Spatio-temporal Representations and Decoding Cognitive Processes from fMRI

Firdaus Janoos[a,b,c,*], Raghu Machiraju[a], Shantanu Singh[a], Istvan Akos Morocz[b,c]

[a] *Dept. of Computer Science and Engineering, Ohio State University, Columbus, USA.*

[b] *Brigham and Women's Hospital, Boston, USA.*

[c] *Harvard Medical School, Boston, USA.*

Abstract

Revealing the spatio-temporal representational space in which different mental states are encoded [14] is an important step towards decoding the cognitive state of the subject from fMRI data in order to study human thought processes. Multi-variate classifiers are commonly used for brain state decoding, however are restricted to simple experimental paradigms with a fixed number of alternatives, and are limited in their representation of the temporal dimension of the task. Moreover, they learn a mapping from the data to experimental conditions, and therefore, do not explain the intrinsic patterns in the data.

In this paper, we present a purely data-driven approach to building a spatio-temporal representation of mental processes using a state-space formalism, without reference to experimental conditions. We propose an intuitive definition of functional similarity based on the distribution of activity on the functional networks of the brain, and from this derive a low-dimensional linear feature-space for the data. Efficient algorithms for estimating the parameters of the model, and a method for model-size selection are also proposed. We applied this method to a study on developmental dyscalculia and compared the models of healthy vs. affected subjects. The method provided quantitative support for hypotheses not available through regular analysis methods.

1. Introduction

In addition to functional localization and integration, the problem of determining whether the data encode some information about the cognitive state of the subject, and if so, how is this information represented has become an important research agenda in functional neuroimaging. This is especially relevant for neurological and psychiatric disorders like dementia, schizophrenia,

*Corresponding Author

Email addresses: firdaus@ieee.org (Firdaus Janoos), raghu@cse.ohio-state.edu (Raghu Machiraju), singhsh@cse.ohio-state.edu (Shantanu Singh), pisti@bwh.harvard.edu (Istvan Akos Morocz)

autism, multiple sclerosis, etc.[5], or common learning disabilities like dyslexia or dyscalculia[18], where group-level inference of spatial activity maps has been inconclusive due the high variation observed between subjects.

A popular approach in this direction has been the use of Multivariate Pattern Recognition (MVPR), which in contrast to the conventional forward methods such as general linear models, learns the reverse relationship from the distributed pattern of activation in an individual brain to the experimental conditions experienced during the scans. Typically linear classifiers, such as correlation-based classifiers [13], single-layer perceptrons[30], linear discriminant analysis [14], linear support vector machines (SVMs)[27] and Gaussian Naive Bayes[26] have been used due to simplicity of interpretation without significant loss of accuracy[17]. Such MVPR methods have been applied mainly to the study of visual (e.g.[13, 17, 30]) processing, but also auditory [24] perception, motor tasks[19], word recogition[27], and to emotional affects such as fear and deception[38]. This is part of a broader trend of machine learning in the analysis of neuroscientific recordings with applications in clinical psychology and cognitive neuroscience[29], brain-machine interfaces[10], real-time biofeedback[19], etc.

The main advantage of the multi-variate approach is its sensitivity to the distributed nature of cognitive processes by integrating information from groups of voxels that individually are weakly activated, but jointly may be highly structured with respect to the task. Also, it obviates the need for spatial smoothing, otherwise required to boost SNR, thereby preserving the information present in fine-grained spatial variations.

However, despite the obvious explanatory and predictive power of MVPR, there are still many open challenges. Firstly, to the best of our knowledge, such methods have only been used in experiments where subjects were presented with fixed number of alternatives not typical of natural perception. Generalization to complex paradigms, and further on, to real world situations poses a significant methodological challenge, given the non-linear nature of brain processing[14]. Also, such methods do not typically account for the temporal variations in mental processes. They make the assumption that all fMRI scans with the same label (i.e. experimental condition) have the same properties, although spatio-temporal information has been incorporated for block-design experiments by considering all the fMRI scans in one block as the feature vector[26]. As we shall show, preserving the temporal dimension of the task provides additional insight into the differences between healthy and affected populations. Moreover, as these methods learn a mapping from fMRI data to labels describing stimulus or subject behavior, their ability to explain the cognitive state of the subject is limited to behavioral correlates, not the unobservable (covert) cognitive states that might be present in the data[14].

In an attempt to address some of these issues, this paper presents a method to to determine the ***intrinsic spatio-temporal patterns*** in the data without reference to the experimental conditions, during a complex cognitive experimental paradigm. Here, mental processes are represented by a Hidden Markov Model (HMM) [4],which captures the concept of the functional brain transitioning through a cognitive state-space over time as it performs a task (c.f. Section 2). The first-order Markov assumption provides a trade-off between computational tractablity and the directed nature of thought, which requires tracking its recent history. The similarity of the (hidden) cognitive state at two different time-points is measured by comparing the patterns in their (unobserved) neural activation, from which the (observed) BOLD signal is generated via the hemodynamic response. The similarity between two activation patterns is defined by the differences in their distributions on the functional connectivity map of the brain, which gives rise to a

low-dimensional linear embedding for the fMRI data (c.f. Section 3). The effects of the hemodynamics on the observed data are modeled by a convolutive hemodynamic response function (HRF) [12], and then removed through marginalization. The proposed model violates the conditions necessary for the standard HMM algorithms, and new algorithms for efficient estimation are provided (c.f. Section 4). The correct model size is determined in an automated fashion by selecting the number of states that best describe the task being performed by the subject. The model is then applied to a study of developmental dyscalculia (DC)[28] to understand the effects of experimental parameters on the spatio-temporal patterns of mental processes for dyscalculics (DCs) vs. controls (c.f.Section 5). The Supplementary Material contains complete proofs, details about the algorithms and discussion of the results.

HMMs have been previously used in fMRI for determining the activation state of individual voxels [9, 15]. Activity detection has also been done with Hidden Markov Multiple Event Sequence Models [11], that pre-process the data into a series of spikes at each voxel to infer neural events. A multi-variate ARMA formalism was used in a Dynamical Components Analysis[39] to extract spatial components and their time-courses from fMRI data, given the experimental stimuli. A Hidden Process Model[16] was used to decompose the fMRI data into a set of spatio-temporal processes and their instantiations, selected from a set of pre-specified configurations of process instances. Dynamic Bayesian Networks[41] have been used to study the time-varying functional integration of a small number of pre-specified regions in the brain, from the interdependency structure of their average time-series.

In contrast to these, in the current work we have built a spatio-temporal representation of the global and instantaneous cognitive state of a subject in an unsupervised fashion solely from the fMRI data. Since this method does not rely on knowledge of the experimental conditions, it can be used for arbitrarily complex paradigms. From this representation, we can draw inferences about task-related effects on the models of individual subjects and groups and compare the models of healthy vs. diseased populations, in terms of complexity, predictability and similarity. This gives us an insight into the information contained in the data, and presumably into brain function, not available through spatial activation-map based analysis methods.

2. State-space Model

The state-space model with K hidden states is parameterized by $\theta = \{\alpha, \pi, \omega, \Sigma_\epsilon\}$ as shown in Fig. 1. Here, $\mathbf{y}_t \in \mathbb{R}^N$, $t = 1 \dots T$ is the observed fMRI data, with the corresponding experimental conditions given by $\mathbf{s}_t$. The underlying mental process is represented as a (hidden) state sequence $X_t \in [1 \dots K]$, for $t = 1 \dots T$. The state marginal distribution is given by $\alpha = (\alpha_1 \dots \alpha_K)$, where $\Pr[X_t = k] = \alpha_k$. The transition probabilities $\Pr[X_{t+1} = k_2 | X_t = k_1] = \pi_{k_1,k_2}$ are given by the $K \times K$ stochastic matrix π. The emission model has a two-level hierarchy to account for the fact that $\mathbf{y}_t$ is the hemodynamic response to the (unobserved) neural activation pattern $\mathbf{z}_t$ corresponding to state X_t. The hemodynamic effect is modeled by the (voxel-wise) convolution $\mathbf{y}_t = \sum_{\tau=0}^{L} \mathbf{h}_\tau \mathbf{z}_{t-\tau} + \epsilon_t$ of the neural activity $\mathbf{z}_t$ with a hemodynamic response function (HRF) $\mathbf{h}$. Here, $\epsilon_t \sim \mathcal{N}(0, \Sigma_\epsilon)$ is a time-stationary noise term. The HRF is a FIR filter of length L given by the difference of two Gamma functions[12], with non-linear parameters γ controlling its delay, dispersion, and ratio of onset-to-undershoot, with prior density $p(\gamma) = \mathcal{N}(\mu_\gamma, \sigma_\gamma)$.

The activity patterns $\mathbf{z}_t$ are transformed into an orthogonal feature-space Ψ, described in Section 3, and when $X_t = k$, is normally distributed in this feature space, i.e. $p(\Psi[\mathbf{z}_t]|X_t = k) \sim \mathcal{N}(\mu_k, \Sigma_k)$. Let $\omega = \{\omega_1 \dots \omega_K\}$, where $\omega_k = \{\mu_k, \Sigma_k\}$, denote the emission parameters of the model. Since the HRF convolution is commutative with respect to Ψ, the convolutive model is applicable to the feature-space transformation $\Psi[\mathbf{y}_t]$, and therefore $\mathbf{z}_t$ and $\mathbf{y}_t$ will stand in for $\Psi[\mathbf{z}_t]$ and $\Psi[\mathbf{y}_t]$, whenever it is clear from the context.

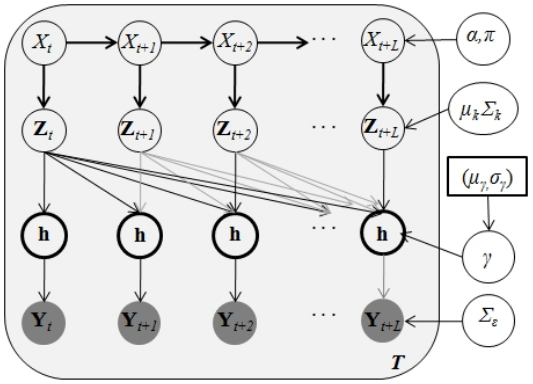

Figure 1: The Markov chain representation of the State-space model.

Also, the following short-hand notation is used through out the paper: $\mathbf{y}_{t_1 \dots t_2} \equiv \{\mathbf{y}_{t_1} \dots \mathbf{y}_{t_2}\}$, $\mathbf{y} \equiv \{\mathbf{y}_1 \dots \mathbf{y}_T\}$, and similarly for $\mathbf{z}$ and X. Also, define $p_\theta(\cdot) = p(\cdot|\theta, \mathbf{h}, K)$.

Expanding out the linear dependence of $\mathbf{y}$ on X through $\mathbf{z}$ and $\mathbf{h}$, the following probability model is obtained:

$$p_\theta(\mathbf{y}_t|X) = \int_{\mathbf{z}_{t-L\dots t}} p_\theta(\mathbf{y}_t|\mathbf{z}_{t-L\dots t}) \prod_{\tau=0}^{L} p_\theta(\mathbf{z}_{t-\tau}|X_{t-\tau}) d\mathbf{z}_{t-L\dots t} = \mathcal{N}\left(\mu_{t-L\dots t}, \Sigma_{t-L\dots t}\right), \tag{1}$$

where $\mu_{t-L\dots t} = \sum_{\tau=0}^{L} \mu_{X_{t-\tau}} \mathbf{h}_\tau$, and $\Sigma_{t-L\dots t} = \Sigma_\epsilon + \sum_{\tau=0}^{L} \Sigma_{X_{t-\tau}} \mathbf{h}_\tau^2$. Thus, the convolution introduces a dependency between states $X_{t-L} \dots X_t$, when conditioned on observation $\mathbf{y}_t$, violating the first-order Markov property required for the classical forward-backward recursions. We present efficient algorithms for estimation of this model in Section 4.

3. Feature Space

The cognitive state at two time points t_1 and t_2 are considered to be similar if their underlying activation patterns are similarly distributed on the functional circuits of the brain. Starting with this axiomatic definition, in this section, we derive a linear embedding which provides a good approximation of similarity. This embedding is a generalization of the linear approximation for the Earth Mover's Distance defined on ℓ_2 metrics[37], to arbitrary distance metrics.

Functional networks are routinely defined by the "temporal correlations between spatially remote neurophysiological events". We have developed an algorithm of computing the functional connectivity that is consistent, sparse and computationally efficient, although any alternative method could also be used (See [22] for a review of this topic). First, the raw fMRI data are spatially smoothed to increase spatial coherence of the connectivity structure, and then spatially proximal voxels are clustered using an agglomerative hierarchical clustering, which reduces dimensionality while utilizing the spatial coherence in the data to increase SNR. Next, regularized covari-

ances between the clusters are estimated using adaptive soft shrinkage. Regularized estimates of voxel-wise correlations are then recomputed from the cluster-wise correlations. If i, j are two cortical voxels, then the functional connectivity map $\mathrm{F}_{i,j} :\to [-1, 1]$ for all $1 \leq i, j \leq N$, obtained by this procedure is consistent and extremely sparse. This method is described in detail in Supplementary Material, Section B.

The difference $\mathrm{TD}\{\mathbf{z}_{t_1}, \mathbf{z}_{t_2}\}$ between two activation patterns $\mathbf{z}_{t_1}$ and $\mathbf{z}_{t_2}$ is quantified by the *transportation distance*[37], i.e. the minimal "transport" of activity $q : x^2 \to \mathbb{R}$ over the functional circuits to convert $\mathbf{z}_{t_1}$ into $\mathbf{z}_{t_2}$, where x is the voxel-grid of the fMRI volume. Specifically, $\mathrm{TD}\{\mathbf{z}_{t_1}, \mathbf{z}_{t_2}\} = \min_q \sum_{i,j \in x} q_{i,j} d_{i,j}$, subject to the constraints: $q_{i,j} \geq 0$, $\sum_i q_{i,j} \leq [\mathbf{z}_{t_1}]_i$, $\sum_j q_{i,j} \leq [\mathbf{z}_{t_2}]_j$, and $\sum_{i,j} q_{i,j} = \min\left\{\sum_i [\mathbf{z}_{t_1}]_i, \sum_j [\mathbf{z}_{t_2}]_j\right\}$. The cost of the transport of $q_{i,j}$ from voxel i to j will depend on the functional distance $d : x^2 \to \mathbb{R}^+$ between the voxels (i.e. measure of functional "disconnectivity") complementary to F, described next. This definition captures the intuitive notion that two activity patterns are functionally more similar if the differences between them are mainly between voxels that are functionally related to each other, indicating the activation of a shared functional network.

The distance metric d arises from the *distortion minimizing embedding*[6] of the graph whose adjacency matrix is given by F, as $\mathbf{f}^* = \arg\inf_{\mathbf{f} \in \mathbb{R}^N} \left\{(\sum_i \sum_j (\mathbf{f}_i - \mathbf{f}_j)^2 \mathrm{F}_{i,j}) / (\sum_i \mathbf{f}_i^2 \mathrm{D}_{i,i})\right\}$, subject to $\mathbf{f} \perp \mathbf{1}$. Here, $\mathbf{f}^*$ will take similar values at voxels that have high functional connectivity and the functional distance between them is $d_{i,j} = |\mathbf{f}_i - \mathbf{f}_j|$. It can be shown that $\mathbf{f}^*$ is the solution to the generalized eigenvalue problem $(\mathrm{D} - \mathrm{F})\mathbf{f} = \lambda \mathrm{D}\mathbf{f}$ subject to $\mathbf{f}'\mathrm{D}\mathbf{1} = 0$, where D is the diagonal degree matrix ($\mathrm{D}_{i,i} = \sum_{j \neq i} \mathrm{F}_{i,j}$ and $\mathrm{D}_{i,j} = 0, \forall i \neq j$). If $\mathbf{u}_1$ is the eigenvector of the normalized graph Laplacian $\mathcal{L} = \mathrm{D}^{-\frac{1}{2}} (\mathrm{D} - \mathrm{F}) \mathrm{D}^{-\frac{1}{2}}$ corresponding to second smallest eigenvalue $\lambda_1 > 0$, then $\mathbf{f}^* = \mathrm{D}^{-\frac{1}{2}} \mathbf{u}_1$.

Through a recursive partitioning of the voxel-grid based on its embedding $\mathbf{f}^*$, we construct an orthogonal basis $\Psi = \{\psi^{(l,m)} \in \mathbb{R}^N\}$ where $l = 0 \ldots 2^m - 1$, $m = 0 \ldots \log_2 N - 1$, as follows. The first basis vector $\psi^{(0,0)} = \mathrm{D}^{-\frac{1}{2}} \mathbf{u}_1$, where $\mathbf{u}_1$ is the eigenvector of $\mathcal{L}^{(0,0)} = \mathcal{L}$ corresponding to the second smallest eigenvalue. The voxel-grid is then partitioned into two disjoint sub-grids based on the sign of $\psi^{(0,0)}$. The residual functional connectivity of the voxels $\widehat{\mathrm{F}} = \mathrm{F} - \lambda_1 \psi^{(0,0)} \psi'^{(0,0)}$ is recomputed. The next two basis vectors $\psi^{(1,1)}$ and $\psi^{(2,1)}$ are the second smallest eigenvectors of the $\mathcal{L}^{(1,1)}$ and $\mathcal{L}^{(2,1)}$, the graph Laplacians of the $\widehat{\mathrm{F}}$ restricted to each sub-grid, respectively. The process may be repeated until only one voxel is left in the partition. However, for numerical stability reasons, the recursive partitioning is terminated when the spectral radius of $\mathcal{L}^{(l,m)}$ drops below a certain tolerance, which also reduces the dimension of the feature-space to $\approx 10^{-3} N$. The algorithm and the derivation of the residual connectivity are explained further in Supplementary Material, Sections C.1, C.2.

The coordinates of $\mathbf{z}_t$ in the feature-space are $\{2^{-m} \tilde{\mathbf{z}}_t^{(l,m)}\}$, $m = 0 \ldots \log_2 N - 1, l = 0 \ldots 2^m - 1$, where $\tilde{\mathbf{z}}_t^{(l,m)} = \langle \mathbf{z}_t, \psi^{(l,m)} \rangle$ and the distance metric $\mathrm{TD}\{\mathbf{z}_{t_1}, \mathbf{z}_{t_2}\}$ is replaced by an approximately equivalent metric $\Delta(\mathbf{z}_{t_1}, \mathbf{z}_{t_2}) = \sum_{l,m} |2^{-m} \left(\tilde{\mathbf{z}}_{t_1}^{(l,m)} - \tilde{\mathbf{z}}_{t_2}^{(l,m)}\right)^2|$.

To examine the equivalence of this linear embedding with respect to $\mathrm{TD}\{\mathbf{z}_{t_1}, \mathbf{z}_{t_2}\}$, consider the dual formulation $\mathrm{TD}\{\mathbf{z}_{t_1}, \mathbf{z}_{t_2}\} = \sup_g \sum_{i=1}^N g_i \cdot S_i$ subject to $g_i - g_j \leq d_{i,j}$ and $\sum_i g_i = 0$. This cost function is an inner product between $g \in \mathbb{R}^N$ and the difference vector $S = \mathbf{z}_{t_1} - \mathbf{z}_{t_2}$,

and since inner products are preserved under orthogonal transformations, we have $\langle g, S\rangle = \langle \Psi[g], \Psi[S]\rangle$. Letting the coefficients of S in the basis Ψ be $\tilde{S}_{l,m} = \langle \psi^{(l,m)}, S\rangle$, the following theorem holds (c.f. Supplementary Material, Section C.3):

Theorem 1. *Let $\tilde{S}_{l,m}$ be coefficients of $S = \mathbf{z}_{t_1} - \mathbf{z}_{t_2}$. Then, there exist constants $M_{0,0} > 0$ and $\widehat{M}_{0,0} > 0$, such that*

$$\widehat{M}_{0,0} \sum_{m=0}^{\log_2 N-1} \sum_{l=0}^{2^m-1} 2^{-m}|\tilde{S}_{l,m}| \;\leq\; \mathrm{TD}\{\mathbf{z}_{t_1}, \mathbf{z}_{t_2}\} \;\leq\; M_{0,0} \sum_{m=0}^{\log_2 N-1} \sum_{l=0}^{2^m-1} 2^{-m}|\tilde{S}_{l,m}| \quad (2)$$

and the tightness of this bound is:

$$\sup_{||\mathbf{z}||_2=1} \left[\sum_{m,l} M_{l,m}|\tilde{\mathbf{z}}_{l,m}| - \sum_{m,l} \widehat{M}_{l,m}|\tilde{\mathbf{z}}_{l,m}|\right] \approx \frac{(M_{0,0} - \widehat{M}_{0,0})}{\sqrt{2}}$$

Therefore, based on the form of these upper and lower-bounds, $\mathrm{TD}\{\mathbf{z}_{t_1}, \mathbf{z}_{t_2}\}$ is approximated by $\Delta(\mathbf{z}_{t_1}, \mathbf{z}_{t_2})$, up to a multiplicative constant. The decay of the coefficients as 2^{-m} implies that the effect of reduced dimensionality of Ψ on the approximation error is small. Please refer to Supplementary Material, Section C.4 for further discussion on this approximation.

4. Model Estimation and Selection

The maximum likelihood (ML) estimate $\theta_{\mathrm{ML}} = \arg\max_\theta \ln p(\mathbf{y}|\theta, \mathbf{h}, K)$ is obtained using the Expectation Maximization (EM) algorithm[4] which involves iterating the following two steps until convergence:

$$\textbf{E-step: } \mathcal{Q}(\theta, \theta^n) = \sum_X p(X|\mathbf{y}, \theta^n, \mathbf{h}, K) \ln p(\mathbf{y}, X, \theta|\mathbf{h}, K), \quad \textbf{M-step: } \theta^{n+1} = \arg\max_\theta \mathcal{Q}(\theta, \theta^n). \quad (3)$$

Because of the inclusion of the FIR filter for the HRF, which violates the first-order Markov property of the state-sequence X when conditioned on an observation $\mathbf{y}_t$, the EM update equations take the following form (c.f. Supplementary Material, Section D.1):

$$\alpha_k^{n+1} = \frac{p_{\theta^{(n)}}(X_1 = k|\mathbf{y})}{\sum_{k'=1}^K p_{\theta^{(n)}}(X_1 = k'|\mathbf{y})} \quad , \quad \pi_{k_1,k_2}^{n+1} = \frac{\sum_{t=2}^T p_{\theta^{(n)}}(X_t = k_1, X_{t+1} = k_2|\mathbf{y})}{\sum_{k'=1}^K \sum_{t=2}^T p_{\theta^{(n)}}(X_t = k_1, X_{t+1} = k'|\mathbf{y})}$$
$$\mu_k^{n+1} = \sum_{k_0 \ldots k_L} \mathbf{H}^-_{k,k_0\ldots k_L} \mu^{n+1}_{k_0\ldots k_L} \quad \text{and} \quad \Sigma_k^{n+1} = \sum_{k_0\ldots k_L} \mathbf{G}^-_{k,k_0\ldots k_L} \Sigma^{n+1}_{k_0\ldots k_L}, \quad (4)$$

where

$$\mathbf{H} = \begin{pmatrix} \mathbf{h}_L + \ldots \mathbf{h}_0 & 0 & \ldots & 0 \\ \mathbf{h}_L + \ldots \mathbf{h}_1 & \mathbf{h}_0 & \ldots & 0 \\ \vdots & \vdots & \ldots & \vdots \\ 0 & 0 & \ldots & \mathbf{h}_{L-1} + \mathbf{h}_0 \\ 0 & 0 & \ldots & \mathbf{h}_L + \mathbf{h}_0 \end{pmatrix} \quad \mathbf{G} = \begin{pmatrix} \mathbf{h}_L^2 + \ldots \mathbf{h}_0^2 & 0 & \ldots & 0 \\ \mathbf{h}_L^2 + \ldots \mathbf{h}_1^2 & \mathbf{h}_0^2 & \ldots & 0 \\ \vdots & \vdots & \ldots & \vdots \\ 0 & 0 & \ldots & \mathbf{h}_{L-1}^2 + \mathbf{h}_0^2 \\ 0 & 0 & \ldots & \mathbf{h}_L^2 + \mathbf{h}_0^2 \end{pmatrix}$$

$$\mu_{k_0 \ldots k_L}^{n+1} = \frac{\sum_{t=1}^{T} p_{\theta^{(n)}}(X_{t-L \ldots t} = k_0 \ldots k_L | \mathbf{y}) \mathbf{y}_t}{\sum_{t=1}^{T} p_{\theta^{(n)}}(X_{t-L \ldots t} = k_0 \ldots k_L | \mathbf{y})},$$
$$\Sigma_{k_0 \ldots k_L}^{n+1} = \frac{\sum_{t=1}^{T} p_{\theta^{(n)}}(X_{t-L \ldots t} = k_0 \ldots k_L | \mathbf{y}) \cdot (\mathbf{y}_t - \mu_{k_0 \ldots k_L}^{n+1})(\mathbf{y}_t - \mu_{k_0 \ldots k_L}^{n+1})'}{\sum_{t=1}^{T} p_{\theta^{(n)}}(X_{t-L \ldots t} = k_0 \ldots k_L | \mathbf{y})}.$$

and $\mathbf{H}^{-}_{k, k_0 \ldots k_L}$ is the $(k,\ k_0 \ldots k_L)$ element of the pseudo-inverse of $\mathbf{H}$, given by $\mathbf{H}^{-} = (\mathbf{H}'\mathbf{H})^{-}\mathbf{H}'$. Even though $\mathbf{H}$ is an $(L+1)^K \times K$ matrix, it is extremely sparse with each column k of $\mathbf{H}$ having only 2^{L+1} non-zero entries corresponding to those $\mu_{k_0 \ldots k_L}^{n+1}$ where any $k_0 \ldots k_L = k$. Therefore, $\mathbf{H}'\mathbf{H}$ is computed in $\mathcal{O}(2^{L+1}K^2)$ time, and is inverted using the SVD pseudo-inverse. Similarly for $\mathbf{G}$.

Using the relationship $p_{\theta^{(n)}}(X|\mathbf{y}) = p_{\theta^{(n)}}(\mathbf{y}, X)/p_{\theta^{(n)}}(\mathbf{y})$ and the fact that $p_{\theta^{(n)}}(\mathbf{y})$ is canceled out by the numerators and denominators of eqn. 4, the conditional densities are replaced by their joint densities $p_{\theta^{(n)}}(\mathbf{y}, X_t), p_{\theta^{(n)}}(\mathbf{y}, X_{t,t+1})$ and $p_{\theta^{(n)}}(\mathbf{y}, X_{t-L \ldots t})$. These are calculated as (c.f. Supplementary Material, Section D.2):

$$\begin{aligned} p_\theta(\mathbf{y}, X_t) &= \sum_{X_{t+1-L \ldots t-1}} a(X_{t+1-L \ldots t}) b(X_{t+1-L \ldots t}) \\ p_\theta(\mathbf{y}, X_{t,t+1}) &= \sum_{X_{t+1-L \ldots t-1}} a(X_{t+1-L \ldots t}) \cdot p_\theta(\mathbf{y}_{t+1} | X_{t+1-L \ldots t+1}) p_\theta(X_{t+1} | X_t) \cdot b(X_{t+2-L \ldots t}) \\ p_\theta(\mathbf{y}, X_{t,t+L}) &= a(X_{t,t-1+L}) \cdot p_\theta(\mathbf{y}_{t+L}, X_{t,t+L}) \cdot p_\theta(X_{t,t+L}) \cdot b(X_{t+1 \ldots t+L}). \end{aligned} \tag{5}$$

where a and b are the forward-backward recursion terms:

$$\begin{aligned} a(X_{t+1-L \ldots t}) = p_\theta(\mathbf{y}_{1 \ldots t}, X_{t+1-L \ldots t}) &= \sum_{X_{t-L}} p_{\theta^{(n)}}(\mathbf{y}_t | X_{t-L \ldots t}) p_{\theta^{(n)}}(X_t | X_{t-1}) \cdot a(X_{t-L \ldots t-1}) \\ b(X_{t+1-L \ldots t}) = p_\theta(\mathbf{y}_{t+1 \ldots T} | X_{t+1-L \ldots t}) &= \sum_{X_{t+1}} p_{\theta^{(n)}}(\mathbf{y}_{t+1} | X_{t+1-L \ldots t+1}) b(X_{t+2-L \ldots t+1}). \end{aligned} \tag{6}$$

The summations (i.e. expectations) over the densities of state-sequences L long of the form $\sum_{X_{t-L \ldots t}} p_{\theta^{(n)}}(\mathbf{y}, X_{t-L \ldots t})[\ldots]$ in eqns. 4 and 5 are replaced with Monte Carlo estimates, by Gibbs sampling from the distribution $p_{\theta^{(n)}}(\mathbf{y}, X_{t-L \ldots t})$ with stochastic forward-backward recursions[34].

The same EM procedure can estimate θ_{ML} given multiple fMRI data-sets corresponding to a group of subjects, with slight modifications to the update equations. The dependence of θ_{ML} on a

specific HRF filter $\mathbf{h}$ is removed by marginalizing out $\mathbf{h}$ under a Laplace approximation to obtain a Bayesian estimate $\theta^* = \int_{\mathbf{h}} \theta_{\mathrm{ML}}(\mathbf{h}) p(\mathbf{h}) d\mathbf{h}$, independent of $\mathbf{h}$. It is computed through Monte Carlo integration by first sampling the parameter γ from $\mathcal{N}(\mu_\gamma, \sigma_\gamma)$, constructing $\mathbf{h}(\gamma)$, finding $\theta_{\mathrm{ML}}(\mathbf{h})$ and then averaging over all samples (c.f. Supplementary Material, Section D.3).

Given a set of parameters θ and observations $\mathbf{y}$, the most probable sequence of states $X^* = \arg\max \ln p_\theta(\mathbf{y}, X)$ is estimated by *backtracking* (c.f. Supplementary Material, Section D.4) through the following recursive system: $\max_X \ln p_\theta(\mathbf{y}, X) = \max_{X_{t-L\ldots T}} \eta_T$, where $\eta_t = \max_{X_{t-1}} [\ln p_\theta(\mathbf{y}_t, X_{t-L\ldots t}) + \ln p_\theta(X_t | X_{t-1}) + \eta_{t-1}]$, and $\eta_1 = \ln p_\theta(\mathbf{y}_1 | X_1) + \ln p_\theta(X_1)$. The initial maximization over $X_{t-L\ldots T}$ is done using Iterated Conditional Modes(ICM)[4], with random restarts.

While it is possible to use information theoretic[20] or MCMC based Bayesian [34] alternatives for model size selection, each criterion will select a different best model, not necessarily related to the experimental task. Instead, we adopt a model-size selection strategy where the experimental conditions $\mathbf{s}_t$ are used to select K that results in a *maximally predictive* model, since we are interested in understanding task related effects, although this step introduces a dependence of the experimental conditions on the model. Let $X^{*,K}$ denote the optimal state-sequence for an fMRI session $\mathbf{y}$ produced by the model with K states and optimal parameters θ^*. And, let $\mathbf{s}_t$ denote the corresponding experimental conditions recorded during the session. The optimal K^* is then selected as $K^* = \arg\min \mathrm{R}\left\{X^{*,K}, f_X(\mathbf{s})\right\}$ where R is the error-rate (i.e. risk) between the optimal state-sequence $X^{*,K}$ and the state sequence predicted by the experimental conditions $f_X(\mathbf{s})$. This prediction is done using a multinomial logisitic regression (MLR) classifier [4], where $f_X(\hat{\mathbf{s}})$ is a vector of probabilities $(\Pr[\hat{X} = 1], \ldots \Pr[\hat{X} = K])$ that the state $\hat{X}$ corresponding to experimental conditions $\hat{\mathbf{s}}$ takes value k for $k = 1 \ldots K$. The error-rate is the average probability of an incorrect prediction over the session $\mathrm{R}\left\{X^{*,K}, f_X(\mathbf{s})\right\} = 1/T \sum_{t=1}^{T} \left[1 - \Pr[\hat{X} = X_t^{*,K}]\right]$ and is computed using cross-validation.

Therefore, the model trained on a data-set $\mathbf{y}$ consists of the tuple (θ^*, K^*, f_X), viz. the optimal model parameters, the optimal number of states, and the prediction function.

5. Results

5.1. Data-set

The method was applied on a study for developmental dyscalculia (DC) [28] consisting of 36 control and 13 DC subjects, who underwent fMRI while judging the incorrectness of multiplication results. In each trial of the *selfpaced, irregular paradigm*, two single-digit numbers (e.g. 4×5) were displayed visually for 2.5s. After an interval of 0.3s an incorrect solution (e.g. 27,23,12) was displayed for 0.8s. Subjects had *up to* 4s to decide, with a button press, if the answer was (a) *close* (within $\pm 25\%$ of the correct answer), (b) *too small* ($< 25\%$) or (c) *too big*. The next trial started after a rest of 1s, and each trial lasted 4–8.6s.

The data were acquired with a GE 3T MRI scanner with a quadrature head coil, using a BOLD sensitized 3D PRESTO pulse sequence with a volume scan time of 2.64s and resolution of $3.75 \times 3.75 \times 3.75\mathrm{mm}^3$. All subjects were scanned in two sessions, with an interval of approximately 30mins between sessions. In each session $\approx$120 multiplication problems were presented and 276

scans acquired. The data were pre-preprocessed to remove imaging artifacts, motion corrected, de-noised, and normalized to an atlas space. All further processing was done in the grey matter. Note that no spatial smoothing was applied. The algorithms were implemented in MATLAB® with Star-P® on an 2.6Hz Opteron cluster with 16 processors and 32GB RAM. Please refer to Supplementary Material, Section E.1 for further details.

5.2. Analysis and Discussion

One aim of the study was to understand the effect of the product size and problem difficulty on the brain response of the two groups, as a trial proceeded and identify if and how the two groups differed.

For each $t = 1 \dots T$, the experimental conditions are described by the vector $\mathbf{s}_t = (\mathrm{Ph}, \mathrm{Len}, \mathrm{LogPs}, \mathrm{LogDiff}, \mathrm{Ans})$, where Ph is the phase within the current trial in 1.2s increments with respect to its start, Len is the length (0–8.6s) of the trial, LogPs quantifies the product size for the presented problem, $\mathrm{LogDiff}$ quantifies the expected difficulty in judging the right answer, and Ans is binary variable indicating if the subject's response was correct.

The following set of models (θ^*, K^*, f_X) were trained: (a) CTRL_SELF: one model per control subject, (b) DYSC_SELF: one model per DC subject, (c) CTRL_GRP: one model per group of 12 out of 36 controls, (d) DYSC_GRP: one model per group of 12 out of 13 DCs. To characterize and compare the models, the following statistics were used: (I) ENT: *Entropy* of the MLR probabilities $f_X(\hat{\mathbf{s}})$. It quantifies how well the brain state can be predicted for a specific experimental condition $\hat{\mathbf{s}}$ and varies between 0 and $\log_2 K^*$, with higher value indicating worse predictability. (II) ERR: Empirical *Error Rate* of $f_X(\hat{\mathbf{s}})$ for the state sequence X generated from previously unseen data $\mathbf{y}$ by a model. It quantifies how well the data from one subject conforms to another model. (III) MI: *Mutual Information* between the state sequences generated for one fMRI session $\mathbf{y}$ by two different models. It quantifies the similarity of the two models, where higher MI indicates better similarity, with a maximum of $\log_2 K^*$. These statistics are defined in detail in Supplementary Material, Section E.2.

The Ph-wise effect of high vs. low LogPs and high vs. low $\mathrm{LogDiff}$ on ENT, ERR and MI for the group-level analysis are shown in Fig. 2. The average (± 1 std.dev.) optimal K^* for the CTRL_SELF models is 22.8 ± 3.21, for DYSC_SELF is 23.5 ± 5.22, while for CTRL_GRP 23.2 ± 3.63 is and DYSC_GRP is 22.1 ± 5.78. It is interesting to note that the variation in model-sizes for the DCs is larger even though their group size is much smaller than the controls.

From these results we observe a greater heterogeneity within the fMRI data of the DC group as compared to the control group, almost on par with the differences between DC vs. control. Also, within individuals for both populations, increase in product size (LogPs) increases the predictability and accuracy of the model, while reducing the correspondence between models. This effect early on in the trial is attributable to a stronger effect of product recall from the rote tables located in the lower left parietal lobe, and strong number size effects in the occipital visual areas. The later effect of LogPs is consistent with strong activation of the working verbal (mute-rehearsal) and visual memories[32]. The effect of problem difficulty ($\mathrm{LogDiff}$) on individuals is similar, and is noticed only after the incorrect result is displayed ($\mathrm{Ph} > 2.8$s), as expected. Also, at the group level for controls, both effects are preserved, indicating spatio-temporal patterns that are shared across their data, while for DCs the effect of LogPs reverses, which indicates

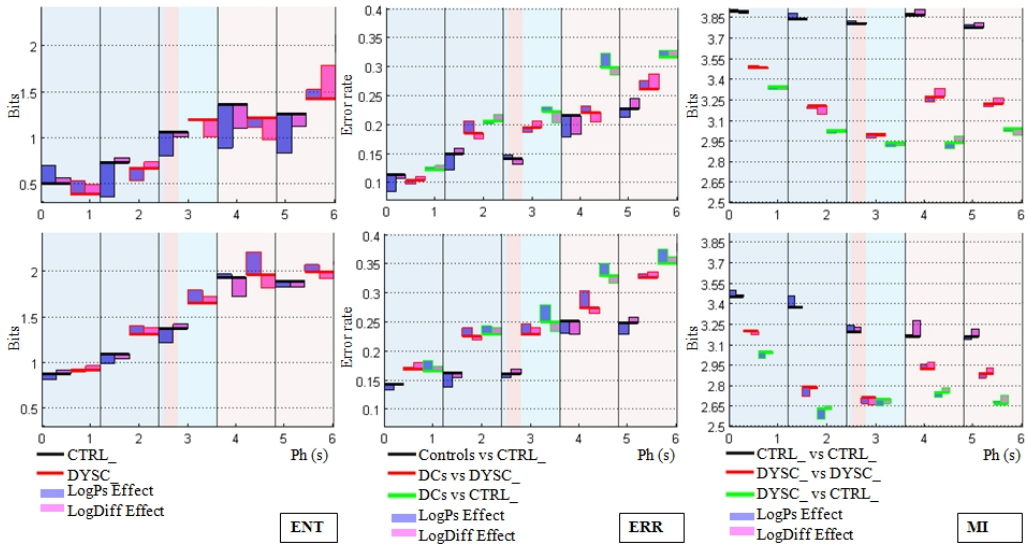

Figure 2: The effect on ENT, ERR, and MI with respect to experiment phase Ph, LogPs and LogDiff. The top row corresponds to the models CTRL_SELF and DYSC_SELF, while the bottom row to CTRL_GRP and DYSC_GRP. The ENT panel shows the effect of Ph (in 1.2s increments indicated by the vertical grid), high minus low LogPs and high minus low LogDiff on the models (θ^*, K^*, f_X) trained on the controls and dyscalculics, either individually or as groups. The ERR panels show the effects on the error rate of predicting a control's data with a CTRL_SELF or CTRL_GRP (black), a DC subject's data with a DYSC_SELF or DYSC_GRP (red), and a DC subject's data with a CTRL_SELF or CTRL_GRP (green). The MI panel shows the effects on the MI between two CTRL_SELF models (black), two DYSC_SELF models (black) and between an CTRL_SELF and DYSC_SELF model (green). Similarly for CTRL_GRP and DYSC_GRP. The background color-coding shows the 2.5s, 0.3s, 0.8s and 0–4s divisions of each trial.

that while at the individual level this parameter creates stronger, more identifiable patterns, at the group level the patterns share fewer similarities. The preservation at the group-level of the LogDiff effect in DCs supports the hypothesis that the difficulty effect is probably attention and conflict resolution related, rather than number related, and therefore is not affected by dyscalculic deficiencies. The later trends in the MI panel, while confirming the above conclusions, also provides evidence towards the theory that the DCs might be replaying the multiplication problem in their minds after they have finished the task [28]. Please refer to Supplementary Material, Section E.3 for a more detailed interpretation.

Every state $k = 1 \dots K$ is associated with emission parameters $\omega_k = \{\mu_k, \Sigma_k\}$, where μ_k is the mean vector given in the coordinates of the (reduced dimensional) feature space Ψ , and therefore the mean spatial map for state k can be reconstructed (only approximately, due to dimensionality reduction). However, exact interpretation of this map in terms of spatial distribution of neural activity is difficult for two reasons. One, since multiple states may occur during a specific experimental condition $\mathbf{s}_t$, with probabilities given by $f_X(\mathbf{s}_t)$, there is no single activation map for a given condition (and vice-versa). But more importantly, the probability that the observed pattern belongs to state k depends not only on μ_k but also on Σ_k, and therefore, interpretation requires taking Σ_k, which is a full covariance matrix, into account. Although, we plan to address these questions in future work, spatial maps corresponding to 0s$\leq$ Ph $\leq$2.4s and 2.4s$<$ Ph $\leq$4.8s constructed with a heuristic method are discussed in Supplementary Material, Section E.4. The

patterns in these maps exhibit strong similarity with the foci known to be activated during this task [28]. This indicates that the model is indeed learning the characteristic distribution of activity (i.e. a template) for a particular mental task.

6. Conclusion

In this paper we presented an unsupervised approach towards decoding and representing the information about mental processes in fMRI data, using a hidden Markov model. The hemodynamic coupling between neural activity and the BOLD response was accounted for by an additional hidden layer. An intuitive definition of the similarity of two activation patterns with respect to the functional connectivity map of the brain was proposed, and a linear feature-space was derived. Efficient estimation algorithms, based on forward-backward recursions and Monte-Carlo sampling, were shown. The effect of the variability in hemodynamics was eliminated through marginalization under a Laplace approximation. Model selection was then performed using a maximally predictive criteria. We applied the method to a group-wise study for developmental dyscalculia, and demonstrated task-related differences between healthy controls and dyscalculics, which were systematically organized in time.

This abstract representation of mental processes built up from fMRI data, by capturing and summarizing spatio-temporal patterns, can help reveal differences between the populations and confirm hypotheses not easily deducible from spatial maps of activity. Moreover, it provides a summary of complex tasks and of the information content in the data that could serve as a starting-point for investigations with regular methods.

We are currently working on addressing the interpretation of the spatial maps acquired from this method. We are also exploring the application of this method to default-state and non task-related fMRI studies. In future work, we also plan to develop a spatially-varying and adaptive hemodynamic response model, along with incorporating Bayesian priors on the model.

Appendix A. Notation

The following table lists the notation used in the main text and Supplementary Material.

Symbol	Definition
T	Total number of time-points acquired in an fMRI session
N	Total number of (cortical) voxels in an fMRI volume
x	The voxel-grid of the volumetric fMRI data with N voxels
$x_i \in \mathbb{R}^3$	Spatial location (in mm) for voxel $i \in x$
$\mathbf{s}_t$	Vector giving the prevailing experimental conditions at time t
$X_t \in [1 \dots K]$	The (hidden) brain state at $1 \leq t \leq T$
X	Defined as $(X_1 \dots X_T)$
$X_{t_1 \dots t_2}$	Defined as $(X_{t_1} \dots X_{t_2})$
α	The multinomial probability vector for the states
π	The transition probability matrix
$\mathbf{z}_t \in \mathbb{R}^N$	Underlying activity pattern at $1 \leq t \leq T$
$\mathbf{z}$	Defined as $(\mathbf{z}_1 \dots \mathbf{z}_T)$
$\mathbf{z}_{t_1 \dots t_2}$	Defined as $(\mathbf{z}_{t_1} \dots \mathbf{z}_{t_2})$
ω_k	The emission parameters (μ_k, Σ_k) for state k
$\mathbf{y}_t \in \mathbb{R}^N$	fMRI scan at $1 \leq t \leq T$
$\mathbf{y}$	Defined as $(\mathbf{y}_1 \dots \mathbf{y}_T)$
$\mathbf{y}_{t_1 \dots t_2}$	Defined as $(\mathbf{y}_{t_1} \dots \mathbf{y}_{t_2})$
$\epsilon \sim \mathcal{N}(0, \Sigma_\epsilon)$	Time-stationary noise in Y_t
$\mathbf{h}$	The linear convolutive hemodynamic response filter of length L
$\gamma \sim \mathcal{N}(\mu_\gamma, \sigma_\gamma)$	The parameters of HRF $\mathbf{h}$
θ	Defined as (α, π, ω)
$p_\theta(\cdot)$	Defined as $p(\cdot\|\theta, \mathbf{h}, K)$
$\mathbb{C} : \mathbb{R}^{N \times N}$	Matrix of covariances between the voxels
$\mathrm{F} : [-1, 1]^{N \times N}$	Functional connectivity (i.e. correlation) map
$d : \mathbb{R}^{+N \times N}$	The distance metric induced by F
$\mathrm{D} : \mathbb{R}^{N \times N}$	The diagonal degree matrix of F
$\mathcal{L}$	The normalized graph Laplacian of F
$\Psi = \{\psi^{(l,m)} \in \mathbb{R}^N\}$	Orthogonal basis functions of the feature space

Appendix B. Functional Connectivity

Let $\mathbf{y}_t$ be the fMRI scan at time t with N voxels, where $[\mathbf{y}_t]_i$ is the time-series data of voxel $i = 1 \dots N$.

The functional connectivity map $\mathrm{F} : [-1, 1]^{N \times N}$ measures the temporal correlation between two voxels. Because the dimension N is much larger than the number of observations T, the standard covariance estimator is badly conditioned, and its eigen-system is inconsistent[3]. Therefore, regularization is required to impose sensible structure on the estimated covariance matrix while being computationally efficient.

To obtain robust estimates of functional connectivity, we use the following procedure: First, the images are smoothed with a Gaussian kernel (FWHM=8mm) to increase spatial coherence of the time-series data.

Next, a hierarchical clustering algorithm is used to cluster closeby voxels with similar time courses (c.f. Algorithm 1) which produces a set of $\widetilde{N}$ spatially contiguous clusters. This procedure has a two-fold benefit of reducing the dimensionality of the estimation problem while simultaneously increasing the SNR of the data through averaging. Typically, we perform agglomeration until $\widetilde{N}$ is $0.25 \times N$. The clusters, after Gaussian smoothing, are $1.5\times$ larger for a given in-cluster variance than without smoothing.

For voxel i, create one cluster c_i of size $n_i = 1$. Each c_i is associated with a time-course $[\mathbf{y}_t]_i$.
repeat
  Find two clusters c_i and c_j that are spatially adjacent to each other and merge them into a new cluster $c_k = (c_i, c_j)$, if and only if $\mathrm{Var}\{c_k\}$ is minimum over all $i,\ j$.
  Remove clusters c_i and c_j from the set of clusters, and add c_k
until Number of clusters reaches the specified value.

Algorithm 1: Hierarchical Clustering

The time-series for the new cluster c_k is defined as $[\mathbf{y}_t]_k = 1/n_k \sum_{c_i \in c_k} [\mathbf{y}_t]_i$, and for a new cluster $c_k = (c_i, c_j)$ can be efficiently updated according to $[\mathbf{y}_t]_k = (n_i[\mathbf{y}_t]_i + n_j[\mathbf{y}_t]_j)/(n_i + n_j)$. The variance of a cluster c_k is $\mathrm{Var}\{c_k\} = (1/n_k T) \sum_{c_i \in c_k} \sum_{t=1}^{T} ([\mathbf{y}_t]_i - [\mathbf{y}_t]_k)^2$, and can be efficiently updated through the variance separation theorem:

$$\mathrm{Var}\{c_k\} = \frac{n_i \mathrm{Var}\{c_i\} + n_j \mathrm{Var}\{c_j\}}{n_i + n_j} - \frac{\sum_{t=1}^{T} ([\mathbf{y}_t]_i - [\mathbf{y}_t]_k)^2}{T(n_i + n_j)}.$$

For a voxel i belonging to cluster c_k, the expected (smoothed) time series, given the cluster average is:

$$\mathrm{E}\{[\mathbf{y}_t]_i | [\mathbf{y}_t]_k\} = \overline{\mathbf{y}}_i + \sigma_{i,k}\sigma_{k,k}^{-1}([\mathbf{y}_t]_k - \overline{\mathbf{y}}_k), \quad \text{and} \quad \mathrm{Var}\{[\mathbf{y}]_i | [\mathbf{y}]_k\} = \sigma_{i,i} - \sigma_{i,k}^2 \sigma_{k,k}^{-1}, \tag{B.1}$$

where $\overline{\mathbf{y}}_i = 1/T \sum_t [\mathbf{y}_t]_i$ and $\sigma_{i,k} = 1/T \sum_t [\mathbf{y}_t]_i [\mathbf{y}_t]_k - \overline{\mathbf{y}}_i \overline{\mathbf{y}}_k$. The correlation coefficient

between the voxel time-series and the cluster average is:

$$\rho_{i,k} = \frac{\sigma_{i,k}}{\sqrt{\sigma_{i,i}\sigma_{k,k}}} \tag{B.2}$$

After hierarchical clustering, the covariance $\widetilde{\mathbb{C}}_{i,j}$ between two clusters c_i and c_j is $\widetilde{\mathbb{C}}_{i,j} = \frac{1}{T}\sum_{t=1}^{T}[\mathbf{y}_t]_i[\mathbf{y}_t]_j - \overline{\mathbf{y}}_i\overline{\mathbf{y}}_j$.The regularized estimate of the covariance is computed using an adaptive soft shrinkage estimator[31] $s_\lambda(\widetilde{\mathbb{C}}_{i,j}) = \mathrm{sgn}(\widetilde{\mathbb{C}}_{i,j})(|\widetilde{\mathbb{C}}_{i,j}| - \lambda|\widetilde{\mathbb{C}}_{i,j}|^{-1})_+$.This estimator has the property that the shrinkage is continuous with respect to $\widetilde{\mathbb{C}}_{i,j}$, but the amount of shrinkage decreases as $\widetilde{\mathbb{C}}_{i,j}$ increases resulting in less bias than the standard soft shrinkage estimator. The threshold parameter λ is selected by minimizing the risk function $R(\lambda) = \mathrm{E}||s_\lambda(\widetilde{\mathbb{C}}) - \widetilde{\mathbb{C}}||_2$. Under certain regularity assumptions about the data, a closed form estimate of the optimal threshold is obtained as[21]:

$$\lambda \approx \frac{\sum_{i\neq j}\mathrm{Var}\{\widetilde{\mathbb{C}}_{i,j}\}}{\sum_{i\neq j}\widetilde{\mathbb{C}}^2_{i,j}}, \quad \text{where} \quad \widehat{\mathrm{Var}}\{\widetilde{\mathbb{C}}_{i,j}\} = \frac{T}{(T-1)^3}\sum_{t=1}^{T}\left([\mathbf{y}_t]_i[\mathbf{y}_t]_j - \sum_{t=1}^{T}[\mathbf{y}_t]_i[\mathbf{y}_t]_j\right)^2. \tag{B.3}$$

This estimator is "sparsistent"[31], that is, in addition to being consistent, it estimates true zeros as zeros and non-zero elements as non-zero with the correct sign, with probability tending to 1.

Regularized estimates of the correlation between two voxels i and j belonging to clusters c_k and $c_{k'}$ respectively are obtained by substituting eqns. B.1 and B.2 to get:

$$\mathrm{F}_{i,j} = \frac{\mathrm{Cov}\left\{\mathrm{E}\{[\mathbf{y}_t]_i|[\mathbf{y}_t]_k\}\mathrm{E}\{[\mathbf{y}_t]_j|[\mathbf{y}_t]_{k'}\}\right\}}{\sqrt{\mathrm{Var}\{[\mathbf{y}]_i|[\mathbf{y}]_k\}\mathrm{Var}\{[\mathbf{y}]_j|[\mathbf{y}]_{k'}\}}} = s_\lambda(\widetilde{\mathbb{C}}_{i,j}).\frac{\rho_{i,k}\rho_{j,k'}}{\sqrt{\left(1-\sigma_{k,k}\rho^2_{i,k}\right)\left(1-\sigma_{k',k'}\rho^2_{j,k'}\right)}}. \tag{B.4}$$

The results of this procedure on the distribution of the functional connectivity estimates are shown in Fig. B.3. Without any regularization, most of the mass of the distribution is in concentrated in small non-zero correlations, while the strong correlations are only a fraction of the total. The smoothing procedure shifts the whole distribution towards the right, by strengthening all correlations, while the hierarchical clustering procedure boosts strong correlations without affecting weak correlations. Finally, the shrinkage step sparsifies the correlation matrix, with most correlations set to zero. It is also easy to see that F is symmetric and positive definite.

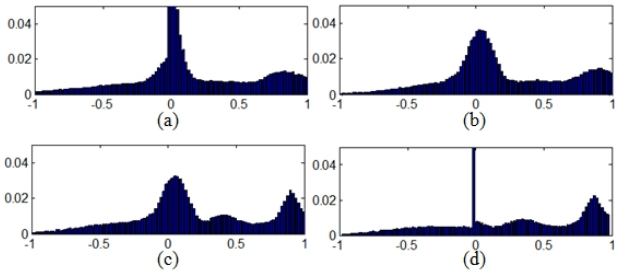

Figure B.3: The normalized histogram of the cross-correlation coefficients (voxel-wise) for the original data (Fig. a), after smoothing (Fig. b), after hierarchical clustering (Fig. c), after shrinkage (Fig. d).

Appendix C. The Linear Feature Space

This section describes the construction of the feature-space (c.f. Section Appendix C.1), along with proofs for the orthogonalization (c.f. Section Appendix C.2) and for the equivalence with the functional similarity metric (c.f. Section Appendix C.3), and concludes with a discussion on the quality of the approximation (c.f. Section Appendix C.4).

Let x denote the voxel-grid of the volumetric fMRI data, and therefore $|x| = N$. If we define the diagonal degree matrix $\mathrm{D} \in \mathbb{R}^{N \times N}$ as $\mathrm{D}_{i,i} = \sum_j \mathrm{F}_{i,j},\ \forall i \neq j$, then the *normalized graph Laplacian* is $\mathcal{L} = \mathrm{D}^{-\frac{1}{2}} (\mathrm{D} - \mathrm{F}) \mathrm{D}^{-\frac{1}{2}}$. Let $\mathbf{u}_1$ be the eigenvector of $\mathcal{L}$ corresponding to its second smallest eigenvalue λ_1.

Appendix C.1. Construction of Orthogonal Basis Functions

Define $x^{(0,0)} \leftarrow x$, $\mathrm{F}^{(0,0)} \leftarrow \mathrm{F}$, $\mathrm{D}^{(0,0)} \leftarrow \mathrm{D}$, and $\mathcal{L}^{(0,0)} \leftarrow \mathcal{L}$
$\psi^{(0,0)}(i) \leftarrow \mathrm{D}^{-\frac{1}{2}} \mathbf{u}_1$ and $\lambda^{(0,0)} \leftarrow \lambda_1$
Set $m \leftarrow 0$
repeat
  for $l \leftarrow 0$ to $2^m - 1$ **do**
    { Here we partition the grid $x^{(l,m)}$ into $x^{(2l,m+1)}$ and $x^{(2l+1,m+1)}$ based on the sign of $\psi^{(l,m)}$ }
    $\widehat{\mathrm{F}} \leftarrow \mathrm{F}^{(l,m)} - \lambda^{(l,m)} \psi^{(l,m)} \psi'^{(l,m)}$
    for $i \leftarrow 1$ to N **do**
      for $j \leftarrow 1$ to N **do**
        if $\psi^{(l,m)}(i) > 0$ AND $\psi^{(l,m)}(j) > 0$ **then**
          $\mathrm{F}^{(2l,m+1)}_{i,j} \leftarrow \widehat{\mathrm{F}}_{i,j}$ and add i, j to $x^{(2l,m+1)}$
        else if $\psi^{(l,m)}(i) \leq 0$ AND $\psi^{(l,m)}(j) \leq 0$ **then**
          $\mathrm{F}^{(2l+1,m+1)}_{i,j} \leftarrow \widehat{\mathrm{F}}_{i,j}$ and add i, j to $x^{(2l+1,m+1)}$
        else
          $\mathrm{F}^{(2l,m+1)}_{i,j} \leftarrow 0$ and $\mathrm{F}^{(2l+1,m+1)}_{i,j} \leftarrow 0$
        end if
      end for
    end for
    Compute the diagonal degree matrices $\mathrm{D}^{(2l,m+1)},\ \mathrm{D}^{(2l+1,m+1)}$ and the normalized graph Laplacians $\mathcal{L}^{(2l,m+1)},\ \mathcal{L}^{(2l+1,m+1)}$ from $\mathrm{F}^{(2l,m+1)}, \mathrm{F}^{(2l+1,m+1)}$ respectively
    Calculate $\psi^{(2l,m+1)},\ \psi^{(2l+1,m+1)}$ from the eigenvectors of $\mathcal{L}^{(2l+1,m+1)},\ \mathcal{L}^{(2l,m+1)}$ corresponding to their second smallest eigenvalues $\lambda^{(2l,m+1)}$ and $\lambda^{(2l+1,m+1)}$, respectively
  end for
  $m \leftarrow m + 1$
until $|x^{(l,m)}| = 1,\ \forall l$

Algorithm 2: Construction of Orthogonal Basis Functions

Appendix C.2. Orthogonalization of F

Since F is a symmetric positive definite kernel, we can consider the functional connectivity $\mathrm{F}_{i,j} = \langle \mathbf{Y}_i, \mathbf{Y}_j \rangle$ in some representation $\mathbf{Y}_i$ and $\mathbf{Y}_j$ at the voxels $i,\ j$[33]. In the definition of Section Appendix B, the regularized correlation coefficient defines this inner-product. If $\mathbf{y} = (\mathbf{Y}_1 \dots \mathbf{Y}_N)$, then $\mathrm{F} = \mathbf{y}'\mathbf{y}$. Consider the SVD of $\mathbf{y} = \mathbf{V}\Lambda^{\frac{1}{2}}\mathbf{U}' = \sum_{n=0}^{N} \lambda_n^{\frac{1}{2}} \mathbf{v}_n \mathbf{u}_n'$. Eliminating the contribution of $\mathbf{v}_1$, the left singular vector corresponding to the second eigenvector $\mathbf{u}_1$, from the functional connectivity, yields $\widehat{\mathbf{y}} = \mathbf{y} - \mathbf{v}_1\mathbf{v}_1'\mathbf{y} = \mathbf{y} - \lambda_1^{\frac{1}{2}}\mathbf{v}_1\mathbf{u}_1'$, and therefore,

$$\widehat{\mathrm{F}} = \widehat{\mathbf{y}}'\widehat{\mathbf{y}} = \mathbf{y}'\mathbf{y} - \lambda_1 \mathbf{u}_1\mathbf{u}_1'. \tag{C.1}$$

Appendix C.3. Proofs for the Linear Approximation

This section contains proofs showing that an orthogonal tranformation $\Phi = \{\phi^{(1)} \dots \phi^{(N)}\}$ where $\sum_i \phi_i^{(l)} = 0$ yields a lower and upper bound to the transportation problem, and that the basis Ψ constructed in Section Appendix C.1 is a tight bound. Define $S = \mathbf{z}_{t_1} - \mathbf{z}_{t_2}$ to be the difference between the two activity patterns $\mathbf{z}_{t_1}$ and $\mathbf{z}_{t_2}$ to be compared. In the discussion that follows, it is assumed without loss of generality[1], that $\sum_i [\mathbf{z}_i]_{t_1} = \sum_i [\mathbf{z}_i]_{t_2}$ i.e. $\sum_i S_i = 0$. The transportation distance is then:

$$\mathrm{TD}\{\mathbf{z}_{t_1}, \mathbf{z}_{t_2}\} = \min_{q \in \mathbb{R}^{N \times N}} \sum_{i,j \in x} q_{i,j} d_{i,j}, \tag{C.2}$$

subject to the constraints:

$$\begin{aligned} q_{i,j} &\geq 0 \\ \sum_i q_{i,j} - \sum_j q_{j,i} &= S_i \end{aligned} \tag{C.3}$$

The dual formulation of the transportation problem is:

$$\mathrm{TD}\{\mathbf{z}_{t_1}, \mathbf{z}_{t_2}\} = \max_{g \in \mathbb{R}^N} \sum_{i \in x} g_i S_i, \tag{C.4}$$

subject to the constraints:

$$\begin{aligned} g_i - g_j &\leq d_{i,j} \\ \sum_i g_i &= 0. \end{aligned}$$

Now, if we define the coefficients of a vector $\upsilon : x \to \mathbb{R}$ in the basis Φ be $\tilde{\upsilon}_l = \langle \phi^{(l)}, \upsilon \rangle$, the following theorem holds:

[1] This condition can be easily satisfied by adding to the optimization problem of eqn. C.2 a dummy node i' called the dump, where $S_{i'} = -\sum_i S_i$ and $d_{i,i'} = 0,\ \forall i \in x$.

Theorem 2. *Consider the optimization problem of eqn. C.4. Let $\tilde{S}_l$ be coefficients S in the basis Φ. Then, there exist constants $M_l > 0$ and $M_l \geq \widehat{M}_l > 0$, such that*

$$\sum_{l=0}^{N} \widehat{M}_l |\tilde{S}_l| \leq \max_g \sum_{i \in x} g_i S_i \leq \sum_{l=0}^{N} M_l |\tilde{S}_l| \tag{C.5}$$

To prove this theorem, the next two lemmas are required. The first lemma will help establish the upper bound property, while the second lemma will be needed to prove the lower bound.

Lemma 1. *If $\sum_i \upsilon_i = 0$ and $|\upsilon_i - \upsilon_j| \leq d_{i,j}$, then there exist constants $M_l, l = 1 \ldots N$, such that $|\tilde{\upsilon}_l| \leq M_l$*

Proof.

$$\begin{aligned}
|\tilde{\upsilon}_l| &= \left| \sum_{i \in x} \upsilon_i \phi_i^{(l)} \right| \\
&= \left| \sum_i (\upsilon_i - \upsilon_{i_0}) \phi_i^{(l)} + \upsilon_{i_0} \sum_i \phi_i^{(l)} \right|, \\
&\leq \sum_i |\upsilon_i - \upsilon_{i_0}| \cdot |\phi_i^{(l)}| \qquad \left(\text{Since,} \sum_i \phi_i^{(l)} = 0 \right) \\
&\leq \sum_i d_{i,i_0} |\phi_i^{(l)}| \\
&\leq \sup_{i,j \in x} d_{i,j} \sum_i |\phi_i^{(l)}| + c \qquad \text{(C.6)} \\
&= M_l \qquad \text{(C.7)}
\end{aligned}$$

□

If we consider the basis vector $\psi^{(l,m)}$ as defined in Section Appendix C.1 the upper bound is:

$$\sum_{i \in x} d_{i,i_0} |\psi_i^{(l,m)}| = \sum_{i \in x^{(l,m)}} d_{i,i_0} |\psi_i^{(l,m)}| \leq \sup_{i,j \in x^{(l,m)}} d_{i,j} \left[\sup_{l,m} \sum_i |\psi_i^{(l,m)}| \right]$$

and, for every level of decomposition m, it can be shown that $\sup_{i,j \in x^{(l,m)}} d_{i,j}$ decays like 2^{-m}. Therefore, the upper bound on coefficients of the function υ decays according to $|\tilde{\upsilon}_{(l,m)}| \leq 2^{-m} M_{0,0}$, in the basis Ψ.

Lemma 2. *There exist positive constants $\widehat{M}_l$, $0 < \widehat{M}_l \leq M_l$, $l = 1 \ldots N$, such that the set of vectors $\{\upsilon \in \mathbb{R}^N\}$, where $|\tilde{\upsilon}_l| \leq \widehat{M}_l$ must satisfy the property $|\upsilon_i - \upsilon_j| \leq d_{i,j}$.*

Proof. If any function satisfies $|\upsilon_i - \upsilon_j| \leq d_{i,j}$ then $\upsilon + c$ will also satisfy this property, for any constant c. Therefore, we shall prove the lemma for the subset of vectors that have the property $\sum_i \upsilon_i = 0$. Now, if $\forall i' \in x$, $|\upsilon_i'| \leq \inf_{i,j} d_{i,j}$ then it must be that $|\upsilon_i - \upsilon_j| \leq d_{i,j}$. Also, because

Φ is an orthogonal basis, it is true that if $|\tilde{\upsilon}_l| \leq \widehat{M}_l$, then

$$\begin{aligned}\sup_{\upsilon}\sup_{i\in x}|\upsilon_i| &= \sup_{\upsilon}\sup_{i\in x}\left|\sum_{l=0}^{N}\tilde{\upsilon}_l\phi_i^{(l)}\right| \\ &\leq \sup_{\upsilon}\sup_{i}\left[\sum_{l=0}^{N}\widehat{M}_l|\phi_i^{(l)}|\right] \\ &= \sup_{i}\left[\sum_{l=0}^{N}\widehat{M}_l|\phi_i^{(l)}|\right]\end{aligned}$$

There exist many combinations of $\{\widehat{M}_l, l = 1\ldots N\}$ such that $\sup_{\upsilon}\sup_{i\in\mathbb{C}}|\upsilon_i| \leq \inf_{i,j} d_{i,j}$. For example, by setting:

$$\widehat{M}_l = \frac{\inf_{i,j} d_{i,j}}{N}\frac{1}{\sum_i |\psi_i^{(l)}|}$$

this property is ensured. □

For the basis Ψ, first observe that, by construction, $\sum_i \psi_i^{(l,m)} = 0, \sum_i |\psi_i^{(l,m)}|^2 = 1$, and therefore, $\sum_{l=0}^{2^m-1}\sum_i |\psi_i^{(l,m)}|^2 = 2^m$. Also, note that for an N dimensional vector υ, $||\upsilon||_1 \leq \sqrt{N}||\upsilon||_2$. Therefore, this bound becomes:

$$\widehat{M}_{l,m} = \frac{\inf_{i,j} d_{i,j}}{N}\frac{1}{\sum_i |\psi_i^{(l,m)}|} \approx 2^{-m}\widehat{M}_{0,0}$$

Using these two lemmas, Theorem 2 is now proved as follows:

Proof. (**Theorem 2**). Since Φ is an orthogonal transformation, $\sum_{i\in x} g_i s_i = \sum_{l=0}^{N}\tilde{g}_l\tilde{\mathbf{z}}_l$,. The upper-bound then follows from Lemma 1.

For the lower bound, assume that g^* is the optimal solution such that $\sum_{i\in x} g_i^* s_i < \sum_{l=0}^{N}\widehat{M}_l|\tilde{\mathbf{z}}_l|$. However, as per Lemma 2, the function $g^+ = \sum_{l=0}^{N}\mathrm{sgn}(\tilde{\mathbf{z}}_l)\widehat{M}_l\phi_l$ is also a feasible solution with cost $\sum_{l=0}^{N}\widehat{M}_l|\tilde{\mathbf{z}}_l|$. Therefore, g^* cannot be the optimal solution, resulting in a contradiction. □

Appendix C.4. Quality of the Approximation

The quality of the approximation is evaluated by the tightness of the bound:

$$\sup_{||\mathbf{z}||_2=1}\left[\sum_{l=0}^{N}M_l|\tilde{\mathbf{z}}_l| - \sum_{l=0}^{N}\widehat{M}_l|\tilde{\mathbf{z}}_l|\right] = \frac{1}{2}\sqrt{\sum_{l=0}^{N}(M_l - \widehat{M}_l)^2}, \tag{C.8}$$

obtained through the method of Lagrange multipliers. For the basis Ψ, eqn. C.8 is approximately equal to $(M_{0,0} - \widehat{M}_{0,0})/\sqrt{2}$.

As is shown by the proofs in Appendix C.3, this approximation is valid for other orthogonal bases Φ defined on x with respect to the distance metric induced by F (if $\sum_i \phi_i^{(l)} = 0$), and their tightness can be (numerically) evaluated. The approximation Ψ defined in Section Appendix C.1 was compared with respect to the following other basis for the functional connectivity maps F over all the subjects in our data-set. The minimum and maximum values of the bound-tightness metric, relative to the average value for Ψ, are listed:

i The delta-basis $\{\delta_i, i = 1 \ldots N\}$, i.e. the original voxel-wise data itself: 8.43 – 11.58

ii The PCA-like basis consisting of the eigenvectors of F: 3.21–4.66

iii The Laplacian eigenmap[2] basis containing the eigenvectors of the normalized graph Laplacian of F : 1.79–2.30 .

iv The basis set containing indicator functions on recursive normalized cuts [36] of the graph defined by F: 2.02–2.95)

v The diffusion wavelet[7] basis induced by F: 0.89–1.13.

vi An orthogonal basis derived from the spatial ICA decomposition[25] of the fMRI data: 3.51–5.87.

One reason for the comparatively tight bound of Ψ is the fast decay of the coefficients in this basis, thereby making their contribution to the error negligible. The relatively similar values of [iii] and [iv] are because they are obtained from a similar set of operations on F, and the basis vectors share a lot of properties in common, such as coefficient decay. Although the diffusion wavelet basis is tighter approximation to the distance metric TD, its marginally better performance is offset by its much greater computational complexity. The high variance of the ICA derived basis could be because it is not directly related to F and also because the coefficients in this basis are not sparse.

It was also observed that the projection of the fMRI data $\mathbf{y}$ on to the feature-space results in a spatial de-correlation of the coefficients roughly as $\mathrm{Correl}\{[\tilde{\mathbf{y}}_{l_1,m_1}]), [\tilde{\mathbf{y}}_{l_2,m_2}]\} \propto |2^{m_1} l_1 - 2^{m_2} l_2|^{-\beta}$, with $\beta \approx 2.3$, which may be explained by the fact that Ψ corresponds to Karhunen-Loéve type decomposition of the spatial correlation in the fMRI data.

Appendix D. Model Estimation

This section is organized as follows: in Section Appendix D.1, the derivation of the EM algorithm for the proposed model is given, and then Section Appendix D.2 explains the forward-backward recursions needed in the **M-step** of the EM algorithm. Next, the procedure to marginalize out the HRF filter from the estimates of the parameters is given in Section Appendix D.3. Finally, the estimation of the optimal state-sequence X^* given model parameters θ and observations $\mathbf{y}$ is described in Section Appendix D.4.

Appendix D.1. Expectation Maximization

The maximum likelihood (ML) estimate $\theta_{\text{ML}} = \arg\max_\theta \ln p(\mathbf{y}|\theta, \mathbf{h}, K)$ can be obtained using the Expectation Maximization (EM) algorithm [8] by decomposing the log-probability into a free-energy and a KL-divergence term as:

$$\ln p(\mathbf{y}|\theta, \mathbf{h}, K) = \sum_X q(X) \ln \frac{p(\mathbf{y}, X, \theta|K)}{q(X)} + \text{KL}(q||p(X|\mathbf{y}, \theta, \mathbf{h}, K)), \tag{D.1}$$

which yields the following two-step iterative algorithm:

$$\textbf{E-step} \qquad \mathcal{Q}(\theta, \theta^n) = \sum_X p(X|\mathbf{y}, \theta^n) \ln p(\mathbf{y}, X|\theta), \tag{D.2}$$

$$\textbf{M-step} \qquad \theta^{n+1} = \arg\max_\theta \mathcal{Q}(\theta, \theta^n). \tag{D.3}$$

The complete log-likelihood term is:

$$\ln p(\mathbf{y}, X|\theta) = \ln p(Y|X, \omega, \Sigma_\epsilon) + \ln p(X|\alpha, \pi) \tag{D.4}$$

$$\text{where} \quad \ln p(X|\alpha, \pi) = \ln \alpha_{X_1} + \sum_{t=2}^{T} \ln \pi_{X_t, X_{t+1}}.$$

Since the relationship between the observations $\mathbf{y}$ and hidden states X is mediated through the underlying activation patterns $\mathbf{z}$ and the hemodynamic response function $\mathbf{h}$, an FIR filter of length L, as per the equation $\mathbf{y}_t = \sum_{\tau=0}^{L} \mathbf{z}_{t-\tau}\mathbf{h}_\tau$, we see that:

$$p_\theta(\mathbf{y}_t|X) = \int_{\mathbf{z}_{t-L\ldots t}} p_\theta(\mathbf{y}_t|\mathbf{z}_{t-L\ldots t}) \prod_{\tau=0}^{L} p_\theta(\mathbf{z}_{t-\tau}|X_{t-\tau}) d\mathbf{z}_{t-L\ldots t} = \mathcal{N}\left(\mu_{t-L\ldots t}, \Sigma_{t-L\ldots t}\right), \tag{D.5}$$

where $\mu_{t-L\ldots t} = \sum_{\tau=0}^{L} \mu_{X_{t-\tau}}\mathbf{h}_\tau$ and $\Sigma_{t-L\ldots t} = \Sigma_\epsilon + \sum_{\tau=0}^{L} \Sigma_{X_{t-\tau}}\mathbf{h}_\tau^2$.

If we consider one particular assignment of $X_{t-L\ldots t} = \{k_0 \ldots k_L\}$ and let $\mu_{k_0\ldots k_L} = \sum_{\tau=0}^{L} \mu_{k_\tau}\mathbf{h}_{L-\tau}$, then any element $\mu^{(i)}_{k_0\ldots k_L}$ of $\mu_{k_0\ldots k_L}$ of $\mu_{k_0\ldots k_L}$ is a linear combination of the corresponding el-

ements of $\mu_1 \dots \mu_K$, as:

$$\begin{pmatrix} \mu^{(i)}_{1\dots1} \\ \mu^{(i)}_{1\dots2} \\ \vdots \\ \mu^{(i)}_{K\dots K-1} \\ \mu^{(i)}_{K\dots K} \end{pmatrix} = \begin{pmatrix} \mathbf{h}_L + \dots \mathbf{h}_0 & 0 & \dots & 0 & 0 \\ \mathbf{h}_L + \dots \mathbf{h}_1 & \mathbf{h}_0 & \dots & 0 & 0 \\ \vdots & \vdots & \dots & \vdots & \vdots \\ 0 & 0 & \dots & \mathbf{h}_L & \mathbf{h}_{L-1} + \mathbf{h}_0 \\ 0 & 0 & \dots & 0 & \mathbf{h}_L + \mathbf{h}_0 \end{pmatrix} \begin{pmatrix} \mu^{(i)}_1 \\ \mu^{(i)}_2 \\ \vdots \\ \mu^{(i)}_{K-1} \\ \mu^{(i)}_K \end{pmatrix}.$$

In matrix notation,

$$\vec{\mu}^{(i)}_{k_0\dots k_L} = \mathbf{H}\vec{\mu}^{(i)}_k \quad \text{and,} \quad \vec{\mu}^{(i)}_k = \mathbf{H}^- \vec{\mu}^{(i)}_{k_0\dots k_L}, \tag{D.6}$$

where $\mathbf{H}^-$ is the pseudo-inverse of $\mathbf{H}$.

Similarly, each element $\Sigma^{(i_1,i_2)}_{k_0\dots k_L}$ of $\Sigma_{k_0\dots k_L}$ is related to the corresponding elements of $\Sigma_1 \dots \Sigma_K$ as:

$$\begin{pmatrix} \Sigma^{(i_1,i_2)}_{1\dots1} \\ \Sigma^{(i_1,i_2)}_{1\dots2} \\ \vdots \\ \Sigma^{(i_1,i_2)}_{K\dots K-1} \\ \Sigma^{(i_1,i_2)}_{K\dots K} \end{pmatrix} = \begin{pmatrix} \mathbf{h}^2_L + \dots \mathbf{h}^2_0 & 0 & \dots & 0 & 0 \\ \mathbf{h}^2_L + \dots \mathbf{h}^2_1 & \mathbf{h}^2_0 & \dots & 0 & 0 \\ \vdots & \vdots & \dots & \vdots & \vdots \\ 0 & 0 & \dots & \mathbf{h}^2_L & \mathbf{h}^2_{L-1} + \mathbf{h}^2_0 \\ 0 & 0 & \dots & 0 & \mathbf{h}^2_L + \mathbf{h}^2_0 \end{pmatrix} \begin{pmatrix} \Sigma^{(i_1,i_2)}_1 \\ \Sigma^{(i_1,i_2)}_2 \\ \vdots \\ \Sigma^{(i_1,i_2)}_{K-1} \\ \Sigma^{(i_1,i_2)}_K \\ \Sigma^{(i_1,i_2)}_\epsilon \end{pmatrix}.$$

In matrix notation,

$$\vec{\Sigma}^{(i_1,i_2)}_{k_0\dots k_L} = \mathbf{G}\vec{\Sigma^{}}^{(i_1,i_2)}_k \quad \text{and,} \quad \vec{\Sigma^{}}^{(i_1,i_2)}_k = \mathbf{G}^- \vec{\Sigma}^{(i_1,i_2)}_{k_0\dots k_L}. \tag{D.7}$$

Furthermore,

$$\begin{aligned} \ln p(\mathbf{y}|X,\omega,\Sigma_\epsilon) &= \sum_{t=1}^{T} \ln p(\mathbf{y}_t|X_{t-L\dots t},\omega,\Sigma_\epsilon) \\ &= -\frac{1}{2}\left[\sum_{t=1}^{T} \ln|\Sigma_{t-L\dots t}| + (\mathbf{y}_t - \mu_{t-L\dots t})'\Sigma^{-1}_{t-L\dots t}(\mathbf{y}_t - \mu_{t-L\dots t})\right] + c. \end{aligned} \tag{D.8}$$

Therefore, by substituting the results of eqns. D.4 and D.8 in eqn. D.2, and interchanging the

order of the summations, the expected complete log-likelihood becomes:

$$\begin{aligned}\mathcal{Q}(\theta,\theta^n) = & \sum_X p(X|\mathbf{y},\theta^n)\left[\ln p(Y|X,\omega,\Sigma_\epsilon) + \ln p(X|\alpha,\pi)\right] \\ = & \; p(X_1|\mathbf{y},\theta^n)\ln\alpha_{X_1} + \sum_{t=2}^{T} p(X_{t-1,t}|\mathbf{y},\theta^n)\ln\pi_{X_{t-1},X_t} \\ & + \sum_{t=1}^{T}\sum_{X_{t-L\ldots t}} p(X_{t-L\ldots t}|\mathbf{y},\theta^n)\ln p(\mathbf{y}_t|X_{t-L\ldots t},\omega,\Sigma_\epsilon). \end{aligned} \tag{D.9}$$

The **M-step** for α, π, constrained to $\sum_{k=1}^K \alpha_k = 1,\ \sum_{k'=1}^K \pi_{k,k'} = 1$, results in:

$$\begin{aligned}\alpha_k^{n+1} &= \frac{p(X_1 = k|\mathbf{y},\theta^n)}{\sum_{k'=1}^K p(X_1 = k'|\mathbf{y},\theta^n)} \\ \pi_{k_1,k_2}^{n+1} &= \frac{\sum_{t=2}^T p(X_t = k_1, X_{t+1} = k_2|\mathbf{y},\theta^n)}{\sum_{k'=1}^K\sum_{t=2}^T p(X_t = k_1, X_{t+1} = k'|\mathbf{y},\theta^n)}. \end{aligned} \tag{D.10}$$

To determine the **M-step** update μ_k^{n+1}, from eqns. D.8,D.9, first observe that:

$$\begin{aligned}&\sum_{t=1}^{T}\sum_{X_{t-L\ldots t}} p(X_{t-L\ldots t}|\mathbf{y},\theta^n)\ln p(\mathbf{y}_t|X_{t-L\ldots t},\omega,\Sigma_\epsilon) \\ &\qquad \propto \sum_{t=1}^{T}\sum_{X_{t-L\ldots t}} p(X_{t-L\ldots t}|\mathbf{y},\theta^n)\left[\ln|\Sigma_{t-L\ldots t}| + (\mathbf{y}_t - \mu_{t-L\ldots t})'[\Sigma_{t-L\ldots t}]^{-1}(\mathbf{y}_t - \mu_{t-L\ldots t})\right]. \end{aligned} \tag{D.11}$$

Maximizing eqn. D.11 with respect to one specific instantiation of states $X_{t-L\ldots t} = k_0 \ldots k_L$ gives:

$$\mu_{k_0\ldots k_L}^{n+1} = \frac{\sum_{t=1}^T p(X_{t-L\ldots t} = k_0\ldots k_L|\mathbf{y},\theta^n)\mathbf{y}_t}{\sum_{t=1}^T p(X_{t-L\ldots t} = k_0\ldots k_L|\mathbf{y},\theta^n)}, \tag{D.12}$$

and from eqn. D.6, we get $\mu_k^{n+1} = \sum_{k_0\ldots k_L} \mathbf{H}^-_{k,k_0\ldots k_L}\mu^{n+1}_{k_0\ldots k_L}$.

Similarly, maximizing eqn. D.11 with respect to a specific $\Sigma_{k_0\ldots k_L}$, gives

$$\Sigma_{k_0\ldots k_L}^{n+1} = \frac{\sum_{t=1}^T p(X_{t-L\ldots t} = k_0\ldots k_L|\mathbf{y},\theta^n)\cdot(\mathbf{y}_t - \mu^{n+1}_{k_0\ldots k_L})(\mathbf{y}_t - \mu^{n+1}_{k_0\ldots k_L})'}{\sum_{t=1}^T p(X_{t-L\ldots t} = k_0\ldots k_L|\mathbf{y},\theta^n)}, \tag{D.13}$$

where as before $\Sigma_k^{n+1} = \sum_{k_0\ldots k_L}\mathbf{G}^-_{k,k_0\ldots k_L}\Sigma^{n+1}_{k_0\ldots k_L}$. A similar relationship applies for Σ_ϵ.

Appendix D.2. Forward Backward Recursions

This section explains the forward-backward recursions to compute the probabilities of the form $p_{\theta^{(n)}}(\mathbf{y}, X_t)$,$p_{\theta^{(n)}}(\mathbf{y}, X_{t-1}, X_t)$ and $p_{\theta^{(n)}}(\mathbf{y}, X_{t-L\ldots t})$, needed in the **M-step** of the EM algorithm.

From the conditional independence structure implied by the model, we observe that:

$$\begin{aligned} p_\theta(\mathbf{y}, X_t) &= \sum_{X_{t+1-L\ldots t-1}} p_\theta(\mathbf{y}, X_{t+1-L\ldots t}) \\ &= \sum_{X_{t+1-L\ldots t-1}} p_\theta(\mathbf{y}_{1\ldots t}, X_{t+1-L\ldots t}) p_\theta(\mathbf{y}_{t+1\ldots T} | X_{t+1-L\ldots t}) \end{aligned} \tag{D.14}$$

$$\begin{aligned} p_\theta(\mathbf{y}, X_{t,t+1}) &= \sum_{X_{t+1-L\ldots t-1}} p_\theta(\mathbf{y}, X_{t+1-L\ldots t-1}, X_{t,t+1}) \\ &= \sum_{X_{t+1-L\ldots t-1}} p_\theta(\mathbf{y}_{1\ldots t}, X_{t+1-L\ldots t}) \cdot p_\theta(\mathbf{y}_{t+1} | X_{t+1-L\ldots t+1}) \\ &\qquad \cdot p_\theta(X_{t+1} | p_\theta(X_t) \cdot p_\theta(\mathbf{y}_{t+2\ldots T} | X_{t+2-L\ldots t+1}), \end{aligned} \tag{D.15}$$

and

$$\begin{aligned} p_\theta(\mathbf{y}, X_{t\ldots t+L}) =& p_\theta(\mathbf{y}_{1\ldots t+L}, X_{t\ldots t+L}) \cdot p_\theta(\mathbf{y}_{t+1+L\ldots T} | X_{t\ldots t+L}) \\ =& p_\theta(\mathbf{y}_{1\ldots t-1+L}, X_{t\ldots t-1+L}) \cdot p_\theta(\mathbf{y}_{t+L}, X_{t\ldots t+L}) \cdot p_\theta(X_{t\ldots t+L}) \\ & \cdot p_\theta(\mathbf{y}_{t+1+L\ldots T} | X_{t+1\ldots t+L}), \end{aligned} \tag{D.16}$$

where L is the length of the FIR filter $\mathbf{h}$.

The forward recursion through this Markov chain is:

$$\begin{aligned} a(X_1) &= p_\theta(\mathbf{y}_1, X_1) \\ &= p_\theta(\mathbf{y}_1 | X_1) p_\theta(X_1), \end{aligned} \tag{D.17}$$

$$\begin{aligned} a(X_{1\ldots 2}) &= p_\theta(\mathbf{y}_{1\ldots 2}, X_{1\ldots 2}) \\ &= p_\theta(\mathbf{y}_2 | X_{1\ldots 2}) p_\theta(\mathbf{y}_1 | X_1) p_\theta(X_2 | X_1) p_\theta(X_1) \\ &= p_\theta(\mathbf{y}_2 | X_{1\ldots 2}) p_\theta(X_2 | X_1) \cdot a(X_1). \end{aligned} \tag{D.18}$$

Similarly, continuing up to,

$$\begin{aligned} a(X_{1\ldots L}) &= p_\theta(\mathbf{y}_{1\ldots L}, X_{1\ldots L}) \\ &= p_\theta(\mathbf{y}_L | X_{1\ldots L}) p_\theta(X_L | X_{L-1}) \cdot a(X_{1\ldots L-1}). \end{aligned}$$

Now, after we have at least L observations $\mathbf{y}$

$$\begin{aligned} a(X_{2\dots L+1}) &= p_\theta(\mathbf{y}_{1\dots L+1}, X_{2\dots L+1}) \\ &= \sum_{X_1} p_\theta(X_{1\dots L+1}, \mathbf{y}_{1\dots L+1}) \\ &= p_\theta(\mathbf{y}_{L+1}|X_{1\dots L+1}) p_\theta(X_{L+1}|X_L) \cdot a(X_{1\dots L}). \end{aligned}$$

And similarly,

$$\begin{aligned} a(X_{3\dots L+2}) &= p_\theta(X_{3\dots L+2}, \mathbf{y}_{1\dots L+2}) \\ &= \sum_{X_2} p_\theta(X_{2\dots L+2}, \mathbf{y}_{1\dots L+2}) \\ &= \sum_{X_2} p_\theta(\mathbf{y}_{L+2}|X_{2\dots L+2}) p_\theta(X_{L+2}|X_{L+1}) \cdot a(X_{2\dots L+1}), \end{aligned}$$

upto,

$$\begin{aligned} a(X_{t+1-L\dots t}) &= \sum_{X(t-L)} p_\theta(X_{t-L\dots t}, \mathbf{y}_{1\dots t}) \\ &= \sum_{X(t-L)} p_\theta(\mathbf{y}_t|X_{t-L\dots t}) p_\theta(X_t|X_{t-1}) \cdot a(X_{t-L\dots t-1}). \end{aligned} \tag{D.19}$$

The backward recursion for this chain is as follows:

$$\begin{aligned} b(X_{T-L\dots T-1}) &= p_\theta(\mathbf{y}_T|X_{T-L\dots T-1}) \\ &= \sum_{X_T} p_\theta(\mathbf{y}_T|X_{T-L\dots T}), \\ b(X_{T-1-L\dots T-2}) &= p_\theta(\mathbf{y}_{T-1\dots T}|X_{T-1-L\dots T-2}) \\ &= \sum_{X_{T-1}} p_\theta(\mathbf{y}_{T-1}|X_{T-1-L\dots T-1}) p_\theta(\mathbf{y}_T|X_{T-L\dots T-1}) \\ &= \sum_{X_{T-1}} p_\theta(\mathbf{y}_{T-1}|X_{T-1-L\dots T-1}) b(X_{T-L\dots T-1}), \end{aligned}$$

and similarly,

$$\begin{aligned} b(X_{t+1-L\dots t}) &= p_\theta(\mathbf{y}_{t+1\dots T}|X_{t+1-L\dots t}) \\ &= \sum_{X_{t+1}} p_\theta(\mathbf{y}_{t+1}|X_{t+1-L\dots t+1}) b(X_{t+2-L\dots t+1}). \end{aligned} \tag{D.20}$$

Therefore, substituting in eqns. D.14 to D.16, the conditional probabilities become:

$$
\begin{aligned}
p_\theta(\mathbf{y}, X_t) &= \sum_{X_{t+1-L\ldots t-1}} a(X_{t+1-L\ldots t}) b(X_{t+1-L\ldots t}), \\
p_\theta(\mathbf{y}, X_{t,t+1}) &= \sum_{X_{t+1-L\ldots t-1}} a(X_{t+1-L\ldots t}) \cdot p_\theta(\mathbf{y}_{t+1} | X_{t+1-L\ldots t+1}) p_\theta(X_{t+1} | X_t) \cdot b(X_{t+2-L\ldots t}), \\
p_\theta(\mathbf{y}, X_{t,t+L}) &= a(X_{t,t-1+L}) \cdot p_\theta(\mathbf{y}_{t+L}, X_{t,t+L}) \cdot p_\theta(X_{t,t+L}) \cdot b(X_{t+1\ldots t+L}).
\end{aligned} \tag{D.21}
$$

Appendix D.3. Marginalizing the HRF Filter $\mathbf{h}$

The EM procedure so far determined θ_{ML} conditioned on a specific HRF filter $\mathbf{h}$. This dependence is removed by marginalizing out $\mathbf{h}$ under a Laplace approximation of the posterior distribution of θ as follows:

Under uninformative priors, the posterior density $p(\theta|\mathbf{y}, \mathbf{h}, K) \propto p(\mathbf{y}|\theta, \mathbf{h}, K)$ and $\theta_{\text{MAP}} = \theta_{\text{ML}}$. Then using a Laplace approximation around θ_{ML} the posterior density is given by:

$$
p(\theta|Y, \mathbf{h}, K) \approx |\frac{1}{2\pi}\nabla^2|^{1/2} \exp\left\{-\frac{1}{2}(\theta - \theta_{\text{ML}})'\nabla^2(\theta - \theta_{\text{ML}})\right\}, \tag{D.22}
$$

where $-\nabla^2$ is the Hessian matrix of $\ln p(\mathbf{y}|\theta, \mathbf{h}, K)$.

Then, the conditional expectation θ^* independent of $\mathbf{h}$ is given by:

$$
\begin{aligned}
\theta^* =& \mathrm{E}[\theta|\mathbf{y}, K] = \mathrm{E}_{\mathbf{h}}\left[\mathrm{E}[\theta|\mathbf{h}\mathbf{y}, K]\right] \\
&= \int_{\mathbf{h}} \left[\int_\theta \theta p(\theta|\mathbf{y}, \mathbf{h}, K) d\theta\right] p(\mathbf{h}) d\mathbf{h} \\
&= \int_{\mathbf{h}} \theta_{\text{ML}}(\mathbf{h}) p(\mathbf{h}) d\mathbf{h},
\end{aligned} \tag{D.23}
$$

and is computed through Monte Carlo integration by first sampling the parameter γ from $\mathcal{N}(\mu_\gamma, \sigma_\gamma)$, constructing $\mathbf{h}(\gamma)$, finding $\theta_{\text{ML}}(\mathbf{h})$ and then averaging over all samples.

Appendix D.4. State-Sequence Estimation

In this section, we explain the procedure to find the most probable set of states $X^* = \arg\max \ln p_\theta(\mathbf{y}, X)$ given a set of model parameters θ and observations $\mathbf{y}$.

Note the following recursive relationship:

$$\begin{aligned}
\max_X \ln p_\theta(\mathbf{y}, X) &= \max_X \left[\ln p_\theta(\mathbf{y}_T|X_{T-L\ldots T}) + \ln p_\theta(\mathbf{y}_{1\ldots T-1}, X_{1\ldots T-1})\right] \\
&= \max_{X_{T-L\ldots T}} \left[\ln p_\theta(\mathbf{y}_T|X_{T-L\ldots T}) + \max_{X_{1\ldots T-1-L}} \ln p_\theta(\mathbf{y}_{1\ldots T-1}, X_{1\ldots T-1})\right] \\
&= \max_{X_{T-L\ldots T}} \left[\ln p_\theta(Y_T|X_{T-L\ldots T}) + \max_{X(T-1-L)} \left[\ln p_\theta(Y_{T-1}|X_{T-1-L\ldots T-1}) + \right.\right. \\
&\qquad \left.\left. \max_{X_{1\ldots T-2-L}} \ln p_\theta(\mathbf{y}_{1\ldots T-2}, X_{1\ldots T-2})\right]\right] \\
&\vdots
\end{aligned}$$

Therefore, if we define:

$$\begin{aligned}
\eta_1 &= \ln p_\theta(\mathbf{y}_1, X_1) &&= \ln p_\theta(\mathbf{y}_1|X_1) + \ln p_\theta(X_1), \\
\eta_2 &= \max_{X_1} \ln p_\theta(\mathbf{y}_{1,2}, X_{1,2}) &&= \max_{X_1} \left[\ln p_\theta(\mathbf{y}_2|X_{1,2}) + \ln p_\theta(X_2|X_1) + \eta_1\right], \\
&\vdots \\
\eta_t &= \max_{X_{t-1}} \left[\ln p_\theta(\mathbf{y}_t, X_{t-L\ldots t}) + \ln p_\theta(X_t|X_{t-1}) + \eta_{t-1}\right], \\
\eta_{t+1} &= \max_{X_t} \left[\ln p_\theta(\mathbf{y}_{t+1}, X_{t+1-L\ldots t+1}) + \ln p_\theta(X_{t+1}|X_t) + \eta_t\right],
\end{aligned}$$

then it can be verified that $\max_X \ln p_\theta(\mathbf{y}, X) = \max_{X_{t-L\ldots T}} \eta_T$.

Let $\varphi_t(X_{t\ldots t+L})$ keep track of the state of X_{t-1} which is a maximum configuration for η_t, given $X_{t\ldots t+L}$. Then, the optimal configuration of states X^* for a particular θ are obtained by backtracking as follows:

$$\begin{aligned}
X^*_{t-L\ldots T} &= \arg \max_{X_{t-L\ldots T}} \eta_T, \\
X^*_{t-L-1} &= \varphi_{T-1}(X^*_{t-L\ldots T}), \\
&\vdots \\
X^*_1 &= \varphi_L(X^*_{2\ldots L}).
\end{aligned} \tag{D.24}$$

Appendix E. Data-set, Analysis and Results

Appendix E.1. Subjects and Paradigm

Thirty-six control subjects and thirteen high-performing (fullscale IQ>95) individuals with pure dyscalculia (DC) [35] participated (controls: 23 female, one female and one male lefthanded, one male ambidextrous, age 21-34 yrs, mean age 25.6 yrs $\pm$ 3.0 yrs; DC: 5 female, 1 male lefthanded, age 22-23yrs). All subjects were free of neurological and psychiatric illnesses, dyslexia, and attention-deficit disorder. All controls denied a history of any calculation difficulties. The layout in Fig. E.4 illustrates the self-paced, irregular paradigm used in these experiments. Subjects were exposed visually to simple multiplication problems with single digit operands, e.g., 4×5 , and had to decide if the incorrect solution subsequently offered was, e.g., *close* for 23 , *too small* for 12 , or *too big* for 27 from the correct result of 20. All solutions were within $\pm 50\%$ of the correct answer. Only one solution was presented at the time. The *close* answer had to be applied for solutions that were within $\pm 25\%$ of the correct result, while the two remaining exceeded this threshold. Subjects answered by pressing a button with the index finger of the dominant hand for *too small*, the middle finger for *close*, and the ring finger for *too big*. Identical operand pairs were excluded. The simplest operand pair was 3×4, while the most demanding pair was 8×9. The order of small vs. large operands was approximately counterbalanced. Presentation times were the following: multiplication problem 2.5s, equal sign (=) 0.3s, solution 0.8s, judgment period up to 4s, and rest condition with fixation point of 1 s until the beginning of a new cycle. Subjects were encouraged to respond as quickly as possible. Stimulus onset asynchrony (SOA) ranged from around 4s to 8.6 s. All subjects were exposed to two different sets of multiplication problems, with an interval of approximately 30min between sessions 1 and 2 during which time they solved other nonnumerical tasks.

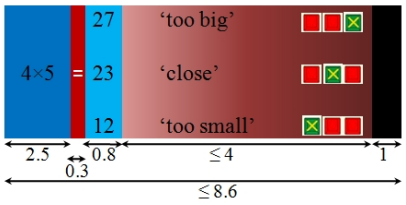

Figure E.4: The five phases of each trial and their associated timings of the paradigm to study arithmetical abilities.

Data were acquired on a General Electric 3-Tesla MRI scanner (vh3) with a quadrature head coil. After localizer scans, a first anatomical, axial-oblique 3D-SPGR volume was acquired. Slab coordinates and matrix size corresponded to those applied during the subsequent fMRI runs using a 3D PRESTO BOLD pulse sequence[40] with phase navigator correction and the following specifications: echo time 40ms, repetition time 26.4ms, echo train length 17, flip angle 17° , volume scan time 2.64s, number of scans 280, session scan time 12:19 min, 3D matrix size $51 \times 64 \times 32$, and isotropic voxel size 3.75mm. At the end of the study, a sagittal 3DSPGR scan was acquired with a slice thickness of 1.2mm and in-plane resolution of 0.94mm. The first four fMRI scans were discarded leaving 276 scans for analysis. Raw data were reconstructed off-line. The structural scans were bias–field corrected, normalized to an MNI atlas space and segmentation into grey and white matter, while the fMRI scans were motion corrected using linear registration and co-registered with the structural scan in SPM8[12]. Further motion correction was performed using Motion Corrected Independent Component Analysis [23]. The fMRI data were then de-noised using a wavelet–based Wiener filter[1] and high-pass filtered to remove artifacts such as breathing, pulsatile effects, and scanner drift. The mean volume of the time-series was then subtracted and white matter masked out.

Appendix E.2. Analysis Parameters

If $R_c = a \times b$ is the correct product for the multiplication problem $a \times b$ and R_d is the displayed incorrect result, then the product size score $\mathrm{LogPs} = \log(R_c)$. The score LogDiff is $\log(|(1.25R_c) - (R_c + |R_c - R_d|)|/|1.25R_c|)$, which measures the closeness of the incorrect result to the $\pm 25\%$ mark and represents the difficulty subjects would have in judging the correct answer as *close* vs. *too big* or *too small*.

The entropy for the MLR probabilities $f_X(\hat{\mathbf{s}}) = (\Pr[\hat{X} = 1], \ldots \Pr[\hat{X} = K])$ corresponding to a given experimental condition $\hat{\mathbf{s}}$ is defined as:

$$\mathrm{ENT}(\hat{\mathbf{s}}) = -\sum_{k=1}^{K} \Pr[\hat{X} = k] \ln \Pr[\hat{X} = k]. \tag{E.1}$$

It measures the predictability of the state of X_t when the prevailing experimental conditions $\mathbf{s}_t$ is equal to $\hat{\mathbf{s}}$, the given condition of interest.

The error rate $\mathrm{ERR}(\hat{\mathbf{s}})$ for new data $\mathbf{y}$, (i.e. data distinct from what the model was trained on) is measured from its optimal state sequence X_t^* by:

$$\mathrm{ERR}(\hat{\mathbf{s}}) = \frac{\sum_{t=1}^{T}(1 - \Pr[\hat{X} = X_t^*]) \cdot \delta(\mathbf{s}_t = \hat{\mathbf{s}})}{\sum_{t=1}^{T} \delta(\mathbf{s}_t = \hat{\mathbf{s}})}. \tag{E.2}$$

It quantifies how well the new data conforms to the given model, for a given experimental condition $\hat{\mathbf{s}}$.

Let $X^{(1)}$ and $X^{(2)}$ be the optimal state sequences for an fMRI session $\mathbf{y}$ obtained from two different models $(\theta^*, K^*, f_X)^{(1)}$ and $(\theta^*, K^*, f_X)^{(2)}$. The mutual information $\mathrm{MI}(\hat{\mathbf{s}})$ between the two models with respect to the fMRI session $\mathbf{y}$ for condition $\hat{\mathbf{s}}$ is:

$$\mathrm{MI}(\hat{\mathbf{s}}) = H^{(1)}(\hat{\mathbf{s}}) + H^{(2)}(\hat{\mathbf{s}}) - H^{(1),(2)}(\hat{\mathbf{s}}) \tag{E.3}$$

where $H^{(1)}(\hat{\mathbf{s}})$ is the empirical entropy of the states of $X^{(1)}$ measured for only those t when $\mathbf{s}_t = \hat{\mathbf{s}}$ as

$$\Pr_k^{(1)} = \frac{\sum_{t=1}^{T} \Pr[\hat{X}^{(1)} = X_t^{(1)}] \delta(X_t^{(1)} = k) \delta(\mathbf{s}_t = \hat{\mathbf{s}})}{\sum_{t=1}^{T} \delta(X_t^{(1)} = k) \delta(\mathbf{s}_t = \hat{\mathbf{s}})}$$

$$H^{(1)}(\hat{\mathbf{s}}) = \sum_{k=1}^{K} \Pr_k^{(1)} \ln \Pr_k^{(1)}$$

Here $\Pr_k^{(1)}$ is the empirical probability of state k for condition $\hat{\mathbf{s}}$ for the data $\mathbf{y}$, given model 1. Similarly for $H^{(2)}(\hat{\mathbf{s}})$. The empirical joint entropy $H^{(1),(2)}(\hat{\mathbf{s}})$ between $X^{(1)}$ and $X^{(2)}$ for

condition $\hat{\mathbf{s}}$ is equivalently defined as:

$$\Pr^{(1),(2)}_{k_1,k_2} = \frac{\sum_{t=1}^{T} \Pr[\hat{X}^{(1)} = X_t^{(1)}]\delta(X_t^{(1)} = k_1)\Pr[\hat{X}^{(2)} = X_t^{(2)}]\delta(X_t^{(2)} = k_2)\delta(\mathbf{s}_t = \hat{\mathbf{s}})}{\sum_{t=1}^{T} \delta(X_t^{(1)} = k_1)\delta(X_t^{(1)} = k_2)\delta(\mathbf{s}_t = \hat{\mathbf{s}})}$$

$$H^{(1),(2)}(\hat{\mathbf{s}}) = \sum_{k_1}\sum_{k_2} \Pr^{(1),(2)}_{k_1,k_2} \ln \Pr^{(1),(2)}_{k_1,k_2}.$$

The Mutual Information between two models is defined in this way, because in general the correspondence between the state labels of two different models is unknown. By comparing the state-sequences of the same data generated by the two models, this correspondence can be determined.

Appendix E.3. Results

The Ph-wise effect of LogPs and LogDiff on ENT, ERR and MI are shown in Fig. E.5. The LogPs effects correspond to the difference in values of these statistics for $\mathbf{s}_t$ corresponding to LogPs > 2.5 minus those for LogPs ≤ 2.5. Similarly, and LogDiff > 3.0 minus LogDiff ≤ 3.0. All other variables of the experimental condition $\mathbf{s}_t$ have been averaged out. These values are

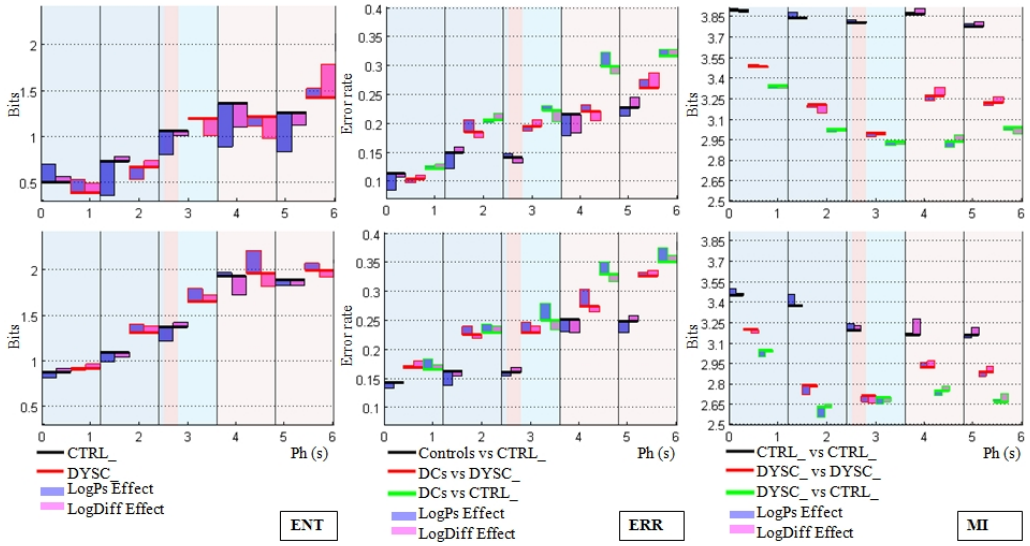

Figure E.5: The effect on ENT, ERR, and MI with respect to experiment phase Ph, LogPs and LogDiff. The top row corresponds to the models CTRL_SELF and DYSC_SELF, while the bottom row to CTRL_GRP and DYSC_GRP. The ENT panel shows the effect of Ph (in 1.2s increments indicated by the vertical grid), high minus low LogPs and high minus low LogDiff on the models (θ^*, K^*, f_X) trained on the controls and dyscalculics, either individually or as groups. The ERR panels show the effects on the error rate of predicting a control's data with a CTRL_SELF or CTRL_GRP (black), a DC subject's data with a DYSC_SELF or DYSC_GRP (red), and a DC subject's data with a CTRL_SELF or CTRL_GRP (green). The MI panel shows the effects on the MI between two CTRL_SELF models (black), two DYSC_SELF models (black) and between an CTRL_SELF and DYSC_SELF model (green). Similarly for CTRL_GRP and DYSC_GRP. The background color-coding shows the 2.5s, 0.3s, 0.8s and 0–4s divisions of each trial, corresponding to Fig. E.4.

also summarized in Table E.1.

Model	CTRL_SELF		DYSC_SELF	CTRL_GRP		DYSC_GRP
Test Data	Controls	Dyscalculics	Dyscalculics	Controls	Dyscalculics	Dyscalculics
Statistic						
K*	22.8 ± 3.21		23.5 ± 5.22	23.2 ± 3.63		22.1 ± 5.78
ENT	1.32 ± 0.11		1.34 ± 0.20	1.58 ± 0.36		1.92 ± 0.42
ERR	0.215 ± 0.06	0.289 ± 0.10	0.227 ± 0.05	0.231 ± 0.09	0.308 ± 0.13	0.264 ± 0.14
MI	3.812 ± 0.12		3.246± 0.15	3.242± 0.34		2.919± 0.40
	2.987± 0.19			2.721± 0.53		

Table E.1: **Statistic averages and ± 1 std-dev for the data-set**.Row 1 shows the optimal K^*, while row 2 shows the ENT for all the CTRL_SELF, DYSC_SELF, CTRL_GRP, and DYSC_GRP models trained. In row 3, the ERR for the following cases is given: (i) control and (ii) DC data vs. CTRL_SELF, (iii) DC data vs. DYSC_SELF, (iv) control and (v) DC data vs. CTRL_GRP and (vi) DC data vs. DYSC_GRP. Row 3 shows the MI between (i) two CTRL_SELF models, (ii) two DYSC_SELF models,(iii) two CTRL_GRP models, (iv) two DYSC_GRP models, while in row 4, the MI of CTRL_SELF vs. DYSC_SELF and CTRL_GRP vs. DYSC_GRP is given.

From these results, we make the following salient observations:

- For the CTRL_SELF and DYSC_SELF models, ENT increases as the task progresses, indicating more unpredictability for the computation phases of the task as compared to the visual presentation phase, during which the visual areas are strongly and commonly recruited. LogPs has an overall negative effect on ENT, indicating higher predictability for problems with larger product sizes. The early effect could be due to a stronger product lookup in the rote tables stored in the left angular gyrus of the lower parietal lobe, and a strong number size response in the visual areas, which would also explain the strong effect in controls as compared to the DCs. Later onset of the effect ($\mathrm{Ph} > 3$s) is strong in controls, consistent with stronger activation of the working verbal (mute-rehearsal) and visual memories[32], but negligible in the DCs, who presumably do not have an intuitive appreciation of product size, and respond similarly to all product sizes. The $\mathrm{LogDiff}$ effect is noticeable after 2.4s, which is expected since it depends on the onset of the incorrect result R_d displayed at 2.8s. The high ENT for $\mathrm{Ph} > 4.8$s and the divergence of effects are because this phase very often corresponds to the post-button-press rest interval before the next trial begins, during which the brain-state is less predictable.
- For the CTRL_GRP models, while the baseline ENT is higher and the LogPs and $\mathrm{LogDiff}$ effects are smaller (most likely due to variations in the activation patterns amongst the subjects), the overall trends are the same as CTRL_SELF. This indicates that controls are well explained by their CTRL_GRP model. However, the picture changes dramatically for the DCs. Not only is the baseline ENT (for $\mathrm{Ph} > 1.2$s) much higher, the direction of the LogPs effect changes. This indicates a failure of DYSC_GRP model to predict the brain-state of the DCs, which is compounded with higher product size. In contrast, $\mathrm{LogDiff}$ has a negative effect for both controls and DCs, and confirms the hypothesis that $\mathrm{LogDiff}$ mainly affects attention and conflict resolution, and is probably not a number-related effect, and therefore is not affected by dyscalculic deficiencies.
- ERR presents a picture similar to that for ENT. For CTRL_SELF, ERR increases with Ph, while LogPs and $\mathrm{LogDiff}$ show negative effects. This indicates that the model for one control is fairly accurate for the data from another control, and gets better with LogPs and $\mathrm{LogDiff}$. Again, the $\mathrm{LogDiff}$ effects starts after $\mathrm{Ph} > 2.4$s. On the other hand,

DYSC_SELF models are not as accurate for the data from other DCs, and get worse with LogPs, while $\mathrm{LogDiff}$ has a negative effect on ERR, which is again probably because $\mathrm{LogDiff}$ is not number-related. This is also the case for CTRL_SELF models vs. DC data (green line). This not only indicates that models built from control data do not accurately represent the DC data, but also that the DCs themselves are not well represented by the models from other DCs.

- The ERR panel for CTRL_GRP and DYSC_GRP shows the same pattern, except at higher baseline error-rates, as expected. Here, it is noteworthy that DCs tested against DYSC_GRP show an error-rate comparable to DCs tested against CTRL_GRP, indicating that at the group level DCs are as dissimilar from each other as controls are from DCs. Again, the high ERR towards $\mathrm{Ph} > 4.8\mathrm{s}$ is due to overlap with the inter-trial rest interval, causing greater inaccuracy in the predictions.
- The MI figures for CTRL_SELF and DYSC_SELF are probably the most revealing. The CTRL_SELF models share high mutual information, throughout the task, with a small positive effect of LogPs and $\mathrm{LogDiff}$, indicating better correspondence with increase in these parameters. The MI between two DYSC_SELF models and between a DYSC_SELF and a CTRL_SELF is significantly lower throughout. The relatively higher MI at the start of the task is probably due to common recruitment patterns in the visual cortices due to the problem presentation. It drops further during mental computation phases, reaching its lowest around Ph =2.4–3.6s, just when the incorrect result R_d is displayed. The DCs align relatively more with each other in the next phase, as compared to with the controls, which could be because the DCs are recapitulating the multiplication problem, after they have finished the task [28] (with strong working memory activation located in the right intra-parietal suclus). This observation is also corroborated by the relatively lower values for ENT and ERR in this phase.
- The MI for CTRL_GRP vs. CTRL_GRP, DYSC_GRP vs. DYSC_GRP and CTRL_GRP vs. DYSC_GRP shows a similar pattern, except for the lower baseline MI.

Appendix E.4. Spatial Maps

Spatial maps for a given experimental condition were obtained using a heuristic procedure. First, for a given condition $\hat{\mathbf{s}}$, the mean and covariance $\mu_{\hat{\mathbf{s}}}, \Sigma_{\hat{\mathbf{s}}}$ are computed as:

$$\begin{aligned} \mu_{\hat{\mathbf{s}}} &= \sum_{k=1}^{K} \Pr[\hat{X} = k]\mu_k \\ \Sigma_{\hat{\mathbf{s}}} &= \sum_{k=1}^{K} \Pr[\hat{X} = k]\Sigma_k \end{aligned} \tag{E.4}$$

where $\Pr[\hat{X} = k]$ is the probability of state k for condition $\hat{\mathbf{s}}$, as given by $f_X(\mathbf{s}_t)$.

Voxel-wise mean and variance spatial maps are then reconstructed from the feature-space Ψ, by projecting $\mu_{\hat{\mathbf{s}}}$ and $\mathrm{diag}(\Sigma_{\hat{\mathbf{s}}})$ on Ψ^{-1}, and a spatial t–score map is created by dividing the mean map by the variance map. Fig. E.6 shows the group-wise t–score maps for $0\mathrm{s} \leq \mathrm{Ph} < 2.4\mathrm{s}$ and $2.4\mathrm{s} \leq \mathrm{Ph} < 4.8\mathrm{s}$ of a multiplication trial.

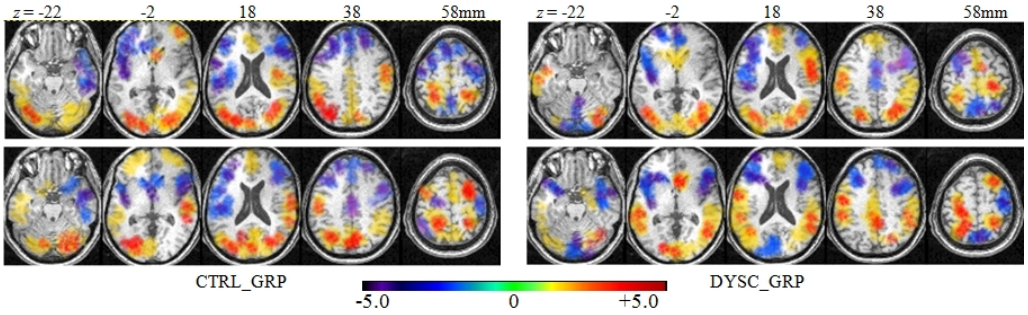

Figure E.6: Axial slices of the t–score maps for CTRL_GRP (left column) and DYSC_GRP (right column). The first row corresponds to 0s $\leq$ Ph $<$ 2.4s while the second row corresponds to 2.4s $\leq$ Ph $<$ 4.8s. Values between ± 1.5 have been masked out for clarity.

Although these maps describe spatially distributed patterns and activation foci per se, we observe numerous regions of high magnitude in these maps that corresponded to brain regions known to be activated for this task[18, 28]. Specifically, in the control group, for 0s$\leq$ Ph $\leq$2.4s, high values are seen in the bilateral occipital extra-striate cortices, the left postcentral area and in the medial frontal gyri, while for 2.4s$<$ Ph $\leq$4.8s, the maps concentrate on both pallida, caudate heads (CdH), left anterior insula (aIn), median frontal gyrus (MFG), the supplementary motor area (SMA) and the left frontoparietal operculum. The left intraparietal sulcus (IPS) shows involvement throughout. In subjects with DC, for 0s$\leq$ Ph $\leq$2.4s both occipital extra-striate cortices show high values, while at 0s$\leq$ Ph $\leq$2.4s high values in the right aIn, both MFG, left IPS, both aIn, the anterior rostral cingulate zone (aRCZ), and the right supramarginal gyrus (SMG) appear.

Early on $0s \leq \mathrm{Ph} \leq 2.67s$, the LogPs effect in controls is concentrated posterior parietooccipital lobe, the bilateral occipital gyri including V1, the right IPS, and at several foci in the frontal lobes. Later on (i.e. $2.67s < \mathrm{Ph} \leq 5.34s$), emphasis shifts to the aIn, CdH, putamina, MFG, IFG and aRCZ. Subjects with DC, in contrast to controls, initially show high values in a few small areas in the left IFG, the right MFG, the left precentral gyrus, and both IPS, and later on, in the right IPS, bilaterally in the MFG, aRCZ, posterior inferior parietum, upper cerebellar lobules, CdH, and right aIn. Smaller focii are observed in both frontal lobes.

As pointed out in the main text, there are many open issues regarding the interpretation of these maps, including the correctness of the heuristic algorithm, the role of the cross-covariance terms of Σ_k, the interpretation of strongly negative values, and the statistical significance of the values. Nevertheless, the strong correspondence of the patterns in these maps with known foci of activation during this task, indicate that the model is learning the characteristic distribution of activity during a particular mental phase.

References

[1] M. Alexander, R. Baumgartner, C. Windischberger, E. Moser, and R. Somorjai. Wavelet domain de-noising of time-courses in MR image sequences. *Magnetic Resonance Imaging*, 18(9):1129–1134, Nov 2000.

[2] M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Comput.*, 15(6):1373–1396, 2003.

[3] P. J. Bickel and E. Levina. Regularized estimation of large covariance matrices. *Annals of Statistics*, 36(1):199–227, 2008.

[4] C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, 1st ed. 2006. corr. 2nd printing edition, October 2007.

[5] R. L. Buckner, J. R. Andrews-Hanna, and D. L. Schacter. The brain's default network: anatomy, function, and relevance to disease. *Annals of NY Academy of Sciences*, 1124:1–38, Mar 2008.

[6] F. Chung. *Lectures on Spectral Graph Theory*. CBMS Regional Conference Series in Mathematics. American Mathematical Society, 1997.

[7] R. R. Coifman and M. Maggioni. Diffusion wavelets. *Applied and Computational Harmonic Analysis*, 21(1):53 – 94, 2006. Diffusion Maps and Wavelets.

[8] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.

[9] R. Duan, H. Man, W. Jiang, and W.-C. Liu. Activation detection on fMRI time series using hidden Markov model. pages 510 –513, 16-19 2005.

[10] A. Eklund, H. Ohlsson, M. Andersson, J. Rydell, A. Ynnerman, and H. Knutsson. Using real-time fMRI to control a dynamical system by brain activity classification. *Med Image Comput Comput Assist Interv*, 12(Pt 1):1000–1008, 2009.

[11] S. Faisan, L. Thoraval, J.-P. Armspach, and F. Heitz. Hidden Markov multiple event sequence models: A paradigm for the spatio-temporal analysis of fMRI data. *Med Image Anal*, 11(1):1–20, Feb 2007.

[12] K. Friston, A. Holmes, K. Worsley, J. Poline, C. Frith, and R. Frackowiak. Statistical parametric maps in functional imaging: A general linear approach. *Human Brain Mapping*, 2(4):189–210, 1995.

[13] J. V. Haxby, M. I. Gobbini, M. L. Furey, A. Ishai, J. L. Schouten, and P. Pietrini. Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, 293(5539):2425–2430, Sep 2001.

[14] J.-D. Haynes and G. Rees. Decoding mental states from brain activity in humans. *Nature Reviews: Neuroscience*, 7(7):523–534, Jul 2006.

[15] P. Højen-Sørensen, L. K. Hansen, and C. E. Rasmussen. Bayesian modelling of fMRI time series. pages 754–760, 2000.

[16] R. A. Hutchinson, R. S. Niculescu, T. A. Keller, I. Rustandi, and T. M. Mitchell. Modeling fMRI data generated by overlapping cognitive processes with unknown onsets using Hidden Process Models. *Neuroimage*, 46(1):87 – 104, 2009.

[17] Y. Kamitani and F. Tong. Decoding the visual and subjective contents of the human brain. *Nature Neuroscience*, 8(5):679–685, May 2005.

[18] K. Kucian, T. Loenneker, T. Dietrich, M. Dosch, E. Martin, and M. von Aster. Impaired neural networks for approximate calculation in dyscalculic children: a functional mri study. *Behavioral and Brain Functions*, 2:31, 2006.

[19] S. M. LaConte, S. J. Peltier, and X. P. Hu. Real-time fMRI using brain-state classification. *Human Brain Mapping*, 28(10):1033–1044, Oct 2007.

[20] A. D. Lanterman. Schwarz, Wallace, and Rissanen: Intertwining themes in theories of model selection. *International Statistical Review / Revue Internationale de Statistique*, 69(2):185–212, 2001.

[21] O. Ledoit and M. Wolf. Honey, I shrunk the sample covariance matrix. *The Journal of Portfolio Management*, 30(4):110–119, 2004.

[22] K. Li, L. Guo, J. Nie, G. Li, and T. Liu. Review of methods for functional brain connectivity detection using fMRI. *Journal of Computerized Medical Imaging and Graphics*, 33(2):131 – 139, 2009.

[23] R. Liao, J. L. Krolik, and M. J. McKeown. An information-theoretic criterion for intrasubject alignment of fMRI time series: motion corrected independent component analysis. *Medical Imaging, IEEE Transactions on*, 24(1):29–44, Jan 2005.

[24] F. D. Martino, G. Valente, N. Staeren, J. Ashburner, R. Goebel, and E. Formisano. Combining multivariate voxel selection and support vector machines for mapping and classification of fMRI spatial patterns. *Neuroimage*, 43(1):44–58, Oct 2008.

[25] M. J. McKeown, S. Makeig, G. G. Brown, T. P. Jung, S. S. Kindermann, A. J. Bell, and T. J. Sejnowski. Analysis of fMRI data by blind separation into independent spatial components. *Human Brain Mapping*, 6(3):160–188, 1998.

[26] T. M. Mitchell, R. Hutchinson, R. S. Niculescu, F. Pereira, X. Wang, M. Just, and S. Newman. Learning to decode cognitive states from brain images. *Machine Learning*, 57(1-2):145–175, 2004.

[27] T. M. Mitchell, S. V. Shinkareva, A. Carlson, K.-M. Chang, V. L. Malave, R. A. Mason, and M. A. Just. Predicting Human Brain Activity Associated with the Meanings of Nouns. *Science*, 320(5880):1191–1195, 2008.

[28] I. Morocz, A. Gross-Tsur, M. von Aster, O. Manor, Z. Breznitz, A. Karni, and R. Shalev. Functional magnetic resonance imaging in dyscalculia: preliminary observations. *Annals of Neurology*, 54(S7):S145,., 2003.

[29] R. A. Poldrack. The role of fMRI in cognitive neuroscience: where do we stand? *Current Opinion in Neurobiology*, 18(2):223–227, Apr 2008.

[30] S. M. Polyn, V. S. Natu, J. D. Cohen, and K. A. Norman. Category-specific cortical activity precedes retrieval during memory search. *Science*, 310(5756):1963–1966, Dec 2005.

[31] A. Rothman, E. Levina, and J. Zhu. Generalized thresholding of large covariance matrices. *Journal of the American Statistical Association (Theory and Methods)*, 104:177–186, 2009.

[32] S. Rotzer, T. Loenneker, K. Kucian, E. Martin, P. Klaver, and M. von Aster. Dysfunctional neural network of spatial working memory contributes to developmental dyscalculia. *Neuropsychologia*, 47(13):2859–2865, Nov 2009.

[33] B. Scholkopf and A. J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond (Adaptive Computation and Machine Learning)*. The MIT Press, 1st edition, December 2001.

[34] S. L. Scott. Bayesian methods for hidden Markov models: Recursive computing in the 21st century. *Journal of the American Statistical Association*, 97(457):337–351, 2002.

[35] R. S. Shalev. Developmental dyscalculia. *Child Neurol*, 19(10):765–771, Oct 2004.

[36] J. Shi and J. Malik. Normalized cuts and image segmentation. page 731, 1997.

[37] S. Shirdhonkar and D. Jacobs. Approximate earth mover's distance in linear time. In *Computer Vision and Pattern Recognition, IEEE Conference on*, pages 1 –8, 23-28 2008.

[38] K. E. Sip, A. Roepstorff, W. McGregor, and C. D. Frith. Detecting deception: the scope and limits. *Trends Cogn Sci*, 12(2):48–53, Feb 2008.

[39] B. Thirion and O. Faugeras. Dynamical components analysis of fMRI data through kernel pca. *Neuroimage*, 20(1):34–49, Sep 2003.

[40] P. van Gelderen, C. W. H. Wu, J. A. de Zwart, L. Cohen, M. Hallett, and J. H. Duyn. Resolution and reproducibility of BOLD and perfusion functional MRI at 3.0 Tesla. *Magnetic Resonance in Medicine*, 54(3):569–576, Sep 2005.

[41] L. Zhang, D. Samaras, N. Alia-Klein, N. Volkow, and R. Goldstein. Modeling neuronal interactivity using Dynamic Bayesian Networks. In Y. Weiss, B. Schölkopf, and J. Platt, editors, *Advances in Neural Information Processing Systems 18*, pages 1593–1600, Cambridge, MA, 2006. MIT Press.