Published in final edited form as:

Psychometrika. 2013 April ; 78(2): 279–307. doi:10.1007/s11336-012-9300-6.

State-Space Analysis of Working Memory in Schizophrenia: An FBIRN Study

Firdaus Janoos, PhD,

Harvard Medical School and Brigham and Women's Hospital, Boston, USA

Istvan A Morocz, PhD,

Harvard Medical School and Brigham and Women's Hospital, Boston, USA

Gregory Brown, PhD, and

University of California, San Diego, USA

William Wells III, PhD

Harvard Medical School and Brigham and Women's Hospital, Boston, USA

Firdaus Janoos: fjanooos@bwh.harvard.edu; Istvan A Morocz: pisti@bwh.harvard.edu; Gregory Brown: gbrown@ucsd.edu; William Wells: sw@bwh.harvard.edu

Abstract

The neural correlates of *working memory* (WM) in schizophrenia (SZ) have been extensively studied using the multi-site fMRI data acquired by the Functional Biomedical Informatics Research Network (fBIRN) consortium. Although univariate and multivariate analysis methods have been variously employed to localize brain responses under differing task conditions, important hypotheses regarding the representation of mental processes in the spatio-temporal patterns of neural recruitment and the differential organization of these mental processes in patients versus controls have not been addressed in this context. This paper uses a multivariate *state-space model* (SSM) to analyse the differential representation and organization of mental processes of controls and patients performing the Sternberg Item Recognition Paradigm (SIRP) WM task. The SSM is able to not only predict the mental state of the subject from the data, but also yield estimates of the spatial distribution and temporal ordering of neural activity, along with estimates of the hemodynamic response. The dynamical Bayesian modeling approach used in this study was able to find significant differences between the predictability and organization of the working memory processes of SZ patients versus healthy subjects. Prediction of some stimulus types from imaging data in the SZ group was significantly lower than controls, reflecting a greater level of disorganization/heterogeneity of their mental processes. Moreover, the changes in accuracy of predicting the mental state of the subject with respect to parametric modulations, such as memory load and task duration, may have important implications on the neuro-cognitive models for WM processes in both SZ and healthy adults. Additionally, the SSM was used to compare the spatio-temporal patterns of mental activity across subjects, in a holistic fashion and to derive a low-dimensional representation space for the SIRP task, in which subjects were found to cluster according to their diagnosis.

1. Introduction

1.1. Background

Understanding the neuro-cognitive processes underlying *working memory* (WM) impairment is a fundamental step towards developing a comprehensive neurobiological theory for schizophrenia (Perlstein, Carter, Noll, & Cohen, 2001) and is an important objective of the Functional Imaging Biomedical Informatics Research Network (fBIRN) consortium (Potkin & Ford, 2009). Although patients with schizophrenia (SZ) exhibit performance deficits in nearly all aspects of brain function, WM has been identified as one most severely affected functional domains (Bilder, 2009). The neural correlates of WM have been extensively studied using the psychometric and neuroimaging data gathered from healthy adults and from patients suffering from schizophrenia or schizoaffective disorder while performing a variant of the Sternberg Item Recognition Paradigm (SIRP) (Brown et al., 2009). The task involved the presentation of a memory set that included 1, 3, or 5 digits followed by 14 probe digits. Participants responded manually to indicate whether or not a probe digit was included in the memory set just studied. Because it encompasses a large number of probe trials, the multi-probe, fixed-set design efficiently assesses the brains response to memory probe events under different memory load conditions (Brown et al., 2009).

Univariate general linear model (GLM) analyses of fMRI data from WM studies have revealed a strong involvement of the dorso-lateral pre-frontal cortex (dlPFC) in WM recall in healthy subjects, while both prefrontal hyper- or hypoactivity has been observed in chronic schizophrenia (Manoach et al., 1997; Manoach, 2003; Potkin et al., 2009). Differences in the recruitment patterns of other regions, including the temporal lobes, superior and inferior parietal lobes and basal ganglia, under varying conditions of memory load, memory encode vs. memory probe and other behavioral indicators, have also been demonstrated using univariate and multivariate functional localization techniques (Mendrek et al., 2005; M. A. Kim et al., 2010). Canonical variates analysis (CVA) of the functional connectivity during WM tasks was used to evince support for the frontal-temporal disconnection hypothesis (Meyer-Lindenberg et al., 2001; Wolf et al., 2007), while independent components analysis (ICA) of the data has shown differences in the functional connectivity of the fronto-striatal and default-mode networks between patients and controls (D. I. Kim et al., 2009). Brown et al. (2009) tested reaction times (RT) against four memory retrieval models and correlated the results with the fMRI time-series. Based on the decoupling of brain response to increasing memory load in patients and on group differences in the correlation of the best fitting models parameters to the brain's response under increasing memory load, they suggested a hypothesis of inefficient or disordered patterns of neural inhibition among individuals with schizophrenia.

1.2. Contribution

Disjoint from the problem of the localization of neural deficits, an important hypothesis regarding the neuropathologies underlying schizophrenia concerns the presence of systematic differences in the neural representation of WM processes, the transformation of this representation at different stages of the processing and *its organization* when compared

to healthy subjects (Egan & Weinberger, 1997), a problem which has been relatively unexplored.

This paper presents an analysis of the representation of these mental processes in the brain and its relative organization in two populations through the ability of a model to explain (predict) observational data, and the statistical complexity of that model (Hastie, Tibshirani, & Friedman, 2003). A *neurophysiologically based multivariate dynamical model of brain function* (Janoos, Machiraju, Singh, & Morocz, 2011; Mórocz et al., 2012) is used, which permits direct assessment of the level of organization of these processes in terms of its accuracy in predicting the cognitive state of the subject and comparison of the spatio-temporal patterns of neural processes between patients and controls.

This *state-space model* (SSM) assumes the presence of a set of abstract “brain-states” that are revisited during the performance of a mental task. Given the goal-oriented and directed nature of human thought, the transition of the brain through this brain-state space not only depends on external stimuli but also exhibits an intrinsic temporal ordering. Each state is associated with a characteristic spatial distribution of neural (metabolic) activity and a realization of the state corresponds to a neural activation derived from this signature. The fMRI data arise from an *unknown and spatially varying* hemodynamic response to these activation patterns, the shape of which is estimated from the data. The high-dimensional fMRI data are transformed into a low-dimensional representation by projecting onto a feature-space derived from a definition of the *functional similarity* between the activation patterns present in the fMRI data at two time-points (Janoos et al., 2011; Mórocz et al., 2012). Dimensionality reduction is performed using a *bootstrap analysis of stability*, that retains only those feature-space basis vectors that are stable across multiple resamples of the fMRI data. This feature-space allows for a multivariate exploration of neural activation patterns spatially distributed over the entire cortex, while providing a high degree of robustness against artifacts of non-biological origin.

The SSM is able to directly and indirectly predict the experimental stimuli presented to subjects from their fMRI data. In most neuroscientific investigations, prediction of the presented stimulus from new fMRI data is not of direct interest, but rather the predictive density or the generalization error-rate (i.e. the error-rate in predicting new samples, typically estimated by crossvalidation) is used in lieu of model-evidence (i.e. the log-likelihood of the model for the observed data) to assess the fit of the model to the data and to detect the presence of an effect of an experimental manipulation (Kriegeskorte & Bandettini, 2007; K. Friston et al., 2008). In the current application, prediction error is used to test how well the spatio-temporal patterns discovered by the model relate to the cognitive states of the subject, as proxied by behavioral indicators. Furthermore, the predictability indicates the degree to which the neural patterns evoked by a specific stimulus occurrence or experimental condition are accordant with those evoked by other occurrences of the same stimulus or condition. The differences in the prediction accuracy of the patient group, with the control group as a baseline, provides a quantification of the *differential organization or regularity* in the neural representation of their mental processes.

The model is elaborated further in Section 2, followed by a brief description of the parameter estimation algorithms in Section 2.1 and hyper-parameter selection procedures in Section 2.2. Section 2.3 describes the low-dimensional feature-space used to represent the fMRI data. The specifics of the experiment and the SIRP task structure followed by the details of the fMRI data-set acquisition, the preprocessing and analysis are provided in Section 3. The ability of the SSM to predict the experimental stimuli is evaluated with respect to alternative methods in Section 4.1, followed by a detailed analysis of the SSM results in Section 4.2.

Comparing brain-function in abstract representation spaces rather than spatial-maps directly has been shown to be a powerful principle in psychology and cognitive neuroscience (Kriegeskorte, Mur, & Bandettini, 2008). The representation of a subject's brain function in terms of its trajectory through the brain-state space permits a holistic comparison of the spatio-temporal patterns of neural recruitment across subjects and the differences between patients and controls, in this abstract representation space, is presented in Section 4.3. The implication of these results and how they pertain to our current understanding of schizophrenia is discussed in Section 5

1.3. Related Work

Multivariate pattern recognition (MVPR) 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 (Haynes & Rees, 2006; Norman, Polyn, Detre, & Haxby, 2006). The main advantage of the multivariate approach is its sensitivity to the distributed nature of cognitive processes by integrating activation information from groups of voxels that individually are weakly activated, but whose joint activations are highly structured with respect to the task. MVPR methods such as linear discriminant analysis (LDA), neural-networks, support vector machines (SVMs) and other types of pattern classifiers (e.g. Haxby et al., 2001; Haynes & Rees, 2006; Mitchell et al., 2008) treat the data as an abstract representation of neural activity without modeling the underlying neurophysiology. Although this approach can be an advantage when facing complete uncertainty about the neurophysiology, judiciously incorporating well-proven assumptions into a model, as in the method used here, can provide the ability to make a wider range of inferences and moreover can permit neurophysiologic interpretation of the results (K. J. Friston, 2009). Furthermore, most MVPR methods make the assumption that all fMRI time-points with the same label (i.e. behavioral state) are equivalent, thereby neglecting temporal variations in mental processes (Norman et al., 2006). The current method, on the other hand, provides a multivariate dynamical system model to understand the spatio-temporal representation of brain function.

Alternative applications of state-space models in fMRI have been concerned with the localization of neural activity in the brain from the observed signal, essentially addressing the problem of independently *deconvolving the hemodynamic response* at each voxel or at a group of voxels. Bayesian hidden Markov models (HMMs) with two latent states have been used to determine the activation state of individual voxels in block-design experiments (Højen-Sørensen, Hansen, & Rasmussen, 2000). Penny, Ghahramani, and Friston (2005)

proposed a Kalman filtering based method, called a bilinear dynamical system (BDS), to deconvolve the hemodynamic response in *individual voxels* that distinguishes between driving experimental manipulations which directly excite neurodynamics and inputs which modulate the neurodynamics. Spatio-temporal methods that incorporate unknown hemodynamics and with spatial regularization of the voxel-wise activity estimates using Markov random field (MRF) priors have also been reported (Makni, Beckmann, Smith, & Woolrich, 2008; Quirós, Diez, & Wilson, 2010; Ou, Wells, & Golland, 2010). Activation detection for known stimuli has also been effected using hidden Markov multiple event sequence models (HMMESM) (Faisan, Thoraval, Armspach, & Heitz, 2007), that pre-process the raw time-series into a series of spikes to infer neural events at each voxel. Univariate testing of an individual voxel's time-series data for the occurrence of small set of pre-specified "neural processes" relative to some external event, called hidden process model (HPM), has also been used in the analysis of fMRI data (Hutchinson, Niculescu, Keller, Rustandi, & Mitchell, 2009). In contrast to these methods, the modeling used here deals with the problem of *inferring a global and temporally evolving "brain-state"* in terms of the patterns present in the distribution of neural activity.

2. State Space Model

The concept of the functional brain transition through a brain-state-space is depicted in Fig. 1. The probability $\Pr[x_t = k]$ that the brain-state x_t at time $t = 1 \dots T$ (in TR units) is $k = 1 \dots K$ depends on not only on the previous state of the brain x_{t-1} but also on the current experimental stimulus described by the vector $\mathbf{s}_t$. In the discussion that follows, the fMRI volume $\mathbf{y}_t$ at each time-point t is represented in a D -dimensional linear feature-space, to be described later in Section 2.3. Also, the $^\top$ symbol indicates matrix transpose throughout the paper.

The multinomial transition probability from $x_{t-1} = i$ to $x_t = j$ is:

$$\pi_{x_{t-1}, x_t}(\mathbf{s}_t) \triangleq p(x_t = j | x_{t-1} = i, \mathbf{s}_t, \mathbf{w}) = \frac{\exp\{\mathbf{s}_t^\top (\omega_j + \mathbf{w}_{i,j})\}}{\sum_{k=1}^K \exp\{\mathbf{s}_t^\top (\omega_k + \mathbf{w}_{i,k})\}} \quad (1)$$

The marginal probability of being in state j at time t , given stimulus $\mathbf{s}_t$ is proportional to $\exp\{\mathbf{s}_t^\top \omega_j\}$, while the probability of transitioning from state i at time $t-1$ to state j at time t is proportional to $\exp\{\mathbf{s}_t^\top \mathbf{w}_{i,j}\}$. The parameter $\mathbf{w}_{i,j}$ is equipped with prior distribution $\mathcal{N}(0, \lambda_{\mathbf{w}}^{-1} \mathbf{I})$ which biases its value towards zero. The precision hyper-parameter $\lambda_{\mathbf{w}}$ implements a trade-off between the influence of the current stimulus $\mathbf{s}_t$ and the previous state x_{t-1} on the probability of the current state x_t . The higher the value of $\lambda_{\mathbf{w}}$, the stronger is the shrinkage of the estimates of $\mathbf{w}_{i,j}$ towards zero, enhancing the influence of ω_j on the current state probability, while reducing the effect of the previous state $x_{t-1} = i$ on $p(X_t = j | X_{t-1} = i)$.

In addition to observed (known) experimental stimuli $\mathbf{s}_t$ it is possible to include epochs, T' frames (i.e. TR units) in duration, of unobserved (hidden/missing) stimuli $\mathbf{u}_t$ which are in turn estimated from the fMRI data. By treating epochs of the experiment as having missing

stimuli in the model, computing their values from the data, and then measuring the error with respect to the known (recorded) stimuli, we equip the model with one level of prediction capability. A second level of prediction, as explained later in Section 4, is possible by correlating the brain-state sequence with the stimuli. These error-rate can then be used to validate different aspects of the modeling process, such as the quality of the fit, whether the model has correctly identified task related spatio-temporal patterns, and whether particular experimental variables demonstrate an effect in the data by changes in the predictability of the corresponding time-points.

Each abstract brain-state $k = 1 \dots K$ is associated with a characteristic spatial distribution of neural¹ activity in the cortex μ_k . Each occurrence x_t of the brain-state k at time t , the (spatially distributed) neural activation $\mathbf{z}_t$ is a realization from the the multivariate normal distribution $\mathcal{N}(\mu_k, \Sigma_k)$, where Σ_k is a covariance matrix describing the possible modes of variation in the realization $\mathbf{z}_t$ from the idealized activation μ_k . Let $\vartheta \triangleq \{\vartheta_1 \dots \vartheta_k\}$, where $\vartheta_k \triangleq (\mu_k, \Sigma_k)$.

The transformation of the activation pattern $\mathbf{z}_t$ into the observed fMRI signal $\mathbf{y}_t$ is modeled by a element-wise linear convolution $\mathbf{y}_t[d] = \sum_l \mathbf{h}_l[d] \mathbf{z}_{t-l} + \varepsilon_t[d]$, for $d = 1 \dots D$, Here D is the dimension of the feature-space and $\mathbf{h}[d] \triangleq (\mathbf{h}_0[d] \dots \mathbf{h}_L[d])$ is a finite impulse response (FIR) filter of length $L+1$, typically set to correspond to an HRF of 32s. Each $\mathbf{h}[d]$, representing the *unknown* hemodynamic response function (HRF) at element d , is equipped with a normal prior $\mathcal{N}(\mu_{\mathbf{h}}, \Sigma_{\mathbf{h}})$ constructed by varying the delay, dispersion and onset parameters of the canonical HRF (Boynton, Engel, Glover, & Heeger, 1996) and computing the mean and variance of these HRFs. This prior enforces domain knowledge by restricting the HRF to the space of physiologically plausible shapes thereby preventing over-fitting of the FIR filter to the data.

The set of HRF parameters is then the $D \times L$ matrix $\mathbf{h} \triangleq (\mathbf{h}[1]^\top \dots \mathbf{h}[D]^\top)^\top$. If we define the matrix $\mathbf{H}_l \triangleq \text{diag}\{(\mathbf{h}_l[1] \dots \mathbf{h}_l[D])^\top\}$, then we can express the fMRI observation vector $\mathbf{y}_t = \sum_l \mathbf{H}_l \mathbf{z}_{t-l} + \varepsilon_t$. Here, $\varepsilon_t \sim \mathcal{N}(0, \Sigma_\varepsilon)$ models an additive spatially normal and temporally i.i.d. noise component in the signal. Note that the HRF $L+1$ TRs long induces a correlation in frames $\mathbf{y}_{t, \dots, t+L}$ due to activation $\mathbf{z}_t$ at time t .

Denoting the set of parameters $\theta \triangleq \{\mathbf{u}, \mathbf{w}, \vartheta, \mathbf{h}, \Sigma_\varepsilon\}$, the full probability model is (§ Fig. 1):

$$p_\theta(\mathbf{y}, \mathbf{z}, x) \triangleq p(\mathbf{y}, \mathbf{z}, x, \mathbf{w}, \mathbf{h} | \mathbf{s}, \mathbf{u}, \vartheta, \sum_\varepsilon) = p(x, \mathbf{w} | \mathbf{s}, \mathbf{u}) p(\mathbf{z} | x, \vartheta) p(\mathbf{y}, \mathbf{h} | \mathbf{z}, \sum_\varepsilon), \quad (2)$$

where

¹More accurately, it is associated with a spatially distributed estimate of the metabolic activity that produces the hemodynamic response measured by the BOLD signal.

$$\begin{aligned}
p(x, \mathbf{w} | \mathbf{s}, \mathbf{u}) &= p(\mathbf{w}) \left[\prod_{t \in \text{observed}} \pi_{x_{t-1}, x_t}(\mathbf{s}_t) \prod_{t \in \text{missing}} \pi_{x_{t-1}, x_t}(\mathbf{u}_t) \right], \\
p(\mathbf{z} | x, \vartheta) &= \prod_{t=1}^T p(\mathbf{z}_t | x_t, \vartheta_{x_t}), \\
p(\mathbf{y}, \mathbf{h} | \mathbf{z}, \Sigma_\varepsilon) &= p(\mathbf{h}) \left[\prod_{t=1}^T p(\mathbf{y}_t | \mathbf{z}_{t-L \dots t}, \mathbf{h}, \Sigma_\varepsilon) \right].
\end{aligned}$$

This layered Markov chain model of a mental process serves three important purposes: a) To enforce a temporal structure on the ordering of the states; b) To model higher order interactions between stimuli and mental states, by aggregating the history of the experiment into the current state; and c) To decouple the stimulus from the fMRI signal through the latent neural activation layer, thereby obviating specification of the exact mathematical relationship between the two. This is in contrast to alternative methods, such as univariate general linear models (Worsley et al., 2002), multivariate linear models (McIntosh & Lobaugh, 2004; K. Friston et al., 2008), and even non-linear models (Krugger, Zysset, & Cramon, 2000; K. J. Friston, Mechelli, Turner, & Price, 2000), which assume a direct relationship between the value of the stimulus and the signal, and which have only a limited ability to capture the modulatory effects of past stimuli on the current mental state.

2.1. Parameter Estimation

As described in Janoos et al. (2011), the parameters of the model $\theta^* = \arg \max_{\theta} p_{\theta}(\mathbf{y})$ are estimated using a generalized Expectation Maximization (GEM) algorithm under the *mean-field approximation*. Introducing a variational density $q(\mathbf{z}, x)$ over the latent variables $\mathbf{z}, x$, the log-probability of eqn. 2 is decomposed into a free-energy and a KL-divergence as: $\ln p_{\theta}(\mathbf{y}) = \mathcal{Q}(q, \theta) + KL(q || p_{\theta})$, where:

$$\begin{aligned}
\mathcal{Q}(q, \theta) &= \sum_x \int_{\mathbf{z}} q(\mathbf{z}, x) \ln \frac{p_{\theta}(\mathbf{y}, \mathbf{z}, x)}{q(\mathbf{z}, x)} d\mathbf{z} \\
\text{and } KL(q || p_{\theta}) &= - \sum_x \int_{\mathbf{z}} q(\mathbf{z}, x) \ln \frac{p_{\theta}(\mathbf{z}, x | \mathbf{y})}{q(\mathbf{z}, x)} d\mathbf{z}. \quad (3)
\end{aligned}$$

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

$$\begin{aligned}
\text{E-step} \quad & q^{(n)} \leftarrow \arg \min_q KL(q || p_{\theta^{(n)}}) \\
\text{M-step} \quad & \theta^{(n+1)} \leftarrow \theta \quad \text{such that} \quad \mathcal{Q}(q^{(n)}, \theta) > \mathcal{Q}(q^{(n)}, \theta^{(n)}). \quad (4)
\end{aligned}$$

Although the exact minimizer of $KL(q || p_{\theta^{(n)}})$ is $p_{\theta^{(n)}}(\mathbf{z}, x | \mathbf{y})$, this conditional density cannot be practically used in the M-step, for the following reason. The linear hemodynamic filter spreads the activation at $\mathbf{z}_t$ over observations $\mathbf{y}_t \dots \mathbf{y}_{t+L}$. Using d -separation rules (Bishop, 2007) to examine the resulting graphical model, we see that when we condition on the observation sequence $\mathbf{y}$, every $(x_t, \mathbf{z}_t)$ latent variable pair becomes dependent on every other $(x_{t'}, \mathbf{z}_{t'})$ pair in a window of $t-L \dots t+L$. Therefore, the posterior density of $p_{\theta^{(n)}}(\mathbf{z}, x | \mathbf{y})$ does not have *any* factorization over the latent variable pairs $(x_t, \mathbf{z}_t)$, $t = 1 \dots T$, and setting

$q(\mathbf{z}, x) = p_{\theta}(\mathbf{z}, x|\mathbf{y})$ would require joint integration over all T variables. Although the $\mathbf{z}$ layer, consisting of multivariate normal variables can be integrated in closed form, integrating the state layer x would require a computationally prohibitive joint summation of $\mathcal{O}(K^T)$ complexity.

The *mean-field approximation* (Bishop, 2007) allows eliminating this expensive marginalization by restricting q to the family of factorizable distributions

$q(\mathbf{z}, x) = \prod_{t=1}^T q_t(\mathbf{z}_t, x_t) = \prod_{t=1}^T q_t(\mathbf{z}_t|x_t)q_t(x_t)$. Under this restriction, the minimizer of $KL(q||p_{\theta})$ can be shown to be $q_t^*(\mathbf{z}_t, x_t) \propto \exp\{\mathbb{E}_{\prod_{t' \neq t} q_{t'}^*}[\ln p_{\theta}(\mathbf{y}, \mathbf{z}, x)]\}$. Each factor $q_t^*(\mathbf{z}_t, x_t)$ evaluates to a product $q_t^*(\mathbf{z}_t|x_t)q_t^*(x_t)$ of a multinomial logistic probability $q_t^*(x_t)$ of state x_t and a normal density $q_t^*(\mathbf{z}_t|x_t)$ of the activation pattern $\mathbf{z}_t$ conditioned on x_t . Consequently, the n -th iteration of the E-step involves computing the factorizable density $q^{(n)}(\mathbf{z}, x) = \prod_{t=1}^T q_t^{(n)}(\mathbf{z}_t, x_t)$ by first initializing q_t with the posterior distribution of $p_{\theta}^{(n-1)}(\mathbf{z}_t, x_t|\mathbf{y})$, and then updating q_t^* until convergence. Using a combination of closed-form and iterative optimization, the M-step is then guaranteed to find estimates of the parameters such that $p_{\theta}^{(n)}(\mathbf{y}) = p_{\theta}^{(n-1)}(\mathbf{y})$, yielding a locally optimal solution θ^* . The iterations are terminated when the updates to θ fall below a pre-specified tolerance (adaptively set at 1% of the absolute value of the parameter in the n -th iteration). The reader is referred to (Janoos et al., 2011) for details on this procedure.

Direct estimation of the most probable sequence of states $x^* = \arg \max_x \ln p_{\theta}(x|\mathbf{y}) = \arg \max_x \ln p_{\theta}(\mathbf{y}, x)$, given an estimate of SSM parameters θ , again requires joint maximization over all T state variables $x_1 \dots x_T$, as the latent $\mathbf{z}$ layer introduces a dependency between all the $\mathbf{y}$ and x variables preventing any factorization of the joint probability density function. To address this computational impossibility, an EM algorithm under a mean field approximation is used. Specifically, the log-probability term $\ln p_{\theta}(\mathbf{y}, x)$ is decomposed into a free-energy and a KL-divergence term $\ln p_{\theta}(\mathbf{y}, x) = \mathcal{Q}(q, x) + KL(q||p_{\theta}(\mathbf{z}|\mathbf{y}, x))$, where

$$\mathcal{Q}(q, x) = \int_{\mathbf{z}} q(\mathbf{z}) \ln \frac{p_{\theta}(\mathbf{y}, \mathbf{z}, x)}{q(\mathbf{z})} d\mathbf{z} \text{ and } KL(q||p_{\theta}(\mathbf{z}|\mathbf{y}, x)) = - \int_{\mathbf{z}} q(\mathbf{z}) \ln \frac{p_{\theta}(\mathbf{z}|\mathbf{y}, x)}{q(\mathbf{z})} d\mathbf{z},$$

by introducing a variational density restricted to the factorizable family $q(\mathbf{z}) = \prod_{t=1}^T q_t(\mathbf{z}_t)$. After a fixed-point evaluation of $q_t^*(\mathbf{z}_t)$ given an estimate of $x^{(n)}$ as per $q_t^*(\mathbf{z}_t) = \exp\{\mathbb{E}_{\prod_{t' \neq t} q_{t'}^*}[\ln p_{\theta}(\mathbf{y}, \mathbf{z}|x^{(n)})]\}$, the M-step takes the form:

$$\begin{aligned} \arg \max_x \mathcal{Q}(q^{(n)}, x) &= \arg \max_x \left\{ \ln p_{\theta}(x) + \int_{\mathbf{z}} q^{(n)}(\mathbf{z}) \ln p_{\theta}(\mathbf{z}|x) d\mathbf{z} \right\} \\ &= \arg \max_x \left\{ \sum_{t=1}^T \ln p_{\theta}(x_t|x_{t-1}) + \int_{\mathbf{z}_t} q^{(n)}(\mathbf{z}_t) \ln p_{\theta}(\mathbf{z}_t|x_t) d\mathbf{z}_t \right\}, \end{aligned}$$

which is identical to the problem of estimating the optimal state-sequence of a first-order HMM, with the difference that the emission (log) probability is replaced by the expected (log) probability under the variational density $q^{(n)}$, and which can be efficiently computed using the Viterbi method (Bishop, 2007). The optimal state sequence $x^{(n)}$ computed in the n -th iteration is then used to refine the mean-field approximation for the $n+1$ -th iteration and the procedure is terminated when the increments $|\ln p_{\theta}(\mathbf{y}, x^{(n+1)}) - \ln p_{\theta}(\mathbf{y}, x^{(n)})|$ fall below a pre-specified tolerance.

2.2. Hyper-Parameter Selection

The hyper-parameters of the SSM are the number of hidden states K and the precision $\lambda_{\mathbf{w}}$ of the prior distribution of the transition weights $\mathbf{w}$ (§ eqn. 1). The hyper-parameter $\lambda_{\mathbf{w}}$ mediates a trade-off between the temporal structure of the fMRI data and that of the stimulus sequence, with a very high value causing the state-transitions to be driven mostly by the current stimulus, and a low value increasing the contribution of the previous state to the transition probability. The other hyper-parameter K controls the complexity of the SSM, where a greater K permits fitting of a larger variety of patterns in the data, at the expense of over-fitting noise.

While model-size selection can be performed using Bayesian or information theoretic criteria (Lanternman, 2001), here an *empirical risk minimization* method for selecting both K and $\lambda_{\mathbf{w}}$ is adopted by leveraging the ability of the SSM to predict missing stimuli $\mathbf{u}$. From the stimulus time-series, blocks of T' consecutive time-points (in TR units) totalling to 25% of the total number of frames, are removed at random to serve as missing stimuli and the optimal SSM parameters θ^* are estimated for a given K and $\lambda_{\mathbf{w}}$. The prediction error (i.e. risk) is then measured as $\mathfrak{R} = \sum_{i \in \text{missing}} \|\mathbf{u}_t - \mathbf{s}_t\|_2$, between the predicted $\mathbf{u}_t$ and their t true values $\mathbf{s}_t$. By selecting hyper-parameters that minimize this empirical estimate of risk $\mathfrak{R}$, the model most relevant to the experiment being conducted is identified.

The optimal value of K is obtained by first stepping through large range of values with large step-sizes and then iteratively refining the step-size. For each setting of K , the optimal $\lambda_{\mathbf{w}}$ is determined by searching over the range $\log_{10} \lambda_{\mathbf{w}} = -3 \dots +3$ for the value that minimizes $\mathfrak{R}$. This allows setting the parameter to effect an optimal compromise between between stimulus driven and previous state driven transitions. We have observed that the risk $\mathfrak{R}$ is relatively insensitive to $\lambda_{\mathbf{w}}$ (§ Section 3.7), and therefore a common value is usually selected across a multi-subject data-set.

It must be noted that the prediction risk is used only as a statistic to select hyper-parameters. The parameters θ themselves, unlike MVPR classifiers, are *not estimated to minimize prediction error* but rather to fit the *generative model* to the data with model-evidence $\ln p_{\theta}(\mathbf{y})$ as the optimization target.

2.3. Feature Space

Each volume of an fMRI study typically contains on the order 10^5 voxels, making dimensionality reduction a prerequisite of any multivariate analysis. Dimensionality reduction via principle components analysis (PCA) has been reported to be unsuitable by

multiple studies (O'Toole et al., 2007; Mitchell et al., 2008), as the largest variance principal components usually correspond to motion and other types of physiological noise. ICA (Calhoun & Adali, 2006), on the other hand, does not provide a natural ordering of the components and therefore no principled way of selecting among them. Alternatives based on using voxels in pre-specified regions-of-interest (ROIs) (Haxby et al., 2001) disregard information contained in the other regions, and are unsuitable for an integrated analysis of the brain as performed here.

The feature-space used here was developed from a different motivation: namely that the SSM is essentially a probabilistic framework for clustering time-points in an fMRI session based on the underlying patterns metabolic activity and therefore like any clustering method, its efficacy depends on the measure of the similarity between the activation patterns at two time-points (Jain, 2010). Janoos et al. (2011) proposed a distance metric called the *functional distance* based on the concept of the cost of minimum “transport” of neural 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, Guo, Nie, Li, & Liu, 2009), and are based on the supposition that brain regions that have higher temporal correlation are functionally more related than regions with lower correlation. The cost of transport of activity from one brain location to the next is (inversely) related to the temporal correlation between these locations. This metric captures the concept that two time-points that exhibit neural activity in brain regions that are more correlated are functionally more similar than two time-points with activity in uncorrelated regions (See Fig. 2 for an illustration.).

Formally, if N is the number of voxels in the cortex, then the functional distance between neural activation patterns $\mathbf{z}_{t_1}$ and $\mathbf{z}_{t_2}$ is:

$$FD(\mathbf{z}_{t_1}, \mathbf{z}_{t_2}) \triangleq \min_f \sum_{i=1}^N \sum_{j=1}^N f[i, j] d_F[i, j], \quad (5)$$

subject to the constraints: $f[i, j] \geq 0$, $\sum_j f[i, j] = \mathbf{z}_{t_1}[i]$, $\sum_i f[i, j] = \mathbf{z}_{t_2}[j]$, and $\sum_{i,j} f[i, j] = \min \{ \sum_i \mathbf{z}_{t_1}[i], \sum_j \mathbf{z}_{t_2}[j] \}$. The transport $f[i, j]$ from voxel i to j is penalized by the “disconnectivity” $d_F[i, j] > 0$ between voxels i and j , complementary to their functional connectivity².

²The functional connectivity of two voxels $0 \leq F[i, j] \leq 1$ is defined by the magnitude of the temporal correlations between their fMRI time-series. The disconnectivity between two voxels i and j is then defined by the difference $d_F[i, j] = |\phi^*[i] - \phi^*[j]|$, where ϕ^* is an 1-dimensional equivalent of the graph represented by F that best preserves its adjacency structure (i.e. its *distortion minimizing 1-d embedding* (Chung, 1997)):

$$\phi^* = \arg \inf_{\phi} \frac{\sum_i \sum_j (\phi[i] - \phi[j])^2 F[i, j]}{\sum_i \phi[i]^2 D_F[i, i]}, \quad \text{subject to } \phi \perp D_F \mathbb{1}.$$

In this equation, $D_F[i, i]$ is the *diagonal degree matrix*, defined as $D_F[i, i] = \sum_j F[i, j]$ and $D_F[i, j] = 0$, $\forall i \neq j$. Intuitively, this means that the cost function $d_F[i, j]$ between voxels i and j was calculated from the connectivity between them, by finding a 1-d vector ϕ^* that assigns more similar values to voxels “closer” to each other (in the sense of functional connectivity) and more different values to voxels “farther” from each other.

This functional distance, which is posed in a linear programming formulation, is tightly approximated by an orthogonal linear basis $\Phi = \{\phi^{(l,m)}\}$ obtained through recursive decomposition³ of the normalized graph Laplacian (Chung, 1997) $\mathcal{L} = D_F^{-\frac{1}{2}} (D_F - F) D_F^{-\frac{1}{2}}$ of the voxel-wise functional connectivity matrix F . Here, the index $m = 0 \dots \log_2 N - 1$ gives the level of decomposition, $l = 0 \dots 2^m - 1$ indexes the basis vectors within level m , and D_F is the diagonal degree matrix of F . The nature of this approximation is given by the following theorem:

Theorem 1—Let $\delta Z[l, m]$ be coefficients of $\delta \mathbf{z} = \mathbf{z}_{t_1} - \mathbf{z}_{t_2}$ in the basis Φ , i.e. $\delta Z[l, m] = 2^{-m} \langle \delta \mathbf{z}, \phi^{(l,m)} \rangle$, $l = 0 \dots 2^m - 1$, $m = 0 \dots \log_2 N - 1$. Then, there exist constants $M_{0,0}$, $\hat{M}_{0,0} > 0$, such that

$$\hat{M}_{0,0} \sum_{m=0}^{\log_2 N - 1} \sum_{l=0}^{2^m - 1} |\delta Z[l, m]| \leq \text{FD}(\mathbf{z}_{t_1}, \mathbf{z}_{t_2}) \leq M_{0,0} \sum_{m=0}^{\log_2 N - 1} \sum_{l=0}^{2^m - 1} |\delta Z[l, m]| \quad (6)$$

and the tightness of this bound is:

$$\sup_{\|\delta \mathbf{z}\|_2=1} \left[M_{0,0} \sum_m \sum_l |\delta Z[l, m]| - \hat{M}_{0,0} \sum_m \sum_l |\delta Z[l, m]| \right] \approx \frac{(M_{0,0} - \hat{M}_{0,0})}{\sqrt{2}} \quad (7)$$

Therefore, the functional distance $\text{FD}(\mathbf{z}_{t_1}, \mathbf{z}_{t_2})$ is approximated (up to a multiplicative constant) by projecting the spatially distributed neural activity patterns $\mathbf{z}_{t_1}$ and $\mathbf{z}_{t_2}$ on to the orthogonal basis vectors (which are also spatial maps) and computing the Euclidean (ℓ_2) distance metric in this space,

$$\Delta(\mathbf{z}_{t_1}, \mathbf{z}_{t_2}) = \left(\sum_{l,m} |\delta Z[l, m]|^2 \right)^{\frac{1}{2}}. \quad (8)$$

The reader is referred to Janoos et al. (2011) for the recursive decomposition algorithm to generate Φ and the proof of the theorem. This feature space has the following benefits: (a) It is derived from a neurophysiologically plausible definition of distance between spatial maps of activity, and therefore is more suitable basis for clustering fMRI data, as compared to a Euclidean metric in voxel space; (b) As a Euclidean space, it has a well defined metric structure with simple definitions of concepts such as linear-space operations, probability densities, expectations and variances. Therefore, it can be easily used in the state-space model framework. (c) Although the feature-space is defined on neural activation patterns (i.e. $\mathbf{z}$), computations can be performed by directly projecting each fMRI time-point $\mathbf{y}_t$ into F . This is because the fMRI data $\mathbf{y}$ is related to neural activity $\mathbf{z}$ through a linear convolution

³The orthogonalization of the graph Laplacian is performed by recursively subdividing the domain of the graph as follows. The first basis vector $\phi^{(0,0)}$ is computed from the eigenvector of $\mathcal{L}^{(0,0)} = \mathcal{L}$ corresponding to the second smallest eigenvalue. Next, the graph is partitioned into two sub-graphs based on the sign of $\phi^{(0,0)}$, and their graph Laplacians $\mathcal{L}^{(1,1)}$ and $\mathcal{L}^{(2,1)}$ are recomputed. The next two basis vectors $\phi^{(1,1)}$ and $\phi^{(2,1)}$ are the second smallest eigenvectors of the $\mathcal{L}^{(1,1)}$ and $\mathcal{L}^{(2,1)}$, respectively. The process is repeated until only one voxel is left in the partition.

filter $\mathbf{h}$ representing hemodynamic response, and linear projection operator commutes with linear convolution.

Dimensionality reduction of the feature space is performed by selecting only stable basis vectors among all $\Phi = \{\phi^{(l,m)}\}$, $m = 0 \dots \log_2 N - 1$, $l = 0 \dots 2^m - 1$, assessed using a bootstrap analysis of stability (Bellec, Rosa-Neto, Lyttelton, Benali, & Evans, 2010). The bootstrap generates an estimate of the sampling distribution (i.e. *bootstrap distribution*) of the functional connectivity matrix F by resampling, with replacement, entire blocks of fMRI scans from a session 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 (Bellec et al., 2010). We found $T/M \approx 5TRs$ to be adequate for our data-sets.

The stability of a particular basis vector $\phi^{(l,m)}$ is defined by its correlation coefficient $\rho^{(l,m)}(r_1, r_2) \triangleq |\langle \phi_{(r_1)}^{(l,m)}, \phi_{(r_2)}^{(l,m)} \rangle|$ across resamples r_1, r_2 of the data. Given the bootstrap distribution of correlations $\Pr_{boot}[\rho^{(l,m)}(r_1, r_2)]$, a vector $\phi^{(l,m)}$ is said to be τ_Φ -stable if $\Pr_{boot}[\rho^{(l,m)}(r_1, r_2) \geq \tau_\Phi] \geq 0.75$, i.e. the correlation between at least 75% of the resamples of $\phi^{(l,m)}$ is greater than the threshold $0 \leq \tau_\Phi \leq 1$. If $\phi^{(l,m)}$ is not τ_Φ -stable, then it is discarded, which also removes all the vectors obtained from the subdivision of $\phi^{(l,m)}$. Therefore, increasing the value of τ_Φ causes an exponential increase in the number of vectors that are removed, as reported in Section 3.6.

Because the basis vectors are derived from a definition of the functional similarity between activation patterns, and dimensionality reduction is performed via bootstrap analysis of stability, this feature-space has been shown to have a high degree of robustness to noise and to artifacts of non biological origin in the signal (Janoos et al., 2011).

2.4. Summary

An overview of the procedure for the SSM and the feature space computation, as described in the preceding sections, is outlined in Fig. 3.

3. Methods and Materials

3.1. Sternberg Item Recognition Paradigm (SIRP)

Each 360-second run of the SIRP task (§ Fig. 4) was composed of six blocks of memory items separated by fixation blocks (Brown et al., 2009). During a run, two memory sets for each memory set size were presented twice. Stimuli were visually presented using the projection method standard at each site. Each memory block was composed of a Prompt, an Encode and a Probe epoch. The prompt cued subjects to learn the subsequent memory set composed of 1, 3, or 5 digits presented during the Encode epoch. The Probe epoch was composed of 7 positive probes (Target_Probe), where the presented digit was a member of the memory set, randomly intermixed with 7 negative probes (Foil_Probe). Each memory

set size was presented twice per run, and three runs were acquired for each subject with varied ordering of the memory set sizes and different encode digits.

3.2. Subjects

In this study, data from 4 of the 8 sites participating in the fBIRN Phase II study were used in order to restrict the analysis to scanners with higher field strength (3T/4T) and to include two GE and two Siemens sites to balance for vendor type. The sites were: (a) Brigham and Women's Hospital in Boston MA (3T GE); (b) Massachusetts General Hospital in Boston, MA (3T Siemens); (c) UCLA in Los Angeles, CA (3T Siemens); and (d) Duke University in Durham, NC (3T/4T GE). At these sites, a cohort of 41 subjects diagnosed with schizophrenia or schizoaffective disorder according to the *Diagnostic and Statistical Manual of Mental Disorders DSM-IV-TR Fourth Edition (Text Revision)* (2000) and 41 healthy comparison subjects, matched group-wise on age and gender to the patients were recruited. Subjects were between the ages of 18 and 70 years, fluent in English, had normal hearing and normal or correctable eyesight. Individuals were excluded from the study if they had a contraindication to MRI, a current major neurological or medical illness, previous head injury with loss of consciousness greater than 30 minutes, substance and/or alcohol dependence, an estimated IQ score less than 75 as measured by the North American Adult Reading Test (NAART), or were being treated for migraines. Schizophrenia patients were required to be clinically stable with no significant changes in their psychotropic medications in the previous two months. Comparison subjects were excluded if they had a current or past psychiatric disorder or if a first-degree family member had a diagnosis of a psychotic illness. All sites received local institutional review board approval for this study.

3.3. Imaging

Images were acquired at the four sites using two 3T Siemens, two 3T General Electric and one 4T General Electric MRI scanners running an imaging protocol prescribed to be as similar as possible. A T_2 -weighted structural scan was obtained in the oblique axial, anterior commissure - posterior commissure (AC-PC) plane using a fast spin echo or turbo spin echo sequence (FOV = 22cm, 27 slices, 4mm slice with 1mm skip, TR/TE = 5000/68 ms, ETL = 12 or 13, 256×192 matrix, total scan time = 1:30 min).

BOLD data were obtained using T_2^* -weighted sequence (EPI or spiral single-shot gradient echo protocol, scan plane = oblique axial AC-PC, FOV = 22cm, 27 slices, 4mm slice thickness with 1mm skip, TR = 2000 ms, TE = 30 ± 2 ms (3T/4T), 40 ± 2 ms (1.5T), flip angle = 90° , 64×64 matrix) while participants performed 3 runs of the SIRP. For each of the three runs, 3 dummy frames were acquired followed by 177 active frames. Refer to Brown et al. (2009) for more details.

3.4. Pre-processing

The data was each subject were pre-processed independently. Pre-processing for motion correction, B_0 field inhomogeneity correction, slice-timing correction along with skull stripping were performed using a developmental version of the fBIRN Image Processing System (FIPS) (Brown et al., 2009), an image analysis pipeline primarily using routines from the FMRIB Software Library (FSL) (Smith et al., 2004). Next, the time-series data

were high-pass filtered (0.5Hz) to remove artifacts due to breathing, blood pressure changes and scanner drift. Residual motion artifacts were removed by performing an ICA analysis of subject's data (Calhoun & Adali, 2006) and removing components highly correlated with the estimated motion parameters. Also, components that could be identified as obvious artifacts through visual inspection were removed. The data of each subject were then normalized into the ICBM-152 atlas space using FSL by first co-registering the mean EPI volume with the high resolution T_2 scan and then registering the T_2 scan with the atlas while keeping the EPI data in co-register. The mean volume of the time-series was subtracted, white matter masked out and all further processing was performed on grey matter voxels.

3.5. Quality Assurance

Quality assurance was performed in three steps. First, subjects were identified as “not engaged in the task” if they did not perform better than chance, failed to respond on several consecutive trials or seemed to have mixed up the response buttons. In the second step, data were screened based on image quality. Two aspects of the EPI images were evaluated: (a) A visual inspection was made to identify coarse visual distortions (eg., a large cyst); (b) the Jaccard similarity index was used to measure the overlap between the normalized EPI image and the ICBM-152 atlas. Low values meant the normalization was inadequate. In the third step, the quality of the time-series data were screened based on two criteria: (a) The time-series exhibited excessive spiking; (b) A large portion ($> 60\%$) of the total energy of the data-set was in ICA components marked as artifactual.

All three dimensions were judged as “adequate” or “inadequate” for each subject across all three runs, and the data marked as inadequate were discarded. Out of the initial group of 82, a total of 69 subjects (36 controls and 33 patients) were retained. The site-wise breakdown of subjects is shown in Table 1.

3.6. Feature-Space Embedding

For each subject $j = 1 \dots 69$, the feature-space basis vectors Φ_j were computed as described in Section 2.3 after concatenating the data from the three runs and then the fMRI volumes were projected into this low-dimensional space.

The effect of the stability threshold τ_Φ on the dimensionality D for one control subject is shown in Fig. 5. Initially there is a steep reduction in dimensionality from $\mathcal{O}(10^5)$ to $\mathcal{O}(10^2)$. However, after a certain value of τ_Φ , the reduction slows down significantly. As shown in Fig. 5, an inflection point (i.e. a maximum in the curvature of the D vs. τ_Φ curve after smoothing with a Gaussian filter of FWHM=0.2) can be observed. Also shown is the largest index m of the vectors $\phi^{(l,m)}$ retained for a given τ_Φ , 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 eqn. 6, implies that the effect of reduced dimensionality of Φ on the approximation error is small, as most of the discarded vectors have a large index m . The inflection point was typically between $0.55 \leq \tau_\Phi \leq 0.65$ across all subject populations and sites, corresponding to a reduction in dimensionality from $N = 110592$ to $390 \leq D \leq 480$. Therefore, τ_Φ was adaptively set such that $D = 440$ for all subjects.

3.7. State Space Modeling

Each run was $T = 177$ TRs long. The stimulus vector was coded as $\mathbf{s}_t = (\text{block_num}, 1)$, where $\text{block_num} = 1 \dots 6$ gives the presentation block number to which the time point $t = 1 \dots 177$ belongs. Under the assumption that the fMRI time-series data $\mathbf{y}^{(r)}$ for a subject to be conditionally independent across each run $r = 1, 2, 3$, i.e. $p(\mathbf{y}^{(r1)}, \mathbf{y}^{(r2)}) = p(\mathbf{y}^{(r1)})p(\mathbf{y}^{(r2)})$, it can be easily seen that the overall parameter estimates are obtained by averaging across the three runs during the estimation steps. The optimal state space model $\mathcal{M}_j = \{\theta, K, \lambda_w\}$ was estimated by first aggregating the parameter estimated across the three runs, and then selecting optimal hyper-parameters, as described in Section 2.2.

Fig. 6 graphs the error in predicting the missing stimulus (i.e. risk $\mathfrak{R}$) with respect to the two hyperparameters K, λ_w , for the control and patient groups at site 4.

From the plot of $\mathfrak{R}$ versus model-size K in Fig. 6(a) the bias-variance tradeoff typical to the model selection problem can be observed. For K much less than the true value, the estimated model is too simple and may not capture the true variation in the data, while for K much greater than the true value, the model may over-fit noise compromising its predictive power. However, relatively shallow basin of the risk curve, with minima occurring in the range $K = 12 \dots 25$ points to the robustness of the model with respect to K . While the curves for the other sites were qualitatively similar, we observed both a group and site-specific effect on the minimum value. This is further discussed in Section 4.2.

The plot of $\mathfrak{R}$ versus precision hyper-parameter λ_w in Fig. 6(b) 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 λ_w) fails to utilize the temporal structure in the mental processes and hence poorly estimates the correct brain-state corresponding to a missing stimulus. On the other hand, the SSM with transitions driven mainly by the previous state (i.e. low λ_w) does not capture the effect of external stimuli on mental processes, compromising its predictive power. However, as in the previous case, $\mathfrak{R}$ shows near optimal levels over a large range of values $10^0 \leq \lambda_w \leq 10^2$, again indicating a high degree of robustness with respect to it. Unlike the case of K , no significant group or site based effect was identified for the optimal λ_w , at $p < 0.3$ (two-sided). The p -value was calculated using Cramér's non-parametric two sample test (Baringhaus & Franz, 2004), and the multiple comparisons were corrected for using the false discovery rate (FDR) method.

4. Results

First, the ability of the state-space model (SSM) to predict the cognitive state of a subject from the fMRI data as compared to alternative MVPR methods is evaluated in Section 4.1. Next, the SSM results and the effects different experimental parameters on them are explored in more depth in Section 4.2. Then in Section 4.3, the ability of the SSM to compare the spatio-temporal patterns of neural activation, in their entirety, across subjects is used to study the similarities between the mental processes of patients and controls by projecting them into an abstract representation space.

4.1. Comparative Evaluation

To establish the baseline in the ability to predict the cognitive state of the subject using MVPR methods, we trained a subject-specific linear support vector machine (SVM) classifier (Pereira, Mitchell, & Botvinick, 2009) for the Target_Probe vs. Foil_Probe conditions. Classification was restricted to only the Probe epoch of each block (i.e. 6×38 secs), and the SVM was trained to predict a Target_Probe vs. Foil_Probe stimulus from the fMRI data deconvolved with the canonical HRF. Among the other classifiers evaluated (viz. GNB, LDA, quadratic and exponential SVM) none significantly outperformed the linear SVM.

In contrast with MVPR methods, the SSM estimates model parameters by maximizing the model-evidence, while prediction of *stimuli included in the model* is used for hyper-parameter selection. In the current case, only information about the presentation block (i.e. block_num) is provided as input to the model, and therefore it can only directly predict block_num. In order to use the SSM to predict a Target_Probe vs. Foil_Probe stimulus, a separate logistic regression classifier was trained on the marginal state distribution vector $\Pr[x_t = k/y, \theta], k = 1 \dots K$ at each time t . Prediction accuracy (i.e. 1 minus error rate) for both the SVM and SSM derived classifiers were computed using a 5-fold cross-validation scheme and are compiled in Table 3.

The site-wise distribution of comparative performance is shown Fig. 7. It can be observed that the overall prediction accuracy of the SSM is much higher, for both subject groups, than the SVM method. This is noteworthy especially since the SVM is specifically optimized for prediction accuracy. However, by treating each fMRI frame as an independent observation, the SVM does not utilize the temporal structure of the data, relying only on spatial patterns for prediction. Additionally, the SVM uses a point-estimate of the neural activation through deconvolution, whereas the SSM accounts for spatially varying and unknown hemodynamics in a probabilistic fashion which contributes to its superior predictive power. A second interesting observation is that the separation between the two groups, in terms of prediction accuracy, is much greater for the SSM method. Using Cramér's non-parametric test the reduction in prediction accuracy of the SSM based classifier was assessed significant at $p = 0.005$ (one-sided), while the separation between the two groups for the SVM classifier was not significant at any $p < 0.11$ (two-sided). For both methods, no site specific differences between their classification performance was observed (not significant at any $p < 0.25$, two-sided, FDR corrected).

The SSM results are analyzed in further detail in the next section.

4.2. SSM Analysis

The complexity of the model learnt for each subject is reflected in the optimal number of states K , which more states indicating a more complex model and possibly more variability (i.e. pattern templates) in the data. Fig. 8 shows the distribution of K with respect to the four sites and the two groups (control vs. patient). Here, an effect of subject type on the distribution of the optimal model complexity can be observed. The number of estimated states for the patients was significantly greater than that of the control group ($p = 0.001$, one-

sided Cramér's test, FDR corrected), while no site specific effect was observed (not significant for the control group at any $p < 0.35$ and patient group at $p < 0.27$, two-sided, FDR corrected).

To assess the fit of the model to the data and its ability to predict the cognitive state of the subject, a multinomial logistic classifier was trained on the marginal state distribution vector $\Pr[x_t = k/\mathbf{y}, \theta]$, $k = 1 \dots K$ at each time-point t with labels representing the different epochs with one presentation block (i.e. Prompt, Encode, Target_Probe and Foil_Probe). **Note:** In contrast to the accuracies reported when comparing the SSM against SVM in Section 4.1 (where prediction was restricted to the Probe epoch only), here prediction accuracy is reported for the entire run. Again, as no site-wise differences were observed in the overall prediction accuracy (plotted in Fig. 9) for both groups (not significant at $p < 0.22$ for control group and $p < 0.43$ for patient group, two-sided Cramér's test, FDR corrected), all further analysis were performed by pooling together the subjects across the sites. When pooled together, a significant reduction was observed in the ability of the model to predict the cognitive state of the patient group ($p = 0.001$, one-sided Cramér's test).

The differential predictability with respect to Probe type, i.e. the accuracy in predicting Target_Probe vs. Foil_Probe stimulus is of specific interest to the hypothesis of the comparatively greater disorganization in the working memory process of subjects with schizophrenia. From the performance of the SSM in predicting Target_Probe or Foil_Probe stimulus in Fig. 10, there are two salient observations: (a) an overall reduction in prediction accuracy for patients compared to controls, (b) a overall reduction in accuracy in predicting the Foil_Probe compared to the Target_Probe, with patients showing a larger reduction (0.15 ± 0.04) as compared to controls (0.13 ± 0.03). The difference between this reduction of prediction accuracy between the two groups was significant at $p = 0.01$ (one-sided Cramér's test).

Examining the effect of memory load (i.e. memory-set size of 1, 3, or 5 digits) on the SSM (§ Fig. 11), a reduction in the prediction accuracy with increase in memory set size was noticed. Moreover, there was no significant difference (for $p < 0.30$) in this reduction between the Target_Probe and Foil_Probe conditions for the control group along with the Target_Probe for the patient group (two-sided t -test on the slope of the regression line). However, the reduction in the prediction accuracy with respect to memory load in the patient group was significantly larger for the Foil_Probe condition ($p = 0.005$, one-sided t -test).

Another facet of working memory, important to the study of its impairment in SZ, is the change in the regularity of its neural representation over time as the task progresses. To study this progression, the 14 probe events in Probe epoch of the tasks were divided into three sets of 5, 5 and 4 events each. The effect of this task phase on prediction accuracy (averaged over three runs consisting of six presentation blocks each) is graphed in Fig. 12.

For the control group, there was a significant increase in prediction accuracy as the task progresses for both Probe types ($p = 0.001$, one-sided t -test) with greater improvement for Target_Probe, in terms of the slope of the regression line ($p = 0.01$, one-sided t -test). This implies that the spatio-temporal patterns during later presentations are more "typical" of the

Probe stimulus resulting in better predictability, while earlier presentations have more heterogeneity and do not conform to a single template. While an improvement was observed for the Target_Probe type ($p = 0.001$, one-sided t -test) in the patient group over task duration, the line is almost flat for the Foil_Probe stimuli (no effect at any $p < 0.23$, two-sided t -test). Therefore, the patterns of controls become more regular as the task progresses whereas patients continue to exhibit the same level of unpredictability when faced with a Foil_Probe.

4.3. Comparing Models

The treatment of mental processes as a sequence of abstract brain-states by the SSM can be used to compare the spatio-temporal patterns of subjects in this abstract representation space. To measure the similarity between the SSMs $\mathcal{M}_i$ and $\mathcal{M}_j$ of two subjects i and j , the *mutual information* (MI) between the optimal state-sequences estimated by the two SSMs was used. Specifically, the MI is measured from the joint histogram of $x^{(i)}$ and $x^{(j)}$, the optimal state-sequences for the same fMRI data $\mathbf{y}$ as labeled by models $\mathcal{M}_i$ and $\mathcal{M}_j$ respectively. The relationship between the state labels of two different SSMs is unknown and by comparing the state-sequences assigned to the same data, this mapping can be determined. A higher MI indicates a higher level of correspondence between the two state-sequences, while an MI of zero indicates no agreement.

From the site-by-site distribution of the pairwise MI between subjects in Fig. 13, the following trends were observed. Firstly, within a single site, the MI between controls was significantly higher than that between patients ($p = 0.001$, Cramér's test, FDR corrected), with an average reduction of 0.34 ± 0.19 units. Although the MI between control and patient SSMs was lower than that between patients (significant at $p = 0.01$, Cramér's test, FDR corrected) the average reduction was much lower at 0.16 ± 0.03 units. This indicates a greater congruence among the spatio-temporal patterns of neural activation of healthy adults than that among the patterns of schizophrenic adults, and a slightly lower similarity between the patterns of patients and controls. Secondly, when comparing subjects across different sites, we observed a similar trend although there is an overall reduction in the MI measures for all cases (significant at $p = 0.001$, Cramér's test, FDR corrected) which may be due to the fact that at each site MRI scanners from different manufacturers and different field strengths were used, which along with differences in the acquisition protocols could have induced site specific signatures into the fMRI signal. Such differences would cause the SSM trained for one subject to be *a priori* biased towards better recognition for the fMRI data from another subject acquired at the same site, resulting in state-sequence labeling with better correspondence.

This procedure when applied to all $\binom{69}{2}$ pairs of subjects yields a *similarity matrix* of pairwise MI, which can be thought of as assigning a position to each subject relative to all other subjects, in a 69-dimensional space. Using the technique of Laplacian eigenmaps (Belkin & Niyogi, 2003), each subject was embedded into a low-dimensional Euclidean space with minimum distortion to the distances implied by this 69×69 matrix. While the

fidelity of the embedding increases with the number of dimensions. reducing the dimensionality to 4 preserved the original distances with less than 5% error.

To study the separation between control and population groups in this abstract representation space, a SVM classifier with a Gaussian kernel was trained with each subject's 4-d Euclidean coordinates as the feature vector and group (Control or Patient) as the label. Generalization error was measured using leave-one-out cross validation. The quality of the discrimination between the two groups was assessed using a non-parametric test for the significance of the generalization error, where the null distribution was estimated by permuting the group labels. The generalization accuracy (i.e. 1 minus error) and its p -value are listed in Table 4. For illustrative purposes, the relative location of each subject in the two most discriminative dimensions of the 4-d embedding space (i.e. the two dimensions of the 4-d space with best classification performance) and the associated SVM classifier boundary (Gaussian kernel) are shown in Fig. 14. The projection error for this 2-d embedding is less than 17%. Please note that the axes are invariant up to a rigid transformation and have no physical significance. Here, the positions of the subjects at each site separately were computed from the Lapacian eigenmap embedding of the pairwise MI matrix restricted to that site (§ Fig. 14(a)–(d)).

Although the two groups are well separated by the SVM classifier boundary, the relatively high p -value of the generalization error was due to the small sample size. The high flexibility afforded by the Gaussian kernel allows the SVM to register high performance even for a permuted labeling. The significance of the separation improves with sample size, as seen in the case of Sites 3 and 4. However, when pooling the subjects from all sites together into a single representation space obtained from the 69×69 pairwise MI matrix, we observe a strong clustering by site and very low generalization accuracy (§Fig. 14(e) and Table 4, column 6). The higher *a priori* MI between subjects at the same site and lower MI between subjects at different sites leads to this site-wise clustering, as visualized in Fig. 14(e).

To control for this effect, a fixed-effects regression on the pairwise MI was performed, as follows. Let $MI^{\text{total}}(i, j)$ be the measured MI between the SSMs of subjects $i, j = 1 \dots 69$ and let $St(i) = 1 \dots 4$, be the site index of subject i . The mutual information $MI(i, j)$ is modeled as a linear combination of the *a priori* site effect $MI_{St(i), St(j)}^{\text{site}}$ and a subject specific effect $MI^{\text{subj}}(i, j)$; namely: $MI^{\text{total}}(i, j) = MI_{St(i), St(j)}^{\text{site}} + MI^{\text{subj}}(i, j)$, assuming $MI^{\text{subj}}(i, j)$ as a zero-mean random variable. After regressing out the site effects $MI_{St(i), St(j)}^{\text{site}}$, the relative position of each subject and the classification boundary is illustrated in Fig. 14(f). As observed visually, and confirmed quantitatively in column 7 of Table 4, there is a significant separation between the controls and patients in this abstract representation space after this correction.

5. Discussion

One of the main observations of the state-space model (SSM) analysis of the fBIRN Phase II data-set for the SIRP task was the reduced overall predictability of experimental stimuli from the fMRI data of schizophrenia (SZ) patients as compared to normal adults. As

mentioned earlier, the predictive power is related to the regularities in spatiotemporal patterns of neural activation and the ability of the SSM to adequately extract these patterns from the data and model their possible variations(?, ?). Therefore, the differential prediction accuracy of the SSM between the two groups provides confirmation towards a hypothesis of a larger heterogeneity or irregularity in neural recruitment patterns of a SZ patient. That the neurocognitive functioning of SZ patients is more heterogeneous than individuals without psychological disorders has been observed in many neuropsychological studies of these patients (Joyce & Roiser, 2007; Palmer, Dawes, & Heaton, 2009). Heterogeneity of brain response to cognitive challenge is also common in SZ and has been attributed to inefficient brain processing, with the level of inefficiency varying by task difficulty and underlying neural involvement (Brown & Thompson, 2010). Previously published univariate GLM analysis of the current imaging data demonstrated more abnormal and inefficient patterns of neural recruitment of the dlPFC in response to increased working memory loads, supporting the present results (Potkin & Ford, 2009; Brown et al., 2009). Although the diminished prediction of the experimental task condition from the patient fMRI data might have been due to greater behavioral movement among patients, studies using movement correction parameters to compare overall or stimulus correlated movement have not usually found differences between SZ patients and controls (Eyler, Olsen, Jeste, & Brown, 2004; Kindermann, Brown, Zorrilla, Olsen, & Jeste, 2004); rather, the neurocognitive heterogeneity in schizophrenia has been hypothesized to be related to the genetic variation of the disorder (Joyce & Roiser, 2007).

In addition to this main effect, the following parametric effects were observed:

- I. **Reduction of prediction accuracy in both groups for the Foil_Probe over the Target_Probe:** This finding implies differences in the underlying neurocognitive architecture of target and foil responding, an implication supported by magnetoencephalographic (MEG) recordings showing greater alpha-suppression duration for negative probes than positive probes (Rojas, Teale, Sheeder, & Reite, 2000). The fMRI difference between foils and targets found in the present study and the MEG differences Rojas et al. (2000) reported would be difficult to account for by recognition models that attribute recognition judgments about targets and foils to the application of a binary decision criterion to a unidimensional memory attribute, such as trace strength or familiarity. Most general theories of recognition memory posit multiple decision and/or memory search mechanisms that are differentially engaged when targets or foils are presented (Atkinson & Juola, 1974; Nosofsky, Little, Donkin, & Fific, 2011). The Atkinson and Juola (1974) model, for example, assumes that memory scanning involves a fast judgment about the familiarity of the memory probe given the experimental context followed by a slower memory scanning process if the familiarity judgment does not lead to a clear yes-no decision. Fits of the Atkinson and Juola (1974) model to response times revealed that on a larger number of trials individuals were bypassing the slower memory scanning stage when probes were targets than when probes were foils. A formal model of the response times applied to data from the present study also combined a memory scanning phase and a familiarity judgment phase that predicted more engagement of both processes on foil trials than on target trials

(Brown et al., 2009). The greater cognitive complexity of foil trials probably contributed to their being less predictable from fMRI data.

II. Reduced prediction accuracy in each group under increasing WM load

conditions: For SZ patients, there was a more marked decrease in prediction accuracy for foils compared to targets (a pattern not observed in controls). That the linkage of task condition to the spatio-temporal patterns underlying fMRI is altered by increasing WM load undermines to some extent the logic of the parametric design of the study. In particular, the assumption that a fixed set of neurocognitive processes become increasingly engaged as WM load increases might be too strong. Instead parametric manipulations might engage increasingly complex neurocognitive processes, albeit in an inconsistent manner, as subjects engage various cognitive processes to optimize performance. This alternative account of the impact of parametric manipulations of memory load is compatible with the multidimensional memory scanning models discussed above. The finding that predictability of the foil condition was especially low among SZ patients was probably due to the increased neurocognitive complexity posited for foil processing coupled with the dysregulation of the brain's WM response in schizophrenia (Jansma, Ramsey, Wee, & Kahn, 2004).

III. Increased accuracy of predictions about task condition from the fMRI data

across probe trials: These results contradict the assumption that task trials provided exchangeable views of the brain's response to WM task demands. Apparently later items provided more systematic spatio-temporal signatures of task demands. Improved memory performance with test practice is well known and is believed to be due the development of consistent retrieval schemas (Zaromb & Roediger, 2010). Interestingly, signal intensity and activation extent have both been shown to increase and to stabilize over repeated trials of a WM task and it has been proposed that the stabilization process may be driven by activity in the medial frontal cortex (Paskavitz et al., 2010). Whatever the ultimate cause, the increased prediction accuracy of the task conditions on later trials was similar among schizophrenia patients and controls only for target trials. Among schizophrenia patients, prediction accuracy of task condition did not improve across trials for foils.

From the MI of the state-space trajectories, we observe that the spatio-temporal patterns of control subjects are relatively similar (when their brains were normalized into the same atlas space). However, the patterns of patients diverge from each other as do the patterns of patients and controls. Furthermore, the spatio-temporal patterns of all subjects inhabit a low-dimensional abstract representation space, and there are clear clustering between the two groups in this space. The results support neurocognitive theories that assume that cognitive tasks usually engage the same abstract brain space in healthy controls and SZ patients but that disordered functioning among patients is manifested by variable patterns of activation and coupling within the same neurocognitive space.

Plots of study participants in the two most discriminating state dimensions showed clustering by diagnostic group and by site. Other studies have found systematic site effects

in fMRI data when MRI scanners at different field strengths are compared as in the present study (Cohen et al., 2004; Friedman et al., 2008). In Fig. 14(e), patients and controls tended to cluster together by site, reinforcing the design principle that patients and controls should be recruited in equal proportions at each site in a multi-site study. After controlling for the site effect, schizophrenia patients and controls could be significantly distinguished in the low dimensional representation space. This layout of subjects in this abstract space, where subjects cluster intrinsically on their population type across sites, provides confirmatory evidence that the effects discovered by the SSM are related to mental condition and not simply artifacts induced by the imaging process. However, unless site is included in the statistical model, site differences can obscure the classification of schizophrenia patients and controls when studies involve different scanners.

As seen from the comparative evaluation of the SSM in Section 4.1, there are clear advantages of modeling the temporal structure in mental processes and including minimal but plausible assumptions about the hemodynamics generating the BOLD signal over a purely “model-free” and temporally agnostic approach such as SVM (Mórocz et al., 2012). The SSM can be used with complex paradigms such as the SIRP task, and the decoupling of the stimulus from the fMRI signal through the intermediate layers allows flexibility in choosing which experimental variables to include in the model and their encoding. The effects of other un-modeled experimental variables can then be tested *post hoc*. This is in contrast to MVPR classifiers that do not, by design, capture the effects of experimental variables against which they have not been trained. However, with simple block design paradigms where the effect of hemodynamics and the temporal structure within a block are insignificant, it has been observed that MVPR classifiers tend to outperform the SSM in predicting the mental state (Janoos et al., 2011).

The analysis presented in this article are exploratory in nature and the results need to be verified on the other 4 sites of the fBIRN study, and replicated in other studies. However, these findings contain hints about the brain abnormalities in schizophrenia in terms of the organization of the neural representation of working memory processes and the effects of experimental conditions on it. These findings might lead to more specific hypotheses and guide in the design of experiments to test them. In particular, the study results showed that the spatio-temporal patterns of MR signals were less tightly regulated by stimulus conditions in patients than in controls. The study also found that the dysregulation observed among SZ patients was especially prominent when stimulus probes were foils. The choice ambiguity associated with the negative memory search during the foil condition appeared especially to disrupt the development of stable patterns of brain response among schizophrenia patients. Further studies will determine whether the failure to develop stable patterns of brain response to memory foils can be explained by abnormal activity of or connectivity to the medial frontal cortex. The findings from this study extend the univariate results previously published by the fBIRN group (Potkin et al., 2009) and reinforce the importance of using multivariate methods to link task conditions to the spatio-temporal data obtained from functional magnetic resonance imaging studies.

Because the SSM approach, in combination with the MI-derived abstract representation spaces, has the potential to uncover subtle differences between functional brain data, that are

both spatially and temporally distinguished, we propose that it may serve for characterizing the gradual distribution of phenotypes of mental conditions that run in families or in a population. For example, there might be valuable correlations between disease parameters such as schizophrenia sub-type, or psychological measures and the SSM results. Alternatively, SSM analysis of intra-individual longitudinal fMRI could be used to study the effects of disease parameters, treatment schedules, response rate, etc. of patients, with respect to controls.

Acknowledgments

This work was partially funded by the National Center for Research Resources to the Function Biomedical Informatics Research Network (National Institutes of Health grant 1 U24 RR021992), NIH P41-RR-013218 and P41-EB-015902 and was supported in part by an allocation of computing time from the Ohio Supercomputer Center. The authors would also like to thank Prof. Jessica Turner (Mind Research Network, Albuquerque, USA) for her help in acquiring and processing the data.

References

- Atkinson, R.; Juola, J. Search and decision processes in recognition memory. In: Krantz, DH.; Atkinson, RC.; Luce, RD.; Suppes, P., editors. Contemporary developments in mathematical psychology: Vol. i. learning, memory, and thinking. W. H. Freeman and Company; San Francisco: 1974. p. 243-293.
- Baringhaus L, Franz C. On a new multivariate two-sample test. *J Multivar Anal.* 2004 Jan;88:190–206.
- Belkin M, Niyogi P. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Comp.* 2003; 15(6):1373–1396.
- Bellec P, Rosa-Neto P, Lyttelton OC, Benali H, Evans AC. Multilevel bootstrap analysis of stable clusters in resting-state fMRI. *Neuroimage.* 2010 Jul; 51(3):1126–1139. [PubMed: 20226257]
- Bilder RM. The neuropsychology of schizophrenia circa 2009. *Neuropsychol Rev.* 2009 Sep; 19(3): 277–279. [PubMed: 19680816]
- Bishop, CM. Pattern recognition and machine learning. 1. Springer; Hardcover: 2007. 2006. Corr. 2nd printing ed
- Boynton GM, Engel SA, Glover GH, Heeger DJ. Linear systems analysis of functional magnetic resonance imaging in human v1. *Neurosci.* 1996 Jul; 16(13):4207–4221.
- Brown GG, McCarthy G, Bischoff-Grethe A, Ozyurt B, Greve D, Potkin SG, et al. Brain-performance correlates of working memory retrieval in schizophrenia: a cognitive modeling approach. *Schizophr Bull.* 2009 Jan; 35(1):32–46. [PubMed: 19023127]
- Brown GG, Thompson WK. Functional brain imaging in schizophrenia: selected results and methods. *Curr Top Behav Neurosci.* 2010; 4:181–214. [PubMed: 21312401]
- Calhoun VD, Adali T. Unmixing fMRI with independent component analysis. *IEEE Eng Med Biol Mag.* 2006; 25(2):79–90. [PubMed: 16568940]
- Chung, F. Lectures on spectral graph theory. Am Math Soc; 1997.
- Cohen ER, Rostrup E, Sidaros K, Lund TE, Paulson OB, Ugurbil K, et al. Hypercapnic normalization of bold fmri: comparison across field strengths and pulse sequences. *Neuroimage.* 2004 Oct; 23(2):613–624. [PubMed: 15488411]
- Diagnostic and statistical manual of mental disorders DSM-IV-TR fourth edition (text revision). 4. Amer Psychiatric Pub. Paperback; 2000.
- Egan MF, Weinberger DR. Neurobiology of schizophrenia. *Curr Opin Neurobiol.* 1997 Oct; 7(5):701–707. [PubMed: 9384552]
- Eyler LT, Olsen RK, Jeste DV, Brown GG. Abnormal brain response of chronic schizophrenia patients despite normal performance during a visual vigilance task. *Psychiatry Res.* 2004 Apr; 130(3):245–257. [PubMed: 15135158]
- Faisan S, Thoraval L, Armspach JP, Heitz F. Hidden Markov multiple event sequence models: A paradigm for the spatio-temporal analysis of fMRI data. *Med Image Ana.* 2007 Feb; 11(1):1–20.

- Friedman L, Stern H, Brown GG, Mathalon DH, Turner J, Glover GH, et al. Test-retest and between-site reliability in a multicenter fmri study. *Hum Brain Mapp*. 2008 Aug; 29(8):958–972. [PubMed: 17636563]
- Friston K, Chu C, Mouro-Miranda J, Hulme O, Rees G, Penny W, et al. Bayesian decoding of brain images. *Neuroimage*. 2008 Jan; 39(1):181–205. [PubMed: 17919928]
- Friston KJ. Modalities, modes, and models in functional neuroimaging. *Science*. 2009 Oct; 326(5951):399–403. [PubMed: 19833961]
- Friston KJ, Mechelli A, Turner R, Price CJ. Nonlinear responses in fMRI: the balloon model, volterra kernels, and other hemodynamics. *Neuroimage*. 2000 Oct; 12(4):466–477. [PubMed: 10988040]
- Hastie, T.; Tibshirani, R.; Friedman, JH. *The elements of statistical learning*. Springer; Hardcover: 2003. Corrected ed
- Haxby JV, Gobbini MI, Furey ML, Ishai A, Schouten JL, Pietrini P. Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*. 2001 Sep; 293(5539):2425–2430. [PubMed: 11577229]
- Haynes JD, Rees G. Decoding mental states from brain activity in humans. *Nature Rev: Neurosci*. 2006 Jul; 7(7):523–534. [PubMed: 16791142]
- Højen-Sørensen, P.; Hansen, LK.; Rasmussen, CE. Bayesian modelling of fMRI time series. In: SS, et al., editors. *Adv neural info proc sys (nips)*. 2000. p. 754-760.
- Hutchinson RA, Niculescu RS, Keller TA, Rustandi I, Mitchell TM. Modeling fMRI data generated by overlapping cognitive processes with unknown onsets using Hidden Process Models. *Neuroimage*. 2009; 46(1):87–104. [PubMed: 19457397]
- Jain AK. Data clustering: 50 years beyond k-means. *Pat Recog Letters*. 2010; 31(8):651–666.
- Janoos F, Machiraju R, Singh S, Morocz I kos. Spatio-temporal models of mental processes from fmri. *Neuroimage*. 2011 Jul; 57(2):362–377. [PubMed: 21440069]
- Jansma JM, Ramsey NF, van der Wee NJA, Kahn RS. Working memory capacity in schizophrenia: a parametric fmri study. *Schizophr Res*. 2004 Jun; 68(2–3):159–171. [PubMed: 15099600]
- Joyce EM, Roiser JP. Cognitive heterogeneity in schizophrenia. *Curr Opin Psychiatry*. 2007 May; 20(3):268–272. [PubMed: 17415081]
- Kim DI, Manoach DS, Mathalon DH, Turner JA, Mannell M, Brown GG, et al. Dysregulation of working memory and default-mode networks in schizophrenia using independent component analysis, an fbirn and mcic study. *Hum Brain Mapp*. 2009 Nov; 30(11):3795–3811. [PubMed: 19434601]
- Kim MA, Tura E, Potkin SG, Fallon JH, Manoach DS, Calhoun VD, et al. Working memory circuitry in schizophrenia shows widespread cortical inefficiency and compensation. *Schizophr Res*. 2010 Mar; 117(1):42–51. [PubMed: 20096539]
- Kindermann SS, Brown GG, Zorrilla LE, Olsen RK, Jeste DV. Spatial working memory among middle-aged and older patients with schizophrenia and volunteers using fmri. *Schizophr Res*. 2004 Jun; 68(2–3):203–216. [PubMed: 15099603]
- Kriegeskorte N, Bandettini P. Analyzing for information, not activation, to exploit high-resolution fmri. *Neuroimage*. 2007 Dec; 38(4):649–662. [PubMed: 17804260]
- Kriegeskorte N, Mur M, Bandettini P. Representational similarity analysis -connecting the branches of systems neuroscience. *Front Syst Neurosci*. 2008; 2:4. [PubMed: 19104670]
- Kruggel F, Zysset S, von Cramon DY. Nonlinear regression of functional MRI data: an item recognition task study. *Neuroimage*. 2000 Aug; 12(2):173–183. [PubMed: 10913323]
- Lanternman AD. Schwarz, Wallace, and Rissanen: Intertwining themes in theories of model selection. *Int Stat Review*. 2001; 69(2):185–212.
- Li K, Guo L, Nie J, Li G, Liu T. Review of methods for functional brain connectivity detection using fMRI. *J Comp Med Imag Graphics*. 2009; 33(2):131–139.
- Makni S, Beckmann C, Smith S, Woolrich M. Bayesian deconvolution of [corrected] fmri data using bilinear dynamical systems. *Neuroimage*. 2008 Oct; 42(4):1381–1396. [PubMed: 18603003]
- Manoach DS. Prefrontal cortex dysfunction during working memory performance in schizophrenia: reconciling discrepant findings. *Schizophr Res*. 2003 Apr; 60(2–3):285–298. [PubMed: 12591590]

- Manoach DS, Schlaug G, Siewert B, Darby DG, Bly BM, Benfield A, et al. Prefrontal cortex fmri signal changes are correlated with working memory load. *Neuroreport*. 1997 Jan; 8(2):545–549. [PubMed: 9080445]
- McIntosh AR, Lobaugh NJ. Partial least squares analysis of neuroimaging data: applications and advances. *Neuroimage*. 2004; 23(Suppl 1):S250–S263. [PubMed: 15501095]
- Mendrek A, Kiehl KA, Smith AM, Irwin D, Forster BB, Liddle PF. Dysfunction of a distributed neural circuitry in schizophrenia patients during a working-memory performance. *Psychol Med*. 2005 Feb; 35(2):187–196. [PubMed: 15841676]
- Meyer-Lindenberg A, Poline JB, Kohn PD, Holt JL, Egan MF, Weinberger DR, et al. Evidence for abnormal cortical functional connectivity during working memory in schizophrenia. *Am J Psychiatry*. 2001 Nov; 158(11):1809–1817. [PubMed: 11691686]
- Mitchell TM, Shinkareva SV, Carlson A, Chang KM, Malave VL, Mason RA, et al. Predicting Human Brain Activity Associated with the Meanings of Nouns. *Science*. 2008; 320(5880):1191–1195. [PubMed: 18511683]
- Mórocz IA, Janoos F, Gelderen P, van Manor D, Karni A, Breznitz Z, et al. Time-resolved and spatio-temporal analysis of complex cognitive processes and their role in disorders like developmental dyscalculia. *Int J Imaging Syst Technol*. 2012 Mar; 22(1):81–96. [PubMed: 22368322]
- Norman KA, Polyn SM, Detre GJ, Haxby JV. Beyond mind-reading: multi-voxel pattern analysis of fMRI data. *Trends Cogn Sci*. 2006 Sep; 10(9):424–430. [PubMed: 16899397]
- Nosofsky RM, Little DR, Donkin C, Fific M. Short-term memory scanning viewed as exemplar-based categorization. *Psychol Rev*. 2011 Apr; 118(2):280–315. [PubMed: 21355662]
- O'Toole AJ, Jiang F, Abdi H, Pnard N, Dunlop JP, Parent MA. Theoretical, statistical, and practical perspectives on pattern-based classification approaches to the analysis of functional neuroimaging data. *J Cog Neurosci*. 2007 Nov; 19(11):1735–1752.
- Ou W, Wells WM, Golland P. Combining spatial priors and anatomical information for fMRI detection. *Med Image Ana*. 2010 Jun; 14(3):318–331.
- Palmer BW, Dawes SE, Heaton RK. What do we know about neuropsychological aspects of schizophrenia? *Neuropsychol Rev*. 2009 Sep; 19(3):365–384. [PubMed: 19639412]
- Paskavitz JF, Sweet LH, Wellen J, Helmer KG, Rao SM, Cohen RA. Recruitment and stabilization of brain activation within a working memory task; an fmri study. *Brain Imaging Behav*. 2010 Mar; 4(1):5–21. [PubMed: 20503110]
- Penny W, Ghahramani Z, Friston K. Bilinear dynamical systems. *Philos Trans R Soc Lond B Biol Sci*. 2005 May; 360(1457):983–993. [PubMed: 16087442]
- Pereira F, Mitchell T, Botvinick M. Machine learning classifiers and fMRI: A tutorial overview. *Neuroimage*. 2009; 45(1, Supplement 1):S199–S209. [PubMed: 19070668]
- Perlstein WM, Carter CS, Noll DC, Cohen JD. Relation of prefrontal cortex dysfunction to working memory and symptoms in schizophrenia. *Am J Psychiatry*. 2001 Jul; 158(7):1105–1113. [PubMed: 11431233]
- Potkin SG, Ford JM. Widespread cortical dysfunction in schizophrenia: the fbirn imaging consortium. *Schizophr Bull*. 2009 Jan; 35(1):15–18. [PubMed: 19023124]
- Potkin SG, Turner JA, Brown GG, McCarthy G, Greve DN, Glover GH, et al. Working memory and dlpc inefficiency in schizophrenia: the fbirn study. *Schizophr Bull*. 2009 Jan; 35(1):19–31. [PubMed: 19042912]
- Quirós A, Diez RM, Wilson SP. Bayesian spatiotemporal model of fMRI data using transfer functions. *Neuroimage*. 2010 Sep; 52(3):995–1004. [PubMed: 20056161]
- Rojas DC, Teale PD, Sheeder JL, Reite ML. Neuromagnetic alpha suppression during an auditory sternberg task. evidence for a serial, self-terminating search of short-term memory. *Brain Res Cogn Brain Res*. 2000 Sep; 10(1–2):85–89. [PubMed: 10978695]
- Smith S, Jenkinson M, Woolrich M, Beckmann C, Behrens T, Johansen-Berg H, et al. Advances in functional and structural MR image analysis and implementation as FSL. *Neuroimage*. 2004; 23(S1):208–219.
- Wolf DH, Gur RC, Valdez JN, Loughhead J, Elliott MA, Gur RE, et al. Alterations of fronto-temporal connectivity during word encoding in schizophrenia. *Psychiatry Research: Neuroimaging*. 2007; 154(3):221–232. [PubMed: 17360163]

- Worsley KJ, Liao CH, Aston J, Petre V, Duncan GH, Morales F, et al. A general statistical analysis for fMRI data. *Neuroimage*. 2002 Jan; 15(1):1–15. [PubMed: 11771969]
- Zaromb FM, Roediger HL. The testing effect in free recall is associated with enhanced organizational processes. *Mem Cognit*. 2010 Dec; 38(8):995–1008.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Figure 1. The state-space model (SSM)

The experimental stimuli are represented by s_t , while the corresponding brain-state is x_t , and the spatially distributed metabolic activation pattern is z_t . The activation pattern is observed in the fMRI data $y_t \dots y_{t+L-1}$ through convolution with the HRF $\mathbf{h}$

Figure 2. Conceptual Illustration of the Functional Distance

A toy example of a functional connectivity network represented as a matrix of the temporal correlations between voxels in Fig.(a) and its graph-based representation in Fig.(b). Figs. (c)–(e) show the neural activity intensity at three time-points from the data-set. The hot color scheme is used in these plots with activation magnitude increasing from left to right.

Figure 3. Overview of Method

Basis vectors of the feature-space Φ are computed from the functional connectivity of the brain estimated from the fMRI data $\mathbf{y}$ and dimensionality reduction is performed by retaining stable basis vectors using bootstrapping (cf. Section 2.3). The data after projecting into the low dimensional feature-space are used to estimate model parameters θ through a generalized EM algorithm. Given a set of model parameters, the optimal state-sequence $\mathbf{x}^*$ are estimated using an EM algorithm that iteratively transforms the problem into a series of first order hidden Markov models (§ Section 2.1). Model hyper-parameters K and λ_w are selected to minimize the error of predicting the stimulus $\mathbf{s}$ (§ Section 2.2).

Figure 4. Sternberg Item Recognition Paradigm (SIRP) task structure

A block was composed of a Prompt; an Encode period during which a memory set of 1, 3, or 5 digits was presented; a Probe period; and interleaved fixation periods. The timings of each are shown in the figure. The prompt was the word “Learn,” which was presented for 1.5 seconds, and was followed by a blank screen of 0.5 seconds. A blank screen lasting 1.0 second appeared between Encode and Probe period. Each of the 14 probes was presented for 1.1 seconds followed by a jittered blank screen interval lasting 1.543 seconds on the average. The minimum duration for the blank screen was 0.6 seconds, and the maximum was 2.486 seconds. Jittered intervals were selected so that the total Probe interval always equaled 38 seconds. A flickering cross-hair appeared during a jittered fixation period, which lasted 12.0 seconds on the average and ranged from 4.0 seconds to 20 seconds. Each run of the task included 6 blocks and each memory set size was presented twice per run.

Figure 5. Dimensionality Reduction – Single Subject

The figure shows the effect of varying the threshold τ_Φ on the dimensionality of the feature-space and on the maximum index m (level of decomposition) of the basis vectors $\phi^{(l,m)}$ retained. Results are for one control subject from site 4. The curves for other subjects were qualitatively similar.

Figure 6. Hyper-Parameters Selection

Fig.(a) shows the risk $\mathfrak{R}$ with respect to model-size K . Fig.(b) graphs $\mathfrak{R}$ with respect to precision hyper-parameter λ_w . Error bars indicate ± 1 std. dev. **Legend.** Blue solid line: control group, Red dashed line: patient group.

Figure 7. Comparative Evaluation – Target vs. Foil Prediction Accuracy

The figure shows the site-wise distribution of the prediction accuracy of the Target_Probe vs. Foil_Probe condition from each fMRI frame, using an SVM and the SSM derived classifiers. The blue box-plot (with notch) shows the SVM classifier results, while the black box-plot (without notch) shows the SSM derived classifier. For each classifier, the outline style box-plot shows the distribution for the control group, with the filled-in style box-plot shows that for the patient group.

Figure 8. Model Complexity

The optimal number of states K of each subject's SSM, distributed by group and by site is shown.

Figure 9. Overall Performance

Displayed is the overall accuracy in predicting whether the current stimulus is a Prompt, Encode, Target_Probe or Foil_Probe from an fMRI frame, distributed by group and by site. Chance-level prediction accuracy is approximately 0.20.

Figure 10. Prediction Accuracy by Probe Type

Plotted are the accuracies in predicting whether the current stimulus is a Target_Probe or Foil_Probe at any time point, for the two groups aggregated across sites. Chance-level accuracy for predicting either probe type is approximately 0.24.

Figure 11. Effect of Memory Load Condition on Prediction Accuracy

The figure shows the effect of varying memory load on the differential accuracy in predicting a Target_Probe or Foil_Probe stimulus from the fMRI data.

Figure 12. Effect of Task Phase on Prediction Accuracy

Shown is the differential accuracy in predicting the Target_Probe or Foil_Probe stimulus from the fMRI data, assessed separately over the first 5, second 5 and last 4 digits presented during the Probe epoch of the task.

Figure 13. Pairwise Mutual Information

Displayed are the pairwise MIs between the optimal state-sequences estimated by the SSMs of control vs. control, patient vs. patient and control vs. patient. The matrix format of the plots shows the MI between subjects on a site-by-site basis.

Figure 14. Low Dimensional Representation Space

Shown in Fig. (a)–(d) is the arrangement of subjects at each site along a 2-d plane of the 4-d representation space. Here, circles indicate control subjects and crosses indicate SZ patients. The solid black contour demarcates the SVM classification boundary between the two subject groups. Fig. (e) shows the arrangement of the subjects when pooled across all 4 sites, color coded according to site. Fig. (f) shows the relative arrangement after controlling for the site effect.

Table 1**Distribution of Subjects**

Listed are the number of healthy adults and subjects diagnosed with schizophrenia or schizoaffective disorder from each of the four sites whose data passed the quality assurance test and were used in this analysis.

	Site 1	Site 2	Site 3	Site 4
Controls	5	5	10	16
Patients	5	9	9	15
Total	10	4	19	31

Table 2
Dimensionality Reduction – Overall

Tabulated are the average (± 1 std.dev.) number of basis vectors of the feature space for the control and patient groups across the four sites, that are retained using the bootstrap analysis of stability.

Group	Site 1	Site 2	Site 3	Site 4
Control	432 $\pm$ 23	437 $\pm$ 24	446 $\pm$ 18	439 $\pm$ 25
Patient	427 $\pm$ 31	422 $\pm$ 29	434 $\pm$ 27	428 $\pm$ 32

Table 3
Comparative Evaluation – Overall Prediction Accuracy

Tabulated are the average prediction accuracies (± 1 std. dev.) for the SVM and SSM based classifiers to predict the Target Probe vs. Foil_Probe stimulus only from the fMRI volumes restricted to the Probe epoch of the task. Also, tabulated are the p -values of the measured prediction accuracy computed using a non-parametric permutation test. “Chance-level” accuracy for both methods is approximately 50%.

Group	Method	Overall	p -value
Control	SVM	0.57 $\pm$ 0.04	0.005
	SSM	0.66 $\pm$ 0.05	< 0.001
Patient	SVM	0.56 $\pm$ 0.05	0.009
	SSM	0.59 $\pm$ 0.05	< 0.001

Table 4

Group Separation

The generalization accuracy of the Gaussian kernel SVM classifier to predict a subject's diagnoses (Control or Patient) from the 4-d Euclidean embedding of the 69×69 MI matrix are compiled. Also included are the p -values, assessed using a permutation test.

	Site 1	Site 2	Site 3	Site 4	All sites	All sites (corrected)
Accuracy	0.72 ± 0.04	0.74 ± 0.05	0.79 ± 0.04	0.78 ± 0.04	0.67 ± 0.06	0.77 ± 0.05
p -value	0.23	0.21	0.14	0.16	0.19	0.09