Bayesian Structure Learning for Stationary Time Series

Alex Tank
University of Washington
alextank@uw.edu

Nicholas J. Foti
University of Washington
nfoti@uw.edu

Emily B. Fox
University of Washington
ebfox@uw.edu

Abstract

While much work has explored probabilistic graphical models for independent data, less attention has been paid to time series. The goal in this setting is to determine conditional independence relations between entire time series, which for stationary series, are encoded by zeros in the inverse spectral density matrix. We take a Bayesian approach to structure learning, placing priors on (i) the graph structure and (ii) spectral matrices given the graph. We leverage a Whittle likelihood approximation and define a conjugate prior—the *hyper complex inverse Wishart*—on the complex-valued and graph-constrained spectral matrices. Due to conjugacy, we can analytically marginalize the spectral matrices and obtain a closed-form marginal likelihood of the time series given a graph. Importantly, our analytic marginal likelihood allows us to avoid inference of the complex spectral matrices themselves and places us back into the framework of standard (Bayesian) structure learning. In particular, combining this marginal likelihood with our graph prior leads to efficient inference of the time series graph itself, which we base on a stochastic search procedure, though any standard approach can be straightforwardly modified to our time series case. We demonstrate our methods on analyzing stock data and neuroimaging data of brain activity during various auditory tasks.

1 INTRODUCTION

Probabilistic graphical models (PGMs)—which compactly encode a set of conditional independence statements—have become a defacto tool for defining probabilistic models over large sets of random variables. When faced with time series, dynamic Bayesian networks (DBNs) are commonly deployed and specify sparse between- and within-time dependencies, often encoded by a *template model* replicated across time to straightforwardly model the growing set of random variables [1]. Learning template models requires specifying the set of dependency lags to be considered [2, 3]. In many applications, one instead aims to infer conditional independence between entire data streams accounting for interactions at all possible lags, represented by a *time series graphical model* (TGM). For example, imagine recording brain activity from multiple regions of the brain over time. Inference of a TGM in this setting would provide insight into the functional connectivity of different brain regions, an object of substantial scientific interest [4, 5]. TGMs have also been applied to intensive care monitoring [6] and financial time series [7].

The pioneering work of Dahlhaus [8] introduced the concept of undirected graphical models for stationary time series. The key insight was to transform the series to the *frequency domain* and express the graph relationships in the resulting spectral representation. For jointly Gaussian stationary time series, Dahlhaus [8] showed that conditional independencies between time series are encoded by zeros in the inverse spectral density matrices. This result is the frequency-domain analog to Gaussian graphical modeling in the i.i.d. (non-time-series) setting, where zeros in the inverse covariance matrix, or *precision matrix*, encode the conditional independencies between observed dimensions [9]. Dahlhaus' insight was first exploited to perform independent hypothesis tests of conditional independence between each pair of time series [8], with more recent work correcting for multiple comparisons [10, 11].

A likelihood-based approach leveraging the *Whittle approximation* [12] has also been introduced [13]. The Whittle approximation casts the likelihood in the frequency domain with terms depending on the spectral density matrices critical to TGM structure learning, and independently so across frequencies. One approach scores graphs using AIC [13]. A recent penalized likelihood variant [14] places a joint graphical lasso [15] across frequencies to enforce a common zero pattern in the spectral density matrices. A penalized likelihood approach restricted to finite vector autoregressive processes has also been considered [7].

We instead consider a Bayesian approach to TGM structure learning, with all the benefits garnered from the Bayesian paradigm, including modeling within a generative framework where information from multiple sources can integrated and combined with available prior knowledge. For example, neural data are notoriously noisy, and robust inferences often rely on integrating time series across multiple trials and individuals or recording platforms (e.g., EEG/MEG). Our approach also leverages the Whittle likelihood. We then introduce a novel hyper Markov law [16], the *hyper complex inverse Wishart* distribution, that serves as a conjugate prior for the spectral density matrices whose inverses have a zero pattern specified by a graph. For decomposable graphs, this formulation leads to a closed-form expression for the marginal likelihood of a multivariate time series given a graph. By placing a prior on graph structures, we achieve a fully Bayesian approach to TGM structure learning for stationary time series. For our graph prior, we consider a multiplicity correcting prior [17]. Our analytic expression for the marginal likelihood is critical to the practicality of our approach since we can avoid inference of the large set of high-dimensional, complex spectral density matrices. In particular, for a length T series of dimension p, there are T $p \times p$ spectral matrices to consider. In the i.i.d. setting, inference of just a single $p \times p$ graph-constrained covariance matrix is challenging; in this setting, inference of the T $p \times p$ matrices is prohibitive.

Hyper Markov laws based on the hyper inverse Wishart are a popular tool for Bayesian graphical model selection in the i.i.d. setting [18, 19]. Indeed, many powerful Bayesian structure learning algorithms based on this framework have been developed, both for decomposable [20, 21] and non-decomposable [22, 23] graphs. By framing TGM structure learning in this common framework, we are able to apply existing state-of-the-art inference machinery for standard structure learning to the time series case. In this paper we use the feature-inclusion stochastic search (FINCS) procedure [20] for inference in decomposable models; however, many other MCMC and search schemes may be used. Importantly, future computational advances in Bayesian inference for i.i.d. graphical models may be easily extended using our framework to the time series case.

We test our methods on data simulated from vector autoregressive models with randomly generated TGMs. Our approach reaches almost perfect TGM recovery as the length of the time series or number of independent replicates increases. We then demonstrate the utility of our methods on a global stock indices dataset and MEG neuroimaging data of auditory attention switching tasks. In both cases we find meaningful, intuitive structure in the data.

Our paper is organized as follows. We provide background on graphical models and stationary time series in Sec. 2. Our proposed TGM method is in Sec. 3, first introduced in the context of multiple independent realizations and then adapted to perform efficient inference of the TGM from only a single realization. In Sec. 5, we discuss how existing Bayesian structure learning methods may be modified to fit our formulation. Simulated results are in Sec. 6, with our stock and MEG analyses in Secs. 7 and 8, respectively.

2 BACKGROUND

2.1 Graphs

Let $G = (V, E)$ be an undirected graph with vertex set $V = \{1, \dots, p\}$ and edge set E, where $E \subset \{(i, j) \in V \times V : i \neq j\}$. Nodes i and j are adjacent, or *neighbors*, if $(i, j) \in E$. A *complete graph* is one having $(i, j) \in E$ for every $i, j \in V$ and complete subgraphs $C \subset V$ are termed *cliques*. A triple of subgraphs (A, S, B) where $V = A \cup B$ and $S = A \cap B$ with S complete is called a *decomposition* if every path from a node in A to a node in B must pass through S, the *separator*. Recursively decomposing A and B in this fashion results in the *prime components* of a graph. If the prime components are complete then the graph is *decomposable*. We let the sets $\mathcal{C} = \{C_1, \dots, C_K\}$ and $\mathcal{S} = \{S_2, \dots, S_K\}$ each denote the prime components and their separators, respectively, generated by the decomposition. For simplicity, we restrict our attention to decomposable graphs but stress that our formulation is extensible to the non-decomposable case (see Sec. 9).

2.2 Hyper Markov distributions

For a given set of of random variables X, with realization $x \in \mathcal{X}$, dimensionality p, and joint density $p(x)$, an undirected graphical model G can be constructed by stating that an edge $(i, j) \notin E$ if X_i and X_j are conditionally independent given the remaining variables, i.e. $X_j \perp\!\!\!\perp X_i | X_{Z_{ij}}$ where $Z_{ij} = V \setminus \{i, j\}$. If the graph is decomposable, the joint density decomposes over cliques and separators:

$$p(x) = \frac{\prod_{C \in \mathcal{C}} p(x_C)}{\prod_{S \in \mathcal{S}} p(x_S)} \tag{1}$$

where $p(x_A)$ for $A \subset V$ denotes the marginal distribution of the set of variables x_A.

A hyper Markov law [16] is a distribution over probability measures that is concentrated on distributions that obey the Markov properties specified by G. Examples include the hyper Wishart and hyper Dirichlet distribution [16, 18]. Such distributions have proven pivotal in Bayesian graphical modeling by serving as conjugate priors for the graph parameters conditioned on the graph structure G. For example, in Gaussian graphical models (GGMs), the hyper inverse Wishart distribution provides a conjugate prior for covariance matrices that obey a zero pattern in the precision, as specified by G. By integrating over the hyper

Markov distribution, one can obtain the *marginal likelihood* of the data conditioned on the structure G alone.

2.3 Stationary time series

Let $X(t) = (X_1(t), ..., X_p(t))^T \in \mathbb{R}^p$ for $t \in \mathbb{Z}$ be a multivariate Gaussian stationary time series such that:

$$E(X(t)) = \mu \quad \forall t \in \mathbb{Z} \tag{2}$$

$$\text{Cov}(X(t), X(t+h)) = \Gamma(h) \quad \forall t, h \in \mathbb{Z}. \tag{3}$$

A time series probabilistic graphical model (TGM), $G = (V, E)$, may be constructed by letting $(i, j) \notin E$ denote that the entire time series $X_i(:)$ and $X_j(:)$ are conditionally independent given the remaining collection of time series $X_{Z_{ij}}$ where $Z_{ij} = V \setminus \{i, j\}$. For the Gaussian stationary series we consider, one can show that conditional independence holds between time series iff [8]

$$\text{Cov}(X_i(t), X_j(t+h) | X_{Z_{ij}}) = 0 \ \ \forall h \in \mathbb{Z}. \tag{4}$$

The *spectral density matrix* of a stationary time series is defined as the Fourier transform of the lagged covariance matrices, $\Gamma(h) = \text{Cov}(X(t), X(t+h))$:

$$S(\lambda) = \sum_{h=-\infty}^{\infty} \Gamma(h) e^{-i\lambda h} \tag{5}$$

for $\lambda \in [0, 2\pi]$ and $S(\lambda) \in \mathbb{C}^{p \times p}$ and Hermitian positive definite. The marginal dependencies between time series are captured by $S(\lambda)$, and from Eq. (5), $S(\lambda)_{ij} = 0$ for all $\lambda \in [0, 2\pi]$ iff $\Gamma(h)_{ij} = 0$ for all $h \in \mathbb{Z}$. Furthermore, conditional independence between Gaussian stationary time series holds iff

$$S(\lambda)^{-1}_{ij} = 0 \ \ \forall \lambda \in [0, 2\pi], \tag{6}$$

implying that inferring zeros in the inverse spectral density matrices across frequencies equates with inferring the TGM structure [8]. More background on the spectral approach to time series is presented in the Supplement.

3 A BAYESIAN APPROACH

There are two standard approaches to Bayesian inference in graphical models: (1) placing a prior that jointly specifies the graph structure and associated parameters or (2) placing a prior on graph structures and then a prior on parameters given a graph; both rely on specifying a likelihood model. We opt for the second approach and describe the various components in this section. At a high level, our methods combine existing Whittle likelihood based methods [13, 14] with the hyper Markov framework to Bayesian graphical modeling [19, 18]. In the context of our TGMs, we introduce a conjugate *hyper complex inverse Wishart* prior on graph-constrained spectral density matrices. By integrating out the spectral density matrices, we obtain a marginal likelihood of the time series given the graph structure, G, allowing us to straightforwardly leverage state-of-the-art computational methods for i.i.d. Bayesian structure learning.

3.1 Whittle likelihood

Let $\mathbf{X} = [X(1), \ldots, X(T)]$, with $x(t) \in \mathbb{R}^p$ a realization of a p-dimensional stationary Gaussian time series observed at T time points, and $\mathbf{X}_{1:N} = \{\mathbf{X}^1, \ldots, \mathbf{X}^N\}$ be the collection of N independent realizations. We move to the frequency domain by transforming each $\mathbf{X}^i$ using a discrete Fourier transform. Let $d_{nk} \in \mathbb{C}^p$ denote the discrete Fourier coefficient associated with the nth time series at frequency $\lambda_k = \frac{2\pi k}{T}$:

$$d_{nk} = \frac{1}{T} \sum_{t=0}^{T-1} x_n(t) e^{-i\lambda_k t}. \tag{7}$$

The Whittle approximation [12] assumes the Fourier coefficients are independent *complex normal random variables* with mean zero and covariance given by the corresponding spectral density matrix $S_k = S(\lambda_k)$:

$$d_{nk} \sim \mathcal{N}_c(0, S_k) \quad k = 0, \ldots, T-1, \tag{8}$$

such that the likelihood of $\mathbf{X}_{1:N}$ is approximated as

$$p(\mathbf{X}_{1:N} | S_{0:T-1}) \approx \prod_{n=1}^{N} \prod_{k=0}^{T-1} \frac{1}{\pi^p |S_k|} e^{-d_{nk}^* S_k^{-1} d_{nk}}, \tag{9}$$

where $\frac{1}{\pi^p |S|} e^{-z^* S^{-1} z}$ is the density of a complex normal distribution, $\mathcal{N}_c(0, S)$, with $S \in \mathbb{C}^{p \times p}$ and Hermitian positive definite. See the Supplement. The Whittle approximation holds asymptotically with large T [24, 25, 12]. This approximation has been used in the Bayesian context in [26, 27]

Recall that conditional independencies are encoded in the off diagonal elements of S_k^{-1}. If time series $X_i(t)$ and $X_j(t)$ are conditionally independent, then the Whittle approximation says that as T gets large the ith and jth elements of the Fourier coefficients d_{nk} are conditional independent across all frequencies. Thus, if G is decomposable, Eq. (9) can be rewritten as

$$p(\mathbf{X}_{1:N} | G, S_{0:(T-1)}) \approx \prod_{k=0}^{T-1} \frac{\prod_{C \in \mathcal{C}} \frac{1}{\pi^{N|C|} |S_{kC}|^N} e^{-\text{tr} P_{kC} S_{kC}^{-1}}}{\prod_{S \in \mathcal{S}} \frac{1}{\pi^{N|S|} |S_{kS}|^N} e^{-\text{tr} P_{kS} S_{kS}^{-1}}} \tag{10}$$

where

$$P_k = \sum_{n=1}^{N} d_{nk} d_{nk}^* \tag{11}$$

is the aggregate *periodogram* over the N time series at frequency $\frac{2\pi k}{T}$. For $A \subset V$, S_{kA} and P_{kA} are the restriction of both matrices to the elements in A and $|A|$ denotes the cardinality of the set A.

3.2 Hyper complex inverse Wishart prior on graph-constrained spectral density matrices

We seek a prior for the spectral density matrices, S_k, whose inverses each have zeros dictated by a graph G. Recall that these S_k matrices are complex-valued and restricted to be Hermitian positive definite. As discussed in Sec. 2.2, the hyper inverse Wishart distribution serves as a prior for real-valued, positive-definite matrices with pre-specified zeros in the inverse, and is a conjugate prior for the covariance of a zero-mean GGM. Motivated by the connection between GGMs and our TGMs, and the analogous structure of our TGM-based Whittle likelihood of Eq. (10) to that of a GGM with N i.i.d. observations, we propose a novel *hyper complex inverse Wishart* prior with density function

$$p(\Sigma|\delta, W, G) =\propto \mathbf{1}_{\Sigma\in M^+(G)}|\Sigma|^{-(\delta+2p)}e^{-\mathrm{tr}W\Sigma^{-1}} \quad (12)$$

for *degrees of freedom* $\delta > 0$, *scale matrix* $W \in \mathbb{C}^{p\times p}$ positive definite and Hermitian, and graph G. We have used an analogous parameterization to that of the hyper inverse Wishart [16]. Here, $\Sigma \in M^+(G)$ denotes that Σ is in the set of all Hermitian positive-definite matrices with $\left(\Sigma^{-1}\right)_{ij} = 0$ for all $(i, j) \notin E$. When G is decomposable, the normalization constant is available and the density decomposes over cliques and separators:

$$\begin{aligned} p(\Sigma|\delta, W, G) &= \frac{\prod_{C\in\mathcal{C}} \mathrm{IW}_c(\Sigma_C|\delta, W_C)}{\prod_{S\in\mathcal{S}} \mathrm{IW}_c(\Sigma_C|\delta, W_C)} && (13)\\ &= \frac{\prod_{C\in\mathcal{C}} B(W_C,\delta)|\Sigma_C|^{-(\delta+2|C|)}e^{-\mathrm{tr}W_C\Sigma_C^{-1}}}{\prod_{S\in\mathcal{S}} B(W_S,\delta)|\Sigma_S|^{-(\delta+2|S|)}e^{-\mathrm{tr}W_S\Sigma_S^{-1}}}, && (14)\end{aligned}$$

where IW_c denotes the complex inverse Wishart [25] detailed in the Supplement with normalizer

$$B(W, \delta) = \frac{|W|^{\delta+p}}{\pi^{\frac{p(p-1)}{2}}\prod_{j=1}^{p}(\delta+p-j)!}. \quad (15)$$

We denote our proposed prior as $HIW_c(\delta, W, G)$ and specify

$$S_k \mid G \sim HIW_c(\delta_k, W_k, G) \quad k = 0, \dots, T-1. \quad (16)$$

In the Supplement, we show that this prior specification is *conjugate* to the TGM-based Whittle likelihood of Eq. (10). Also note that the graph, G, is shared across all frequencies.

3.3 Marginal likelihood

Due to conjugacy of our proposed hyper complex inverse Wishart prior, the marginal likelihood of the time series $\mathbf{X}_{1:N}$ given a decomposable graph G, integrating out the spectral density matrices $S_{0:T-1}$, has a closed form which is derived in the Supplement and given by

$$p(\mathbf{X}_{1:N}|G) \approx \pi^{-NTp}\prod_{k=0}^{T-1}\frac{h(W_k,\delta_k,G)}{h(W_k^*,\delta_k^*,G)}. \quad (17)$$

Here, $\delta_k^* = \delta_k + N$, $W_k^* = W_k + P_k$, and

$$h(W, \delta, G) = \frac{\prod_{C\in\mathcal{C}} B(W_C,\delta)}{\prod_{S\in\mathcal{S}} B(W_S,\delta)}. \quad (18)$$

From the definition of δ_k^*, we see that N, the number of time series, acts as the effective number of observations in this case. For the i.i.d. GGM, N represents the number of independent vector-valued observations; in our TGM, N plays the same role, but represents the number of independent *time series* observations. Likewise, as in standard inverse Wishart based modeling of covariances for i.i.d. Gaussian data, based on a set of N i.i.d. complex normal observations of Fourier coefficients d_{nk} with covariance S_k (see Eq. (9)), we update the prior scale matrix W_k with the outer product $P_k = \sum_{n=1}^{N} d_{nk}d_{nk}^*$, which is the aggregate *periodogram* (see Eq. (11)).

Having an analytic marginal likelihood of the time series given a PGM allows us to perform inference directly over graphs, sidestepping any thorny issues with inference directly on the T $p\times p$ spectral density matrices themselves. This is a critical feature of the practicality of our approach.

3.4 Fractional priors for model selection

Marginal likelihoods used for model comparison [28] are notoriously sensitive to the choice of prior parameters, or *hyperparameters*. In our case, the marginal likelihood in Eq. (17) depends strongly on the hyper complex inverse Wishart scale matrix, W_k. Since the scale and shape of the spectral density matrices are not known a priori, and vary dramatically across frequencies, we employ *fractional priors* [29] over each S_k. Fractional priors effectively hold out some fraction of the data, and utilize that fraction to determine an adequate hyperparameter setting for each model. The rest of the data are then used for model comparison. Fractional priors have been deployed for graphical model selection in i.i.d. graphs and have a number of desirable properties such as information consistency and demonstrated robustness [20]. In our case, under a fractional prior with parameter $g \in (0, 1)$, the fractional marginal likelihood is

$$p(\mathbf{X}_{1:N}|g, G) = \pi^{-NTp}\prod_{k=0}^{T-1}\frac{h(gP_k, gN, G)}{h(P_k, N, G)}. \quad (19)$$

Here, we see that g controls the fraction of data used for prior formulation versus model comparison. Importantly, we now have just a single, scalar, and interpretable parameter g to tune. Default settings are suggested in [29, 20].

3.5 Graph prior

There are two common approaches in the literature to specifying a prior distribution on graphs. The first approach places a uniform distribution on the space of all possible graphs [18, 30, 31]. As noted in [32], this prior puts high weight on graphs with a medium number of edges and significantly less weight on graphs with small or many edges. In response to this problem, it has been proposed to place a prior directly on the size of the graph and then consider a conditionally uniform prior on all graphs of the same size [32, 33, 19]. We follow this later approach and place a binomial distribution on the number of edges, k:

$$p(G) \propto r^k(1-r)^{m-k}, \tag{20}$$

where r is the prior probability that each of $m = \frac{p(p-1)}{2}$ possible undirected edges $(i, j) \in V \times V$ is included. Since r is unknown, we further place a Beta(a, b) prior over r. Integrating out r gives the marginal prior over graphs

$$p(G) \propto \frac{\beta(a+k, b+m-k)}{\beta(a,b)} \tag{21}$$

where $\beta(.,.)$ is the beta function. As explored in [20], this is a multiplicity correcting prior [34] over graphs with the desirable property of diminishing false positive edge discoveries as extra unconnected nodes are added to the graph.

4 METHODS FOR SINGLE TIME SERIES

In some applications of interest one observes only a single multivariate time series, $N = 1$, from which the graph must be inferred. Two challenges arise in this setting: (1) the effective number of observations informing Eq. (17) is just one and (2) the periodogram used in computing W_k^* is noisy regardless of the length of the series, T. The periodogram is a notoriously poor estimator of the spectral density, and when the spectral density itself is of primary interest, a common frequentist method is to smooth the periodogram to obtain a consistent spectral density estimator [14, 13, 8]. One could imagine using the smoothed periodogram as a plug-in estimator in Eq. (17), scaled by the effective degrees of freedom (see the Supplement for more details on this plug in estimator for our formulation). An alternative variance-reduction technique is the Bartlett method [35], that divides the length T series into M shorter series of length $\frac{T}{M}$ and averages the resulting M periodograms, but at the cost of reduced resolution (i.e., number of considered frequencies). This approach mimics the implicit smoothing that occurs when we compute the periodogram based on N truly independent series each of length T, as in Eq. (11).

In contrast to a plug-in estimator, a natural Bayesian approach enforces smoothing across frequencies via a prior distribution over the set of spectral densities [26]. Previous approaches have coupled elements of a Cholesky decomposition of each spectral density matrix across frequencies, however this approach is unsuitable to our case since 1) it does not enforce sparsity in the inverse spectral density and 2) a prior of this form will remove the simple marginal likelihood structure in Eq. (17) that we harness for efficient inference. Motivated by our aims to both share information across frequencies and maintain the form of the marginal, we utilize a piecewise constant prior over spectral densities given a graph, G. We partition the interval $[0, 2\pi]$ into M intervals $w_1 = \left[0, \frac{2\pi}{M}\right), \ldots, w_j = \left[\frac{2\pi(j-1)}{M}, \frac{2\pi j}{M}\right), \ldots, w_M = \left[\frac{2\pi(M-1)}{M}, 2\pi\right]$ and then draw a separate positive definite Hermitian matrix from a HIW_c distribution for each interval:

$$\tilde{S}_j \sim HIW_c(\delta, W_j, G) \quad j = 1, \ldots, M. \tag{22}$$

Our resulting spectral density is simply

$$S(\lambda) = \sum_{j=1}^{M} \mathbf{1}_{\lambda \in w_j} \tilde{S}_j \quad \forall \lambda \in [0, 2\pi]. \tag{23}$$

Under this prior, the marginal likelihood for the single $(N = 1)$ time series becomes

$$p(\mathbf{X}|G) \approx \pi^{-Mp} \prod_{j=1}^{M} \frac{h(W_j, \delta_j, G)}{h(W_j^*, \delta_j^*, G)} \tag{24}$$

where $\delta_j^* = \delta_j + \sum_{k=0}^{T-1} \mathbf{1}_{\lambda_k \in w_j}$ and $W_j^* = W_j + \sum_{k=0}^{T-1} \mathbf{1}_{\lambda_k \in w_j} P_k$. By setting $M = \lfloor \sqrt{T} \rfloor$, we obtain an asymptotically approximate nonparametric prior distribution over continuous spectral density matrices: for T large enough the prior puts positive support on spectral density matrices arbitrarily close to any continuous spectral density over $[0, 2\pi]$. Furthermore, under this setting as $T \to \infty$, the number of Fourier frequencies, and thus number of samples $\sum_{k=0}^{T-1} \mathbf{1}_{\lambda_k \in w_j}$, within each interval grows as $\sqrt{T}$.

5 INFERENCE

Bayesian structural learning algorithms for decomposable graphs come in two flavors: MCMC samplers and stochastic search procedures [20, 22]. By placing decomposable graphical inference for time series in the same framework as for the i.i.d. case via our analytic $p(\mathbf{X}_{1:N} \mid G)$, we can easily modify both types of existing methods for the time series case.

Classical MCMC samplers for decomposable graphs sample from the posterior over graphs via Metropolis-Hastings (MH) by proposing single edge addition and deletion moves that keep the graph decomposable [18, 32]. While it is possible to obtain any decomposable graph from any other decomposable graph via a sequence of edge additions and deletions, the path may be hard to reach leading to prohibitive converge times for even a moderate number of vertices p. More recent graph samplers add more global moves by either randomly generating new decomposable graphs [36] or by generating from a Markov chain over a junction tree representation of the graph [21]. To compute the MH acceptance ratio, these samplers rely on computing ratios of present and proposed marginal likelihoods. For simple edge additions and deletions, this ratio simplifies into a function of only the cliques and separators that change between moves. For our case, the ratio expands into a product over frequencies of the same affected cliques and separators, allowing simple modifications to the existing implementations of these samplers to handle TGMs.

All current MCMC samplers struggle to scale to even moderate numbers of nodes. In contexts where point estimates suffice, we can instead consider stochastic search procedures. We utilize a modification of the efficient feature-inclusion stochastic search (FINCS) [20] for inference in our TGMs. FINCS interleaves three moves: 1) single edge addition and deletion moves for local changes to the graph, 2) global sampling moves where edges are added independently to an empty graph and the final graph is triangulated to maintain decomposability, and 3) resampling at step t a full graph from a list of past visited models, $\{G_1, G_2, \dots, G_{t-1}\}$, in proportion to their posterior probabilities. In steps 1) and 2), to enforce exploration of high probability regions, edge additions that tend to continually improve the model probability are preferentially selected in proportion to a current heuristic estimate of the posterior edge probability

$$\hat{q}_{ij}(t) = \frac{\sum_{k=1}^{t} 1_{\{i,j\} \in E_t} p(X_{1:N}|G_t) p(G_t)}{\sum_{k=1}^{t} p(X_{1:N}|G_t)}, \qquad (25)$$

where E_t is the current edge set. Edge deletions are performed proportional to $\hat{q}_{ij}^{-1}(t)$. As in MCMC samplers [18, 32], the junction tree representation of the graph can be efficiently updated after each local move since the two graphs only differ by a single clique and its corresponding separators, allowing a quick computation of the marginal likelihood of a proposed graph in Eq. (17). Importantly, the FINCS algorithm depends on the data only through the marginal likelihoods of the cliques C—used to compute the full graph marginal likelihood—which in our TGM case is a product over T frequencies:

$$\prod_{k=0}^{T-1} \frac{B(W_{k,C}, \delta)}{B(W_{k,C}^*, \delta^*)}. \qquad (26)$$

That is, our implementation simply modifies the original FINCS definition of the clique marginal likelihood.

6 SIMULATIONS

To test our TGM methods, we consider simulated setups for both $N > 1$ and $N = 1$ time series each generated from an order-1 vector autoregressive process, denoted VAR(1), for $p = 20$ dimensions. Specifically, we simulated data from the model

$$x(t) = Ax(t-1) + \epsilon(t), \qquad (27)$$

where $x(t) \in \mathbb{R}^p$, $A \in \mathbb{R}^{p \times p}$, and $\epsilon(t) \sim N(0, I_{p \times p})$. The inverse spectral density of a VAR(1) process is given by [7]

$$S(\lambda)^{-1} = I + A^T A + e^{-i\lambda} A + e^{i\lambda} A^T. \qquad (28)$$

Random sparse TGMs were generated by first restricting A to be upper triangular. Following the simulated setup in [7], we set the diagonal elements to a constant $A_{ii} = .5$ and sample the upper diagonal elements as $a_{ij} \sim .5\delta_{ij}$, where $\delta_{ij} \sim \text{Binomial}(\rho)$ with $\rho = .2$ for all simulations. The graph G was then determined by identifying the zeros in $S(\lambda)^{-1}$ using Eq. (28). Proposed A matrices were accepted when both the absolute value of all eigenvalues of A were less than one, making the series stationary, and the graph G determined by A was decomposable.

We note that since our formulation reduces to a standard structure learning problem, our emphasis is less on assessing performance with respect to p, which should follow from whichever structure learning algorithm is selected; instead, our focus is on N and T, which are specific to the time series and spectral analysis. For example, in the FINCS algorithm [20], it is quoted that the method can handle graphs with up to roughly $p = 100$ nodes.

6.1 Multiple time series

To analyze how our TGM structure learning performance varies with the number of time series replicates, N, we simulated data for $N \in \{20, 50, 100, 150, 200, 250, 300, 350\}$ and $T \in \{25, 50, 100500, 1000, 1500, 2000\}$. This process was repeated 200 times for each combination of N and T. Each time series is first decomposed into its discrete Fourier components. We then ran 10,000 iterations of the FINCS algorithm using the fractional marginal likelihood in Eq. (19) with $g = \frac{4}{N}$, a default setting [29, 20]. Our graph prior followed the multiplicity correcting form in Eq. (21) with $a = b = 1$. The graph visited with highest posterior probability was then selected and true and false positive rates were computed. Results are displayed in Fig. 1. Across T, the true positive rate increases quickly with the number of series, N, achieving an almost perfect true positive rate by about $N = 150$. We also see that the rate of increase in the true positive rate increases with the length of the series T, which relates to the number of considered Fourier frequencies. It is interesting to note that for all T under consideration, the false positive rate tends to start very low ($\approx .005$) for $N = 20$ replicates then spike at $N \in \{50, 100\}$ before declining again. This occurs due to the fact that at low N, very few edges are introduced at all, perhaps due to an Occam's razor type effect of marginal likelihoods penalizing model complexity. As N starts to increase, more edges are introduced, both correct and incorrect, and as N further increases, the false edges are pruned and true edges are retained, leading to a decline in the false positive rate. Note that the false positive spike tends to be more pronounced for time series of

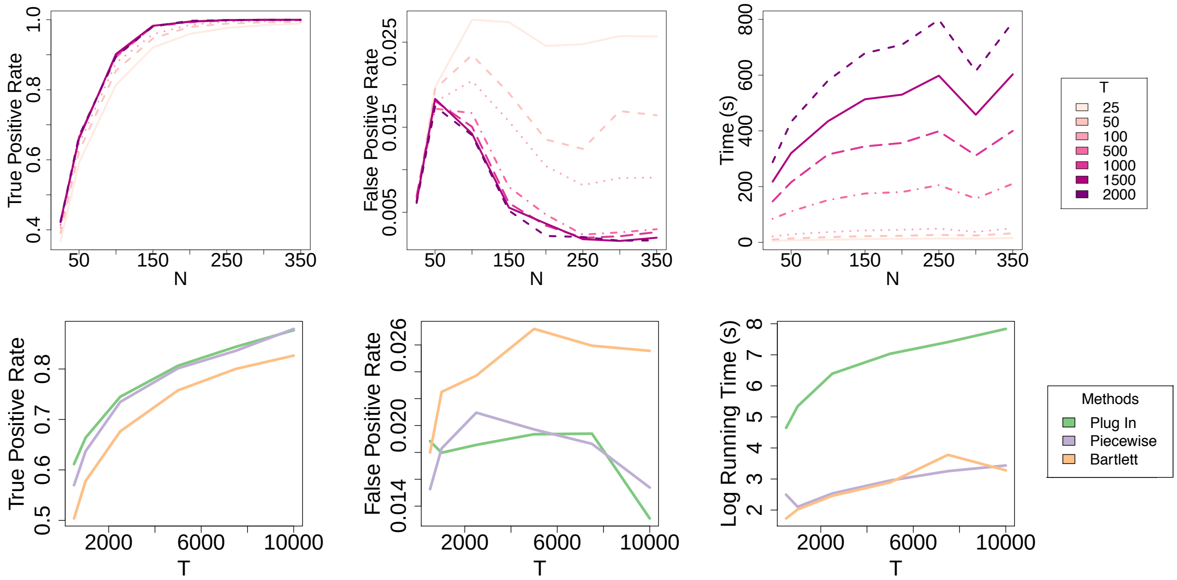

Figure 1: **Top:** As a function of the number of time series N, and plotted for various values of their length T, (*left*) mean true positive rate, (*middle*) median false positive rate, and (*right*) mean running time computed across the 200 replicates. Standard error bars are small relative to the scale of the plots and are omitted for clarity. **Bottom:** Same plots as a function of T for a single time series ($N = 1$), and plotted for various periodogram smoothing techniques.

smaller length, $T \in \{25, 50\}$. One would expect to see significant improvements, especially for small N, by leveraging the piecewise constant prior of Sec. 4 and explored in Sec. 6.2 where we show that we are able to learn graphs from just $N = 1$ time series. However, we chose not to include this prior in this analysis so as not to confound its effect with our performance. Here, the noisy periodogram is smoothed implicitly by averaging over N.

Finally, in Fig. 1 we see that runtime increases as a function of T due to the dependence on T in the marginal likelihood computation of Eq. (17), though significant cost reductions can be achieved through parallelizations leveraging the product form.

6.2 Single time series: comparison of methods

To assess the performance of our single-time-series methods outlined in Sec. 4, we simulated a time series with $T \in \{500, 1000, 2500, 5000, 7500, 10000\}$. For the piecewise constant prior method, we use $M = \lfloor\sqrt{T}\rfloor$ pieces. We compare against the Bartlett time-series-splitting approach with the number of splits set to $\lfloor\sqrt{T}\rfloor$. We also examine a smoothed plug-in estimator of the spectral density using a Daniell smoother outlined in the Supplement with $m = \lfloor\frac{\sqrt{T}}{2}\rfloor$ for a total window size of $2\lfloor\frac{\sqrt{T}}{2}\rfloor + 1 \approx \lfloor\sqrt{T}\rfloor$. For each method, the FINCS algorithm was run for 10,000 iterations and the highest scoring graph was selected and used to compute true and false positive rates. This process was repeated 200 times with results displayed in Fig. 1 with a replicate representative of our median performance shown in Fig. 2. The true positive rate increases for all three methods as a function of T, achieving a final value of about .9 for both the plug-in and piecewise constant prior methods and .79 for the Bartlett method at $T = 10000$. All methods maintain a low false positive rate around .02. Overall, the Bartlett method performs uniformly worse in terms of both true and false positive performance, while the piecewise prior method performs on par with the plug-in method, but at a fraction of the computational cost. Further experimental simulations are given in the Supplement.

7 GLOBAL STOCK INDICES

We explore the utility of our method in discovering conditional independencies between countries inherent in the global financial system. A similar experiment was conducted in [7] using a penalized-likelihood approach to learn TGMs, but restricted to finite-order VAR models with pre-specified order. (Recall that our method only assumes Gaussian stationarity, which includes the class of possibly infinite order VAR processes.) Using `www.globalfinancialdata.com`, we acquired the daily closing prices of 17 stock indices in US dollars for various countries around the world (see the Supplement for the full list) from June 3, 1997 to June 30, 1999. Missing prices were backfilled and only days where all exchanges traded were considered which resulted in time series of length 542. Following standard practice when analyzing stock prices, we converted the closing prices, p_t, on day t to log-returns according to

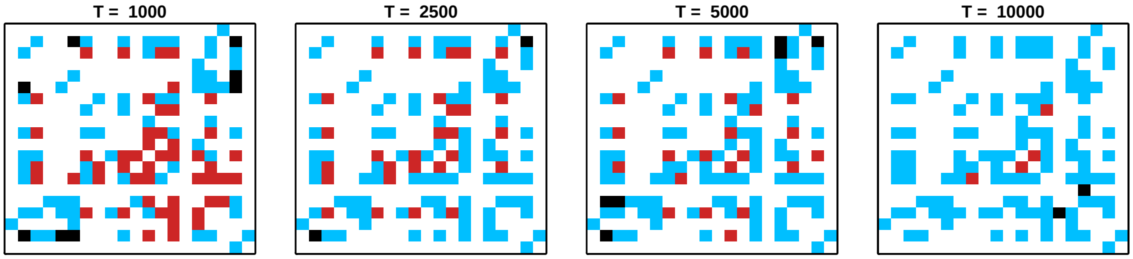

Figure 2: Example evolution of error types for the piecewise prior method as a function of series length, $T \in \{1000, 2500, 5000, 10000\}$ and $N = 1$, for a selected graph. Blue, red, **black**, and white entries indicate true positives, false negatives, false positives, and true negatives, respectively. The graph was selected by choosing the graph out of 200 replications with median true positive rate at $T = 2500$.

$$r_t = 100 \log(p_t / p_{t-1}).$$

We compare the graphical models inferred under two settings: (i) treating the log-returns as independent (as in [20]) and (ii) using our methods to learn a TGM treating the log-returns as a time series. The best graphical models learned in each scenario are depicted in Fig. 3.

For our TGM algorithm, we computed the periodogram for the 17-dimensional time series, resulting in 542 complex-valued matrices of dimension 17×17. Since we only have one realization of the time series, we smoothed the periodogram using the techniques and settings discussed in Sec. 6.2. We then ran the FINCS algorithm for 100,000 iterations. We compare the resulting highest-probability graph (see Fig. 3) to that learned treating the time series as independent based on the model in [20], again using 100,000 iterations of the FINCS algorithm, but in its originally proposed form for non-temporal data.

In Figure 3, we see that in both cases we recover some geographical relationships between countries. However, the independent model returns a significantly denser graph than that learned by our TGM approach. Since the independent model is not taking the temporal nature of the data into account, some edges are likely spurious due to random correlations. The TGM, on the other hand, provides an interpretable and intuitive structure with strong geographic connections. For example, there is a distinct United Kingdom / eurozone cluster of Germany ‘DE’, Finland ‘FI’, Netherlands ‘NL’, Belgium ‘BE’, Switzerland ‘CH’, Austria ‘AT’, Spain ‘ES’, Italy ‘IT’, Portugal ‘PT’, and the United Kingdom ‘UK’. Another distinct cluster includes the United States ‘US’, Canada ‘CA’, Hong Kong ‘HK’ (whose currency is linked to the USD), and Australia ‘AU’ (whose currency is correlated with the US S&P), with Japan ‘JP’ hanging off this cluster. One perhaps strange missing link is between Ireland ‘IE’ and the UK, though the US and Ireland have a long history of economic connections possibly explaining why Ireland is included in the separator between these two distinct clusters.

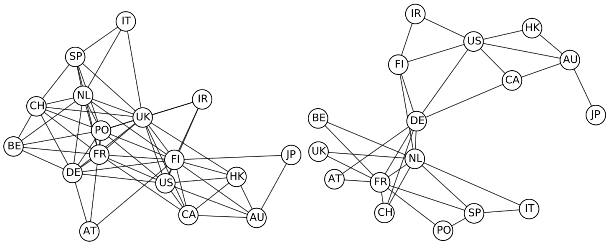

Figure 3: Graphical models with the highest posterior probability for the stock index data. **Left:** Treating the log-returns as independent. **Right:** Using our TGM algorithm. In both cases, we see regional connections, but our TGM algorithm results in a sparser and more interpretable graph.

In the Supplement, we include (i) a comparison of our learned graph with that of Songsiri et. al. [7], and (ii) further details on the stock data itself.

8 MAGNETOENCEPHALOGRAPHY DATA

Next we learn TGMs to capture the structure of underlying cortical dynamics from magnetoencephalography (MEG) data collected from ten subjects who were asked to perform a task while maintaining focus on an audio stream and then again while switching focus [37]. Our goal is to discover differences in the underlying TGMs between the non-switching and switching attention conditions. Such differences provide further understanding into the neural underpinnings of auditory selective selection, an important constituent to communication.

The data were collected for each subject performing the experiment in the *switching* (S) and *non-switching* (N) attention conditions. For both S and N conditions, each subject performed the task under an auditory condition of *high* (U) and *low* (D) pitch, and spatial conditions of *left* (L) and *right* (R) attending. For each of the eight possible conditions, MEG recordings were collected resulting in a 150-dimensional time series of length 992 where each dimen-

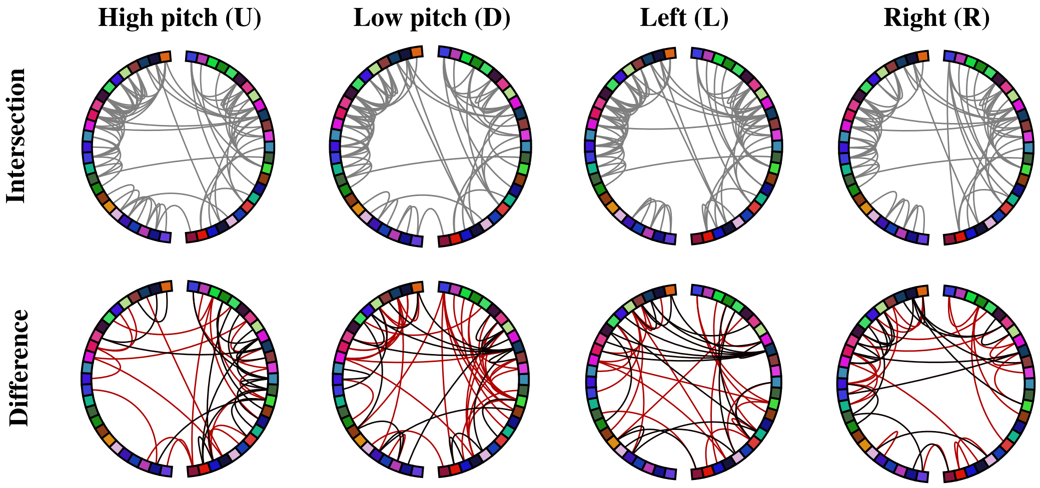

Figure 4: Learned TGMs for different MEG conditions. Each node on the periphery represents a brain region with location indicating anatomical location. **Top:** Intersection of learned edges between switching and non-switching conditions. **Bottom:** *Black* edges indicating those in the non-switching condition but not in the switching and *red* vice versa.

sion corresponds to a localized region of the brain. We have between 17 and 30 trials for each subject, resulting in about 200 replicate time series per condition.

Often with MEG data, many of the dimensions are dominated by noise due to limited brain activity in that region. We reduced the number of brain regions we studied from 150 to 50 by only considering those with largest variance. In particular, for each trial we mean-centered all of the time-series and computed the variance and retained the top 50 most volatile regions.

We computed the periodogram for each trial and averaged across trials within each condition, resulting in eight periodograms. We ran our spectral TGS version of the FINCS algorithm on these periodograms for 100,000 iterations with fractional prior parameter $4/N_c$, where N_c is the number of trials for condition $c \in \{\text{S, N}\} \times \{\text{U, D, L, R}\}$. We also ran the algorithm for 1.7 million iterations and saw no difference in the resulting graphs.

In Figure 4, we depict the intersections and differences between the learned graphs for each experimental condition. We see in the top row that there are a lot of shared connections between the switching and non-switching conditions for each auditory condition. In the bottom row, the differences between the switching and non-switching conditions are depicted where red edges are those in the switching condition but not the non-switching, and black edges are the reverse. The difference plots show that there seems to be substantial "rewiring" for many of the conditions with many edges connecting frontal to back regions. Interestingly, we again see consistencies in these rewirings across conditions. Additionally, we reliably uncover local connections between adjacent brain regions across experimental conditions. Such observations provide guidance for developing experiments and methods to discern the underlying mechanisms that give rise to these different structures.

9 DISCUSSION

We introduced a Bayesian approach to graphical model structure learning for time series. In particular, we propose a prior—the *hyper complex inverse Wishart* distribution—for the spectral density matrices in a Whittle likelihood approximation. For decomposable graphs, this prior is conjugate and leads to a closed-form expression of the marginal likelihood of the time series given the graph, marginalizing the spectral density matrices across frequencies. Being able to integrate out this large collection of complex matrices—one for each time point—is critical to developing a practical and scalable inference algorithm. For this, exploiting the fact that our marginal likelihood is analogous to that for i.i.d. Gaussian graphical models [19] but with a product over the number of Fourier frequencies, allows us to deploy straightforward modifications to existing MCMC and stochastic search algorithms. Our simulations show that when many time series are observed, our method recovers the correct graph. When a single time series is observed, we proposed a method to increase robustness of our graph estimation using a piecewise constant prior. Our results on the stock and MEG datasets demonstrated our ability to discover intuitive and interpretable structure in these datasets, importantly leveraging the temporal dependencies.

Extensions to non-decomposable graphs are possible using the i.i.d. graph approaches in both [31] and [22]. A Laplace approximation to the marginal likelihood for non-decomposable graphs is proposed in [22], which we could similarly utilize to approximate the frequency-specific marginal at each term in Equation (17). Parallelizing the Laplace approximation computation across frequencies would lead to a scalable method for inference in non-decomposable time series graphs.

Acknowledgements: This work was supported by DARPA Grant FA9550-12-1-0406 negotiated by AFOSR, ONR Grant N00014-10-1-0746, NSF CAREER Award IIS-1350133, and AFOSR Grant FA9550-12-1-0453.

References

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

[2] B. D. Ziebart, A. K. Dey, and J. A. Bagnell. Learning selectively conditioned forest structures with applications to dbns and classification. *UAI*, 2007.

[3] M. R. Siracusa and J. W. Fisher III. Tractable Bayesian inference of time-series dependence structure. In *AISTATS*, 2009.

[4] O. Sporns. *Networks of the Brain*. MIT Press, 2010.

[5] T. Medkour, A. T. Walden, Burgess A. P., and Strelets V. B. Brain connectivity in positive and negative syndrome schizophrenia. *Neuroscience*, 169(4):1779 – 1788, 2010.

[6] U. Gather, M. Imhoff, and R. Fried. Graphical models for multivariate time series from intensive care monitoring. *Statistics in Medicine*, 21(18), 2002.

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

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

[9] S. L. Lauritzen. *Graphical Models*. Oxford University Press, 1996.

[10] Y. Matsuda. A test statistic for graphical modelling of multivariate time series. *Biometrika*, 93(2):pp. 399–409, 2006.

[11] R. J. Wolstenholme and A. T. Walden. An efficient approach to graphical modeling of time series. *ArXiv e-prints*, 2015.

[12] P. Whittle. The analysis of multiple stationary time series. *JRSS(B)*, 15(1):125–139, 1953.

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

[14] A. Jung, G. Hannak, and N. Görtz. Graphical LASSO based model selection for time series. *ArXiv e-prints*, 2014.

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

[16] A. P. Dawid and S. L. Lauritzen. Hyper Markov laws in the statistical analysis of decomposable graphical models. *Ann. Statist.*, 21(3):1272–1317, 1993.

[17] C. M. Carvalho and J. G. Scott. Objective Bayesian model selection in Gaussian graphical models. *Biometrika*, 96(3):497–512, 2009.

[18] P. Giudici and P. J. Green. Decomposable graphical Gaussian model determination. *Biometrika*, 86(4):785–801, 1999.

[19] B. Jones, C. Carvalho, A. Dobra, C. Hans, C. Carter, and M. West. Experiments in stochastic computation for high-dimensional graphical models. *Statistical Science*, 20(4):388–400, 2005.

[20] J. G. Scott and C. M. Carvalho. Feature-inclusion stochastic search for Gaussian graphical models. *J. Comp. Graph. Stat.*, 17(4):790–808, 2008.

[21] P. J. Green and A. Thomas. Sampling decomposable graphs using a Markov chain on junction trees. *Biometrika*, 100(1):91–110, 2013.

[22] M. Baback, E. Khan, K. M. Murphy, and B. M. Marlin. Accelerating Bayesian structural inference for non-decomposable Gaussian graphical models. In *NIPS*, pages 1285–1293, 2009.

[23] A. Mohammadi and E. C. Wit. Bayesian structure learning in sparse Gaussian graphical models. *Bayesian Analysis*, 10(1):109–138, 2015.

[24] P. J. Brockwell and R. A. Davis. *Time Series: Theory and Methods*. Springer-Verlag, New York, NY, 1991.

[25] D.R. Brillinger. *Time Series: Data Analysis an Theory*. Holden-Day, 2001.

[26] O. Rosen and D. S. Stoffer. Automatic estimation of multivariate spectra via smoothing splines. *Biometrika*, 94(2):335–345, 2007.

[27] R. T. Krafty, O. Rosen, D. S. Stoffer, D. J. Buysse, and M. H. Hall. Conditional spectral analysis of replicated multiple time series with application to nocturnal physiology. *ArXiv e-prints*, 2015.

[28] R. E. Kass and A. E. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90:773–795, 1995.

[29] A. O'Hagan. Fractional Bayes factors for model comparison. *JRSS(B)*, 57(1):99–138, 1995.

[30] P. Dellaportas, P. Giudici, and G. Roberts. Bayesian inference for nondecomposable graphical Gaussian models. *Sankhy: The Indian Journal of Statistics*, 65(1):43–55, 2003.

[31] A. Roverato. Hyper inverse Wishart distribution for non-decomposable graphs and its application to Bayesian inference for Gaussian graphical models. *Scand. J. Stat*, 29(3):391–411, 2002.

[32] H. Armstrong, C. Carter, K. Wong, and R. Kohn. Bayesian covariance matrix estimation using a mixture of decomposable graphical models. *Statistics and Computing*, 19(3):303–316, 2009.

[33] A. Dobra, C. Hans, B. Jones, J.R. Nevins, Joseph R., G. Yao, and M. West. Sparse graphical models for exploring gene expression data. *Journal of Multivariate Analysis*, 90(1):196–212, 2004.

[34] J. G. Scott and J. O. Berger. An exploration of aspects of Bayesian multiple testing. *Journal of Statistical Planning and Inference*, 136:2144–2162, 2006.

[35] M. S. Bartlett. Smoothing periodograms from time series with continuous spectra. *Nature*, 161(4096):686–687, 1948.

[36] H. Zhu, N. Strawn, and D. B. Dunson. Bayesian graphical models for multivariate functional data. *ArXiv e-prints*, 2014.

[37] E. Larson and A.K.C. Lee. Switching auditory attention using spatial and non-spatial features recruits different cortical networks. *NeuroImage*, 84:681–687, 2014.