VARIATIONAL BAYES LEARNING OF DYNAMIC GRAPHICAL MODELS

Hang Yu and Justin Dauwels

School of Electrical and Electronics Engineering, Nanyang Technological University, Singapore, 639798

ABSTRACT

We propose a tuning-free Bayesian approach to learn a set of sparse graphical models, in which adjacent graphs share similar structures. This model can be applied to estimating dynamic networks that evolve smoothly with regard to (w.r.t.) a covariate (e.g., time). Specifically, a novel structured spike and slab prior is constructed. This prior allows covariate-varying sparsity pattern by smoothing the spike probabilities across the covariate using a Gauss-Markov chain. Efficient variational Bayes (VB) algorithm is then derived to learn the model, and comparisons are made to related frequentist methods in the literature. We further leverage the proposed mechanism to learn graphical models for multivariate time series in the frequency domain. As an example, we analyze scalp electroencephalograms (EEG) recordings of patients at early stages of Alzheimer disease (AD), and show the loss of synchrony in comparison with control subjects.

***Index Terms*—** Dynamic graphical models, structure, sparse Bayesian learning, time series

1. INTRODUCTION

The recent decades have witnessed a rapid development of graphical models, since they provide a refined language to describe complicated systems and further facilitate the derivation of efficient learning and inference algorithms [1]. While an extensive literature revolves around learning sparse graphical models from independent and identically distributed (i.i.d.) data (see [2, 3] and references therein), little has been done towards observations drawn from a distribution that varies with a covariate (e.g. time or space).[1] For instance, during epileptic seizures, functional brain networks are shown to evolve across time through a sequence of distinct topologies [4]. Estimating such networks can show how the dysrhythmia of the brain propagates and may help the treatment of epilepsy.

Existing works on learning covariate-dependent networks can be categorized into three groups. The first one [5]-[8] considers the temporal dependence by smoothing the sufficient statistics across time using kernels. Given the temporally dependent sufficient statistics, the sparse graphical models are then estimated individually at each time point. The estimation problem can be solved by maximizing the likelihood with an ℓ_1 penalty on the graph sparsity. However, such methods are only applicable in the settings where graphs continuously and smoothly evolve over time [9], whereas in practice the smoothness assumption can be broken at certain time points, such as the onset of seizure in the above example. Moreover, unexpected variability may arise between two adjacent networks since each network are estimated independently [10]. To mitigate these issues, the second group of literature [9, 11, 12] model the temporal dependence by enforcing ℓ_1 or ℓ_2 norm constraints on the difference of two consecutive graphs. The resulting problem is usually convex, and can be solved by the alternating directions method of multipliers (ADMM). Finally, the third group combines both two aforementioned approaches [10, 13].

Unfortunately, the dynamic graphical models inferred by all the three methods are sensitive to the tuning parameters, including the kernel bandwidth and the penalty parameters. Typical data-driven methods for selecting these parameters are cross validation (CV) and Akaike information criterion (AIC) [10]. Nevertheless, heavy computational burden comes along with these methods; the learning algorithm needs to be run once for every possible value of the parameters in a predefined candidate set (which may be large) before the one associated with the largest AIC (or CV) score is chosen. Moreover, it has been demonstrated in [14] that neither CV nor AIC yields satisfying results for graphical model selection.

In this paper, we take a Bayesian approach to construct graphical models varying with a covariate. The major benefit of the proposed algorithm is that all parameters can be estimated in an automatic manner from the data without tuning. In particular, we focus on Gaussian graphical models that characterize the conditional independence (i.e., absence of edges) by zero entries in the precision (inverse covariance) matrix. As a result, our objective is to infer the covariate-varying precision matrix. To this end, we propose a novel structured spike and slab prior. Specifically, each off-diagonal entry of the precision matrix can be factorized as the product of a Bernoulli and a Gaussian distributed variable, and these two variables are further coupled across the covariate via Gauss-Markov chains (i.e., thin-membrane models [15]).

[1] To provide a more readable presentation, we regard time as the covariate unless specified otherwise.

We then develop an efficient VB algorithm to learn the model. Numerical results show that when compared with the frequentist method [11]-[12], the proposed method achieves better performance in terms of estimation accuracy with significantly less amount of computational time.

Interestingly, the proposed model can be directly applied to infer graphical models between multiple time series in a certain frequency band [16]. We therefore investigate the performance of our model on multichannel scalp EEG signals, and clearly show that graphical models for patients at early stages of AD are more sparse than those for healthy control subjects. Note that this effect is not always easily detectable, especially for patients in the pre-symptomatic phase. Thus, the proposed model may offer another useful tool for early detection of AD.

This paper is structured as follows. We first present our Bayesian model in Section 2, and then derive the VB algorithm in Section 3. In Section 4, we show the numerical results for both synthetic and real data. Finally, we close this paper with conclusions in Section 5.

2. DYNAMIC GRAPHICAL MODELS

In this section, we first introduce the proposed structured spike and slab prior, and subsequently construct the Bayesian model for dynamic graphical models. Finally, we explain how to exploit the proposed model to infer graphical models for stationary time series in the frequency domain.

2.1. Structured Spike and Slab Prior

Suppose that $K_{ij}^{1:T}$ is off-diagonal entry (i, j) of the $P \times P$ precision matrix K with covariate from 1 to T. A spike and slab prior on K_{ij}^t can be defined as [17]:

$$K_{ij}^t \sim \eta_{ij}^t \mathcal{N}(\mu_{ij}^t, \nu_{ij}^t) + (1 - \eta_{ij}^t)\delta_0, \tag{1}$$

where $\mathcal{N}(\mu_{ij}^t, \nu_{ij}^t)$ is a Gaussian distribution with mean μ_{ij}^t and variance ν_{ij}^t, δ_0 is a Kronecker delta function, and $\eta_i j^t \in [0, 1]$ determines probability of $K_{ij}^t = 0$ (i.e., the spike probability). By increasing η_{ij}^t to 1, this prior would shrink K_{ij}^t to 0, thus encouraging sparsity in K. The above expression can also be equivalently written as [18]:

$$K_{ij}^t = s_{ij}^t J_{ij}^t \tag{2}$$

$$J_{ij}^t \sim \mathcal{N}(\mu_{ij}^t, \nu_{ij}^t), \tag{3}$$

$$s_{ij}^t \sim \text{Ber}(\eta_{ij}^t), \tag{4}$$

where $\text{Ber}(\eta_{ij}^t)$ is a Bernoulli distribution with successful probability η_{ij}^t. To obtain K^t that changes smoothly with t, we need to impose smoothness priors on both s_{ij}^t and J_{ij}^t. First, let us focus on s_{ij}^t. One tempting approach is to assume $s_{ij}^{1:T}$ forms a binary Markov chain as in [18, 19]. However, as pointed out in [20], the resulting hidden Markov model (HMM) may induce unrealistically rapid switching between states. This problem can be eliminated by introducing a self-transition bias, giving rise to a sticky HMM [20]. Unfortunately, VB learning of the sticky HMM is prone to local maxima [21]. Instead of using binary Markov chains, Ren *et al.* [21] assume that $\eta_{ij}^t = g(\beta_{ij}^t)$, where $g(\cdot)$ is the standard logistic function, and smooth β_{ij}^t across t based on kernels. Although this method is amenable to the VB framework, the bandwidth of the kernels can only be estimated via Monte Carlo methods, thus slowing down the entire learning process. An alternative approach is proposed in [22], in which β_{ij}^t is drawn from a Gaussian process and $g(\cdot)$ is the standard Gaussian cumulative distribution function. The parameters of the Gaussian process, however, have to be defined in advance; otherwise the computational cost of learning these parameters is $\mathcal{O}(T^3)$. Here, we specify $g(\cdot)$ to be the standard logistic function due to its utility in the VB framework [21]. We further assume β_{ij}^t forms a Gauss-Markov chain, in particular, a thin-membrane model [15]. The resulting prior on s_{ij}^t can be expressed as:

$$p(s_{ij}^t|\beta_{ij}^t) = \text{Ber}(s_{ij}^t; g(\beta_{ij}^t)), \tag{5}$$

$$g(\beta_{ij}^t) = \frac{1}{1 + \exp(-\beta_{ij}^t)}, \tag{6}$$

$$p(\beta_{ij}^{1:T}) \propto \lambda^{\frac{T-1}{2}} \exp\Big(-\frac{\lambda}{2} \sum_{t=2}^{T} \big(\beta_{ij}^t - \beta_{ij}^{t-1}\big)^2 \Big), \tag{7}$$

where λ is the smoothness parameter controlling the smoothness of β_{ij}^t across t. We further impose a conjugate Gamma prior on λ:

$$p(\lambda) = \text{Gamma}(\lambda; a_0, b_0) \propto \lambda^{a_0 - 1} \exp(-b_0 \lambda). \tag{8}$$

The parameters a_0 and b_0 are set to be small (e.g., $a_0 = b_0 = 10^{-10}$) such that the prior is non-informative. As λ increases, the thin-membrane model reduces the difference between consectuve β_{ij}^t and β_{ij}^{t-1}, and the resulting prior favors that $s_{ij}^t = s_{ij}^{t-1}$. Furthermore, the shape of logistic function allows piecewise-constant segments in $s_{ij}^{1:T}$. Similarly, to promote the smoonth variation of J_{ij}^t across t, we impose a thin-membrane model prior with smoothness parameter γ on J_{ij}^t, and a Gamma prior $\text{Gamma}(c_0, d_0)$ on γ.

2.2. Covariate-Varying Graphical Models

Let $\boldsymbol{x}_{1:P}^{1:T}$ denote the observations of the Gaussian graphical models from 1 to t. Then,

$$p(\boldsymbol{x}^t|K^t) \propto \det(K^t)^{\frac{1}{2}} \exp\Big(-\frac{1}{2} \boldsymbol{x}^{t\prime} K^t \boldsymbol{x}^t \Big), \tag{9}$$

where $\boldsymbol{x}^{t\prime}$ is the (conjugate) transpose of $\boldsymbol{x}^t$. In our model, to facilitate variational inference, we further relax the equality constraint in Eq. (2) as a Gaussian distribution:

$$p(K_{ij}^t|s_{ij}^t, J_{ij}^t, \alpha) \propto \sqrt{\alpha} \exp\Big(-\frac{\alpha}{2} \big(K_{ij}^t - s_{ij}^t J_{ij}^t\big)^2 \Big).$$

This resembles the Lagrangian multiplier in the frequentist methods [9, 12, 10]. As the learning algorithm proceeds, α will take a very large value, and the above Gaussian distribution degenerates to a Kronecker delta function $\delta_0(K_{ij}^t -$

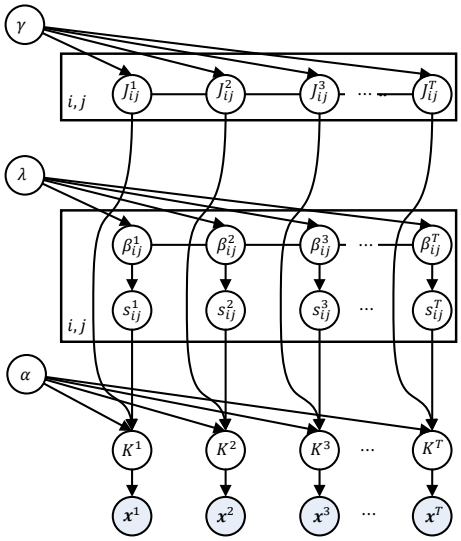


Fig. 1: Bayesian Covariation-Varying Graphical Models

$s_{ij}^t J_{ij}^t$). As before, a conjugate Gamma prior Gamma(e_0, f_0) is put on α. In summary, the overall Bayesian model can be factorized as:

$$p(\boldsymbol{x}^{1:T}, K^{1:T}, J^{1:T}, \boldsymbol{s}^{1:T}, \boldsymbol{\beta}^{1:T}, \alpha, \boldsymbol{\gamma}, \boldsymbol{\lambda})$$
$$= p(\alpha)p(\lambda)p(\gamma)\prod_{t=1}^{t}\Big[p(\boldsymbol{x}^t|K^t)\prod_{i=1}^{P}\prod_{j=i}^{P}p(K_{ij}^t|s_{ij}^t, J_{ij}^t, \alpha)\times$$
$$p(s_{ij}^t|\beta_{ij}^t)\Big]\prod_{i=1}^{P}\prod_{j=i}^{P}\Big[\prod_{t=2}^{T}p(\beta_{ij}^t|\beta_{ij}^{t-1}, \lambda)p(J_{ij}^t|J_{ij}^{t-1}, \gamma)\Big].$$

The graphical model representation of the proposed model is depicted in Fig 1. Note that for diagonal elements in K^t, we assume $s_{ii}^t = 1$ for all t, and the remaining priors are the same with the off-diagonal elements.

2.3. Graphical Models for Time Series

In this section, we discuss how to exploit the proposed model to learn interactions among P univariate stationary Gaussian processes (i.e., time series) $\boldsymbol{x}_{1:P}^t$, where t denotes time. Under this setting, $\boldsymbol{x}^t$ and $\boldsymbol{x}^{t-1}$ are still correlated given K^t and K^{t-1}. As such, the proposed model is not directly applicable to analyzing $\boldsymbol{x}^{1:T}$ in the time domain. Notice that time is not the covariate in this section.

A graphical model $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ for $\boldsymbol{x}^t$ can be constructed by letting an edge $(i, j) \notin \mathcal{E}$ denote that the two entire time series $x_i^{1:T}$ and $x_j^{1:T}$ are conditionally independent given the rest collection of time series $x_{\mathcal{V}|i,j}^{1:T}$, where $\mathcal{V}|i, j$ denotes all nodes in $\mathcal{V}$ except i and j, that is [23]:

$$COV(x_i^t, x_j^{t+\tau}|x_{\mathcal{V}|i,j}^{1:T}) = 0, \quad \forall\tau. \tag{10}$$

In other words, the lagged conditional covariance equals 0 for all time tag τ. On the other hand, the conditional dependence can also be defined in the frequency domain of the time series. Concretely, we first define the spectral density matrix as the Fourier transform of the lagged covariance matrix $COV(\boldsymbol{x}^t, \boldsymbol{x}^{t+\tau})$:

$$S^\omega = \sum_\tau COV(\boldsymbol{x}^t, \boldsymbol{x}^{t+\tau})\exp(-i\omega\tau), \tag{11}$$

for $\omega \in [0, 2\pi]$. Let $K^\omega = (S^\omega)^{-1}$, the conditional independence between x_i and x_j holds if and only if

$$K_{ij}^\omega = 0, \quad \forall\omega \tag{12}$$

This suggests that one common zero entry in the inverse spectral density matrices across a certain frequency band equates with the conditional independence between the corresponding two time series in this frequency band. Therefore, for a multivariate time series, we aim to infer the inverse spectral density matrices K^ω.

Here, we follow the state-of-the-art Whittle approximation framework [24], and assume that the discrete Fourier transform coefficients $\boldsymbol{f}_{1:P}^\omega$ are independent complex Gaussian random variables with mean zero and precision matrix given by the inverse spectral density matrix K^ω at the same frequency. As a result, we can learn K^ω that changes smoothly with ω from $\boldsymbol{f}^\omega$ using the proposed model in Section 2.2.

3. VARIATIONAL BAYES LEARNING

In this section, we derive a mean field VB algorithm to learn the above Bayesian model. VB algorithm seeks to approximate the intractable posterior distribution $p(\Theta|\boldsymbol{x}^{1:T})$ with a tractable variational distribution $q(\Theta)$ by minimizing the KL divergence between them $\mathrm{KL}(q|p) = \int q\log(q/p)$, where $\theta = \{K^{1:T}, J^{1:T}, \boldsymbol{s}^{1:T}, \boldsymbol{\beta}^{1:T}, \alpha, \boldsymbol{\gamma}, \boldsymbol{\lambda}\}$. This is equivalent to maximizing a lower bound of the data evidence. Concretely, we factorize the variational distribution as:

$$q(\theta) = q(\alpha)q(\lambda)q(\gamma)\prod_i\prod_j q(J_{ij}^{1:T})q(\beta_{ij}^{1:T})\prod_t q(K_{ij}^t)q(s_{ij}^t).$$

The VB update rules can be then derived as follows. For the precision matrices K^t,

$$q(K^t) \propto \exp(\langle\log p(\boldsymbol{x}^t|K^t)\rangle + \langle\log p(K^t|J^t, s^t, \alpha)\rangle)$$
$$\propto \exp\left(\frac{1}{2}\log\det K^t - \frac{1}{2}\boldsymbol{x}^{t\prime}K^t\boldsymbol{x}^t\right.$$
$$\left. - \frac{1}{2}\langle\alpha\rangle\langle\|K^t - s^t \odot J^t\|_2^2\rangle\right), \tag{13}$$

where $\langle\cdot\rangle$ denotes expectation, $\|\cdot\|_2$ denotes ℓ_2 norm, and $\odot$ is a Hadamard product. Due to the non-conjugacy between $p(\boldsymbol{x}^t|K^t)$ and $p(K^t|J^t, s^t, \alpha)$, it is intractable to compute the normalization constant of the above expression. Instead, we specify $q(K^t)$ as a product of independent Gaussian distributions $\prod_i\prod_j q(K_{ij}^t)$, whose mean $[\mu_K^t]_{ij}$ and variance $[\nu_K^t]_{ij}$ can be determined by Laplace approximation [25]. Specifically, equating to 0 the gradient of the exponential part in (13) w.r.t. K^t yields an equation of the mean matrix μ_K^t:

$$\frac{1}{2}{\mu_K^t}^{-1} - \langle\alpha\rangle\mu_K^t = \frac{1}{2}\boldsymbol{x}^t\boldsymbol{x}^{t\prime} - \langle\alpha\rangle\langle s^t\rangle \odot \langle J^t\rangle. \tag{14}$$

Algorithm 1 VB Learning of Covariate-Varying Graphical Models

Input: $\boldsymbol{x}^{1:T}$;
Output: $q(K^{1:T})$, $q(s^{1:T})$;
Initialize the varitional parameters and iteration number $\kappa = 1$;
Repeat
1. Update $q(K_{ij}^t) = \mathcal{N}(K_{ij}^t, [\mu_K^t]_{ij}, [\nu_K^t]_{ij})$:
   1.1. $VDV' = \frac{1}{2}\boldsymbol{x}^t\boldsymbol{x}^{t\prime} - \langle\alpha\rangle\langle s^t\rangle \odot \langle J^t\rangle$,
   1.2. $\tilde{D}_{ii} = -D_{ii} + \sqrt{D_{ii}^2 + 2\langle\alpha\rangle}/2\langle\alpha\rangle$,
   1.3. $\mu_K^t = V\tilde{D}V'$,
   1.4. $\nu_K^t = \frac{1}{2}\text{diag}({\mu_K^t}^{-1})\text{diag}({\mu_K^t}^{-1})' + \langle\alpha\rangle$.
2. Update $q(s_{ij}^t) = \text{Ber}(s_{ij}^t; \tilde{\eta}_{ij}^t)$:
   2.1. $\tilde{\eta}_{ij}^t = g\big(\langle\beta_{ij}^t\rangle - \langle\alpha\rangle(\langle {J_{ij}^t}^2\rangle - 2\langle K_{ij}^t\rangle\langle J_{ij}^t\rangle)\big)$.
3. Update $q(\beta_{ij}^{1:T})$:
   3.1. for $t = 1$ or T, $\phi(\beta_{ij}^t) \propto \exp(-(h(\zeta_{ij}^t) + \langle\lambda\rangle/2){\beta_{ij}^t}^2 + (\langle s_{ij}^t\rangle - 0.5)\beta_{ij}^t)$,
   3.2. for $2 < t < T-1$, $\phi(\beta_{ij}^t) \propto \exp(-(h(\zeta_{ij}^t) + \langle\lambda\rangle){\beta_{ij}^t}^2 + (\langle s_{ij}^t\rangle - 0.5)\beta_{ij}^t)$,
   3.3. $\phi(\beta_{ij}^t, \beta_{ij}^{t-1}) \propto \exp(-\langle\lambda\rangle\beta_{ij}^t\beta_{ij}^{t-1})$,
   3.4. Compute mean $[\mu_\beta^t]_{ij}$ and variance $[\nu_\beta^t]_{ij}$ via belief propagation in the Gauss-Markov chain with unary potential $\phi(\beta_{ij}^t)$ and pairwise potential $\phi(\beta_{ij}^t, \beta_{ij}^{t-1})$.
4. Update $\zeta_{ij}^t = \sqrt{[\mu_\beta^t]_{ij}^2 + [\nu_\beta^t]_{ij}}$.
5. Update $q(J_{ij}^{1:T})$ in a similar manner as $q(\beta_{ij}^{1:T})$.
6. Update $q(\alpha)$, $q(\lambda)$, and $q(\gamma)$.
Until convergence criterion is met

Letting VDV' denote the eigendecomposition of $\frac{1}{2}\boldsymbol{x}^t\boldsymbol{x}^{t\prime} - \langle\alpha\rangle\langle s^t\rangle \odot \langle J^t\rangle$, the solution of the above equation is given by $\mu_K^t = V\tilde{D}V'$, where $\tilde{D}$ is the diagonal eigenvalue matrix with

$$\tilde{D}_{ii} = \frac{-D_{ii} + \sqrt{D_{ii}^2 + 2\langle\alpha\rangle}}{2\langle\alpha\rangle}. \tag{15}$$

It follows from (15) that $\tilde{D}_{ii}$ is guaranteed to be positive, and therefore, μ_K^t is always positive definite during the inference process. On the other hand, the variance of $q(K_{ij}^t)$ is given by the negative Hessian at $[\mu_K^t]_{ij}$, that is,

$$\nu_K^t = \frac{1}{2}\text{diag}({\mu_K^t}^{-1})\text{diag}({\mu_K^t}^{-1})' + \langle\alpha\rangle, \tag{16}$$

where $\text{diag}(\cdot)$ is the diagonal of a matrix.

Next, we turn our attention to the update rule of β_{ij}^t. Its likelihood involves the logistic function, which spoils the conjugate-exponential structure. We overcome this limitation by utilizing a variational lower bound for $g(\beta_{ij}^t)$ based on bounding log convex functions [26]:

$$g(\beta_{ij}^t) \geq g(\zeta_{ij}^t)\exp\left(\frac{\beta_{ij}^t - \zeta_{ij}^t}{2} - h(\zeta_{ij}^t)\big({\beta_{ij}^t}^2 - {\zeta_{ij}^t}^2\big)\right),$$

where $h(\zeta_{ij}^t) = \tanh(\zeta_{ij}^t/2)/(4\zeta_{ij}^t)$, and ζ_{ij}^t is a variational parameter to be estimated. The bound is exact at $\zeta_{ij}^t = \pm\beta_{ij}^t$. Given ζ_{ij}^t, the Gaussian prior on $\beta_{ij}^{1:T}$ is now conjugate to the likelihood of β_{ij}^t, and the mean $[\mu_\beta^t]_{ij}$ and variance $[\nu_\beta^t]_{ij}$ of $q(\beta_{ij}^t)$ can be obtained via belief propagation in the Gauss-Markov Chain. After obtaining $q(\beta_{ij}^t)$, the variational parameter ζ_{ij}^t can be updated as:

$$\zeta_{ij}^t = \sqrt{[\mu_\beta^t]_{ij}^2 + [\nu_\beta^t]_{ij}}. \tag{17}$$

For the rest parameters $s_{ij}^{1:T}$, $J_{ij}^{1:T}$, λ, γ, and α, the corresponding prior and likelihood are within the conjugate-exponential family, and hence, we omit the detailed derivations of their variational distributions. We summarize the VB algorithm in Algorithm 1. The resulting computational complexity of the proposed algorithm is $\mathcal{O}(TP^3)$, indicating that this model is applicable to problems with large T.

4. EXPERIMENTAL RESULTS

In this section, we validate the proposed model via both synthetic data and real data. Particularly, we benchmark our model (referred to as VB-CVGM) with 1) the frequentist method for learning covariate-varying graphical models (referred to as OPT-CVGM) [9, 12], 2) the graphical lasso (i.e., glasso) for learning sparse graphical models from i.i.d. data [2]. The penalty parameters in the latter two methods are selected by AIC [10] and stability selection [27] respectively.

4.1. Synthetic Data

We simulate synthetic Gaussian distributed data from both relatively smoothly and abruptly changing precision matrix. The dimension of the data is P and the sample size is T. For both scenarios, we first randomly partition all off-diagonal elements into two sets: 10% of non-zeros and 90% of zeros. Next, in the first case, we randomly select n elements from each set. For the n elements chosen from the non-zero set, we gradually shrink them to 0 as a linear function of t. Similarly, for the n elements selected from the zero set, we increase them from zero at a rate linear in t. On the other hand, under the second scenario, the precision matrix only changes at $t = t_p, 2t_p, \cdots$, where t_p is a predefined period. At every t_p time step, we randomly choose n elements from the non-zero set and set them to zero, while picking n elements from the zero set and set them to non-zero. In our simulations, $T = 3000$, $P = 20$, n equals 9 and 4 respectively in the first and second scenario, and $t_p = 500$.

For both cases, we compare the performance of the three models in terms of estimation accuracy of graphical model, model fitting and computational time. More specifically, for accuracy of graph model estimation, we consider four criteria, including precision, recall, F_1-score as well as normalized mean squared error (NMSE). Precision and Recall are respectively defined as the proportion of correctly estimated edges to all the edges in the estimated graph and the true graph; F_1-score is defined as 2·precision·recall/(precision+recall), which is a weighted average of the precision and recall; NMSE can be computed as $\sum_t\sum_i\sum_j(K_{ij}^t - \hat{K}_{ij}^t)^2 / \sum_t\sum_i\sum_j(K_{ij}^t)^2$, where K_{ij}^t and $\hat{K}_{ij}^t$ are true and estimated value respectively. For model

Table 1: Performance of the three models on synthetic data

Scenario	Model	Precision	Recall	F_1-score	NMSE	NegLogLLH	Prm No.	AIC	Running Time
Smooth Case	VB-CVGM	1.00	0.74	0.85	1.71×10^{-2}	-4.92×10^{3}	83.89	-9.67×10^{3}	79.49
	OPT-CVGM	0.48	0.56	0.47	1.65	-3.71×10^{4}	146.18	-7.39×10^{4}	3.18×10^{4}
	glasso	1.00	0.91	0.95	9.44×10^{-3}	-3.95×10^{3}	77.62	-7.75×10^{3}	47.05
Abrupt Case	VB-CVGM	0.92	0.90	0.91	3.34×10^{-2}	-4.37×10^{3}	92.37	-8.55×10^{3}	73.73
	OPT-CVGM	0.45	0.70	0.49	1.68	-3.50×10^{4}	142.90	-6.97×10^{4}	3.24×10^{4}
	glasso	0.83	0.90	0.86	3.38×10^{-2}	-2.36×10^{3}	71.43	-4.58×10^{3}	36.22

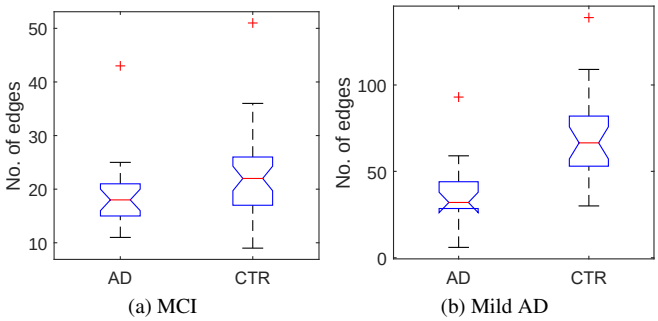
fitting, we evaluate the negative log-likelihood (negLogLLH), the number of parameters (Prm No.) in the model, and the AIC score. The results averaged over 100 data sets in each scenario are listed in Table 1.

As observed from the table, OPT-CVGM cannot well recover the true graphical model, although it introduces the largest number of parameters. The smallest NegLogLLH further suggests that this approach overfits the data. Indeed, as as reported in [14], graphical model selection via AIC typically generates over-dense graphs. It is therefore imperative to find a proper method to select the penalty parameters, and this again highlight the significance of our work. In addition, OPT-CVGM is very time-consuming, and can be prohibitive to large-scale data in practice. On the other hand, the performance of glasso for model fitting is the worst among the three models, since it fails to capture the temporal variations in K^t. Furthermore, by comparing the results of two case studies, we can tell that this issue is exacerbated when K^t changes more drastically. In contrast, the proposed VB-CVGM successfully strike a balance between data fitting and graph recovery. Moreover, the model is free of tuning and the computational time is several magnitudes shorter than that of OPT-CVGM.

4.2. Scalp EEG of AD patients

In this section, we consider inferring functional brain networks from scalp EEG recordings. Specifically, we analyze two data sets. The first one contains 22 patients with mild cognitive impairment (MCI, i.e., the first stage of AD) and 38 healthy control subjects [28]. The second one consists of 17 patients with mild AD (i.e., the second stage of AD) and 24 control subjects [29]. Although AD cannot be cured, current symptons-delaying medications are proven to be more effective at early stages of AD [28]. On the other hand, scalp EEG recording systems are inexpensive and potentially mobile, thus making it a useful tool to screen a large population for the risk of AD. As a result, it is crucial to identify MCI or Mild AD from scalp EEG signals.

As introduced in Section 2.3, we can learn graphical models from scalp EEG signals by estimating the inverse spectral density matrices K^ω in the frequency domain. Here, we use the Fourier coefficients $\boldsymbol{f}^\omega$ in the frequency band $4-30$Hz, as suggested by previous works on the same data sets [28, 29]. After obtaining K^ω, we further consider several components within the frequency range: $4-8$Hz, $8-12$Hz, and $12-30$Hz. For each smaller frequency band, we infer the corresponding graphical models by finding the common zero patterns of all



(a) MCI (b) Mild AD

Fig. 2: Boxplots for the number of edges resulting from VB-CVGM.

K^ω in this band. Finally, we count the number of edges in the graphical models, which can be regarded as a measure of synchrony. We observe that graphical models in $8-12$Hz can best distinguish between patients and controls for the MCI data, while $4-8$Hz for the Mild AD data. We depict in Fig. 2 the boxplots of the number of edges in the graphical models for both AD patients and control subjects. Clearly, the graphical models for AD patients are more sparse than of healthy people, and this phenomenon becomes more pronounced for Mild AD patients. Such findings are consistent with the loss of synchrony within the EEG signals for AD patients as reported in the literature [28, 29]. We further conduct Mann-Whitney test. The resulting p-values for the two data sets are respectively 3.53×10^{-2} and 2.09×10^{-4}, which are statistically significant. As a comparison, we filter EEG signals with a bandpass filter (i.e., $8-12$Hz for MCI and $4-8$Hz for Mild AD), and utilize glasso to learn graphical models in the time domain. The resulting p-values are 0.15 for the MCI data and 0.82 for the Mild AD data. Obviously, the proposed model can better describe the perturbations in the EEG synchrony for AD patients.

5. CONCLUSION AND FUTURE WORK

In this paper, we reformulate the problem of estimating covariate-varying graphical models from a Bayesian perspective. A VB algorithm is then developed to learn the model without tuning. We further apply the proposed model to learn graphical models between multiple time series in the frequency domain. Numerical results from real data show that the proposed model may help diagnose AD at an early stage from scalp EEG.

6. ACKNOWLEDGMENT

This research was supported by MOE (Singapore) ACRF Tier 2 grant M4020187.

7. REFERENCES

[1] D. Koller and N. Friedman, "Probabilistic Graphical Models," *The MIT Press*, 2009.

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

[3] C.-J. Hsieh, M. A. Sustik, I. S. Dhillon, P. Ravikumar, and R. A. Poldrack, "BIG & QUIC: Sparse Inverse Covariance Estimation for a Million Variables," *Proc. NIPS*, 2013.

[4] M. A. Kramer, U. T. Eden, E. D. Kolaczyk, R. Zepeda, E. N. Eskandar, and S. S. Cash, "Coalescence and Fragmentation of Cortical Networks during Focal Seizures", *J. Neurosci.*, **30** (30): 10076-10085, 2010.

[5] S. Zhou, J. Lafferty, and L. Wasserman, "Time varying undirected graphs," *Mach. Learn.*, **80**: 295-319, 2010.

[6] M. Kolar, L. Song, A. Ahmed, and E. P. Xing, "Estimating Time-Varying Networks," *Ann. Appl. Math.*, **4** (1): 94-123, 2010.

[7] M. Kolar and E. P. Xing, "On Time Varying Undirected Graphs," *J. Mach. Learn. Res.*, **15** W&CP, 2011.

[8] H. Qiu and F. Han, "Joint estimation of multiple graphical models from high dimensional time series," *J. R. Statist. Soc.* B, **78** (2): 487-507, 2016.

[9] A. J. Gibberd and J. D. B. Nelson, "Estimating Dynamic Graphical Models from Multivariate Time-Series Data," *Proc. AALTD*, 2015.

[10] R. P. Monti, P. Hellyer, D. Sharp, R. Leech, C. Anagnostopoulos, G. Montana, "Estimating time-varying brain connectivity networks from functional MRI time series," *NeuroImage*, **103**: 427-443, 2014.

[11] A. Ahmed and E. P. Xing, "Recovering time-varying networks of dependencies in social and biological studies," *Proc. Natl. Acad. Sci. USA*, **106** (29): 11878-11883, 2009.

[12] S. Yang, Z. Lu, X. Shen, P. Wonka, and J. Ye, "Fused Multiple Graphical Lasso," *SIAM J. Optim.*, **25** (2): 916-943, 2015.

[13] A. J. Gibberd and J. D. B. Nelson, "High Dimensional Changepoint Detection with a Dynamic Graphical Lasso," *Proc. ICASSP*, 2014.

[14] H. Liu, K. Roeder, and L. Wasserman, "Stability Approach to Regularization Selection (StARS) for High Dimensional Graphical Models," *Proc. NIPS*, 2010.

[15] H. Yu, J. Dauwels, and P. Johnathan, "Extreme-Value Graphical Models with Multiple Covariates," *IEEE Trans. Signal Process.*, **62** (21): 5734-5747, 2014.

[16] A. Jung, G. Hannak, and N. Goertz, "Graphical LASSO based Model Selection for Time Series," *IEEE Signal Process. Lett.*, **22** (10): 1781-1785, 2015.

[17] H. Ishwaran and J. S. Rao, "Spike and slab variable selection: Frequentist and Bayesian strategies," *Ann. Stat.*, **33** (2): 730-773, 2005.

[18] J. Ziniel and P. Schniter, "Dynamic Compressive Sensing of Time-Varying Signals via Approximate Message Passing," *IEEE Trans. Signal Process.*, **61** (21): 5270-5284, 2013.

[19] V. Cevher, M. F. Duarte, C. Hegde, and R. G. Baraniuk, "Sparse signal recovery using Markov random fields," *Proc. NIPS*, 2008.

[20] E. B. Fox, E. B. Sudderth, M. I. Jordan, and A. S. Willsky, "A Sticky HDP-HMM with Applications to Spearker Diarization," *Ann. Appl. Stat.*, **5** (2A): 1020-1056, 2011.

[21] L. Ren, L. Du, L. Carin, and D. B. Dunson, "Logistic Stick-Breaking Process," *J. Mach. Learn. Res.*, **12**: 203-219, 2011.

[22] M. R. Andersen, O. Winther, and L. K. Hansen, "Bayesian Inference for Structured Spike and Slab Priors," *Proc. NIPS*, 2014.

[23] J. Dauwels, H. Yu, X. Wang, F. Vialatte, C. Latchoumane, J. Jeong, and A. Cichocki, "Inferring Brain Networks through Graphical Models with Hidden Variables", *Mach. Learn. & Interpretation in Neuroimaging, Lecture Notes in Comput. Sci., Springer*, pp. 194-201, 2012.

[24] P. Whittle, "The analysis of multiple stationary time series," *J. R. Statist. Soc.* B, **15** (1): 125-139, 1953.

[25] C. Wang and D. B. Blei, "Variational Inference in Nonconjugate Models," *J. Mach. Learn. Res.*, **14**: 899-925, 2013.

[26] C. M. Bishop and M. Svensén, "Bayesian hierarchical mixture of experts," *Proc. UAI*, 2003.

[27] S. Li, L. Hsu, J. Peng and P. Wang, "Bootstrap inference for network construction with an application to a breast cancer microarray study," *Ann. Appl. Stat.*, **7** (1): 391-417, 2013.

[28] J. Dauwels, F. Vialatte, T. Musha, and A. Cichocki, "A comparative study of synchrony measures for the early diagnosis of Alzheimer's disease based on EEG," *NeuroImage*, **49**: 668-693, 2010.

[29] G. Henderson, E. Ifeachor, N. Hudson, C. Goh, N. Outram, S. Wimalaratna, C. Del Percio, and F. Vecchio, "Development and assessment of methods for detecting dementia using the human electroencephalogram," *IEEE Trans. Biom. Eng*, **53**: 1557-1568, 2006.