Computational Statistics and Data Analysis 178 (2023) 107596

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

www.elsevier.com/locate/csda

Fast Bayesian inference on spectral analysis of multivariate stationary time series ☆

Zhixiong Hu *, Raquel Prado

Department of Statistics, University of California, Santa Cruz, 1156 High St, Santa Cruz, 95064, CA, USA

ARTICLE INFO

Article history:
Received 17 November 2021
Received in revised form 9 August 2022
Accepted 12 August 2022
Available online 19 August 2022

Keywords:
Multivariate time series
Spectral analysis
Stochastic gradient variational Bayes
Global-local shrinkage prior

ABSTRACT

Spectral analysis discovers trends, periodic and other characteristics of a time series by representing these features in the frequency domain. However, when multivariate time series are considered, and the number of components increases, the size of the spectral density matrix grows quadratically, making estimation and inference rather challenging. The proposed novel Bayesian framework considers a Whittle likelihood-based spectral modeling approach and imposes a discounted regularized horseshoe prior on the coefficients that define a spline representation of each of the components of a Cholesky factorization of the inverse spectral density matrix. The proposed prior structure leads to a model that provides higher posterior accuracy when compared to alternative currently available approaches. To achieve fast inference that takes advantage of the massive power of modern hardware (e.g., GPU), a stochastic gradient variational Bayes approach is proposed for the highly parallelizable posterior inference that provides computational flexibility for modeling high-dimensional time series. The accurate empirical performance of the proposed method is illustrated via extensive simulation studies and the analysis of two datasets: a wind speed data from 6 locations in California, and a 61-channel electroencephalogram data recorded on two contrasting subjects under specific experimental conditions.

1. Introduction

Spectral analysis has been widely used to characterize the properties of one or more time series in the frequency domain. Accurate inference of spectral density matrices is critical for understanding the structure underlying the components of a given multivariate temporal process, and for revealing potential relationships across its components. However, inference of spectral density matrices suffers from the curse of dimensionality. This paper develops models and methods to obtain scalable and accurate inference on the spectral density matrix, and functions of this matrix, for high-dimensional stationary multivariate time series under a Bayesian framework.

Several frequency domain methods have been developed for Bayesian spectral analysis of stationary multivariate time series. Rosen and Stoffer (2007) proposed a Bayesian approach that uses Markov chain Monte Carlo (MCMC) techniques to obtain posterior inference in a model that considers smoothing splines to represent the real and imaginary components of the modified complex Cholesky decomposition of the inverse spectral density matrix. In earlier related work,

☆ Supplementary materials with code and dataset are available online.

* Corresponding author.
E-mail addresses: zhu95@ucsc.edu (Z. Hu), raquel@soe.ucsc.edu (R. Prado).

https://doi.org/10.1016/j.csda.2022.107596
0167-9473/

Dai and Guo (2004) used a Cholesky factorization of the spectral density matrix. More recently, Krafty and Collinge (2013) derived an approach that considers a penalized Whittle log-likelihood to incorporate regularization on the multivariate power spectra, allowing for varying levels of smoothness among power spectral components. Meier et al. (2020) modeled the spectral density matrix with matrix-valued mixture weights induced by a Hermitian positive definite Gamma process, effectively extending the univariate Bernstein-Dirichlet process prior approach of Choudhuri et al. (2004) to the multivariate case. There are also a few approaches that adapt the aforementioned stationary models to handle nonstationary multivariate spectral analysis (Zhang, 2016; Li and Krafty, 2019; Li et al., 2021).

The approaches for spectral analysis of multivariate stationary and nonstationary time series discussed above are sophisticated and very flexible, however, they incur very high computational costs and hence, lack of scalability in high-dimensional and even in relatively low or moderate-dimensional settings that involve joint analysis of a collection of time series components. Here we focus on providing accurate and scalable inference in the case of stationary spectral analysis of a p-dimensional time series, where a p-by-p spectral matrix needs to be inferred. Following an approach similar to that of Rosen and Stoffer (2007), we model each component in the Cholesky decomposition of the inverse spectral density matrix via a smoothing spline with M bases. In this case, the total number of unknown parameters in the model is proportional to p^2M, which grows quadratically with the number of time series p, and linearly on the number of basis functions M. Using low-rank factor representations such as those proposed by Li et al. (2021) considerably reduces the number of model parameters and improves scalability, however, it does not fully solve the problem in practical settings where several time series need to be jointly analyzed. In particular, using iterative simulation-based techniques, such as MCMCs, to obtain posterior samples of all the unknown parameters is rather challenging if both p and M are large, resulting in a tremendous computational overhead.

We develop and illustrate a novel and fast Bayesian modeling and inference framework for spectral analysis of multivariate stationary time series. A key feature of our approach is that it uses variational inference (VI), leading to fast approximate posterior inference in cases where the number of time series is large. VI casts inference in terms of the optimization of a surrogate distribution that approximates the intractable posterior distribution (Blei et al., 2017), and can generally be facilitated and scaled up by stochastic optimization (Robbins and Monro, 1951). For instance, Kingma and Welling (2013) proposed a stochastic gradient variational Bayes (SGVB) approach that utilizes the gradient of the unnormalized log-posterior distribution throughout optimization. Assuming a multivariate Gaussian variational distribution, SGVB relies on the reparameterization trick to yield a simple differentiable unbiased and low-variance gradient estimator (Xu et al., 2019; Domke, 2019). In comparison to MCMC, VI tends to be much faster and scalable for large-scale inference problems.

Our proposed approach uses a SGVB scheme that is highly parallelizable and compatible with GPU acceleration, providing a computationally efficient framework for spectral modeling and inference of high-dimensional time series. As some of the approaches discussed above, we use a rich set of spline bases to represent the inverse spectral density matrix, preserving model flexibility. In addition, in order to allow for model flexibility while also avoiding overfitting, we consider a modified version of the regularized horseshoe priors of Piironen and Vehtari (2017), referred to as discount regularized horseshoe priors, on the parameters that define the spline representation of the inverse of the spectral density matrix. Our discounted regularized horseshoe (DRH) prior includes a pre-specified discount factor on the hyper-parameter that controls the local shrinkage parameters on spline coefficients. The proposed prior varies the degree of regularization according to the smoothness of the basis functions. In addition to providing various levels of smoothness for the spectral components, the global and local shrinkage parameters in the proposed prior add significant model flexibility.

The remainder of the article is organized as follows. Section 2 specifies the model and priors used for spectral analysis of stationary multivariate time series. Section 3 describes the proposed stochastic gradient variational Bayes scheme for fast and scalable posterior inference. Section 4 reports results of extensive simulation studies that illustrate the accuracy and scalability of the proposed approach. Section 5 applies the proposed framework to the analysis of multi-location wind speed time series data from the Iowa Environmental Mesonet (IEM) database (Todey et al., 2002). This section also presents the results of the analysis of a multi-channel electroencephalogram dataset from the UCI Machine Learning Repository (Dua and Graff, 2017).

2. Model specification

2.1. The modified complex Cholesky decomposition

Let $\{\mathbf{x}_t\}$ denote a p-dimensional time series process. Consider a realization $\{\mathbf{x}_1, ..., \mathbf{x}_n\}$ of such process, or $\mathbf{x}_{1:n}$. These data can be represented in the frequency domain via the discrete Fourier transform (DFT), $\boldsymbol{y}(\nu_k) = n^{-\frac{1}{2}} \sum_{t=1}^{n} \mathbf{x}_t \, e^{-2\pi \, i \, \nu_k t}$, where $\nu_k = k/n$ denote Fourier frequencies, $k = 0, 1, ..., n-1$. Because the discrete Fourier transform is an even function of ν, there are only $[n/2]$ distinct $\boldsymbol{y}(\nu_k)$. If we assume $\{\mathbf{x}_t\}$ to be stationary with $p \times p$ autocovariance matrix, $\boldsymbol{\Gamma}(h) = \{\gamma_{il}(h)\}$, satisfying $\sum_{h=-\infty}^{\infty} |\gamma_{il}(h)| < \infty$ for all $i, l = 1, ..., p$, there exists a $p \times p$ spectral density matrix $\boldsymbol{f}(\nu) = \sum_{h=-\infty}^{\infty} \boldsymbol{\Gamma}(h) \, e^{-2\pi \, i \, \nu h}$, where $\nu \in [-0.5, 0.5]$ is a frequency measured in cycles per unit time. Note that $\boldsymbol{f}(\nu) = \boldsymbol{f}^*(-\nu)$ where $*$ denotes the conjugate transpose. Additionally, $\boldsymbol{f}(\nu)$ is also positive definite. The diagonal entries of $\boldsymbol{f}(\nu)$, $\boldsymbol{f}_{jj}(\nu)$ for $j = 1, ..., p$, correspond to the spectral densities of each of the p components of the time series $\mathbf{x}_t$. The off-diagonal elements of the spectral matrix can be combined to define the squared coherence $\rho_{il}^2(\nu) = \frac{|\boldsymbol{f}_{il}(\nu)|^2}{\boldsymbol{f}_{ii}(\nu)\boldsymbol{f}_{ll}(\nu)}$, $\rho_{il}^2(\nu) \in [0, 1]$, where $i, l = 1, ..., p$ with $i \neq l$. The coherence

is a frequency-domain analog of a the correlation between two components. It measures the strength of the association of the different time series components for a given frequency ν. Squared coherence values closer to one indicate strong association, while values close to zero show little or no association between components i and l.

Following an approach similar to that in Rosen and Stoffer (2007), we use a Cholesky factorization of the inverse of the spectral density matrix and then represent the components of this factorization via smoothing splines, as follows. Let $\mathbf{y}_k \equiv \boldsymbol{y}(\nu_k)$ and $\mathbf{f}_k \equiv \boldsymbol{f}(\nu_k)$ with $\boldsymbol{f}$ the matrix-valued spectral density function. Then, the multivariate extension of the Whittle likelihood approximation (Whittle, 1957) is given by $L(\mathbf{y}_{1:N}; \mathbf{f}_{1:N}) \approx \prod_{k=1}^{N} \det(\mathbf{f}_k)^{-1} \exp\left(-\mathbf{y}_k^* \mathbf{f}_k^{-1} \mathbf{y}_k\right)$, with $N = [n/2]$. To ensure that $\boldsymbol{f}$ is positive definite, $\mathbf{f}_k^{-1}$ can be modeled through the modified complex Cholesky factorization:

$$\mathbf{f}_k^{-1} = \mathbf{T}_k^* \mathbf{D}_k^{-1} \mathbf{T}_k, \tag{1}$$

where $\mathbf{D}_k = \operatorname{diag}\left(\delta_{1k}^2, \ldots, \delta_{pk}^2\right)$, and $\mathbf{T}_k$ is a complex unit lower triangular matrix such that

$$\mathbf{T}_k = \begin{pmatrix} 1 & & & & \\ -\theta_{21}^{(k)} & 1 & & & \\ -\theta_{31}^{(k)} & -\theta_{32}^{(k)} & 1 & & \\ \vdots & \vdots & & \ddots & \\ -\theta_{p1}^{(k)} & -\theta_{p2}^{(k)} & \ldots & -\theta_{p,p-1}^{(k)} & 1 \end{pmatrix}. \tag{2}$$

Consequently, the likelihood can be rewritten as:

$$L(\mathbf{y}_{1:N}; \mathbf{f}_{1:N}) \propto \prod_{k=1}^{N} \prod_{j=1}^{p} \left[\delta_{jk}^{-2} \exp\left(- \frac{|y_{jk} - \sum_{l=1}^{j-1} \theta_{jl}^{(k)} y_{lk}|^2}{\delta_{jk}^2} \right) \right], \tag{3}$$

with y_{jk} denoting the j-th dimension of $\mathbf{y}_k$, and $|\cdot|$ denoting the absolute value. Note that (3) decomposes the likelihood into p components. For $j = 1, \ldots, p$, each component is written as

$$L_j(\mathbf{y}_{1:N}; \mathbf{f}_{1:N}) = \prod_{k=1}^{N} \left[\delta_{jk}^{-2} \exp\left(- \frac{|y_{jk} - \sum_{l=1}^{j-1} \theta_{jl}^{(k)} y_{lk}|^2}{\delta_{jk}^2} \right) \right], \tag{4}$$

such that $L(\mathbf{y}_{1:N}; \mathbf{f}_{1:N}) \propto \prod_{j=1}^{p} L_j(\mathbf{y}_{1:N}; \mathbf{f}_{1:N})$ and so, posterior inference can be obtained in parallel over j.

2.2. Prior specification

Rosen and Stoffer (2007) used a basis representation to model the entries of the $\mathbf{D}_k$'s and $\mathbf{T}_k$'s. Our approach considers the same representation, which is briefly described below, but uses a different prior structure on the parameters that define this representation. More specifically, the Demmler-Reinsch basis functions (Eubank, 1999) are used to represent each $\log \delta_{jk}^2$ and the real and imaginary parts of each $\theta_{jl}^{(k)}$ in terms of M-truncated smoothing splines: $\log \delta_{jk}^2 = \gamma_{j,0} + \gamma_{j,1} \nu_k + \sum_{s=1}^{M-1} \psi_s(\nu_k) \gamma_{j,s+1}$, $\Re(\theta_{jl}^{(k)}) = \alpha_{jl,0} + \alpha_{jl,1} \nu_k + \sum_{s=1}^{M-1} \psi_s(\nu_k) \alpha_{jl,s+1}$, and $\Im(\theta_{jl}^{(k)}) = \beta_{jl,0} + \beta_{jl,1} \nu_k + \sum_{s=1}^{M-1} \psi_s(\nu_k) \beta_{jl,s+1}$, where $\psi_s(\nu_k) = \sqrt{2} \cos(s\pi \nu_k)$. M is a pre-specified constant determining the number of basis functions to be included and therefore the model flexibility. Let $\mathbf{X}_k = (1, \nu_k, \psi_1(\nu_k), \ldots, \psi_{M-1}(\nu_k))$, $\boldsymbol{\gamma}_j = (\gamma_{j,0}, \ldots, \gamma_{j,M})'$, $\boldsymbol{\alpha}_{jl} = (\alpha_{jl,0}, \ldots, \alpha_{jl,M})'$, and $\boldsymbol{\beta}_{jl} = (\beta_{jl,0}, \ldots, \beta_{jl,M})'$. Then, the aforementioned smoothing splines can be rewritten as:

$$\log \delta_{jk}^2 = \mathbf{X}_k \boldsymbol{\gamma}_j, \quad \Re(\theta_{jl}^{(k)}) = \mathbf{X}_k \boldsymbol{\alpha}_{jl}, \quad \Im(\theta_{jl}^{(k)}) = \mathbf{X}_k \boldsymbol{\beta}_{jl}. \tag{5}$$

For notation brevity, let $\boldsymbol{\alpha}_{j\cdot}$ denote $\{\boldsymbol{\alpha}_{j1}, \ldots, \boldsymbol{\alpha}_{jj-1}\}$ and $\boldsymbol{\beta}_{j\cdot}$ denote $\{\boldsymbol{\beta}_{j1}, \ldots, \boldsymbol{\beta}_{jj-1}\}$. By plugging (5) into (4), $L_j(\mathbf{y}_{1:N}; \mathbf{f}_{1:N})$ can be reparameterized as $L_j(\mathbf{y}_{1:N}; \boldsymbol{\gamma}_j, \boldsymbol{\alpha}_{j\cdot}, \boldsymbol{\beta}_{j\cdot})$, where for every $j = 1, \ldots, p$:

$$\log L_j(\mathbf{y}_{1:N}; \boldsymbol{\gamma}_j, \boldsymbol{\alpha}_{j\cdot}, \boldsymbol{\beta}_{j\cdot}) = -\sum_{k=1}^{N} \left[\mathbf{X}_k \boldsymbol{\gamma}_j + \frac{|y_{jk} - \sum_{l=1}^{j-1} \left(\mathbf{X}_k \boldsymbol{\alpha}_{jl} + \mathrm{i}\, \mathbf{X}_k \boldsymbol{\beta}_{jl}\right) y_{lk}|^2}{\mathrm{e}^{\mathbf{X}_k \boldsymbol{\gamma}_j}} \right]. \tag{6}$$

Note that in the case of $k = 1$, L_1 depends solely on $\boldsymbol{\gamma}_1$.

Our goal is to obtain posterior estimates of the $\boldsymbol{\gamma}_j$'s, $\boldsymbol{\alpha}_{jl}$'s, and $\boldsymbol{\beta}_{jl}$'s. Hence, priors on spline coefficients are needed. One option is to use normal priors as in Rosen and Stoffer (2007); Zhang (2019). We show later via simulation study that such priors can be inadequate to handle cases in which some of the time series components require a large number of bases in their representation, while others require a small number, thus failing to provide a flexible modeling framework that can adequately capture this situation. An ideal prior setting should allow the selection of enough basis terms to achieve a good fit, while also avoiding overfitting for each component. To achieve this, our prior setting is based upon the so-called

regularized horseshoe prior (Piironen and Vehtari, 2017). This is a global-local shrinkage prior with a global parameter that provides shrinkage towards zero for all the components sharing this parameter, and local, or component-specific parameters that allow some of individual components to escape from the shrinkage. Xie (2018) first explored the use of this prior in spectral analysis of univariate stationary time series. Here we extend this approach to develop a novel model and related inference procedure for accurate and computationally efficient Bayesian multivariate spectral analysis.

The proposed prior distribution structure is as follows. First, we impose no shrinkage on $\gamma_{j,0}$ and $\gamma_{j,1}$ and assume that $\gamma_{j,s} \sim N(0, 10)$ for all j and $s = 0, 1$. This is based on the idea that the individual spectral densities will have at the very least a baseline basis representation that is non-zero on the intercept and slope terms (e.g., the spectral density of a white noise process is a linear constant function over the frequencies), and possibly a more sophisticated structure that can be captured by the regularized horseshoe prior on the remaining parameters as explained below. We have found that a prior variance of 10 (or any value within $[10, 10^5]$ that leads to a non-informative prior) for these terms typically works well in practice. Then, for $s = 2, \ldots, M$ we assume $\gamma_{j,s} \mid \lambda_{js}, \tau_j \sim \mathrm{N}\left(0, \frac{c^2\tau_j^2\lambda_{js}^2}{c^2+\tau_j^2\lambda_{js}^2}\right)$, where c is a constant. Piironen and Vehtari (2017) recommend placing an inverse-gamma prior on c. However, placing a prior distribution on c breaks the possibility of achieving parallel posterior computation since c is shared for all j. Hence, in order to maintain computational efficiency we fix c and discuss the selection of c below. Similarly, for $s = 0, \ldots, M$, $\alpha_{jl,s} \mid \lambda_{jls,(re)}, \tau_{jl} \sim \mathrm{N}\left(0, \frac{c^2\tau_{jl}^2\lambda_{jls,(re)}^2}{c^2+\tau_{jl}^2\lambda_{jls,(re)}^2}\right)$, and $\beta_{jl,s} \mid \lambda_{jls,(im)}, \tau_{jl} \sim \mathrm{N}\left(0, \frac{c^2\tau_{jl}^2\lambda_{jls,(im)}^2}{c^2+\tau_{jl}^2\lambda_{jls,(im)}^2}\right)$. The parameters τ_j and τ_{jl} above control the overall roughness of the spectral density matrix for each component j as a function of the frequency. We assume $\tau_j, \tau_{jl} \sim \mathrm{C}^+(0, c_\tau)$, where C^+ denotes Half-Cauchy distribution and c_τ is a fixed constant. We also have local parameters λ_{js}, $\lambda_{jls,(re)}$ and $\lambda_{jls,(im)}$ which adjust the roughness as a function of the frequency according to individual Demmler-Reinsch basis with $\lambda_{js}, \lambda_{jls,(re)}, \lambda_{jls,(im)} \sim \mathrm{C}^+(0, c_s)$. Here, c_s is a pre-specified "discount effect" function varying with s. The choices of c_τ and c_s are described below.

The prior structure above has the following features. First, as it was previously mentioned, it assumes that the diagonal entries of a spectral density matrix are all non-zero, and so it does not impose a shrinkage prior on $\gamma_{j,0}$ and $\gamma_{j,1}$. On the other hand, a global-local shrinkage prior is proposed on the rest of $\gamma_{j,s}$'s.

Second, the sparsity of $\mathbf{f}_k^{-1}$'s is directly controlled by θ_{jl}, whose real and imaginary parts are modeled by splines with coefficients $\boldsymbol{\alpha}_{jl}$ and $\boldsymbol{\beta}_{jl}$ accordingly. To detect sparsity patterns on the off-diagonal entries in the spectral matrix, group shrinkage effects are proposed on each pair of $\{\boldsymbol{\alpha}_{jl}, \boldsymbol{\beta}_{jl}\}$ by letting every coefficient pair, $\alpha_{jl,s}$ and $\beta_{jl,s}$, share the same global parameter, τ_{jl}. In practice, the lack of association between the components of the multivariate time series will be inferred in terms of the shrinkage of the corresponding estimates of the $\{\boldsymbol{\alpha}_{jl}, \boldsymbol{\beta}_{jl}\}$ pairs towards zero. We have found results to be robust to the choice of hyper-parameters, with $c \in (1, 10^5)$ and $c_\tau \in (10^{-5}, 10^{-1})$ providing indistinguishable estimates in all empirical examples considered. Thus, we simply set $c = 2$ and $c_\tau = 0.01$ by default.

Third, the discount effect function c_s is used to heuristically gain model flexibility. Previous studies do not provide a general criterion on how to optimally choose the number of basis, M. Rosen et al. (2009, 2012); Zhang (2016); Krafty et al. (2017) and Zhang (2019) suggest using $M = 10$. However, such choice is based on their specific prior structure and simulation studies, which lacks generalization and may not be accurate in some settings. We want our approach to incorporate more basis functions when needed in order to preserve model flexibility in general situations. The cosine bases, $\psi_s(\nu_k)$, allow for modeling more volatile behavior with larger s, which is generally less likely to appear in practical settings. Therefore, in addition to the regularization obtained from the regularized horseshoe prior above, we further penalize the appearance of bases according to the value of s. Coefficients of basis terms with larger s will be more likely to be shrunk towards zero, unless the need to include them in the model is strongly supported by the data. In this way, one can simply choose a relatively large M and the proposed prior will then automatically preserve the informative bases and will shrink towards zero those that are not needed. Inspired by the shape of the sigmoid function $\mathrm{Sig}(x) = [1 + \exp(-x)]^{-1}$, we choose the form of c_s as $c_s = \mathrm{Sig}(-as + b)$ such that c_s is a monotone decreasing function of s and is bounded between 0 and 1. Here a and b are fixed constants defining the shape of the function. a determines how fast the shrinkage level grows. If a is small, c_s decreases slowly, meaning the shrinkage level grows slowly. On the other hand, a larger a indicates a faster growth of the shrinkage level. Meanwhile, b controls when the growth starts. A larger b causes later decrease of c_s, or equivalently, later growth of the shrinkage level. We proposed a default choice of $a = 1$ and $b = M/2$. This choice is justified through experimentation via simulation studies and real data analyses.

We refer to the prior structure above as the discounted regularized horseshoe (DRH), and the DRH model to the Bayesian smoothing splines model for multivariate spectral analysis that assumes the DRH prior. We show later via simulation studies that DRH can easily handle more than $M = 30$ bases, which results in much richer representations than those considered in previous studies. It should be noted that this notion of increasing the penalization as M increases was also used by Li and Krafty (2019); Li et al. (2021). The key difference is that our approach is based on a regularized horseshoe prior structure while that of Li and Krafty (2019) uses normal priors without a global-local shrinkage structure, and the prior in Li et al. (2021) is designed for a stationary factor model representation of the spectral density matrix that is totally different from the modified complex Cholesky representation used here. Section 4 includes a comparison between our approach and other currently available approaches in extensive simulation studies. In summary, the DRH joint prior for the j-th likelihood component can be written as:

$$\pi^{(j)}_{\text{DRH}}(\cdot) = \prod_{s=0}^{1} p(\gamma_{j,s}) \prod_{s=2}^{M} p(\gamma_{j,s} \mid \lambda_{js}, \tau_j)\, p(\lambda_{js})\, p(\tau_j) \prod_{l=1}^{j-1} \prod_{s=0}^{M} p(\alpha_{jl,s} \mid \lambda_{jls,(re)}, \tau_{jl})\, p(\lambda_{jls,(re)}) \prod_{l=1}^{j-1} \prod_{s=0}^{M} p(\beta_{jl,s} \mid \lambda_{jls,(im)}, \tau_{jl})\, p(\lambda_{jls,(im)})\, p(\tau_{jl}). \tag{7}$$

We further take the log transformation on all λ and τ to satisfy the positive constraint of these parameters.

3. Posterior computation

In this section, we describe our approach to adapting variational Bayes techniques to obtain posterior inference of the proposed DRH spectral model. We take advantage of the power of modern computational resources, such as graphic processing units (GPUs), in order to provide a flexible, scalable and computationally efficient inference scheme for high-dimensional time series. Our posterior inference scheme is based on the stochastic gradient variational Bayes (SGVB) approach (Kingma and Welling, 2013). As explained above, the likelihood and priors can be decomposed into p components so we consider inference on each component in parallel. Let $\mathbf{v}_j$ denote all the model parameters for the jth component. The joint log density for the jth component can be written as:

$$\log p(\mathbf{v}_j, \mathbf{y}_{1:N}) = \log L_j\left(\mathbf{y}_{1:N}\,;\, \mathbf{v}_j\right) + \log \pi(\mathbf{v}_j), \tag{8}$$

where $\log L_j(\cdot)$ is defined in (6). $\pi(\cdot)$ is the proposed prior in (7). The true posterior is given by $p(\mathbf{v}_j \mid \mathbf{y}_{1:N}) = p(\mathbf{v}_j, \mathbf{y}_{1:N}) / p(\mathbf{y}_{1:N})$, with $p(\mathbf{y}_{1:N})$ an intractable normalizing constant, thus $p(\mathbf{v}_j \mid \mathbf{y}_{1:N})$ also becomes intractable. The goal of variational inference is to find a surrogate distribution that is most similar to the true posterior. Let $q_{\boldsymbol{\phi}_j}(\mathbf{v}_j)$ denote the surrogate distribution of $p(\mathbf{v}_j \mid \mathbf{y}_{1:N})$, where $\boldsymbol{\phi}_j$ are unknown learnable parameters. The so-called evidence lower bound (ELBO) between $q_{\boldsymbol{\phi}_j}(\mathbf{v}_j)$ and $p(\mathbf{v}_j \mid \mathbf{y}_{1:N})$ is defined as $\mathcal{L}\left(p_j, q_{\boldsymbol{\phi}_j}\right) = \mathrm{E}_{\mathbf{v}_j \sim q_{\boldsymbol{\phi}_j}(\mathbf{v}_j)}\left[\log p(\mathbf{v}_j, \mathbf{y}_{1:N}) - \log q_{\boldsymbol{\phi}_j}(\mathbf{v}_j)\right]$. The goal of variational inference is to find the values of $\boldsymbol{\phi}_j$ that maximize ELBO.

We consider $q_{\boldsymbol{\phi}_j}(\mathbf{v}_j) = \mathrm{N}(\mathbf{v}_j \mid \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$ as the default surrogate distribution for $p(\mathbf{v}_j \mid \mathbf{y}_{1:N})$. $\boldsymbol{\Sigma}_j$ is a diagonal matrix with $\mathrm{diag}(\boldsymbol{\Sigma}_j) = \boldsymbol{\sigma}_j^2$. This is one of the most popular choices of a surrogate distribution for SGVB, since the diagonal structure of the variance-covariance matrix significantly simplifies computations. We show in experiments that the SGVB approximate inference works well in terms of covering the underlying true spectra. Having said this, the proposed framework is friendly to user-specified surrogate distributions, meaning that we can consider more sophisticated choices if needed in practice. For instance, the Gaussian variational approximation with a factor covariance structure (Ong et al., 2018) is an alternative that we have also implemented and that leads to improved uncertainty quantification in low-dimensional and moderate-dimensional practical settings as illustrated in Section 4.1.

Next, to improve run-time convergence, we utilize a reparameterization (Kingma and Welling, 2013) on $\mathbf{v}_j$ to obtain more accurate estimates of the gradient of the variational objective (Xu et al., 2019; Domke, 2019). More specifically, we set $\mathbf{v}_j = \boldsymbol{\mu}_j + \boldsymbol{\sigma}_j \odot \boldsymbol{\epsilon}_j$, such that $\boldsymbol{\epsilon}_j \sim \mathbf{N}(\mathbf{0}, \mathbf{I})$ where $\mathbf{I}$ denotes the identity matrix, and $\odot$ denotes element-wise product. To guarantee that the components $\boldsymbol{\sigma}_j^2$ are positive, we take the log transformation, i.e., $\boldsymbol{\zeta}_j = \log \boldsymbol{\sigma}_j^2$. Accordingly, $\boldsymbol{\phi}_j = \{\boldsymbol{\mu}_j, \boldsymbol{\zeta}_j\}$, and the ELBO between $p(\mathbf{v}_j \mid \mathbf{y}_{1:N})$ and $q_{\boldsymbol{\phi}_j}(\mathbf{v}_j)$ can be written as:

$$\mathcal{L}\left(p_j, q_{\boldsymbol{\phi}_j}\right) = \mathrm{E}_{\boldsymbol{\epsilon}_j \sim \mathbf{N}(\mathbf{0}, \mathbf{I})}\left[\log p(\boldsymbol{\mu}_j + \exp\frac{\boldsymbol{\zeta}_j}{2} \odot \boldsymbol{\epsilon}_j\,, \mathbf{y}_{1:N}) - \log q_{\boldsymbol{\phi}_j}(\boldsymbol{\mu}_j + \exp\frac{\boldsymbol{\zeta}_j}{2} \odot \boldsymbol{\epsilon}_j)\right]. \tag{9}$$

With one sample $\boldsymbol{\epsilon}_j^{(1)} \sim \mathbf{N}(\mathbf{0}, \mathbf{I})$, the SGVB gradient estimator of (9) is given by:

$$\nabla_{\boldsymbol{\phi}_j} \mathcal{L}\left(p_j, q_{\boldsymbol{\phi}_j}\right) \simeq \nabla_{\boldsymbol{\phi}_j} \log p(\boldsymbol{\mu}_j + \exp\frac{\boldsymbol{\zeta}_j}{2} \odot \boldsymbol{\epsilon}_j^{(1)}\,, \mathbf{y}_{1:N}) - \nabla_{\boldsymbol{\phi}_j} \log q_{\boldsymbol{\phi}_j}(\boldsymbol{\mu}_j + \exp\frac{\boldsymbol{\zeta}_j}{2} \odot \boldsymbol{\epsilon}_j^{(1)}). \tag{10}$$

We found that (10) led to a fast convergence in our experiments. More discussions on the convergence of SGVB in our studies are included in supplementary materials Section S.1. Alternatively, in the Gaussian variational approximation with a factor covariance structure of Ong et al. (2018), the surrogate posterior for a m-dimensional parameter vector $\mathbf{v}$ is defined as $\mathbf{v} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma} = \mathbf{B}\mathbf{B}^T + \mathbf{I}\boldsymbol{\sigma}^2$. $\mathbf{B}$ is a $m \times q$ full rank matrix with $q << m$. The reparameterization trick for the factor covariance structure can be conducted by calculating $\mathbf{v} = \boldsymbol{\mu} + \mathbf{B}\boldsymbol{\xi} + \boldsymbol{\sigma} \odot \boldsymbol{\epsilon}$ with $\boldsymbol{\xi} \sim \mathbf{N}(\mathbf{0}, \mathbf{I}_q)$ and $\boldsymbol{\epsilon} \sim \mathbf{N}(\mathbf{0}, \mathbf{I}_m)$, where $\mathbf{I}_q, \mathbf{I}_m$ denote the q, m-dimensional identity matrices. The factor covariance structure allows the variational Gaussian approximation to also capture dependency among the parameters, however, estimating $\mathbf{B}$ can be computational intense with longer run time and more parameter storage when m and q are large. A comparison between the aforementioned two surrogate choices is presented in Section 4.1.

In practice, we propose the so-called three-phase Variational Bayes (TPVB) for posterior inference, as illustrated in Algorithm 1, to stabilize the inference process. Note that Algorithm 1 presents the version with the default Gaussian surrogate distribution with a diagonal covariance matrix, but this algorithm can be easily modified to incorporate the surrogate distribution with a factor covariance structure. In particular, Phase 1 uses (8) as the objective function to obtain a point approximation, $\tilde{\mathbf{v}}_j = \arg\max_{\mathbf{v}_j} \log p(\mathbf{v}_j, \mathbf{y}_{1:N})$ by gradient ascent. Phase 2 fixes $\boldsymbol{\mu}_j = \tilde{\mathbf{v}}_j$ and only updates $\boldsymbol{\zeta}_j$ by SGVB to maximize (9). Finally, both previously updated $\boldsymbol{\mu}_j$ and $\boldsymbol{\zeta}_j$ are fine-tuned at Phase 3 by SGVB to maximize (9). Results from our extensive simulation studies below show that our proposed TPVB achieves high numerical stability. We built the model and implemented the experiments in `Python 3.7` with `Tensorflow-Probability` and `Tensorflow` (Abadi et al., 2015) packages. All gradients are computed via auto-differentiation modules in `Tensorflow`. At each iteration, we use `Adam` (Kingma and Ba, 2017) to update model parameters via gradient ascent.

Algorithm 1: Three-Phase Variational Bayes (TPVB) Posterior Inference.

Input : DFT transformed observations of p-dimensional time series, $\mathbf{y}_{1:N}$.
Output: variational posteriors $q_{\boldsymbol{\phi}_j}(\mathbf{v}_j)$, $j = 1, ..., p$.
Params: (parameters) $\boldsymbol{\phi}_1, ..., \boldsymbol{\phi}_p$, with $\boldsymbol{\phi}_j = \{\boldsymbol{\mu}_j, \boldsymbol{\zeta}_j\}$, $j = 1, ..., p$.

for $j \in \{1, ..., p\}$ **do**
  Phase 1–Point Approximation: maximizes (8) w.r.t $\mathbf{v}_j$, **while** *in iteration* **do**
    Compute the gradients of $\mathbf{v}_j$;
    Update $\mathbf{v}_j$ via gradient ascent;
  end
  Fix $\boldsymbol{\mu}_j$ equal to the last updated $\mathbf{v}_j$ in Phase 1.
  Phase 2–Uncertainty Quant: maximize (9) w.r.t $\boldsymbol{\zeta}_j$, **while** *in iteration* **do**
    Draw $\boldsymbol{\epsilon}_j \sim \mathbf{N}(\mathbf{0}, \mathbf{I})$ and compute the SGVB gradients of $\boldsymbol{\zeta}_j$ using (10);
    Update $\boldsymbol{\zeta}_j$ via gradient ascent;
  end
  Phase 3–Fine-tuning (Optional): **while** *in iteration* **do**
    Draw $\boldsymbol{\epsilon}_j \sim \mathbf{N}(\mathbf{0}, \mathbf{I})$ and compute the SGVB gradients of both $\boldsymbol{\mu}_j$ and $\boldsymbol{\zeta}_j$ using (10);
    Update $\boldsymbol{\mu}_j$ and $\boldsymbol{\zeta}_j$ via gradient ascent;
  end
end

4. Simulation studies

In this section, we show the accuracy and efficiency of our methods through simulations. In all cases below, for all smoothing splines, we set the number of basis to $M = 30$. This provides much richer representations in comparison to those used in previous studies where $M = 10$ regardless of the time series length or their underlying spectral complexity. We show that the DRH spectral model can effectively handle a large M and lead to accurate spectral inference for various spectral density shapes. During the TPVB inference process, the learning rate is used as a hyper-parameter to control the rate at which the algorithm updates the parameter estimates. The learning rates in `Adam` optimizers are set to be 0.0005, 0.05 and 0.005, respectively for each phase, to obtain gradient descent updates for trainable parameters. We decide the number of iterations by monitoring the values of the objective function, and stop the training when a typical convergent pattern is reached. That leads to approximately 5000, 500, and 500 iterations required for each of the 3 phases. The convergence of the TPVB algorithm is monitored via plots of the log-posterior and the evidence lower bound, as shown in for in Section S.1 of the Supplementary Material. The simulation experiments are executed on a x64-based PC with a 2.60-GHz Intel® Core™ i7-9750H CPU and a Nvidia GeForce GTX 1660 Ti GPU card. We illustrate the accurate performance of the proposed framework for both low and large dimensional spectral analysis, provide comparisons to other available approaches, and discuss model scalability.

4.1. Simulation study 1

In this study, we show that the inference obtained by the proposed TPVB is an excellent approximation to the full posterior inference obtained from Markov Chain Monte Carlo (MCMC) methods. We simulate 100 datasets containing $\mathbf{x}_t = (x_{1,t}, x_{2,t})'$, for $t = 1, \cdots, 1024$, from the following bivariate vector autoregressive–moving-average VARMA(2,2) process:

$$\mathbf{x}_t = \begin{bmatrix} 0.2 & 0.5 \\ 0 & -0.2 \end{bmatrix} \mathbf{x}_{t-1} + \begin{bmatrix} 0 & 0 \\ 0.5 & -0.2 \end{bmatrix} \mathbf{x}_{t-2} + \mathbf{w}_t + \begin{bmatrix} 0.6 & 0 \\ 0.2 & -0.5 \end{bmatrix} \mathbf{w}_{t-1} + \begin{bmatrix} 0.3 & 0 \\ 0 & 0.3 \end{bmatrix} \mathbf{w}_{t-2}, \tag{11}$$

where $\mathbf{w}_t$ are independent zero-mean bivariate Gaussian random variables whose components have unit variance and pairwise correlation 0.8. The average squared error (ASE) was used to numerically summarize the performance of the estimators of the spectral components. The ASE of a given posterior spectral estimator is obtained by averaging squared errors across

Table 1
Means (standard deviations) of the run time and ASE values of posterior estimates for spectral components obtained from the DRH model using TPVB and HMC inference schemes with 10000 posterior samples based on 100 simulated datasets of the VARMA process (11). In TPVB, the Gaussian variational surrogate distributions with diagonal covariance (default) and factor covariance structures under $q = 5, 10$ are compared.

	$\boldsymbol{f}_{11}$	$\boldsymbol{f}_{22}$	ρ^2_{12} ($\times 10^3$)	Runtime [sec]
HMC	0.44 (0.37)	0.28 (0.23)	0.26 (0.18)	382 (65)
TPVB (q=10)	0.44 (0.36)	0.29 (0.21)	0.27 (0.16)	72 (16)
TPVB (q=5)	0.46 (0.32)	0.28 (0.18)	0.26 (0.12)	62 (15)
TPVB (default)	0.45 (0.31)	0.27 (0.16)	0.26 (0.10)	46 (12)

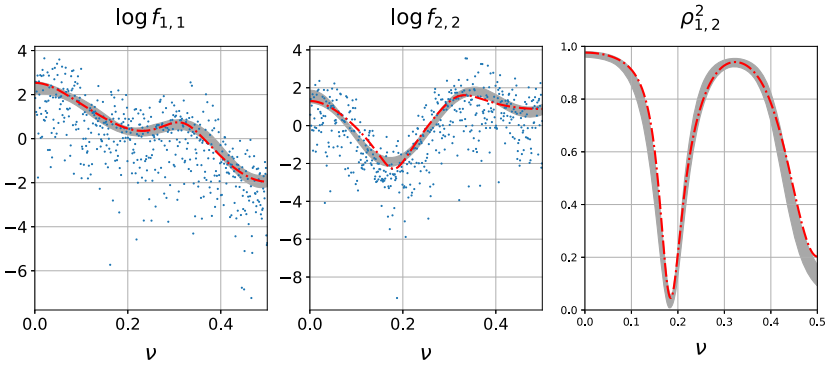

Fig. 1. Posterior inference of log spectral densities and squared coherence for the VARMA process (11) given by the DRH TPVB framework. Grey regions are 95% posterior intervals. Lines are true values. Dots represent log periodogram.

an equally spaced grid of 500 frequencies. For example, the ASE of $\hat{f}_{11}(\nu)$, a posterior mean estimator of $f_{11}(\nu)$, is computed as: $\text{ASE} = \frac{1}{500}\sum_{k=1}^{500}\left[\hat{f}_{11}(\nu_k) - f_{11}(\nu_k)\right]^2$, with $\nu_k = k/1000$.

For each of the 100 datasets, we fit the same DRH model, with the same prior structure, and compare the posterior estimates given by Hamiltonian Monte Carlo (HMC; Neal, 2011), TPVB with the default Gaussian surrogate having a diagonal covariance structure (default) defined in Section 3, and TPVB with Gaussian surrogate having a factor covariance structure (Ong et al., 2018) defined in Section 3 with $q = 5$ and $q = 10$ factors.

We averaged the ASEs for each inference scheme over the 100 datasets to obtain means and standard deviations of ASEs for the two spectral densities, $\boldsymbol{f}_{11}$ and $\boldsymbol{f}_{22}$, and for the squared coherence, ρ^2_{12}. Table 1 shows the ASE and run time results obtained by all the methods. In general, we note that all three TPVB approximations slightly underestimate the posterior variability, as expected, however, the results are pretty similar to those obtained under the HMC scheme. The incorporating a factor covariance structure in the variational surrogate distribution mitigates the uncertainty underestimation issue in variational inference, with posterior variability closer to HMC as q increases. We also compared the GPU running time for all schemes. TPVB (default) takes on average 42 seconds to run per dataset. By utilizing the factor covariance structure in the variational surrogate distribution, the average run time for TPVB increases to 62 seconds when $q = 5$, and 72 seconds when $q = 10$, indicating that even though such factor covariance structure provides more accurate estimates on posterior variability, it also requires higher computation costs than the default choice. Note that, the run time of TPVB does not significantly change by the posterior sample size. The run time for HMC, however, grows with respect to the required posterior sample size. On average, HMC takes roughly 380 seconds per dataset to collect ten thousand posterior samples. Given that we do not know in advance how many posterior samples are required, and usually more than ten thousand samples are required for HMC convergence, TPVB ends up being much faster in practice.

Fig. 1 displays the posterior estimates for the individual spectral densities and the coherence given by TPVB (default). It can be seen that the estimates approximate the true curves smoothly. In comparison to the factor covariance structure, TPVB (default) is more computational efficient, and preserves similar accuracy regarding to power spectrum estimates (similar mean ASEs). Considering that the main focus of this paper is the high-dimensional model scalability, we focus on the TPVB (default) in the following sections, noting that the variational Bayeses approximation can be improved by using the factor covariance structure in the surrogate distribution.

Similar results in terms of the accuracy of the variational approximations were obtained in another simulation study that considers a VAR(2) bivariate structure (see supplementary materials S.3 for details).

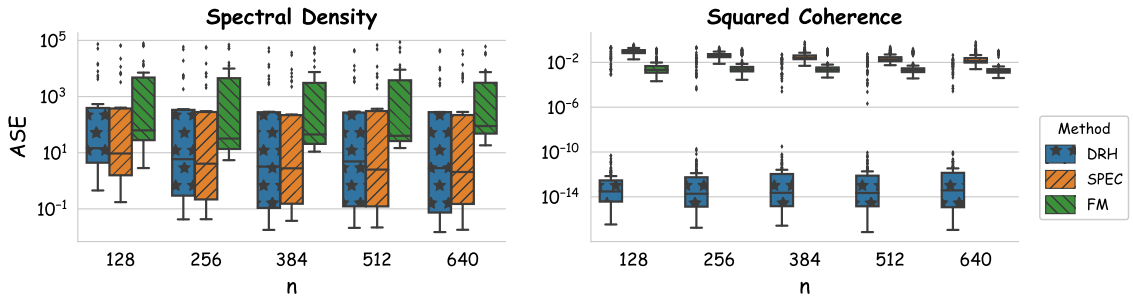

Fig. 2. Log-scaled boxplots of averaged squared errors (ASEs) of spectral densities (left frame) and squared coherence estimates (right frame) given by our DRH model, the SPEC model, and factor model (FM), all fitted by TPVB for simulated 21-dimensional time series with length n.

4.2. Simulation study 2

The computation efficiency gained by TPVB allows us to analyze higher dimensional time series in the spectral domain. We simulated 20 datasets containing 21-dimensional time series where each multivariate time series was constructed by combining 7 groups of 3-dimensional time series. We generated each group of time series from one of 7 3-dimensional VAR processes of order 3. The explicit data generation procedure is detailed in Section S.2 of the Supplementary Material.

Each of the 20 simulated datasets consists of 21 time series components, $\mathbf{x}_t = (x_{1,t}, x_{2,t}, ..., x_{21,t})'$ for $t = 1:n$, where components within each of the 7 groups allow non-zero coherence, but components that are not in the same group have zero coherence. We considered different scenarios with n set to 128, 256, 384, 512, and 640. For each n, we calculate the ASEs for our DRH model, the SPEC model of Rosen and Stoffer (2007), and the factor model (FM) for stationary time series of Li et al. (2021), all fitted by TPVB on every replication. We set $M = 30$ for all models. Note that, the SPEC model does not consider a global-local regularization in the prior, and instead assigns normal priors to spline coefficients. Unlike DRH and SPEC which model the modified complex Cholesky factorized components of the inverse spectral matrix, the FM tries to learn a low-rank decomposition of the spectral matrix.

Considering that there are 21 spectral densities and 210 squared coherences in total, it is impractical to list all the results individually. Instead, we visualize the ASEs for all the spectral densities and squared coherences in the two boxplots shown in Fig. 2. This figure presents the boxplots of ASEs varying by the length n of the time series. It can be seen that in comparison to SPEC and FM, DRH provides much better estimates of pairwise squared coherences (see right plot in Fig. 2), and still offers competitive results in terms of the inference for spectral densities for all n (left plot in the same figure). To further illustrate the accuracy of the proposed approach, we also provide inference results for some spectral densities and some coherences. Fig. 3 shows DRH inference for the first 8 spectral densities in one of the 21-dimensional datasets of length $n = 640$, while Fig. 4 displays DRH inference for 8 randomly chosen squared coherence terms. In general, we see that the proposed model adequately captures the behavior of the spectral densities and coherences (both zero and non-zero terms). We note that obtaining results using the TPVB approximation for the proposed DRH for a 21-dimensional dataset with a length of 640 observations took less than 9 minutes in the same hardware platform described above.

4.3. Simulation study 3: model scalability for high dimensions

This section discusses model scalability as time series dimension considerably increases ($p > 100$). None of the alternative approaches reviewed in the Introduction can be feasibly scaled to consider spectral analysis of multivariate time series of this dimension. For this study, we independently generated $[p/3]$ times from the 3-dimensional VAR(3) process used in Section 4.2. The resulting $[p/3]$ sets of 3-dimensional times series were combined into a p-dimensional time series, and were modeled jointly. Specific details of the data generation procedure are included in supplementary materials Section S.2. We consider values of $p = 30, 45, 60, 75, 90$ and 105. Fig. 5 shows the boxplots of ASE values for the spectral densities and coherencies for different values of p for datasets of length $n = 1024$. It can be seen that the ASEs for the spectral densities remain in the same range of values as the dimension of the time series increases. For the pairwise squared coherences, the median ASE, indicated by the dark horizontal line in each boxplot, stays mostly constant, but there is slightly more variation as the number of time series increases, which is expected as the number of pairwise comparisons increases from 435 for $p = 30$ to 5,460 for $p = 105$.

To illustrate the time efficiency of the DHR TPVB, we also evaluate how the model runtime is affected by the length, n, and the dimension, p of the multivariate time series. Fig. 6 shows that the runtime increases as n and p increase, but is still quite affordable: our framework can analyze a 105-dimensional time series with 1024 observations in less than 24 mins. Considering that the inference process can use parallel computing and we have so far only used one GPU for computation, the increase of run time for even larger dimensional analyses could be made up by using multiple GPUs, or Tensor Processing Units (TPUs). One of our future goals is developing reliable code for effective multi-GPU utilization.

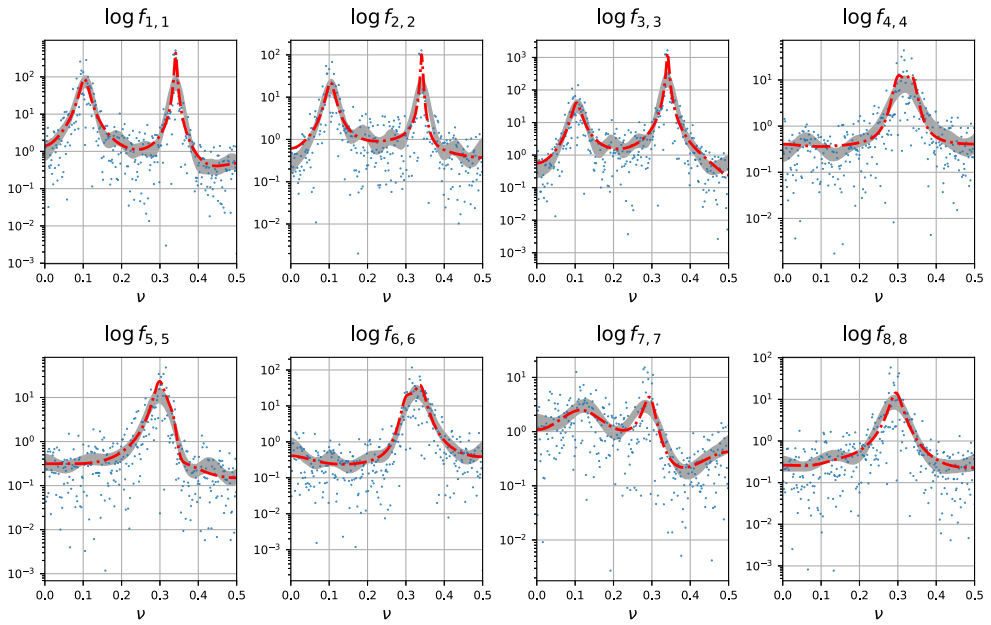

Fig. 3. Posterior inference of the first 8 log spectral densities given by the proposed DRH TPVB framework for a 21-dimensional dataset of length $n = 640$. Grey regions correspond to 95% posterior intervals. Lines are true values. Dots represent log periodograms.

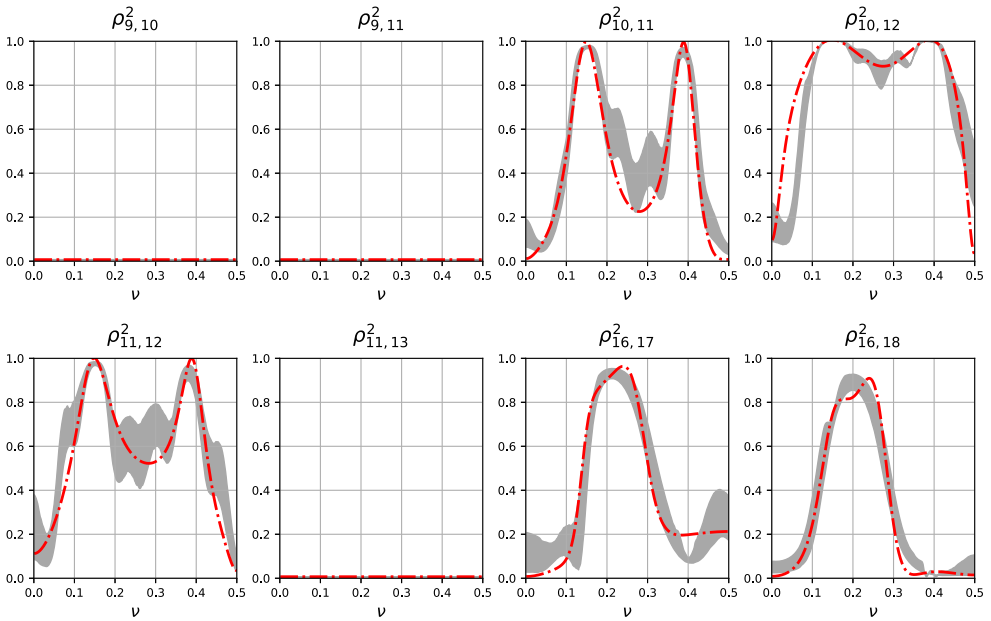

Fig. 4. Posterior inference of 8 randomly chosen squared coherences obtained from the proposed DRH TPVB framework for a 21-dimensional dataset of length $n = 640$. Grey regions are 95% posterior intervals. Lines are true values.

5. Data analysis

5.1. Analysis of California wind profile

We test our method on data from the Iowa State University Environmental Mesonet (IEM) Automated Surface System (ASOS) database (Todey et al., 2002; Mannarano, 1998). This is a publicly available repository of automated airport weather observations and general basic weather reports from the National Weather Service (NWS), the Federal Aviation Administration (FAA) and the Department of Defense (DOD). We consider wind speed data taken from 6 airports in California: EDU (Davis), SAC and SMF (Sacramento), MRY (Monterey), SNS (Salinas), and WVI (Watsonville). Note that EDU, SAC and SMF are

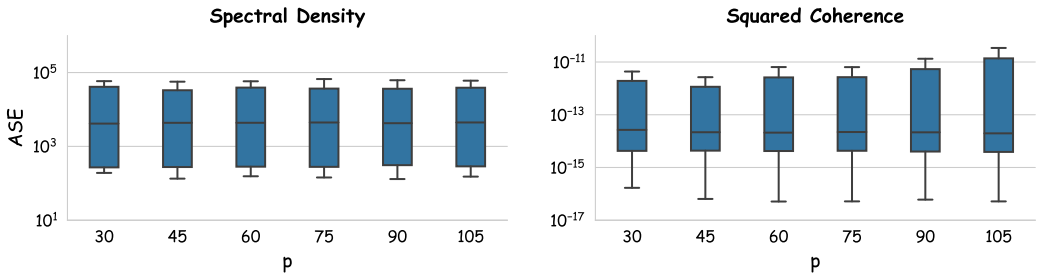

Fig. 5. Log-scaled boxplots of averaged squared errors (ASEs) of spectral densities (left frame) and squared coherences (right frame) given by the DRH TPVB framework fitted on multivariate time series of dimension p, with fixed length $n = 1024$.

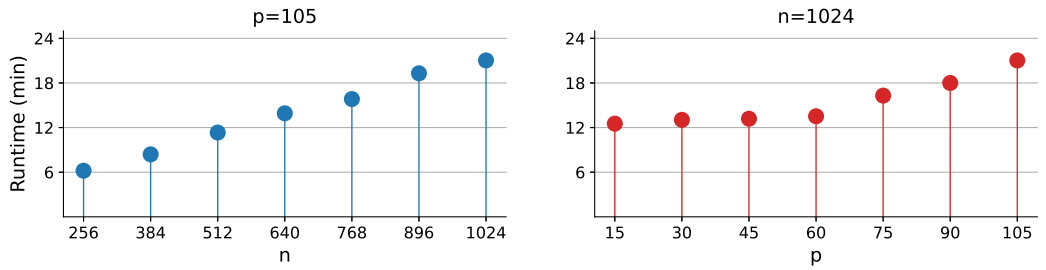

Fig. 6. DRH TPVB framework runtime (in minutes) per dataset, where n denotes length of observed time series, and p denotes the multivariate time series dimension. Left: runtime change by n with $p = 105$. Right: runtime change by p with $n = 1024$.

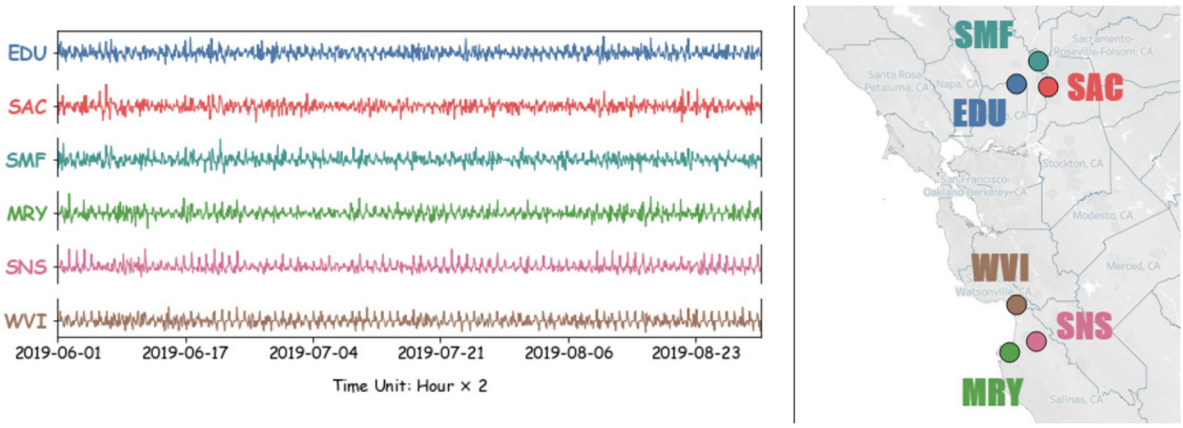

Fig. 7. Standardized first differences of median wind speed measurements every two hours between 06/01/2019 12:00 am and 08/31/2019 11:59 pm from selected California airports.

located in Sacramento area, while MRY, SNS and WVI are located near the Monterey Bay in the Monterey and Santa Cruz counties. Our goal is to infer 6 spectral densities and 15 pairs of squared coherences in order to provide insight into the temporal relationships across different locations over a particular period of time. We consider the median wind speed every 2 hours, starting from 06/01/2019 12:00 am to 08/31/2019 11:59 pm, for each location. Given that our method is based on the assumption of stationarity, we consider a period of time within the summer months to avoid extreme values and non-stationarities related to rainfall and storms that occur in other months. We note that there is essentially no rainfall during this period within these locations. Prior analyses of wind profiles in some of these locations have shown quasi-periodic patterns every 24-hours (Garcia et al., 2020). Here we focus on spectral inference on frequency values that are away from zero, therefore, in order to remove any local trends we jointly analyze 6-dimensional data corresponding to the first order differences for each time series. We further standardized each detrended time series by subtracting its mean and dividing by its standard deviation to enhance computation stability. The resulting 6-dimensional time series data with 1,104 observations along with their locations are shown in Fig. 7. Additional stationarity diagnostics of the differenced time series are included in Section S.4 of the Supplementary Materials. For this analysis we set $M = 30$. The prior hyperparameters and algorithmic settings for the DHR TPVB inference are the same as those discussed in Section 4.

Figs. 8 and 9 show the posterior estimates and 95% posterior intervals given by TPVB under the proposed DHR model. It can be seen that our estimates smoothly follow the trajectories of the observed log periodograms. Note that all the

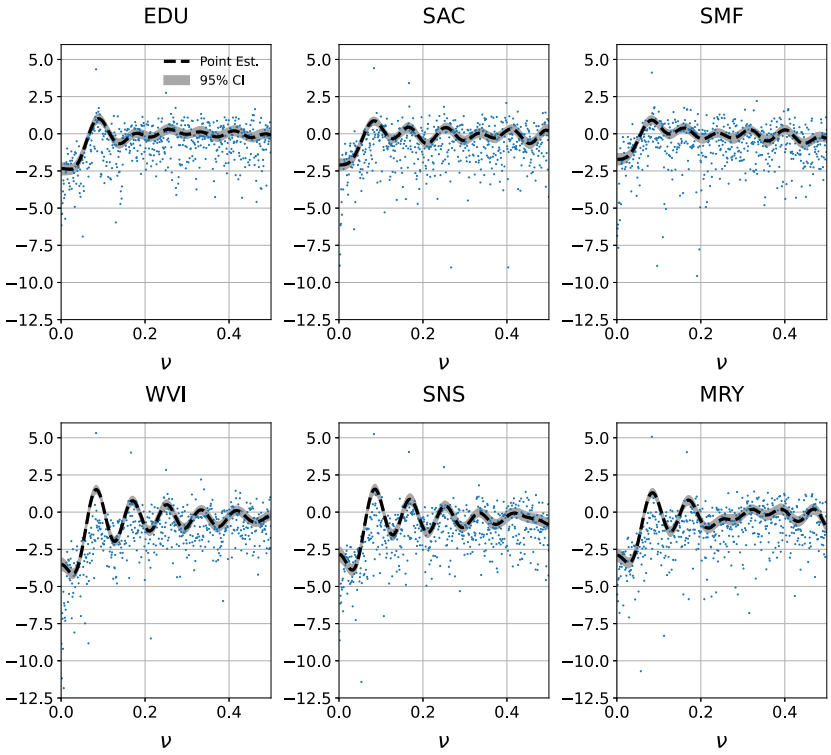

Fig. 8. Results of the wind profile analysis: Posterior inference of log spectral densities given by the proposed DRH TPVB framework. Lines are posterior mean estimates. Grey regions are 95% posterior intervals. Dots represent observed log periodograms.

estimated spectral densities have a peak around frequency 0.08 (approximately 1/12), indicating a strong daily periodic behavior at each location, with higher power around 0.08 for those locations closer to coastal areas (MRY, SNS, WVI). This is consistent with the quasi-periodic nature of the daily wind profile during summer months in such locations (Garcia et al., 2020). Fig. 9 shows that squared coherences are inferred be non-zero only for pairs of sites that are geographically close to each other, i.e., non-zero squared coherences between pairs in the group (EDU, SAC, SMF), and between pairs in the group (MRY, SNS, WVI), but zero coherence between pairs in which the two series are in different groups. This shows that locations that are next to each other share strong coherences in their wind profile patterns at certain frequencies, while this is not the case for locations that are far apart. We also see that the coherence is higher for lower frequencies and, for most pairs, the highest coherence occurs near the 0.08 frequency (daily quasi-periodicity).

5.2. Analysis of multi-channel EEG data

We demonstrate the proposed methodology through the analysis of a multi-channel EEG data set available at the UCI Machine Learning Repository (Dua and Graff, 2017). These data arise from a large study aiming to examine EEG correlates of genetic predisposition to alcoholism. The dataset includes measurements from 61 electrodes placed on the scalp of each participant (Zhang et al., 1995) sampled at 256 Hz for 1 second. The full dataset contains a total of 122 subjects with two groups of subjects, alcoholic subjects and control subjects. Each subject completed 120 trials under different stimuli. During each trial, each subject was exposed to either a single stimulus (S1) or to two stimuli (S1 and S2) which were pictures of objects chosen from the 1980 Snodgrass and Vanderwart picture set (Snodgrass and Vanderwart, 1980).

In order to illustrate the applicability of our spectral methodology in a large-dimensional setting, we consider data from the first trial for each of the subjects under a single stimulus, consisting of two 256-length, 61-dimensional time series, one for an alcoholic subject and another one for a control subject. We removed local trends by taking the 1-st order differences. We fit our proposed DRH multivariate spectral model with $M = 10$ and the same hyperparameters and TPVB inference setting discussed in Section 4 to the two 61-dimensional time series separately. We obtain fast results leading to approximate posterior inference of the spectral density matrix for the two 61-dimensional time series.

We summarize the results in terms of posterior estimates of the spectral densities for the individual channels for the alcoholic and control subjects, and also look at summary measures related to the coherence across channels for a specific frequency band. More specifically, we focus on computing frequency-band collapsed measures at the beta band (16–31 Hz), as this frequency band has been found to provide important neurophysiological information (Ferrarelli et al., 2019; Li et al., 2021). Frequency-band collapsed measures can be computed as integrals of the power spectra (Li et al., 2021). The beta-

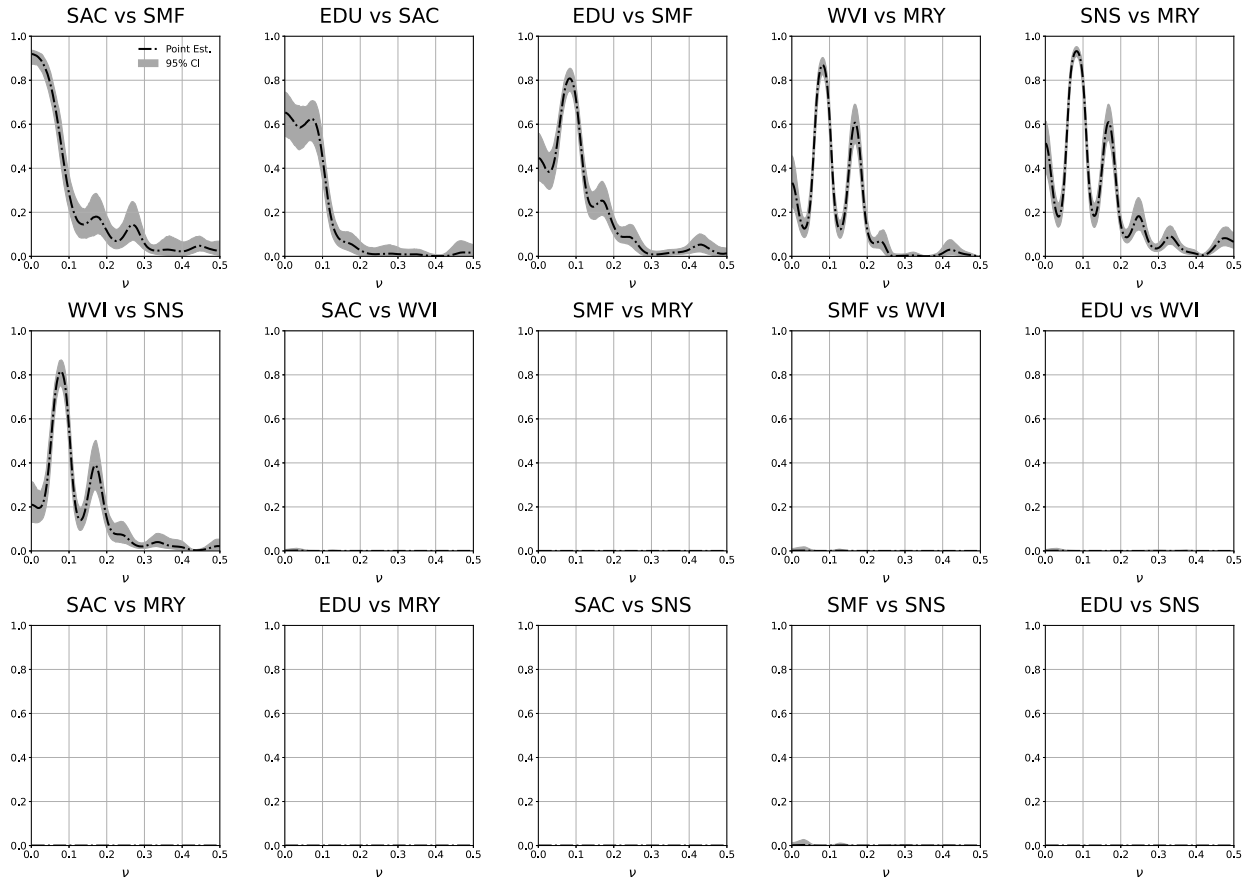

Fig. 9. Result of the wind profile analysis: posterior inference of squared coherences given by the proposed DRH TPVB framework. Lines are posterior mean estimates. Grey regions are 95% posterior intervals.

band collapsed squared coherence between channels j and l is defined as $\rho_{jl}^{2,\beta} = \left|\boldsymbol{f}_{jl}^{\beta}\right|^2 / \left\{\boldsymbol{f}_{jj}^{\beta}\boldsymbol{f}_{ll}^{\beta}\right\}$, where $\boldsymbol{f}^{\beta} = \int_{16}^{31} \boldsymbol{f}(\nu)\, d\nu$ defines beta-band collapsed spectral matrix. Fig. 10 shows the estimated beta-band log spectral densities of 61 channels for alcoholic and control subjects under the same stimulus. Fig. 11 presents the corresponding estimated beta-band squared coherences. We see that, in comparison to the control subject, the alcoholic subject shows a smaller number of non-zero coherences among channels over beta band. Fig. 12 presents the estimated top 20 largest beta-band squared coherences for each of the groups. We see that most of the actively connected channel pairs differ between groups. For instance, the estimated beta-band squared coherence of the channel pair, C1-C3, is relatively high in the alcoholic subject, but essentially zero in the control subject. In contrast, channel pairs such as AF1-AF2 and C4-FC4 are estimated to have large coherences in the control subject, while showing zero coherence in the alcoholic subject. We note that even though many of the large coherence values occur between nearby locations, some non-negligible coherence values occur between channels that are not very close to each other. For instance, it can be seen from Fig. 11 that, for the alcoholic subject, channel CZ has the strongest beta-band coherences with Fp1 and FT7, regardless of the fact that Cz is not in the close proximity of FP1 and FT7. One the other hand, the beta-band coherences between CZ and its nearby channels, such as CP2 and CP4, are very close to zero (for a schematic representation of the channel locations see Section S.5 of the supplementary material).

This analysis illustrates that our proposed spectral method leads to fast spectral inference that can offer helpful insight to neuroscientists that need to analyze large-dimensional brain signals recorded in experimental settings that involve multiple trials and/or subjects.

6. Conclusion and future work

The proposed novel modeling and inference framework provides accurate and computationally fast approximate Bayesian spectral inference for multivariate time series. We showed that the discounted regularized horseshoe prior leads to accurate inference of spectral densities and squared coherences in several simulation settings involving multivariate time series. Furthermore, the proposed inference scheme, that utilizes stochastic gradient variational Bayes, is highly scalable and efficient, providing solid computational support for large-dimensional spectral analysis.

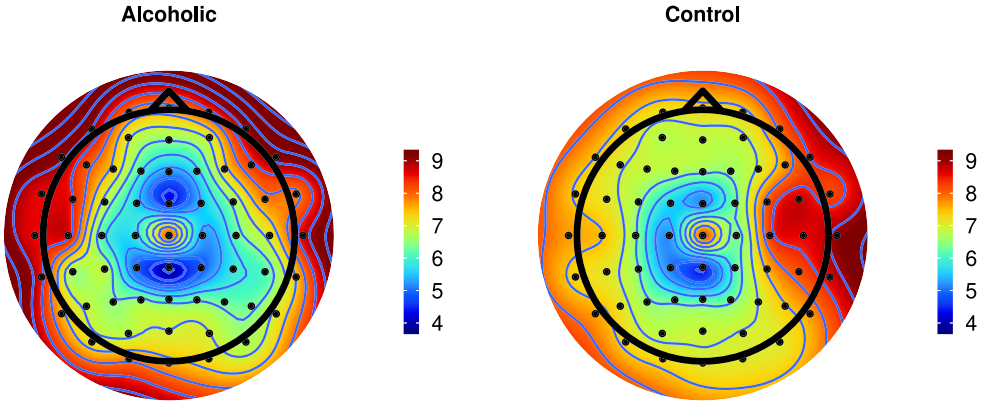

Fig. 10. Result of EEG analysis: posterior estimates of log beta-band collapsed spectral densities for alcoholic (left) and control (right) subjects.

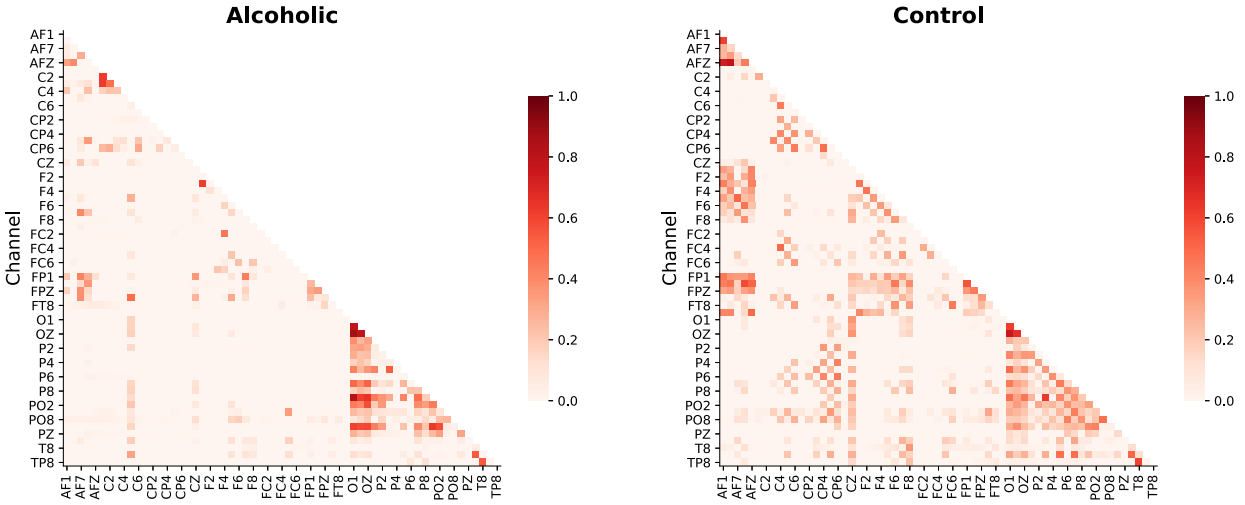

Fig. 11. Result of EEG analysis: posterior estimates of beta-band collapsed squared coherences for alcoholic (left) and control (right) subjects.

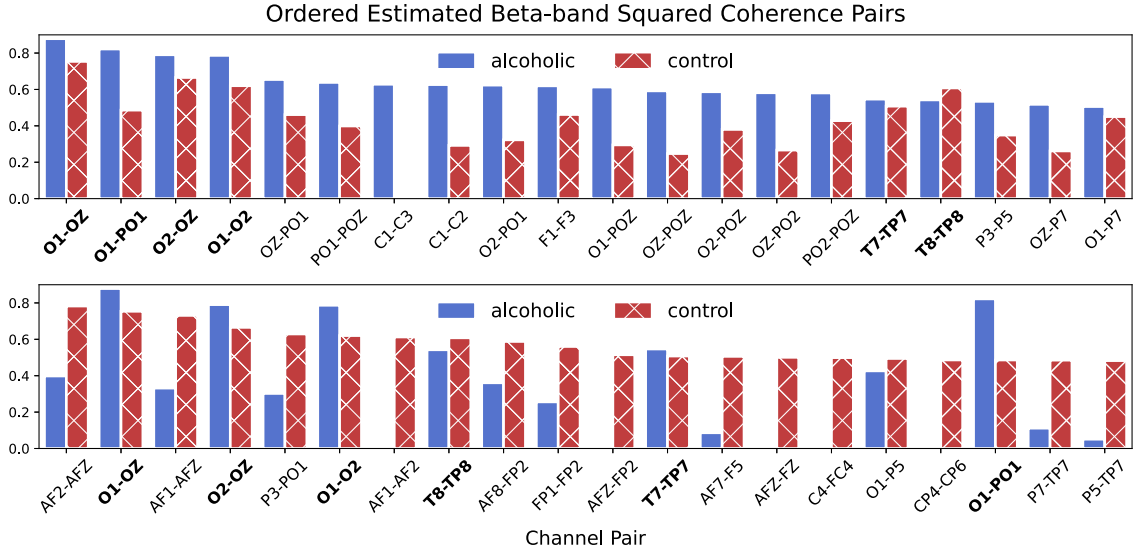

Fig. 12. Top: top-20 largest estimated beta-band collapsed squared coherence pairs in alcoholic group and their contrasts in control group. Bottom: top-20 largest estimated beta-band collapsed squared coherence pairs in control group and their contrasts in alcoholic group. There are only limited common pairs (bold) existing in both top-20 groups.

Future work involves extending the current prior setting. Possible extensions include considering shrinkage priors such as the normal-gamma prior by Griffin and Brown (2010) and its modifications by Huber and Feldkircher (2019), as well as exploring the benefits of introducing additional sparsification procedures similar to those of Huber et al. (2021), where a parameter sparsification step is proposed after obtaining the posterior estimates using global-local shrinkage priors to further reduce storage load and improve performance.

More relevant extensions in terms of increasing the applicability of the proposed modeling framework and related inferential procedures include considering multivariate spectral analysis for non-stationary time series, as well as hierarchical model formulations that allow us to jointly analyze multi-trial data.

Funding

Professor Raquel Prado was partially funded by NSF award SES 1853210.

Appendix A. Supplementary material

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.csda.2022.107596.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., Zheng, X., 2015. TensorFlow: large-scale machine learning on heterogeneous systems. https://www.tensorflow.org/. Software available from tensorflow.org.

Blei, D.M., Kucukelbir, A., McAuliffe, J.D., 2017. Variational inference: a review for statisticians. J. Am. Stat. Assoc. 112, 859–877.

Choudhuri, N., Ghosal, S., Roy, A., 2004. Bayesian estimation of the spectral density of a time series. J. Am. Stat. Assoc. 99, 1050–1059.

Dai, M., Guo, W., 2004. Multivariate spectral analysis using Cholesky decomposition. Biometrika 91, 629–643.

Domke, J., 2019. Provable gradient variance guarantees for black-box variational inference. Adv. Neural Inf. Process. Syst. 32.

Dua, D., Graff, C., 2017. UCI machine learning repository. School of Information and Computer Sciences, University of California, Irvine. http://archive.ics.uci.edu/ml. Available online.

Eubank, R.L., 1999. Nonparametric Regression and Spline Smoothing. CRC Press.

Ferrarelli, F., Kaskie, R.E., Graziano, B., Reis, C.C., Casali, A.G., 2019. Abnormalities in the evoked frontal oscillatory activity of first-episode psychosis: a TMS/EEG study. Schizophr. Res. 206, 436–439.

Garcia, I., Huo, S., Prado, R., Bravo, L., 2020. Dynamic Bayesian temporal modeling and forecasting of short-term wind measurements. Renew. Energy 161, 55–64. https://doi.org/10.1016/j.renene.2020.05.182.

Griffin, J.E., Brown, P.J., 2010. Inference with normal-gamma prior distributions in regression problems. Bayesian Anal. 5, 171–188.

Huber, F., Feldkircher, M., 2019. Adaptive shrinkage in Bayesian vector autoregressive models. J. Bus. Econ. Stat. 37, 27–39.

Huber, F., Koop, G., Onorante, L., 2021. Inducing sparsity and shrinkage in time-varying parameter models. J. Bus. Econ. Stat. 39, 669–683.

Kingma, D.P., Ba, J., 2017. Adam: a method for stochastic optimization. arXiv:1412.6980.

Kingma, D.P., Welling, M., 2013. Auto-encoding variational Bayes. arXiv:1312.6114.

Krafty, R.T., Collinge, W.O., 2013. Penalized multivariate Whittle likelihood for power spectrum estimation. Biometrika 100, 447–458.

Krafty, R.T., Rosen, O., Stoffer, D.S., Buysse, D.J., Hall, M.H., 2017. Conditional spectral analysis of replicated multiple time series with application to nocturnal physiology. J. Am. Stat. Assoc. 112, 1405–1416.

Li, Z., Krafty, R.T., 2019. Adaptive Bayesian time–frequency analysis of multivariate time series. J. Am. Stat. Assoc. 114, 453–465.

Li, Z., Rosen, O., Ferrarelli, F., Krafty, R.T., 2021. Adaptive Bayesian spectral analysis of high-dimensional nonstationary time series. J. Comput. Graph. Stat. 30, 794–807.

Mannarano, D., 1998. Automated Surface Observing System (ASOS) User's Guide. NOAA, DoD, FAA, US Navy. https://www.weather.gov/media/asos/aum-toc.pdf.

Meier, A., Kirch, C., Meyer, R., 2020. Bayesian nonparametric analysis of multivariate time series: a matrix gamma process approach. J. Multivar. Anal. 175, 104560.

Neal, R.M., 2011. MCMC using Hamiltonian dynamics. In: Handbook of Markov Chain Monte Carlo.

Ong, V.M.H., Nott, D.J., Smith, M.S., 2018. Gaussian variational approximation with a factor covariance structure. J. Comput. Graph. Stat. 27, 465–478.

Piironen, J., Vehtari, A., 2017. Sparsity information and regularization in the horseshoe and other shrinkage priors. Electron. J. Stat. 11, 5018–5051.

Robbins, H., Monro, S., 1951. A stochastic approximation method. Ann. Math. Stat. 22, 400–407.

Rosen, O., Stoffer, D.S., 2007. Automatic estimation of multivariate spectra via smoothing splines. Biometrika 94, 335–345.

Rosen, O., Stoffer, D.S., Wood, S., 2009. Local spectral analysis via a Bayesian mixture of smoothing splines. J. Am. Stat. Assoc. 104, 249–262.

Rosen, O., Wood, S., Stoffer, D.S., 2012. Adaptspec: adaptive spectral estimation for nonstationary time series. J. Am. Stat. Assoc. 107, 1575–1589.

Snodgrass, J.G., Vanderwart, M., 1980. A standardized set of 260 pictures: norms for name agreement, image agreement, familiarity, and visual complexity. J. Exp. Psychol. Hum. Learn. Mem. 6, 174.

Todey, D.P., Herzmann, D.E., Takle, E.S., 2002. The Iowa environmental mesonet—combining observing systems into a single network. In: Sixth Symposium on Integrated Observing Systems.

Whittle, P., 1957. Curve and periodogram smoothing. J. R. Stat. Soc., Ser. B, Methodol. 19, 38–47.

Xie, Y., 2018. A Bayesian Model for Spectral Density Estimation. Open Access Theses & Dissertations. University of Texas at El Paso.

Xu, M., Quiroz, M., Kohn, R., Sisson, S.A., 2019. Variance reduction properties of the reparameterization trick. In: The 22nd International Conference on Artificial Intelligence and Statistics. PMLR, pp. 2711–2720.

Zhang, S., 2016. Adaptive spectral estimation for nonstationary multivariate time series. Comput. Stat. Data Anal. 103, 330–349.

Zhang, S., 2019. Bayesian copula spectral analysis for stationary time series. Comput. Stat. Data Anal. 133, 166–179.

Zhang, X.L., Begleiter, H., Porjesz, B., Wang, W., Litke, A., 1995. Event related potentials during object recognition tasks. Brain Res. Bull. 38, 531–538.