Matrix-normal models for fMRI analysis

Michael Shvartsman[1]
ms44@princeton.edu

Narayanan Sundaram[2]
narayanan.sundaram@intel.com

Mikio Aoi[1]
maoi@princeton.edu

Adam Charles[1]
adamsc@princeton.edu

Theodore L. Willke[2]
ted.wilke@intel.com

Jonathan D. Cohen[1]
jdc@princeton.edu

[1]Princeton Neuroscience Institute
Princeton University

[2]Parallel Computing Lab
Intel Corporation

Abstract

Multivariate analysis of fMRI data has benefited substantially from advances in machine learning. Most recently, a range of probabilistic latent variable models applied to fMRI data have been successful in a variety of tasks, including identifying similarity patterns in neural data, combining multi-subject datasets, and mapping between brain and behavior. Although these methods share some underpinnings, they have been developed as distinct methods, with distinct algorithms and software tools. We show how the matrix-variate normal (MN) formalism can unify some of these methods into a single framework. In doing so, we gain the ability to reuse noise modeling assumptions, algorithms, and code across models. Our primary theoretical contribution shows how some of these methods can be written as instantiations of the same model, allowing us to generalize them to flexibly modeling structured residual covariances. Our formalism permits novel model variants and improved estimation strategies for SRM and RSA using substantially fewer parameters. We empirically demonstrate advantages of our two new methods: for MN-RSA, we show up to 10x improvement in runtime, up to 6x improvement in RMSE, and more conservative behavior under the null. For MN-SRM, our method grants a modest improvement to out-of-sample reconstruction while relaxing the orthonormality constraint of SRM. We also provide a software prototyping tool for MN models that can flexibly reuse residual covariance assumptions and algorithms across models.

1 Introduction

Functional magnetic resonance imaging (fMRI) analysis is a challenging problem for statistics and machine learning: signal-to-noise ratio for extracting scientifically meaningful information is low, and physiological and instrumentation noise creates correlations in space and time that can mask signal and magnify false alarms. Recent methods have been developed in the statistics and machine learning community to address these challenges, including for dimension reduction and subject-to-subject mapping [7, 16]), estimation of patterns of neural similarity within and across subjects [5, 21], and mapping from brain to behavior via latent cognitive models [25].

These models are similar in that they seek a latent, typically low-rank, structure in fMRI data using multivariate gaussian models. Yet they are different in the quantity they attempt to estimate, and in the estimation methods they use. Furthermore, these techniques are restricted to modeling only either temporal or spatial correlation (or neither), even though both spatial and temporal noise structure exists in the data. These differences make it difficult to share insights and advances across techniques. In this work we show that matrix-variate (MN) normal models provide a powerful formalism for understanding and developing fMRI data analysis methods in a unified way.

Specifically, we show that many existing methods can be derived from the MN framework. MN variants of these methods are not restricted in their noise model and can simultaneously capture spatial and temporal

Proceedings of the 21^{st} International Conference on Artificial Intelligence and Statistics (AISTATS) 2018, Lanzarote, Spain. PMLR: Volume 84. Copyright 2018 by the author(s).

noise. Furthermore, the shared mathematical structure enables the creation of an MN development software framework that admits flexible swapping between various covariance models—a task that otherwise involves substantial engineering effort.

In addition to showing the formal connection between existing methods, we also use the formalism to develop two novel analyses, MN-RSA and MN-SRM. MN-RSA outperforms the previous a state-of-the-art method in both speed and accuracy, and MN-SRM method improves on SRM in reconstruction accuracy.

Our contributions are as follows:

1. Motivation for MN models as a unifying mathematical model for fMRI analysis, illustrating its wide applicability with examples from both regression (via RSA) and factor analysis (via SRM).
2. A toolkit for developing MN models using Tensorflow [1], and implementations of RSA and SRM variants that can model both spatial and temporal covariance. (§4).
3. An expectation-conditional-maximization (ECM) algorithm for fitting MN-SRM, which removes the orthonormality constraint of SRM and in which the number of parameters does not scale with the number of subjects, unlike SRM (§4).
4. Demonstration that MN-RSA is approximately an order of magnitude faster than the previous state of the art method, can be up to 6x more accurate at SNRs as low as 0.08 and thousands of voxels, and is most conservative under the null hypothesis.
5. Demonstration that MN-SRM can improve on SRM performance in terms of reconstruction (§5).

The remainder of the paper is organized as follows: we discuss background and related work in §2. §3 provides motivation for our formalism, and derives MN-RSA and MN-SRM. §4 discusses our software implementation and the challenges involved therein. We show the results of our experiments in §5 and conclude in §6 with some discussion as to how other cutting edge analyses fall into our framework.

2 Background

fMRI uses the magnetic properties of oxygenated blood to measure blood flow in the brain as a proxy for neural computation. fMRI data exhibits temporal and spatial correlations due to blood flow dynamics, acquisition constraints, and the spatially distributed temporally evolving mental computation itself. With computational and theoretical advances, Multi-Voxel Pattern Analysis (MVPA; 17) has leveraged successes in machine learning for decoding more sophisticated representations and processes from fMRI data. A number of recent analyses pipelines have relied on gaussian latent variable models due to their wide applicability and computational tractability. We focus on two here – SRM and RSA – due to their wide use and broad applicability, though we treat two additional models in the supplementary material.

2.1 Representational Similarity Analysis (RSA)

The goal of RSA [13, 14, 29] is to use distances between correlations or other (dis)similarity metrics between responses to stimuli in the fMRI dataset to theoretically predicted distances. Due to the isomorphism between correlation and regression, the standard RSA estimator is biased when applied to within-run data, but the state of the art empirical Bayes method based on maximum marginal likelihood (BRSA) mitigates this particular bias [5].

2.2 Shared response mapping

A challenge in analyzing grouped data (e.g. coming from multiple subjects) is that while we expect subjects to have similar mental responses to a given stimuli, these responses may be idiosyncratically realized in the neural signal. For classification-based decoding and other discriminative analyses that fall under the rubric of MVPA, managing this is called the hyperalignment problem. Hyperalignment models project all subjects into a shared space that is used for analysis. SRM [7] is one recent factor-analytic hyperalignment method that linearly projects all subjects' data into a shared, low-dimensional functional timecourse. SRM can also be used for feature selection to enable state of the art decoding performance.

3 Matrix Normal Models for fMRI

Conventional multivariate fMRI analysis methods choose whether to model noise covariance in space or in time, while assuming independence in the other dimension, an assumption violated in real fMRI data as noted above.

MN models, also known as kronecker-separable covariance models, provide a formalism addressing the problem of multivariate data analysis [e.g. 3, 28]. The matrix-normal distribution is defined as:

$$\begin{aligned}\mathbf{X} \sim& \mathcal{MN}_{mn}(\mathbf{M}, \mathbf{R}, \mathbf{C}) && (1)\\ \log p(\mathbf{X} \mid \mathbf{M}, \mathbf{R}, \mathbf{C}) =& -2\log mn - m \log|\mathbf{C}| && (2)\\ & - n\log|\mathbf{R}| - \mathrm{Tr}\left[\mathbf{C}^{-1}(\mathbf{X}-\mathbf{M})^\top \mathbf{R}^{-1}(\mathbf{X}-\mathbf{M})\right].\end{aligned}$$

The intuition behind the kronecker separability is that if $\mathbf{Y} \sim \mathcal{MN}(\mathbf{M}, \mathbf{R}, \mathbf{C})$ then $\mathrm{vec}(\mathbf{Y}) \sim \mathcal{N}(\mathrm{vec}(\mathbf{M}), \mathbf{C} \otimes \mathbf{R})$, where $\otimes$ is the kronecker product operator and vec is the vectorization operator. In the case of fMRI, a kronecker-separable covariance assumes that spatial covariance is the same at every time, and temporal covariance is the same for every voxel. The covariance between any two voxels at two times is a product of their space and time covariance.

In this section, we show how the popular representational similarity analysis (RSA; Cai et al. 5, Kriegeskorte 14) and shared response mapping (SRM; Chen et al. 7) methods in neuroimaging can be written as matrix-normal models. We perform similar derivations for intersubject functional connectivity (ISFC; Simony et al. 21) and joint modeling (JM; Turner et al. 25) in the supplement. We begin with RSA [14]. Standard correlation-based RSA estimates stimulus-by-stimulus distances in brain activity space. If the distance matrix is a correlation matrix, this process is equivalent to encoding the predicted process model components (e.g. Markov states for reinforcement learning, or neural network activations) in a design matrix $\mathbf{X}$, under the linear model [5]:

$$\mathbf{y}_i \mid \mathbf{X}, \beta_i, \tau^2 \sim \mathcal{N}(\mathbf{X}\beta_i, \tau^{-2}\mathbf{I}), \tag{3}$$

where τ^2 is the residual precision, $\mathbf{y}_i$ is the (centered) timecourse of the ith voxel, and the coefficient vector β_i is the response pattern of each voxel to the modeled stimulus. The empirical row correlation of the β's is the RSA correlation matrix. If one uses point estimates of β to compute the RSA correlation matrix the estimator is biased, and this bias can inject structure from the design matrix into the estimate [5]. Bayesian RSA (BRSA; [5]) instead marginalizes over β:

$$\begin{aligned}\beta &\sim \mathcal{N}(0, \mathbf{U}) && (4)\\ \mathbf{y}_i \mid \mathbf{X}, \beta_i, \tau^2 &\sim \mathcal{N}(0, \tau^{-2}\mathbf{I} + \mathbf{X}\mathbf{U}\mathbf{X}^\top), && (5)\end{aligned}$$

and performs MAP estimation on $\mathbf{U}$ by gradient descent, mitigating this bias.

Now we derive matrix-variate RSA. To write a matrix-normal density for RSA (MN-RSA), we stack all of the $\mathbf{y}_i$ vectors into a matrix $\mathbf{Y}$, and stack all of the regression weights β into a matrix $\mathbf{W}$:

$$\begin{aligned}\mathbf{Y} \mid \mathbf{W}, \Sigma_t, \Sigma_v &\sim \mathcal{MN}(\mathbf{X}\mathbf{W}, \Sigma_t, \Sigma_{v,\mathbf{Y}}) && (6)\\ \mathbf{W} \mid \mathbf{U}, \Sigma_v &\sim \mathcal{MN}(0, \mathbf{U}, \Sigma_{v,\mathbf{W}}) && (7)\end{aligned}$$

This model can no longer model voxel-specific temporal correlations, but in return can model the residual spatial covariance Σ_v. This tradeoff will play out differently in different datasets. In this multilinear regression form, this problem appears similar to a number of models used recently in the multi-task learning literature [e.g. 4, 9, 18, 22, 23]. However, unlike those settings, the estimation target in RSA is the covariance $\mathbf{U}$ rather than predicted data in new tasks.

In previous work on estimating kronecker-separable covariances, both the signal and noise spatial covariances are assumed to be different, i.e. $\Sigma_{v,\mathbf{Y}} \neq \Sigma_{v,\mathbf{W}}$. As a result, the marginal covariance (marginalizing over $\mathbf{W}$) is a sum of kronecker factors, which previous work had estimated using Permuted Rank-penalized Least Squares [9–11] or gradient descent exploiting the compatibility between diagonalization and the kronecker product for efficient likelihood computation [18, 20, 23].

In fMRI, spatiotemporal covariance is driven by a combination of physiological factors (blood flow), instrument constraints, and task-related shared structure. We assume that the covariance is dominated by task-irrelevant factors, i.e. we assume $\Sigma_v := \Sigma_{v,\mathbf{Y}} = \Sigma_{v,\mathbf{W}}$. This gives us a convenient matrix-normal marginal likelihood (see supplement for derivation):

$$\mathbf{Y} \mid \mathbf{U}, \Sigma_t, \Sigma_v \sim \mathcal{MN}(0, \Sigma_t + \mathbf{X}\mathbf{U}\mathbf{X}^\top, \Sigma_v), \tag{8}$$

which we term the *MN-RSA* model.

3.1 Matrix-variate shared response model

Consider the following factor analysis model for fMRI data for multiple subjects:

$$\mathbf{y_{jk}} \mid \mathbf{W}_k, \mathbf{s}_j, \Sigma_v, \tau_k \sim \mathcal{N}(\mathbf{W}_k\mathbf{s}_j, \tau_k^{-2}\Sigma_v), \tag{9}$$

where $\mathbf{y}_{jk}$ is a mean-centered vector containing all voxel activities at a single timepoint (rather than $\mathbf{y}_i$, the single voxel's time series in Eq. 3). We have also added indexing by subject k, since SRM (unlike RSA) is a multi-subject method. $\mathbf{s}_j$ is a latent spatial map *for all subjects* for that particular time point, and $\mathbf{W}$ is subject-specific a projection matrix from the shared map into that subject's data. Σ_v is a shared *spatial* residual covariance as in MN-RSA above, scaled by a subject-specific precision τ_k^2. To make a matrix-variate factor model, we row-stack $\mathbf{y}_{jk}{}^\top$ into $\mathbf{Y}_k{}^\top$, stack $\mathbf{s}_j$ into $\mathbf{S}$, and obtain the following model:

$$\mathbf{Y}_k{}^\top \mid \mathbf{W}_k, \mathbf{S}, \mu, \Sigma_v, \tau_k \sim \mathcal{MN}_{v,n}(\mathbf{W}_k\mathbf{S}^\top, \tau_k^{-2}\Sigma_v, \mathbf{I}) \tag{10}$$

This factor analysis model now has the exact same form as the regression model above, except that $\mathbf{X}_k$ is observed and $\mathbf{S}$ is latent. Both of these matrix-variate models now have the exact same form: a mean that is an intercept plus a product of two matrices, one fully-specified covariance, and an identity covariance.

We drop the subject indices by row-stacking all of the subject timecourses $\mathbf{X}_j$ into $\mathbf{X}$ and weights $\mathbf{W}_j$ into $\mathbf{W}$, and introducing a subject covariance $\rho := \mathrm{diag}(\tau_1^{-2}, \tau_1^{-2}, \ldots, \tau_n^{-2})$:

$$\mathbf{S} \sim \mathcal{MN}(0, \Sigma_s, \Sigma_t) \tag{11}$$

$$\mathbf{W} \sim \mathcal{MN}(0, \rho \otimes \Sigma_v, \Sigma_w) \tag{12}$$

$$\mathbf{X} \mid \mathbf{W}, \mathbf{S}, \Sigma_t, \Sigma_v, \rho \sim \mathcal{MN}(\mathbf{WS}, \rho \otimes \Sigma_v, \Sigma_t), \tag{13}$$

giving us the *MN-SRM* model[1]. The covariances Σ_w, Σ_s are both set to $\mathbf{I}$ to regularize the model. The model implies that all subjects share a latent timecourse $\mathbf{S}$, as well as temporal and spatial noise covariances Σ_v, Σ_t that are scaled independently for each subject[2]. In practice, we restrict the form of both the spatial and temporal residuals to be diagonal or autoregressive, since estimating unconstrained Σ_v, Σ_t is still intractable at fMRI scale.

If $\Sigma_t = \Sigma_v = \mathbf{I}$ and $\mathbf{W}_k{}^\top \mathbf{W}_k = \mathbf{I} \quad \forall k$, this MN-SRM model is exactly the SRM model. However, in the MN formulation, we see that we have two marginalization choices: the first is marginalizing over the shared timecourse $\mathbf{S}$, as SRM does, and the second is marginalizing over the mappings $\mathbf{W}$ instead. We choose the latter marginalization, which only estimates tk parameters (t timepoints, k features) rather than the vnk parameters estimated by the original SRM method, which is appealing because for whole-brain analyses $v \gg t$. It also replaces the strong orthonormal constraint on $\mathbf{W}$ with a weaker Gaussian prior. This is a theoretically desirable property for the following reason: if the true data is not generated with orthonormal $\mathbf{W}$ per subject, forcing an orthonormal $\mathbf{W}$ makes $\mathbf{S}$ counter-rotate against it. With a single subject $\mathbf{W}^\top \mathbf{W} = \mathbf{I}$ w.l.o.g., but with multiple subjects, the best $\mathbf{S}$ for each subject is rotated differently to maintain orthonormality for that subject, giving a worse group $\mathbf{S}$. We later validate this intuition empirically in our reconstruction task.

4 Estimation and the `matnormal` prototyping tool

We leverage the shared structure of MN models to develop a unified framework for estimation using Python and the Tensorflow library. The implementation is flexible in the specification of residual covariances: for a (spatial or temporal) covariance Σ, the API only requires implementations of $\Sigma^{-1}\mathbf{X}$ and $\log|\Sigma|$ given $\mathbf{X}$ for efficient computation of marginal likelihoods marginalizing in either the row or column direction. This gives users the ability to choose the noise model complexity relative to the size of their data – or the ability to explore a large number of models with simple noise quickly before selecting a more complex noise model for later analysis.

All other routines are automatically derived from these, including marginalization that automatically leverages efficiencies derived from non-marginal covariance structures using the Woodbury and Sylvester lemmas. The amount of shared code allows, for example, MN-RSA to be implemented in only 50 lines of python code. We have implemented isotropic, diagonal, full rank, AR(1) kernel, squared exponential kernel, and Kronecker factored covariances. That is, we can further factor the spatial covariance Σ_v into $\Sigma_z \otimes \Sigma_y \otimes \Sigma_x$, where x, y, z are the spatial dimensions. Using kronecker-factored spatial covariances in fMRI is challenging because, since the brain is not a perfect cube, the masking of voxels that do not contain brain partially violates the kronecker structure. We address this challenge by developing a fast algorithm for the inverse and determinant of a masked kronecker-factored covariance (detailed in the supplement), increasing the toolbox's utility to a wider variety of fMRI datasets. We also include example implementations of MN (multilinear) regression, MN factor analysis MN-RSA, and MN-SRM.

Using this toolbox, we can perform maximum marginal likelihood estimation using gradient descent, leveraging gradients automatically computed with Tensorflow and our covariance API for rapid prototyping. In practice, this is sufficient for single-subject estimates, and our results for MN-RSA below are all using gradient descent.

Directly maximizing the marginal likelihood for group models such as MN-SRM is substantially slower because RSA is fit to a single subject whereas SRM is fit to a ten subjects or more, an order of magnitude more data. To mitigate this issue, we derive an efficient expectation conditional maximization (ECM) algorithm for learning MN-SRM, which estimates the sufficient statistics of $\mathbf{S}$ in the E-step and performs conditional maximization updates of the remaining parameters in the M-step. When estimating MN-SRM using ECM in our toolkit, we can still impose structure on Σ_v and Σ_t, however only certain constraints allow closed-form covariance updates. Due to space constraints we delegate the ECM derivation to the supplementary mate-

[1] we omit the spatial mean μ_j in the derivation for brevity, though not in the implementation

[2] Since a kronecker-structured covariance is determined only up to a constant, the scale on Σ_t and Σ_v is isomorphic, except that by scaling Σ_v we make the remainder of the derivation more straightforward.

rial. The algorithm does not exploit any special properties of MN-SRM relative to other MN models, and should be applicable to them with minor changes.

5 Results

We validate the MN framework for fMRI by exploring the behavior of MN-RSA and MN-SRM in simulations and real data. To demonstrate the accuracy of MN-RSA, we explore its performance on synthetic data. Our focus on synthetic data is because neither out-of-sample prediction nor real-data ground truth for RSA is well-defined in the literature, with the standard measure for evaluating RSA methods comparing the estimated covariance to a behaviorally relevant matrix. Since RSA matrices consistent with behavior can arise due to estimator bias alone [5], this metric is not useful. We do show MN-RSA performance on real data to verify that it does not recover spurious correlations when the design matrix and brain data are unrelated, and highlight the need for the field to develop better predictive validation metrics for covariance estimation.

For MN-SRM, we perform two experiments. The first is an out-of-sample reconstruction experiment testing whether the shared response we recover can reconstruct a new subject's data. The second is using MN-SRM for feature extraction with the goal of classification. We fit all of the models using the Brain Imaging Analysis Kit (BrainIAK, `http://brainiak.org`), which also includes implementations of our methods.

5.1 MN-RSA

For the MN-RSA experiment we compare BRSA [5] against MN-RSA with diagonal spatial covariance temporal covariance consisting of an AR(1) component, plus a low-rank matrix with the rank set to 15. For the spatial variance, the intuition is that by learning the variance of each voxel we can better tune SNR. For temporal covariance, AR(1) is simple, expressive and comparable to BRSA. We excluded naive RSA from this comparison because of its known bias, and since BRSA has been shown to achieve superior performance [5].

Experiment 1: synthetic data We generated synthetic data using the BRSA example in the `brainiak` package. This synthetic dataset has AR(1) noise in the temporal domain, spatial noise generated from a gaussian process, and a number of design-irrelevant timecourses included. Thus, it violates both models' assumptions, and includes spatial structure that is challenging for non-matrix-variate models to handle. The synthetic datasets included two different SNR levels, three different numbers of timepoints known in the field as 'TRs' (equivalent to 1, 2, and 4 runs of the experiment), and two different numbers of voxels (2500, 10000) to show how the algorithms scale with noise, time, and space. We replicated each combination 10 times. Each model was run and timed separately on a full node of a compute cluster with two Intel® Xeon® E5-2670 processors at 2.6 GHz with hyper-threading enabled. The deviation from ceiling performance on simple synthetic training data suggests that these methods are not overfitting. MN-RSA is up to 10x faster than the reference implementation of BRSA (estimated using BFGS) on the largest problems (figure in supplement).

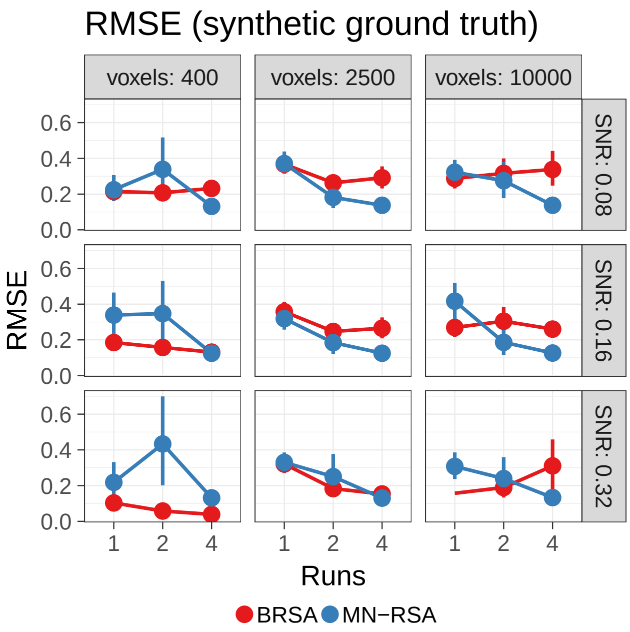

Figure 1: **MN-RSA performs better at larger numbers of voxels and lower SNR.** For smaller datasets (e.g. 400 voxels; not shown) and larger SNRs, BRSA performs better. The improved performance of MN-RSA is enabled by not modeling temporal noise independently for each voxel.

Fig. 1 shows estimated root-mean-squared-error (RMSE) performance against the true correlation matrix for the different numbers of voxels, TRs, and SNRs. MN-RSA obtains lower RMSE than BRSA in most settings, with the difference being particularly stark at larger problem sizes. In addition, because it estimates fewer parameters, MN-RSA can be up to 10x faster on the same hardware for large-scale problems.

Experiment 2: null data As noted above, there is no ground truth evaluation or RSA, and has been shown to recover spurious results under the null hypothesis when there is structure in the design matrix (for conventional RSA) or model mismatch in noise covariance (for all RSA methods) [5]. Consequently, an important evaluation of RSA methods is their behavior under the null hypothesis, i.e. where no signal is expected to exist.

In this experiment, we use a real resting state dataset [26] in which subjects are not given a task, with a random temporally contiguous window of 186 TRs selected from each participant and a lateral occipital cortex region of interest (ROI). With this resting state dataset, we used the same design matrix as in the experiment above, which is completely unrelated to the dataset. We show two example subjects' RSA covariance matrices under all three methods in Fig. 5.1, and the remainder in the supplement. Since MN-RSA estimates both the low-rank temporal structure and the $\mathbf{U}$ matrix simultaneously, it is capable of assigning next to zero variance to $\mathbf{U}$ if the design matrix is unrelated to the data. This feature means that for most subjects under the null hypothesis, $\mathbf{U}$ correctly approaches 0, and the RSA correlation matrix is clearly degenerate, in contrast to RSA and BRSA, both of which produce the appearance of structure. In the supplement, we also show the distribution of the elements of the estimated RSA matrix for each subject, with a clear spike at zero in a majority of subjects only for MN-RSA, showing that it is the most conservative method under the null hypothesis.

5.2 MN-SRM

We test two variants of MN-SRM. The first sets $\Sigma_v = \Sigma_t = \mathbf{I}$, differing from SRM only in the marginalization direction and the removal of the orthonormality constraint on $\mathbf{W}$. Since it has the same relationship to SRM that dual probabilistic PCA [15] has to PCA, we call it dual probabilistic SRM (DP-SRM). The second variant, MN-SRM, uses the diagonal Σ_v and AR(1) Σ_t we used for MN-RSA, above. We compare these models to SRM, as well as to ICA as a naive baseline. For these experiments, we use the *raider* [13] and *sherlock* [6] datasets (see Tab. 1for detail).

Experiment 3: out of sample reconstruction To test each model's ability to recover the shared latent time-course, we perform a held-out reconstruction experiment. We fit the factorization methods on all but one subject with 10, 30, or 50 features, and then learn a projection from the shared time-course into that new subject. Our loss metric is the reconstruction error of the held out subject's data using the estimated shared time-course, and the new subject's map. In both, we use the portion of the dataset where subjects watched the same movie (Raiders of the Lost Ark, and an episode of BBC's Sherlock).

Fig. 3 shows that the reconstruction error of both MN methods is consistently lower than that of SRM in the *raider* dataset, and lower in all but the smallest numbers of features in the *sherlock* dataset. The improvement of MN methods over SRM validates our assertion that we should be able to more effectively fit our model by marginalizing over a larger number of parameters, and shows that benefit of MN models' flexibility in removing the orthonormality constraint on $\mathbf{W}$. The relative performance between the MN methods on the two datasets is also interesting: on *raider*, adding the noise covariance modeling improves performance, whereas on *sherlock* it does not. The ultimate reason for this is an interesting scientific question, and provides validation for our approach of flexible noise covariance modeling: there may not be a one-size-fits-all hyperalignment method.

Experiment 4: feature extraction for classification One of the primary use-cases for SRM is as feature extraction method for classification. For this reason, and because classification performance was previously used to compare hyperalignment methods [7], we report performance on this task next. In both *raider* and *sherlock*, subjects begin by watching a movie clip, and then perform a cognitive task. SRM and similar methods can be used to learn a projection into a shared space while subjects are watching the same movie stimulus, and then use that learned mapping to project fMRI data recorded during the cognitive task. For *raider* this task was viewing one of 7 possible images, and for *sherlock* it was free-recalling scenes in the movie. We train a linear SVM to discriminate between the images subjects viewed in *raider*, and between the scenes in *sherlock*.

Fig 4 shows that in spite of our methods' better performance on the reconstruction task, their ability to extract features useful for linear classification lags behind the original SRM method. Note however that neither method is designed for post-hoc linear classification. Rather, both methods are designed to do unsupervised learning of a latent space that explains the shared variance across subjects based on the training data. By that metric our method outperforms SRM and differences in classification performance must be incidental.

The fact that SRM does better on classification suggests that there is movie-specific variability that exists across subjects but is nevertheless unrelated to the classification task. It is possible that the orthonormality constraint regularizes the solution for W away from subspaces explaining variability specific to the training movie, leaving the proportion of variability explained lower, while providing a higher SNR for information relevant to classification. However, this need not be

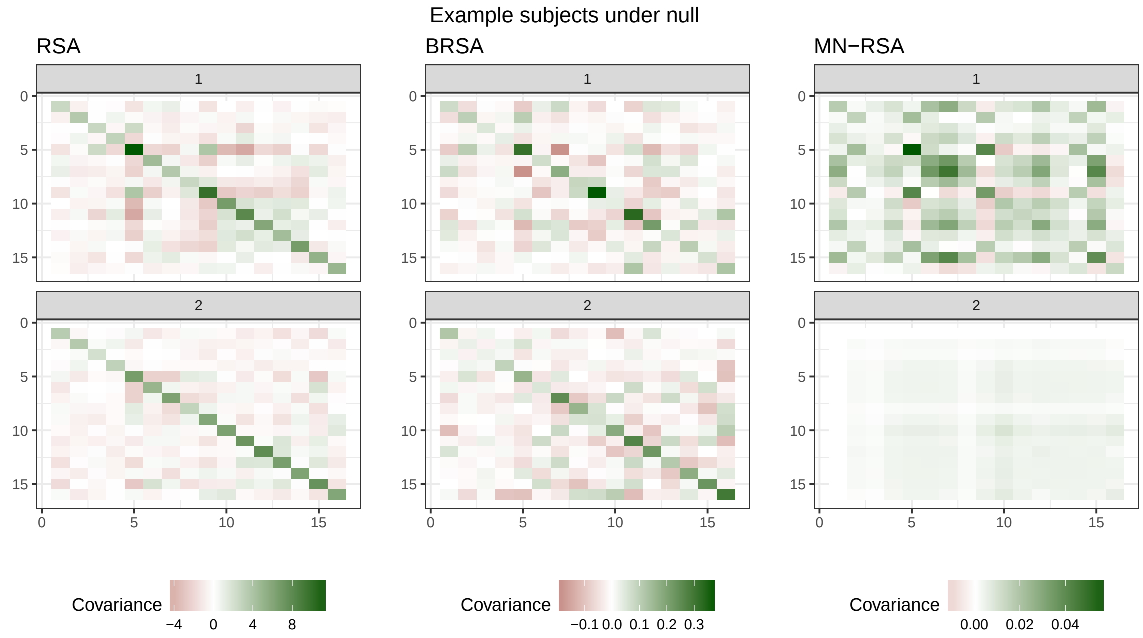

Figure 2: **MN-RSA is the only method that delivers obviously degenerate results under the null.** Estimates for all but three of the subjects look like subject 2 for MN-RSA, BRSA estimates all look homogeneous, and naive RSA estimates are about evenly split between finding low rank structure like subject 1 and not finding it like subject 2.

Dataset	Subjs.	TRs	Region of Interest	Voxels
sherlock [6]	16	1976	Posterior Medial Cortex	813
raider [13]	10	2203	Ventral Temporal Cortex	1000
HCP [26]	29	186	Lateral Occipital Cortex	2000

Table 1: fMRI dataset properties used for experiments 2, 3, and 4. We thank the authors for sharing their data.

the case and it is an empirical question as to why the orthogonality constraint confers this property.

6 Discussion and conclusion

Probabilistic multivariate analyses of fMRI data are a promising direction of research, combining the interpretability previously associated with univariate analyses with the power of multivariate approaches. However, advances tend to proceed independently of each other, with distinct methods and algorithms for different problems. At face value this is not surprising as they have substantial differences: SRM and TFA are unsupervised, while BRSA and ISFC are supervised; BRSA and ISFC are somewhat unusual in seeking the correlation matrices of latent variables, whereas SRM is more conventionally concerned with latent space projection, and TFA with inferring brain networks. In addition, they all use distinct fitting techniques: gradient-based maximum marginal likelihood for BRSA, expectation-maximization for SRM, and variational inference for TFA.

We showed how some of these methods can be viewed as closely related matrix-variate models, and how the matrix-variate view allows us to simultaneously model spatial and temporal noise covariances in both methods. In neuroscience, such models have been applied to MEG/EEG data [19], as well as non-latent models for fMRI data [12], with some evidence that a separable covariance is a reasonable approximation to fMRI data even though voxel temporal correlations vary with spatial location. Our work contrasts with this previous work both in its unification of distinct methods, and in bringing matrix-variate latent variable models to fMRI analysis more broadly.

In the MN view, we can show the relationship of some supervised fMRI analysis methods (RSA and ISFC) to multi-task regression and more broadly to kronecker-structured covariance models. Such models have been applied in areas as diverse as recommendation systems [2], environmental science [8, 24], MIMO chan-

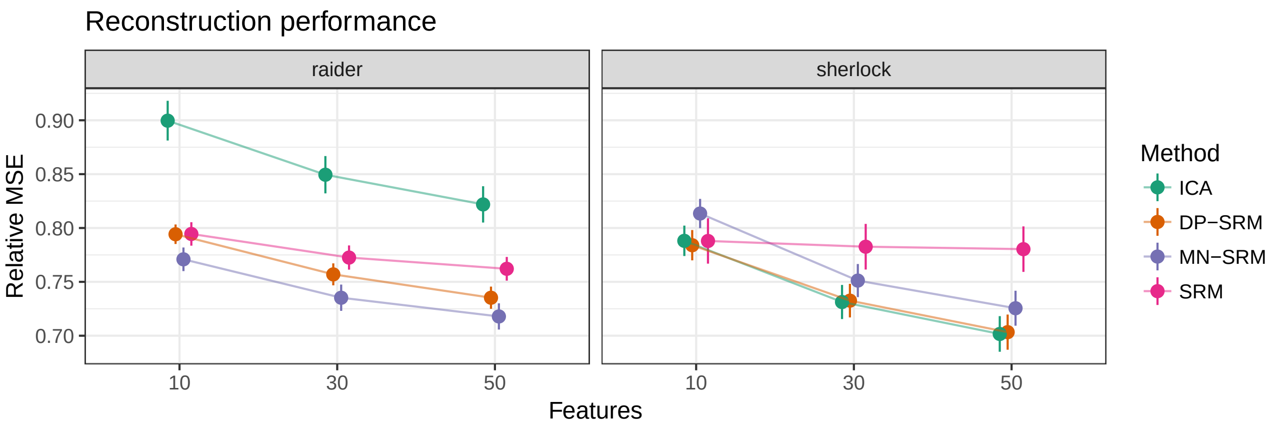

Figure 3: **MN-SRM and DP-SRM reconstruct the same or better than SRM and ICA.** All models are trained on n-1 subjects, and the shared timecourse used to reconstruct the nth subject. Plotted are means and standard error across subjects.

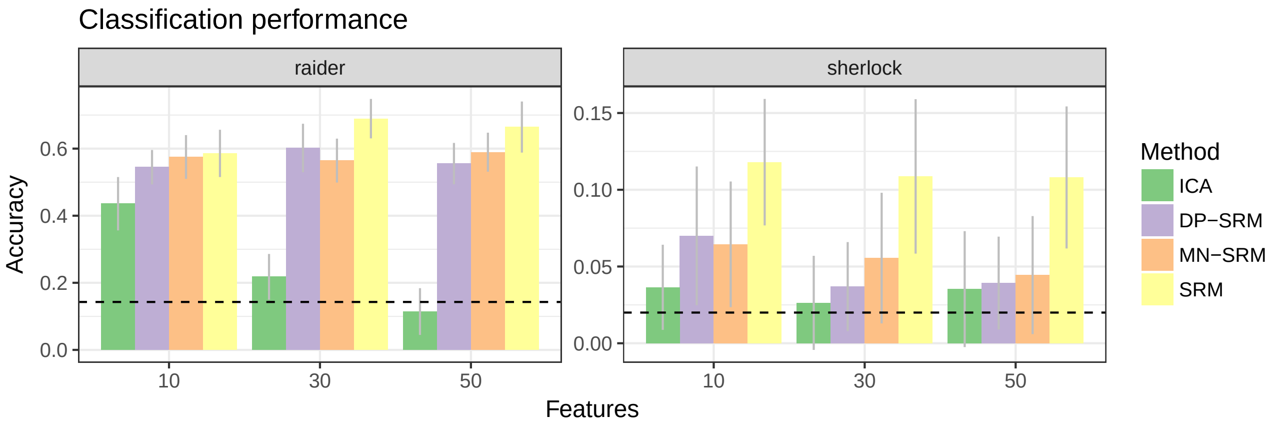

Figure 4: **MN-SRM and DP-SRM approach SRM performance in feature extraction, while relaxing the orthonormality constraint on W.** We train the SRM on all subjects watching a movie, and project the other task data into a shared space for classification. Plotted are means and error bars of out-of-sample prediction across subjects. The dashed line is chance performance.

nel behavior [27, 28], collaborative filtering [31], compiler performance prediction and student test score modeling [4], video understanding [11], and genomics [18, 23, 30]. However, in contrast to this existing work (and especially [18, 23], which is closest to our contribution), the nature of fMRI noise admits simpler noise covariance assumptions that in turn yield different techniques for efficient likelihood computation, and a novel expectation-conditional-maximization algorithm.

In addition to our theoretical contribution, we provided a software package for estimating the above models that allows for flexible assumptions about noise covariance, and provided evidence that for best performance, noise covariance assumptions may need to be adjusted for different datasets and tasks.

Our experiments also revealed opportunity for future work. For example, MN-RSA performed worse than the previous method at larger numbers of TRs and smaller numbers of voxels (figure not shown), we suspect partially because of our method's inability to model different noise covariances for each voxel. Alternatively, it may exploit the connection between RSA and multi-task regression apparent in the matrix-variate formalization to bring techniques from multi-task regression to this latent covariance estimation problem. Likewise, while MN-SRM performed better at reconstruction than SRM, it did not produce features that improved classification performance. A broader exploration of noise models may help here, but we suspect that the true next gain may come from using the matrix-variate view to bring SRM and RSA even closer together into a unified formalism. Regardless, our toolkit will enable rapid prototyping as we progress in this domain.

References

[1] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng. {TensorFlow}: Large-Scale Machine Learning on Heterogeneous Systems, 2015.

[2] G. I. Allen and R. Tibshirani. Transposable regularized covariance models with an application to missing data imputation. *The Annals of Applied Statistics*, 4(2):764–790, jun 2010.

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

[4] E. Bonilla, K. M. Chai, and C. Williams. Multi-task Gaussian Process Prediction. *Nips*, 20 (October):153–160, 2008.

[5] M. B. Cai, N. W. Schuck, J. W. Pillow, and Y. Niv. A Bayesian method for reducing bias in neural representational similarity analysis. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 4951—-4959. Curran Associates, Inc., sep 2016.

[6] J. Chen, Y. C. Leong, C. J. Honey, C. H. Yong, K. A. Norman, and U. Hasson. Shared memories reveal shared structure in neural activity across individuals. *Nature Neuroscience*, 20(1):115–125, dec 2016.

[7] P.-H. C. Chen, J. Chen, Y. Yeshurun, U. Hasson, J. Haxby, and P. J. Ramadge. A Reduced-Dimension fMRI Shared Response Model. *Neural Information Processing Systems Conference (NIPS)*, pages 460–468, 2015.

[8] M. G. Genton. Separable approximations of space-time covariance matrices. *Environmetrics*, 18(7): 681–695, nov 2007.

[9] K. Greenewald and A. O. Hero. Robust Kronecker Product PCA for Spatio-Temporal Covariance Estimation. *IEEE Transactions on Signal Processing*, 63(23):6368–6378, dec 2015.

[10] K. Greenewald, T. Tsiligkaridis, and A. O. Hero. Kronecker sum decompositions of space-time data. In *2013 5th IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, number 2, pages 65–68. IEEE, dec 2013.

[11] K. H. Greenewald and A. O. Hero. Kronecker PCA based spatio-temporal modeling of video for dismount classification. page 90930V, jun 2014.

[12] N. V. Hartvig. A Stochastic Geometry Model for Functional Magnetic Resonance Images. *Scandinavian Journal of Statistics*, 29(3):333–353, 2002.

[13] J. V. Haxby, J. S. Guntupalli, A. C. Connolly, Y. O. Halchenko, B. R. Conroy, M. I. Gobbini, M. Hanke, and P. J. Ramadge. A Common, High-Dimensional Model of the Representational Space in Human Ventral Temporal Cortex. *Neuron*, 72 (2):404–416, oct 2011.

[14] N. Kriegeskorte. Representational similarity analysis connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2 (November):4, 2008.

[15] N. D. Lawrence. Probabilistic non-linear Principal Component Analysis with Gaussian Process Latent Variable Models. *Journal of Machine Learning Research*, 6:1783–1816, 2005.

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

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

[18] B. Rakitsch, C. Lippert, K. Borgwardt, and O. Stegle. It is all in the noise: Efficient multi-task Gaussian process inference with structured residuals. *Advances in Neural Information Processing Systems*, pages 1466–1474, 2013.

[19] B. Roś, F. Bijma, M. de Gunst, and J. de Munck. A three domain covariance framework for EEG/MEG data. *NeuroImage*, 119:305–315, oct 2014.

[20] Y. Saatchi. *Scalable inference for structured Gaussian process models*. PhD thesis, University of Cambridge, 2011.

[21] E. Simony, C. J. Honey, J. Chen, O. Lositsky, Y. Yeshurun, A. Wiesel, and U. Hasson. Dynamic reconfiguration of the default mode network during narrative comprehension. *Nature Communications*, 7(May 2015):12141, jul 2016.

[22] G. Skolidis and G. Sanguinetti. Bayesian Multi-task Classification With Gaussian Process Priors. *IEEE Transactions on Neural Networks*, 22(12): 2011–2021, dec 2011.

[23] O. Stegle, C. Lippert, J. Mooij, N. D. Lawrence, and K. Borgwardt. Efficient inference in matrix-variate Gaussian models with iid observation noise. *Advances in Neural Information Processing Systems 24 (NIPS 2011)*, pages 630–638, 2011.

[24] T. Tsiligkaridis and A. O. Hero. Covariance Estimation in High Dimensions Via Kronecker Product Expansions. *IEEE Transactions on Signal Processing*, 61(21):5347–5360, nov 2013.

[25] B. M. Turner, C. A. Rodriguez, T. M. Norcia, S. M. McClure, and M. Steyvers. Why more is better: Simultaneous modeling of EEG, fMRI, and behavioral data. *NeuroImage*, 128:96–115, mar 2016.

[26] D. Van Essen, K. Ugurbil, E. Auerbach, D. Barch, T. Behrens, R. Bucholz, A. Chang, L. Chen, M. Corbetta, S. Curtiss, S. Della Penna, D. Feinberg, M. Glasser, N. Harel, A. Heath, L. Larson-Prior, D. Marcus, G. Michalareas, S. Moeller, R. Oostenveld, S. Petersen, F. Prior, B. Schlaggar, S. Smith, A. Snyder, J. Xu, and E. Yacoub. The Human Connectome Project: A data acquisition perspective. *NeuroImage*, 62(4):2222–2231, oct 2012.

[27] K. Werner and M. Jansson. Estimation of kronecker structured channel covariances using training data. *European Signal Processing Conference*, (Eusipco):1201–1205, 2007.

[28] K. Werner, M. Jansson, and P. Stoica. On Estimation of Covariance Matrices With Kronecker Product Structure. *IEEE Transactions on Signal Processing*, 56(2):478–491, feb 2008.

[29] D. L. K. Yamins, H. Hong, and C. Cadieu. Hierarchical Modular Optimization of Convolutional Networks Achieves Representations Similar to Macaque IT and Human Ventral Stream. *Advances in Neural Information Processing Systems*, (October):1–9, 2013.

[30] J. Yin and H. Li. Model selection and estimation in the matrix normal graphical model. *Journal of Multivariate Analysis*, 107:119–140, may 2012.

[31] K. Yu, J. Lafferty, S. Zhu, and Y. Gong. Large-scale collaborative prediction using a nonparametric random effects model. In *Proceedings of the 26th Annual International Conference on Machine Learning - ICML '09*, pages 1–8, New York, New York, USA, 2009. ACM Press.