Sparse plus low-rank graphical models of time series to infer functional connectivity from MEG recordings

Rahul Nadkarni Nicholas J. Foti Adrian KC Lee Emily B. Fox

Abstract

A fundamental problem in neuroscience is inferring the functional connections underlying cognitive behaviors such as vision, speech, and audition. Magnetoencephalography (MEG) has become a popular neuroimaging technique for studying functional connectivity; however, unlike typical approaches, we treat the MEG signals as time series rather than i.i.d. observations. We represent the functional connectivity network through conditional independence statements between the cortex-localized signals, encoded via a graphical model of time series. Importantly, we incorporate a low-rank component that accounts for latent signals that would otherwise lead to inferring spurious connections. We develop an ADMM algorithm to learn the model parameters, and evaluate the model on synthetic data as well as real MEG data collected from an auditory attention task.

1 Introduction

Inferring interactions between brain regions underlying cognitive behaviors is a central problem in neuroscience. Such brain connectivity is thought to underlie many disorders such as autism and a basic understanding of this connectivity could potentially lead to effective diagnosis or new treatments. In this work we address the challenge of inferring connections in the brain from noisy, high-dimensional neuroimaging recordings.

In particular, our focus is on deciphering so called *functional connectivity* underlying auditory attention from magnetoencephalography (MEG) data. MEG recordings consists of measurements of the small fluctuations in the magnetic field given off by a subject's brain while performing a cognitive task. MEG offers a unique microscope to study brain interactions since it provides excellent temporal resolution along with good spatial resolution. This is in contrast to functional magnetic resonance imaging (fMRI) that measures blood oxygenation levels and so has limited temporal (but excellent spatial) resolution. Thus, MEG allows us to infer brain interactions that are impossible to discover from fMRI data.

The neuroscience community distinguishes functional connectivity from two other notions of connectivity. *Anatomical connectivity* refers to the physical neuronal connections in the brain. Alternatively, *effective connectivity* refers to particular brain regions driving other regions during cognitive operation. The causal nature of effective connectivity means that it is always directed, and is extremely challenging to learn except in specific situations. On the other hand, functional connectivity describes the temporal correlation between two brain regions. We note that functional connectivity measures can either be directed or undirected [1], however we consider the undirected definition of functional connectivity in this work.

Specifically, functional connectivity is based on temporal correlations between signals in two different brain regions when the linear effects of all other regions have been removed [2]. This definition is precisely that of *conditional independence* under a Gaussian likelihood. Such conditional independencies are equivalently captured by zeros in the precision (i.e., inverse covariance) matrix, which encodes the structure in a Gaussian graphical model.

Learning Gaussian graphical models (i.e., sparse precision matrices) from high-dimensional data is a broadly popular tool in machine learning and statistics to discern the underlying conditional independence structure between random variables. Beyond providing a natural definition of functional connectivity in neuroscience [3], inferring such structure has far reaching applications, such as allowing us to quantify risk between financial instruments [4]. However, most existing approaches to learning Gaussian graphical models assume that the observations are independent and identically distributed (i.i.d.) Gaussian random vectors. In many applications of interest, such as our MEG recordings, the data represent a multivariate time series. Simply treating the observations as i.i.d. across time (cf., [5]) ignores important information contained in the dynamics that we might want to account for when assessing conditional independencies.

Recently, there has been work to try to learn graphs of Gaussian stationary time series [6, 7, 8, 9]. Most of this work builds on an elegant result of Dahlhaus [10], which states that zeros at all frequencies in the *inverse spectral density* matrices characterize conditional independencies between time series [10]. This represents a direct analog of the standard sparse precision result for Gaussian i.i.d. random variables. Focusing on likelihood-based approaches to structure learning, recent work has considered transforming the time series into the frequency domain and specifying Bayesian priors on the inverse spectral density matrices [9], or a penalized likelihood approach using a group-lasso penalty [8]. The latter builds on the common ℓ_1-based *graphical lasso (glasso)* approach for Gaussian graphical model structure learning [11], but with a group structure to capture shared zero patterns across frequencies in the spectral domain. Interestingly, the frequency domain representation, which enables straightforward encoding of conditional independencies between time series, is also commonly used in studies of electrophysiological data, making it natural for the neuroscience community. Furthermore, connectivity studies are often focused on particular frequency bands, which is straightforward to handle in this framework.

Although one expects task-related brain networks to be sparse, absent idealized experimental conditions, the MEG signals are often corrupted by activity from other sources on the cortex and by imaging artifacts. These confounding factors can lead to spurious edges being learned. Due to costly experiments for verifying networks, such spurious edges are undesirable. Likewise, densely connected graphs are challenging to interpret. To address this challenge, we extend the frequency-domain-based graphs of time series framework to account for latent processes. The goal is to separate the corrupted brain activity into an interpretable sparse component and confounding dense connections.

Accounting for latent processes has an elegant formulation in the i.i.d. setting through a *sparse plus low-rank* decomposition of the inverse covariance matrix for the observed variables [5]. A challenge arises when applying this approach to graphs of time series since the zeros have to be consistent across frequencies. We use a penalty that encourages our inverse spectral density matrices of the observed process to decompose into sparse and low rank components per frequency, while having shared sparsity structure across frequencies.

The modification to the objective functions of existing methods—either sparse plus low-rank for i.i.d. data [5] or spectral graphical models for time series [8]—is straightforwardly specified. However, existing algorithms to solve such sparse plus low-rank models, like `logdetPPA` [12], do not scale to the problems we consider due to the rapidly increasing number of parameters with frequencies analyzed. To address this computational issue, we develop an efficient ADMM algorithm that can be parallelized over frequencies.

We apply the methods to synthetic data and real MEG data collected during an experiment on auditory attention, and demonstrate that the latent component of our specification aids in producing more interpretable connectivity structures.

2 Background

Gaussian Graphical Models Let $X \in \mathbb{R}^p$ be a random variable distributed according to a multivariate Gaussian distribution, $X \sim \mathrm{N}(0, \Sigma)$, and let $G = (V, E)$ be a graph with vertices $V = \{1, \ldots, p\}$ and edge set $E \subset \{(i, j) \in V \times V : i \neq j\}$. When $\Sigma^{-1}_{ij} \neq 0$ for all pairs $(i, j) \in E$, we say that X respects the graph G and that X follows the *Gaussian graphical model* specified by the precision matrix Σ^{-1} [13].

Given a sample covariance estimate $\hat{\Sigma}$ based on a set of observations $X_1, \ldots, X_N$, a common approach to inferring a Gaussian graphical model is the *graphical lasso* [11] given by the convex program:

$$\operatorname*{arg\,min}_{\Omega \in \mathbb{S}^p_{++}} - \log\det\Omega + \operatorname{tr}\{\Omega\hat{\Sigma}\} + \lambda \left|\left|\Omega\right|\right|_1 , \tag{1}$$

where $\mathbb{S}^p_{++}$ is the cone of $p \times p$ symmetric positive-definite matrices and $||\Omega||_1 = \sum_{i<j} |\Omega_{ij}|$ is the 1-norm of the matrix encouraging a sparse solution for Ω.

Graphical Models of Stationary Time Series Let $X_t = (X_{1t}, X_{2t}, \ldots, X_{pt})^T \in \mathbb{R}^p$ for $t \in \mathbb{Z}$ be a mean-zero, stationary multivariate Gaussian time series so that

$$\mathbb{E}\left[X_t, X_{t+h}\right] = \Gamma(h), \ \forall t \in \mathbb{Z} \tag{2}$$

where $\Gamma(h)$ is the *autocovariance function* at lag h, a $p \times p$ symmetric positive definite matrix for all $h \in \mathbb{Z}$. Under stationarity, we have $\sum_{-\infty}^{\infty} ||\Gamma(h)||_2 < \infty$, where $||\cdot||_2$ is the spectral norm. This implies that the *spectral density* of the time series is given by the discrete Fourier transform of the autocovariance function:

$$S(\omega) = \frac{1}{2\pi} \sum_{h=-\infty}^{\infty} \Gamma(h) e^{-ih\omega} \tag{3}$$

for $\omega \in [0, 2\pi]$ and $S(\omega) \in \mathbb{C}^{p \times p}$ a $p \times p$ Hermitian positive definite matrix. The spectral density matrix is a key quantity for analyzing stationary time series as it compactly describes both the inter- and intra-series dynamics in the frequency domain. In particular, let d_k denote the Fourier coefficients of x_t (a realization of X_t) at frequency $\omega_k = 2\pi k/T$ for $k \in \{0, \dots, T-1\}$,

$$d_k = \frac{1}{\sqrt{T}} \sum_{t=0}^{T-1} x_t e^{-i\omega_k t}. \tag{4}$$

One can then show that d_k is approximately distributed according to a complex-valued Gaussian distribution with mean zero and complex-covariance $S(\omega_k)$, i.e. the spectral density matrix at frequency ω_k [2].

The idea of a Gaussian graphical model is naturally extended to Gaussian stationary time series: Instead of conditional independencies being encoded as zeros in the inverse covariance matrix, such statements are encoded as zeros in the inverse spectral density [10]. Specifically, if $S(\omega)^{-1}_{ij} = 0$ for all $\omega \in [0, 2\pi]$ then X_i and X_j are independent conditioned on the entire trajectories of the other series.

One approach to learning such structure is through a penalized likelihood approach. In the frequency domain, the likelihood of the time series can be efficiently computed using the *Whittle likelihood* approximation [2]. Specifically, the log-likelihood is given by

$$\begin{aligned} &\log p(X_1, \dots, X_T) \approx \\ &-\frac{1}{2} \sum_{k=0}^{T-1} \Big[\log |S(\omega_k)| + \mathrm{tr}[S(\omega_k)^{-1} \hat{S}(\omega_k)] \Big] - \frac{Tp}{2} \log 2\pi. \end{aligned} \tag{5}$$

Here, $\hat{S}(\omega_k)$ is the spectral density estimate at frequency $\omega_k = 2\pi k/T$. A commonly used sample-based estimate of the spectral density matrix is given by the *periodogram*

$$I(\omega_k) = \frac{1}{2\pi} d_k d_k^*, \tag{6}$$

for d_k as in Eq. (4). It is well known, however, that the periodogram does not provide a consistent estimator of the spectral density [2]. For consistency, it is common to instead smooth the periodogram. We use a *multitaper* approach described in the Supplement.

In the context of the Whittle likelihood approximation, it is natural to apply a *group-lasso* penalty [14], extending the vanilla graphical lasso approach to capture shared zero patterns across frequencies. This idea was recently explored in [8]. Let $\mathbf{\Psi}[\cdot] = (\Psi[1], \dots, \Psi[F]) : \Psi[f] \succ \mathbf{0}, \forall f \in [F]$ be a sequence of $p \times p$ Hermitian positive definite matrices. Also, let $\hat{S}[f] = \hat{S}(\omega_f) \in \mathbb{C}^{p \times p}$ for $f = 1, \dots, F$. The goal is to solve the following convex optimization problem:

$$\arg\min_{\mathbf{\Psi}[\cdot]} \sum_{f=1}^{F} -\log |\Psi[f]| + \Big\langle \Psi[f], \hat{S}[f] \Big\rangle + \lambda \left|\left|\mathbf{\Psi}[\cdot]\right|\right|_1, \tag{7}$$

where $\langle A, B \rangle = \mathrm{tr}\{AB^*\}$ with B^* indicating element-wise complex-conjugation, $\Psi[f] := S(\omega_f)^{-1}$, and $\left|\left|\mathbf{\Psi}[\cdot]\right|\right|_1 = \sum_{i<j} \sqrt{\sum_{f=1}^{F} \Psi[f]_{ij}}$ is the group-lasso penalty. The first two terms in Eq. (7) are the Whittle likelihood of the data represented by the smoothed periodogram, $\hat{S}[f]$. The group-lasso term, $\lambda \left|\left|\mathbf{\Psi}[\cdot]\right|\right|_1$, encourages zeros to be shared across frequencies. We refer to the above method as Sglasso for *spectral* glasso.

We note that moving to the frequency domain is doubly beneficial. First, the Whittle likelihood requires computing T inverses of $p \times p$ matrices in contrast to the naive method of inverting a $Tp \times Tp$ matrix in the time domain likelihood computation. Additionally, via [10], the structure learning problem is straightforward to define in the frequency domain.

3 Graphs of Time Series with Latent Structure

The formulation of Eq. (7) assumes that all series of interest have been observed. Often in practice this is not the case and interactions between pairs of unrelated observed series can be inferred due to an unobserved, or latent, series. For i.i.d. observations, a sparse plus low-rank decomposition of the underlying precision matrix has proven useful to discern the connections between observed variables from those arising from an unobserved variable [5]. Informally, to disentangle these signals, the latent variables must each connect to many of the sparsely connected observed variables. These ideas have been used to learn sparse autoregressive processes in the presence of latent variables in the time domain [15].

Let $x_t = [y_t, u_t]^T \in \mathbb{R}^{p+r}$ be a complete stationary time series at times $t \in \{0, \ldots, T-1\}$ where $y_t \in \mathbb{R}^p$ are the observed components and $u_t \in \mathbb{R}^r$ are the latent components. Throughout, we assume that $r \ll p$. Using this decomposition of x_t, we can write the inverse spectral density matrix of x_t at frequency ω as

$$S(\omega)^{-1} := \Psi(\omega) = \left[\begin{array}{cc} \Psi_{YY}(\omega) & \Psi_{YU}(\omega) \\ \Psi_{UY}(\omega) & \Psi_{UU}(\omega) \end{array} \right]. \tag{8}$$

The marginal inverse spectral density of the observed time series y_t is then

$$\Psi_Y(\omega) = \underbrace{\Psi_{YY}(\omega)}_{\text{sparse}} - \underbrace{\Psi_{YU}(\omega)\Psi_{UU}(\omega)^{-1}\Psi_{UY}(\omega)}_{\text{rank = } r \ll p}. \tag{9}$$

Following [5, 16], we make the assumption that $\Psi_{YY}(\omega)$ is *sparse*, as in a graphical model of fully observed series, and $\Psi_{YU}(\omega)\Psi_{UU}(\omega)^{-1}\Psi_{UY}(\omega)$ is *low-rank* due to the presence of a few latent components. From Eq. (9) we see how ignoring the contribution of the latent processes results in losing the sparse structure in $\Psi_{YY}(\omega)$ when considering $\Psi_Y(\omega)$. To disentangle the sparse and low-rank components, we propose to optimize the following adaptation of Eq. (7), which we refer to as the *LVSglasso* method:

$$\begin{aligned} \underset{\mathbf{\Psi}[\cdot], \mathbf{L}[\cdot]}{\arg\min} \sum_{f=1}^{F} & - \log |\{\Psi[f] - L[f]\}| + \left\langle \Psi[f] - L[f], \hat{S}[f] \right\rangle \\ & + \lambda_\Psi \, ||\mathbf{\Psi}[\cdot]||_1 + \lambda_L \operatorname{tr}\{\mathbf{L}[\cdot]\} \\ \text{s.t. } & \Psi[f] - L[f] \succ \mathbf{0}_{p \times p}, L[f] \succcurlyeq \mathbf{0}_{p \times p}, \ \forall f \in [F] \end{aligned} \tag{10}$$

Here, $\Psi[f]$ represents the sparse component of the marginal inverse spectral density while $L[f]$ captures the low-rank structure of $\Psi_{YU}(\omega)\Psi_{UU}(\omega)^{-1}\Psi_{UY}(\omega)$. In addition to the group-lasso penalty $||\mathbf{\Psi}[\cdot]||_1$, we add $\operatorname{tr}\{\mathbf{L}[\cdot]\} = \sum_f \operatorname{tr}\{L[f]\}$. Each trace term is a surrogate for the rank function, in our case the sum of the eigenvalues. This penalty adapts that of [5] to handle multiple frequencies, and the structure across frequencies.

In order to solve Eq. (10) for problems on the scale of MEG data, we develop a consensus alternating directions method of multipliers (ADMM) algorithm [17].

3.1 Consensus ADMM Algorithm

A commonly used algorithm for convex optimization problems like Eq. (10) that contain a smooth objective function and non-smooth penalties is ADMM. ADMM works by splitting the original problem into two problems that can be easily solved when the proximal operators of the penalties can be computed efficiently [17].

In particular, we solve Eq. (10) using a consensus ADMM algorithm [18]. First, we introduce an auxiliary variable $\mathbf{R}[\cdot] := \mathbf{\Psi}[\cdot] - \mathbf{L}[\cdot]$ which allows us to rewrite Eq. (10) as

$$\begin{aligned} \underset{\mathbf{\Psi}[\cdot], \mathbf{L}[\cdot], \mathbf{R}[\cdot]}{\arg\min} & \sum_{f=1}^{F} - \log |R[f]| + \left\langle R[f], \hat{S}[f] \right\rangle + \lambda_\Psi \, ||\mathbf{\Psi}[\cdot]||_1 \\ & + \lambda_L \sum_{f=1}^{F} \left(\operatorname{tr}\{L[f]\} + \mathbb{I}[L[f] \succ \mathbf{0}] \right) \\ & \text{s.t. } \mathbf{R}[\cdot] - \mathbf{\Psi}[\cdot] + \mathbf{L}[\cdot] = \mathbf{0}[\cdot]. \end{aligned} \tag{11}$$

Naively applying ADMM to Eq. (11) is not straight-forward since the convergence of ADMM for three blocks of variables is only guaranteed for a small class of functions [18]. Instead, we rewrite the problem in terms of two blocks of variables $Z = [\mathbf{R}[\cdot], \mathbf{\Psi}[\cdot], \mathbf{L}[\cdot]]$ and $\tilde{Z} = [\tilde{\mathbf{R}}[\cdot], \tilde{\mathbf{\Psi}}[\cdot], \tilde{\mathbf{L}}[\cdot]]$ as

$$\begin{aligned} &\arg\min_{Z,\tilde{Z}} \; g(Z) + h(\tilde{Z}), \quad \text{s.t. } Z - \tilde{Z} = [\mathbf{0}[\cdot], \mathbf{0}[\cdot], \mathbf{0}[\cdot]] \\ &g(Z) = \sum_{f=1}^{F} -\log|R[f]| + \left\langle R[f], \hat{S}[f] \right\rangle \\ &\qquad + \lambda_\Psi \, ||\mathbf{\Psi}[\cdot]||_1 + \lambda_L \sum_{f=1}^{F} \operatorname{tr}\{L[f]\} \\ &h(\tilde{Z}) = \mathbb{I}[\tilde{\mathbf{R}}[\cdot] - \tilde{\mathbf{\Psi}}[\cdot] + \tilde{\mathbf{L}}[\cdot] = \mathbf{0}[\cdot]]. \end{aligned} \tag{12}$$

See the Supplement for detailed update equations. We use a standard ADMM stopping scheme based on primal and dual residuals, $r_k = \sum_f R[f] - \Psi[f] + L[f]$ and $s_k = \tilde{Z}^{(k+1)} - \tilde{Z}^{(k)}$, respectively [17]. We monitor

$$\begin{aligned} \epsilon_{\text{pri}} &= \sqrt{p^2 F}\epsilon_{\text{abs}} + \epsilon_{\text{rel}} \sum_f \max\left(||Z||_F \, , \left|\left|\tilde{Z}\right|\right|_F \right) \\ \epsilon_{\text{dual}} &= \sqrt{p^2 F}\epsilon_{\text{abs}} + \epsilon_{\text{rel}} \sum_f ||\Lambda||_F \end{aligned}$$

and stop when $||r_k||_2 < \epsilon_{\text{pri}}$ and $||s_k||_2 < \epsilon_{\text{dual}}$.

3.2 Choosing the regularization parameters

Appropriate choice of the regularization parameters for the LVSglasso model is important, since different values can result in inferring very different graph structures due to the interaction between the sparse and low-rank terms of the optimization problem. Determining the regularization parameters for sparse plus low-rank models is an unsolved problem. In this work we use the Akaike information criteria (AIC) to score the solutions to the LVSglasso problem and use the values of the regularization parameters that result in the lowest AIC. Specifically, the AIC is defined as

$$\text{AIC} = -2N\mathcal{L} + 2Fk \tag{13}$$

where $\mathcal{L}$ is the log-Whittle likelihood from Eq. (5), N is the number of replicate series, F is the total number of frequencies in the spectral density estimate, and k is the number of edges in the graphical model. Notice that the learned rank does not explicitly appear in our model selection criteria. The reason for this is that we are only interested in selecting the edge structure that provides the optimal trade-off between explaining the data and parsimony. This approach is similar to that used in previous work learning sparse plus low rank graphical models of VAR processes [19].

A well known property of penalized likelihood approaches such as LVSglasso is that the solution ($\mathbf{\Psi}[\cdot]$, $\mathbf{L}[\cdot]$) for given regularization parameters λ_Ψ and λ_L will be shrunk towards zero. The shrunken parameter estimates lead to biased AIC computations in Eq. (13) and inaccurate model selection. To address this issue, after finding a solution to LVSglasso for given λ_Ψ and λ_L we determine the edges present in the graph, $\mathcal{G}$, and then reinflate the parameters. The reinflation is done by maximizing the Whittle likehood, Eq. (5), for $S(\omega_k)$, $k = 1, \ldots, F$, constrained so that the inverse spectral density follows $\mathcal{G}$ learned by LVSglasso. This is equivalent to finding the maximum likelihood estimate (MLE) of the spectral density with inverse following $\mathcal{G}$. The graph-constrained maximum likelihood problem is

$$\arg\min_{\mathbf{\Psi}[\cdot]} \sum_{f=1}^{F} -\log|\Psi[f]| + \left\langle \hat{S}[f], \Psi[f] \right\rangle \tag{14}$$

$$\text{s.t. } \mathbf{\Psi}[f]_{ij} = 0, \, \forall f, \, \forall (i,j) \notin \mathcal{G},$$

which we solve using projected gradient descent. We plug the MLE $\hat{\mathbf{\Psi}}[\cdot]$ of $\mathbf{\Psi}[\cdot]$ into Eq. (13) and select the graph that produces the *minimum* AIC.

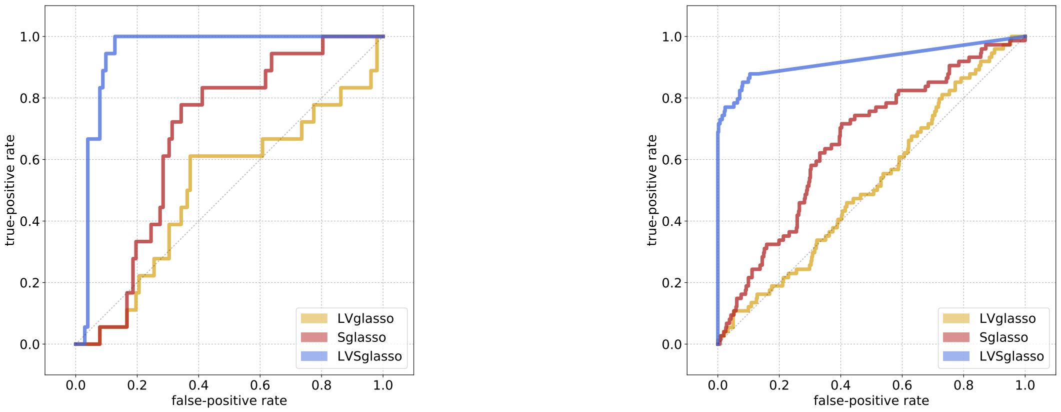

Figure 1: ROC plots comparing the ability of LVglasso, Sglasso, and LVSglasso to recover conditional independencies between the components of simulated time series with latent confounders, using $p = 16$, $r = 1$ (left) and $p = 32$, $r = 2$ (right). Accounting for both latent components and temporal dynamics improves our ability to uncover the true graph structure as compared to models that only account for one of these.

4 Related work

Utilizing the frequency domain representation, latent processes, and graphical models each have a history in the study of functional connectivity from neuroimaging data. In [20], connectivity structure was determined by estimating the spectral density matrix with a smoothed periodogram, $\hat{S}[f]$, and then numerically inverting to obtain $\hat{S}[f]^{-1}$. From the inverse spectral density matrices the authors determined the partial coherence matrices $P(f) = D^{-\frac{1}{2}}\hat{S}[f]^{-1}D^{-\frac{1}{2}}$, where $D = \mathrm{diag}(\hat{S}[f])$. To determine whether an edge is present, the authors threshold the maximum partial coherences over all frequencies, i.e. there is an edge between series i and j if $\max_f |P_{ij}(f)| > \epsilon$ for some $\epsilon > 0$. The authors found that the discovered networks reproduce existing knowledge about functional connections. We utilize the same procedure in our experiments below, except replacing $\hat{S}[f]^{-1}$ with the estimates obtained from LVSglasso.

Although not applied in the frequency domain, a similar group penalty to Eq. (7) was developed to account for inter-subject variability in the graphical lasso for fMRI data, treating the data as i.i.d. [21]. A convex formulation to learn Gaussian graphical models specifically for vector autoregressive (VAR) processes was developed and applied to fMRI data in [7]. This approach was then extended by incorporating a low-rank component to account for unobserved effects in neuroimaging data, however the method is still only applicable to the restrictive class of VAR processes [19].

We also distinguish the current work from previous work on inferring connectivity in neuroscience. The *dynamic causal model* (DCM) is one of the most widely known methods in neuroscience to learn effective connectivity [22]. The DCM assumes fixed dynamics and perturbs the inputs to the system to discern how regions drive one another. The structure of the dynamics are usually chosen by enumerating all possible model structures and scoring them. As such, DCM is only applicable to very small numbers of regions and latent confounders are not handled in the framework. A method based on Granger causality to infer effective connectivity that accounts for latent confounders was proposed in [23] and was used to analyze fMRI observations. We do not compare against these methods in this work as our goal is not to make any claims of effective connectivity since the neuroscience community is still exploring the mechanistic aspects related to the MEG data we analyze.

We leverage the flexibility to capture a broad family of dynamic processes, the computational efficiency, and the convenient encoding of conditional independence that working in the frequency domain entails. We likewise make important computational strides that enable new scaling of the technique to larger numbers of brain regions and the number of frequencies.

5 Experiments

5.1 Synthetic Data

Accounting for latent confounders We first evaluate our proposed sparse plus low-rank LVSglasso method on a synthetic data set with known graphical model structure that includes spurious connections introduced via the presence of latent processes. We do so by constructing a synthetic inverse spectral density matrix that is the sum of a sparse component with known graph structure and a low-rank component that models marginalization over latent processes.

In order to demonstrate the performance of our model on datasets of varying size and latent dimensionalities, we generate synthetic data with an observed dimensionality of $p = 16$ and a latent dimensionality of $r = 1$, as well as a dataset with observed dimensionality $p = 32$ and latent dimensionality $r = 2$.

To construct our synthetic inverse spectral density, we first construct a sparse matrix $A \in \mathbb{C}^{p \times p}$ and a low-rank matrix $B \in \mathbb{C}^{p \times p}$ to serve as the base matrices for our inverse spectral density. A is constructed by selecting a random 15% of its entries and setting them to the complex value $a + b\,i$, where a and b are set to one of $\{-0.5, 0.5\}$ independently and with equal probability. A is then adjusted by adding a multiple of the identity to ensure that it is positive-definite. The entries of B are drawn as $B_{ij} \sim \mathcal{N}(0, 1)$, and B is made low-rank and positive semidefinite by taking the absolute value of its eigenvalues and retaining only the top r eigenvalues. Finally, we scale B by 0.75, and again add a multiple of the identity to A to ensure that the difference $A - B$ is also positive-definite. We construct the sparse and low-rank components of the inverse spectral density matrix for frequencies $f = \{1, \ldots, F\}$ as

$$S[f] = A \cdot x[f] \cdot \exp\left(-0.5x[f]\right)$$
$$L[f] = B \cdot x[f] \cdot \exp\left(-0.5x[f]\right),$$

where $x[\cdot]$ is the sequence of F real values spaced evenly between 1 and 8. This is modeled after the structure of decaying power over frequencies that can be observed in the inverse spectral density matrix of a VAR(1) process. Our synthetic spectral density matrix is then constructed by computing $\Psi[f] = (S[f] - L[f])^{-1}$ at each frequency, which we then use to simulate 250 realizations x_t, where $t = \{1, \ldots, 256\}$. From these series, we compute an estimate of the spectral density matrix using the multitaper method [24], and decimate down to 8 frequencies to speed up computations.

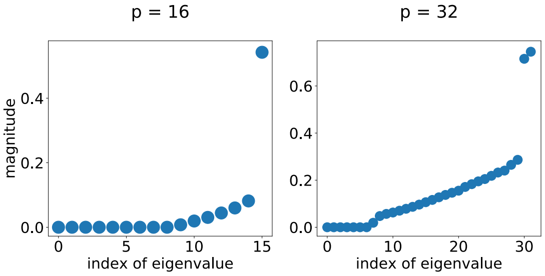

Figure 2: Eigenvalues in increasing order of magnitude for the low-rank component of the minimum-AIC solution from the synthetic experiments. Shown are the eigenvalues for the $p = 16$, $r = 1$ simulated dataset (left) and the $p = 32$, $r = 2$ simulated dataset (right), with error bars indicating standard deviation of magnitude across frequency (error bars too small to see). These plots indicate that while the learned latent components are numerically high-rank, they are well-approximated by a low-rank solution that has the true latent dimensionality of $r = 1$ (left) and $r = 2$ (right).

We run Sglasso and our new LVSglasso model on the estimated spectral density matrix. We also compare both of these models to one that we term *LVglasso*, which models a latent component but assumes that data points are i.i.d., corresponding to the model proposed in [5]. We report the ROC curves for each of these three models in Fig. 1. The ROC results indicate that Sglasso's incorporation of the dynamics allows it to learn fewer false positives than LVglasso. However, we see that LVSglasso is able to outperform both of the other models by incorporating temporal dynamics *and* latent variables.

Our empirical evaluations show that typical solutions learned by LVSglasso will have a latent component that is numerically high-rank. However, we find upon further examination that these solutions have a number of dominant eigenvalues corresponding to the true latent dimensionality. These results can be observed in Fig. 2, where we see that there is one dominant eigenvalue in the case of observed dimensionality $p = 16$ and two dominant eigenvalues for the

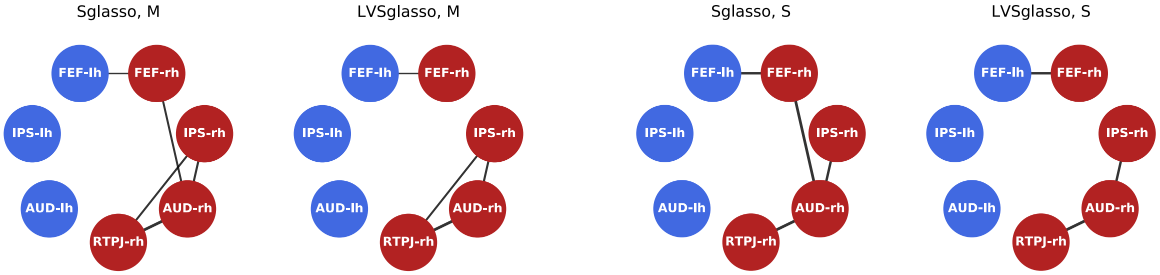

Figure 3: Graphs of connections learned by Sglasso and LVSglasso for 7 regions of interest in an auditory attention experiment, examining the M (left) and S (right) experimental conditions. Sglasso learns an additional edge that is likely to be spurious in each condition, whereas accounting for latent interactions allows LVSglasso to remove this edge while still retaining the scientifically relevant connections. Both methods are able to differentiate between the two conditions by the edges they learn.

$p = 32$ dataset, corresponding to the true latent dimensionalities of $r = 1$ and $r = 2$, respectively. The fact that the solutions are numerically high-rank while being statistically low-rank is likely a result of attempting to model a noisy realization of the true underlying process.

5.2 Auditory Attention MEG Recordings

Magnetoencephalography (MEG) uses 306 sensors (accelerometers and gradiometers) at 102 locations around the head to measure the small deviations in the magnetic field produced by the brain [25]. The signal at each of these sensors captures the activity from many neurons which can be distributed across the cortical surface. Thus, an imperative step in analyzing functional connectivity is to map these *sensor-space* recordings to *source-space*, i.e. map the recordings of the magnetic field to signals of neural activity on the cortical surface itself. This problem is known as the *inverse-problem* and a variety of methods have been proposed to solve it [25]. In this work, we use the *minimum-norm estimate* (MNE) which uses a penalized regression approach to solve the ill-posed inverse-problem that maps the 306-dimensional sensor-space signals to a high-resolution tesselation of the cortical surface with approximately 10,000 locations [26].

We apply LVSglasso and Sglasso to real MEG recordings collected from 15 individuals while they performed an auditory attention task. Each subject was presented with an auditory stream and was instructed to either maintain attention to that stream or switch to a different stream after presented with an additional auditory cue. The two stimuli had different spatial profiles, specifically one stream came from the left and the other from the right. This resulted in two experimental conditions to analyze, *maintain* (M) vs. *switch* (S).

For each subject, the data consists of time series of length 1,000 at each of 10,000 locations on the cortical surface after the MNE inverse mapping. Despite the large number of cortical surface locations for which time series data is present, a typical study of this kind will look at a small subset of hand-selected regions of interest (ROIs) that are assumed to be related to the task, aggregating all series within each ROI. However, by only considering a specific set of ROIs, we run the risk of finding spurious connections between these ROIs as a result of excluding all other regions of activity from the analysis. In such cases, we expect the latent component in our model to help address this issue by removing some of these spurious edges.

We consider two setups for analyzing this MEG data, one using a small number of regions and the second with a larger number of regions. First, we perform the analysis using 7 predefined ROIs that neuroscientists think are important to this auditory attention task and average the signals within each region based on an existing parcellation. We also perform a similar analysis where we augment the 7 ROIs suggested by neuroscientists with 20 additional regions from a standard parcellation that exhibit high signal variance. The purpose of this additional analysis is to discover potentially interesting interactions beyond those in the small subset of pre-determined ROIs. In both cases, the regions we consider do not cover the entire cortical surface, so we expect possible spurious edges caused by unobserved regions.

Small ROI Analysis For the small number of ROIs, the data set consists of 850 7-dimensional time series over all subjects for the M condition and 550 series over all subjects for the S condition. The stimuli patterns per condition are repeatedly presented to the subject, representing a single trial. For each presentation, the data are marked with a *stimulus*

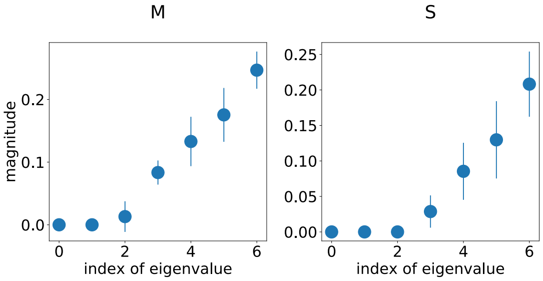

Figure 4: Eigenvalues in increasing order of magnitude for the latent component of the solutions found by AIC in the small ROI analysis. Results are presented for both the M (left) and S (right) experimental conditions in the auditory attention task. Error bars indicate standard deviation in magnitude across frequencies. The plots show an effective learned rank of ∼4 for each solution.

period and *gap* period. For each trial, we extract the short time segment directly following the stimulus of interest. We then discard the stimulus period to allow transient brain activity to subside, and only retain the so-called gap period (approximately 575 milliseconds each). Although the MEG signal is clearly non-stationary, it is reasonable to assume that it is approximately stationary over such small, homogeneous segments (i.e., the signal is *locally stationary*).

We compute an estimate of the spectral density matrix from each trial using an adaptive multitaper method [24]. The MEG signals were low-pass filtered at about 60 Hz as a pre-processing step (prior to us obtaining them) so we truncate the spectral density estimate to ignore higher frequencies with very low power, considering frequencies up to about 60 Hz. We further attempt to reduce noise in the estimate by applying a Daniell smoother with a bandwidth of 20 Hz. The smoothed spectral density estimate is then decimated down to 16 frequencies and fed to Sglasso and LVSglasso, with the optimal solutions for each model chosen using our model selection procedure described in Sec. 3.2.

We plot the graphs learned by both Sglasso and LVSglasso on the set of 7 ROIs in Fig. 3. The cycle formed by IPS-rh, AUD-rh, and RTPJ-rh is interesting to neuroscientists as these are three regions that are known to be active while employing auditory attention and LVSglasso picks up evidence in the data that they are functionally connected.

The edge between FEF-lh and FEF-rh, the frontal eye fields, is most likely spurious as the regions are close on the medial wall making them hard to disambiguate. Their activity during a trial is likely due to the visual indicators during the experiment, and the difficulty in differentiating these regions is expected to lead to a spurious connection for both experimental conditions and models, as we observe in Fig 3. However, we note that Sglasso picks up an additional edge between FEF-rh and AUD-rh. Though this edge is also likely to be spurious as a result of visual indicators, it could mistakenly be marked as an interesting edge, since it connects regions that are not spatially co-located. By modeling the effect of interactions between latent and observed regions, LVSglasso is able to prune out this spurious connection. The ability of LVSglasso to remove this connection is particularly significant when analyzing neuroimaging data, since each false positive is another potential hypothesis that needs to be scrutinized and may lead to expensive and time-consuming future experiments.

When comparing the two experimental conditions, we also observe an edge between IPS-rh and RTPJ-rh that both models learn in the M condition but is absent in the S condition. This indicates a connection that may be indicative of differences in neural processing when switching vs. maintaining spatial auditory attention. This edge is therefore a potential connection of interest for neuroscientists studying auditory processing in the brain.

In addition to presenting the graphs of the learned sparse component for each model, we also present the eigenvalues of the latent components learned by LVSglasso in Fig. 4. The eigenvalue plots indicate that ∼4 underlying latent processes were found in the recordings for the 7 ROIs being analyzed. While this may seem like a high rank to learn for an observed dimensionality of 7 regions, the MEG recordings are both noisy and very complex, with an underlying structure that could require a rank-4 latent component in order to fully explain the data.

An interesting result of this small-scale analysis is that while LVSglasso demonstrates that there is a latent component of rank 4, both Sglasso and LVSglasso give solutions of comparable sparsity. What this likely indicates is that for these particular ROIs, most spurious interactions induced by interactions with other regions are weak, such that Sglasso is able to prune most of them just as well as LVSglasso. This is unsurprising in this setup, as we have restricted our analysis to regions that are deemed to be of significance to the auditory attention task, and will therefore exhibit the strongest interactions with each other relative to interactions with other regions. However, this may not be the case in a more exploratory analysis where someone is interested in arbitrary regions that may not be directly related to the experiment. In this setting, modeling the effects of an unobserved latent component is crucial to removing spurious connections between the regions being studied.

Large ROI Analysis We also perform a large ROI analysis, using the 7 ROIs mentioned previously as well as 20 additional ROIs exhibiting the highest signal variance in the time segment of interest. The resulting graphs identified by our model selection criteria result in a sparser solution for LVSglasso (125 edges) as compared to Sglasso (146 edges). However, further analysis is necessary to determine the scientific validity of the edges learned by both methods.

6 Discussion

We have presented a method to learn graphical models of time series that capture the conditional independence structure between Gaussian stationary time series while accounting for latent processes. In doing so, we combine modern techniques from machine learning and statistics with fundamental ideas of analyzing time series in the frequency domain. This confluence of (i) parsimoniously modeling the underlying dynamics in the spectral domain and (ii) the computational advances of structure learning as convex optimization provides a practical framework for learning graphs of temporal data in a wide variety of applications. Since graphical models of time series provide a natural characterization of *functional connectivity*, we focus on the application of learning networks of interactions between brain regions from MEG recordings. In doing so, we demonstrate the ability of our proposed models to learn interesting functional interactions underlying auditory attention from MEG data, with the latent component adding additional robustness. However, this same analysis could be just as useful applied to data in other domains, such as high-dimensional financial series, recordings from large-scale sensor networks, or long-term measurements from patients in a medical setting.

There are still many avenues of research, including statistical, computational, and in neuroscience, to further develop our approach. For instance, selecting the hyperparameters in sparse plus low-rank models (both for i.i.d. and time-series data) is an open problem that could be approached by formally developing (extended) information criteria for the models. Additionally, the statistical properties of consistency and sample complexity in various scaling regimes of our sparse plus low-rank model need to be determined.

In terms of the applications to neuroscience, there are other interesting phenomena related to functional connectivity and frequency-domain analysis that are of interest to researchers in that field and could be addressed with extentions to our model. One of these is the notion of *dynamic functional connectivity*, the idea that functional connections between regions can vary over time. Though our model captures functional connections that exist throughout a particular time period of interest in the data, extensions to allow for graphs that vary over time could be made to model dynamic connectivity over longer time ranges. The neuroscience community also has interest in properties of connectivity that are specific to particular frequency bands of interest, and while the MEG analysis in this work focused on broadband connectivity, an examination of band-specific connectivity could be done with the methods we have developed.

References

[1] B. Horwitz. The elusive concept of brain connectivity. *Neuroimage*, 19(2):466–470, 2003.

[2] D. R. Brillinger. *Time Series: Data Analysis an Theory*. Holden-Day, 1981.

[3] T. Medkour, A. T. Walden, and A. Burgess. Graphical modelling for brain connectivity via partial coherence. *Journal of Neuroscience Methods*, 180(2):374–383, 2009.

[4] H. Liu, F. Han, and C.-h. Zhang. Transelliptical graphical models. In *Advances in Neural Information Processing Systems*, 2012.

[5] V. Chandrasekaran, P. A. Parrilo, and A. S. Willsky. Latent variable graphical model selection via convex optimization. *The Annals of Statistics*, 40(4):1935–1967, 2012.

[6] F. R. Bach and M. I. Jordan. Learning graphical models for stationary time series. *IEEE Trans. Signal Process.*, 52(8):2189–2199, 2004.

[7] J. Songsiri and L. Vandenberghe. Topology selection in graphical models of autoregressive processes. *Journal of Machine Learning Research*, 11:2671–2705, 2010.

[8] A. Jung, G. Hannak, and N. Goertz. Graphical LASSO based model selection for time series. *IEEE Signal Processing Letters*, 22(10):1781–1785, 2015.

[9] A. Tank, N. J. Foti, and E. B. Fox. Bayesian structure learning for stationary time series. In *Uncertainty in Artificial Intelligence*, 2015.

[10] R. Dahlhaus. Graphical interaction models for multivariate time series. *Metrika*, 51(2):157–172, 2000.

[11] J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

[12] C. Wang, D. Sun, and K-C Toh. Solving log-determinant optimization problems by a Newton-CG primal proximal point algorithm. *SIAM J. Optimization*, 20:2994–3013, 2010.

[13] S. Lauritzen. *Graphical Models*. Oxford University Press, 1999.

[14] P. Danaher, P. Wang, and D. M. Witten. The joint graphical lasso for inverse covariance estimation across multiple classes. *JRSS: Series B (Statistical Methodology)*, 76(2):373–397, 2014.

[15] A. Jalali and S. Sanghavi. Learning the dependence graph of time series with latent factors. In *International Conference on Machine Learning*, 2012.

[16] J. Dauwels, H. Yu, X. Wang, F. Vialatte, C. Latchoumane, J. Jeong, and A. Cichocki. *Machine Learning and Interpretation in Neuroimaging: International Workshop*, pages 194–201. 2012.

[17] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Found. Trends Mach. Learn.*, 3(1):1–122, 2011.

[18] S. Ma, L. Xue, and H. Zou. Alternating direction methods for latent variable Gaussian graphical model. *Neural Computation*, 25(8):2172–2198, 2013.

[19] R. Liégeois, B. Mishra, M. Zorzi, and R. Sepulchre. Sparse plus low-rank autoregressive identification in neuroimaging time series. In *54th IEEE Conference on Decision and Control*, pages 3965–3970, 2015.

[20] R. Salvador, J. Suckling, C. Schwarzbauer, and E. Bullmore. Undirected graphs of frequency-dependent functional connectivity in whole brain networks. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, 360(1457):937–946, 2005.

[21] G. Varoquaux, A. Gramfort, J.-b. Poline, and B. Thirion. Brain covariance selection: Better individual functional connectivity models using population prior. In *Advances in Neural Information Processing Systems*, pages 2334–2342. 2010.

[22] K. J. Friston, L. Harrison, and W. Penny. Dynamic causal modelling. *Neuroimage*, 19(4):1273–1302, 2003.

[23] M. Eichler. A graphical approach for evaluating effective connectivity in neural systems. *Phil. Trans. R. Soc. B*, 360:953–967, 2005.

[24] D. Thomson. Spectrum estimation and harmonic analysis. *Proceedings of the IEEE*, 70:1055–96, 1982.

[25] P. C. Hansen, M. L. Kringelbach, and R. Salmelin. *MEG: An Introduction to Methods*. Oxford University Press, 2010.

[26] MNE Python. `http://martinos.org/mne/stable/index.html`, 2010.