Neuropsychologia 144 (2020) 107500

Contents lists available at ScienceDirect

Neuropsychologia

journal homepage: www.elsevier.com/locate/neuropsychologia

Incorporating structured assumptions with probabilistic graphical models in fMRI data analysis

Ming Bo Cai[a,b,*,1], Michael Shvartsman[c,1], Anqi Wu[d,1], Hejia Zhang[e,1], Xia Zhu[f,1]

[a] International Research Center for Neurointelligence (WPI-IRCN), UTIAS, The University of Tokyo, Japan
[b] Princeton Neuroscience Institute, Princeton University, United States
[c] Facebook Reality Labs, United States
[d] Mortimer B. Zuckerman Mind Brain Behavior Institute, Columbia University, United States
[e] Department of Electrical Engineering, Princeton University, United States
[f] Intel Corporation, United States

ARTICLE INFO

Keywords:
Probabilistic graphical model
Bayesian
fMRI
Cognitive neuroscience
Big data
Factor model
Matrix normal

ABSTRACT

With the wide adoption of functional magnetic resonance imaging (fMRI) by cognitive neuroscience researchers, large volumes of brain imaging data have been accumulated in recent years. Aggregating these data to derive scientific insights often faces the challenge that fMRI data are high-dimensional, heterogeneous across people, and noisy. These challenges demand the development of computational tools that are tailored both for the neuroscience questions and for the properties of the data. We review a few recently developed algorithms in various domains of fMRI research: fMRI in naturalistic tasks, analyzing full-brain functional connectivity, pattern classification, inferring representational similarity and modeling structured residuals. These algorithms all tackle the challenges in fMRI similarly: they start by making clear statements of assumptions about neural data and existing domain knowledge, incorporate those assumptions and domain knowledge into probabilistic graphical models, and use those models to estimate properties of interest or latent structures in the data. Such approaches can avoid erroneous findings, reduce the impact of noise, better utilize known properties of the data, and better aggregate data across groups of subjects. With these successful cases, we advocate wider adoption of explicit model construction in cognitive neuroscience. Although we focus on fMRI, the principle illustrated here is generally applicable to brain data of other modalities.

1. Introduction

Functional magnetic resonance imaging (fMRI) (Ogawa et al., 1990; Belliveau et al., 1991) is a powerful tool to study the brain's activity and functions. The fluctuation of the fMRI signal is related to the fluctuation of the concentrations of the oxygenated and deoxygenated hemoglobin in the blood, which follows the increase or decrease of local neuronal activity with a delay (Buxton, 2013; Heeger and Ress, 2002). This relation to the neural activity, together with its non-invasive nature, full brain coverage and reasonable balance between spatial and temporal resolution, makes fMRI a widely used brain imaging technique for studying the neural correlates of perceptual and cognitive processes in humans.

However, deriving insights about neural information processing from fMRI data can be challenging in many situations, because (1) fMRI signals only indirectly relate to neural activity (Buxton, 2013; Friston et al., 1995); (2) the data typically contain noise and unknown physiological signals with complex spatial and temporal correlation, and various artifacts (Triantafyllou et al., 2005; Bright and Murphy, 2015; Zarahn et al., 1997); (3) the number of brain volumes scanned in each experiment is much smaller than the number of voxels (high dimensionality of data in contrast to small sample size); and (4) there is large variation in detailed brain anatomical structures and functional organization across people (Finn and Constable, 2016; Suárez et al., 2020), making it harder to aggregate data across people. Analysis tools need to take these factors into account in order to obtain fruitful insight from data. One promising approach to address these challenges is that of probabilistic graphical models. Probabilistic graphical models (PGM) are frameworks used to create probabilistic models of complex data distributions and represent them in compact graphical representation (Koller and Friedman, 2009). PGMs have been widely used in many different fields (Friedman, 2004; Ahelegbey, 2016; Ji, 2019). In behavioral studies of perception and cognition, PGM is not only a

[*] Corresponding author at: International Research Center for Neurointelligence (WPI-IRCN), UTIAS, The University of Tokyo, Japan.
E-mail address: mingbo.cai@ircn.jp (M.B. Cai).
[1] All authors contributed equally.

https://doi.org/10.1016/j.neuropsychologia.2020.107500
Received 31 October 2019; Received in revised form 9 May 2020; Accepted 15 May 2020
Available online 17 May 2020

framework for describing the computational processes in the brain (Ma, 2012; Griffiths et al., 2008; Geisler, 2011), but also a framework for testing different models against behavior (Kruschke, 2010; Shiffrin et al., 2008; Etz and Vandekerckhove, 2018). However, except in few domains, its value for neural imaging analysis has not been fully appreciated. With PGMs, we can explicitly assume the relations and dependencies between quantities of interest and the data, construct hierarchical generative models that posit how fMRI data is generated from mental processes, incorporate structural assumptions about the data and domain knowledge into the models, and take into account the uncertainty/noise in fMRI data explicitly. These are benefits that are hard to achieve in conventional fMRI analysis tools, which provide a universal set of analyses whose baseline assumptions may or may not apply to specific datasets. Fig. 1 describes the scheme of building PGMs for fMRI study. In this paper, we illustrate how to build PGMs to solve neural imaging analysis problems following this scheme.

As illustrated in Fig. 1, an approach which relies on explicit PGM typically involves four major steps. The first step is defining the problem: deciding what question is asked or what problem needs to be solved, and deciding what quantity allows one to answer the question or to characterize certain aspect of the brain. This is essentially the hypothesis generation step of hypothesis-driven science, but we emphasize it as a distinct step because the hypothesis needs to be precise enough to translate into a model in subsequent steps.

After the question is clearly defined, the second step is to make explicit assumptions of how the quantity of interest and experimental manipulations directly or indirectly contribute to the data to be analyzed, and to make assumptions of how variables of no interest (nuisance factors) may jointly impact the data. If there is domain knowledge of the properties of fMRI data that can help construct models of the data, it should be clearly stated at this step as well.

The third step is to translate these assumptions and domain knowledge into a computational model. Such computational models can often be described by probabilistic graphical models (PGM) (Koller and Friedman, 2009) composed of nodes and directed edges between nodes. The nodes and edges together form a graph. When building PGMs, the data, experimental manipulation, quantity of interest and nuisance factors that are considered in the assumptions of the previous step all become variables (either known or unknown) and are each represented by a node in the graph. The hypothesized relations between variables in the models are expressed as conditional probability of one variable given one or more other variables, and are represented by directed edges. Each edge is directed from one variable to another variable that is conditioned on it (i.e., the distribution of the variable at the head of an edge (arrow) depends on the variable at the tail of the edge). The domain knowledge is either captured by the prior distribution of certain variables in the graph, or in the form of the conditional dependencies. The probabilistic nature of such models makes them a natural choice for capturing the noise properties in the system and the potential uncertainty in the estimates of parameters from the researchers' perspective.

Once the PGM is built, the fourth step is to deploy computational techniques to estimate the unknown variables of interest in the model. This step essentially inverts the model by inferring variables of interest at the source of the directed edges in the graphical model. In some cases, inferring these variables serves to answer the original question by providing characterization of some aspect of brain activity. In other cases, when the scientific question is to test competing hypotheses, the competing hypotheses should be translated into PGMs that differ in either the range of values of some key variables or in the structures of the models. The selection of the winning model can be either based on classical statistical tests of the inferred values of the key variables (when each value is inferred independently), or based on the likelihood that each model can give rise to the data (model evidence), marginalizing unknown variables (MacKay and Mac Kay, 2003; Jeffreys, 1998). To approximate posterior distributions of latent variables (variables that are directly or indirectly causal to the nodes representing observable data) in the probabilistic graphical models given the observed data, techniques such as Markov Chain Monte Carlo (MCMC) (Metropolis et al., 1953; Hastings, 1970) or variational Bayes (Jordan et al., 1999) are often employed (Gelman et al., 2013). In certain cases, when the posterior distribution of these latent variables can be analytically derived, exact inference of the posterior distribution or the *maximum a posteriori* values of the variables can often be achieved. A full discussion of the inference methods is out of the scope of this paper. Interested readers may refer to tutorials such as Chapter 8 of Bishop (2006) or part II of Koller and Friedman (2009).

Because a PGM is explicitly built, it is easy to evaluate whether the inference procedure can reliably recover the variables of interest in the model, by simulating data according to the model and comparing the recovered values of those variables with the values used in the simulation. In contrast, traditional approaches without building an explicit model of the data generating process lack the ability to simulate data in accordance with its (implicit) assumptions. Without simulating data, it becomes impossible to verify that an analysis can yield correct results, because researchers are only left with real neural data of which the generative process and the ground truth of the variables of interest are not known. Thus, there is no guarantee that the quantity extracted by analysis methods without explicit assumptions of data generating process bears direct relation to what the researchers are interested in.

In addition to transparency and verifiability, PGMs offer the flexibility to combine the advantage of various pieces of domain knowledge of the brain (for example, brain activation patterns tend to be spatially smooth). This is because domain knowledge can be translated into a prior distribution of certain form over some latent variables in the PGM. With the PGM as a backbone, different prior distributions may act as add-on parts that can be plugged in at different places of the model, depending on what domain knowledge is proper for the purpose of analysis. For example, in 2.2, we show that the smoothness assumption of the brain activity and the similarity of brain networks across people are incorporated as the 3D Gaussian shape of the spatial basis for brain patterns and the Gaussian distribution of node location across subjects, respectively. In 2.3, we show that two types of prior knowledge about fMRI decoding weights, smoothness and sparsity, can also be incorporated together by assuming a Gaussian process prior on the joint distribution of the fMRI decoding weights of all voxels.

In the following, we select example analysis methods developed in different research domains to illustrate how the PGM approach to neural imaging data can be applied to discovering shared neural dynamics across participants doing the same task, modeling the functional connectivity among brain regions, improving the performance of decoding mental contents and obtaining more biologically informed decoding weights, reducing the bias in estimating similarity among activation patterns, and providing more comprehensive model of the noise in fMRI data. These methods together illustrate how the PGM (Koller and Friedman, 2009) can accommodate domain knowledge and known properties of the data and facilitate aggregating information over larger datasets. These features allow us to mitigate the limitations in fMRI data: high dimensionality (many voxels), low sample size in single participant, heterogeneity across participants and high noise. Although introducing a PGM is not the only way to overcome these limitations, it is one of the simplest ways to achieve transparency, verifiability and interpretability, due to explicit modeling.

Because the focus is on illustrating the principle of PGMs, this paper can by no means provide a thorough review of all PGM-based methods of fMRI analysis. For example, the use of dynamic causal modeling (Friston et al., 2003; Stephan et al., 2010) to infer the interactive relations among brain regions, the construction of encoding models (Naselaris et al., 2009; Nishimoto et al., 2011; Naselaris et al., 2011; Dumoulin and Wandell, 2008) to understand the features encoded by a brain region or to reconstruct perceived sensory inputs, and the development of a probabilistic event segmentation model (Baldassano

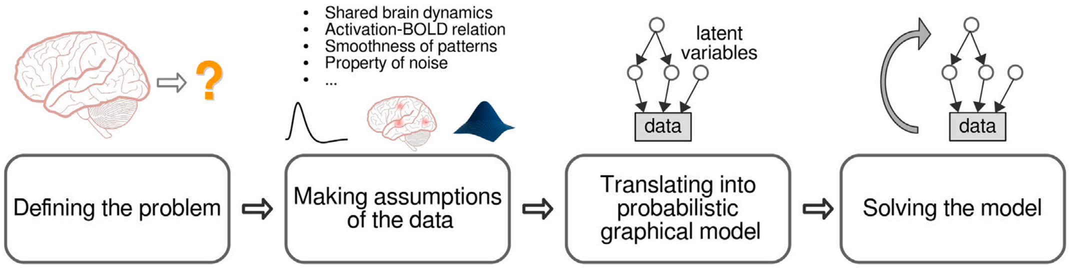

Fig. 1. The PGM-based approach to analyze neural data. In general, this involves four steps: (1) clearly defining the problem to solve or the question being asked; (2) making assumptions about the property of the data, including domain knowledge about data and causal relation between latent variables and measured data; (3) translating these assumptions to a probabilistic graphical model which expresses how latent variables together generate measured data. The model uses conditional probability distributions between variables to capture their causal relations; (4) solving the model to infer latent variables or to draw conclusion for the question being asked in the first step.

et al., 2017) to discover distinctive and sustained brain states, are all good illustrations of the four-step procedure above. Readers are encouraged to also refer to several other reviews (e.g., Woolrich (2012), Woolrich et al. (2009)) on Bayesian approaches to fMRI for a more comprehensive understanding of other existing PGM tools that share the advantages illustrated here.

While we advocate for explicit construction of probabilistic models in fMRI analysis, PGM-based methods do still have limitations. For example, performing efficient inference on PGMs can still scale poorly in both computation and memory, limiting their use on large-scale data without specially-tuned algorithms and approximations. Furthermore, deriving the inference algorithm for any specific model has historically required extensive knowledge in computer science or statistics, and it is not always easy to establish how robust any specific model is to mismatches between the assumptions made in the models and the true data properties. But the overall outlook is encouraging: algorithmic improvements have enabled scaling up models previously considered impossible (e.g. Wang et al. (2019)), general tools for probabilistic inference have blurred the lines between practitioner and methods developer (e.g. Carpenter et al. (2017), Salvatier et al. (2016), Bingham et al. (2019)), and advances in these aspects of PGM are an area of active research that will continue within and outside the neuroscientific domain.

2. Examples of PGM-based analysis methods for fMRI data

2.1. Discovering latent neural dynamics for naturalistic task

- Defining the problem: aggregating multi-subject fMRI data

fMRI datasets with naturalistic stimuli, such as movies or audiobooks, usually have limited number of samples per subject. In general, fMRI datasets not only have a large number of voxels, but also tend to have a small number of time points due to the limitation of samples per experiment session as a result of the slowness of the hemodynamic response and limited sample rate of the scanner. In fMRI datasets with naturalistic stimuli, it is also infeasible to collect many samples from a single subject when the experiments require the natural stimulus to be fresh to the subjects, so each subject could only be exposed to the same stimulus once. Therefore, to improve analysis sensitivity, we need to aggregate data from multiple subjects with the same stimulus effectively. The idea is similar to repeated-measures designs in neuroscience where the same variable is measured multiple times, but here the repetition is over different subjects. In our fMRI analysis application, we want to find what is common across subjects. The challenge is that the anatomical and functional structures between subjects are not aligned (Talairach, 1988). For example, when listening to the same music, a musician and a person without any music training will probably have different responses. Some early attempts applied pipelines such as averaging the fMRI data from all subjects after anatomical alignment, which assumes voxels of different brains have one-to-one correspondence (Talairach, 1988; Mazziotta et al., 2001). In contrast, the Shared Response Model (SRM) (Chen et al., 2015) is a Bayesian factor analysis model that finds the shared latent neural dynamics across subjects in a multi-subject fMRI dataset after anatomical alignment, without assuming one-to-one voxel correspondence.

- Making assumptions: temporally-aligned stimulus

SRM assumes that the stimulus in a naturalistic task dataset is temporally-aligned. That is, all the subjects receive the same stimulus at the same time point in the task. Therefore, we assume that all the subjects share the same low-dimensional latent representation within a dataset, called "shared response." On the other hand, to account for the differences between subjects, SRM assumes that each subject has a subject-specific spatial basis for generating the observed fMRI data from the shared response.

- Translating assumptions to a graphical model: shared response as a latent variable

To translate the assumptions above into a computational model, let us look at the deterministic SRM first and then the probabilistic version. The deterministic SRM factorizes the transpose of each subject's brain image data $\mathbf{X}_m^T$ into a subject-specific spatial basis $\mathbf{W}_m$ and the shared response $\mathbf{S}$ with the orthogonal constraint $\mathbf{W}_m^T\mathbf{W}_m = I$ (Fig. 2), where $\mathbf{X}_m \in \mathbb{R}^{T\times V_m}$ is the brain image data of subject m, $\mathbf{W}_m \in \mathbb{R}^{V_m\times K}$ is the subject-specific spatial basis of subject m, $\mathbf{S} \in \mathbb{R}^{K\times T}$ is the shared response across subjects, V_m is the number of voxels of subject m, T is the number of time points, and K is the number of features (the dimensionality of the shared response). K is a tunable hyper-parameter which is usually much smaller than T. More formally, deterministic SRM minimizes the Frobenius norm of reconstruction error

$$\left\|\mathbf{X}_m^T - \mathbf{W}_m\mathbf{S}\right\|_F^2 \tag{1}$$

under the constraint $\mathbf{W}_m^T\mathbf{W}_m = I$. This simple model is then extended to a probabilistic setting, as shown in Fig. 2. Here $x_{mt} \in \mathbb{R}^{V_m}$ denotes the observed brain image data of subject m at time t, $s_t \in \mathbb{R}^K$ denotes a shared latent random vector with

$$s_t \sim \mathcal{N}(0, \Sigma_s). \tag{2}$$

The distribution of x_{mt} conditioned on s_t is then

$$x_{mt}|s_t \sim \mathcal{N}(\mathbf{W}_m s_t + \mu_m, \rho_m^2 I), \tag{3}$$

where the subject-specific average μ_m accounts for non-zero mean and $\rho_m^2 I$ is the subject dependent isotropic noise covariance (for non-isotropic noise covariance in SRM, see 2.5). In the probabilistic version, the orthogonal constraint still holds. A constrained expectation–maximization (EM) algorithm is used to solve this model.

- Applications: identified shared responses and extensions

The SRM identifies the shared and individual responses in a multi-subject fMRI dataset with naturalistic tasks. The explicit structure of SRM makes the fMRI data it is applied to more interpretable: the extracted shared responses allow us to aggregate information from

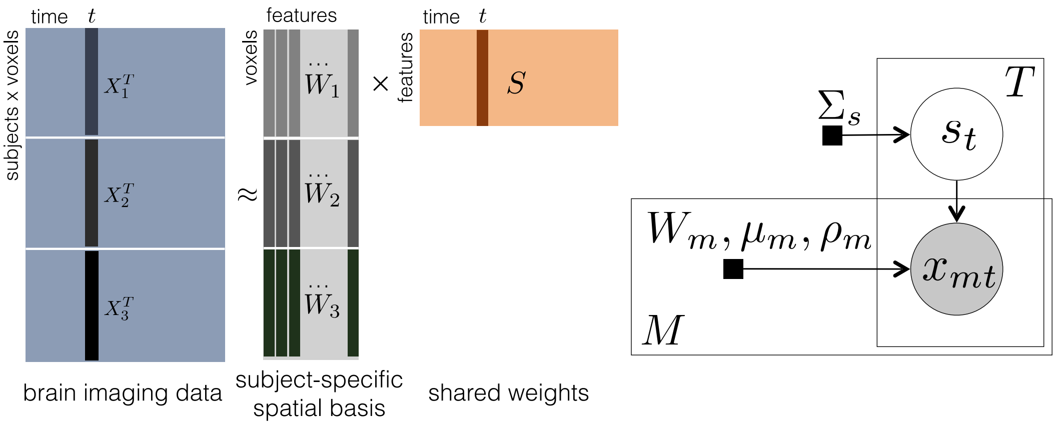

Fig. 2. Left: Illustration of deterministic SRM for three subjects. **Right:** Graphical model for SRM with M subjects and T time points (adapted from Chen et al. (2015)). Brain image data $x_{mt} \in \mathbb{R}^{V_m}$ (V_m voxels) is observed from subject m at time t, $t = 1 : T, m = 1 : M$. Each observation x_{mt} is a linear combination of subject-specific orthogonal basis (columns of $\mathbf{W}_m$) using the weights specified by s_t. The two plates are repeated T and M times, respectively. Shaded nodes: observations, unshaded nodes: latent variables, and black squares: parameters.

multiple subjects, and the individual responses could be used to identify what is unique for each subject. SRM shows improved performance in various tasks, such as image-viewing fMRI data classification, using shared and individual responses, as described in Chen et al. (2015), and movie scene classification (Vodrahalli et al., 2018).

Compared with hyperalignment (HA) (Haxby et al., 2011), SRM also has a built-in dimensionality reduction mechanism with a tunable number of features, where HA is an earlier multi-subject alignment algorithm with the objective to minimize

$$\left\|\mathbf{W}_m^T\mathbf{X}_m^T - \mathbf{S}\right\|_F^2 \tag{4}$$

under the constraint $\mathbf{W}_m^T\mathbf{W}_m = I$, $\mathbf{W}_m \in \mathbb{R}^{V_m \times V_m}$. Note that $\mathbf{W}_m$ is a square matrix here because HA aims to rotate each subject's $\mathbf{X}_m$ to match a global template S. More importantly, if $\mathbf{W}_m$ in HA is set to $\mathbb{R}^{V_m \times K}$ as in SRM, then it sometimes learns uninformative S. As illustrated in Chen et al. (2015), when performing an image stimulus classification experiment, HA with $\mathbf{W}_m \in \mathbb{R}^{V_m \times K}$ shows much lower testing accuracy than SRM.

Furthermore, SRM already has several extensions which make it more useful. For example, searchlight SRM (Zhang et al., 2016) combines SRM with searchlight analysis, which enables the localization of shared responses. Multi-dataset multi-subject analysis (MDMS) (Zhang et al., 2018) extends SRM to the multi-dataset setting where the model can aggregate information across subjects and datasets with different stimuli. Semi-supervised SRM (Turek et al., 2017) combines SRM with an additional multinomial logistic regression objective, such that the model can leverage partially labeled data. Matrix-normal SRM (Shvartsman et al., 2018), discussed below, makes different choices in modeling the residuals and the constraints on $\mathbf{W}_m$. All of these recent development of SRM illustrate the expandability of PGM-based method and the flexibility of adapting one PGM to different experimental settings and research purposes. In all these developments, the effectiveness of inverting a PGM can be verified by simulation, which is difficult for algorithms that do not explicitly build PGMs.

2.2. Discovering full-brain functional connectivity from fMRI

- Defining the problem: discovering full-brain functional connectivity

Recent research suggests that the functional connectivity (networks) in human brain, commonly represented by the spatial covariance structure of fMRI data, can change during different cognitive states (Turk-Browne, 2013). To estimate functional connectivity during a particular cognitive state (or an experimental condition) from fMRI data, one approach is to compute the correlation between the time series of pairs of voxels (Rubinov and Sporns, 2010). Because of the computational time and memory demanded by this voxel-based approach, most researchers focus their analysis on pre-selected regions of interest (ROIs). But this requires anatomically predefined ROIs which may not correctly capture the voxel-wise correlation. Voxel-based methods such as independent component analysis (Beckmann et al., 2005; Calhoun and Adali, 2012) generate statistically independent spatial maps and they are useful for applications that assume statistical independence between different neural sources. But each component discovered in this approach, or their combination, cannot be easily used to analyze spatially overlapping but functionally distinct activity patterns. Topographic Factor Analysis (TFA) (Manning et al., 2014) and Hierarchical Topographic Factor Analysis (HTFA) are Bayesian factor analysis models that can be used to efficiently analyze full-brain functional connectivity in large multi-subject neuroimaging datasets (Manning et al., 2018). Further, one of properties of HTFA is to generate spatially compact factors that partially overlap, and this property can help analyze and disentangle the contributions of activity patterns that are functionally distant but spatially overlapping.

- Making assumptions: spatial function-based latent factors

Both TFA (Manning et al., 2014) and HTFA (Manning et al., 2018) cast each subject's brain images as a linear combination of latent factors, where each latent factor is modeled as a parameterizable spatial function. Each latent factor can be interpreted as a node in a simplified representation of the brain's network. A subject's matrix of the changing weights on the nodes over time may be viewed as a low-dimensional embedding (or representation) of the original brain data. The pairwise correlations between each factor's weights over time further reflect the signs and strengths of the node-to-node connections (i.e. the functional connectivity). Both TFA and HTFA approximate each subject's functional connectivity by firstly representing each brain image in terms of the activities of a set of localized network nodes, and then computing the covariance of the activity. Furthermore, HTFA (Manning et al., 2018) is a multi-subject extension of TFA (Manning et al., 2014), and attempts to discover the network nodes that are common across a group of subjects. HTFA estimates a global template as well as each individual's subject-specific template. The global template describes where each common network node is placed, how wide it is and how active it tends to be. Each subject-specific template is a particular instantiation of the common network nodes and the subject's node activities.

- Translating assumptions to a graphical model: global and subject specific template

HTFA is formulated as a probabilistic latent variable model. Let $\mathbf{X}_m \in \mathbb{R}^{T_m \times V_m}$ represent subject m's data as a matrix with T_m fMRI

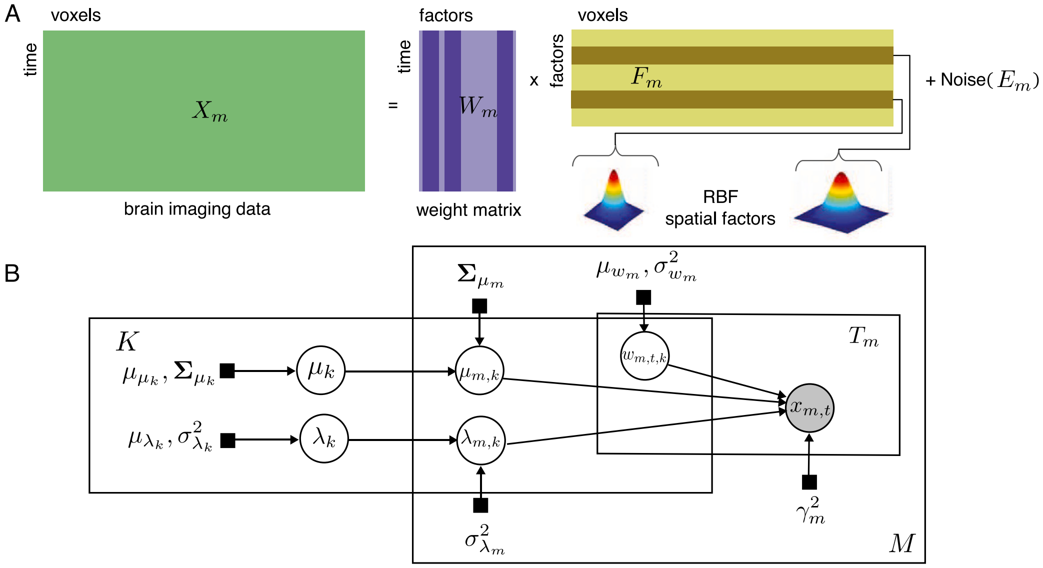

Fig. 3. (A)(H)TFA factor model. fMRI data $\mathbf{X}_m$ is decomposed into weight matrix $\mathbf{W}_m$ and factor matrix $\mathbf{F}_m$. Each factor is a RBF function. (B) Graphical model for HTFA, adapted from Manning et al. (2018) figure 2. Brain image data $x_{m,t} \in \mathbb{R}^{V_m}$ (V_m voxels) is observed from subject m at time t, $t = 1 : T_m, m = 1 : M$. Each observation $x_{m,t}$ is a linear combination of K subject-specific latent factors, using the weights specified by $w_{m,t,k}$. Each latent factor (row of F_m) is a spatial function of $\boldsymbol{\mu}_{m,k}$ and $\lambda_{m,k}$). The three plates are repeated K, T_m and M times, respectively. Shaded nodes: observations, unshaded nodes: latent variables, and black squares: parameters.

samples of the activity of V_m voxels, each sample being vectorized as one row in $\mathbf{X}_m$. Then, each subject is approximated with a factor analysis model

$$\mathbf{X}_m = \mathbf{W}_m\mathbf{F}_m + \mathbf{E}_m, \tag{5}$$

where $\mathbf{W}_m \in \mathbb{R}^{T_m \times K}$ are the weights of $\mathbf{F}_m \in \mathbb{R}^{K \times V_m}$, the latent factors. $\mathbf{E}_m$ is the noise term. Each latent factor (row of $\mathbf{F}_m$) is a radial basis function (RBF) with center at $\mu_{m,k}$ and width $\lambda_{m,k}$

$$f_{v,m,k}\left(\mathbf{p}_v; \boldsymbol{\mu}_{m,k}, \lambda_{m,k}\right) = \exp\left\{-\frac{\|\mathbf{p}_v - \boldsymbol{\mu}_{m,k}\|_2^2}{\lambda_{m,k}}\right\}, \tag{6}$$

in positions $\mathbf{p}_v \in \mathbb{R}^3$ for all the voxels in the three-dimensional voxel space of the brain. HTFA defines the local factors in $\mathbf{F}_m$ as perturbations of the factors of a global template in $\mathbf{F}$. Therefore, the factor centers $\boldsymbol{\mu}_{m,k}$ for all subjects are obtained from a multivariate normal distribution with mean μ_k and covariance $\boldsymbol{\Sigma}_{\mu_m}$. The mean μ_k represents the center of the global kth factor, while $\boldsymbol{\Sigma}_{\mu_m}$ determines the distribution of the possible distance between the global and the local center of the factor. Similarly, the widths $\lambda_{m,k}$ for all subjects are drawn from a normal distribution with mean λ_k, the width of the global kth factor, and variance $\sigma^2_{\lambda_m}$. The model defines multivariate Gaussian prior $\mathcal{N}\left(\mu_{\mu_k}, \boldsymbol{\Sigma}_{\mu_k}\right)$ for the global parameters μ_k and Gaussian prior $\mathcal{N}\left(\mu_{\lambda_k}, \sigma^2_{\lambda_k}\right)$ for λ_k, respectively. In addition, the columns of the weight matrices $\mathbf{W}_m$ are modeled with a $\mathcal{N}\left(\mu_{w_m}, \sigma^2_{w_m}\right)$ distribution and the elements in the noise term $\mathbf{E}_m$ are assumed to be independent with a $\mathcal{N}\left(0, \gamma_m^2\right)$ distribution (for one approach to non-independent noise, see 2.5). The associated graphical model is shown in Fig. 3.

- *Solving the model*

The *maximum a posteriori* (MAP) probability estimation procedure is used to solve the HTFA model. The method consists of a global and local step that iteratively update the parameters (Anderson et al., 2016). The global step updates the parameters of the K distributions in the global template. The local step updates for each subject m the weight matrices $\mathbf{W}_m$, the local centers $\boldsymbol{\mu}_{m,k}$ and the widths $\lambda_{m,k}$ of each latent factor. To update the parameters of the factors in $\mathbf{F}_m$, the local step solves the following problem, where ϕ_m is a subsampling coefficient. Optimized implementations of TFA and HTFA (Anderson et al., 2016) can be found in BrainIAK (Kumar et al., 2019).

$$\begin{aligned}\left\{\hat{\boldsymbol{\mu}}_{m,k}, \hat{\lambda}_{m,k}\right\}_k = \underset{\left\{\boldsymbol{\mu}_{m,k}, \lambda_{m,k}\right\}_k}{\operatorname{argmin}} \, & [\frac{1}{2\gamma_m^2}\|\mathbf{X}_m - \mathbf{W}_m\mathbf{F}_m\|_F^2 \\ & + \frac{1}{2\phi_m}\sum_{k=1}^{K}\left(\boldsymbol{\mu}_{m,k} - \hat{\boldsymbol{\mu}}_k\right)\boldsymbol{\Sigma}_{\mu_m}^{-1}\left(\boldsymbol{\mu}_{m,k} - \hat{\boldsymbol{\mu}}_k\right)^T \\ & + \frac{1}{2\phi_m\sigma^2_{\lambda_m}}\sum_{k=1}^{K}\left(\lambda_{m,k} - \hat{\lambda}_k\right)^2]\end{aligned} \tag{7}$$

Eq. (7) consists of the reconstruction error, the Mahalanobis distance between global and local centers, and the Euclidean distance between global and local widths. Due to its non-linearity, the latent factors of each subject are computed using a non-linear least squares solver (Kumar et al., 2019), and implemented with a trust-region reflective method (Coleman and Li., 1996). The weight matrix is solved with a closed-form solution of the form of ridge regression. The hyperparameters of the global template are updated given the local estimates and under the assumption that the posterior has a conjugate prior with multivariate normal and normal distribution for centers and width, respectively.

- *Advantages*

Because the number of network nodes is typically substantially smaller than the number of fMRI voxels, one obvious advantage of HTFA is that it can be orders of magnitude more efficient than traditional voxel-based functional connectivity approaches. Compared to other dimensionality reduction methods, HTFA provides additional advantages: (a) it provides estimation of both global and subject-specific templates, and builds connections between them; (b) modeling the latent factors as spatially smooth allows them to be overlapping rather than distinct, as would be the case of functional connectivity based on anatomically defined brain region segmentation; (c) it provides a natural means of determining how many network nodes (latent factors) should be used for a given dataset (further details about determining K can be acquired from Manning et al. (2018); and (d) because HTFA decomposes brain images into sums of spatial functions, it supports seamless mapping between images of different resolutions and potentially different imaging modalities.

- *Applications*

HTFA can be applied to different tasks with multi-subject fMRI datasets, for example, inferring dynamic full-brain inter-subject functional connectivity when participants are listening to a story or watching a television show (Manning et al., 2018). The functional connectivity of each subject can be estimated by the correlation between the column of $\mathbf{W}_m$. Since the global template of HTFA makes sure the columns of the $\mathbf{W}_{1\ldots M}$ correspond to the same network nodes across the different subjects, the ISFC can be computed by the correlation between the columns of $\mathbf{W}_{1\ldots M}$ across subjects. A recent study showed both HTFA-derived activities and HTFA-derived ISFC can be used to reliably decode which moments in the story or show the participants were experiencing. A decoder with the combination of these two types of patterns outperformed decoders with either activity or connectivity patterns alone (Manning et al., 2018).

2.3. Obtaining biologically informed decoding weights on fMRI patterns

- *Defining the problem: fMRI decoding with sparse weights*

A primary research problem neuroscientists have been studying with fMRI is brain decoding or inverse inference (O'Toole et al., 2007; LaConte et al., 2005; Cox and Savoy, 2003). The goal of a decoding task is to understand how brain activity represents task-related variables, e.g. the orientation of a grating (Haynes and Rees, 2005) or the category of an object (Haxby et al., 2001). Researchers often use linear classification and regression methods to identify the brain regions or voxels that are most closely related to these task-related variables by inspecting the decoding weights.

A piece of domain knowledge in fMRI decoding is that different regions of the brain are specialized for different functions, implying that only few small regions of the brain are specifically activated during an individual task. In the linear regression methods that are common in the field, this assumption is equivalent to assuming that the weights mapping fMRI to task-related variables are mostly zeros with a few non-zero values, which is referred to as "sparsity". This model assumption is also reasonable from a statistical standpoint, since the task variable is linked to fMRI data with usually tens of thousands of voxels, but the number of fMRI volumes with valid task labels is far smaller, e.g. a few hundred. We need to estimate tens of thousands of coefficients to map a full brain pattern down to a single task variable given only a few hundred observations. This is referred to as a high-dimensional and small-sample issue, where the linear regression model would fit seemingly predictive information from noise instead of the underlying brain signal, and thus would not generalize well to new data. To address this issue, one can reduce the number of coefficients. With the sparsity assumption, we effectively regularize the linear decoding model by restricting the weight parameter space to a much smaller one, thus mitigating the issue.

- *Making assumptions: region sparsity*

Sparse decoding has already been exploited in the previous literature (Carroll et al., 2009; Michel et al., 2011; Grosenick et al., 2013). However, the non-zero coefficients are not randomly distributed throughout the brain, but tend to arise in clusters, and are therefore not independent a priori. Sets of voxels allowing to discriminate between different brain states are expected to form small localized and connected areas. If one voxel encodes information related to the task, its neighboring voxels should carry similar information, given that contiguous brain regions of shared functions extend over multiple adjacent voxels. This type of sparsity is referred to as "region sparsity" (Wu et al., 2019). By considering such region sparsity, one can impose a structured sparsity regularization over the decoding weights which further constrains the parameter space to search and thus eases the decoding weights optimization task. Wu et al. (2019) developed a Bayesian framework that incorporated such region sparsity into brain decoding for fMRI analysis and showed superior decoding performance and more biologically informed decoding weights for three brain imaging datasets.

- *Translating assumptions to a graphical model: building a region sparsity prior over the brain weights* (The model proposed in Wu et al., 2019) is referred to as "Dependent Relevance Determination" (DRD). It builds a Bayesian hierarchical model that imposes a sparsity prior over the decoding weights. Unlike previous work with sparsity assumptions, DRD also assumes that nearby sparse voxel-activations should be correlated to each other based on their spatial locations.

Formally, the fMRI decoding problem can be formulated in a linear regression setting: at time t, consider a scalar response $y_t \in \mathbb{R}$ linked to an input vector $\mathbf{x}_t \in \mathbb{R}^V$ via the linear model:

$$y_t = \mathbf{x}_t^{\mathsf{T}}\mathbf{w} + \epsilon_t, \quad \text{for} \quad t = 1, 2, \ldots, T, \tag{8}$$

with observation noise $\epsilon_t \sim \mathcal{N}(0, \sigma^2)$, where T is the number of time points and V is the number of voxels. The regression (linear weight) vector $\mathbf{w} \in \mathbb{R}^V$ is the quantity of interest. We can denote the fMRI data matrix by $\mathbf{X} \in \mathbb{R}^{T\times V}$, where each row of $\mathbf{X}$ is the tth input vector $\mathbf{x}_t{}^T$ and $T \ll V$, and the observation vector by $\mathbf{y} = [y_1, \ldots, y_T]^T \in \mathbb{R}^T$. Since the noise is Gaussian, it can be written as

$$\mathbf{y}|\mathbf{X}, \mathbf{w}, \sigma^2 \sim \mathcal{N}(\mathbf{y}|\mathbf{X}\mathbf{w}, \sigma^2\mathbf{I}). \tag{9}$$

DRD imposes a zero-mean multivariate normal prior on $\mathbf{w}$:

$$\mathbf{w}|\boldsymbol{\theta} \sim \mathcal{N}(0, C(\boldsymbol{\theta})), \tag{10}$$

where the prior covariance matrix $C(\boldsymbol{\theta})$ is a function of hyperparameters $\boldsymbol{\theta}$. One can specify $C(\boldsymbol{\theta})$ based on prior knowledge on the regression vector, e.g. sparsity (Tipping, 2001; Faul and Tipping, 2002; Wipf and Nagarajan, 2008), smoothness (Sahani and Linden, 2003; Schmolck, 2008), or both (Park and Pillow, 2011). Ridge regression assumes $C(\theta) = \theta^{-1}I$ where θ is a scalar for precision and I is the identity matrix. Automatic relevance determination (ARD) (Neal, 2012) uses a diagonal prior covariance matrix with a distinct hyperparameter θ_i for each element of the diagonal, thus $C_{ii} = \theta_i^{-1}$. DRD is an extension of ARD by imposing dependency between θ_i.

Given the general Bayesian linear regression setting, DRD aims to construct a covariance $C(\theta)$ which generates the region-sparse $\mathbf{w}$. This is achieved by introducing a latent variable $\mathbf{u} \in \mathbb{R}^V$. $\mathbf{u}$ is from a Gaussian process (GP) prior, i.e.

$$\mathbf{u} \sim \mathcal{N}(b\mathbf{1}, k). \tag{11}$$

A Gaussian process (Rasmussen, 2003) is a stochastic process whose realizations are draws from a multivariate normal distribution, but whose mean b and covariance k can be functions of another input (e.g. spatial locations). For example, by defining a Gaussian process with covariance (kernel) that is a function of spatial distances, we can constrain the samples drawn from the Gaussian process distribution to exhibit spatial correlation based on the kernel. Most commonly in GPs, the squared exponential kernel is used, which constrains the draws from the multivariate normal to be smooth over space, i.e. $k(\chi, \chi') = \rho\ \exp(-\frac{\|\chi-\chi'\|^2}{2l^2})$ where χ and χ' are the spatial location of any two voxel. Functions sampled from such a GP are smooth functions. The smoothness is determined by the length scale $l \in \mathbb{R}$ and the magnitude of the functions is determined by $\rho \in \mathbb{R}$. These three hyperparameters in the DRD prior are jointly denoted by $\boldsymbol{\theta} = \{b, \rho, l\}$.

By imposing a GP prior over the latent $\mathbf{u}$, DRD effectively captures dependencies in $\mathbf{u}$. Given such latent, Wu et al. formulate the covariance of $\mathbf{w}$ with

$$C = \mathrm{diag}[\exp(\mathbf{u})]. \tag{12}$$

The exponential function here ensures the non-negativity of values on the diagonal of C, which makes it a valid covariance. When the mean b is very negative, $\exp(\mathbf{u})$ has many close-to-zero values that result in soft-sparsity (since their prior mean is zero and the variance is nearly zero as well). Note that the spatial smoothness of $\mathbf{u}$ induces dependencies

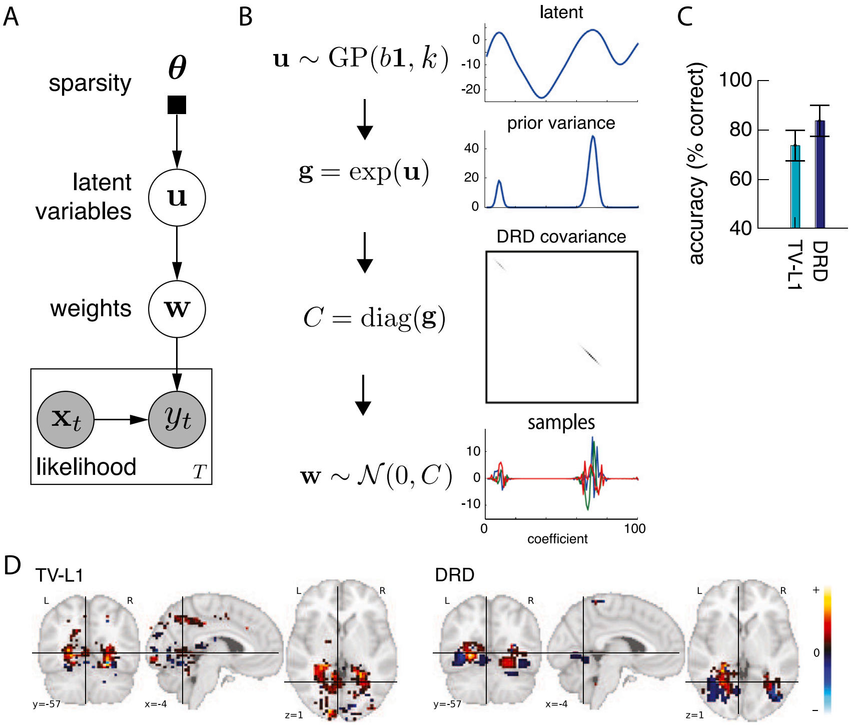

Fig. 4. (A) Probabilistic graphical model for DRD. The rectangular box indicates a graph for each time point. Each fMRI volume $\mathbf{x}_t$ at time t is mapped to the experimental response y_t together with a global variable $\mathbf{w}$ (Eq. (8)). The decoding weight vector $\mathbf{w}$ is conditioned on a latent variable $\mathbf{u}$ (Eqs. (10) and (12)). The latent variable $\mathbf{u}$ is generated from some hyperparameters in θ (Eq. (11)). (B) The generating process for region-sparse decoding weight $\mathbf{w}$. (C) Accuracy comparison between TV-L1 and DRD. The accuracy value is averaged over all pairs of objects. (D) Decoding weight map for the house vs bottle pair using TV-L1 (left) and DRD (right). Yellow indicates very positive values and light blue indicates very negative values. Black means small values. Adapted from Wu et al. (2019).

between the variances of nearby voxels, that is, the prior variance changes slowly between neighboring coefficients. If the ith coefficient of $\mathbf{u}$ has a large prior variance, then probably the coefficients of its adjacent voxels are large as well.

Fig. 4A and B show the probabilistic graphical model of DRD and the process to generate region-sparse samples for $\mathbf{w}$.

- *Solving the model*

In the paragraphs above, we show how to build a generative model for DRD to generate region-sparse decoding weights. When using DRD, one can apply it to fMRI decoding problems where we have the imaging data $\mathbf{X}$ and prediction targets $\mathbf{y}$, and we aim to infer the decoding weight vector $\mathbf{w}$. To solve this problem, we need to reverse the generating process using some inference methods. Exact Bayesian inference is infeasible with a DRD prior. However, approximate inference can be carried out efficiently using both Laplace approximation and Markov Chain Monte Carlo (MCMC) sampling. Further details regarding inference can be acquired from Wu et al. (2019).

- *Application: classification on a visual recognition task*

The visual recognition dataset (Haxby et al., 2011) is from a study on object representation in human ventral temporal cortex. In the object recognition experiment, 6 subjects were asked to recognize 8 different types of objects (bottles, houses, cats, scissors, chairs, faces, shoes and scrambled control images). Wu et al. (2019) examined this dataset to learn the weights mapping the fMRI brain activity to object categories for each subject. They cast the multi-category classification problems into multiple binary classification problems for each pair of categories. Wu et al. employed the same linear regression model as in Eq. (8) for training the model. When making predictions, they took the sign of the output y as the discrete binary labels $(+1/-1)$.

They showed that DRD achieved the highest accuracies for most of the binary classifications compared with other state-of-art sparse decoding methods (Michel et al., 2011; Grosenick et al., 2013). Fig. 4C shows a comparison of accuracies between DRD and a baseline model, total variation L1 (TV-L1) (Gramfort et al., 2013). More specifically, DRD is able to find more biologically informed decoding weight maps for many pairs compared with TV-L1. By saying biologically informed, we mean that only small regions of voxels are correlated to a specific task and nearby voxels are more likely to be activated together compared with LASSO. Fig. 4D presents the brain map estimation for the house-vs-bottle pair for TV-L1 and DRD. DRD weights have significant positive regions in the parahippocampal place area (PPA) (responding more strongly to scenes depicting places) (Epstein et al., 1999) and clustered negative weights in the lateral occipital complex (LOC) (responding to objects in human occipito-temporal cortex) (Eger et al., 2008). By comparison, TV-L1 weights in LOC are not very clustered and do not show negative activations.

We describe the DRD model here in a generative way. The brain decoding weights are generated from a DRD prior, but the application is a discriminative model, i.e. mapping fMRI data to experimental variables. Because the DRD prior was proposed to learn region-sparse brain weights regardless of whether a model is discriminative or generative, it can also be inserted to generative models such as the factor analysis models in SRM in essentially the same way.

2.4. *Inferring representational similarity between neural patterns*

- *Defining the problem: neural pattern similarity*

As sensory inputs get processed in the brain, each neural population of one brain region performs nonlinear computation of the input from neurons of other regions. The representation of the same external object thus changes from one region to another. One fundamental question in neuroscience is how these representations are transformed, in service for deciding the right actions to take (DiCarlo et al., 2012; Marr and Nishihara, 1978). One way to describe representation is in terms of what stimuli are encoded closer and what are encoded farther apart. Beyond studying representation of external stimuli, the same question

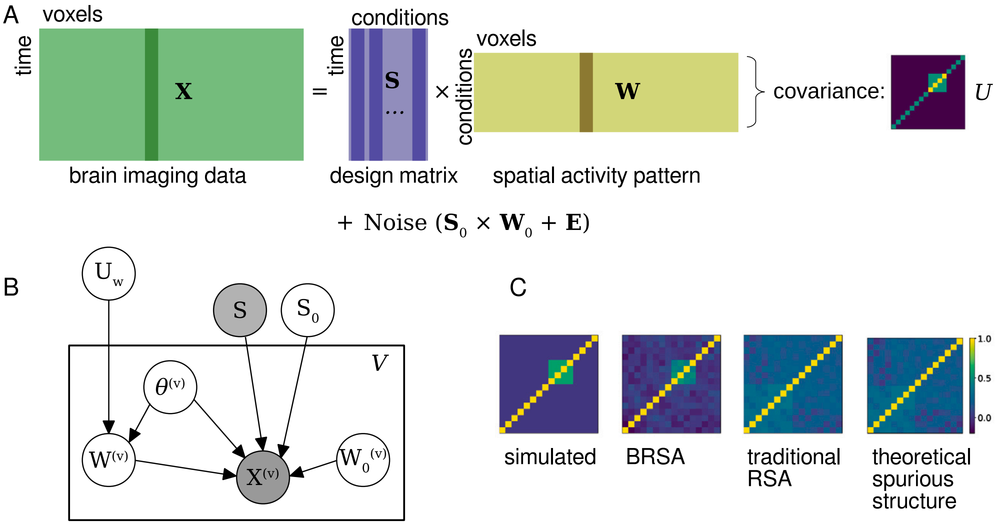

Fig. 5. (A) BRSA assumes a similar factor model as SRM and (H)TFA. To capture both spatial and temporal correlation in residual noise, the noise is further modeled by a factor decomposition of spatially correlated noise plus spatially independent noise. Additionally, each column of the weight matrix **W** (activation patterns) are assumed to share the same covariance structure, which underlies the similarity between patterns. (B) Probabilistic graphical model for BRSA. The rectangular plate is repeated for each voxel. Variables within the plate are voxel-specific and those outside the plate are shared by all voxels. U_W is the target to estimate but is indirectly related to **X** through unknown patterns **W**. To infer U_W, other unknown variables are either marginalized or (in the case of $\mathbf{S}_0$) determined through an iterative fitting procedure (see Cai et al. (2019).) (C) The simulated similarity structure, the similarity structures recovered by BRSA, by correlation of point estimates of **W** (data-mining approach) and the theoretical spurious structure expected to be introduced by the design matrix **S** when estimating $\hat{\mathbf{W}}$. (B) and (C) are adapted from Cai et al. (2019).

can also be asked about different cognitive states: which states are represented closer in a brain region?

Early behavioral studies investigated representations of objects by asking people to judge how similar a pair of stimuli are to each other (Shepard and Chipman, 1970). The structure of the similarity matrix, composed of the judged degrees of similarity between all pairs of tested stimuli, reflects the geometry of the internal representational space being used to encode stimuli. Such approach is limited to representations accessible for conscious report (Ericsson and Simon, 1980). To overcome this and to compare computational models against multiple types of neural data, Kriegeskorte et al. (2008) proposed Representational Similarity Analysis (RSA), which utilizes neural recordings to understand the structure of representations. This analysis assumes that the similarity between the neural patterns elicited by each pair of stimuli in a brain region reflects the similarity between the representations of these stimuli in that region. Because it does not rely on subjective judgment, RSA can be applied to studying representation in any stage of sensory processing (Connolly et al., 2012; Iordan et al., 2015). Measuring similarity between neural activity patterns evoked by sensory stimuli or cognitive states is its central goal.

- Making assumptions: relations of representational structure, neural patterns and fMRI data

In order to infer the similarity between neural activity patterns, one needs to first make assumptions about the relations between the neural patterns and the recorded neural data, and between the similarity structure and the patterns.

The neural activity of a region[2] during a task can be considered as being generated by the sum of various spatial patterns, each being modulated by different time courses. In this sense, the basic assumption of fMRI data underlying RSA is also a factor model, as in SRM and (H)TFA (Fig. 5A). The difference here is that at least a subset of the modulation time courses are explicitly tied to when and how much the brain is engaged in each task condition, which are pre-defined by the researchers. The spatial pattern being modulated by each time course is the relative degree by which different voxels are activated by the task condition. In addition to the activity explained by the temporal modulation of these patterns, the data also contain unexplained fluctuation with both spatial and temporal correlation. Therefore, the similarity matrix one seeks to estimate is only indirectly related to the noisy fMRI data through unknown neural activity patterns and their modulation time courses predicted by the task.

There are many ways to define similarity. One way is based on the cosine of the angle between the vectors corresponding to activity patterns in the space spanned by the voxel activation levels, which is adopted by the algorithm of Bayesian RSA (BRSA) (Cai et al., 2019, 2016). Other common ways include correlation between demeaned patterns, and Euclidean distance or Mahalanobis distance between patterns (as measures of dissimilarity) (Kriegeskorte et al., 2008; Diedrichsen et al., 2016; Ramírez, 2017; Nili et al., 2014). Here we focus on cosine of angle between patterns, which can be alternatively considered as correlation without demeaning patterns.

- Translating assumptions to a graphical model: two-stage model of fMRI data with representational structure as latent variable

Since the time course of a task is known, the modulation time course (so-called design matrix) can be constructed based on the timing of the task conditions and the shape of the smooth delayed response (the hemodynamic response function, HRF) in fMRI signals following neuronal activity. We denote the design matrix as $\mathbf{S} \in \mathbb{R}^{T\times K}$, where T is the total time points and K is the number of task conditions in an experiment. Then, the factor model of fMRI data can be expressed as

$$\mathbf{X} = \mathbf{S}\mathbf{W} + \mathbf{S}_0\mathbf{W}_0 + \mathbf{E}. \tag{13}$$

Here, $\mathbf{X} \in \mathbb{R}^{T\times V}$ is the time by voxel matrix of the fMRI time series in a region of interest, where V is the total number of voxels in that region. $\mathbf{W} \in \mathbb{R}^{K\times V}$ is the unknown activation patterns associated with all the task conditions. $\mathbf{S}_0\mathbf{W}_0$ captures spatially correlated fluctuation unrelated to the task. $\mathbf{E}$ denotes the residual spatially independent noise, but it can have temporal autocorrelation, which may be modeled with an auto-regressive (AR) process such as AR(1) (for an alternate approach to the residual noise in RSA, see 2.5). Generally, researchers do not have full knowledge of $\mathbf{S}_0$ or $\mathbf{W}_0$, but may have regressors (such as the head motion time course) which accounts for some variance in $\mathbf{S}_0$. Assuming that $\mathbf{E}$ is random variable drawn from the noise distribution, Eq. (13) implicitly defines the conditional probability of

[2] RSA typically focuses on single brain region instead of the whole brain.

the data in each voxel given $\mathbf{S}$, $\mathbf{W}$, $\mathbf{S}_0$, $\mathbf{W}_0$ and the parameters θ of the AR process, i.e. $p(X^{(v)}|\mathbf{S}, W^{(v)}, \mathbf{S}_0, W_0^{(v)}, \theta^{(v)})$ for voxel v.

When cosine angle $\alpha_{i,j}$ is used as a measure of similarity between patterns w_i and w_j (row vectors of $\mathbf{W}$), $\cos\alpha_{i,j} = \frac{w_i w_j^T}{\sqrt{w_i w_i^T}\sqrt{w_j w_j^T}}$. If the activation profile of each voxel $w^{(v)}$ is a sample from a multivariate distribution, then $\mathbb{E}[w_i w_j^T]$ is the covariance between the dimensions i and j of this distribution (Diedrichsen and Kriegeskorte, 2017). By estimating the covariance structure U_W of $\mathbf{W}$, one can obtain the cosine angle between patterns as a similarity measure. Therefore, the relation between unknown neural patterns and their similarity is modeled by assuming that each column of $\mathbf{W}$ is a sample drawn from a multivariate distribution with its covariance matrix being U_W:

$$w^{(v)} \sim N(0, U_W) \tag{14}$$

This specifies the form of conditional probability of $w^{(v)}$ given U_W: $p(w^{(v)}|U_W)$. The two-stage generative model from covariance structure through activity patterns to fMRI data is depicted in Fig. 5B.

- Solving the model: inferring covariance structure of unknown neural patterns directly from data

After the probabilistic graphical model is built and the conditional probability distribution corresponding to each edge in the graphical model is specified, one can derive the likelihood $p(\mathbf{X}|U_W)$. This can be achieved by marginalizing the intermediate variables such as $\mathbf{W}$ and other unknown quantities that $\mathbf{X}$'s distribution is conditioned on ($\mathbf{S}_0$ is determined through an iterative fitting procedure as in Cai et al. (2019)). Marginalization in probability refers to removing a variable in the expression of probability density by integrating the joint or conditional distribution over the variable, an important procedure in applying Bayesian models that allows the analyst to remain agnostic about the value of 'nuisance' variables unimportant to the main analysis. For example, for any two variables A and B, $p(A) = \int P(A, B)dB = \int P(A|B)p(B)dB$. For any three variables A, B and C, $p(A|C) = \int p(A|B, C)p(B|C)dB$. In our case, A, B and C can be replaced by $\mathbf{X}$, $\mathbf{W}$ and U_W: we are agnostic about the specific mapping of the design matrix to the measurements, we are only interested in its implied covariance. In practice, the integration over several unknown variables have closed-form solution due to the assumption of Gaussian distributions in both $p(x^{(v)}|\mathbf{S}, w^{(v)}, \mathbf{S}_0, w_0^{(v)}, \theta^{(v)})$ and $p(w^{(v)}|U_W)$, which makes the computation simple. Other unknown variables can be marginalized by numerical approximation. After obtaining the formula of the marginal likelihood $p(\mathbf{X}|U_W)$, maximizing its logarithm with respect to U_W yields the maximum likelihood estimation $\widehat{U_W}$ of U_W. Finally, the cosine angles between $\mathbf{W}$ can be obtained as the correlation matrix corresponding to the covariance matrix $\widehat{U_W}$.

- Application: reducing spurious similarity structure

Maximizing the likelihood $p(\mathbf{X}|U_W)$ while marginalizing unknown intermediate variables and uninteresting variables is a principled approach to infer the latent variable U_W based on PGM. An alternative non-PGM approach is to instead first calculate $\hat{\mathbf{W}}$ as estimates of the unknown patterns $\mathbf{W}$ from the data by regressing $\mathbf{X}$ against $\mathbf{S}$, and then calculate the similarity among rows of $\hat{\mathbf{W}}$. This approach, however, has been shown (Alink et al., 2015; Cai et al., 2016, 2019; Henriksson et al., 2015) to introduce spurious similarity structure unrelated to the neural activity corresponding to the task of interest. The reason is that although the regression provides unbiased estimates $\hat{\mathbf{W}}$ of the neural patterns, the covariance of $\hat{\mathbf{W}}$ is not the same as the covariance of $\mathbf{W}$: $\hat{\mathbf{W}}$ is contaminated by noise with specific covariance structure introduced by the regression procedure. The noise itself originates from the task-unrelated fluctuation in fMRI data. The regression procedure, at the same time of disentangling $\mathbf{W}$ from $\mathbf{X}$, also "entangles" the noise into each row of $\hat{\mathbf{W}}$ in a way that depends on the correlational structure between different columns of $\mathbf{S}$. The covariance structure of the noise in $\hat{\mathbf{W}}$ can dominate the estimated similarity structure when signal-to-noise ratio is low (Cai et al., 2016, 2019) (Fig. 5C). BRSA takes into account both the property of noise and uncertainty of intermediate variables $\mathbf{W}$, thus avoiding analyzing $\hat{\mathbf{W}}$ with structured noise.

Instead of directly inferring U_W from $\mathbf{X}$, one can alternatively assume that U_W is composed of the sum of a few theoretically-motivated candidate covariance structures, and estimate the mixture coefficient of each component covariance structure. This method is called Pattern Component Modeling (PCM) (Diedrichsen et al., 2011, 2018). One can even impose a hyperprior on the mixing coefficients, and use variational Bayesian technique to infer them (Friston et al., 2019). The introduction of a hyperprior can incorporate additional prior assumptions or knowledge of the data. Although not directly aimed at reducing statistical bias, these methods are both developed based on clear PGMs. It is worth pointing out that in using these methods, in order to overcome the spurious similarity structure introduced by the design matrix $\mathbf{S}$, one still needs to either directly model the data $\mathbf{X}$ as in BRSA, or to model $\hat{\mathbf{W}}$ while explicitly taking into account the structure of the non-independent noise it carries.

Even if one takes an approach without explicitly relying on a PGM, a correct understanding of the confounding effect of noise by analyzing a PGM is helpful for developing a better non-Bayesian algorithm. For example, one can still approximate the covariance or distance structure based on the noisy patterns estimated from separate runs of experiment (Alink et al., 2015; Diedrichsen et al., 2016). This is because the noises in the patterns estimated separately from different fMRI runs come from independent sources and have zero correlation. Therefore, the covariance between estimated patterns from separate runs is an unbiased estimation of the covariance of the true unknown patterns. However, it is worth pointing out that the correlation derived from such cross-run covariance is still biased, because to calculate correlation we have to divide the covariance by the estimated standard deviations of each pattern across voxels, which is inflated by the existence of noise. To see these, one needs to understand how data $\mathbf{X}$ is generated from $\mathbf{W}$ (13) and how the noise in this data generating process impacts $\hat{\mathbf{W}}$. Therefore, regardless of whether a researcher directly uses an analysis tool based on PGM, analyzing the data generating process and the interactions between the analysis procedures and noise is always important for realizing and avoiding any unintended consequence introduced by the chosen analysis methods.

2.5. Modeling structured residuals

- Defining the problem: modeling spatiotemporal residuals in fMRI data

fMRI data has structure in both the spatial and temporal dimension, and this spatiotemporal consistency needs to be exploited (or at least, managed) in order to contend with this high-dimensional and noisy data. This spatiotemporal structure exists both in the neural components corresponding to the effects of interest, and in the *residual* components corresponding to everything else going on. In the context of supervised regression models for fMRI, practitioners tend to worry about temporal structure in both signal (by convolving the predictors with a synthetic hemodynamic response function) and residual (by performing generalized least squares, or GLS, estimation wherein the temporal structure of the residuals is modeled, (e.g. Mumford and Nichols, 2006). More recent factor-analytic unsupervised approaches likewise assume the signal of interest itself is spatially or temporally structured due to their low-rank structure, for example the case of TFA (above) modeling brain networks as a linear combination (in time) of spatially contiguous factors. Most of these methods leave handling the residual to preprocessing stages, but this is not the only possible choice — another is to model the spatial or temporal structure in the residuals explicitly. One example of this is modeling the residual temporal autocorrelation per-voxel, as in the case of BRSA. Another, and the one we discuss here, is modeling spatiotemporable separable residuals.

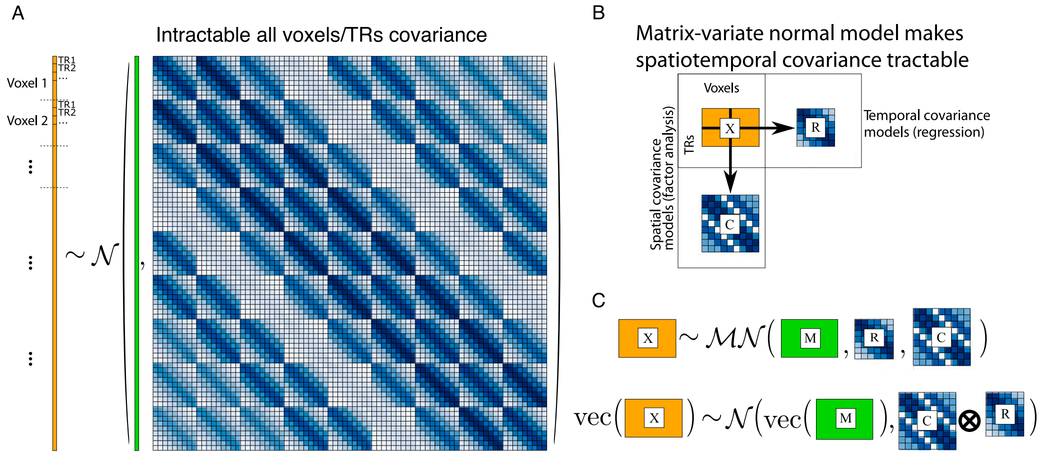
A
Intractable all voxels/TRs covariance
Voxel 1
TR1
TR2
...
Voxel 2
TR1
TR2
...
~ 𝒩
B
Matrix-variate normal model makes spatiotemporal covariance tractable
Voxels
TRs
X
R
Temporal covariance models (regression)
Spatial covariance models (factor analysis)
C
X ~ 𝓜𝒩(M , R , C)
vec(X) ~ 𝒩(vec(M), C ⊗ R)

Fig. 6. Matrix normal models simultaneously model spatial and temporal residuals. [A]: a schematic view of a vectorized data matrix, where each voxel's time series is vertically concatenated (in orange, on the left), and the covariance of every voxel at every timepoint with every other voxel at every other timepoint is modeled. Modeling all of these elements independently is intractable, and some structure needs to be imposed — in this case, kronecker-separable structure. [B]: the un-vectorized data matrix (X; orange rectangle), and its spatial and temporal covariances on the right and bottom. [C]: A matrix-normal distribution with the mean M and row/column covariances R, C is equivalent to the large structure in [A], but can be much more tractable to estimate. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

- *Making assumptions: structured, separable residuals*

As noted above, both the fMRI signal and residual are autocorrelated in both space and time; thus, modeling the residual structure in both dimensions is needed. This is not tractable in the general case, as it effectively means modeling the covariance between every voxel at every timepoint with every other voxel at every other timepoint. A simplifying assumption that permits modeling residuals in both space and time is that the spatial residuals of all time points have the same distribution, and the temporal residuals of all voxels likewise have the same distribution (for an illustration, see Fig. 6). This *separable* residuals assumption has been made in a GLS framework by Katanoda et al. (2002) and factor-analytic framework by Shvartsman et al. (2018). A similar approach has been taken to modeling the entire dataset (rather than residuals only) in both neuroimaging (e.g. Bijma et al., 2005; Roś et al., 2014) and elsewhere in the *multitask learning* community (e.g Bonilla et al., 2008; Skolidis and Sanguinetti, 2011; Stegle et al., 2011; Rakitsch et al., 2013; Greenewald and Hero, 2015). Once separability is assumed, theoretically motivated structure could be placed on the individual spatial and temporal residual covariances, for example autoregressive in time (as in BRSA, above) and smooth in space (as in DRD, above).

- *Translating assumptions to a graphical model: matrix-normal*

The informal claim of separability above is denoted by defining Σ_{all} to be equal to the kronecker product of a spatial and temporal residual covariance, $\Sigma_{all} := \Sigma_t \otimes \Sigma_v$. The kronecker product is a generalization of the vector outer product to matrices, and precisely performs the weighted tiling illustrated in Fig. 6. Using this notation, we define the matrix-variate normal distribution, a distribution over matrices parameterized by a mean matrix and (separable) row and column covariances. We denote matrices drawn from this distribution as $\mathbf{X} \sim \mathcal{MN}_{m,n}(\mathbf{M}, \mathbf{R}, \mathbf{C})$, with mean $\mathbf{M} \in \mathbb{R}^{m\times n}$, row covariance $\mathbf{R} \in \mathbb{R}^{m\times m}$ and column covariance $\mathbf{C} \in \mathbb{R}^{n\times n}$. It has the following log-likelihood:

$$\log p(\mathbf{X} \mid \mathbf{M}, \mathbf{R}, \mathbf{C}) = -2\log mn - m\log|\mathbf{C}| \\ - n\log|\mathbf{R}| - \mathrm{Tr}\left[\mathbf{C}^{-1}(\mathbf{X}-\mathbf{M})^T\mathbf{R}^{-1}(\mathbf{X}-\mathbf{M}M)\right]. \tag{15}$$

The above notation is equivalent to denoting $\mathrm{vec}(\mathbf{X}) \sim \mathcal{N}(\mathrm{vec}(\mathbf{M}), \mathbf{C}\otimes\mathbf{R})$, where $\otimes$ is the kronecker product and vec is the vectorization operator. If the column covariance $\mathbf{C}$ is the identity matrix (i.e. the columns are independent), the expression reduces to the log-likelihood of the multivariate normal distribution summed over the columns. We can use this notation to write, for example, a separable-residual model SRM model:

$$\mathbf{S} \sim \mathcal{MN}(0, \boldsymbol{\Sigma}_\mathbf{s}, \mathbf{I}) \tag{16}$$

$$\mathbf{X}^T{}_m \mid \mathbf{S} \sim \mathcal{MN}(\mathbf{W_m S} + \mathbf{M}_m, \boldsymbol{\Sigma}_\mathbf{v}, \boldsymbol{\Sigma}_\mathbf{t}), \tag{17}$$

where $\boldsymbol{\Sigma}_\mathbf{v}$ and $\boldsymbol{\Sigma}_\mathbf{t}$ are spatial and temporal residual covariances and the remaining parameters are as defined above. In contrasting the diagram in Fig. 7 one can see the disappearance of the plate iterating over timepoints, since now temporal residuals are modeled. In this view, we can also see a similar model in which the prior on $\mathbf{W}_m$ is modeled instead:

$$\mathbf{W_m} \sim \mathcal{MN}(0, \mathbf{I}, \boldsymbol{\Sigma}_\mathbf{w}) \tag{18}$$

$$\mathbf{X}^T{}_m \mid \mathbf{W_m} \sim \mathcal{MN}(\mathbf{W_m S} + \mathbf{M}_m, \boldsymbol{\Sigma}_\mathbf{v}, \boldsymbol{\Sigma}_\mathbf{t}). \tag{19}$$

In this view, which Shvartsman et al. labeled dual probabilistic SRM (DP-SRM) by analogy to dual probabilistic PCA (Lawrence, 2005), $\mathbf{W}_m$ can no longer be modeled as orthonormal but can now be integrated over with a Gaussian prior, estimating substantially fewer parameters. Similar modeling of residual covariance can be performed on other factor models (Shvartsman et al., 2018), including all of the generative models in this paper, or a generative variant of the structured sparsity (DRD) model, and others such as ISFC (Simony et al., 2016). It is not obviously applicable to discriminative models, whose residuals are in the space of predictors and not the space of voxels.

- *Solving the model*

While simply estimating all parameters by gradient descent is theoretically possible, a more practical approach is to marginalize over nuisance parameters, and estimate only the parameters of interest. Marginalization in the multivariate normal setting with Gaussian priors is well-known (Bishop, 2006), but the separable covariance formulation introduces some new inference challenges: marginalization yields a non-separable marginal likelihood, naive computation of which would require inverting a matrix of dimension $vt \times vt$ for v voxels and t time points, which is intractable for fMRI data. However, Rakitsch et al. (2013) provided an efficient method for computing this likelihood by exploiting the compatibility between diagonalization and the kronecker product. If the spatial residual matrix itself needs to be separable (e.g. for efficiently modeling whole-brain spatial residuals by separating them in the x, y, and z dimensions), Shvartsman et al. (2018) show that particular assumptions about prior covariances can likewise render the marginal separable (and thus tractable). Once the marginal likelihood

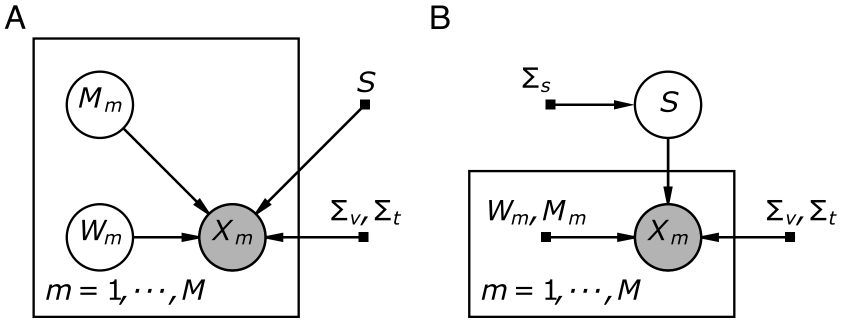

Fig. 7. Plate diagrams for matrix-normal shared response model. In the matrix-normal notation one can see that there are two possible formulations for an SRM-type model: one which integrates over the shared timecourse (as SRM does), and one which integrates over the subject-specific weightings while removing the orthonormality assumption on $\mathbf{W_m}$ (this is termed 'dual probabilistic SRM' or DP-SRM by analogy to dual probabilistic PCA, which makes the same extension to PCA (Lawrence, 2005). In both cases, the brain image data $X_m \in \mathbb{R}^{T \times V}$ is observed from subject m, $m = 1 : M$. As in conventional SRM, each observation (now represented as the full data matrix) is a linear combination of subject-specific latent factors. In regular MN-SRM (A, left), the time-course $\mathbf{S}$ is treated as a latent variable that is integrated over and the mean $\mathbf{M}_m$ and weight vector $\mathbf{W}_m$ are treated as (hyper)parameters that need to be estimated. In DP-SRM (B, right), the weight vector and mean matrix are treated as latent variables and integrated over whereas the shared timecourse is treated as a (hyper)parameter to estimate. Note how in contrast to Fig. 2, there is no plate denoting independence between timepoints, since their covariance is now modeled. Shaded nodes: observations, unshaded nodes: latent variables, and black squares: parameters.

can be computed efficiently, standard gradient-based techniques can be used for estimation. For even greater speed, Shvartsman et al. (2018) derived an expectation-conditional-maximization algorithm for maximizing the marginal likelihood by coordinate ascent (though they only did so for matrix-normal SRM; matrix-normal RSA was estimated by gradient ascent).

- *Applications and benefits*

Similarly to the other models in the paper, we are not advocating any specific spatiotemporal covariance model here, nor its specific application to any specific method. Rather, we highlight the explicit modeling approach and its ability to incorporate a class of structure assumptions into other models, as long as they are linear Gaussian regression or factor models (which includes many models in the literature). That said, specific empirical benefits of introducing a separable residual covariance to other models have been realized. In the case of the GLM for fMRI, Katanoda et al. (2002) validated the separable-residuals model on synthetic data, as well as on a finger-tapping experiment. There, they demonstrated that the separable model recovers larger activations more closely focused around the expected motor regions. Additionally, the separable model provided a higher goodness of fit to experimental data than models that included temporal residual structure only, or no residual structure at all. In the case of factor models, Shvartsman et al. (2018) show that the separable model can be substantially faster to estimate than a model that includes voxel-specific temporal residuals (as in the case of BRSA vs MN-RSA) and can achieve lower error while retaining BRSA's conservative behavior under the null. A separable variant of SRM achieves lower out-of-sample reconstruction error for new subjects than conventional SRM, though this reduced error does not seem to translate to improved feature extraction for brain decoding. The matrix-normal modeling toolkit (under review for inclusion in the BrainIAK analysis package (Kumar et al., 2019)) makes it possible to prototype inclusion of separable covariances into other models.

3. Discussion

In this paper, we use five computational tools developed for different goals in fMRI research to illustrate how to build probabilistic graphical models (PGMs) to address important questions arising in neuroimaging studies. These methods also illustrate how the PGM (Koller and Friedman, 2009), which is central to the methods reviewed, can accommodate domain knowledge and known properties of the data and facilitate aggregating information over larger datasets. These features allow us to mitigate the limitations in fMRI data: high dimensionality (many voxels), low sample size in single subject, heterogeneity across subjects and complex noise with high magnitude. The PGM-based approach helps ensure the faithfulness of an algorithm to its original purpose and provides flexibility in model building.

To tackle the limits of high dimensionality and low sample size, SRM (Chen et al., 2015) uses the existence of a shared latent response as its core assumption, which allows aggregating data from multiple subjects; HTFA (Manning et al., 2018) uses a hierarchical model across subjects to discover common nodes in many brains. By utilizing big data across many subjects, both methods essentially increase the sample size to discover common structure in the data. In addition, the low-rank factor model underlying both methods reduces the model complexity, thus mitigating overfitting.

In aggregating data, both SRM and (H)TFA tolerate the heterogeneity of data across subjects, but in slightly different ways: SRM assumes different spatial weight matrices across subjects while HTFA allows the spatial location of the same node in different subjects to vary. Similarly, an extension of BRSA, the Group BRSA (Cai et al., 2019) allows spatial patterns to differ across subjects while assuming the same similarity matrix is shared by subjects.

An alternative way to mitigate high dimensionality and low sample size is to introduce domain knowledge which trades off between bias and variance in parameter estimation. The three-dimensional Gaussian kernel in (H)TFA (Manning et al., 2014, 2018) can be considered as adopting the belief that fMRI activations are smooth and local. DRD (Wu et al., 2019) introduces similar domain knowledge (region sparsity) to tackle the problem by using a Gaussian Process prior on the logarithm of decoding weight variance. This prior allows the weights to have more flexible spatial patterns than Gaussian blobs. Although not reviewed in this article, the method of estimating population receptive field (Dumoulin and Wandell, 2008) and more generally, the encoding model approach (Naselaris et al., 2011) essentially also bring in domain knowledge of neural tuning properties in modeling fMRI data.

Aggregating more data and introducing domain knowledge both essentially reduce the impact of high noise in fMRI data. BRSA and kronecker-separable factor model variants (Shvartsman et al., 2018) go one step further by explicitly modeling the spatial and temporal correlation structure in noise. BRSA separates the spatially correlated and independent noise components and models the former with a factor model, allowing for a more complex correlation structure. The matrix-normal formalism assumes separability of the residual covariance structure into one corresponding to spatial covariance and one corresponding to temporal covariance, largely reducing the number of free parameters while still being able to capture the major structure in noise. Explicitly modeling the noise structure helps reduce bias in estimation arising from the mismatch between an overly simplified noise assumption and the complex property of noise in the data.

In addition to tackling the limitation in fMRI to increase the power for discovering meaningful information in the data, one advantage of the PGM-based approach is its faithfulness to the original goal of a research. This is illustrated in the case of BRSA, where an approach without explicitly examining the data generating and analyzing process may overlook the difference between the output of an early-stage analysis procedure and the true quantity of data that the procedure attempts to estimate, and may introduce spurious results. PGMs allow for simulation of data according to the model and verification of the inference algorithm. This is an advantage not easily achieved by analysis procedures developed without an explicit model. In fact, during sequential applications of analysis or filtering steps, later steps may reintroduce artifacts intended to be removed by early steps (Lindquist et al., 2019), and variation in complex pipelines may vary the results (Carp, 2012). In functional connectivity analysis, various denoising procedures can introduce spurious brain network correlational structures (Chen et al., 2017; Murphy and Fox, 2017; Murphy et al., 2009; Saad et al., 2012; Leonardi and Van De Ville, 2015). These are often due to the interaction between the preprocessing procedures and later-stage analysis, which is hard to foresee without building and analyzing explicit generative models.

The PGM-based approach to neuroimaging analysis also offers the flexibility of combining advantages of different models and tailoring models for new application domains. This has been illustrated by the extensions of SRM to several variants that utilize partial labels of data (Turek et al., 2017) or datasets with partially overlapping subjects (Zhang et al., 2018). Likewise, it is illustrated in the development of separable-covariance variants of existing models (Shvartsman et al., 2018). It is an interesting future research direction to develop new tools that combine the advantage of the existing PGM-based methods, including the models reviewed here. Understanding the commonality among models is the first step towards integrating them. This is the reason we intentionally use the same notation and matrix orientation of the data matrices in this paper to help readers see the commonality among these methods. Furthermore, several of the tools in this article are available in the same open source package Brain Imaging Analysis Kit (BrainIAK) (Kumar et al., 2019), which makes it easier for tool developers to understand how the computational models and inference algorithms ultimately turn into functioning code and to draw inspiration from these tools (examples for the usage of most algorithms can be found at https://github.com/brainiak/brainiak/tree/master/examples)

Although PGM-based methods come with the aforementioned advantages, they are not without limitations. The first limitation is the speed of computation. Because such models need to consider uncertainty of unknown variables, marginalization of unknown variables is involved, which often requires inverting relatively large matrices, a time- and memory-consuming computation. However, with the advance of parallel computing techniques and code optimization (Anderson et al., 2016; Gardner et al., 2018), these limitations are gradually being resolved. Second, although the integrative approach of PGM-based methods reduces the chance of obtaining spurious outcomes due to interaction between different stages of data processing, it does reduce the flexibility provided by traditional approach which concatenates many modular analysis tools as a pipeline (Esteban et al., 2019). Traditional pipeline approaches allow for fast reanalysis when more data (e.g., a new subject) are added, while some of the PGM-based methods may need to redo the analysis on all the data in such situation, or at least require deriving new model update equations. The third limitation is the potential sensitivity of these methods to the correctness of the prior assumptions and generative models used in such methods. For example, noise are often assumed to follow variants of multivariate Gaussian distributions for the easiness of inverting models. But more investigations are needed on the impact of such assumptions when the data distributions in fact violate the assumptions, for example, by having heavier tails than Gaussian distributions. If a model is not a good description of the true generative process of data, then the analysis result may not reflect the ground truth underlying the data. Some assumptions may still be overly simplified compared to the complex nature of true fMRI noise (here, noise may include intrinsic neural activity). This is a challenge but also an opportunity of the PGM-based approach, because making all assumptions explicit makes it easier to examine the impact of the assumptions being made. One may ask how to check whether an assumption of the noise property is correct. Although it is hard to know the true model for data, PGMs at least offer the ability to calculate the likelihood of data given any assumption of noise property. By comparing which assumptions about noise gives higher likelihood for the data, one can decide what assumptions are most appropriate for the data acquired. Finally, although PGM-based approaches aim to consider as much of the noise property as possible, they are still not end-to-end, in the sense that various preprocessing procedures, such as slice-timing correction, motion correction, and spatial distortion correction, are still performed separately prior to employing these methods. The generative process of motion-induced artifacts is typically not modeled in these PGM-based methods. Building a PGM which directly models the raw data straight out of fMRI scanners is still too complex a process. One needs to find a balance between the advantage from the explicit and probabilistic nature of PGM and the complexity introduced by modeling every detail of the data.

Beyond the properties of fMRI data that have been discussed in this paper, there are many more complexities in the properties of fMRI data. These complexities may all influence the conclusions one can draw from the data, depending on how much they are taken into account. For example, the temporal profile of hemodynamic response can differ not only across regions, but also across brains. PGM-based methods in fact played an important role in modeling and quantifying this variation (Woolrich et al., 2004). Future work that incorporates this variation in methods such as BRSA may improve the power of the algorithm and help users evaluate how sensitive their analyses are to such variations. Behavioral contingencies such as reaction time can also influence the fMRI response. Currently their impact is usually modeled deterministically when building design matrices in traditional regression analysis and in BRSA (Cai et al., 2019). Future work may incorporate the probabilistic impact of such behavioral contingency on fMRI responses, or treat these behavioral measures as additional observations that can be probabilistically predicted by a PGM. The signal-to-noise ratio of data can be influenced by the speed of fMRI acquisition, together with other factors. So there is a trade-off between data quality and the spatial and temporal resolution of the data. Therefore, there is a need for developing new PGM which simultaneously incorporates multiple types of noise and artifacts, and estimate their relative magnitudes due to the choice of data acquisition protocol. Such work holds promise to improve over existing methods (Ellis et al., 2020) that estimate different noise separately for evaluating the power of an fMRI study. However, as stated above, there is a tradeoff between the amount of details a model can capture and the difficulty of making inference based on a complex model.

In the new era of big data for neuroscience (Landhuis, 2017; Kandel et al., 2013; Van Horn and Toga, 2014), facilitating data sharing is obviously one of the most important effort for making big data analysis possible (Poldrack and Gorgolewski, 2014; Choudhury et al., 2014; Gorgolewski et al., 2017, 2016). One step further, developing computational models that derive insights from big data is another key for the field of neuroscience to benefit from increasing data size, which should also be in synergy with developing theories of the essence of the neural computation (Cohen et al., 2017; Sejnowski et al., 2014; Bzdok and Yeo, 2017). We suggest that future method development places model building at the center of its focus.

CRediT authorship contribution statement

Ming Bo Cai: Conceptualization, Writing - original draft, Writing - review & editing. **Michael Shvartsman:** Conceptualization, Writing -

original draft, Writing - review & editing. **Anqi Wu:** Conceptualization, Writing - original draft, Writing - review & editing. **Hejia Zhang:** Conceptualization, Writing - original draft, Writing - review & editing. **Xia Zhu:** Conceptualization, Writing - original draft, Writing - review & editing.

Acknowledgments

MBC is supported by National Institute of Drug Abuse award R01DA042065, USA and World Premier International Research Center Initiative (WPI), MEXT, Japan.

References

Ahelegbey, D.F., 2016. The Econometrics of Bayesian Graphical Models: A Review with Financial Application. University Ca'Foscari of Venice, Dept. of Economics Research Paper Series No 13.

Alink, A., Walther, A., Krugliak, A., van den Bosch, J.J., Kriegeskorte, N., 2015. Mind the drift - improving sensitivity to fMRI pattern information by accounting for temporal pattern drift. bioRxiv 032391.

Anderson, M.J., Capota, M., Turek, J.S., Zhu, X., Willke, T.L., Wang, Y., Chen, P.-H., Manning, J.R., Ramadge, P.J., Norman, K.A., 2016. Enabling factor analysis on thousand-subject neuroimaging datasets. In: 2016 IEEE International Conference on Big Data (Big Data). IEEE, pp. 1151–1160.

Baldassano, C., Chen, J., Zadbood, A., Pillow, J.W., Hasson, U., Norman, K.A., 2017. Discovering event structure in continuous narrative perception and memory. Neuron 95 (3), 709–721.

Beckmann, C.F., DeLuca, M., Devlin, J.T., Smith, S.M., 2005. Investigations into resting-state connectivity using independent component analysis. Philos. Trans. R. Soc. B 360 (1457), 1001–1013.

Belliveau, J., Kennedy, D., McKinstry, R., Buchbinder, B., Weisskoff, R., Cohen, M., Vevea, J., Brady, T., Rosen, B., 1991. Functional mapping of the human visual cortex by magnetic resonance imaging. Science 254 (5032), 716–719.

Bijma, F., De Munck, J.C., Heethaar, R.M., 2005. The spatiotemporal MEG covariance matrix modeled as a sum of Kronecker products. NeuroImage 27 (2), 402–415.

Bingham, E., Chen, J.P., Jankowiak, M., Obermeyer, F., Pradhan, N., Karaletsos, T., Singh, R., Szerlip, P., Horsfall, P., Goodman, N.D., 2019. Pyro: Deep universal probabilistic programming. J. Mach. Learn. Res. 20, 1–6.

Bishop, C.M., 2006. Pattern Recognition and Machine Learning. Springer-Verlag New York, Inc., Secaucus, NJ, USA.

Bonilla, E.V., Chai, K.M., Williams, C., 2008. Multi-task gaussian process prediction. In: Advances in Neural Information Processing Systems. pp. 153–160.

Bright, M.G., Murphy, K., 2015. Is fMRI "noise" really noise? Resting state nuisance regressors remove variance with network structure. NeuroImage 114, 158–169.

Buxton, R.B., 2013. The physics of functional magnetic resonance imaging (fMRI). Rep. Progr. Phys. 76 (9), 096601.

Bzdok, D., Yeo, B.T., 2017. Inference in the age of big data: Future perspectives on neuroscience. NeuroImage 155, 549–564.

Cai, M.B., Schuck, N.W., Pillow, J.W., Niv, Y., 2016. A Bayesian method for reducing bias in neural representational similarity analysis. In: Advances in Neural Information Processing Systems. pp. 4951–4959.

Cai, M.B., Schuck, N.W., Pillow, J.W., Niv, Y., 2019. Representational structure or task structure? Bias in neural representational similarity analysis and a Bayesian method for reducing bias. PLoS Comput. Biol. 15 (5), e1006299.

Calhoun, V.D., Adali, T., 2012. Multisubject independent component analysis of fMRI: a decade of intrinsic networks, default mode, and neurodiagnostic discovery. IEEE Rev. Biomed. Eng. 5, 60–73.

Carp, J., 2012. On the plurality of (methodological) worlds: estimating the analytic flexibility of fMRI experiments. Front. Neurosci. 6, 149.

Carpenter, B., Gelman, A., Hoffman, M.D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., Riddell, A., 2017. Stan: a probabilistic programming language. Journal of statistical software 76 (1).

Carroll, M.K., Cecchi, G.A., Rish, I., Garg, R., Rao, A.R., 2009. Prediction and interpretation of distributed neural activity with sparse models. NeuroImage 44 (1), 112–122.

Chen, P.-H.C., Chen, J., Yeshurun, Y., Hasson, U., Haxby, J., Ramadge, P.J., 2015. A reduced-dimension fMRI shared response model. In: Advances in Neural Information Processing Systems. pp. 460–468.

Chen, J.E., Jahanian, H., Glover, G.H., 2017. Nuisance regression of high-frequency functional magnetic resonance imaging data: denoising can be noisy. Brain Connect. 7 (1), 13–24.

Choudhury, S., Fishman, J.R., McGowan, M.L., Juengst, E.T., 2014. Big data, open science and the brain: lessons learned from genomics. Front. Hum. Neurosci. 8, 239.

Cohen, J.D., Daw, N., Engelhardt, B., Hasson, U., Li, K., Niv, Y., Norman, K.A., Pillow, J., Ramadge, P.J., Turk-Browne, N.B., et al., 2017. Computational approaches to fMRI analysis. Nature Neurosci. 20 (3), 304.

Coleman, T., Li., Y., 1996. An interior, trust region approach for nonlinear minimization subject to bounds. SIAM J. Optim. 6, 418–445.

Connolly, A.C., Guntupalli, J.S., Gors, J., Hanke, M., Halchenko, Y.O., Wu, Y.-C., Abdi, H., Haxby, J.V., 2012. The representation of biological classes in the human brain. J. Neurosci. 32 (8), 2608–2618.

Cox, D.D., Savoy, R.L., 2003. Functional magnetic resonance imaging (fMRI) "brain reading": detecting and classifying distributed patterns of fMRI activity in human visual cortex. NeuroImage 19 (2), 261–270.

DiCarlo, J.J., Zoccolan, D., Rust, N.C., 2012. How does the brain solve visual object recognition? Neuron 73 (3), 415–434.

Diedrichsen, J., Kriegeskorte, N., 2017. Representational models: A common framework for understanding encoding, pattern-component, and representational-similarity analysis. PLoS Comput. Biol. 13 (4), e1005508.

Diedrichsen, J., Provost, S., Zareamoghaddam, H., 2016. On the distribution of cross-validated mahalanobis distances. ArXiv preprint arXiv:1607.01371.

Diedrichsen, J., Ridgway, G.R., Friston, K.J., Wiestler, T., 2011. Comparing the similarity and spatial structure of neural representations: a pattern-component model. NeuroImage 55 (4), 1665–1678.

Diedrichsen, J., Yokoi, A., Arbuckle, S.A., 2018. Pattern component modeling: a flexible approach for understanding the representational structure of brain activity patterns. NeuroImage 180, 119–133.

Dumoulin, S.O., Wandell, B.A., 2008. Population receptive field estimates in human visual cortex. NeuroImage 39 (2), 647–660.

Eger, E., Ashburner, J., Haynes, J.-D., Dolan, R.J., Rees, G., 2008. fMRI activity patterns in human LOC carry information about object exemplars within category. J. Cogn. Neurosci. 20 (2), 356–370.

Ellis, C.T., Baldassano, C., Schapiro, A.C., Cai, M.B., Cohen, J.D., 2020. Facilitating open-science with realistic fMRI simulation: validation and application. PeerJ 8, e8564.

Epstein, R., Harris, A., Stanley, D., Kanwisher, N., 1999. The parahippocampal place area: Recognition, navigation, or encoding? Neuron 23 (1), 115–125.

Ericsson, K.A., Simon, H.A., 1980. Verbal reports as data. Psychol. Rev. 87 (3), 215.

Esteban, O., Markiewicz, C.J., Blair, R.W., Moodie, C.A., Isik, A.I., Erramuzpe, A., Kent, J.D., Goncalves, M., DuPre, E., Snyder, M., et al., 2019. fMRIPrep: a robust preprocessing pipeline for functional MRI. Nature Methods 16 (1), 111.

Etz, A., Vandekerckhove, J., 2018. Introduction to Bayesian inference for psychology. Psychon. Bull. Rev. 25 (1), 5–34.

Faul, A.C., Tipping, M.E., 2002. Analysis of sparse bayesian learning. In: Advances in Neural Information Processing Systems. pp. 383–389.

Finn, E.S., Constable, R.T., 2016. Individual variation in functional brain connectivity: implications for personalized approaches to psychiatric disease. Dialogues Clin. Neurosci. 18 (3), 277.

Friedman, N., 2004. Inferring cellular networks using probabilistic graphical models. Science 303 (5659), 799–805.

Friston, K.J., Diedrichsen, J., Holmes, E., Zeidman, P., 2019. Variational representational similarity analysis. NeuroImage 115986.

Friston, K.J., Frith, C.D., Turner, R., Frackowiak, R.S., 1995. Characterizing evoked hemodynamics with fMRI. NeuroImage 2 (2), 157–165.

Friston, K.J., Harrison, L., Penny, W., 2003. Dynamic causal modelling. NeuroImage 19 (4), 1273–1302.

Gardner, J., Pleiss, G., Weinberger, K.Q., Bindel, D., Wilson, A.G., 2018. Gpytorch: blackbox matrix-matrix Gaussian Process inference with GPU acceleration. In: Advances in Neural Information Processing Systems. pp. 7576–7586.

Geisler, W.S., 2011. Contributions of ideal observer theory to vision research. Vis. Res. 51 (7), 771–781.

Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A., Rubin, D.B., 2013. Bayesian Data Analysis. Chapman and Hall/CRC.

Gorgolewski, K.J., Auer, T., Calhoun, V.D., Craddock, R.C., Das, S., Duff, E.P., Flandin, G., Ghosh, S.S., Glatard, T., Halchenko, Y.O., et al., 2016. The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. Sci. Data 3, 160044.

Gorgolewski, K., Esteban, O., Schaefer, G., Wandell, B., Poldrack, R., 2017. OpenNeuro—A Free Online Platform for Sharing and Analysis of Neuroimaging Data, vol. 1677. Organization for Human Brain Mapping, Vancouver, Canada.

Gramfort, A., Thirion, B., Varoquaux, G., 2013. Identifying predictive regions from fMRI with TV-L1 prior. In: 2013 International Workshop on Pattern Recognition in Neuroimaging. IEEE, pp. 17–20.

Greenewald, K., Hero, A.O., 2015. Robust Kronecker product PCA for spatio-temporal covariance estimation. IEEE Trans. Signal Process. 63 (23), 6368–6378.

Griffiths, T.L., Kemp, C., Tenenbaum, J.B., 2008. Bayesian Models of Cognition. Carnegie Mellon University.

Grosenick, L., Klingenberg, B., Katovich, K., Knutson, B., Taylor, J.E., 2013. Interpretable whole-brain prediction analysis with GraphNet. NeuroImage 72, 304–321.

Hastings, W.K., 1970. Monte Carlo Sampling Methods Using Markov Chains and Their Applications. Oxford University Press.

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

Haxby, J.V., Guntupalli, J.S., Connolly, A.C., Halchenko, Y.O., Conroy, B.R., Gobbini, M.I., Hanke, M., Ramadge, P.J., 2011. A common, high-dimensional model of the representational space in human ventral temporal cortex. Neuron 72 (2), 404–416.

Haynes, J.-D., Rees, G., 2005. Predicting the orientation of invisible stimuli from activity in human primary visual cortex. Nature Neurosci. 8 (5), 686–691.

Heeger, D.J., Ress, D., 2002. What does fMRI tell us about neuronal activity? Nat. Rev. Neurosci. 3 (2), 142.

Henriksson, L., Khaligh-Razavi, S.-M., Kay, K., Kriegeskorte, N., 2015. Visual representations are dominated by intrinsic fluctuations correlated between areas. NeuroImage 114, 275–286.

Iordan, M.C., Greene, M.R., Beck, D.M., Fei-Fei, L., 2015. Basic level category structure emerges gradually across human ventral visual cortex. J. Cogn. Neurosci. 27 (7), 1427–1446.

Jeffreys, H., 1998. The Theory of Probability. OUP Oxford.

Ji, Q., 2019. Probabilistic Graphical Models for Computer Vision. Academic Press.

Jordan, M.I., Ghahramani, Z., Jaakkola, T.S., Saul, L.K., 1999. An introduction to variational methods for graphical models. Mach. Learn. 37 (2), 183–233.

Kandel, E.R., Markram, H., Matthews, P.M., Yuste, R., Koch, C., 2013. Neuroscience thinks big (and collaboratively). Nat. Rev. Neurosci. 14 (9), 659.

Katanoda, K., Matsuda, Y., Sugishita, M., 2002. A spatio-temporal regression model for the analysis of functional MRI data. NeuroImage 17 (3), 1415–1428.

Koller, D., Friedman, N., 2009. Probabilistic Graphical Models: Principles and Techniques. MIT Press.

Kriegeskorte, N., Mur, M., Bandettini, P.A., 2008. Representational similarity analysis-connecting the branches of systems neuroscience. Front. Syst. Neurosci. 2, 4.

Kruschke, J.K., 2010. What to believe: Bayesian methods for data analysis. Trends Cogn. Sci. 14 (7), 293–300.

Kumar, M., Ellis, C.T., Lu, Q., Zhang, H., Capota, M., Willke, T.L., Ramadge, P.J., Turk-Browne, N., Norman, K., 2019. BrainIAK Tutorials: User-Friendly Learning Materials for Advanced Fmri Analysis. OSF Preprints.

LaConte, S., Strother, S., Cherkassky, V., Anderson, J., Hu, X., 2005. Support vector machines for temporal classification of block design fMRI data. NeuroImage 26 (2), 317–329.

Landhuis, E., 2017. Neuroscience: Big Brain, Big Data. Nature Publishing Group.

Lawrence, N.D., 2005. Probabilistic non-linear principal component analysis with Gaussian process latent variable models. J. Mach. Learn. Res. 6, 1783–1816.

Leonardi, N., Van De Ville, D., 2015. On spurious and real fluctuations of dynamic functional connectivity during rest. NeuroImage 104, 430–436.

Lindquist, M.A., Geuter, S., Wager, T.D., Caffo, B.S., 2019. Modular preprocessing pipelines can reintroduce artifacts into fMRI data. Hum. Brain Mapp. 40 (8), 2358–2376.

Ma, W.J., 2012. Organizing probabilistic models of perception. Trends Cogn. Sci. 16 (10), 511–518.

MacKay, D.J., Mac Kay, D.J., 2003. Information Theory, Inference and Learning Algorithms. Cambridge University Press.

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

Manning, J., Zhu, X., Willke, T., Ranganath, R., Stachenfeld, K., Hasson, U., Blei, D., Norman, K., 2018. A probabilistic approach to discovering dynamic full-brain functional connectivity patterns. NeuroImage 180, 243–252.

Marr, D., Nishihara, H.K., 1978. Representation and recognition of the spatial organization of three-dimensional shapes. Proc. R. Soc. Lond. Biol. 200 (1140), 269–294.

Mazziotta, J., Toga, A., Evans, A., Fox, P., Lancaster, J., Zilles, K., Woods, R., Paus, T., Simpson, G., Pike, B., et al., 2001. A probabilistic atlas and reference system for the human brain: International Consortium for Brain Mapping (ICBM). Philos. Trans. R. Soc. London Biol. 356 (1412), 1293–1322.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., Teller, E., 1953. Equation of state calculations by fast computing machines. J. Chem. Phys. 21 (6), 1087–1092.

Michel, V., Gramfort, A., Varoquaux, G., Eger, E., Thirion, B., 2011. Total variation regularization for fMRI-based prediction of behavior. IEEE Trans. Med. Imaging 30 (7), 1328–1340.

Mumford, J.A., Nichols, T., 2006. Modeling and inference of multisubject fMRI data. IEEE Eng. Med. Biol. Mag. 25 (2), 42–51.

Murphy, K., Birn, R.M., Handwerker, D.A., Jones, T.B., Bandettini, P.A., 2009. The impact of global signal regression on resting state correlations: are anti-correlated networks introduced? NeuroImage 44 (3), 893–905.

Murphy, K., Fox, M.D., 2017. Towards a consensus regarding global signal regression for resting state functional connectivity MRI. NeuroImage 154, 169–173.

Naselaris, T., Kay, K.N., Nishimoto, S., Gallant, J.L., 2011. Encoding and decoding in fMRI. NeuroImage 56 (2), 400–410.

Naselaris, T., Prenger, R.J., Kay, K.N., Oliver, M., Gallant, J.L., 2009. Bayesian Reconstruction of natural images from human brain activity. Neuron 63 (6), 902–915.

Neal, R.M., 2012. Bayesian Learning for Neural Networks, vol. 118. Springer Science & Business Media.

Nili, H., Wingfield, C., Walther, A., Su, L., Marslen-Wilson, W., Kriegeskorte, N., 2014. A toolbox for representational similarity analysis. PLoS Comput. Biol. 10 (4), e1003553.

Nishimoto, S., Vu, A.T., Naselaris, T., Benjamini, Y., Yu, B., Gallant, J.L., 2011. Reconstructing visual experiences from brain activity evoked by natural movies. Curr. Biol. 21 (19), 1641–1646.

Ogawa, S., Lee, T.-M., Kay, A.R., Tank, D.W., 1990. Brain magnetic resonance imaging with contrast dependent on blood oxygenation. Proc. Natl. Acad. Sci. 87 (24), 9868–9872.

O'Toole, A.J., Jiang, F., Abdi, H., Pénard, N., Dunlop, J.P., Parent, M.A., 2007. Theoretical, statistical, and practical perspectives on pattern-based classification approaches to the analysis of functional neuroimaging data. J. Cogn. Neurosci. 19 (11), 1735–1752.

Park, M., Pillow, J.W., 2011. Receptive field inference with localized priors. PLoS Comput. Biol. 7 (10), e1002219.

Poldrack, R.A., Gorgolewski, K.J., 2014. Making big data open: data sharing in neuroimaging. Nature Neurosci. 17 (11), 1510.

Rakitsch, B., Lippert, C., Borgwardt, K., Stegle, O., 2013. It is all in the noise: efficient multi-task gaussian process inference with structured residuals. In: Advances in Neural Information Processing Systems. pp. 1466–1474.

Ramírez, F.M., 2017. Representational confusion: the plausible consequence of demeaning your data. bioRxiv 195271.

Rasmussen, C.E., 2003. GaussIan processes in machine learning. In: Summer School on Machine Learning. Springer, pp. 63–71.

Roś, B., Bijma, F., de Gunst, M., de Munck, J., 2014. A three domain covariance framework for EEG/MEG data. NeuroImage 119, 305–315.

Rubinov, M., Sporns, O., 2010. Complex network measures of brain connectivity: uses and interpretations. Neuroimage 52, 1059–1069.

Saad, Z.S., Gotts, S.J., Murphy, K., Chen, G., Jo, H.J., Martin, A., Cox, R.W., 2012. Trouble at rest: how correlation patterns and group differences become distorted after global signal regression. Brain Connect. 2 (1), 25–32.

Sahani, M., Linden, J.F., 2003. Evidence optimization techniques for estimating stimulus-response functions. In: Advances in Neural Information processing systems. pp. 317–324.

Salvatier, J., Wiecki, T.V., Fonnesbeck, C., 2016. Probabilistic programming in Python using PyMC3. PeerJ Comput. Sci. 2016 (4), 1–24.

Schmolck, A., 2008. Smooth Relevance Vector Machines (Ph.D. thesis). University of Exeter.

Sejnowski, T.J., Churchland, P.S., Movshon, J.A., 2014. Putting big data to good use in neuroscience. Nature Neurosci. 17 (11), 1440.

Shepard, R.N., Chipman, S., 1970. Second-order isomorphism of internal representations: Shapes of states. Cogn. Psychol. 1 (1), 1–17.

Shiffrin, R.M., Lee, M.D., Kim, W., Wagenmakers, E.-J., 2008. A survey of model evaluation approaches with a tutorial on hierarchical Bayesian methods. Cogn. Sci. 32 (8), 1248–1284.

Shvartsman, M., Sundaram, N., Aoi, M., Charles, A., Willke, T.L., Cohen, J.D., 2018. Matrix-normal models for fMRI analysis. In: International Conference on Artificial Intelligence and Statistics, AISTATS 2018, pp. 1914–1923, http://proceedings.mlr.press/v84/shvartsman18a.html.

Simony, E., Honey, C.J., Chen, J., Lositsky, O., Yeshurun, Y., Wiesel, A., Hasson, U., 2016. Dynamic reconfiguration of the default mode network during narrative comprehension. Nature Commun. 7 (May 2015), 12141.

Skolidis, G., Sanguinetti, G., 2011. Bayesian Multitask classification with Gaussian process priors. IEEE Trans. Neural Netw. 22 (12), 2011–2021.

Stegle, O., Lippert, C., Mooij, J.M., Lawrence, N.D., Borgwardt, K., 2011. Efficient inference in matrix-variate Gaussian models with iid observation noise. In: Advances in Neural Information Processing Systems. pp. 630–638.

Stephan, K.E., Penny, W.D., Moran, R.J., den Ouden, H.E., Daunizeau, J., Friston, K.J., 2010. Ten simple rules for dynamic causal modeling. NeuroImage 49 (4), 3099–3109.

Suárez, L.E., Markello, R.D., Betzel, R.F., Misic, B., 2020. Linking structure and function in macroscale brain networks. Trends Cogn. Sci.

Talairach, J., 1988. Co-planar stereotaxic atlas of the human brain-3-dimensional proportional system. In: An Approach to Cerebral Imaging. Georg Thieme Verlag.

Tipping, M.E., 2001. Sparse Bayesian learning and the relevance vector machine. J. Mach. Learn. Res. 1, 211–244.

Triantafyllou, C., Hoge, R.D., Krueger, G., Wiggins, C.J., Potthast, A., Wiggins, G.C., Wald, L.L., 2005. Comparison of physiological noise at 1.5 T, 3 T and 7 T and optimization of fMRI acquisition parameters. NeuroImage 26 (1), 243–250.

Turek, J.S., Willke, T.L., Chen, P.-H., Ramadge, P.J., 2017. A semi-supervised method for multi-subject fMRI functional alignment. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP. IEEE, pp. 1098–1102.

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

Van Horn, J.D., Toga, A.W., 2014. Human neuroimaging as a "Big Data" science. Brain Imaging Behav. 8 (2), 323–331.

Vodrahalli, K., Chen, P.-H., Liang, Y., Baldassano, C., Chen, J., Yong, E., Honey, C., Hasson, U., Ramadge, P., Norman, K.A., et al., 2018. Mapping between fMRI responses to movies and their natural language annotations. NeuroImage 180, 223–231.

Wang, K.A., Pleiss, G., Gardner, J.R., Tyree, S., Weinberger, K.Q., Wilson, A.G., 2019. Exact Gaussian processes on a million data points, no. 1. pp. 1–13.

Wipf, D.P., Nagarajan, S.S., 2008. A new view of automatic relevance determination. In: Advances in Neural Information Processing Systems. pp. 1625–1632.

Woolrich, M.W., 2012. Bayesian Inference in FMRI. NeuroImage 62 (2), 801–810.

Woolrich, M.W., Jbabdi, S., Patenaude, B., Chappell, M., Makni, S., Behrens, T., Beckmann, C., Jenkinson, M., Smith, S.M., 2009. Bayesian Analysis of neuroimaging data in FSL. NeuroImage 45 (1), S173–S186.

Woolrich, M.W., Jenkinson, M., Brady, J.M., Smith, S.M., 2004. Fully Bayesian spatio-temporal modeling of fMRI data. IEEE Trans. Med. Imaging 23 (2), 213–231.

Wu, A., Koyejo, O., Pillow, J., 2019. Dependent relevance determination for smooth and structured sparse regression. J. Mach. Learn. Res. 20 (89), 1–43.

Zarahn, E., Aguirre, G.K., D'Esposito, M., 1997. Empirical analyses of BOLD fMRI statistics. NeuroImage 5 (3), 179–197.

Zhang, H., Chen, P.-H., Chen, J., Zhu, X., Turek, J.S., Willke, T.L., Hasson, U., Ramadge, P.J., 2016. A searchlight factor model approach for locating shared information in multi-subject fMRI analysis. ArXiv preprint arXiv:1609.09432.

Zhang, H., Chen, P.-H., Ramadge, P., 2018. Transfer learning on fMRI datasets. In: International Conference on Artificial Intelligence and Statistics, pp. 595–603, http://proceedings.mlr.press/v84/zhang18b.html.