Spatio-temporal models of mental processes from fMRI

Firdaus Janoos a,b,c,*, Raghu Machiraju a, Shantanu Singh a, Istvan Ákos 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

ARTICLE INFO
Article history:
Received 1 October 2010
Revised 15 March 2011
Accepted 17 March 2011
Available online 24 March 2011

Keywords:
fMRI
Spatio-temporal
Multivariate
Markov
Brain-state
Semi-supervised
Mean field

ABSTRACT
Understanding the highly complex, spatially distributed and temporally organized phenomena entailed by mental processes using functional MRI is an important research problem in cognitive and clinical neuroscience. Conventional analysis methods focus on the spatial dimension of the data discarding the information about brain function contained in the temporal dimension. This paper presents a fully spatio-temporal multivariate analysis method using a state-space model (SSM) for brain function that yields not only spatial maps of activity but also its temporal structure along with spatially varying estimates of the hemodynamic response. Efficient algorithms for estimating the parameters along with quantitative validations are given. A novel low-dimensional feature-space for representing the data, based on a formal definition of functional similarity, is derived. Quantitative validation of the model and the estimation algorithms is provided with a simulation study. Using a real fMRI study for mental arithmetic, the ability of this neurophysiologically inspired model to represent the spatio-temporal information corresponding to mental processes is demonstrated. Moreover, by comparing the models across multiple subjects, natural patterns in mental processes organized according to different mental abilities are revealed.

In contrast to localization based methods in fMRI, Lehmann et al. (1998) demonstrated the presence of intrinsic microstates in EEG recordings corresponding to characteristic distributions of electric activity in the brain, each ranging from 70 ms to 150 ms. Hypothesized to reflect the activation of different neuro-cognitive networks, these may be the “atoms of thought” that constitute the seemingly continual “stream of consciousness”.

Inspired by these findings, we demonstrated the presence of characteristic distributions of neural activity in fMRI data (Janoos et al., 2010a) by defining a metric for the functional distance between two activity patterns and then clustering functionally similar fMRI volumes. Identified intrinsically without reference to experimental variables, these spatially distributed and temporally varying signatures were shown to correspond to relatively longer lasting and more high-level internal mental states of the subject, such as visual perception, motor planning, conflict-resolution, arithmetical processing, etc. However, one of the drawbacks of this purely unsupervised methods used there was the problem of selecting spatio-temporal patterns related to the mental task. In addition to task-related mental processes, fMRI data contain traces of background (i.e. default-state) mental processes along with confounds such as respiration, heart-beat, head-motion and scanner drift (Logothetis, 2008). Identifying task-related patterns amongst this multitude involves determining the correct solution in a non-convex optimization landscape with multiple local minima.

In this paper, we present a fully multivariate spatio-temporal model that represents the brain transitioning through an abstract state-space as it performs mental task. This representation delineates not only the
spatial distribution of activity but also its temporal ordering. Here, information about the experimental task is used in a semi-supervised fashion to stabilize estimation and select a model of interest to the investigator, without precluding discovery of new and un-modeled patterns. Importantly, the use of a neurophysiologically inspired model allows comparison of the spatio-temporal patterns of mental processes of subjects, in their entirety.

**Road map**

The state-space model (SSM), represented by a first order Markov chain (Bishop, 2007), assumes the presence of a set of abstract mental states that are revisited during the performance of a mental task. Given the goal-oriented and directed nature of human thought these brain-states not only exhibit a temporal ordering but also respond to external stimuli.

Each state is associated with a characteristic spatial distribution of neural/metabolic activity and an occurrence of the state corresponds with an activation pattern based on this signature. The observed fMRI data arise from an unknown and spatially varying hemodynamic response to these activation patterns. The model is elaborated further in the State-Space Model (SSM) section. The Markov chain of brain-states serves two purposes: a) To enforce a temporal ordering on the states, and b) To decouple the stimulus from the fMRI signal, thereby obviating specification of the exact mathematical relationship between the two. This second aspect makes it a semi-supervised method as it uses the stimuli to guide estimation but does not preclude discovery of new patterns in the data and investigation of effects not explicitly encoded in the experimental variables.

The SSM can predict the value of experimental stimuli at new frames and is able to estimate a spatially varying hemodynamic response function (HRF) from the data.

In the Feature-space section, a novel linear low-dimensional feature-space is derived from the functional distance (Janoos et al., 2010a) between the activation patterns present in the fMRI data at two time-points. This functional distance is measured by the amount of change between their activity distributions over the functional networks of the brain. The feature-space, derived from this measure of similarity between brain-states, allows an exploration of hidden patterns in the data.

Then in the Estimation section, an efficient algorithm based on a mean field approximation of expectation maximization (EM) (Bishop, 2007) is presented to estimate the model parameters, the activation maps, the hemodynamic filter and the optimal state sequence, along with predictions of unobserved stimuli. A method to determine the correct model size and other hyper-parameters in an automated fashion to stabilize estimation and select a model of interest to the investigator, without precluding discovery of new and un-modeled patterns is its ability to integrate information in groups of voxels that individually are weakly activated, but jointly may be highly structured.

The advantages of using the state-space model representation to compare the mental processes of different subjects in their entirety is also demonstrated. Finally, we conclude with some remarks and observations on the method in the Conclusion section.

**Contribution**

In contrast to other methods for studying the mental state of the subject from fMRI data (cf. Related work section), this paper present a novel data-driven multivariate method for the dynamical analysis of mental processes in a time-resolved fashion using a phenomenological model for brain function that: (a) reveals not only the spatial distribution but also the temporal ordering of activity during a task; (b) can be applied to arbitrarily complex paradigms, not just experiments with fixed alternatives; (c) does not require specification of the mathematical relationship between experimental variables and fMRI signal; (d) does not assume a known and spatially fixed hemodynamic response; (e) uses experimental information to guide estimation towards of patterns of interest, unlike unsupervised methods; (f) permits testing for the effects of experimental variables against which the model was not trained, unlike supervised methods; and (g) allows neurophysiological interpretation of the parameters and comparison of the spatio-temporal patterns between subjects in their entirety.

**Related work**

**Multivariate methods**

Multivariate methods for fMRI analysis have made major contributions in studying the representation of information in the brain i.e. how distributed neuronal responses encode the sensorial or cognitive state of the subject. The main advantage of the multivariate approach is its ability to integrate information in groups of voxels that individually are weakly activated, but jointly may be highly structured.
with respect to the task. Communication among neurons as well as larger functional units is the main basis of neural computation, and by not disregarding their interactions, multivariate methods are able to reveal more about the "neural code" (O'Toole et al., 2007).

Supervised or confirmatory approaches such as multivariate linear models (MVLM) (Friston et al., 2008) and multivariate pattern recognition (MVPR) (Haynes and Rees, 2006) learn the mathematical relationship from the distributed patterns of activation contained in the data to the experimental variables. Multivariate unsupervised or exploratory approaches such as principal components analysis (PCA) (Multiple, 2007), independent components analysis (ICA) (Calhoun and Adali, 2006) and cluster analysis (Bauermann et al., 2000) have also been widely applied to the analysis of fMRI, especially resting-state and non task-related data. Semi-supervised multivariate analysis come in two flavors: (a) where unlabeled data (e.g. resting-state data) are used to regularize the regression of data with labels (e.g. task-related data) (Blaschko et al., 2009) or (b) where information about the task is used to guide a clustering process (Friman et al., 2003) or a matrix-based decomposition of the data (e.g. partial least squares (PLS) (Mcintosh and Lobaugh, 2004), functional PCA (Ghebreab and Smeulders, 2010) or constrained ICA (Lin et al., 2010)).

Despite the success of these multivariate methods, there are nevertheless many open challenges. Because supervised methods learn a fixed mapping from fMRI data to regressors/labels describing stimuli or subject behavior, their ability to explain the cognitive state of the subject is limited to behavioral correlates and they cannot discover intrinsic patterns that might be present in the data.

MVLM methods such as canonical correlation analysis (CCA) (Multiple, 2007) and multivariate Bayesian decoding (MVB) (Friston et al., 2008) model a linear relationship between regressors, formed by convolving the stimuli with a hemodynamic response function (HRF), and the observed data. This requires that the mathematical relationship between experimental variables and fMRI signal be known a priori, which may be hard to define, especially in experiments for higher level cognition. Equally problematic is the assumption of spatially and temporally constant hemodynamics in these models. Multiple studies (Logothetis, 2008) have shown a large variation in the HRF across subjects, across brain sites within the same subject, and even at the same brain site of the same subject across time. Furthermore, these methods are extremely sensitive to the design of the experiment in terms of event-timing, orthogonality, estimability and the layout of the design matrix.

On the other hand, MVPR methods such as linear discriminant analysis (LDA), Gaussian naive Bayes (GBN), neural-networks, support vector machines (SVMs) and other types of classifiers (Haxby et al., 2001; Haynes and Rees, 2006; O'Toole et al., 2007; Mitchell et al., 2008) do not specify a neurophysiological model and instead treat the data as an abstract representation of mental activity. Therefore, neuro-scientific interpretation of results and comparison across subjects becomes difficult. A significant limitation of MVPR classifiers is their applicability only to studies where subjects are presented with a fixed number of alternatives (e.g. faces vs. objects (Haxby et al., 2001)). Also, most methods make the assumption that all fMRI scans with the same label (i.e. behavioral state) are equivalent, thereby neglecting the temporal structure in mental processes. By using all the scans in one block as a feature vector through temporal embedding (Mourão-Miranda et al., 2007), the temporal structure of only block design experiments with fixed length blocks has been incorporated in a MVPR framework. Generalization to complex cognitive paradigms with interval-valued parameters, event related designs and further on to real world situations poses a fundamental challenge for MVPR (Haynes and Rees, 2006).

Similarly, many unsupervised and semi-supervised approaches also suffer from the problem of interpretability due to lack of a neurophysiological model. More importantly, purely unsupervised methods are blind to the structure and parameters of the experiment and fail to provide quantifiable links to experimental variables (O’Toole et al., 2007).

The reader is referred to the excellent article by Friston et al. (2008) on MVLMs and the reviews by Haynes and Rees (2006) and O’Toole et al. (2007) on MVPR for a better understanding of these various alternatives and their trade-offs.

State-space methods

Hidden Markov models (HMM) have been previously used in fMRI for determining the activation state of individual voxels in a univariate fashion from their time-series data (Hajen-Sorensen et al., 2000). Activation detection associated with known stimuli has also been done with hidden Markov multiple event sequence models (HMMESM) (Faisal et al., 2007), that pre-process the raw time-series into a series of spikes to infer neural events at each voxel. A hidden process model (HPM) (Hutchinson et al., 2009) was used for univariate testing of each voxel’s time-series data for the occurrence of small set of pre-specified “neural processes” relative to some external event. Bayesian spatio-temporal models that estimate the activation state of voxels using Markov random fields as a spatial prior have also been proposed (Quiros et al., 2010). Dynamic Bayesian networks (Zhang et al., 2006) and dynamic causal models (Stephan and Friston, 2010) have been applied to study the time-varying functional integration of a few preselected functional modules, from the interdependency structure of their average time-series.

The state-space model (SSM)

The concept of the functional brain transition through a mental state-space is depicted in Fig. 2. The probability \( P(x_t = k) \) that the brain-state \( x_t \) at time \( t = 1 \ldots T \) (in TR units) is \( k = 1 \ldots K \) depends on not only on the previous state of the brain but also on the current experimental stimulus described by the vector \( s_t \). The multinomial transition probability from \( x_{t-1} = 1 \) to \( x_t = j \) is (Bishop, 2007):

\[
\pi_{x_{t-1}, x_t}(s_t) \delta_p(x_t = j | x_{t-1} = 1, s_t, w) = \frac{\exp \left( \omega_j + w_{ij} \right)}{\sum_k^{K} \exp \left( \omega_k + w_{jk} \right)}
\]

The \( \delta \) symbol indicates matrix transpose throughout the paper. The probability of being in state \( j \) at any instant is parameterized by the vector \( \omega_j \). The probability of transitioning from state \( i \) at time \( t-1 \) to state \( j \) at time \( t \) is parameterized by \( w_{ij} \) which has a normal prior \( \mathcal{N}(0, \lambda_{\omega}) \) with precision hyper-parameter \( \lambda_{\omega} \). All these transitions are driven by the stimulus vector \( s_t \), introducing an additional element in the stimulus vector \( s_t \) set to 1 allows modifying the transition probability to include a term independent of the current stimulus. Though the experiment maybe have combination of interval and categorical valued stimuli, they are converted into standardized normal variables \( s_t \) through a probit transformation of their cumulative distribution functions. The hyper-parameter \( \lambda_{\omega} \) controls the trade-off between the influence of the current stimulus \( s_t \) and the previous state \( x_{t-1} \) on the probability of the current state \( x_t \). A low value biases the estimates of \( w_{ij} \) towards its mean value \( \omega_j \) reducing the influence of the previous state \( x_{t-1} = 1 \) on \( p(X_t = j | X_{t-1} = 1) \) and increasing the influence of the \( s_t \) on the transition.

The SSM allows estimating the value of unobserved or missing stimuli at a subset of the time-points \( U \neq \{1 \ldots T\} \), represented by the hidden variables \( u_t, t \in U \). This feature enables prediction of stimuli from data at these time-points \( t \in U \).

Each abstract brain-state is realized by a characteristic distribution of neural (metabolic) activity in the cerebral cortex, denoted by \( Z_t \) in voxel-space and by \( z_t = \Phi[Z_t] \) in the D-dimensional \( (D \ll N, \text{the number of voxels}) \) feature-space \( \Phi \) (cf. Feature-space section). If \( x_t = k, k = 1 \ldots K \), then \( z_t \) is modeled \( \mathcal{N}(\mu_k, \Sigma_k) \) in feature-space.
impulse response associated with an $z$ variable. The full probability model is (cf. Fig. 2):

Here, $H$ is the HRF, modeled as a finite impulse response (FIR) filter $h[d] = (h_{1[d]} ... h_{l[d]})$ of length $L + 1$. Each $h[d]$ has a normal prior $\mathcal{N}(\mu_h, \Sigma_h)$ constructed by varying the delay, dispersion and onset parameters of the canonical HRF of SPM8. The HRF $h$ in voxel-space and $y[t]$ is feature-space. For example, it does not provide a closed-form expression for the state-space model (SSM). The experimental parameters are represented by $\theta = (\beta, \Sigma_u, \mu_u, \Sigma_y)$. Each element $d = 1...D$ of the $D$-dimensional feature-space is associated with an unknown HRF, modeled as a finite impulse response (FIR) filter $h[d] = (h_{1[d]} ... h_{l[d]})$ of length $L + 1$. Each $h[d]$ has a normal prior $\mathcal{N}(\mu_h, \Sigma_h)$ constructed by varying the delay, dispersion and onset parameters of the canonical HRF of SPM8 (Multiple, 2007) and computing their mean and variance. The length $L + 1$ is typically set to 32 s. The set of HRF parameters is then the $D \times L$ matrix $h = (h_{1[1]} ... h_{l[D]})^\top$. The fMRI data $Y_t$ in voxel-space and $y[t]$ is feature-space is generated as $y[t] = \Phi(Y_t)$ in feature-space arises through an element-wise convolution $y[t] = \sum_l H_{x, l} + \zeta[t]$ of the metabolic activity $z[t]$ with the HRF $h$. Here, $H_t = \text{diag}(h_{1[1]} ... h_{l[D]})$ and $\zeta[t] \sim \mathcal{N}(0, \Sigma_u)$ is temporally i.i.d. noise. The HRF $L + 1$ TRs long induces a correlation in the scans $y[t]$ for each $t$. The instantaneous activation pattern is observed at time $t$. Note that the convolution is actually in voxel-space, but since the feature-space transform is commutative with convolution, it can be performed directly in feature-space.

Therefore, denoting the set of parameters $\theta = (\beta, \Sigma_u, \mu_u)$, the full probability model is (cf. Fig. 2):

$$p(\theta, y, z, x) \propto p(y \mid h, z, x) p(z \mid x, \theta) p(x \mid s, u, w)$$ \hspace{1cm} (2)

where

$$p(x \mid s, u, w) = p(w) \prod_{t \in [T]} \prod_{v \in \Omega} \pi_{x_{v-1}, x_{v}}(s_{v}) \prod_{i \in \Omega} \pi_{x_{v-1}, x_{v}}(u_{i})$$

$$p(z \mid x, \theta) = \prod_{t = 1}^{T} p(z_{t} \mid x_{t}, \theta_{x_{t}})$$

and

$$p(y \mid h, z, x) = p(h) \prod_{t = 1}^{T} p(y_{t} \mid z_{t-1:T}, h, x_{t})$$

The SSM hyperparameters $K$ and $\lambda_w$ are selected using an automatic data-driven procedure described in the SSM hyper-parameter selection section.

**Feature-space**

The SSM effectively provides a temporally coherent formalism for clustering the fMRI scans based on their patterns of spatially distributed metabolic activity. Like any clustering algorithm, its efficacy depends on the metric quantifying the difference/similarity between the activity patterns at two time-points (Jain, 2010). In the Functional distance section, we describe a novel functional distance based on the concept of the cost of minimum “transport” of activity between two time-points measured over the functional networks of the brain. Functional networks are routinely defined by the temporal correlations between the fMRI time-series of the voxels (Li et al., 2009). Technical ?? presents an algorithm of computing the functional connectivity (i.e. correlations) $F[i, j] \in [-1, 1]$ between all pairs of voxels $i, j$ that is consistent, sparse and computationally efficient, although any equivalent method can be used. In previous work (Janoos et al., 2010a), we had demonstrated the efficacy of this method to determine functionally meaningful clusters in the space of activation patterns.

Because this metric, related to the earth mover's distance (EMD) (Shirdhonkar and Jacobs, 2008) between distributions, is posed as an optimization problem, it does not have a well-understood metric structure. For example, it does not provide a closed-form expression for the mean or variance of a cluster. As a result, determining the statistical properties of clusters obtained under this metric is not straightforward. A strategy for circumventing this problem is to embed the fMRI scans in an Euclidean space obtained from the pair-wise distance matrix through an embedding such as the graph Laplacian (Chung, 1997), and performing analysis in this space (Janoos et al., 2010a). This, however, requires an estimation of the functional distance between all pairs of volumes in an fMRI session, which with a worst-case running time of $O(N^2 \log N)$ per pair (Shirdhonkar and Jacobs, 2008), is prohibitively expensive. Therefore, in the Orthogonal basis construction section we adopt an alternative strategy, wherein each fMRI scan is embedded in a linear space obtained by an orthogonalization of the matrix of voxel-wise functional connectivities. Similar to the approximation of EMD by a wavelet transform (Shirdhonkar and Jacobs, 2008), the Euclidean metric in this feature-space provides a good approximation of the functional distance.

Since the number of voxels and therefore the number of linear basis vectors $N \sim O(10^6)$ is orders of magnitude larger than the number of scans $T \sim O(10^5)$, dimensionality reduction is required to prevent over-fitting. Using a bootstrap analysis of stability (Bellec et al., 2010), basis vectors that are not "stable" across resamples of the data are discarded, causing a significant reduction in dimensionality. This procedure is laid out in the Feature-selection section.

Other feature-spaces in fMRI include purely unsupervised ones like PCA (Multiple, 2007), ICA (Calhoun et al., 2009) and functional parcellations of the cortex (Thirion et al., 2006), or supervised ones like PLS and regions-of-interest (ROIs) identified either by a univariate GLM analysis or through manual demarcation (Haynes and Rees, 2006). Dimensionality reduction is achieved either in an unsupervised fashion, for example by selecting principle components (PCs) that explain a certain percentage of variance or in a supervised fashion by selecting features correlated with or predictive of the experimental variables (Mitchell et al., 2008). For the aims of the methodology presented here, supervised feature-spaces are unsuitable as they are inherently biased towards the experimental variables against which they were selected and may not capture other patterns in the data. At the other extreme, unsupervised feature-spaces do not necessarily describe patterns of interest. For example, in our data-sets, we observed that the largest variance principal components corresponded to motion and physiological noise such as respiration and pulsatile activity. In contrast to these, the feature-space developed here is motivated by an intuitive definition of the functional distance between activation patterns and feature selection is performed using a criterion of stability. A
quantitative evaluation of these various alternatives is reported in the fMRI data-set: mental arithmetic task section.

**Functional distance**

The functional distance $\text{FD}(Z_t, Z_o)$ between the activation patterns $Z_t$ and $Z_o$ (in voxel-space) at two time-points is quantified by the *transportation distance*, i.e. the minimal "transport" $f : N \times N \rightarrow \mathbb{R}$ of activity over the functional circuits to convert $Z_t$ into $Z_o$ (Janoos et al., 2010a). Specifically,

$$\text{FD}(Z_t, Z_o) = \min \sum_{i,j} F[i,j] d_t[i,j],$$

subject to the constraints: $f[i,j] \geq 0$, $\sum_i f[i,j] \leq Z_t[i]$, $\sum_j f[i,j] \leq Z_o[j]$, and $\sum_i \sum_j f[i,j] = \sum_i Z_t[i]$, $\sum_j Z_o[j]$. The cost of the transport of $f[i,j]$ from voxel $i$ to $j$ will depend on a measure of the "disconnectivity" $d_t = N \times N \rightarrow \mathbb{R}^+$ between the voxels complementary to $F$, defined in the *Cost metric* section.

This definition captures the intuitive notion that two activity patterns are functionally more similar if the differences between them are mainly in voxels that are functionally related to each other indicating the activation of a shared functional network.

**Cost metric**

Treating $F$ as the adjacency matrix of a graph with $N$ vertices, a cost metric $d_t[i,j] = |\phi^*[i] - \phi^*[j]|$ is induced via a distortion minimizing embedding $\phi^*: N \rightarrow \mathbb{R}$ of the graph (Chung, 1997):

$$\phi^* = \operatorname{arg\,inf}_{\phi, D_t} \sum_{i,j} \left| \phi^*[i] - \phi^*[j] \right|^2 F[i,j] / \sum_{i,j} F[i,j] D_t[i,j]$$

where $D_t$ is the diagonal degree matrix of the adjacency matrix $F$.

Here, the embedding $\phi^*$ will take similar values at voxels that have high functional connectivity and the functional distance between them is $d_t[i,j] = |\phi^*[i] - \phi^*[j]|$. The constraint $\phi^* D_t = 1$ is to prevent $\phi^*$ from taking a value at each vertex proportional to its degree, which is the trivial minimizer of Eq. (4). It can be shown (Chung, 1997) that $\phi^*$ is the solution to the generalized eigenvalue problem $(D_t - \lambda F) \phi^* = \Lambda F \phi^*$ subject to the constraint $\phi^* D_t = 1 = 0$. If $\eta_1$ is the eigenvector $\eta_1 = \eta_1 C_0 \eta_1$ of the normalized graph Laplacian $L = D_t^{-1/2}(F - D_t^{-1})$ corresponding to the second smallest eigenvalue $\lambda_1 > 0$, then $\phi^* = D_t^{-1/2} \eta_1$.

**Orthogonal basis construction**

Through a recursive partitioning of the voxel-grid based on its embedding $\phi^*$, we construct an orthogonal basis $\Phi = \{\phi^{(m)}: \mathbb{R} \rightarrow \mathbb{R}\}$ where the index $m = 0, \ldots, \log_2 N - 1$ gives the level of decomposition, while $l = 0, \ldots, 2^m - 1$ indexes the basis vectors at level $m$. The first basis vector $\phi^{(0)} = D_t^{-1/2} \eta_1$, and their graph Laplacians $L^{(1)}$ and $L^{(2)}$ are computed. The next two basis vectors $\phi^{(1)}$ and $\phi^{(2)}$ are the second smallest eigenvectors of the $L^{(1)}$ and $L^{(2)}$, respectively. The process is repeated until only one voxel is left in the partition. The details of this algorithm are provided in Technical Appendix.

Let the projection of $Z_t$ on this orthogonal basis be denoted as $\tilde{Z}_t[i,m] = m = 0, \log_2 N - 1, l = 0, 2^m - 1$, where $Z_t[i,m] \leq 2^{-m} Z_t$ and $\phi^{(m)}$. The following theorem asserts that the representation in this basis $\Phi$ of the difference $\hat{Z} = Z_t - Z_o$ between the activity patterns at two time-points $t_1$ and $t_2$ provides a tight bound on the functional distance.

**Theorem 1.** Let $\hat{Z}[i,m]$ be coefficients of $\hat{Z} = Z_t - Z_o$ in the basis $\Phi$. Then, there exist constants $M_0 > \hat{M}_0 > 0$, such that

$$\hat{M}_0 \sum_m \sum_i |\hat{Z}[i,m]| \leq \text{FD}(Z_t, Z_o) \leq M_0 \sum_m \sum_i |\hat{Z}[i,m]|$$

and the tightness of this bound is:

$$\sup_{\|\hat{Z}\|_2 = 1} \left[ M_0 \sum_m \sum_i |\hat{Z}[i,m]| - \hat{M}_0 \sum_m \sum_i |\hat{Z}[i,m]| \right] \approx \frac{(M_0 - \hat{M}_0)}{\sqrt{2}}$$

Therefore, the functional distance $\text{FD}(Z_t, Z_o)$ is approximated (up to a multiplicative constant) by the $\ell_2$ distance metric in this space.

$$\hat{d}(Z_t, Z_o) = \left( \sum_m \left( \sum_i |Z_t[i,m] - Z_o[i,m]|^2 \right) \right)^{1/2}.$$  

The quality of the approximation $M_0 - \hat{M}_0$ is examined at the end of the next section.

**Feature-selection**

Basis vectors for the feature-space are selected using a bootstrap analysis of stability (Belloc et al., 2010). The bootstrap generates a non-parametric estimate of the sampling distribution of a statistic (i.e. bootstrap distribution) from a single sample of the data by creating multiple surrogate samples of same size as the original sample, by resampling with replacement from the original sample. Bootstrap estimates of the functional connectivity matrix $F$ are obtained by resampling entire blocks of fMRI scans from a session $Y = \{Y_1, \ldots, Y_T\}$ to create a surrogate session. The presence of serial correlations in the time-series data necessitates a block bootstrap method, wherein the $T$ scans are divided into $M$-blocks and a resample is created by randomly selecting $M$-blocks from this set, with replacement. Although the block length $T/M$ needs to be adapted to the range of temporal dependencies present in the data, the correlation structure in the data is faithfully reproduced over a fairly wide range of lengths (Belloc et al., 2010). We found $T/M \approx 5T/R$ to be adequate for our data-sets.

The stability of a particular basis vector $\phi^{(m)}$ is defined by its correlation coefficient $\rho^{(m)}(r_1, r_2) = \rho(\phi^{(m)}_1, \phi^{(m)}_2)$ across resamples $r_1, r_2$ of the data. Given the bootstrap distribution of correlations $Pr_{\text{boot}}[\rho^{(m)}(r_1, r_2)]$, a vector $\phi^{(m)}$ is said to be $\tau_{\rho}$-stable if $Pr_{\text{boot}}[\rho^{(m)}(r_1, r_2) \geq \tau_{\rho}] \geq 0.75$, i.e. the correlation between at least 75% of the resamples of $\phi^{(m)}$ is greater than the threshold $0 \leq \tau_{\rho} \leq 1$. If $\rho^{(m)}$ is not $\tau_{\rho}$-stable, then it is discarded, which also removes all the vectors obtained from the subdivision of $\phi^{(m)}$. Therefore, increasing the value of $\tau_{\rho}$ causes an exponential increase in the number of vectors that are removed.

The effect of $\tau_{\rho}$ on the dimensionality $D$ is shown in Fig. 3. Initially there is a steep reduction in dimensionality from $O(10^2)$ to $O(10^1)$. However, after a certain value of $\tau_{\rho}$, the reduction slows down significantly. This knee-point usually occurred at $D = 500$ corresponding to $\tau_{\rho} = 0.4-0.5$ in our data-sets. Therefore, $\tau_{\rho}$ was adaptively set for each fMRI session such that $D = 500$. The figure also shows the largest index $m$ of the vectors $\phi^{(m)}$ retained for a given $\tau_{\rho}$, indicating that the stability of basis vectors reduces as the level of decomposition $m$ increases. This observation, along with the $2^{-m}$ decay of the coefficients in Eq. (5), implies that the effect of the reduced dimensionality of $\Phi$ on the approximation error is small, as most of the discarded vectors have a large index $m$. 

---

1. $D_t[i,j] = \sum_{i,j} F[i,j]$ and $D_t[i,j] = 0$, $\forall i \neq j$. 

---

F. Janoos et al. / Neuroimage 57 (2011) 362–377
A comparison of the relative logarithmic error in the approximation of $\mathcal{F}(z_n, z_t)$ using the reduced $\mathcal{P}$ versus the full basis is shown in Fig. 4. It can be observed that the linear approximation $\hat{a}(z_n, z_t)$ provided by the full basis $\mathcal{P}$ is typically within $2.5 \times \mathcal{F}$ the transportation distance $\mathcal{F}(z_n, z_t)$, while the distance in reduced dimensionality is within $3 \times \mathcal{F}(z_n, z_t)$. Setting empirical validation of Eq. (6). We see that reducing the dimensionality by an order of $O\left(10^3\right)$ increases the approximation error by less than 20%, on average.

Estimation

In this section, a generalized expectation–maximization (GEM) algorithm (Bishop, 2007) to estimate the parameters $\theta^* = \arg\max Q(\theta) / C_0 / C_1$ is presented. Introducing a variational density $q(z, x)$ over the latent variables $z, x$, the log-probability of Eq. (2) is decomposed into a free-energy and a KL-divergence as $\ln p_d(y) = Q(q, \theta) + KL(q||p_d)$, where,

$$
Q(q, \theta) = \sum_i \int q(z, x) \ln \frac{p_i(y, z, x) / C_0 / C_1}{q(z, x) / C_0 / C_1} dz \text{ and } KL(q||p_d)
$$

$$
= - \sum_i \int q(z, x) \ln \frac{p_d(z, x) / C_0 / C_1}{q(z, x) / C_0 / C_1} dz.
$$

Starting with an initial estimate $\theta^{(0)}$, the GEM algorithm finds a local maxima of $\ln p_d(y)$ by iterating the following two steps:

**E-step**

$$q^{(n)} \leftarrow \arg\min_{q} KL(q||p_d(n))$$

**M-step**

$$\theta^{(n+1)} \leftarrow \theta \text{ such that } Q\left(q^{(n)}, \theta\right) = \arg\max Q\left(q^{(n)}, \theta\right).$$

The iterations are terminated when the updates to $\theta$ fall below a pre-specified tolerance (adaptively set at 1% of the absolute value of the parameter in the $n^{th}$ iteration), yielding a locally optimal solution $\theta^*$. The expressions and detailed derivations for the E-Step estimates of all the parameters in the M-Step are provided in Technical ?? and ?? respectively.

**E-step**

Although the minimizer of $KL(q||p_d)$ is $q(z, x) = p_d(z, x)$, the HRF introduces a dependency structure between $x_t - L, \ldots, x_t + L$ and $z_{t-L} \ldots z_{t+L}$ when conditioned on the measurement $y_t$. Therefore, evaluation of $Q(p_{d, \theta}(z, x|y, \theta)$ in the M-step would require marginalization over sequences of $2L + 1$ variables, resulting in a computational complexity of $O\left(T \times K^2\right)$ for parameter estimation (as compared to $T \times K^2$ for first-order HMMs). To avoid this expensive computation, we restrict $q$ to the family of factorizable distributions $q(z, x) = \prod_{t=-L}^{+L} q_t(z_t, x_t|x_{-t}/q(x_t)$. This is known as the mean field approximation in statistical physics and it can be shown (Bishop, 2007) that if $q^n(z, x) = \prod_{t=-L}^{+L} q^n_t(z_t, x_t|x_{-t}/q^n(x_t)$, then $q^n_t(z_t, x_t) = \arg\max_{q_t} KL(q||p_d(z_t, x_t))$.

As shown in Technical ??, each factor of the mean field approximation is a product $q^n_t(z_t, x_t) = q^n_t(z_t|x_{-t}/q^n_t(x_t)$. Thus, a multimodal logistic probability $q^n(x_t)$ and a normal density $q^n_t(z_t|x_{-t}/q^n_t(x_t)$. Therefore, under this approximation, the $n^{th}$ iteration of E-step involves computing the factorizable density $q^n(z, x) = \prod_{t=-L}^{+L} q^n_t(z_t, x_t|x_{-t}/q^n_t(x_t)$.

Since the approximation is a product $q^n_t(z_t, x_t) = q^n_t(z_t|x_{-t}/q^n_t(x_t)$, and $\Theta$ is not jointly concave in $\mathbf{w}$ and $\mathbf{u}$, we decouple the problem into two concave problems, by first maximizing $\Theta$ with respect to $\mathbf{w}$ setting $\mathbf{u} = \mathbf{u}(n)$, and then maximizing with respect to $\mathbf{u}$ setting $\mathbf{w} = \mathbf{w}(n + 1)$.

Although, $\mathbf{w}$ can be estimated using the iteratively re-weighted least squares (IRLS) method, it involves an expensive inversion of the Hessian $\nabla^2 \Theta Q$ at each iteration. This inversion is avoided using a bound optimization method (Krishnapuram et al., 2005) that iteratively maximizes a surrogate function $\mathcal{L}(\mathbf{w}^{n+1}) = \arg\max_{\mathbf{w}} \mathcal{Q}'(\mathbf{w}|\mathbf{w}^{n+1})$. The index $n$ marks the iterations of the bound maximization of $\Theta$ with respect to $\mathbf{w}$, during one iteration of the M-step indexed by $n$. This inner maximization loop is initialized with $\mathbf{w}(0) \rightarrow \mathbf{w}(n)$ and terminates when the update $||\mathbf{w}(n+1) - \mathbf{w}(n)||$ falls below a certain tolerance, and $\mathbf{w}(n+1)$ is the new value for the M-step iteration. This tolerance can be fairly loose (typically 10% of the absolute value $||\mathbf{w}(n)||$), as the GEM algorithm only requires an increase in the value of $\Theta$ with respect to its parameters and not necessarily the maximization of $\Theta$. The surrogate function used here is the quadratic function with constant Hessian $\mathbf{B}$ such that $\nabla^2 \Theta Q - \mathbf{B}$ is negative-definite. Although bound optimization takes more iterations to converge than IRLS, on the whole it is much faster since it precludes inverting the Hessian (of the order of the size of $\mathbf{w}$) at each step (Krishnapuram et al., 2005). Detailed proofs for the gradient and Hessian of $\Theta$ with respect to $\mathbf{w}$ are elaborated in Technical ?? followed by the bound optimization procedure in Technical ??.

After estimating $\mathbf{w}(n+1), \mathcal{L}(\mathbf{w}(0), \theta)$ is then maximized with respect to $\mathbf{u}$, for all $t \in U$ by setting $\mathbf{w} = \mathbf{w}(n+1)$. Again, the Hessian $\nabla^2 \Theta Q$ is negative-definite and therefore $\Theta$ is concave in $\mathbf{u}$, with a unique global maximum. Since $\nabla^2 \Theta Q$ is of the dimension of the stimulus vector and is usually easily invertible, this maximization is done using IRLS because of its faster convergence. The expressions for the gradient and Hessian of $\Theta$ with respect to $\mathbf{u}$ are given in Technical ??.

The estimates of emission parameters $\phi_t = (\phi_k, \phi_z)$ in the $n^{th}$ iteration of the M-step can be computed in closed-form, as per the formulæ detailed in Technical ??, setting $\mathbf{h}[d] = (\mathbf{h}[d], \ldots, \mathbf{h}[d])$ denote the $d + 1$-tap HRF FIR filter corresponding to the $d^{th}$ element of the D-dimension feature-space. As shown in Technical ??, the gradient $\partial \Theta / \partial \mathbf{h}[d]$ for the FIR filter at the $d$-th element depends on the values of $\mathbf{h}[d]$ at all the other $d' \neq d$ of the D-dimensional space. Setting $\partial \Theta / \partial \mathbf{h}[d] = 0$, for
all $d = 1 \ldots D$, results in a linear system of $D \times L$ equations in $D \times L$ unknowns. The unique solution $h^{(n+1)}$ is computed using conjugate gradient descent (Golub and Van Loan, 1996) initialized at $h^{(0)}$, and its iterations are terminated when the update $||h^{(n+1)} - h^{(n)}||_2$ falls below a pre-specified tolerance (set adaptively at 10% of $||h^{(0)}||_2$). The closed-form estimation of the noise variance $\Sigma$ is given in Technical app.

Spatial activation maps

The activation pattern for a specific value of the experimental variables $s_i$ is obtained by first computing the invariant distribution $p(x|w, s_i)$ as the first eigenvector of the state-transition matrix $\pi(s_i)$, and then computing the mean activation pattern as $\mu_k = \sum_i \pi_k(s_i = k|w, s_i) \mu_i$ and its variance as $\Sigma_k = \sum_i \pi_k(s_i = k|w, s_i) \Sigma_i$. The z-score map for the activation pattern corresponding to $s_i$ is given by $\Sigma_i^{-1/2} \mu_i$ in feature-space, which can then be transformed back into a voxel-wise spatial map of activity.

Estimating the optimal state-sequence

Direct estimation of the most probable sequence of states $x^t = arg \max_y p(x^t|y) = \arg \max_{y} p(h, y, x^t)$, given an estimate of SSM parameters, requires joint maximization over all $T$ state variables $x_1, \ldots, x_T$ since the hidden layer $z$ introduces a dependency between all the $y$ and $x$ variables preventing factorization of the graphical model. As the size of the search space increases exponentially with $T$ with a complexity of $O(T^K)$ for the whole chain, exhaustive search soon becomes infeasible and an approximation such as iterated conditional modes (ICM) is required (Bishop, 2007).

To avoid this computational burden, we developed an EM algorithm for optimal state-sequence estimation under a mean field approximation that iteratively transforms the problem into a series of first order HMMs. Each M-step can then be computed using the Viterbi algorithm with $O(T \times K^2)$ complexity (Bishop, 2007). The EM iterations terminate when the increments $|\ln p(y, x^{(n+1)}| - \ln p(y, x^{(n)})|$ fall below a pre-specified tolerance, typically set to 0.00099 corresponding to <1% increase in the probability. Technical app. contains the details of this state-sequence estimation procedure.

SSM hyper-parameter selection

The hyper-parameters of the SSM are the number of hidden states $K$, the precision $\lambda_w$ of the prior distribution of the transition weights $w$, and the parameters $\mu_h, \Sigma_h$ of the prior model of the HRF $h$. The values of $\mu_h$ and $\Sigma_h$, determined from the canonical HRF of SPM8, are used to enforce domain knowledge by restricting the HRF to the space of physiologically plausible shapes. This provides an optimal trade-off between allowing a spatially varying and unknown HRF against over-fitting the FIR filter to the data.

The hyper-parameter $\lambda_w$ determines the variance in the weights $w$, and implements a trade-off between the effect of the stimulus versus the previous state on the current state probability and mediates a complex set of interactions between the temporal structure of the fMRI data and of the stimulus sequence. A very high value of $\lambda_w$ causes the state-transitions to be driven mostly by the current stimulus, while a low value increases the contribution of the previous state to the transition probability. It therefore cannot be practically provided as a user-tunable parameter. On the other hand, model-size (i.e. $K$) selection is typically done using Bayes factors (Kass and Raftery, 1995), information theoretic criteria (McGory and Titterington, 2009) or reversible jump MCMC based methods (Scott, 2002). Implicitly these methods require an a priori notion about the complexity of a given model.

Here instead, we adopt an automated method for selecting both $K$ and $\lambda_w$ based on a maximally predictive criterion leveraging the ability of the SSM to predict missing stimuli. From the stimulus time-series, blocks of $T$ consecutive time-points (in TR units) totalling to 25% of the total number of scans, are removed at random to serve as missing stimuli $U\{t_1, t_2, T - 1, \ldots, T - M + T - 1\}$ and the optimal SSM parameters $\theta^*$ are estimated for a given $K$ and $\lambda_w$. The prediction error is then measured as $\text{ERR}_{\text{missing}} = \sum_{t = 1}^T ||y_t - \hat{y}_t||_2$ between the predicted $\hat{y}_t$ and their true values $y_t$. The hyper-parameters are then selected to minimize this error-rate. The optimal value of $K$ is obtained by first stepping through different values of $K$ with large step-sizes and then iteratively refining the step-size. The advantage of this procedure is that it allows selecting a model most relevant to the experiment being conducted.

For each setting of $K$, the optimal $\lambda_w$ is determined by searching over the range $\log_{10} \lambda_w = -3 \ldots +3$, and selecting the value that minimizes $\text{ERR}_{\text{missing}}$. This allows setting the parameter to effect an optimal compromise between stimulus driven and previous state driven transitions. We observed that the prediction error is relatively insensitive to $\lambda_w$ (cf. Results section), and therefore a common value can be selected across a multi-subject data-set for one study.

The reader will observe that prediction error is used merely as a statistic (Friston et al., 2008) to select hyper-parameters. The parameters themselves, unlike MVP classifiers, are not estimated to minimize prediction error but rather to fit a model of brain function to the data. It is this distinction that allows interpretation of the estimated parameters (in terms of the underlying neurophysiological model) in contrast to MVP classifiers. The effect of these hyper-parameters and the length $T$ of a missing-stimulus block on the model estimation is evaluated in the next section.

Results

This section starts off with a quantitative validation of the model and estimation algorithms using a synthetic data-set. Then the results of a method applied to a multi-subject fMRI study for mental arithmetic processing are presented, including the ability of the method to discover new patterns in the data, a comparative evaluation with respect to other analysis methods and feature-spaces and followed by the group-level analysis of the data-set made using this spatio-temporal generative model.

The algorithms were implemented in MATLAB® with Star-P® on a 2.6 Hz Opteron cluster with 16 processors and 32 GB RAM.
Discussion

The relative error $\text{ERR}_{\text{estimate}}$ in the parameter estimates $\theta^*$, the relative error $\text{ERR}_K$ of model-size estimates $K^*$ and the prediction error $\text{ERR}_{\text{missing}}$ (cf. SSM hyper-parameter selection section) for the various experiments are charted in Fig. 5.

One of the main observations is that the MCMC algorithm requires almost thrice the total running-time (including searching for the optimal hyper-parameter values) for the same estimation error $\text{ERR}_{\text{estimate}}$ as the mean field EM method (MF-GEM), while the prediction error of MF-GEM is within 20% of the best $\text{ERR}_{\text{missing}}$ as measured by MCMC:CNV. While reducing SNR does not affect running-time significantly, its effect on the errors is large. Reducing the SNR from 30 to 20 dB caused prediction error to increase from $<10$% to $\approx 30$%. Furthermore, for SNR $\leq$ 20dB the estimate for model-size using the maximally predictive criteria is within 10% of the true $K$.

Although all the parameters $\theta$ are important in determining the accuracy of the model, of special interest are the $\mu_k$, $k = 1 \ldots K$, as they correspond to the spatial distribution of activity representative of each state. The average estimation error, defined as $\text{ERR}_{\text{spatial}} = 1/K \sum_k (\mu_k - \mu_k^*)$, of these parameters for the MF-GEM and MCMC:CNV cases are listed in Table 2. It can be observed that for the 20 dB case, the estimated spatial patterns are within $\approx 0.25$ standard deviations (given by $\Sigma_k$) of the true $\mu_k$.

Examining effect of model-size $K$ on prediction-error in Fig. 6 (black solid line), the classical bias-variance trade-off can be observed. For $K$ much less than the true value, the estimated model is too simple and may fail to account for all the patterns in the data, while for $K$ much greater than the true value, the model may start to fit noise (over-fitting) thereby compromising its ability to predict the driving stimulus.

The plot of $\text{ERR}_{\text{missing}}$ versus precision hyper-parameter $\lambda_w$ in Fig. 6 (red dashed line) shows how it determines the balance between the temporal structure of the stimulus-sequence and the state-sequence. A model with transitions highly determined by current stimulus (i.e. high $\lambda_w$) fails to learn the temporal structure of the state-sequence and hence fails to estimate the correct state for a time-point with missing stimulus. Similarly, one with transitions driven mainly by the previous state (i.e. low $\lambda_w$) fails to leverage any task-related structure in the data, thereby reducing its prediction power for missing stimuli. However, $\text{ERR}_{\text{missing}}$ achieves near optimal levels over a large range of values of $\lambda_w$ (e.g. $10^{-1.5} \leq \lambda_w \leq 10^1$) indicating a high degree of robustness with respect to it.

The effect of the block size $T$ (cf. SSM hyper-parameter selection section) on prediction error, for different FWHMs of the Gaussian filter applied to $s$, is shown in Fig. 7. The prediction error for small values of $T$ is extremely low, increases as the length of the missing stimulus block increases and then stabilizes. This increase followed by stabilization happens at block lengths proportional to the FWHMs, typically $3 \times \text{FWHM}$. For block lengths $T \leq 3 \times \text{FWHM}$, the high accuracy is due to the strong temporal regularity in the stimulus sequence over small durations. However, as block length increases, this temporal regularity becomes weaker and prediction is driven primarily by the spatio-temporal patterns in the data after $T \geq 3 \times \text{FWHM}$ resulting in stable, albeit higher, error-rates.

fMRI data-set: mental arithmetic task

The method was applied to an fMRI study (Morocz et al., 2003; Shalev, 2004) for number processing capabilities in healthy, dyscalculic and dyslexic individuals. Using this data-set, this section provides quantitative comparisons of the SSM and the feature-space $\Phi$, an investigation of the various parameters of the model and a neurophysiological discussion of the results. Also the ability of the SSM to compare and contrast the spatio-temporal patterns across groups of subjects is demonstrated.

Background

Dyscalculia (DC) is a specific learning disability affecting the acquisition of mathematical skills in children with otherwise normal
general intelligence and has been attributed to either abnormalities in the numerical (non-verbal) quantity processing system localized in the left and right intraparietal regions, or to impairments of the verbally encoded mathematical-fact retrieval system localized in the frontal and temporal (language) regions (Dehaene et al., 2003).

Dyslexia (DL) is a reading disorder defined as a selective inability to build a visual representation of a word used in subsequent language processing, in the absence of general visual impairment or speech disorders. Certain but not all types of DL may be caused by infarcts in the left middle cerebral hemisphere affecting the temporo-parietal and frontal regions or by atrophy in the anterolateral temporal lobe or by damage to the left occipito-temporal regions (Price and Mechelli, 2005).

Neuroimaging studies of both developmental disabilities have failed to pin-point their neuronal substrates. In the case of DC, difficulties in accurate diagnosis, heterogeneity of arithmetic difficulties and the frequent association with DL and attention disorders have impeded spatial localization. Indeed, a great variety of nonspecific problems including slow speed of processing, poor working memory span, attentional disorders and deficits in the long-term retrieval of arithmetic facts, may influence arithmetic performance (Molko et al., 2003). While, it has been proposed that DL is more generally characterized by a disconnection syndrome of the reading network, the neural correlates of these putative reading pathways are not well understood (Price and Mechelli, 2005).

**Methods and materials**

In each trial of the self-paced, irregular paradigm, two single-digit numbers (e.g. $4 \times 5$) were displayed visually for 2.5 s. After an interval of 0.3 s an incorrect solution (e.g. 27, 23, 12) was displayed for 0.8 s. Subjects had up to 4 s to decide, with a button press, if the answer was (a) close (within ±25% of the correct answer), (b) too small or (c) too big.

The next trial started after a rest of 1 s, and each trial lasted 4–8.6 s. For each $t = 1 \ldots T$, the experimental conditions were described by:

- $[\text{Ph}]$: Indicates if $t$ is (1) multiplication or (2) subtraction/judgement or (3) decision-making phase of the experiment
- $[\text{LogPs}]$: Quantifies the product size of the multiplication problem ($1 \leq \text{LogDiff} \leq 10$)
- $[\text{LogDiff}]$: Quantifies the expected difficulty in judging the right answer$^3$ ($1 \leq \text{LogDiff} \leq 5$).

$^3$ 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 is scored as $\text{LogPs} = \log(R_c)$. The score $\text{LogDiff}$ is $\log([|1.25R_c| - (R_c + |R_c - R_d|)/1.25R_c])$, which measures the closeness of the incorrect result to the ±25% mark and represents the difficulty in judging the correct answer.

<table>
<thead>
<tr>
<th>SNR (dB)</th>
<th>MF-GEM</th>
<th>MCMC:RT</th>
<th>MCMC:EE</th>
<th>MCMC:CNV</th>
</tr>
</thead>
<tbody>
<tr>
<td>30</td>
<td>0.151±0.06</td>
<td>0.126±0.05</td>
<td>0.126±0.05</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>0.223±0.09</td>
<td>0.212±0.09</td>
<td>0.212±0.09</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>0.361±0.13</td>
<td>0.357±0.14</td>
<td>0.357±0.14</td>
<td></td>
</tr>
</tbody>
</table>

![Fig. 5. Simulation results. The GEM method under mean field approximation (MF-GEM) is compared against an MCMC based estimation algorithm matched in terms of equal running time (MCMC:RT), equal estimation error (MCMC:EE) and MCMC run until convergence of the estimates (MCMC:CONV). The experiments were repeated for SNR = 10, 20 and 30 dB. Plotted are total running time (Fig. (a)), relative estimation error (Fig. (b)), prediction error (Fig. (c)) and relative error in estimating the correct $K$ (Fig. (d)) for the different experiments. Error bars indicate ±1 standard deviations.](image-url)
Twenty control subjects, thirteen high-performing (full-scale IQ:95) individuals with pure dyscalculia (DC) and nine with dyslexia (DL) participated. All subjects were free of neurological and psychiatric illnesses and attention-deficit disorder. All controls denied a history of calculation and reading difficulties. In order to balance the group sizes, group-level analysis was done by selecting 8 subjects at random from each group and computing the statistics over multiple resamples.

Each subject participated in two fMRI sessions each of 12.19 min, with 240 trials and 552 scans in total. The data were acquired with a GE 3 T MRI scanner with quadrature head coil, using a BOLD sensitized 3D PRESTO pulse sequence with volume scan time of 2.64s and resolution of 3.75×3.75×3.75 mm³. High resolution anatomical scans were also acquired, bias-field corrected, normalized to an MNI atlas and segmented into gray and white matter regions. The fMRI scans were motion corrected using linear registration and co-registered with the structural scans using SPM8 (Multiple, 2007). Next, the time-series data were high-pass filtered (0.5Hz) to remove artifacts due to breathing, blood pressure changes and scanner drift. The mean volume of the time-series was subtracted, white matter artifacts due to breathing, blood pressure changes and scanner drift. Registered with the structural scans using SPM8 (Multiple, 2007).

The mean volume of the time-series was subtracted, white matter artifacts due to breathing, blood pressure changes and scanner drift. Registered with the structural scans using SPM8 (Multiple, 2007). The mean volume of the time-series was subtracted, white matter artifacts due to breathing, blood pressure changes and scanner drift. Registered with the structural scans using SPM8 (Multiple, 2007).

The mean volume of the time-series was subtracted, white matter artifacts due to breathing, blood pressure changes and scanner drift. Registered with the structural scans using SPM8 (Multiple, 2007). The mean volume of the time-series was subtracted, white matter artifacts due to breathing, blood pressure changes and scanner drift. Registered with the structural scans using SPM8 (Multiple, 2007).

The mean volume of the time-series was subtracted, white matter artifacts due to breathing, blood pressure changes and scanner drift. Registered with the structural scans using SPM8 (Multiple, 2007).

The mean volume of the time-series was subtracted, white matter artifacts due to breathing, blood pressure changes and scanner drift. Registered with the structural scans using SPM8 (Multiple, 2007). The mean volume of the time-series was subtracted, white matter artifacts due to breathing, blood pressure changes and scanner drift. Registered with the structural scans using SPM8 (Multiple, 2007).

The mean volume of the time-series was subtracted, white matter artifacts due to breathing, blood pressure changes and scanner drift. Registered with the structural scans using SPM8 (Multiple, 2007). The mean volume of the time-series was subtracted, white matter artifacts due to breathing, blood pressure changes and scanner drift. Registered with the structural scans using SPM8 (Multiple, 2007).

The mean volume of the time-series was subtracted, white matter artifacts due to breathing, blood pressure changes and scanner drift. Registered with the structural scans using SPM8 (Multiple, 2007). The mean volume of the time-series was subtracted, white matter artifacts due to breathing, blood pressure changes and scanner drift. Registered with the structural scans using SPM8 (Multiple, 2007).
whereas the SSM accounts for spatially varying and unknown hemodynamics in a probabilistic fashion which contributes to its ability to predict the mental state of the subject.

Using only information about the phase $\Phi$ to train SSMs increased the error as compared to SSM:FULL, but only slightly ($\approx 1$ SEM). But removing all experimental information (SSM:NONE) caused a dramatic increase in error ($\text{ERR}_{\text{SSM:NONE}} \approx 0.48$). This implies that the semi-supervised SSM can detect the effect of experimental variables from patterns in the data (namely LogPs and LogDiff in the case of SSM:PH) against which it was not explicitly trained. Including some cues (namely $\Phi$) about the experiment guides this discovery, stabilizes estimation and precludes the model from learning spatio-temporal patterns that are not relevant to the task and which may be due to artifacts (unlike SSM:NONE). It is nevertheless interesting that, despite not using any stimulus information, SSM:NONE had a prediction error much better than chance ($\text{ERR}_{\text{Chance}} \approx 0.87$) which implies that it has discovered the mental states of the subject in a purely unsupervised fashion, validating some of the neurophysiological assumptions behind the SSM.

The unsupervised PCA feature-space (FS:PCA-NONE) exhibited performance worse than FS:$\Phi$ in all cases. The poor accuracy of PCA has also been documented in other fMRI studies (O’Toole et al., 2007) and can be attributed to the lack of a specific relationship between the task selectivity of a PC and its variance. In contrast, as FS:$\Phi$ is obtained from the correlation matrix, it describes the structure of the inter-relationships between the voxel time-series and not their magnitudes. Using PCs selected against the stimuli (FS:PCA-FULL) deteriorates the performance of the SSMs even further. This is because the exact coupling between stimuli and fMRI signal (and therefore PC time-courses) is unknown and may be non-linear, and selecting PCs linearly correlated with HRF-convolved stimuli may not preserve a large proportion of the spatio-temporal patterns in the data. In contrast, the SVM, based on optimizing prediction error, has best overall performance with this feature-selection strategy. The limitations, however, of supervised feature-selection are apparent in the case of PCA:PH. Although the SVM predicts Ph with high accuracy ($\text{ERR}_{\text{SVM:PH}} \approx 0.08$ for FS:PCA-PH vs. $\approx 0.11\%$ for FS:PCA-FULL), its ability to predict any other stimulus is severely degraded with an overall error of $\approx 0.50$. The SSMs have similarly poor performance, due to the loss of information about spatio-temporal patterns in this basis.

### SSM parameter estimates

This section further investigates the parameters as estimated by SSM:PH trained with FS $\Phi$ and $s_i = (\Phi,1)$. The prediction error $\text{ERR}_{\text{SSM:PH}}$ exhibits a relatively shallow basin with respect to model-size $K$ for all three groups in Fig. 9(a), with minima occurring in the range $K \approx 18 \ldots 28$. This points to the robustness of the SSM estimation with respect to $K$ for each subject. The robustness of $\text{ERR}_{\text{SSM:PH}}$ with respect to $\lambda_w$, shown in Fig. 9(b) was comparable to that of the simulation study, with the curve of $\text{ERR}_{\text{SSM:PH}}$ almost flat for $10^{-2} \leq \lambda_w \leq 10^{-1}$.

From the plot of $\text{ERR}_{\text{SSM:PH}}$ versus the dimensionality $D$ of the feature-space $\Phi$ in Fig. 9(c), it can be noticed that initially $\text{ERR}_{\text{SSM:PH}}$ drops as $D$ increases with $\Phi$ explaining more of the information in the data and bottoms out at $400 \leq D \leq 600$, across the three groups. It then begins to slowly rise as a larger number of unstable basis vectors are included, capturing an increasing percentage of the noise in the data.

Fig. 9(d) graphs $\text{ERR}_{\text{SSM:PH}}$ versus the length $T$ of the missing stimulus block used in the assessment of $\text{ERR}_{\text{missing}}$ (cf. SSM hyper-parameter selection section). Here, we observe a very low error for small $T$ as prediction is driven primarily by the strong temporal regularity of the stimulus presentation sequence over short durations. However, as in the case of the simulation, it increases with $T$ and stabilizes at a block length of 2 trials ($T \approx 5$ TRs) after which point there is no structure in the stimulus sequence and prediction is driven mainly by the patterns in the data.

Table 3 compares the prediction error of SSMs with: [HRF:NONE] no HRF FIR filter; [HRF:CONST] a spatially constant HRF of length $L + 1 = 32 s$; [HRF:UNCON] spatially varying HRF of length $L + 1 = 32 s$ without any constraints; and [HRF:PRIOR] the SSM with spatially varying and unknown HRF of length $L + 1 = 32 s$ but constrained by the prior density $N(\mu_s, \Sigma_s)$ (cf. The state-space model (SSM) section). Here, the advantage of the spatially-varying but physiologically constrained HRF (HRF:PRIOR) in dealing with the variable hemodynamics of the brain and accurately predicting the mental state can be seen.

---

**Fig. 9.** Effect of hyper-parameters on $\text{ERR}_{\text{SSM:PH}}$. Fig. (a): $\text{ERR}_{\text{SSM:PH}}$ with respect to model-size $K$. Fig. (b): $\text{ERR}_{\text{SSM:PH}}$ with respect to precision hyper-parameter $\lambda_w$. Fig. (c): $\text{ERR}_{\text{SSM:PH}}$ with respect to dimension $D$ of $\Phi$. Fig. (d): $\text{ERR}_{\text{SSM:PH}}$ with respect to missing stimulus block length $T$. Error bars indicate $\pm 1$ SEM. Legend. Blue solid line: control group. Red dashed line: DC group. Green dot-dashed line: DL group.
Removing the HRF altogether from the model (HRF:NONE), thereby not accounting for the lag in the fMRI signal due to hemodynamics, leads to the largest deterioration in performance. Although the inclusion of a spatially constant HRF (HRF:CONST) causes some reduction in accuracy, the spatially varying HRF with a prior obtained from canonical HRF of SPM8 (HRF:PRIOR) results in even worse performance due to over-fitting of noise.

Fig. 10 shows the estimates of the spatially varying but constrained HRF FIR filter (HRF:PRIOR) for each group, averaged in regions-of-interest (ROI) selected in the left primary motor cortex (BA3, BA4) and the bilateral intraparietal sulcus (IPS) (BA40). A qualitative difference in the estimated HRFs is apparent in terms of their rise-time, peak value and dispersion. The prolonged and repeated recruitment of the IPS in this task may explain the dispersed shape of its HRF as compared to the motor cortex. No significant differences in HRF estimates were observed between the three groups.

The group-wise spatial maps corresponding to the three phases of each trial are shown in Fig. 11.

Discussion

The average optimal model size $K^*$ and prediction error $\text{ERR}_{\text{SSM:PH}}$ for the three groups are shown in Table 4. Here, we notice the variation in model-sizes for the DC group is larger than the controls while that for DL group is almost of the same order. This points to a greater heterogeneity in the DC data necessitating models with different sizes.

Also, the consistently higher error-rate of the DC population indicates the relative inaccuracy of the models for their mental processes, as compared to the other two groups.

These observations concur with the theory (Dehaene et al., 2003) that not only are dyscalculics different from each other in their arithmetic strategies, their lack of an intuitive notion of numerical size maybe compensated for by shifting mental strategies resulting in the poor fit of a single model for a subject.

The effect of Ph, LogPs and LogDiff on the error $\text{ERR}_{\text{SSM:PH}}$ is shown in Fig. 12. Note that LogPs and LogDiff were not used to train the model, and therefore the influence of these parameters on the mental patterns of the subjects was effectively discovered by the method.

To measure the similarity between the SSMs $M_i$ and $M_j$ of two subjects $i$ and $j$, the *mutual information* ($\text{MI}$) between the optimal state-sequences estimated by the two SSMs was used. Specifically, the MI was derived from the joint histogram of $X^{(i)}$ and $X^{(j)}$, the optimal state-sequences for the same fMRI data $Y$ as labeled by models $M_i$ and $M_j$ respectively. In general the mapping between the state labels of two different SSMs is unknown and by comparing the state-sequences for the same data, this correspondence can be determined. A higher MI indicates a higher level of correspondence, while an MI of zero indicates no agreement. This procedure applied to all ($\binom{\text{2}}{\text{2}}$) pairs of subjects yielded a similarity matrix of pair-wise MI that was then visualized with *multidimensional scaling* (MDS) (Edelman et al., 1999) in Fig. 13. The specification of the SSM in terms of abstract mental-states allows comparing the spatio-temporal patterns between subjects in their entirety in this abstract representation (Kriegeskorte et al., 2008).

Fig. 13(a) shows a clustering of subjects in the MDS space with respect to their group (control, DL or DC) along the vertical axis, while along the horizontal axis we see a slight, but not significant, organization dictated by gender. Since this labeling is applied after plotting all the subjects in the MDS space, an *intrinsic organization* in the spatio-temporal patterns of the subjects in each group has been identified. Interestingly, there are a few DC subjects that cluster along with the DL group, at the top of Fig. 13(a). This is not surprising, given that often times dyscalculia is comorbid with dyslexia (Molko et al., 2003) and these DC subjects may exhibit dyslexic deficits during this task.

Table 3

| Prediction error versus different HRF models. $\text{ERR}_{\text{SSM:PH}}$ ($\pm 1$ SEM) for the SSM model with no HRF (HRF:NONE), spatially constant HRF (HRF:CONST), spatially varying and unconstrained HRF (HRF:UNCON) and the spatially varying HRF with a prior obtained from canonical HRF of SPM8 (HRF:PRIOR). |
|------------------|--------|--------|
|                  | Control | DL     |
| HRF:NONE         | 0.62±0.11 | 0.68±0.13 | 0.64±0.10 |
| HRF:CONST        | 0.36±0.08 | 0.46±0.11 | 0.40±0.09 |
| HRF:UNCON        | 0.54±0.13 | 0.59±0.15 | 0.55±0.16 |
| HRF:PRIOR        | 0.31±0.05 | 0.40±0.09 | 0.33±0.05 |

Fig. 10. Estimated HRF FIR filter $h$. (a): The locations of the ROIs (in the left hemisphere). (b-d): The estimated FIR filter coefficients ($\pm 1$ std.dev.) for each group averaged in the ROI in the left motor cortex and the left and right IPS. Legend: Blue solid line: control group, Red dashed line: DC group, Green dot-dashed line: DL group.
The separation between the MDS clusters for each group can be quantified using Cramér test (Baringhaus and Franz, 2004) which provides a non-parametric measure of the $p$-value of the distance between the means of two samples through a permutation method. The $p$-values of the group-wise differences are compiled in Table 5.

From the results in Figs. 11–13 and Table 5 the following observations can be made.

**Multiplication phase.** The error rate for the DL group is much higher than that for the controls (cf. Fig. 12). An increase in product-size causes a large ($>1.5$ SEM) reduction in ERR$_{SSM,PH}$ for controls, while the effect for the DC and DL groups is less pronounced ($>1$ SEM). Also, there is a clear separation between the DL and control groups in the MDS space and product-size increases the separation between the DC and control groups. For the control subjects high values are seen in the bilateral occipital extra-striate cortices, the left postcentral area, the left angular gyrus (lAG), the medial frontal gyri (MFG), and the left intra-parietal sulcus (IPS). The DC subjects show lower activation in the bilateral IPS, while the DL subjects show increased activation in their left fronto-parietal and left medial frontal gyral (lMFG) regions as compared to controls.

These results may be due to the greater difficulty and conflict experienced by the DL subjects and multiplicity of mental strategies adopted during the reading phase of the task. The higher error of the DC subjects may be due to irregular patterns in accessing the verbally encoded rote multiplication tables located in the lAG. The reduction in error-rates of all subjects with increase in product-size may be due to increased organization of their mental processes as their multiplication memory is stressed, while the increased separation between the groups could indicate greater divergence of the mental patterns of the DC individuals from the controls.

**Judgement phase.** ERR$_{SSM,PH}$ for the DC group increases drastically, while that for the DL and control groups match up. The DC subjects may experience difficulty in judging the difference between the size of the correct and incorrect results and may resort to a greater variety of mental strategies. Not surprisingly, as the reading phase of the experiment has ended, the patterns of the DL individuals begin to resemble that of the controls and the separation between these groups reduces in MDS space, while the separation of the DC group increases. The control and DL subjects exhibit high values in the left and right IPS, both pallida, caudate heads (CdH), left anterior insula (aIn), lMFG, the supplementary motor area (SMA) and the left fronto-parietal operculum, while the map for the DC group activates in both aIn, both MFG, left IPS, the anterior rostral cingulate zone (aRCZ), and the right supramarginal gyrus (SMG). Although LogDiff reduces the error-rate of the control and DL subjects, it has the opposite effect on the DC group as increased conflict may recruit new functional circuits. The effect of LogPs is consistent with strong activation of the working verbal (mute rehearsal) and visual memories.
Third phase. This phase involves decision-making and conflict-resolution and is highly variable between repetitions and subjects, causing increased inaccuracy during this phase. Also, due to the self-paced nature of the task, it very often contained the button-press and inter-trial rest interval. The spatial-maps for the three groups show increased foci in the pre-frontal and motor areas. The left IPS region in the DC group is also strongly activated during this phase, which may point to irregular storage and retrieval of the number size using spatial attributes typically processed in this region (Morocz et al., 2003).

Conclusion

This paper has described a time-resolved analysis of the spatio-temporal patterns contained in fMRI data that correspond to the metabolic traces of neural processes using a state-space formalism. The model incorporated information about the experiment to guide the estimation towards patterns of relevance to the task. The data were represented in a low-dimensional feature-space derived from a physiologically motivated definition of functional distance. Efficient estimation algorithms using a variational formulation of generalized EM under the mean field approximation were developed and quantified with a simulation study. The HRF of the brain is known to be highly variable (Logothetis, 2008) and by using a spatially varying but unknown FIR filter, the state-space model (SSM) was able to compensate for this variability. Model hyper-parameters were selected in an automated fashion using a maximally predictive criterion.

The hidden layers in the SSM decouple the stimulus from the data, and therefore neither does the stimulus need to be convolved with an HRF nor does the exact mathematical relationship between the stimulus and the fMRI signal need to be specified. This allows flexibility in choosing which experimental variables to include and their encoding, without having to worry about statistical issues like the orthogonality of the experiment, the estimability of the design matrix and omitted variable bias. But classical issues like confounding variables will still affect inference and must be addressed through appropriate experimental designs.

As demonstrated by the mental arithmetic study, this method can be used with arbitrarily complex paradigms, where the investigator can decide which stimuli to provide as input thereby choosing a trade-off between data driven and model (i.e. stimulus) driven estimation of parameters. The effects of other un-modeled experimental variables on the model can then be tested, post hoc. This is in contrast to supervised methods that cannot, by design, capture the effects of experimental variable against which they have not been modeled. However, with simple block design paradigms where the effect of hemodynamics and the temporal structure within a block are insignificant, we observed that MVPR classifiers tended to outperform the SSM in predicting the mental state. Also, its application to default-state and non-task related fMRI studies would require an alternative model-size selection procedure that does use prediction error as a criterion.

The SSM parameters are estimated through a fitting criterion and consequently have a well-defined interpretation implied by the underlying neurophysiological model. Here prediction error is used as a...
statistic to select between models and to infer an effect of the experimental variables on the data, which implicitly involves selecting between alternative hypotheses (Friston et al., 2008). For example, the ability to predict mental states at “much better than chance” levels adds evidence against the null-hypothesis that the SSM does not explain the data. A similar argument applies for the predictability of experimental variables that were not included during the training of the SSM. The SSM, however, due the lack of a parametric form of the null distribution the prediction error and the prohibitively high cost of a non-parametric permutation test, cannot measure the confidence level (i.e. a p-value) in a hypothesis test.

Comparing brain-function in abstract representation spaces rather than the spatial-maps directly has been shown to be a very powerful abstract state-space representation was used to compare the spatio-temporal signatures of mental processes in their entirety. Systematic differences in the cascade of recruitment of the functional modules between subject populations were shown indicating the necessity of retaining the temporal dimension. The MDS plots derived from the MI between subject pairs enabled a succinct assessment of the relationship between alternative hypotheses (Friston et al., 2008). For example, the abstract state-space representation was used to compare the spatio-temporal signatures of mental processes in their entirety. Systematic differences in the cascade of recruitment of the functional modules between subject populations were shown indicating the necessity of retaining the temporal dimension. The MDS plots derived from the MI between subject pairs enabled a succinct assessment of the relationship between different groups with respect to experimental parameters. This ability to reveal and study the group-wise structure in the spatio-temporal patterns could guide in the design of more specific experiments to test interesting effects.

Therefore, given its advantages and disadvantages with respect to other analysis methods, we believe that it is a complementary tool in an investigator’s arsenal providing a new and different insight into mental processes.

Acknowledgments

We would like to thank Prof. Dana Brook, Prof. William Wells (III) and the anonymous reviewers for their suggestions and constructive comments, and Dr. Ruth Shalev for identifying and scanning subjects with dyscalculia and dyslexia.

This work was partially funded by NSF grant CISE:DC-SMALL-0916196 and was supported in part by an allocation of computing time from the Ohio Supercomputer Center.

Appendix A. Supplementary data

Supplementary data to this article can be found online at doi:10.1016/j.neuroimage.2011.03.047.

References


