bioRxiv preprint doi: https://doi.org/10.1101/106690; this version posted February 7, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

A probabilistic approach to discovering dynamic full-brain functional connectivity patterns

Jeremy R. Manning[1,*], Xia Zhu[2], Theodore L. Willke[2], Rajesh Ranganath[3], Kimberly Stachenfeld[3], Uri Hasson[3], David M. Blei[4], and Kenneth A. Norman[3]
[1]Dartmouth College, Hanover, NH
[2]Intel Corporation, Hillsboro, OR
[3]Princeton University, Princeton, NJ
[4]Columbia University, New York, NY
[*]Address correspondence to jeremy.r.manning@dartmouth.edu

February 7, 2017

Abstract

Recent work indicates that the covariance structure of functional magnetic resonance imaging (fMRI) data – commonly described as *functional connectivity* – can change as a function of the participant's cognitive state (for review see [32]). Here we present a technique, termed *hierarchical topographic factor analysis* (HTFA), for efficiently discovering full-brain networks in large multi-subject neuroimaging datasets. HTFA approximates each subject's network by first re-representing each brain image in terms of the activations of a set of localized *nodes*, and then computing the covariance of the activation time series of these nodes. The number of nodes, along with their locations, sizes, and activations (over time) are learned from the data. Because the number of nodes is typically substantially smaller than the number of fMRI voxels, HTFA can be orders of magnitude more efficient than traditional voxel-based functional connectivity approaches. In one case study, we show that HTFA recovers the known connectivity patterns underlying a synthetic dataset. In a second case study, we illustrate how HTFA may be used to discover dynamic full-brain activity and connectivity patterns in real fMRI data, collected as participants listened to a story. In a third case study, we carried out a similar series of analyses on fMRI data collected as participants viewed an episode of a television show. In these latter case studies, we found that both the HTFA-derived activity and connectivity patterns may be used to reliably decode which moments in the story or show the participants were experiencing. Further, we found that these two classes of patterns contained partially non-overlapping information, such that classifiers trained on combinations of activity-based and dynamic connectivity-based features performed better than classifiers trained on activity or connectivity patterns alone.

bioRxiv preprint doi: https://doi.org/10.1101/106690; this version posted February 7, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Introduction

The most common approaches for analyzing functional Magnetic Resonance Imaging (fMRI) data involve relating, in individual images, the activity of individual voxels or multi-voxel spatial patterns of brain activity to the subject's cognitive state [12, 13, 25, 37]. In contrast, functional connectivity analyses correlate the time series of activations *across* images of pairs of voxels [28]. Functional connectivity analyses have already led to important new insights into how the brain's correlational structure changes during different experimental conditions [32].

Because the size of the full-brain functional connectivity matrix grows with the square of the number of voxels, both filling in its entries and storing it in memory can become intractable for fMRI images with tens of thousands of voxels. For example, each connectivity matrix for a series of 50,000 voxel images occupies approximately 5 GB of memory (assuming only the upper triangle is stored, using single precision floating point entries). Storing many such matrices in memory (e.g. to compare different subjects and/or experimental conditions) can be impractical on modern hardware. Further, many of the algorithms used to relate multivariate patterns of voxel activations in individual images to cognitive states or experimental conditions (e.g. [21, 25]) use polynomial time and memory with respect to the number of features; this makes it impractical to use the same techniques to examine correlational data (although see [33] for another promising approach using massive parallelization on a specialized computing cluster).

Previous work has circumvented this scaling issue by using functionally [10, 18, 27, 35] or anatomically [2, 11] defined voxel clusters or regions of interest. These approaches segment the brain into discrete components, and then examine interactions or correlations between the activity patterns exhibited by those components (rather than attempting to examine every voxel-to-voxel interaction). In other words, these approaches encapsulate an intuition about how our brains work– specifically, that our brains are composed of a relatively small number of network nodes that interact with each other. Each of these approaches provides its own heuristic for defining those nodes and examining how they interact.

Here we propose *hierarchical topographic factor analysis* (HTFA), a Bayesian solution to this problem. Like previous methods, HTFA provides a compact means of representing full brain connectivity patterns that scales well to large datasets. But HTFA goes beyond these methods in that it (a) provides a natural means of determining how many network nodes should be used to describe a given dataset, (b) allows those nodes to be overlapping rather than forcing nodes to be fully distinct, and (c) constrains the nodes to be in similar (but not necessarily identical) locations across people.

HTFA casts each subjects' brain images as linear combinations of latent *factors* [Gaussian radial basis functions (RBFs)]. Each RBF can be interpreted as a spherical node in a simplified representation of the brain's networks. (The number of factors, K, is determined from the data.) In this way, HTFA is a true spatial model, in that these nodes reflect structures localized in 3D space whose activity patterns influence the observed voxel activations. This is conceptually different from approaches that interpret voxel activations directly— HTFA provides an explicit model for how voxels relate to each other according to their relative locations in space (see also [16]).

Applying HTFA to an fMRI dataset reveals the locations and sizes of these factors (i.e. the

bioRxiv preprint doi: https://doi.org/10.1101/106690; this version posted February 7, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

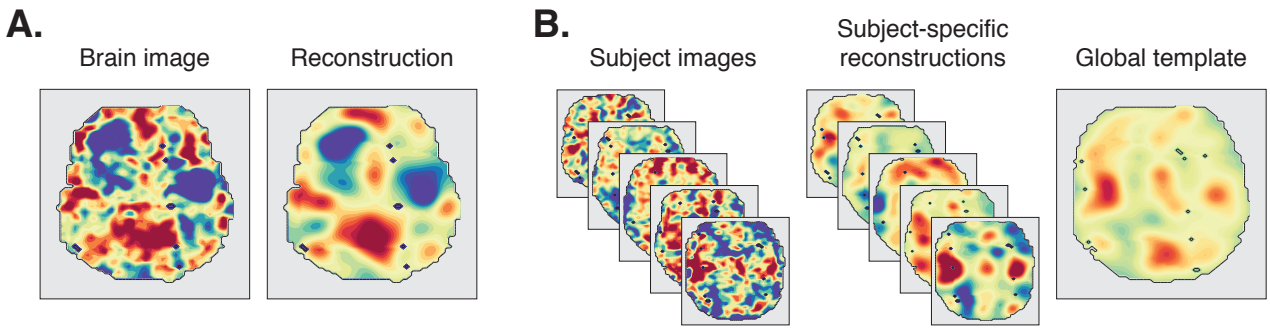

Figure 1. Hierarchical topographic factor analysis. A. A brain image and its associated reconstruction. The left sub-panel displays a single horizontal slice from a single subject; the right sub-panel displays its associated HTFA reconstruction, which we obtain by summing together the weighted images of the subject's RBF factors. **B. Explaining data across subjects.** The left and middle sub-panels display example images from five subjects (left sub-panel), and their associated reconstructions (middle sub-panel). The right sub-panel displays the approximation of all of the single-subject images in the left sub-panel, obtained by setting the weights of the global template's factors to their average weights in the subject-specific reconstructions. The locations of the RBFs in the global template reflect commonalities across subjects, whereas the single-subject RBF locations reflect the associated subject-specific idiosyncrasies.

centers and widths of their RBFs), as well as the per-image factor weights. If a given subject has contributed N images to the dataset, then the subject's N by K factor weights matrix may be viewed as a low-dimensional embedding of their original data. Further, the pairwise correlations between columns of this weight matrix reflect the signs and strengths of the node-to-node connections (just as the pairwise correlations between voxel time series reflect the corresponding connectivity in voxel-based functional connectivity analyses).

We designed HTFA to naturally account for data from multiple subjects. The model comprises a *global template*, which describes how data look in general (e.g. for the "typical" subject), as well as *subject-specific templates* which describe how each individual subject differs from that global template (Fig. 1). This hierarchical design [15] constrains the RBF centers and widths to be similar across subjects (without forcing them to be the same). Further, since every subject-specific template has an instantiation of every RBF in the global template, the model provides a natural means of performing across-subject hypothesis testing on functional connectivity patterns. This hierarchical structure also highlights another advantage of modeling 3D nodes, rather than considering voxel activations directly. Specifically, HTFA accounts for both anatomical and functional differences across individuals. For example, if the same anatomical or functional node is located in different locations in different people's brains, HTFA provides an explicit model for relating those nodes and comparing their activity and connectivity patterns across individuals.

In the next section we provide an overview and formal description of the HTFA model, and in *Materials and methods* we describe how one can efficiently fit the model to a multi-subject

bioRxiv preprint doi: https://doi.org/10.1101/106690; this version posted February 7, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

fMRI dataset using maximum *a posteriori* inference. To demonstrate our approach, we first generated a synthetic dataset for which the underlying activation patterns, node locations and sizes, etc. were known, and we show that HTFA recovers these known patterns. Then we apply HTFA to two (real) fMRI datasets collected as participants listened to an audio recording of a story (Case Study 2) and watched an episode of a television show (Case Study 3). We show that the moment-by-moment patterns uncovered by HTFA may be used to decode which moments in the story or show the participants were experiencing.

Model description

Overview

HTFA is a member of a family of models, called *factor analysis* models, that includes Topographic Factor Analysis (TFA) [24], Topographic Latent Source Analysis (TLSA) [16], Principal Components Analysis (PCA) [26], Exploratory Factor Analysis (EFA) [30], and Independent Components Analysis (ICA) [8, 23], among others. If we have organized our collection of images (from a single subject) into an N by V data matrix $\mathbf{Y}$ (where N is the number of images and V is the number of voxels), then factor analysis models decompose $\mathbf{Y}$ as follows:

$$\mathbf{Y} \approx \mathbf{W}\mathbf{F}, \tag{1}$$

where $\mathbf{W}$ is an N by K weight matrix (which describes how each of K factors are activated in each image), and $\mathbf{F}$ is a K by V matrix of factor images (which describes how each factor maps onto the brain). Note that, in the general case, this decomposition is underspecified– in other words, there are infinitely many solutions for $\mathbf{W}$ and $\mathbf{F}$ that approximate the data equally well. What differentiates factor analysis models is the particular constraints they place on what form $\mathbf{W}$ and/or $\mathbf{F}$ should take (i.e. by changing the function being optimized in order to settle on a specific choice of $\mathbf{W}$ and $\mathbf{F}$). We may then use $\mathbf{W}$ as a low-dimensional embedding of the original data (e.g. to facilitate interpretability or improve computational tractability), or we may choose to examine the factor images in $\mathbf{F}$ to gain insights into the dataset.

In related approaches such as PCA and ICA, the entries of $\mathbf{W}$ and $\mathbf{F}$ can be any real number. Specifically, In PCA, each row of $\mathbf{F}$ is an eigenvector of the data covariance matrix, and $\mathbf{W}$ is chosen to minimize the reconstruction error (i.e. to make $\mathbf{WF}$ as close as possible to $\mathbf{Y}$ in terms of mean squared error). In ICA, the goal is to minimize the statistical dependence between the rows of $\mathbf{F}$ while also adjusting $\mathbf{W}$ to minimize the reconstruction error. In this way, the factor images (the rows of $\mathbf{F}$) obtained using PCA and ICA are unstructured images (i.e. activation patterns) of the same complexity as observations in the original dataset: each factor is parameterized by V numbers (1 parameter per voxel).

In TFA (and TLSA), each row of $\mathbf{F}$ is parameterized by the center parameter, μ, and the width parameter, λ, of an RBF. If an RBF has center μ and width λ, then its activation $\mathrm{RBF}(\mathbf{r}|\mu, \lambda)$ at location $\mathbf{r}$ is:

$$\mathrm{RBF}(\mathbf{r}|\mu, \lambda) = \exp\left\{-\frac{||\mathbf{r} - \mu||^2}{\lambda}\right\}. \tag{2}$$

bioRxiv preprint doi: https://doi.org/10.1101/106690; this version posted February 7, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

The factor images are filled in by evaluating each RBF, defined by the corresponding parameters for each factor, at the location of each voxel. In contrast to the factors obtained using PCA or ICA, TFA's more constrained factors are easily interpretable: each factor corresponds to the structure or group of structures in the brain over which the factor spreads its mass (which is governed by μ and λ). As highlighted above, TFA's factors correspond to nodes located in 3D space whose activity patterns influence the observed voxel activations. While constraining the factors in this way reduces reconstruction accuracy relative to PCA or ICA (assuming the same number of factors), it substantially improves interpretability. For example, PCA's factors need not be concentrated in one brain region, and therefore reconstruct the data well but are difficult to interpret without additional processing.

HTFA works similarly to TFA, but places an additional constraint over the factors to bias all of the subjects to exhibit similar factors. In this way, whereas TFA attempts to find the factors that best explain an individual subject's data, HTFA also attempts to find the factors that are common across a group of subjects (Fig. 1). This is an important advance, because it allows the model to jointly consider data from multiple subjects (thereby facilitating across-subject analyses, etc.).

The model handles multi-subject data by defining a *global template*, which describes in general where each RBF is placed, how wide it is, and how active its factor tends to be. In addition to estimating how factors look and behave in general (across subjects), HTFA also estimates each individual's *subject-specific template*, which describes each subject's particular instantiations of each RBF (i.e. that subject's RBF locations and widths) and the factor activations (i.e. the activations of each of that subject's factors in each of that subject's images). These factor activations, in turn, are used to estimate each subject's full-brain functional connectivity patterns. Further, because the subject-specific templates are related to each other (via the global template), a given factor's RBF will tend to be located in about the same location, and be about as large, across all of the subject-specific templates. This property allows us to estimate an equivalent full-brain functional connectivity pattern in each subject, facilitating across-subject analyses.

Formal definition and notation

We formulate HTFA as a probabilistic latent variable model, which can be represented in graphical model notation. In the graphical model (Fig. 2), variables associated with the subject-specific template are found in the yellow plate. These include the subject-specific RBF centers ($\mu_{1\ldots K,1\ldots S}$), RBF widths ($\lambda_{1\ldots K,1\ldots S}$), and per-image factor weights ($w_{1\ldots N,1\ldots K,1\ldots S}$), as well as the observed images ($\mathbf{y}_{1\ldots N,1\ldots S}$). Variables associated with the global template are found outside of the yellow plate; these include the global RBF centers ($\hat{\mu}_{1\ldots K}$) and the global RBF widths ($\hat{\lambda}_{1\ldots K}$). The subject-specific templates are conditioned on the global template, thereby associating data from different subjects. (This interaction between the subject-specific and global templates occurs where the yellow and blue plates overlap.)

The structure of the graphical model specifies the conditional dependencies in HTFA:

$$p(\mathbf{Y}_{1\ldots S}, \mathbf{\Omega}) = \psi\gamma, \tag{3}$$

bioRxiv preprint doi: https://doi.org/10.1101/106690; this version posted February 7, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

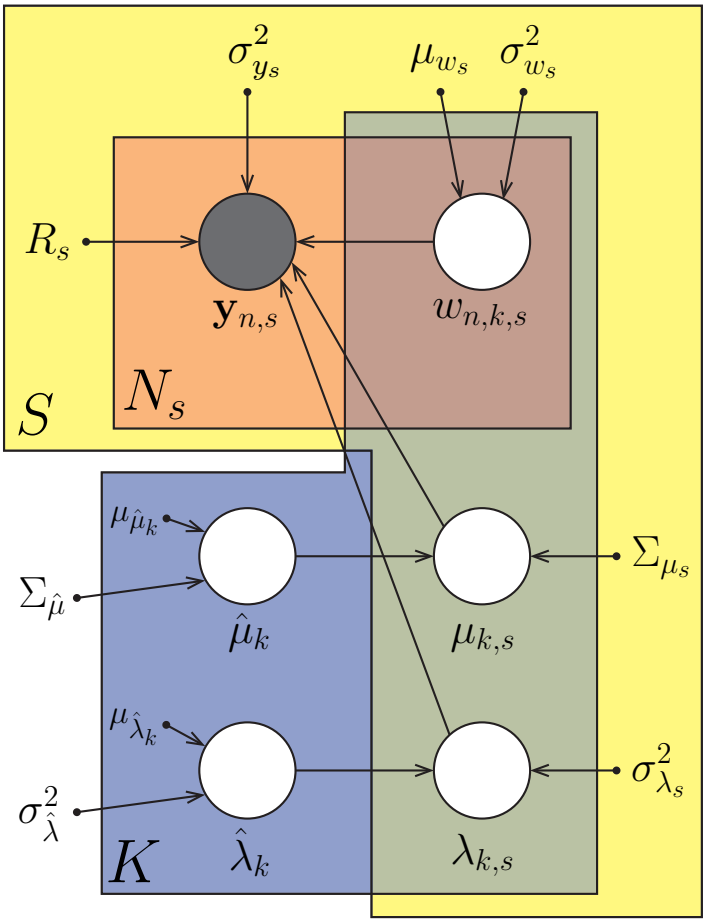

Figure 2. Graphical model for HTFA. Each variable in the model appears as a circle; hidden variables are unshaded and observed variables are shaded. Hyperparameters are denoted by dots. Arrows denote conditional dependence, originating at terms that appear on the right sides of conditionals and pointing towards terms that appear on the left sides. Rectangular plates denote repeated structure, where the number of copies is indicated within each plate (e.g. N_s, S, or K). For a comprehensive introduction to graphical models see [3]. Variables are defined in Algorithm 1 and in the text.

bioRxiv preprint doi: https://doi.org/10.1101/106690; this version posted February 7, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

where $\mathbf{Y}_s$ is the N_s by V_s matrix of images from subject s,[1] $\mathbf{\Omega}$ is the set of hidden variables in the model

$$\mathbf{\Omega} = \{w_{1\ldots N,1\ldots K,1\ldots S}, \mu_{1\ldots K,1\ldots S}, \lambda_{1\ldots K,1\ldots S}, \hat{\mu}_{1\ldots K}, \hat{\lambda}_{1\ldots K}\}, \tag{4}$$

the probability of the set of subject-specific templates ψ is

$$\psi = \prod_{s=1}^{S}\prod_{n=1}^{N_s} p(\mathbf{y}_{n,s}|w_{n,1\ldots K,s}, \mu_{1\ldots K,s}, \lambda_{1\ldots K,s}) \prod_{k=1}^{K} p(w_{n,k,s})p(\mu_{k,s}|\hat{\mu}_k)p(\lambda_{k,s}|\hat{\lambda}_k), \tag{5}$$

and where the probability of the global template γ is

$$\gamma = \prod_{k=1}^{K} p(\hat{\mu}_k)p(\hat{\lambda}_k). \tag{6}$$

(Note that the hyperparameters have been omitted to simplify the notation; see Algorithm 1 for the full specification.)

Another useful way to describe HTFA is through the *generative process* it implies. HTFA's generative process is an algorithm that, when run, generates a single sample from HTFA's joint distribution (Eqn. 3), yielding one value for each hidden variable and a multi-subject fMRI dataset. HTFA's generative process is detailed in Algorithm 1. The generative process starts by drawing a set of global RBF parameters, and from there draws subject-specific RBF parameters and factor weights, and from there draws the data. We emphasize that this "algorithm" represents the imaginary generative process from which the model assumes the data arises. When we fit the model by applying HTFA to a dataset, we "reverse" the generative process by starting with each subject's data, which we use to estimate their subject-specific RBFs and weights, which we

[1]Note that both the number of images from subject s, N_s, and the number of voxels from subject s, V_s, may vary across subjects.

bioRxiv preprint doi: https://doi.org/10.1101/106690; this version posted February 7, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

use in turn to estimate the global RBFs.

Algorithm 1: HTFA's generative process. Here $\mathbf{F}_s$ is the K by V_s factor image matrix for subject s, which depends on their subject-specific RBF centers ($\mu_{1...K,s}$), RBF widths ($\lambda_{1...K,s}$), and their V_s by 3 voxel location matrix ($\mathbf{R}_s$).

for $k = 1$ **to** K **do**
  Draw factor k's global RBF center $\hat{\mu}_k \sim \mathcal{N}(\mu_{\hat{\mu}_k}, \Sigma_{\hat{\mu}})$;
  Draw factor k's global RBF width $\hat{\lambda}_k \sim \mathcal{N}\left(\mu_{\hat{\lambda}_k}, \sigma^2_{\hat{\lambda}}\right)$;
end
for $s = 1$ **to** S **do**
  for $k = 1$ **to** K **do**
    Draw the kth factor's RBF center for subject s: $\mu_{k,s} \sim \mathcal{N}(\hat{\mu}_k, \Sigma_{\mu_s})$;
    Draw the kth factor's RBF width for subject s: $\lambda_{k,s} \sim \mathcal{N}\left(\hat{\lambda}_k, \sigma^2_{\lambda_s}\right)$;
    for $n = 1$ **to** N_s **do**
      For image n from subject s, draw the kth factor's weight: $w_{n,k,s} \sim \mathcal{N}\left(\mu_{w_s}, \sigma^2_{w_s}\right)$;
    end
  end
  for $n = 1$ **to** N_s **do**
    Draw brain image n for subject s: $\mathbf{y}_{n,s} \sim \mathcal{N}\left(w_{n,s}\mathbf{F}_s, \sigma^2_{y_s}\mathbf{I}^{V_s}\right)$;
  end
end

Our goal in applying HTFA to a dataset is to compute the probability distribution over the model's hidden variables (e.g. RBF centers and widths, and factor weights) given the dataset. The posterior distribution $p(\mathbf{\Omega}|\mathbf{Y}_{1...S})$ tells us how likely each hidden variable is to be set to a particular value, given the data and our prior assumptions about what these values should be. In theory, we could use Bayes' rule to compute this posterior:

$$p(\mathbf{\Omega}|\mathbf{Y}_{1...S}) = \frac{p(\mathbf{Y}_{1...S}|\mathbf{\Omega})p(\mathbf{\Omega})}{p(\mathbf{Y}_{1...S})}. \tag{7}$$

However, computing the denominator (as for most models) is intractable:

$$p(\mathbf{Y}_{1...S}) = \int_{\mathbf{\Omega}} p(\mathbf{Y}_{1...S}, \mathbf{\Omega}) d\mathbf{\Omega}. \tag{8}$$

Notice that solving for $p(\mathbf{Y}_{1...S})$ requires integrating over all possible values of each of the hidden variables in the model, for which there is no analytic solution. Therefore, instead of computing the full posterior distribution, here we have developed an efficient *maximum a posteriori* (MAP) algorithm for estimating the most probable values of the hidden variables under the posterior. We provide a high-level description of how we can use MAP inference to efficiently apply HTFA to large multi-subject fMRI datasets in *Materials and methods*; a full description may be found in [1]. We have also released an open source toolbox for applying HTFA to fMRI data, available for download here.

bioRxiv preprint doi: https://doi.org/10.1101/106690; this version posted February 7, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Results

We applied HTFA to three datasets. The first dataset was synthetic, and we used it to test HTFA's ability to recover known patterns in the data. We used the second and third datasets to demonstrate HTFA's performance on (real) fMRI datasets where participants listened to a story (Case Study 2) and watched an episode of a television show (Case Study 3).

Case study 1: examining network recovery from synthetic data

Our synthetic dataset includes 100 brain images from each of 10 simulated subjects. Each image comprises a $25 \times 25 \times 25$ rectangular block of 15,625 $4 \times 4 \times 2$ mm voxels. To generate the voxel activations in each image, we first randomly placed 50 RBF centers in a "template" brain volume, and randomly assigned a positive width to each RBF. We then generated each subject's RBF centers, widths, and activations by adding a small amount of Gaussian noise to the corresponding parameters in the template. Next, we set the time-varying activations of each simulated subject's RBF factors to exhibit a pre-defined sequence of activation patterns (Fig. 3C and D).

Our primary goal in examining the synthetic dataset was to assess whether the patterns revealed by applying HTFA to the dataset (i.e. the RBF factors and factor weights) corresponded to the known patterns in the data. We first compared the global RBF centers and widths recovered by the model to the RBF centers and widths in the original template image. Our inference procedure (see *Materials and methods*) determined that 80 RBF factors were needed to adequately explain the synthetic data. Note that the data were generated using 50 RBF factors, so our inference procedure over-estimated the number of required factors. We assessed how well the underlying parameters were recovered by assigning each recovered factor (in the inferred posterior) to the nearest true factor (in the global template we used to generate the dataset). The assignments were made by linking each recovered factor with the true factor whose center parameter was closest (in terms of Euclidean distance).

We found that the RBF centers recovered by applying HTFA to the subject data closely matched the template's RBF centers (Fig. 3A), and that the recovered RBF widths were reliably correlated with the template RBF widths (Fig. 3B; $r = 0.38, p = 0.007$). HTFA tended to over-estimate the true RBF widths, and the degree to which a given RBF's width was over-estimated was correlated with the Euclidean distance between the true and estimated RBF center ($r = 0.39, p = 0.005$). This positive correlation may be explained by the following intuition: if an RBF's center is mis-estimated, then to explain the variability in the data governed by the corresponding RBF in the template, HTFA compensates by growing the width of the estimated RBF to encompass the original (true) RBF.

We next asked whether HTFA accurately recovered the correlational patterns in the synthetic data. We found that both the mean voxel-to-voxel correlation matrix ($r = 0.62, p < 10^{-10}$; Fig. 3C) and the mean across-image correlation matrix ($r = 0.95, p < 10^{-10}$; Fig. 3D) recovered by HTFA were strongly correlated with the true patterns we embedded into the synthetic data. Taken together, these analyses show that HTFA is able to accurately infer (from synthetic data) the locations and sizes of the underlying RBF factors, the correlation patterns within the images, and the correlations across images. We next turn to a series of analogous analyses on a real

bioRxiv preprint doi: https://doi.org/10.1101/106690; this version posted February 7, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

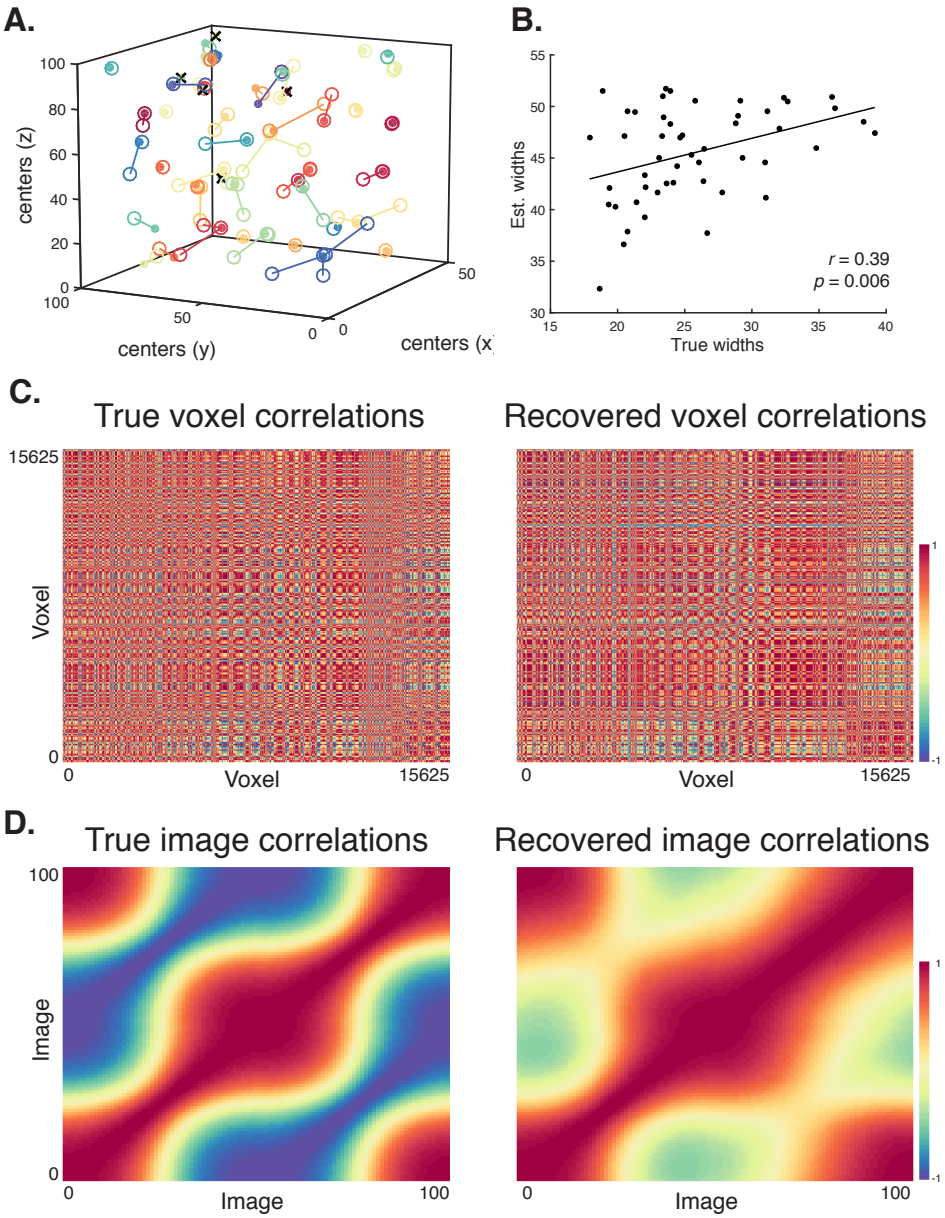

Figure 3. Recovered structure from synthetic data. A. RBF centers. Each color denotes a different factor's RBF in the synthetic template. The filled circles indicate the true locations of the RBF centers, and the open circles indicate the RBF centers recovered by HTFA. To facilitate visual comparison between the true and recovered locations, we have drawn a line between each recovered RBF center and the closest matching RBF center in the synthetic template. RBFs that have not been assigned any factors by HTFA (as determined by using this matching technique) are denoted by ×s. **B. RBF widths.** Each dot denotes an RBF's true and recovered widths. The correlation reported in the panel is between the true and estimated RBF widths. To enable us to make comparisons in this panel, we have assigned each recovered factor to the factor in the original template with the closest RBF center (in terms of Euclidean distance). **C. Voxel-to-voxel correlation matrices (within image).** The panels display the true and recovered voxel-to-voxel correlations. **D. Across-image correlation matrices.** The panels display the true and recovered across-image correlation matrices.

bioRxiv preprint doi: https://doi.org/10.1101/106690; this version posted February 7, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

fMRI dataset.

Case study 2: full-brain networks are modulated by story listening

We examined a large fMRI dataset collected as 115 participants listened to an audio recording of a story (*intact* condition; 36 participants), listened to time-scrambled recordings of the same story (18 participants in the *paragraph-scrambled* condition listened to the paragraphs in a randomized order and 25 in the *word-scrambled* condition listened to the words in a randomized order), or lay resting with their eyes open in the scanner (*rest* condition; 36 participants). We sought to demonstrate how HTFA may be used to efficiently discover and examine functional connectivity patterns in (real) multi-subject fMRI datasets. This story listening dataset was collected as part of a separate study, where the full imaging parameters, image preprocessing methods, and experimental details may be found [29]. The dataset is available here.

In contrast to the synthetic dataset we examined in Case Study 1, in real datasets there are no "ground truth" parameter values to compare to the recovered estimates. Instead, we sought to explore how well the patterns HTFA discovered could be used to decode which specific moments in the story participants were listening to. We also sought to explore whether (and how) decoding performance varied with the properties of the stimuli the participants experienced. Specifically, we constructed a series of timepoint decoders that take brain patterns and estimate which moment in the story those patterns correspond to. The decoders make their estimates by comparing different people's brain patterns during different moments in the story (i.e., how similar is the pattern in one participant's brain to all of the timepoint-specific patterns from other participants who listened to the same story). Prior work with this dataset has shown that different participants' brain patterns are most similar for participants in the intact condition, less so for the paragraph-scrambled condition, even less for the word-scrambled condition, and least during the rest condition [29]. We therefore expected that our decoders should perform best on data from the intact condition, worst on data from the rest condition, and so on.

We first used a cross-validation procedure to determine the optimal number of RBF sources for efficiently representing the data while still capturing the relevant structure (see *Materials and methods*). We used all of the data from all of the experimental conditions (intact, paragraph-scrambled, word-scrambled, and rest) in this procedure– in other words, all of the experimental conditions were lumped together into a single dataset. The analysis indicated that $K = 700$ sources were optimal.

The fitted model provided estimates for each participants' RBF locations, widths, and time-varying activations. Further, because the global model connects the subject-specific models, we were able to compare different participants' activation patterns, even if their underlying RBF sources were not in exactly the same locations. We reasoned that, on one hand, this aspect of HTFA might improve our ability to capture cognitively relevant patterns (relative to examining the voxel activations directly). On the other hand, representing the brain images in the lower-dimensional space captured by HTFA necessarily results in information loss relative to the original voxel activation data. Effectively, HTFA blurs out high spatial frequency details from the images, where the precise amount of blurring depends on how large the RBF sources are and how many sources there are overall.

bioRxiv preprint doi: https://doi.org/10.1101/106690; this version posted February 7, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

We next evaluated how well HTFA is able to capture cognitively relevant brain patterns. We performed a decoding analysis, using cross-validation to estimate (using other participants' data) which parts of the story each HTFA-derived brain activity pattern corresponded to (see *Materials and methods*). We note that our primary goal was not to achieve perfect decoding accuracy, but rather to use decoding accuracy as a benchmark for determining whether different neural features specifically capture cognitively relevant brain patterns.

Separately for each experimental conditions, we divided participants into two groups. We then computed the average activity for each group, for each of 241 overlapping 90 s (60 TR) time windows. (The 90 s window length we used in our analyses followed [29].) This yielded one activity pattern for each group of participants and across each time window. Next, for each time window, we correlated the group 1 activity patterns in that window with the group 2 activity patterns. Using these correlations, we labeled the group 1 timepoints using the group 2 timepoints with which they were most highly correlated; we then computed the proportion of correctly labeled group 1 time windows. (We also performed the symmetric analysis whereby we labeled the group 2 timepoints using the group 1 timepoints as a template.) We repeated this procedure 100 times (randomly re-assigning participants to the two groups each time) to obtain a distribution of decoding accuracies for each experimental condition. (There were 241 time windows, so chance performance on this decoding test is $\frac{1}{241}$.) Further, the decoding test we used is more conservative than those used in some previously reported timepoint decoding studies (e.g. [22]) because we count a timepoint label as incorrect if it is not an exact match, even if it overlaps substantially with the correct label. For example, if our decoder matches the 0 – 60 TR window from group 1 with the 1 – 61 TR window from group 2 (i.e. 88.5 of the 90 seconds are overlapping), our performance metric considers this to be a decoding failure, indistinguishable (performance-wise) from if the group 1 and group 2 windows had not overlapped at all; by contrast, [22] used a more liberal procedure where they only compared the correct time window (e.g., 0 – 60 TR) to the exactly matching window or non-overlapping time windows (e.g., 61 – 120 TR), but not to partially overlapping windows (e.g., 1 – 61 TR). We chose to use this conservative test because our decoders attained over 99% accuracy for both voxel-based and HTFA-derived neural features (on data from the intact condition of the experiment) when we used the more liberal matching procedure. This made it difficult to achieve our goal of comparing and distinguishing between neural features.

As a baseline, we first used the participants' full-brain voxel activity patterns (44,415 voxels; $\mathbf{Y}_{1...S}$) to decode story timepoints. These voxel-based classifiers achieved reliably above-chance performance on data from all four experimental conditions ($ts(99) > 22, ps < 10^{-20}$; Fig. 4), with the best average performance in the intact condition (11.2% accuracy) and the worst average performance in the rest condition (1.5% accuracy). This shows that the original data (that we applied HTFA to) contained information about the story times that our correlation-based decoders could pick up on.

We compared the performance of the voxel-based classifiers to the decoding performance achieved using classifiers trained on the HTFA factor weights – i.e. the inferred timepoint-by-timepoint activations of the 700 nodes derived for each participant ($\mathbf{W}_{1...S}$). Like the voxel-based classifiers, these factor activation-based classifiers achieved reliably above-chance performance on data from all four experimental conditions ($ts(99) > 20, ps < 10^{-20}$; Fig. 4). However, although

bioRxiv preprint doi: https://doi.org/10.1101/106690; this version posted February 7, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

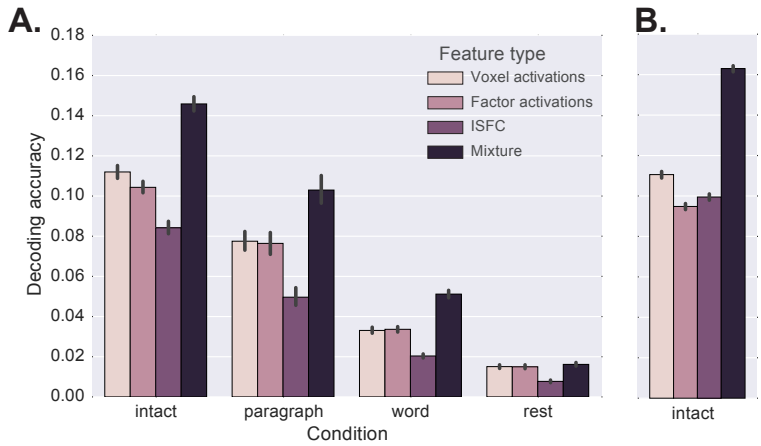

Figure 4. Decoding accuracy. Bars of each color display cross-validated decoding performance for classifiers trained using different sets of neural features: whole-brain patterns of voxel activations, whole-brain patterns of HTFA factor activations, HTFA-derived inter-subject functional connectivity (ISFC) patterns, and a 50-50 mixture of the classifiers trained using HTFA-derived activity and ISFC patterns. Panel **A.** displays the decoding performance for Case Study 2 and panel **B.** displays the decoding performance for Case Study 3.

decoding performance of the factor-based and voxel-based classifiers were similar (see panels A and B in Fig. 4), the voxel-based classifiers reliably out-performed the factor-based classifiers in the intact condition ($t(99) = 3.4, p = 0.007$; mean accuracy .77% higher for the voxel-based classifiers). (The performances of the factor-based and voxel-based classifiers were not statistically distinguishable in the other experimental conditions; ts$(99) < 0.6, p$s> 0.5.) This suggests that, although the HTFA-derived node activations still contain information about the story times, there seems to have been a small amount of information loss in the transformation from voxel activations to node activations, leading to slightly lower classification accuracy.

Although the above analysis shows that some cognitively relevant information is lost by applying HTFA to the data, the strength of HTFA is in its enabling efficient computations that involve functional connectivity patterns (whereas these computations are in some cases intractable in the original voxel space). Following the logic of [29], we reasoned that brain activations during story listening should capture two sources of information. First, some activity should reflect the story itself. Because every participant (within each condition) listened to the same stimulus, this story-driven activity should be similar across people. Second, some activity might reflect idiosyncratic thoughts or physiological processes specific to each individual, independent of the story. This non-story-driven activity should not be similar across people. To home in on the former (story-driven) contribution to functional connectivity, we used inter-subject functional connectivity (ISFC) [29]. This approach measures the correlations between brain regions *of different individuals* in each of several sliding windows (see *Materials and methods* for details). The resulting ISFC patterns are analogous to standard within-brain functional connectivity patterns (which reflect the correlational structure, across brain regions, within an individual's brain), but they should reflect only activity that is specifically stimulus-driven. Decoders trained

bioRxiv preprint doi: https://doi.org/10.1101/106690; this version posted February 7, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

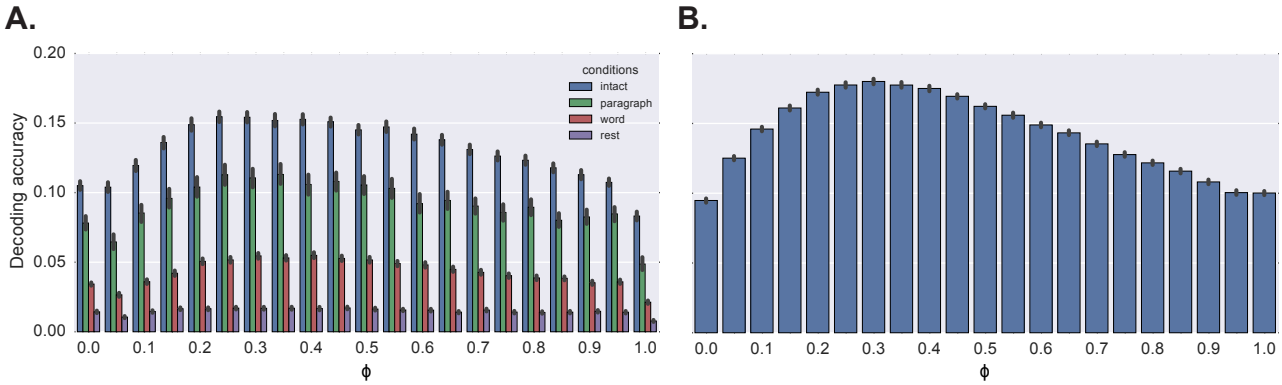

Figure 5. Decoding accuracy as a function of mixing proportion (ϕ). The bar heights indicate the decoding accuracy achieved for each value of the mixing parameter, ϕ, where $\phi = 0$ corresponds to activity-based decoding; $\phi = 1$ corresponds to ISFC-based decoding; and $0 < \phi < 1$ reflects a weighted mixture of activity- and ISFC-based decoding. Panel **A.** displays the decoding performance for Case Study 2 (colors denote the experimental condition) and panel **B.** displays the decoding performance for Case Study 3.

and tested on these HTFA-derived ISFC patterns achieved reliably above-chance performance on data from all experimental conditions ($ts(99) > 10, ps < 10^{-15}$; Fig. 4). Although these ISFC-based decoders were out-performed by the voxel-based and factor activity-based decoders (compare bars of the first three colors), it is important to note that the correlation-based features that the ISFC-based decoders utilize are fundamentally different than node activity patterns. Therefore, we hypothesized that the factor activation-based decoders and ISFC-based decoders might be picking up on partially non-overlapping sources of cognitively relevant information. If so, classifiers trained on a mix of activity-based and ISFC-based features might outperform classifiers trained on only a single class of features.

We therefore designed a fourth set of classifiers whose predictions were a 50-50 mix of the factor activity-based classifiers and the ISFC-based classifiers; see *Materials and methods* for additional details. These hybrid classifiers reliably out-performed the other classifiers we examined on all experimental conditions except the rest condition (intact, paragraph, and word conditions: $ts(99) > 14, ps < 10^{-8}$; Fig. 4A). This finding is consistent with the notion that activity-based and ISFC-based patterns contain *different* information about the story moments people were listening to.

To follow up on this result, we set out to determine the optimal mixing proportion of the two types of features. We defined a mixing parameter, ϕ, where $\phi = 0$ corresponds to activity-based decoding; $\phi = 1$ corresponds to ISFC-based decoding; and $0 < \phi < 1$ reflects a weighted mixture of activity- and ISFC-based decoding. We then re-ran the "mixture" decoding analysis in Figure 4A using 21 linearly spaced values of ϕ ranging between 0 and 1, inclusive. As shown in Figure 5A, the decoding accuracy (for the intact condition) peaked for $\phi = 0.25$, corresponding to a mix of 75% activity-based and 25% ISFC-based features.

These analyses indicate that, although HTFA results in a slight loss of cognitively relevant information, it enables efficient implementation of more sophisticated decoders that would other-

bioRxiv preprint doi: https://doi.org/10.1101/106690; this version posted February 7, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

wise be computationally expensive. These decoders (that incorporate correlational information, or a mix of activity-based and correlation-based information) out-perform decoders trained only on raw voxel activity patterns.

Case Study 3: full-brain networks are modulated by movie viewing

As a secondary test of our results from Case Study 2, we ran an analogous set of analyses on data collected as 17 participants viewed an episode from the BBC television show *Sherlock* [7]. (Experimental and imaging methods may be found in [7].) As in Case Study 2, we used our cross validation procedure to determine the optimal number of network nodes for the dataset (see *Materials and methods*). This procedure revealed that $K = 900$ nodes were optimal. We then applied HTFA to the full dataset, which comprised (for each participant) 43,371 voxel activations over 1976 TRs. The dataset is available here.

As shown in Figure 4B, all four classes of features we examined in Case Study 2 could also be used for this movie viewing dataset to reliably decode which timepoint participants were viewing: voxel activations ($t(99) = 131.6, p < 10^{-112}$; mean accuracy: 11.1%), HTFA factor activations ($t(99) = 120.4, p < 10^{-108}$; mean accuracy: 9.5%), HTFA-derived ISFC patterns ($t(99) = 125.0, p < 10^{-110}$; mean accuracy: 9.9%), and a 50-50 mixture of HTFA-derived factor activations and ISFC patterns ($t(99) = 197.4, p < 10^{-129}$; mean accuracy: 16.3%). Following our approach in Case Study 2, we divided the experiment into 60 TR sliding windows. Therefore chance performance on this test is $\frac{1}{1917}$ ($1976 - 60 + 1 = 1917$ unique sliding windows).

We also compared the relative performance of classifiers trained on these different classes of neural features. We found that voxel activation-based classifiers outperformed the HTFA factor activation-based classifiers ($t(99) = 13.7, p < 10^{-29}$) and HTFA-derived ISFC-based classifiers ($t(99) = 9.6, p < 10^{-17}$). However, classifiers that leveraged mixtures of activation-based and ISFC-based features reliably outperformed voxel activation-based classifiers ($t(99) = 44.7, p < 10^{-104}$). We also carried out an exploratory analysis to estimate the optimal mixing proportions of activity-based and ISFC-based features (Fig. 5B). The classifiers with $\phi = 0.3$ (reflecting a mixture of 70% activity-based and 30% ISFC-based features) performed best. These analyses provide additional evidence that the HTFA-derived activity and connectivity patterns contain partially non-overlapping information about the moments in the movie the participants were viewing.

Discussion

We proposed HTFA, a probabilistic approach to discovering and examining full-brain patterns of functional connectivity in multi-subject fMRI datasets. In Case Study 1, we used a synthetic dataset to demonstrate HTFA's ability to recover known patterns in synthetic data, and in Case Studies 2 and 3 we used a real fMRI dataset to demonstrate how HTFA may be used to efficiently find and analyze full-brain connectivity patterns in real data.

bioRxiv preprint doi: https://doi.org/10.1101/106690; this version posted February 7, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Related approaches

Although most modern computer systems are capable of computing and storing (in RAM) a single subject's voxel-to-voxel connectivity matrix, most commonly available systems cannot store many such matrices. Researchers have used several techniques to reduce the computational load. The most straightforward methods entail pre-selecting a small number of ROIs (e.g. motor cortex; [4]) or a seed voxel (e.g. a single voxel within the posterior cingulate; [19]). This reduces the connectivity matrix from a $V \times V$ matrix (where, as above, V is the total number of voxels in each brain volume) to a much smaller $V_{ROI} \times V$ matrix (where V_{ROI} is the number of seed or ROI voxels). However, reducing the connectivity matrix in this way precludes finding connectivity patterns unrelated to the ROI or seed region. For example, if the analysis is limited to connectivity patterns between motor cortex and the rest of the brain, this precludes finding patterns of connectivity that do not involve the motor cortex (e.g. connectivity between prefrontal cortex and the hippocampus).

A related technique that does not require pre-selecting ROIs or seed regions is to compute the full voxel-to-voxel connectivity matrix, and then to threshold the connection strengths such that one only examines the most reliable connections [6]. This yields a sparse voxel-to-voxel connectivity matrix that may be easily manipulated (provided that it is sufficiently sparse). One drawback to this approach is that it is not always clear how to set the connectivity strength threshold; for example, setting too high a threshold will leave out potentially important patterns, whereas setting too low a threshold will not substantially reduce the computational burden (as compared with examining the original voxel-to-voxel connectivity matrix). More generally, it is not clear that the strongest connections are necessarily the most informative; for example, a sub-threshold connection may carry information that could have been used to gain insights into the subject's cognitive state.

Other approaches have focused on reducing the dimensionality of the connectivity patterns (see [34] for a comparison between threshold–based and dimensionality-reduction–based approaches). For example, clustering-based approaches attempt to group together voxels that exhibit similar activation patterns over time, across subjects [9]. HTFA takes the approach of describing each image as a weighted sum of factors (where the number of factors is typically much smaller than the number of voxels in the original images). One may then examine connectivity patterns between the factors rather than between the voxels (this approach was originally developed in the positron emission tomography literature using PCA [14]).

Unlike standard approaches such as PCA and ICA, HTFA factors are constrained to be RBFs. Constraining HTFA factors to be RBFs has deeper geometric implications that we elaborate on in the next sub-section. Having RBF factors also makes HTFA especially well-suited to representing connectivity patterns, since each factor is spatially compact. In this respect, HTFA is qualitatively similar to spatial ICA [5]. However, unlike spatial ICA, HTFA further constrains factors to be in similar locations across subjects, providing a natural means of combining or comparing connectivity across individuals.

bioRxiv preprint doi: https://doi.org/10.1101/106690; this version posted February 7, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Benefits and costs of our approach

HTFA was designed to develop efficient ways of examining full-brain connectivity in multi-subject datasets. What would it have taken to study the same ISFC and patterns we examined in *Case study 2* using voxel-based methods? After masking out non-gray matter voxels and warping every subject's data to a common brain space (see *Materials and methods*), each fMRI volume comprises 44,415 voxels. Therefore each voxelwise ISFC matrix contains roughly 986 million unique entries ($\frac{V^2-V}{2}$). Assuming single precision storage (4 bytes per number), the ISFC matrix for each timepoint would require roughly 3.9GB of memory. By contrast, each HTFA-derived ISFC matrix for 700 sources requires fewer than 250 thousand unique entries (1 MB). In general, for decoders that combine activity-based and ISFC-based features, the computational savings gained by using HTFA-derived networks (of K factors) rather than the original voxel data is given by $\frac{V^2+V}{K^2+K}$. For our Case Study 2, the HTFA-derived ISFC analyses in Figure 4A were approximately 4,000 times more computationally efficient than the corresponding voxel-based analyses would have been. Following this logic, we achieved a similar computational efficiency savings for Case Study 3; Fig. 4B. Although HTFA facilitates efficient computations of full-brain correlation patterns, we note that with sufficient computing power, optimized approaches have been developed for computing these matrices directly [33].

Beyond computational efficiency, the fact that HTFA's parameters exist in real space, rather than being tied to the voxel locations, brings additional benefits. As an illustrative example, suppose that (in some brain region), two subjects exhibit identical patterns of activity across images. Now suppose that we were to randomly perturb all of the odd voxels from the first subject and all of the even voxels from the second subject. Because basic voxel-based methods consider only the voxel *labels* rather than the voxel *locations*, perturbing the data in this way would perfectly disrupt the correlation between the two subjects' activity patterns in that brain region. In contrast, from the perspective of location-based models like HTFA, the two subjects' activation patterns would still look similar. Even though HTFA biases the subject-specific RBFs to be in similar locations, they are still allowed some flexibility in order to explain subject-specific idiosyncrasies (Fig. 1). Similar intuitions have led researchers to apply spatial smoothing to fMRI data (typically by convolving the images with a Gaussian kernel). However, it is not always clear *a priori* how much smoothing one should apply, nor is it clear that the amount of smoothing that would be "best" in one part of the brain would also be the best in another part of the brain. Another approach is to use shared response models such as *hyperalignment* [20] to identify the affine transformations that bring different subjects' voxel data into a common representational space. This approach assumes that all participants viewed the same stimulus (or sequence of stimuli); if so, hyperaligning the "odd" and "even" subjects in the above example would align their responses appropriately. HTFA addresses these issues by automatically determining the appropriate degree of effective smoothing in each part of the brain (via the RBF widths), without requiring all participants to have the same number of images or to have viewed the same stimuli.

Another benefit of HTFA's separation from voxel space is its ability to naturally fill in missing observations (a property we exploit to determine the optimal number of factors). Techniques like probabilistic PCA [31] can fill in missing voxel activations using the data covariance matrix, provided that we observe at least *some* activations from those missing voxels (in other images).

bioRxiv preprint doi: https://doi.org/10.1101/106690; this version posted February 7, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

However, suppose that *all* activations from a given voxel were missing– or more realistically, suppose that we wish to estimate what the activations would have been at any arbitrary point in space. Because PCA does not explicitly represent the voxels' spatial locations, neither PCA nor probabilistic PCA can accurately predict activation patterns at these never-observed voxels. HTFA, by contrast, naturally predicts the missing data by simply evaluating each factor's RBF at the corresponding location in space.

These missing data examples also provide insights into other benefits of allowing factors to exist in real space rather than considering only the set of voxel locations. For example, HTFA allows for different subjects' data to be sampled at different resolutions, or to contain different numbers of voxels. In principle, although we have not explored this possibility formally in this paper, different subjects' data may even come from different recording modalities (e.g. one subject may contribute fMRI data and another may contribute EEG data). In this way, HTFA provides a common framework for describing neural data in general that transcends the specifics of the recording (modality, spatial or temporal resolution, etc.). For additional discussion of the benefits of spatial-based (rather than voxel-based) factors see [16], and for an example of how similar models may be used to analyze EEG data see [17].

We also note that using HTFA to examine connectivity patterns may not always out-perform voxel-based approaches. In particular, to the extent that the relevant patterns are high spatial frequency (at the level of single voxels), those patterns will be better described by voxel-based approaches than RBF factors. (Representing brain images as sums of RBF factors effectively blurs out the images in space, where the amount of blurring is inversely proportional to the number of factors.) To address this issue, we developed an algorithm for estimating the minimum number of sources required to reliably describe the data up to a desired level of precision (see *Materials and methods*).

Concluding remarks

By providing a highly efficient means of examining full-brain functional connectivity, HTFA makes it possible to examine connectivity patterns using a wide range of techniques. The possibility of applying polynomial time and space algorithms (such as pattern classification algorithms) to connectivity data is particularly exciting.

Materials and methods

Applying HTFA to multi-subject fMRI datasets

We use a *maximum a posteriori* (MAP) estimation procedure to compute the most probable RBF factors and factor weights. The procedure has three basic steps: initialization (during which we set the prior over each factor), fitting subject-specific parameters for each subject (given the prior), and updating the global template (using the subject-specific parameters). When we carry out the full inference procedure, we perform an initial cross-validation procedure (described next) to determine the optimal number of factors, K. We then randomly initialize K RBF factors by drawing from the prior distribution (Alg. 1). Finally, we iterate between

bioRxiv preprint doi: https://doi.org/10.1101/106690; this version posted February 7, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

updating the subject-specific parameters (using the current global template as the prior) and the global template (using the latest estimates of the subject-specific parameters) until the largest change in any parameters value from the previous iteration to the current iteration is less than a pre-determined threshold value, ϵ. (We typically set ϵ to be the length of the longest voxel dimension.)

Each hidden variable, $x \in \mathbf{\Omega}$, in the model (unshaded circles in Fig. 2) comes from a Gaussian with a mean parameter μ_x and either a covariance parameter Σ_x (for the RBF centers) or variance parameter σ_x^2 (for the RBF widths and factor weights). The covariance and variance parameters are hyperparameters (initialized to a fixed value).

At a descriptive level, the algorithm works by alternating between two steps: (1) update the subject-specific centers and widths given the global centers and widths, and (2) update the global centers and widths given the (new) subject-specific centers and widths. These two steps repeat until the global centers and widths stop changing (i.e. the maximum change across all latent variables from one round updates to the next is less than a pre-defined threshold). To update the subject-specific centers and widths, we perform an additional two alternating steps, whereby in the first step we update the per-image factor weights (holding the centers and widths fixed) and in the second we update the centers and widths (holding the factor weights fixed). Once the global centers and widths have stopped changing, we run an additional step whereby we re-compute the per-image factor weights for each subject. We developed a highly efficient implementation of our algorithm for applying HTFA to large multi-subject fMRI datasets [1]; additional details may be found there.

Estimating the optimal number of factors

When K (the number of factors) equals V (the number of voxels), HTFA can exactly recover the data by setting the RBF widths to be very small and the per-image factor activations equal to the voxel activations. Therefore setting $K = V$ represents one logical extreme whereby HTFA loses no information about the original data but also achieves no gains in how efficiently the data are represented. At the other extreme, when $K = 1$, HTFA achieves excellent compression but poor reconstruction accuracy (only a single activation value may be represented for each image). In practice, we will typically want to set K to some value between these extremes. Specifically, we want to choose the minimum K that is expected to explain the data up to a pre-defined level of precision, q.

Given a multi-subject dataset, we first select $s_{training}$ subjects at random (from the full set) to participate in the cross validation procedure. For each of those subjects, we select (at random) a set of $n_{training}$ images from each subject's data (we set this number to be 70% of the subject's images). We fit HTFA to these randomly selected images to estimate the subject-specific centers ($\mu_{1...K,1...S}$) and widths ($\lambda_{1...K,1...S}$). Next, of the remaining $n_{test} = N_s - n_{training}$ images for each subject, we select (at random) 70% of the voxels to estimate the per-image factor weights, $\mathbf{w}_{1...N,1...K,1...S}$. Finally, we use the estimated centers, widths, and weights to reconstruct the voxel activations for the remaining 30% of the voxels in those n_{test} images. The mean squared error between the reconstructed and true (observed) voxel activations provides an error signal that we can use to optimize K. In particular, starting from a minimum value of $K = \delta_K$, we

bioRxiv preprint doi: https://doi.org/10.1101/106690; this version posted February 7, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

use the above procedure to compute the mean squared error for δ_K. We then increase K (in increments of δ_K) until the mean squared error is less than our pre-defined threshold, q. (In this paper we set $\delta_K = 10$ for synthetic data, $\delta_K = 100$ for real data and $q = 0.25$.)

Estimating full-brain inter-subject functional connectivity

The above parameter inference procedure yields, for each subject, an N_s by K matrix, $\mathbf{W}_s$, of per-image factor weights. We can estimate the functional connectivity between each pair of factors by computing the correlation between the columns of $\mathbf{W}_s$. This approach is analogous to standard voxel-wise techniques for estimating functional connectivity [4]. Further, because the columns of $\mathbf{W}_{1...S}$ correspond to the same factors across the different subjects, since all of the factors are linked through the global template, the set of these weight matrices provide a convenient means of testing hypotheses related to the connectivity strengths.

In our analyses for Case Study 2 and 3, we used ISFC [29] to isolate the time-varying correlational structure (functional connectivity patterns) that was specifically driven by the story participants listened to. We first applied HTFA to the fMRI dataset to obtain a time series of source activations for every participant (where $K = 700$ for Case Study 2 and $K = 900$ for Case Study 3). We sought to obtain ISFC matrices for each of a series of overlapping temporal windows (*sliding windows*). To do so, we performed the following analysis for each sliding window (which contained a time series of source activations for each source and participant). For each participant, we computed the correlation between the activations of each source from that participant (during that sliding window) and the average activations of every source during the same window (where the average was taken across all of the other participants). The result, $C_{s,t}$ was a K by K correlation matrix for a single participant (s), during a single sliding window (t). We computed the ISFC matrix (across participants) during time t as:

$$\bar{C}_t = R\left(\frac{1}{2S}\sum_{s=1}^{S} Z\left(C_{s,t}\right)^{\mathrm{T}} + Z\left(C_{s,t}\right)\right), \tag{9}$$

where Z is the Fisher z-transformation [36]:

$$Z(r) = \frac{\log(1+r) - \log(1-r)}{2} \tag{10}$$

and R is the inverse of Z:

$$R(z) = \frac{\exp(2z-1)}{\exp(2z+1)}. \tag{11}$$

For additional details and discussion of ISFC see [29].

Decoding analysis

We sought to estimate whether the moment-by-moment HTFA-derived ISFC patterns we identified in Case Studies 2 and 3 were reliably preserved across participants, and (in Case Study 2) whether

bioRxiv preprint doi: https://doi.org/10.1101/106690; this version posted February 7, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

the degree of agreement across participants was modulated according to the cognitive salience of the stimuli participants experienced. For example, prior work has shown that different participants exhibit similar responses while experiencing richly structured stimuli (such as story listening), whereas participants exhibit less stereotyped responses while experiencing less structured stimuli (such as resting with their eyes open in the scanner) [29].

To study these phenomena, we randomly divided participants in Case Study 2 into two groups, for each experimental condition: intact (i.e., participants who listened to the original story recording), paragraph-scrambled (i.e., participants who listened to an altered recording where the paragraphs occurred in a randomized order), word-scrambled (i.e., participants who listened to an altered recording where the words occurred in a randomized order), and rest (i.e., participants who rested with their eyes open in the scanner, without listening to any story). Participants within each condition experienced the same auditory stimuli, but the cognitive salience (i.e. how meaningful the stimuli were) varied systematically across these experimental conditions. (For the experiment presented in Case Study 3, every participant experienced the equivalent of the "intact" condition.)

For each experimental condition, we computed the mean voxel or source activations within each sliding window (this resulted in either V-dimensional or K-dimensional vector for each moment of the story). For each group of participants in turn, we used correlations between activity patterns to estimate the story times each pattern corresponded to. In the activity-based analyses shown in Figure 4 (first two bar colors), we used these activity vectors to decode which moments of the story participants were listening to. Specifically, we asked, for each sliding window (t): what are the correlations between the first group's activity pattern at time t and the second group's activity patterns in *every* sliding window (this yielded one correlation value per sliding window). We used the best-matching pattern (i.e. the activity pattern with the strongest positive correlation) to estimate which story time sliding window t corresponded to.

We used a similar approach to examine moment-by-moment ISFC patterns for each of the two groups of participants (i.e., for each condition, we obtained one ISFC pattern for each sliding window, for each of the two groups). For the ISFC analysis shown in Figure 4, we reshaped these ISFC patterns into vectors, and used the same correlation-based technique to label each group's sliding windows according to how well they matched the ISFC patterns in the other group's sliding windows.

Finally, we carried out a mixed activity-based and correlation-based decoding analysis by combining the estimates of the two above decoders. Specifically, for each sliding window (from one group of participants), we computed the correlations between the ISFC patterns from each sliding window from the other group, and the correlations between the activity patterns from each sliding window of the other group. We used the average of these two correlations to label each group's timepoints in the "mixed" decoding analysis shown in Figure 4 (rightmost bar color). In Figure 5 we extended this analysis by changing the relative weights of the activity-based and ISFC-based decoders using a mixing parameter, ϕ, where $\phi = 0$ corresponds to activity-based decoding; $\phi = 1$ corresponds to ISFC-based decoding; and $0 < \phi < 1$ reflects a weighted mixture of activity- and ISFC-based decoding

bioRxiv preprint doi: https://doi.org/10.1101/106690; this version posted February 7, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Acknowledgements

We acknowledge useful discussions with Jonathan Cohen, Janice Chen, Justin Hulbert, Talia Manning, Peter Ramadge, Erez Simony, and Nicholas Turk-Browne. This work was supported by the NSF/NIH Collaborative Research in Computational Neuroscience Program, grant number NSF IIS-1009542; NSF EPSCoR Award Number 1632738; and a grant from the Intel Corporation. The content is solely the responsibility of the authors and does not necessarily represent the official views of our supporting organizations.

References

[1] M. J. Anderson, M. Capota, J. S. Turek, X. Zhu, T. L. Willke, Y. Wang, P.-H. Chen, J. R. Manning, P. J. Ramadge, and K. A. Norman. Enabling factor analysis on thousand-subject neuroimaging datasets. In L. L. X. H. R. A. Y. X. W. X. A.-H. S. S. R. L. U. P. S. Y. R. G. T. S. James Joshi, George Karypis, editor, *Proceedings of the IEEE International Conference on Big Data*, pages 1242–1251, 2016.

[2] R. E. Betzel, J. D. Medaglia, L. Papadopoulos, G. Baum, R. Gur, R. Gur, D. Roalf, T. D. Satterthwaite, and D. S. Bassett. The modular organization of human anatomical brain networks: accounting for the cost of wiring. *Network Neuroscience*, page Advance online publication. doi:10.1162/netn_a_00002., 2017.

[3] C. Bishop. *Pattern recognition and machine learning*. Springer, 2006.

[4] B. Biswal, F. Z. Yetkin, V. M. Haughton, and J. S. Hyde. Functional connectivity in the motor cortex of resting human brain using echo-planar MRI. *Magnetic Resonance in Medicine*, 34(4):537 – 541, 1995.

[5] V. D. Calhoun, T. Adali, G. D. Pearlson, and J. J. Pekar. Spatial and temporal independent component analysis of functional MRI data containing a pair of task-related waveforms. *Human Brain Mapping*, 13:43–53, 2001.

[6] J. Cao and K. J. Worsley. The geometry of correlation fields, with an application to functional connectivity of the brain. *Ann. Appl. Probab.*, 9:1021–1057, 1999.

[7] J. Chen, Y. C. Leong, K. A. Norman, and U. Hasson. Shared experience, shared memory: a common structure for brain activity during naturalistic recall shared experience, shared memory: a common structure for brain activity during naturalistic recall. *Nature Neuroscience*, 20(1):115–125, 2017.

[8] P. Comon, C. Jutten, and J. Herault. Blind separation of sources, part II: Problems statement. *Signal Processing*, 24(1):11 – 20, 1991.

[9] D. Cordes, V. M. Haughton, K. Arfanakis, J. D. Carew, and K. Maravilla. Hierarchical clustering to measure connectivity in fMRI resting-state data. *Proc. Intl. Soc. Mag. Reson. Med.*, 20(4):305–317, 2002.

bioRxiv preprint doi: https://doi.org/10.1101/106690; this version posted February 7, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

[10] R. C. Craddock, G. A. James, I. P E Holtzheimer, X. P. Hu, and H. S. Mayberg. A while brain fmri atlas generated via spatially constrained spectral clustering. *Human Brain Mapping*, 33(8):1914–1928, 2012.

[11] M. D. Fox and M. Greicius. Clinical applications of resting state functional connectivity. *Frontiers in Systems Neuroscience*, 4(10):1–13, 2010.

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

[13] K. Friston, C. Price, P. Fletcher, C. Moore, R. Frackowiak, and R. Dolan. The trouble with cognitive subtraction. *NeuroImage*, 4:97 – 104, 1996.

[14] K. J. Friston, C. D. Frith, P. F. Liddle, and R. S. J. Frackowiak. Functional connectivity: the principal-component analysis of large (PET) data sets. *Journal of Cerebral Blood Flow and Metabolism*, 13:5–14, 1993.

[15] A. Gelman and J. Hill. *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press, 2007.

[16] S. Gershman, K. Norman, and D. Blei. A topographic latent source model for fMRI data. *NeuroImage*, 57:89 – 100, 2011.

[17] S. J. Gershman, D. M. Blei, K. A. Norman, and P. B. Sederberg. Decomposing spatiotemporal brian patterns into topographic latent sources. *NeuroImage*, 98:91–102, 2014.

[18] J. Gonzalez-Castillo, C. W. Hoy, D. A. Handwerker, M. E. Robinson, L. C. Buchanan, Z. S. Saad, and P. A. Bandettini. Tracking ongoing cognition in individuals using brief, whole-brain functional connectivity patterns. *Proceedings of the National Academy of Science USA*, 112(28):8762–8767, 2016.

[19] M. D. Greicius, B. Krasnow, A. L. Reiss, and V. Menon. Functional connectivity in the resting brain: a network analysis of the default mode hypothesis. *Proceedings of the National Academy of Science USA*, 100(1):253–258, 2003.

[20] J. S. Guntupalli, M. Hanke, Y. O. Halchenko, A. C. Connolly, P. J. Ramadge, and J. V. Haxby. A model of representational spaces in human cortex. *Cerebral Cortex*, 2016.

[21] J. V. Haxby, M. I. Gobbini, M. L. Furey, A. Ishai, J. L. Schouten, and P. Pietrini. Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, 293:2425–2429, 2001.

[22] J. V. Haxby, J. S. Guntupalli, A. C. Connolly, Y. O. Halchenko, B. R. Conroy, M. I. Gobbini, M. Hanke, and P. J. Ramadge. A common, high-dimensional model of the representational space in human ventral temporal cortex. *Neuron*, 72:404–416, 2011.

bioRxiv preprint doi: https://doi.org/10.1101/106690; this version posted February 7, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

[23] C. Jutten and J. Herault. Blind separation of sources, part I: An adaptive algorithm based on neuromimetic architecture. *Signal Processing*, 41(1):1 – 10, 1991.

[24] J. R. Manning, R. Ranganath, K. A. Norman, and D. M. Blei. Topographic factor analysis: a Bayesian model for inferring brain networks from neural data. *PLoS One*, 9(5):e94914, 2014.

[25] K. A. Norman, S. M. Polyn, G. J. Detre, and J. V. Haxby. Beyond mind-reading: multi-voxel pattern analysis of fMRI data. *Trends in Cognitive Sciences*, 10(9):424–430, 2006.

[26] K. Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2:559–572, 1901.

[27] J. D. Power, A. L. Cohen, S. M. Nelson, G. S. Wig, K. A. Bames, J. A. Church, A. C. Vogel, T. O. Laumann, F. M. Miezin, B. L. Schlaggar, and S. E. Peterson. Functional network organziation of the humna brain. *Neuron*, 72:665–678, 2011.

[28] M. Rubinov and O. Sporns. Complex network measures of brain connectivity: uses and interpretations. *NeuroImage*, 52:1059 – 1069, 2010.

[29] E. Simony, C. J. Honey, J. Chen, and U. Hasson. Uncovering stimulus-locked network dynamics during narrative comprehension. *Nature Communications*, 2016.

[30] C. Spearman. General intelligence, objectively determined and measured. *Americal Journal of Psychology*, 15:201 – 293, 1904.

[31] M. E. Tipping and C. M. Bishop. Probabilistic principal component analysis. *Journal of Royal Statistical Society, Series B*, 61(3):611–622, 1999.

[32] N. B. Turk-Browne. Functional interactions as big data in the human brain. *Science*, 342:580–584, 2013.

[33] Y. Wang, J. D. Cohen, K. Li, and N. B. Turk-Browne. Full correlation matrix analysis (FCMA): a high-performance toolbox and case study for unbiased functional connectivity in human brain imaging. *Submitted*, 2014.

[34] K. J. Worsley, J.-I. Chen, J. Lerch, and A. C. Evans. Comparing functional connectivity via thresholding correlations and singular value decomposition. *Philosophical Transactions of the Royal Society B*, 360:913–920, 2005.

[35] B. T. T. Yeo, F. M. Krienen, J. Sepulcre, M. R. Sabuncu, D. Lashkari, M. Hollinshead, J. L. Roffman, J. W. Smoller, L. Zollei, J. R. Polimieni, B. Fischl, H. Liu, and R. L. Buckner. The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *Journal of Neurophsiology*, 106(3):1125–1165, 2011.

[36] J. H. Zar. *Biostatistical analysis*. Prentice-Hall/Pearson, 2010.

[37] E. Zarahn, G. Aguirre, and M. D'Esposito. A trial-based experimental design for fMRI. *NeuroImage*, 6:122 – 138, 1997.