boldsymbol{D}_j^i$ on $\boldsymbol{d}, \boldsymbol{\omega}, \boldsymbol{l}$ and weights of h, f, g, k for readability. The second term is the KL loss term on the observation model due to the deviation magnitudes,

$$\mathbf{KL} = D_{KL}(q(\boldsymbol{d})||p(\boldsymbol{d})) = \sum_{i=1}^{N}\sum_{j=1}^{J} D_{KL}(q(d_j^i)||p(d_j^i)), \tag{10}$$

where $D_{KL}(q(.)||p(.))$ is the Kullback–Leibler divergence ([29]) and we approximate the posterior distribution $p(\boldsymbol{d}|\boldsymbol{x})$ with a mean field approximation:

$$q(\boldsymbol{d}) = \prod_{i=1}^{N}\prod_{j=1}^{J} q(d_j^i) = \prod_{i=1}^{N}\prod_{j=1}^{J}\Gamma(d_j^i|\alpha_j^i,\beta_j^i), \tag{11}$$

where the shape α_j^i and the rate β_j^i parameters of the Gamma distributions Γ are learnable parameters. We can write down the analytic solution for **KL** and its gradients, and we used the reparameterisation trick ([27]) to get an estimate of the gradient of **LL** with respect to $\boldsymbol{\alpha} = \{\alpha_j^i\}_{i,j}$ and $\boldsymbol{\beta} = \{\beta_j^i\}_{i,j}$. During training, we minimise **F** with the Adam optimiser ([26]). The complete inference framework including forward pass and backward update is summarised in Figure 4.

Choosing the embedding dimension n_ω

We need to pre-define the length n_ω of the embedding vectors $\boldsymbol{\omega}^i$. A smaller n_ω will lead to a more parsimonious model but there could be a risk of over-regularisation. On the other hand, a larger n_ω will lead to a more flexible model but could lead to overfitting. In practice, we use the following strategy. Firstly, we define a set of candidate values for n_ω, say $\{n_\omega^1, \ldots, n_\omega^M\}$ where $n_\omega^k < n_\omega^{k+1}, \forall k \in [M-1]$. Then we train the model with each of these values a number of times independently and for each run, we get the variational free energy of the model on the training data. Starting from $k = 1$, we would then do progressive t-tests with the alternative hypothesis that the variational free energy of the model with n_ω^{k+1} is lower than the variational free energy of the model with n_ω^k. We would stop when the null hypothesis is not rejected and choose the model with n_ω^k as the final model, or the null hypothesis is never rejected until the largest candidate n_ω^M is reached, in which case we choose the model with n_ω^M as the final model. See Section A.5 for details.

Initialisation and annealing during training

A good initialisation is paramount in training deep neural networks to avoid slow and unstable training. The variational parameters $\boldsymbol{\alpha}, \boldsymbol{\beta}$ are initialised by calculating the session-specific variations of static covariance of the data.

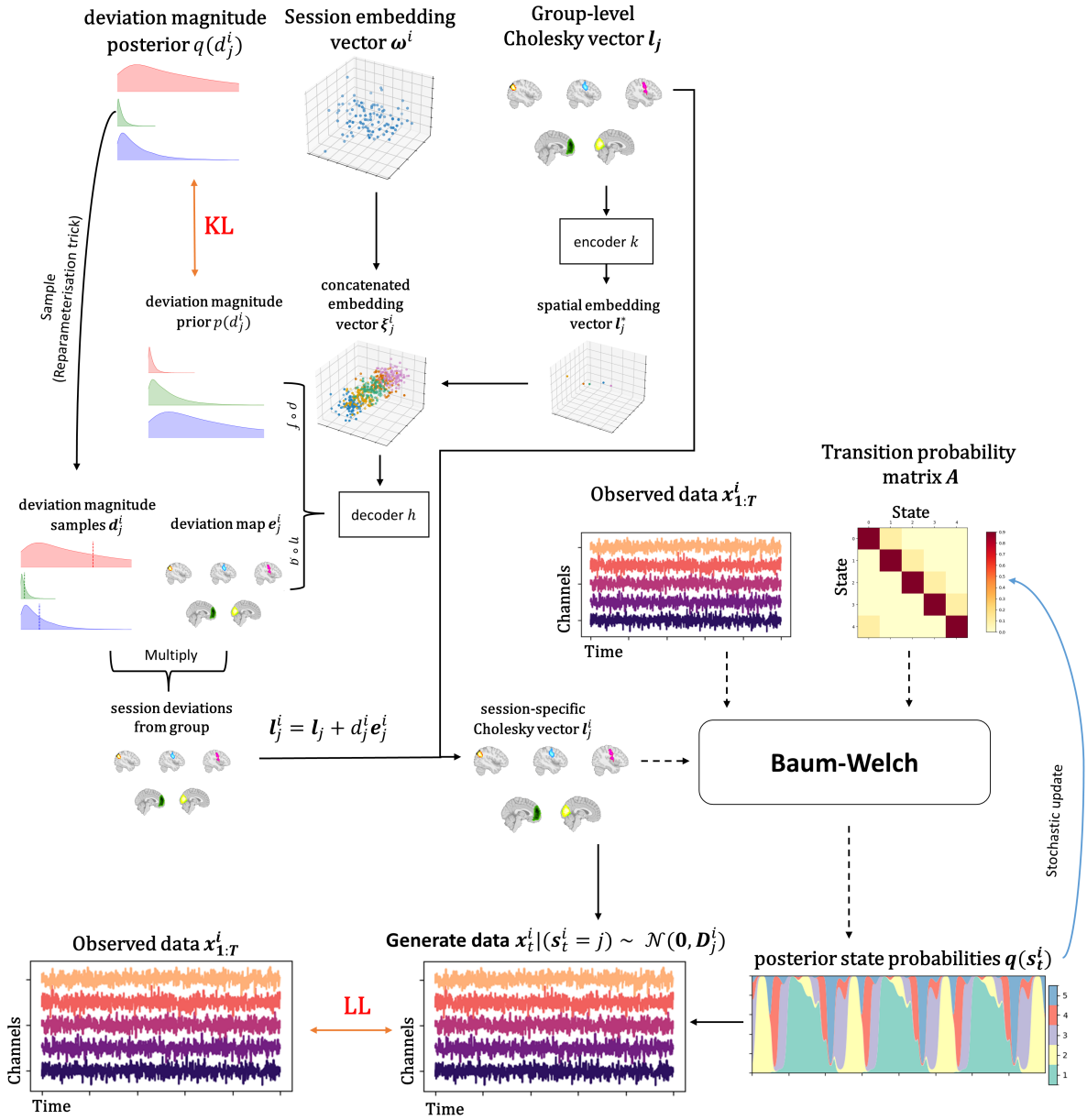

Figure 4: **The full HIVE inference framework**. The session embedding vector $\boldsymbol{\omega}^i$, group-level Cholesky vector $\boldsymbol{l}_j$ are used to generate the deviation map $\boldsymbol{e}_j^i$ and the deviation magnitude prior distribution $p(d_j^i)$. We sample d_j^i from the deviation magnitude posterior distribution $q(d_j^i)$ using the reparameterisation trick and use the drawn samples, which give the session-specific Cholesky vector $\boldsymbol{l}_j^i$, together with the deviation map $\boldsymbol{e}_j^i$ and group-level Cholesky vector $\boldsymbol{l}_j$. The prior and posterior distributions $p(d_j^i), q(d_j^i)$ are used to calculate the **KL** term in the loss function. Subsequently, we get the posterior state probabilities $q(\boldsymbol{s}_t^i)$ with the Baum-Welch algorithm which takes the observed data $\boldsymbol{x}_{1:T}^i$, the transition probability matrix $\boldsymbol{A}$ and the session-specific Cholesky vectors $\boldsymbol{l}_j^i$ as inputs. Finally, we calculate the **LL** term in the loss function with the observed data $\boldsymbol{x}_{1:T}^i$. The black solid arrows in the plot show connections where gradient flows through during back-propagation and the dashed black arrows show connections where gradient does not flow through. The blue arrow shows the stochastic update step for the transition probability matrix $\boldsymbol{A}$.

During training, we employed two annealing techniques to help find the global minimum. The first one is KL annealing ([6]) where we start the training without the KL term in the loss and gradually increase the contribution of the KL term, which is a common practice for training recurrent neural network-based variational autoencoders ([6]). The second one is that we anneal the sampling process of the variational distribution of deviation magnitudes $q(\boldsymbol{d})$ when applying the reparemeterisation trick:

$$\hat{\boldsymbol{d}} = (1-\kappa)\bar{\boldsymbol{d}} + \kappa \boldsymbol{d}^*, \tag{12}$$

where $\bar{\boldsymbol{d}}$ is the expectation of $q(\boldsymbol{d})$, $\boldsymbol{d}^*$ is a sample from $q(\boldsymbol{d})$ and κ is the annealing factor that is zero at the beginning of the training and gradually increases to one. We find this significantly improves the convergence of the model and we believe this serves as an exploration-exploitation mechanism, which is extensively studied in the fields of reinforcement learning ([44]) and Bayesian Optimisation ([24]) . At the beginning of the training, only the expectation is used and the gradient will intend to push the mean deviations to the correct position. This is the exploration phase. As training progresses, we gradually take into account the variance and higher moments of the variational distribution, which allows the training to fine tune the variational parameters. This is the exploitation phase.

2.6 Datasets

2.6.1 Simulated datasets

In the simulation studies, we simulate data with an HMM. To do this, we specify a transition probability matrix and randomly simulate covariance matrices for each state (sessions can have different covariances). Subsequently, a Markov chain is simulated with the pre-specified transition probability matrix. Lastly, at each time point, data is simulated using a multivariate Gaussian distribution with the state covariance matrix which is active at the given time. Gaussian noise is also added to the final time series data. The simulated data for the 3 simulation studies are described below.

- **Simulation 1**. We simulate data with $N = 10, T_i = 25,600, \forall i \in [N], N_c = 11, J = 5$. Here the state covariance matrices of sessions 1 and 2 are altered and all other sessions have the same unaltered group-level covariances - session 1 has increased variance in channel 1 and decrease variance in channel 2 whilst session 2 has decreased variance in channel 1 and increased variance in channel 2. For both session 1 and 2, the increased variances are 5 times and the decreased variances are 20% of the group-level variance.

- **Simulation 2**. Here $N = 100, T_i = 3,000, \forall i \in [N], N_c = 40, J = 5$. Sessions are assigned into 3 groups and an embedding vector for each session is simulated according to the session's assigned group. Session-specific covariances are simulated based on the simulated embedding vectors. (See A.2 for details).

- **Simulation 3**. In this study 4 datasets are simulated with $N \in \{5, 10, 50, 100\}$ and $T_i = 3000, \forall i \in [N], N_c = 40, J = 5$. Session-specific covariances are simulated in the same way as in Simulation 2.

2.6.2 Real MEG data

We demonstrate the use cases of the proposed model with 3 publicly available MEG datasets, including two resting-state and one visual task dataset. The datasets are source-reconstructed and parcellated to 38 regions, which matches well with how MEG data is typically inputted into the vanilla HMM. We describe the steps of data processing before model training below.

- **Raw data**. The first resting-state dataset ([37], we refer to this dataset as the **Cam-CAN** dataset) contains eyes-closed data of 612 healthy participants. These data were collected using an Elekta Neuromag Vectorview 306 scanner at a sampling frequency of 1kHz. A highpass filter of 0.03Hz and MaxFilter were applied. In the visual task MEG dataset ([43], we refer to this dataset as the **Wakeman-Henson** dataset), each of the 19 health participants were scanned 6 times, during which 3 types of visual stimuli were shown to the participants. The data were also collected using an Elekta Neuromag Vectorview 306 scanner. The second resting-state dataset was collected suing a 275-channel CTF scanner. This dataset (we refer to this dataset as the **Nottingham** dataset) contains eyes-closed data of 64 healthy participants, collected at Nottingham University, UK as part of the MEGUK partnership.

- **Preprocessing**. The Cam-CAN and Nottingham datasets were preprocessed with the same pipeline using the OHBA software library (OSL). The data were band-pass filtered between 0.5Hz and 125Hz, followed by a notch filter at 50Hz and 100Hz. Then the data were downsampled to 250Hz before automated bad segment and bad channel detection were applied to remove abnormally noisy segments and channels of the recording. Finally, an independent component analysis (ICA) with 64 components was applied to identify artifacts. The preprocessing of the Wakeman-Henson dataset is the same as above except that in the end, ICA with 40 components was applied.

- **Source reconstruction and parcellation.** Coregistration and source reconstruction were done using OSL. Structural data were coregistered with the MEG data using an iterative close-point algorithm and digitised head points acquired with a Polhemous pen were matched to individual subject's scalp surfaces extracted with FSL's BET tool ([25], [36]). The nose was not included in the coregistration as the structural MRI images were defaced. Preprocessed sensor data were source reconstructed onto an 8mm isotropic grid using a linearly constrained minimum variance beamformer ([38]). Voxels were then parcellated into 38 anatomically defined regions of interest, before the symmetric multivariate leakage reduction described in [10] was applied.

- **Data preparation.** Before model training, we follow the preparation steps described in [19]. The data were time-delay embedded with ± 7 lags. Then principal component analysis (PCA) was used to reduce the dimensionality to 80 channels before a standardisation step (z-transform) was applied to make sure each channel has zero mean and variance of one.

3 Results

3.1 Simulation

Here we show the results of training HIVE on the 3 simulated datasets described in Section 2.6.1 and compare these results with those from HMM-DE.

Simulation 1: Variability encoding block learns session-specific covariance deviations

We aim to test if patterns of deviations across different channels are learnt by the generative model, in particular, the variability encoding block, during inference. This is an important feature in the sense that we want the generative model to act as a prior that regularises how each session can deviate from the group. Firstly we can see from Figure 5a that both HMM-DE and HIVE can infer the group-level covariances correctly. What HIVE offers in addition is shown in Figure 5b where the inferred session embedding vectors are plotted. We can clearly see that the embedding vectors for sessions 1 and 2 are far away from the cluster formed by other sessions which have unaltered covariances, which matches well with the ground truth. The model's capability to encode deviation pattern across different channels in the generative model is illustrated in Figure 5c, in which the simulated patterns of deviation for sessions 1 and 2 can be generated from the trained variability encoding block. HMM-DE, though being capable of infer session-specific deviations, cannot generate the patterns of deviations. This is because intrinsically HMM-DE is a group-level model and its generative model is not capable of learning session-specific deviations.

Simulation 2: Underlying subpopulation structure is inferred by HIVE

In this simulation study, we aim to show that HIVE can recover the ground truth underlying population structure (simulated embedding vectors of sessions). This is demonstrated in Figure 6a, where the ground truth grouping of sessions is recovered by the inferred session embedding vectors. In particular, sessions which are close together in the simulated space (e.g. sessions 7 and 45) stay close in the inferred space. Figure 6b shows both HMM-DE and HIVE can infer the state time courses perfectly. Furthermore, we can see from Figure 6c that both approaches can recover the session-pairwise relationship between session-specific covariances, though HMM-DE overestimates the pairwise distance due to noise added to the data, whereas HIVE has the preferable behaviour of underestimating the pairwise distance due to the regularising effect of the prior on the deviation magnitude. This prior makes the inferred session-specific covariances more similar to the group average, when there is insufficient evidence available in the data to do otherwise.

Simulation 3: HIVE performs more accurate inference than HMM-DE.

Now we focus on comparing HIVE and HMM-DE in terms of accuracy of inferred session-specific covariances. Both HIVE and HMM-DE are trained 10 times to get a more accurate and stable estimation of the accuracy. From Figure 7 we can see HIVE can infer more accurate session-specific covariances than HMM-DE. Crucially, with the number of samples per session being kept the same, the accuracy for HIVE improves with more sessions in the dataset while that for HMM-DE does not have a significant improvement. This shows

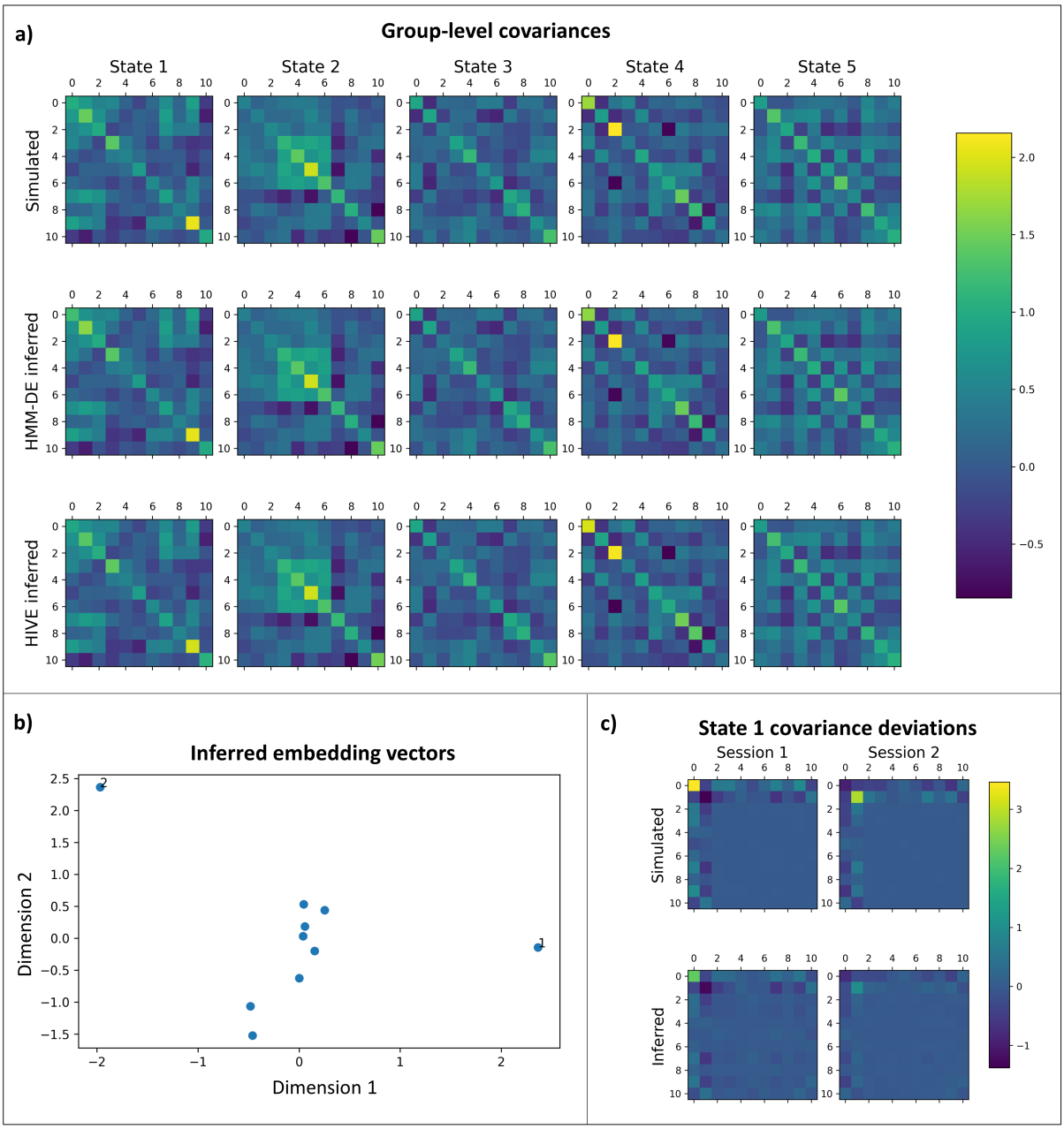

Figure 5: **Variability encoding block learns session-specific covariance deviations.** *Results obtained using Simulation 1.* a) Simulated (top row), HMM-DE inferred (middle row) and HIVE inferred (bottom row) group-level covariances. The columns show covariances of different states. b) Inferred session embedding vectors. Sessions 1 and 2 are annotated. c) Simulated deviations (top row) and deviations from trained generative model of HIVE (bottom row) for state 1 for session 1 (left column) and for session 2 (right column).

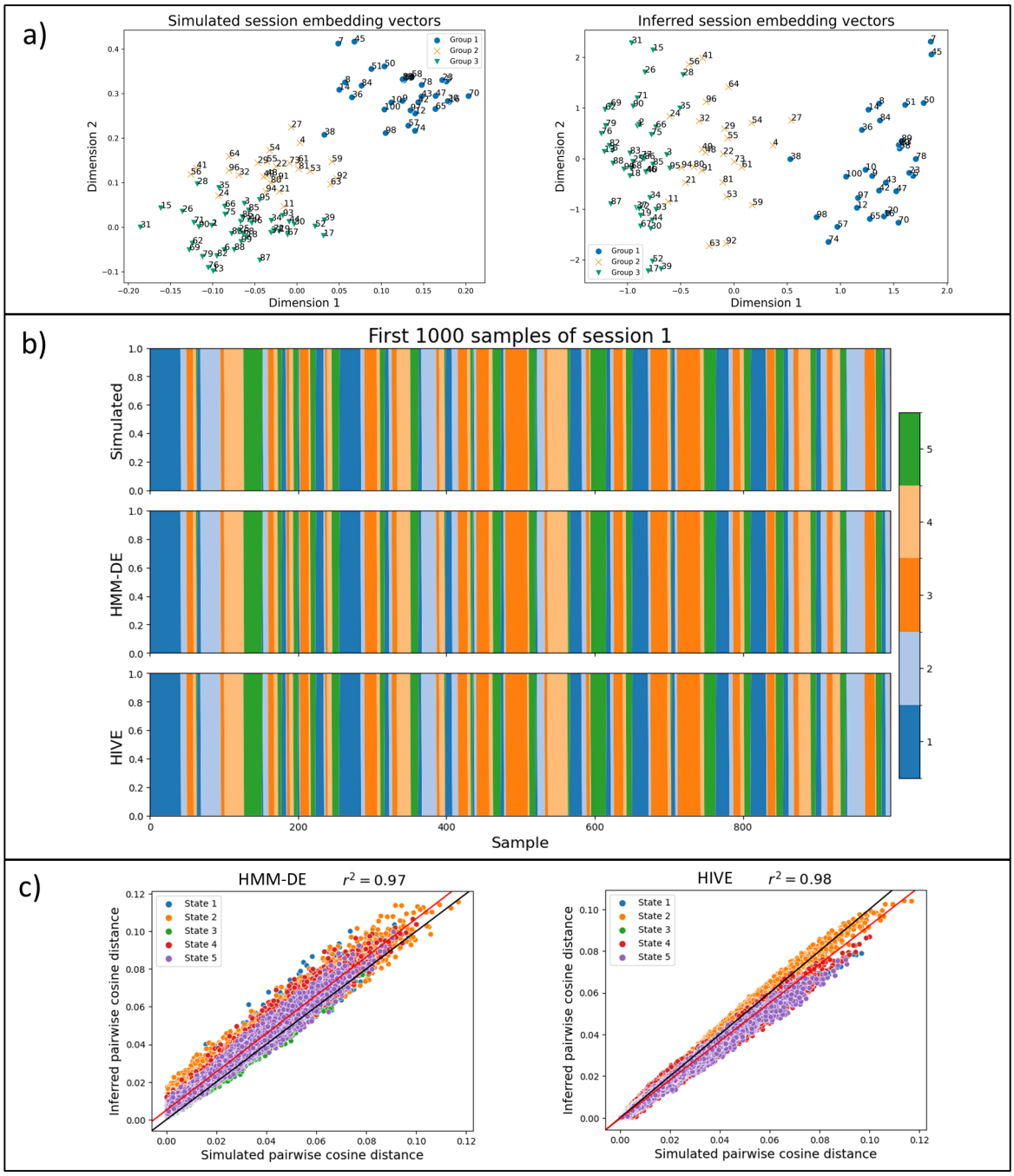

Figure 6: **HIVE infers the underlying subpopulation structure.** *Results obtained using Simulation 2.* a) Simulated (left) and inferred (right) session embedding vectors. Each point is marked and coloured by the ground truth group assignment, and is annotated by the session number. b) Simulated (top), HMM-DE inferred (middle) and HIVE inferred (bottom) state time courses. c) Session-pairwise cosine distance of simulated (x-axis) against inferred (y-axis) covariances from HMM-DE (left) and HIVE (right). The black line shows the $y = x$ line, which corresponds to optimal performance, and the red line is a fitted line through the points, with the coefficient of determinant r^2 displayed in the title.

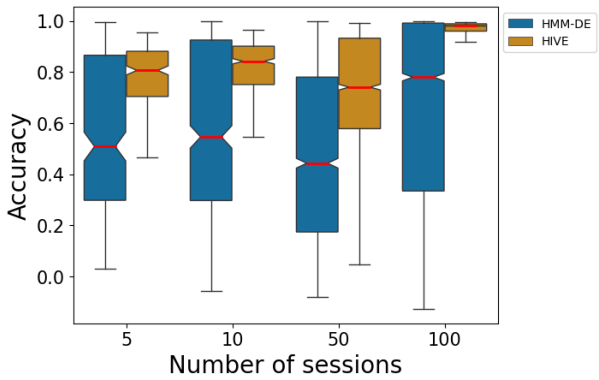

Figure 7: **HIVE performs more accurate inference than HMM-DE.** *Results obtained using Simulation 3.* Accuracy of session-specific covariances (y-axis) plotted against number of sessions in the dataset (x-axis) for HMM-DE (blue) and HIVE (orange). Here the number of samples per session is always the same. Correlation between inferred and simulated session-specific covariances is used as a metric for accuracy.

the advantage of the variability encoding block in making use of data of heterogeneous sessions to help infer on every single session.

3.2 Real MEG data

3.2.1 HIVE reveals similarities and differences between MEG recordings

The data we use in this section is the Wakeman-Henson dataset described in Section 2.6. The embedding vectors in our model are not only limited to subjects, they can also be applied to any type of variability we want to characterise (e.g. session variability). Ideally, we would expect the sessions for a subject to be more similar than sessions for different subjects. In this case, we assign each session an embedding vector and train HIVE on this data with $n_\omega = 10$ (See Section A.5). The session-pairwise cosine distances of embedding vectors are plotted in Figure 8a, which shows a clear block diagonal structure - session embedding vectors from the same subject are closer together than those from different subjects. This shows the model is able to identify certain recordings which have similar deviations from the group, despite HIVE being trained in an unsupervised manner with no knowledge of which sessions belong to which subjects.

To compare with the traditional approaches, we train both HIVE and HMM-DE on this dataset. In Figure 8b, the session-pairwise L2 distances of inferred covariances from both approaches are plotted. Although the same block diagonal structure can be seen in both approaches, it is much clearer in the case of HIVE. One can observe there is a particular session (session 3 of subject 8) that has much higher distance with all other sessions, especially with HMM-DE. This is related to the fact that this particular session has very different profiles from other sessions (See Section A.3). In order to quantify the advantages with HIVE, we employ 3 different metrics - Silhouette score ([35]), Davies-Bouldin score ([13]) and Calinski-Harabasz score ([8]) for assessing how well the inferred covariances form distinct and well-separated subject clusters. Figure 8c shows that with all three metrics, HIVE inferred covariances form tighter and more distinct clusters than dual estimated covariances from HMM-DE.

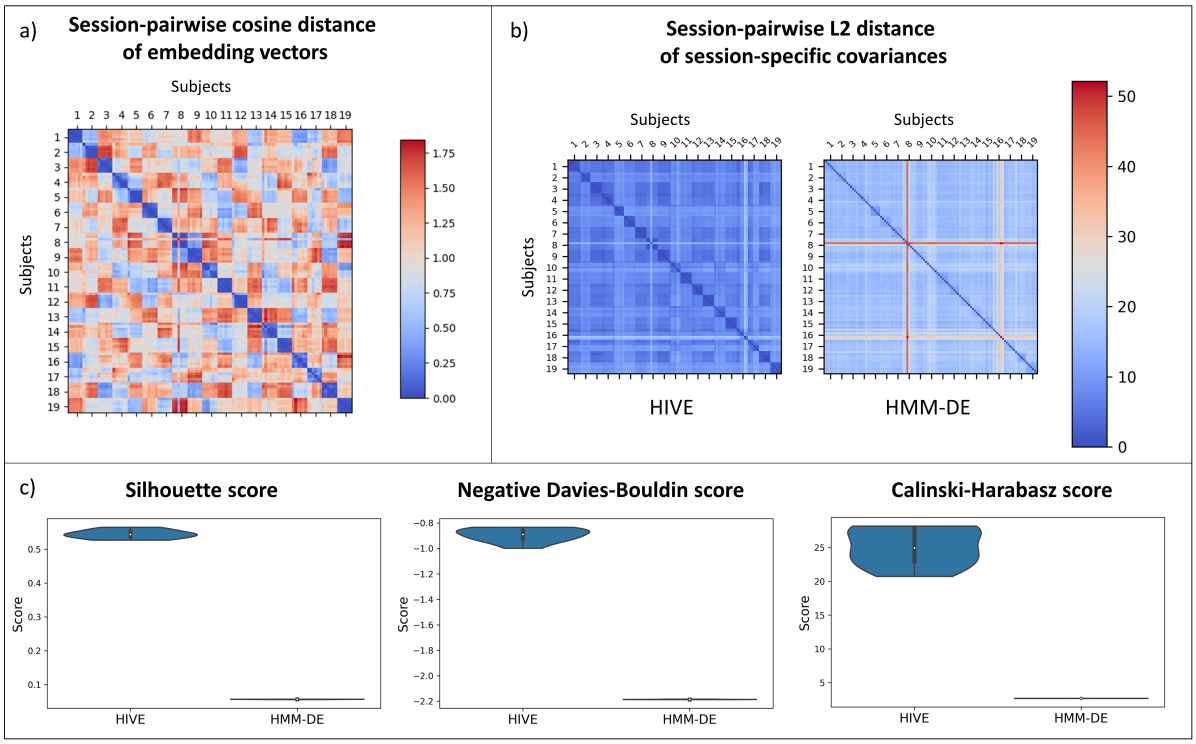

Figure 8: **HIVE reveals similarities and differences between MEG recordings.** *Results obtained on the Wakeman-Henson dataset.* a) Session-pairwise cosine distance of inferred embedding vectors. Embedding vectors for the same subject are grouped together and have smaller distance btween them. b) Session-pairwise L2 distance of inferred covariances from HIVE (left) and HMM-DE (right). c) Clustering metrics - Silhouette score (left), negative Davies-Bouldin score (middle), Calinski-Harabasz score (right), based on subject labels for 10 independent runs of both approaches. Higher values for these metrics indicate better clustering.

3.2.2 Embedding vectors reflect meaningful variability regarding scanner

In this study, we want to test if our model is able to differentiate data acquired by different scanners. The data we train on include data from two different resting-state MEG datasets (Nottingham and Cam-CAN) described in Section 2.6. There are 128 subjects in total and 64 subjects from each of the datasets. To avoid the possibility that the model is biased towards any of the datasets, we match the age group and sex profiles, i.e. in each of the age groups and sex, there is same number of subjects from each dataset. Here we choose $n_\omega = 20$ (See Section A.5). We can clearly see two clusters of embedding vectors inferred in Figure 9a, and at the same time an age gradient in the embedding space. This means scanner type and age information are simultaneously encoded (in different directions in the embedding space) by the embedding vectors, despite the fact that HIVE is trained unsupervised with no knowledge that the sessions were from different scanner types. We can also see a block diagonal structure in Figure 9b where subjects from the same dataset have smaller pairwise cosine distances of their inferred embedding vectors. With the clustering metrics described in Section 3.2.1, we see from Figure 9c that subject-specific covariances from HIVE form better-defined clusters based on scanners/sites.

HIVE also provides a way to summarise the differences in state-specific spectral content between scanner types. Ten nearest neighbours in the embedding space are selected for each of the dataset centroids and are used to give a representation of spectral contents of both datasets. In Figure 9d, we can see these differences between the "centroids" of the datasets. In particular, subjects from the Nottingham dataset generally have lower power than those from Cam-CAN, especially in posterior regions and alpha band.

3.2.3 HIVE infers age structure in the population more accurately than dual estimation

Here we use the large dataset Cam-CAN, described in Section 2.6, that consists of 612 healthy subjects and choose $n_\omega = 50$ (See Section A.5). From Figure 10a, We can clearly see an age gradient from darker to lighter dots in both plots. This means after training, age information is encoded by the embedding vectors of the subjects, despite the fact that HIVE is trained unsupervised with no knowledge of the subjects' ages. In order to see if the learnt hidden representation given by the embedding vectors helps improve the model's ability to distinguish between subject demographics, we try to predict age with the inferred subject-specific covariances. Here HMM-DE and HIVE with increasing embedding dimensions $3, 5, 10, 20, 50, 100$ are trained on this dataset. Each of the 7 models (one HMM-DE and six HIVE models) are trained 10 times independently and the best model is chosen based on free energy. Next the inferred subject-specific covariances are projected by PCA, whose first few PCs are used as regressors in a ridge regression to predict subject age. A 20-fold nested cross validation (See A.4 for details) is used to select the best hyperparameters and to give a distribution of prediction accuracy by each of the models, and the results are shown in Figure 10b. The coefficient of determinant (r^2) is used as a measure of prediction accuracy and it gradually increases with the embedding dimension until there is a significant improvement over HMM-DE when using embedding dimension of 50. There is a slight drop in accuracy when using embedding dimension of 100, which could be due to overfitting. Notice that during embedding dimension selection, we do not see a significant decrease in free energy if we increase n_ω from 50 to 100 (Figure 12).

Centroids of subjects of age groups < 36 and > 72 are used as representatives of

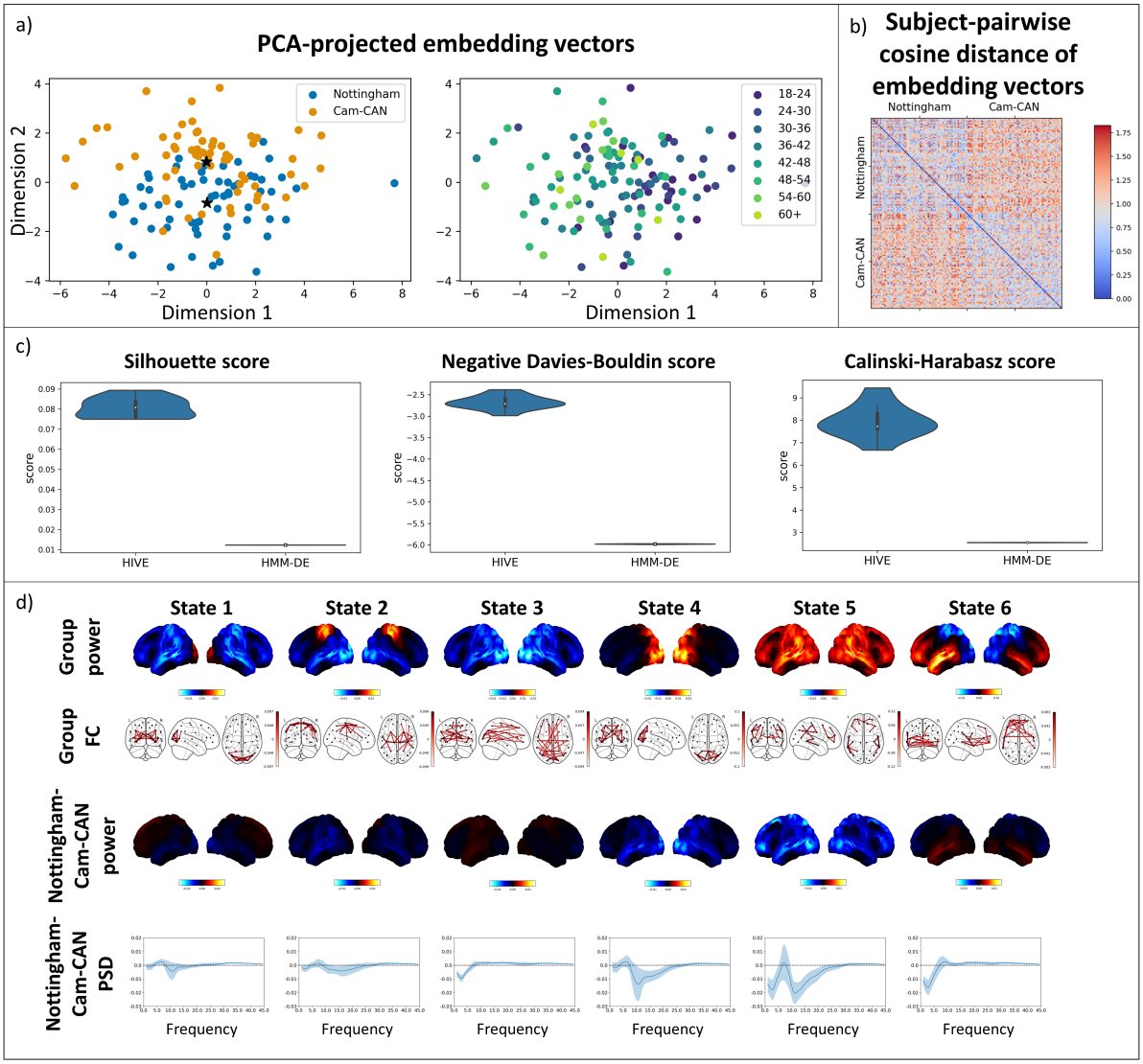

Figure 9: **Embedding vectors reflect meaningful variability regarding scanner/site**. *Results obtained with a combination of the Nottingham and Cam-CAN datasets.* a) PCA-projected embedding vectors coloured with different datasets (left, black stars indicate centroids of embedding vectors from the two datasets) and different age groups (right, older subjects are coloured with lighter colours). b) Subject-pairwise cosine distance of embedding vectors. c) Clustering metrics - Silhouette score (left), negative Davies-Bouldin score (middle), Calinski-Harabasz score (right), based on subject labels for 10 independent runs for both approaches. d) 6 states are inferred with HIVE. The first and second row shows the group level power (red areas show above average and blue areas show below average power across states) and FC maps (top 3% edges are plotted). The third row shows power difference (red areas show higher and blue areas show lower power in Nottingham subjects than Cam-CAN subjects) for each state between centroids of both datasets. The bottom row shows the difference in PSDs (solid lines show the means and shaded areas show one standard deviation across the channels) across frequencies between datasets.

young and old subjects. Similar to the analysis in Section 3.2.2, we can find the nearest neighbours (here we choose 20 subjects) for both representatives and get the power maps, PSDs for young and old subjects. We show the results in Figure 10c. We can conclude that HIVE can discover meaningful age patterns in the data.

4 Discussion

The results above show several key highlights of the proposed model. Firstly, from simulated data studies in Section 3.1, we show that during inference, the variability encoding block learns session-specific deviations across different channels. We also see that the model can accurately infer the pairwise relationships between sessions, i.e. recover the structure within a population. Furthermore, by learning the underlying population structure through the embedding vectors, the model can perform more accurate inference than dual estimation on individual sessions.

Secondly, similar to the approach taken in PROFUMO ([22, 21]), session deviations from the group-level estimates are generated through a Bayesian prior. The additional feature of embedding vectors allows the model to find subpopulations in the group. By training HIVE on the Wakeman-Henson dataset, we observe that inter-subject variability is much greater than inter-session variability (Section 3.2.1). This was also found in fMRI literature ([20]). Although the HIVE inferred covariances have slightly underestimated deviations from the group due to the regularising effect of the Bayesian prior, it is robust to noise in the dataset. Furthermore, the covariances from this approach forms better-separated subject clusters due to this de-noising effect compared to the traditional HMM-DE.

Thirdly, in Section 3.2.2, we can observe systematic differences between scanners by training HIVE on data from different sites. The model is trained in an unsupervised fashion and is unaware of the scanners/sites. By finding the centroids of embedding vectors of both datasets and pooling over nearest neighbours, we are able to find the power maps, FC maps and PSD profiles of representatives from both datasets. This allows us to study the systematic differences between scanners. Moreover, HIVE inferred covariances form better-separated clusters than HMM-DE inferred ones based on dataset.

Finally, in Section 3.2.3, by training on a large-scale dataset, we show that the inferred embedding vectors contains age information and there is a clear age gradient. By training classifiers to predict subject age with inferred subject-specific covariances, we see that accuracy increases with embedding dimension and at some point, HIVE inferred features are able to outperform HMM-DE inferred ones. By picking representatives of young and old subjects, and again pooling over information of nearest neighbours, we can acquire brain networks and PSD profiles of young and old brains. This could be potentially useful in aging studies.

Future work. In this paper we present results of HIVE, but the novel variability encoding block is also compatible with another popular model - DyNeMo ([19]), for inferring dynamic functional networks. We believe HIVE provides a way to learn the underlying population structure, which can be used to regularise the estimates of subject/session-specific measures (e.g. spectral content). This could be potentially useful in boutique studies which has relatively small number of subjects and contains subjects with specific demographics. By training our proposed model on a large scale dataset and using the learned population structure, subject-specific estimates could be better inferred on the small dataset. Additionally, we believe this model has the potential to be trained on fMRI

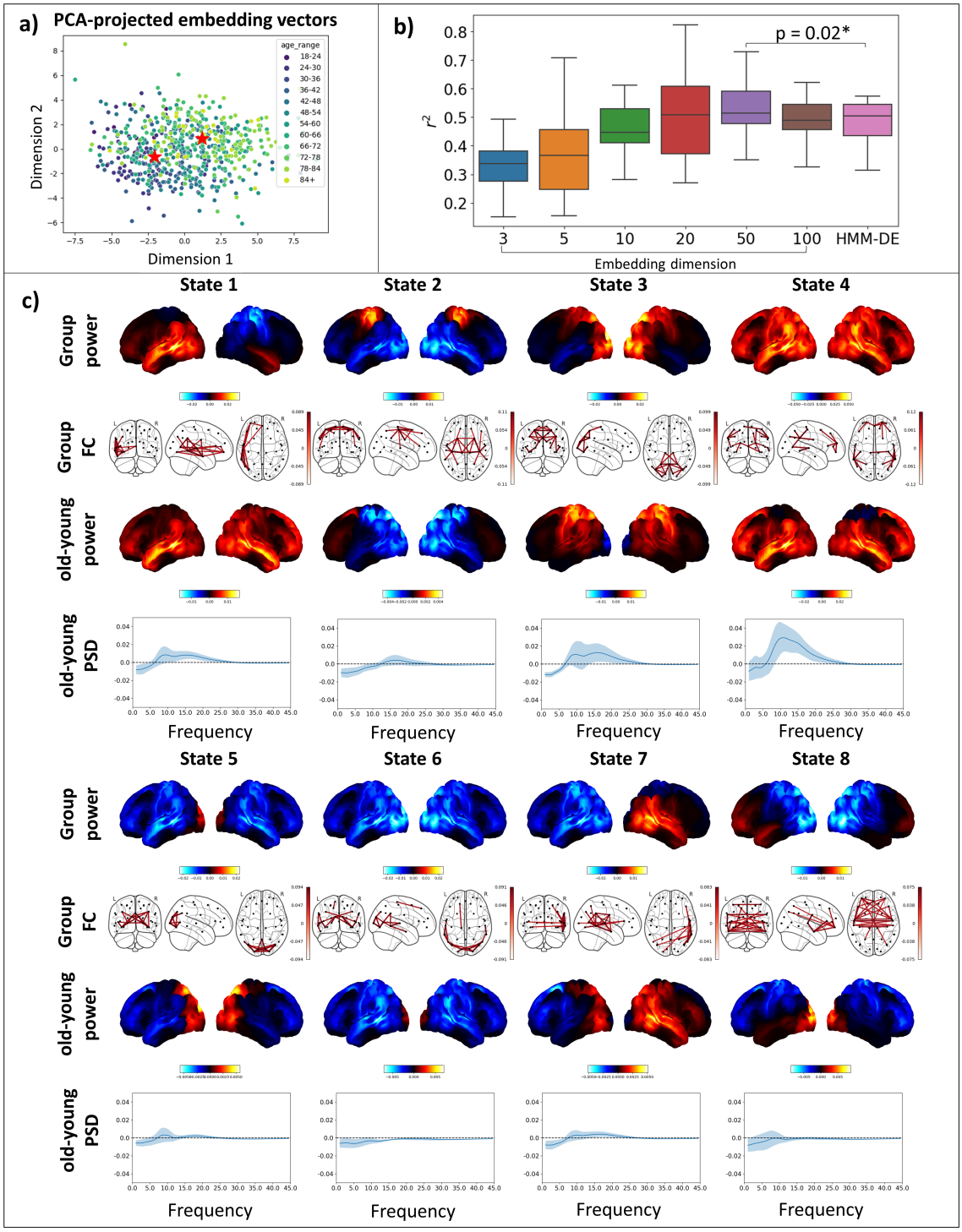

Figure 10: **HIVE infers age structure in the population more accurately than dual estimation**. *Results obtained using the Cam-CAN dataset.* a) Inferred embedding vectors projected to 2 dimensions with PCA. Darker-coloured points are younger participants and lighter-coloured ones are older participants. The two red stars show centroids of subjects with age < 36 and > 72. b) 20-fold cross validated age prediction accuracy for HIVE with different embedding dimensions and HMM-DE. c) 8 states are inferred by HIVE. Group level power (red areas show above average and blue areas show below average power across states) and FC maps (top 3% edges are plotted) are shown as well as the differences in power (red areas show higher and blue areas show lower power in older subjects than younger subjects) and PSDs (solid lines show the means and shaded areas show one standard deviation across the channels) between old and young subjects.

data where the number of subjects is much larger but number of samples per subject is much smaller. Furthermore, we have shown that HIVE gives better clustering results than dual estimation. A next step will be to investigate the possibility of utilising this advantage in demographics and disease prediction. Lastly, the model provides a natural way to provide subject/session information (e.g. task, age, sex, etc) via embedding vectors in a hierarchical manner. This could be potentially useful when there is an intrinsic subpopulation structure in the group (e.g. multiple sessions for different subjects and both task and resting-state data are available for the subjects).

Hyperparameters. Due to the additional complexity of the model, it has more hyper-parameters than HMM and DyNeMo, including the number of layers, number of neurons per layer in the decoders of the variability encoding block. However in practice, we find that results are very robust to the choice of these hyper-parameters.

The number of states/modes is an important hyper-parameter in HMM and DyNeMo that could affect the interpretation of the data. Previous studies show the variational free energy keeps decreasing with increasing number of state/modes ([2], [19]). We hypothesise that this could be due to variability in the data (See A.6). Modelling the variability in the data could be a potential solution to this problem.

5 Conclusion

We proposed a novel generative model that explicitly models variability within a population in a principled way. We provide a way to perform efficient inference on the model parameters and the algorithm is readily scalable to large amount of data. With a Bayesian prior, the model pools information across individuals for how they may deviate from the group average. The additional feature of embedding vectors allows the model to group together similar data and help the interpretation of more details in a population. This is an important step towards making use of the large amount of datasets collected with different protocols and we believe the above results demonstrate that the proposed model provides a novel perspective in population modelling, inference of dynamic brain networks and analysis of MEG data.

6 Credit authorship contribution statement

RH: Conceptualisation, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing - original draft, Writing - review and editing, Visualisation. CG: Conceptualisation, Methodology, Software, Data curation, Writing - review and editing. MW: Conceptualisation, Methodology, Data curation, Writing - review and editing, Supervision.

7 Acknowledgements

This research was supported by the National Institute for Health Research (NIHR) Oxford Health Biomedical Research Centre. The Wellcome Centre for Integrative Neuroimaging is supported by core funding from the Wellcome Trust (203139/Z/16/Z). RH is supported by the EPSRC Centre for Doctoral Training in Health Data Science (EP/S02428X/1). CG is supported by the Wellcome Trust (215573/Z/19/Z). MW is supported by the Wellcome

Trust (106183/Z/14/Z, 215573/Z/19/Z), the New Therapeutics in Alzheimer's Diseases (NTAD) study supported by UK MRC, the Dementia Platform UK (RG94383/RG89702) and the NIHR Oxford Health Biomedical Research Centre (NIHR203316). The views expressed are those of the author(s) and not necessarily those of the NIHR or the Department of Health and Social Care.

References

[1] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[2] Adam P Baker, Matthew J Brookes, Iead A Rezek, Stephen M Smith, Timothy Behrens, Penny J Probert Smith, and Mark Woolrich. Fast transient networks in spontaneous human brain activity. *elife*, 3:e01867, 2014.

[3] Christian F Beckmann, Marilena DeLuca, Joseph T Devlin, and Stephen M Smith. Investigations into resting-state connectivity using independent component analysis. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360(1457):1001–1013, 2005.

[4] Christian F Beckmann, Clare E Mackay, Nicola Filippini, Stephen M Smith, et al. Group comparison of resting-state fmri data using multi-subject ica and dual regression. *Neuroimage*, 47(Suppl 1):S148, 2009.

[5] David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017.

[6] Samuel R Bowman, Luke Vilnis, Oriol Vinyals, Andrew M Dai, Rafal Jozefowicz, and Samy Bengio. Generating sentences from a continuous space. *arXiv preprint arXiv:1511.06349*, 2015.

[7] Matthew J Brookes, Mark Woolrich, Henry Luckhoo, Darren Price, Joanne R Hale, Mary C Stephenson, Gareth R Barnes, Stephen M Smith, and Peter G Morris. Investigating the electrophysiological basis of resting state networks using magnetoencephalography. *Proceedings of the National Academy of Sciences*, 108(40):16783–16788, 2011.

[8] Tadeusz Caliński and Jerzy Harabasz. A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, 3(1):1–27, 1974.

[9] Omar Chehab, Alexandre Defossez, Jean-Christophe Loiseau, Alexandre Gramfort, and Jean-Remi King. Deep recurrent encoder: A scalable end-to-end network to model brain signals. *arXiv preprint arXiv:2103.02339*, 2021.

[10] Giles L Colclough, Matthew J Brookes, Stephen M Smith, and Mark W Woolrich. A symmetric multivariate leakage correction for meg connectomes. *Neuroimage*, 117:439–448, 2015.

[11] Michael W Cole, Danielle S Bassett, Jonathan D Power, Todd S Braver, and Steven E Petersen. Intrinsic and task-evoked network architectures of the human brain. *Neuron*, 83(1):238–251, 2014.

[12] Richard Csaky, Mats WJ Van Es, Oiwi Parker Jones, and Mark Woolrich. Group-level brain decoding with deep learning. *Human Brain Mapping*, 44(17):6105–6119, 2023.

[13] David L Davies and Donald W Bouldin. A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence*, (2):224–227, 1979.

[14] Francesco De Pasquale, Stefania Della Penna, Abraham Z Snyder, Laura Marzetti, Vittorio Pizzella, Gian Luca Romani, and Maurizio Corbetta. A cortical core for dynamic integration of functional networks in the resting human brain. *Neuron*, 74(4):753–764, 2012.

[15] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society: series B (methodological)*, 39(1):1–22, 1977.

[16] Pascal Fries. Rhythms for cognition: communication through coherence. *Neuron*, 88(1):220–235, 2015.

[17] Paul A Gagniuc. *Markov chains: from theory to implementation and experimentation*. John Wiley & Sons, 2017.

[18] C Gohil, R Huang, E Roberts, MWJ van Es, AJ Quinn, D Vidaurre, and MW Woolrich. osl-dynamics: A toolbox for modelling fast dynamic brain activity. *eLife*, 12, 2023.

[19] Chetan Gohil, Evan Roberts, Ryan Timms, Alex Skates, Cameron Higgins, Andrew Quinn, Usama Pervaiz, Joost van Amersfoort, Pascal Notin, Yarin Gal, et al. Mixtures of large-scale dynamic functional brain network modes. *NeuroImage*, 263:119595, 2022.

[20] Caterina Gratton, Timothy O Laumann, Ashley N Nielsen, Deanna J Greene, Evan M Gordon, Adrian W Gilmore, Steven M Nelson, Rebecca S Coalson, Abraham Z Snyder, Bradley L Schlaggar, et al. Functional brain networks are dominated by stable group and individual factors, not cognitive or daily variation. *Neuron*, 98(2):439–452, 2018.

[21] Samuel J Harrison, Janine D Bijsterbosch, Andrew R Segerdahl, Sean P Fitzgibbon, Seyedeh-Rezvan Farahibozorg, Eugene P Duff, Stephen M Smith, and Mark W Woolrich. Modelling subject variability in the spatial and temporal characteristics of functional modes. *NeuroImage*, 222:117226, 2020.

[22] Samuel J Harrison, Mark W Woolrich, Emma C Robinson, Matthew F Glasser, Christian F Beckmann, Mark Jenkinson, and Stephen M Smith. Large-scale probabilistic functional modes from resting state fmri. *NeuroImage*, 109:217–231, 2015.

[23] Cameron Higgins, Yunzhe Liu, Diego Vidaurre, Zeb Kurth-Nelson, Ray Dolan, Timothy Behrens, and Mark Woolrich. Replay bursts in humans coincide with activation of the default mode and parietal alpha networks. *Neuron*, 109(5):882–893, 2021.

[24] Ali Jalali, Javad Azimi, and Xiaoli Fern. Exploration vs exploitation in bayesian optimization. *CoRR*, 2012.

[25] Mark Jenkinson, Mickael Pechaud, Stephen Smith, et al. Bet2: Mr-based estimation of brain, skull and scalp surfaces. In *Eleventh annual meeting of the organization for human brain mapping*, volume 17, page 167. Toronto., 2005.

[26] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

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

[28] John K Kruschke and Wolf Vanpaemel. Bayesian estimation in hierarchical models. *The Oxford handbook of computational and mathematical psychology*, pages 279–299, 2015.

[29] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

[30] Henry Luckhoo, Joanne R Hale, Mark G Stokes, Anna Christina Nobre, Peter G Morris, Matthew J Brookes, and Mark William Woolrich. Inferring task-related networks using independent component analysis in magnetoencephalography. *Neuroimage*, 62(1):530–541, 2012.

[31] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[32] Lisa D Nickerson, Stephen M Smith, Döst Öngür, and Christian F Beckmann. Using dual regression to investigate network shape and amplitude in functional connectivity analyses. *Frontiers in neuroscience*, 11:115, 2017.

[33] Marius-Constantin Popescu, Valentina E Balas, Liliana Perescu-Popescu, and Nikos Mastorakis. Multilayer perceptron and neural networks. *WSEAS Transactions on Circuits and Systems*, 8(7):579–588, 2009.

[34] Andrew J Quinn, Diego Vidaurre, Romesh Abeysuriya, Robert Becker, Anna C Nobre, and Mark W Woolrich. Task-evoked dynamic network analysis through hidden markov modeling. *Frontiers in neuroscience*, 12:603, 2018.

[35] Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.

[36] Stephen M Smith. Fast robust automated brain extraction. *Human brain mapping*, 17(3):143–155, 2002.

[37] Jason R Taylor, Nitin Williams, Rhodri Cusack, Tibor Auer, Meredith A Shafto, Marie Dixon, Lorraine K Tyler, Richard N Henson, et al. The cambridge centre for ageing and neuroscience (cam-can) data repository: Structural and functional mri, meg, and cognitive data from a cross-sectional adult lifespan sample. *neuroimage*, 144:262–269, 2017.

[38] Barry D Van Veen and Kevin M Buckley. Beamforming: A versatile approach to spatial filtering. *IEEE assp magazine*, 5(2):4–24, 1988.

[39] Diego Vidaurre, Romesh Abeysuriya, Robert Becker, Andrew J Quinn, Fidel Alfaro-Almagro, Stephen M Smith, and Mark W Woolrich. Discovering dynamic brain networks from big data in rest and task. *NeuroImage*, 180:646–656, 2018.

[40] Diego Vidaurre, Alberto Llera, Stephen M Smith, and Mark W Woolrich. Behavioural relevance of spontaneous, transient brain network interactions in fmri. *Neuroimage*, 229:117713, 2021.

[41] Diego Vidaurre, Andrew J Quinn, Adam P Baker, David Dupret, Alvaro Tejero-Cantero, and Mark W Woolrich. Spectrally resolved fast transient brain states in electrophysiological data. *Neuroimage*, 126:81–95, 2016.

[42] Diego Vidaurre, Stephen M Smith, and Mark W Woolrich. Brain network dynamics are hierarchically organized in time. *Proceedings of the National Academy of Sciences*, 114(48):12827–12832, 2017.

[43] Daniel G Wakeman and Richard N Henson. A multi-subject, multi-modal human neuroimaging dataset. *Scientific data*, 2(1):1–10, 2015.

[44] Haoran Wang, Thaleia Zariphopoulou, and Xunyu Zhou. Exploration versus exploitation in reinforcement learning: a stochastic control approach, 2018. *International Center for Decision and Risk Analysis, Jindal School of Management, The University of Texas at Dallas, TX75080, USA, and School of Data Science, City University Hong Kong E-mail address: alain. bensoussan@ utdallas. edu, abensous@ cityu. edu. hk*, 1812.

[45] Fanny Yang, Sivaraman Balakrishnan, and Martin J Wainwright. Statistical and computational guarantees for the baum-welch algorithm. *The Journal of Machine Learning Research*, 18(1):4528–4580, 2017.

A Appendix

A.1 Dual estimation of HMM

Let γ_{jt} be the state probability of state j at time t, then the session-specific means $\{\hat{\mu}_j\}_{j=1}^J$ and covariances $\{\hat{D}_j\}_{j=1}^J$ are estimated by

$$\begin{aligned} \hat{\mu}_j &= \frac{\sum_{t=1}^T \gamma_{jt} x_t}{\sum_{t=1}^T \gamma_{jt}} \\ \hat{D}_j &= \frac{\sum_{t=1}^T \gamma_{jt}(x_t - \hat{\mu}_j)(x_t - \hat{\mu}_j)^T}{\sum_{t=1}^T \gamma_{jt}} \end{aligned} \tag{13}$$

A.2 Simulating session-specific covariances

Group level covariances D_j are generated randomly. Each embedding vector is assigned a group uniformly randomly from a set of groups and each embedding vector is generated with a multivariate Gaussian distribution with the same variance but different means depending on the group assignment. Next the encoder and two decoders in Figure 2 are randomly generated linear transformations and the subject deviation magnitudes d_j^i are

directly generated by applying the linear transformation on the concatenated embeddings ξ_j^i. Following the flow chart of Figure 2, we have generated session-specific covariances from different groups of sessions.

A.3 Outlier in the dataset

We see from Figure 11 that run 3 of subject 8 has a very different static PSD profile from the other runs. In particular the spike at around 26 Hz. Furthermore it has a very different session-pairwise correlation of the static covariances from the other runs.

A.4 Classifier for predicting age with subject-specific covariances

The subject age data are separated into 20 folds. For each fold, a classifer is trained on 19 folds and tested on the remaining 1 fold. During training, a 5-fold cross-validated grid search is used to select the best dimension of PCA projection and the regularisation strength of the ridge regression. PCA projection dimension is chosen from $\{5, 20, 50, 100\}$ and regularisation strength is chosen from $\{10^i : i \in \{-5, -4, -3, -2, -1, 0, 1, 2, 3\}\}$.

A.5 Choosing the embedding dimension n_ω on the real datasets

We choose n_ω with the procedure described in Section 2.5. For the model trained on the Wakeman-Henson dataset in Section 3.2.1, the decrease in free energy is insignificant when comparing $n_\omega = 10$ and $n_\omega = 20$. Hence for this dataset, we choose $n_\omega = 10$. For the model trained on the combined dataset in Section 3.2.2, the tests are all significant and we choose $n_\omega = 20$. For the model trained on the Cam-CAN dataset in Section 3.2.3, the tests are all significant until $n_\omega = 50$. Hence we choose $n_\omega = 50$.

A.6 Variability in data causes ever-decreasing free energy with number of states.

We simulate two datasets, one without variability, one with variability. The ground truth number of states for both datasets is 4. HMM is trained 5 times independently on both datasets and the free energy on training data is plotted in Figure 13. We see that when there is no variability, the free energy decreases with increasing number of states until it plateaus at the ground truth number of states. However, when there is variability, the free energy keeps decreasing with increasing number of states. This is because the model can keep finding more states to explain the variability in the data.

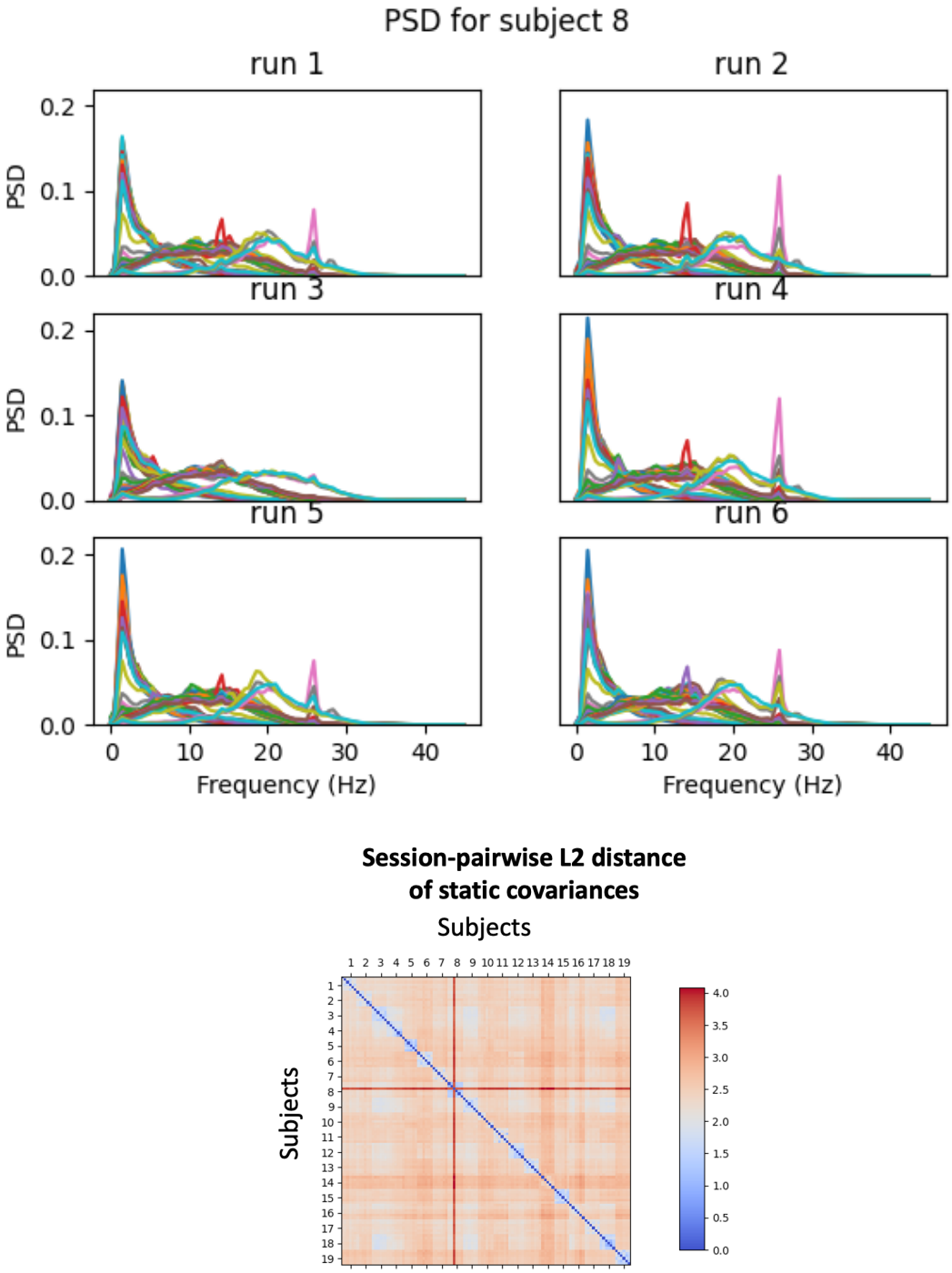

Figure 11: **Outlier in the dataset**. Top: PSD for the prepared data used for training is plotted for different sessions/runs of subject 8. Bottom: The session-pairwise correlation of the static covariances is plotted for the prepared data used for training. The outlier run is highlighted in red

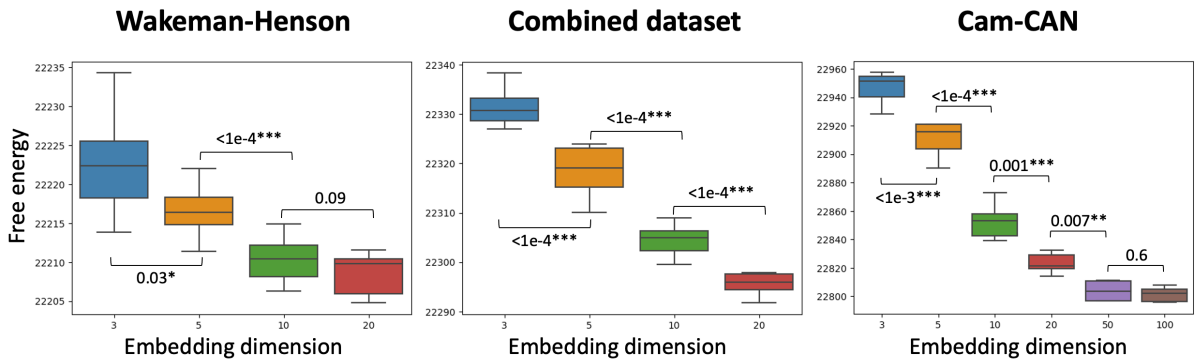

Figure 12: **Free energy is used to choose** n_ω. For each of the real dataset results, different models with different candidate embedding dimensions are trained 10 times independently. The free energy on training data is plotted against embedding dimension. The p-values of the progressive t-tests are plotted between the boxplots.

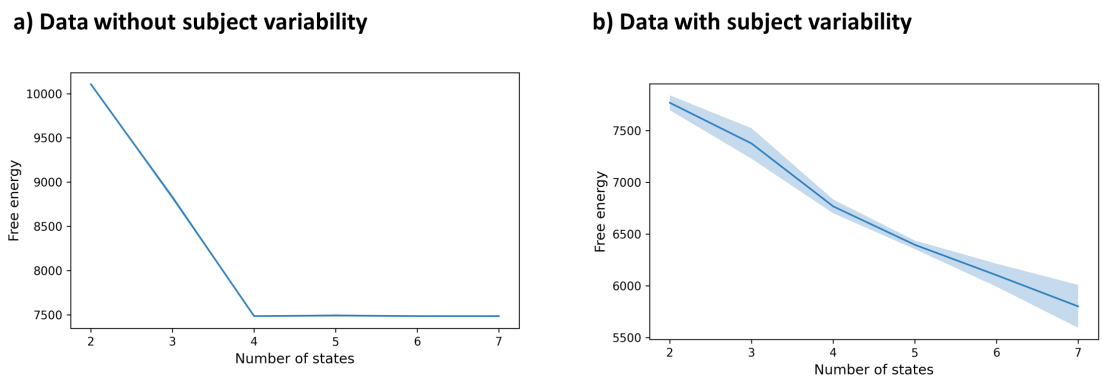

Figure 13: **Variability in the data can cause ever-decreasing free energy in HMM.** a) There is no subject variability in the data and variational free energy is plotted against number of states. b) There is subject variability in the data and variational free energy is plotted against number of states.