18th International Conference on Information Fusion
Washington, DC - July 6-9, 2015

Variational Inference for Graphical Models of Multivariate Piecewise-Stationary Time Series

Hang Yu and Justin Dauwels
School of Electrical and Electronics Engineering,
Nanyang Technological University, 50 Nanyang Avenue, Singapore 639798

***Abstract*—Graphical models provide a powerful formalism for statistical modeling of complex systems. Especially sparse graphical models have seen wide applications recently, as they allow us to infer network structure from multiple time series (e.g., functional brain networks from multichannel electroencephalograms). So far, most of the literature deals with stationary time series, whereas real-life time series often exhibit non-stationarity. In this paper, we focus on multivariate piecewise-stationary time series, and propose novel Bayesian techniques to infer the change points and the graphical models of stationary time segments. Concretely, we model the time series as a hidden Markov model whose hidden states correspond to different Gaussian graphical models. As such, the transition between different states represents a change point. We further impose a stick-breaking process prior on the hidden states and shrinkage priors on the inverse covariance matrices of different states. We then derive an efficient stochastic variational inference algorithm to learn the model with sublinear time complexity. As an important advantage of the proposed approach, the number and position of the change points as well as the graphical model structures are inferred in an automatic manner without tuning any parameters. The proposed method is validated through numerical experiments.**

I. Introduction

Inferring sparse graphical models is currently en vogue, since such models can represent the dependence between a great number of variables in a succinct manner [1], [2]. For example, given a large collection of genes, a sparse graphical model (a.k.a a gene regulatory network in this example) can be used to automatically detect the gene pairs with strong correlation, thus greatly reducing the time and effort for further experimental analysis. As a result, there is substantial literature on learning graphical models from various types of data, such as Gaussian [3], [4], non-Gaussian [5]-[7], and discrete [8].

While the previous works are limited to estimating a single static graphical model from independent and identically distributed (*i.i.d*) samples, real data are often associated with non-stationarity, and proper consideration of it will greatly help interpret the data. In [9], for instance, the functional brain network is shown to evolve through a distinct topological progression during the seizure, thus providing new insights into the mechanisms of seizures and novel intervention strategies. However, the authors simply defined a time window with fixed length and inferred a network for each window, introducing artifacts to the analysis. To resolve the issue, change points detection is required, yet modeling changing dependency structure in multivariate time series has only received limited attention so far.

Below, we present a brief review of multiple change points detection for multivariate time series. Xuan *et al* [10] extended the Bayesian change point detection approaches for univariate time series to the multivariate setting: they adopt a geometric prior on the time segment lengths, and then iterate between MAP segmentation and graphical model inference. Despite the time-consuming Monte Carlo Markov chain (MCMC) method used in the paper, the main restriction is that the graph for all segments must be decomposable. As an alternative, a greedy binary segmentation scheme is proposed in [11]. A change point is inserted such that the Bayesian information criterion (BIC) of the two graphical models of the data before and after the change point is minimized; this procedure is repeated until no further splits reduce the BIC score. Unfortunately, besides the high computational complexity, the binary segmentation can be misleading and overestimate the number of change points, as pointed out in [12]. To overcome the problem, dynamic programming is applied in [13], leading to joint estimation of all the change points. However, the method has computational complexity of order $\mathcal{O}(T^3)$ in the number of fixed points T, which is impractical for most real-life time series with length of hundreds or thousands. In our previous work [14], we formulate the change point detection problem as maximizing the log-likelihood of all segments with a penalty on the number of change points. The optimization problem is then solved using a pruned dynamic programming method with linear time complexity [15]. Graphical models associated with each segments are inferred via convex optimization techniques proposed in [3], [4]. We also put forward adaptive methods to choose the penalty parameters for both change points detection and graphical model inference. This method is still computational demanding in practice though, since the algorithm has to be run on every possible choice of the penalty parameters in order to find the best ones.

To address the abovementioned concerns, we propose in this paper a novel variational Bayes method to infer the abruptly changing graphical models for multivariate piecewise-stationary time series. Specifically, we describe the piecewise-stationary time series using a hidden Markov model (HMM), in which the emission distribution of each state is given by a Gaussian graphical model (GGM). We further assume that the transition matrix is upper triangular such that the resulting HMM is equivalent to the classical change point detection models in which data of different time segments are independent of each other. Consequently, the number of

states equals the number of change points in the time series, and a transition between different states represents a change point between two time segments. In order to infer the number of hidden states, we impose a stick-breaking process on the transition probabilities. Such a prior automatically selects a proper number of hidden states to express the data. On the other hand, to obtain a sparse graphical model for each hidden state, we put a shrinkage prior on the corresponding precision (inverse covariance) matrix of the GGM. The resulting Bayesian model is then learnt using a stochastic variational inference approach [16]. To be more explicit, we borrow the idea from [17] and compute the stochastic (natural) gradients in each iteration based on a minibatch of subchains sampling from the HMM. Therefore, the time complexity can be reduced to sublinear, and the resulting model can be applicable to time series with length of thousands or millions. Note that Wulsin *et al.* [18] utilizes a similar framework when describing changing correlations in brain recordings. However, in their model, the structure of the graphical model is fixed to be the neighboring structure of the electrodes. Furthermore, in order to automatically infer the number of hidden sates, a hierarchical Dirichlet process (HDP) prior is leveraged. Due to the lack of conjugacy between the two levels of Dirichlet process, it is not straightforward to derive fast variational Bayes algorithms. Instead, they apply MCMC methods, and hence this method is computationally intensive. On the other hand, in [19], a HMM with an upper triangular transition matrix is integrated in a Bayesian framework to identify change points in univariate time series. The model is inferred by MCMC methods. Different from these models, the proposed model deals with abruptly changing graphical models of multivariate time series. Moreover, we develop low-complexity stochastic variational inference algorithms to learn the model such that the proposed model is applicable to large scale data.

Experimental results show that the proposed algorithm can infer the number and position of the change points as well as the sparse graphical model of each state in an automatic manner.

This paper is structured as follows. In Section II, we present the proposed graphical models for multivariate piecewise-stationary time series in length. We then derive the stochastic variational inference algorithm in Section III. Numerical results for both synthetic data are presented in Section IV. We close the paper by offering concluding remarks in Section V.

II. Graphical Models of Multivariate Pieicewise-Stationary Time Series

Let us suppose that we have an ordered time sequence of data $\boldsymbol{y} = (\boldsymbol{y}_t)$, where $t = 1, \cdots, T$ and $\boldsymbol{y}_t \in \mathbb{R}^P$. We aim to partition the T samples into K stationary time segments, thus introducing $K-1$ change points $\boldsymbol{\tau}_{1:K-1} = (\tau_1, \cdots, \tau_K - 1)$. Each change point is an integer between 1 and $T-1$. We further define $\tau_0 = 0$ and $\tau_K = T$, therefore, the k-th segment is given by $\boldsymbol{y}^{(1:P)}_{(\tau_{k-1}+1:\tau_k)}$, where $k = 1, \cdots, K$.

A hidden Markov model (HMM) defines a probability distribution of $\boldsymbol{y}$ by introducing another sequence of hidden states $\boldsymbol{s} = (s_t)_{t=1}^T$, where $s_t \in \{1, \cdots, N_s\}$ and $N_s \leq K$ is the number of states. The sequence of hidden states is a Markov process. Given the state s_t at time t, the observed $\boldsymbol{y}_t$ is independent of other variables in the model. As a consequence, the model is well defined by three sets of parameters, including the initial distribution $p(s_1)$, the transition matrix A such that $A_{ij} = p(s_{t+1} = j | s_t = i)$, and the emission distribution $p(\boldsymbol{y}_t | s_t = i) = \mathcal{N}(\boldsymbol{y}_t; \boldsymbol{0}, (J^i)^{-1})$, where we assume that the mean is zero and J^i is the precision matrix (inverse covariance matrix) characterizing the Gaussian graphical model (GGM) of the P variables $\boldsymbol{y}_t^{(1:P)}$ indexed by the state at time t (i.e., $s_t = i$). We further define $p(s_1)$ to be a uniform distribution over all possible states for simplicity. Note that the HMM introduces a change point automatically when $s_{t+1} \neq s_t$.

In previous works [10]-[14], the product partition model (PPM) is often utilized to identify change points, in which data is independent across different time segments. In other words, in the PPM, we can never enter an old state once we have left the corresponding time segment. Therefore, to resemble the PPM, we assume that the transition matrix of the HMM is upper triangular. As a result, in our model, the number of hidden states equals the number of change points plus one.

Here, our objective is to infer the state variables $\boldsymbol{s}$, thereby obtaining the change points, as well as the precision matrix J^i corresponding to all states $s_t = i$. For the problem of change points detection, we aim to use the smallest number of change points (i.e., number of states) that can well explain the piecewise-stationary property of the time series. As mentioned in Section I, imposing a nonparametric hierarchical Dirichlet process prior on the transition matrix has proven effective to infer the number of states automatically [18]. However, this prior does not extend to fast variational algorithms due to the non-conjugate issue. Instead, we resort to the stick-breaking process [20] that is conjugate to the transition probabilities. For the upper triangular transition matrix, the stick breaking process can be defined as follows:

$$A_{ij} = V_{ij} \prod_{k=i}^{j-1} (1 - V_{ik}), \tag{1}$$

$$V_{ij} \sim \text{Beta}(\alpha_{ij}, \beta_{ij}). \tag{2}$$

The process can be interpreted as iteratively breaking the portion of V_{ij} from the remaining of a unit-length stick $\prod_{k=i}^{j-1}(1 - V_{ik})$. According to the definition, the process is infinite, that is, we may allow a countably infinite number of hidden states as the length of the time series increases. Moreover, although the state space is infinite, the resulting posteriors $p(A_{i:}|\boldsymbol{y})$ will only have "large" probabilities in a finite number of states that are useful in explaining the observed data, whereas all others are nearly equal to zero [20]. As such, the stick-breaking process prior can effectively select a proper number of states. We can find from (2) that the stick-breaking process requires an infinite parameterizations. To simplify this, we follow [20] to fix $\alpha_{ij} = 1$ and put a conjugate gamma prior $\text{Gamma}(\beta_{ij}; a, b)$ on β_{ij}. The hyper parameters a and b are set to be 10^{-6} and 0.1 respectively,

such that the state transition is properly encouraged to detect the entire state structure [20].

On the other hand, we wish to infer a sparse J^i for every possible state i, thus we can unveil the changing network structure of the observed data. We thus associate the off-diagonal elements J^i_{jk} of J^i with Gaussian priors with zero means and precisions λ^i_{jk}, i.e.,

$$p(J^i_{jk}|\lambda^i_{jk}) \propto \sqrt{\lambda^i_{jk}} \exp\Big(-\frac{1}{2}\lambda^i_{jk} {J^i_{jk}}^2\Big), \tag{3}$$

for all $j > k$. As shown in our numerical experiments, many of the precisions λ^i_{jk} will take very large values during the learning process, and consequently, the prior can successfully shrink most elements of J^i to zero, and yield sparse graphical models. We further impose conjugate Gamma hyperprior on the precisions λ^i_{jk}:

$$p(\lambda^i_{jk}) = \text{Gamma}(\lambda^i_{jk}; c, d) \propto {\lambda^i_{jk}}^{a-1} \exp(-b\lambda^i_{jk}). \tag{4}$$

The parameters c and d are set to small values (e.g., 10^{-10}) to obtain a flat non-informative prior. Note that

$$\int p(J^i_{jk}|\lambda^i_{jk}) p(\lambda^i_{jk}) d\lambda^i_{jk} = \frac{\Gamma\big(c+\frac{1}{2}\big)}{\Gamma(c)\sqrt{2\pi d}} \Big(\frac{1}{1+\frac{1}{2d}{J^i_{jk}}^2}\Big)^{c+\frac{1}{2}},$$

which is a t distribution. Therefore, we essentially put a t prior on J^i_{jk}. Such shrinkage prior is often used in the Bayesian framework to promote sparsity [22], [23]. Note that in the literature of learning graphical models [3], Laplace priors are often used since they amount to ℓ_1 norm penalties on the precision matrix and the resulting optimization problem is convex. Although Laplace priors can also be regarded as a scale mixture of Gaussian, the hyperprior on precisions λ^i_{ij} is the inverse Gamma distribution that is not conjugate to the Gaussian distributions parameterized by precisions [24]. As a result, we employ t prior here since it is more tractable for Bayesian inference.

We now turn our attention to the overall model, which can be expressed as:

$$\begin{aligned}
&p(\boldsymbol{y}, \boldsymbol{s}, V, \boldsymbol{\alpha}, \boldsymbol{\beta}, J, \boldsymbol{\lambda}) \\
&= p(\boldsymbol{y}|\boldsymbol{s}, J) p(V|\boldsymbol{\beta}) p(\boldsymbol{\beta}) p(J|\boldsymbol{\lambda}) p(\boldsymbol{\lambda}) \\
&= p(s_1) \prod_{t=2}^{T} p(s_t|s_{t-1}, V) \prod_{t=1}^{T} p(y_t|s_t, J_{s_t}) \times \\
&\prod_{i=1}^{\infty}\prod_{j=1}^{\infty} \big[p(V_{ij}|\beta_{ij}) p(\beta_{ij})\big] \prod_{i=1}^{\infty}\prod_{k=1}^{P}\prod_{j=k+1}^{P} \big[p([J^i]_{jk}|\lambda^i_{jk}) p(\lambda^i_{jk})\big].
\end{aligned} \tag{5}$$

III. Stochastic Variational Inference

In this section, we derive a stochastic variational inference algorithm [16] to learn the model parameters. Concretely, we seek a variational distribution $q(\boldsymbol{s}, V, \boldsymbol{\alpha}, \boldsymbol{\beta}, J, \boldsymbol{\lambda})$ that maximizes the evidence lower bound L:

$$\log p(\boldsymbol{y}) \geq E_q[\log p(\boldsymbol{y}, \boldsymbol{s}, V, \boldsymbol{\alpha}, \boldsymbol{\beta}, J, \boldsymbol{\lambda})] - E_q[\log q] = L. \tag{6}$$

Maximizing L is equivalent to minimizing the KL divergence between the variational distribution q and the intractable posterior $p(\boldsymbol{s}, V, \boldsymbol{\alpha}, \boldsymbol{\beta}, J, \boldsymbol{\lambda}|y)$ as measured by $\text{KL}(q|p) = \int q \log(q/p)$. Here, we apply the mean-field approximation, and therefore, the variational distribution can be factorized as:

$$\begin{aligned}
q(\boldsymbol{s}, V, \boldsymbol{\alpha}, \boldsymbol{\beta}, J, \boldsymbol{\lambda}) =& q(\boldsymbol{s}) \prod_{i=1}^{K}\prod_{j=i}^{K-1} \big[q(V_{ij}) q(\beta_{ij})\big] \\
&\times \prod_{i=1}^{K} q(J^i) \prod_{i=1}^{K}\prod_{k=1}^{P}\prod_{j=k+1}^{P} q(\lambda^i_{jk}),
\end{aligned} \tag{7}$$

where

$$q(V_{ij}) = \text{Beta}(V_{ij}; W_{1ij}, W_{2ij}), \tag{8}$$

$$q(\beta_{ij}) = \text{Gamma}(\beta_{ij}; W_{3ij}, W_{4ij}), \tag{9}$$

$$q(J^i) = \prod_{j=1}^{P} \delta(J^i_{j:P,j} - {J^i_{j:P,j}}^*), \tag{10}$$

$$q(\lambda^i_{jk}) = \text{Gamma}(\lambda^i_{jk}; {W_5}^i_{jk}, {W_6}^i_{jk}), \tag{11}$$

$\delta(J^i_{j:P,j} - {J^i_{j:P,j}}^*)$ is a delta function which equals 1 when $J^i_{j:P,j} = {J^i_{j:P,j}}^*$ and 0 otherwise, and $J^i_{j:P,j}$ denotes the jth to Pth elements in the jth column. Since $p(J|\boldsymbol{\lambda})$ is not conjugate to $p(\boldsymbol{y}|J)$ in the proposed model (5), there is no closed-form variational distribution $q(J)$ in the framework of mean-field variational inference. Instead, it is convenient to use a point estimate of J^i (10) as in [23], [24]. Furthermore, as the algorithm proceeds, many of the precisions λ^i_{jk} will become very large, and then the delta functions can well approximate the true posterior distribution.

Additionally, in expression (7), we truncate the variational stick-breaking process to yield K levels, since the infinite large state space is computationally intractable for variational inference. In other words, the variational transition probability $\tilde{A}$ is given by:

$$\tilde{A}_{ij} = \begin{cases} V_{ij}\prod_{k=i}^{j-1}(1-V_{ik}) & \text{for } j < K \\ \prod_{k=i}^{j-1}(1-V_{ik}) & \text{for } j = k \\ 0 & \text{for } j > K \end{cases}, \tag{12}$$

$$V_{ij} \sim q(V_{ij}). \tag{13}$$

We emphasize that using the truncation level is quite different from setting a finite state-space in a statistical perspective. The proposed model is still a full stick-breaking process and is not truncated. K should be sufficiently large to ensure the accuracy of the approximation.

Finally, note that making a full mean-field approximation of the latent states $q(\boldsymbol{s}) = \prod_{t=1}^{T} q(s_t)$ would lose critical information about the hidden Markov chain required for accurate inference. Instead, we will infer a joint variational distribution $q(\boldsymbol{s})$ over the states of all time points.

The stochastic variational inference algorithm aims to find the variational parameters W_i $(i = 1, \cdots, 6)$ and J^* that maximizes the evidence lower bound L. To this end, the algorithm proceeds by updating the variational parameters in

the direction of the stochastic natural gradient. More precisely, in each iteration, W_i, for example, can be updated as:

$$W_i^{(\kappa+1)} = W_i^{(\kappa)} + \rho_\kappa \tilde{\nabla}_{W_i} L(W_i^{(\kappa)}), \tag{14}$$

where $\tilde{\nabla}_{W_i} L(W_i^{(\kappa)})$ denotes the stochastic natural gradients of L $w.r.t.$ W_i at the value of $W_i^{(\kappa)}$. In the sequel, we first list the natural gradients of all variational parameters, and further elaborate on the stochastic version of the gradients.

Natural gradients have a convenient form if the prior and the complete-data likelihood corresponding to the variational distribution are a conjugate pair of exponentials family distributions. This condition is satisfied by $q(V)$, $q(\boldsymbol{\beta})$, and $q(\boldsymbol{\lambda})$. Specifically,

1) For the variational parameters of V,

$$\begin{aligned}
&\nabla_{W_{1ij}} L({W_1}_{ij}^{(\kappa)}) \\
=&1 + E_{q(s)}[\sum_{t=2}^{T} \log p(s_t|s_{t-1})] - {W_1}_{ij}^{(\kappa)} \\
=&1 + \sum_{t=2}^{T} q(s_{t-1}=i, s_t=j) - {W_1}_{ij}^{(\kappa)}, \qquad (15) \\
&\nabla_{W_{2ij}} L({W_2}_{ij}^{(\kappa)}) \\
=&\frac{{W_3}_{ij}^{(\kappa)}}{{W_4}_{ij}^{(\kappa)}} + E_{q(s)}[\sum_{t=2}^{T} \log p(s_t|s_{t-1})] - {W_2}_{ij}^{(\kappa)} \\
=&\frac{{W_3}_{ij}^{(\kappa)}}{{W_4}_{ij}^{(\kappa)}} + \sum_{t=2}^{T} q(s_{t-1}=i, s_t>j) - {W_2}_{ij}^{(\kappa)}. \qquad (16)
\end{aligned}$$

2) For the variational parameters of $\boldsymbol{\beta}$,

$$\nabla_{W_{3ij}} L({W_3}_{ij}^{(\kappa)}) = a + 1 - {W_2}_{ij}^{(\kappa)}, \tag{17}$$

$$\nabla_{W_{3ij}} L({W_3}_{ij}^{(\kappa)}) = b - E_{q(V_{ij})}[\log(1-V_{ij})] - {W_3}_{ij}^{(\kappa)}, \tag{18}$$

3) For the variational parameters of $\boldsymbol{\lambda}$,

$$\nabla_{{W_5}_{jk}^{i}} L({W_5}_{jk}^{i\ (\kappa)}) = c + \frac{1}{2} - {W_5}_{jk}^{i\ (\kappa)}, \tag{19}$$

$$\nabla_{{W_6}_{jk}^{i}} L({W_6}_{jk}^{i\ (\kappa)}) = d + \frac{{J_{jk}^{i}}^2}{2} - {W_6}_{jk}^{i\ (\kappa)}. \tag{20}$$

Note that natural gradients are closely related to traditional variational Bayes (VB) algorithms [25]. By setting the natural gradients to zero in each iteration, we obtain the update rules of the VB algorithm. In other words, the natural gradients can be regarded as the difference between two consecutive VB updates. This approach is employed to compute the gradient of L $w.r.t$ J^*. In this case, the corresponding prior $p(J|\boldsymbol{\lambda})$ is not the conjugate to the data likelihood $p(\boldsymbol{y}|\boldsymbol{s}, J)$. Specifically, we first sequentially set the gradient of the L $w.r.t$ ${J_{j:P,j}^{i}}^*$ to zero, as in the VB framework. As such, we can update ${J_{j:P,j}^{i}}^*$

TABLE I: SVI of Graphical models for Multivariate Piecewise-Stationary Time Series.

Input: observed multivariate time series $\boldsymbol{y}$, maximum number of possible change points K, number $M = 50$ and length $L_s = 2$ of subchains drawn from the HMM in each iteration
Iterate the following steps until convergence.

1) draw M subchains from the entire Markov chain, each with length L
2) For each subchain $\boldsymbol{s}_m^{\text{sub}}$, compute $q(s_t)$ and $q(s_{t-1}, s_t)$ for $t \in \boldsymbol{s}_m^{\text{sub}}$ as follows:
   a) Initialize $q^{\text{old}}(s_t)$ ($t \in s_m^{\text{sub}}$) by running the forward-backward in $\boldsymbol{s}_m^{\text{sub}}$. $u = 1$.
   b) Augment $\boldsymbol{s}_m^{\text{sub}}$ in each direction by u observations and compute $q^{\text{new}}(s_t)$ ($t \in s_m^{\text{sub}}$) using the augmented subchain.
   c) If $\|q^{\text{new}}(s_t) - q^{\text{old}}(s_t)\| \leq \epsilon$, return $q(s_t) = q^{\text{new}}(s_t)$; otherwise, set $u = u + 1$, $q^{\text{old}}(s_t) = q^{\text{new}}(s_t)$ and go back to Step 2b.
3) Compute the (natural) gradients for all parameters of variational distributions following Eq. (15) to Eq. (23).
4) Update the parameters following Eq. (14).

as:

$$\begin{aligned}
{J_{j+1:P,j}^{i}}^{*(\kappa+1)} = &-\Bigg[S_{jj}^{i}\Big[{J_{-j,-j}^{i}}^{*-1}\Big]_{j:P-1,j:P-1} \\
&+ \text{diag}(E_{q(\boldsymbol{\lambda})}[\boldsymbol{\lambda}_{j+1:P,j}^{i}])\Bigg]^{-1}\Bigg(S_{j+1:P,j}^{i} \\
&+ S_{jj}^{i}\Big[{J_{-j,-j}^{i}}^{*-1}\Big]_{j:P-1,1:j-1} {J_{1:j-1,j}^{i}}^{*}\Bigg), \qquad (21) \\
{J_{jj}^{i}}^{*(\kappa+1)} = &\ N^i/[S^i]_{jj} + {J_{j,-j}^{i}}^{*} {J_{-j,-j}^{i}}^{*-1} {J_{-j,j}^{i}}^{*}. \qquad (22)
\end{aligned}$$

where $N^i = \sum_{t=1}^{T} q(s_t = i)$ and $S^i = N^i E_{q(s)}[\boldsymbol{y}_t \boldsymbol{y}_t^T]$. Next, we can calculate the corresponding gradient as:

$$\nabla_{J^{i*}} L(J^{i*(\kappa)}) = J^{i*(\kappa+1)} - J^{i*}. \tag{23}$$

Note that J^i is always positive definite during the update procedure given that its initial value is positive definite, as proven in [24].

We can tell from Eqs. (15), (16), (21), and (22) that in order to compute the accurate natural gradient for W_1, W_2 and J, we need to run forward-backward algorithm on the entire Markov chain to get $q(s_t)$ and $q(s_{t-1}, s_t)$. Although the time complexity of the message passing algorithm is linear, it may still be prohibitive for very long time series. Therefore, we instead borrow the idea from [17] and compute noisy stochastic natural gradients using the $q(s_t)$ and $q(s_{t-1}, s_t)$ from subchains $\boldsymbol{s}^{\text{sub}}$ containing consecutive observations. As a consequence, the time complexity of the proposed model is sublinear. In order to consider the forward and backward messages passing into the subchain, when computing the local beliefs $q(s_t)$ ($t \in \boldsymbol{s}^{\text{sub}}$), the subchain is augmented adaptively to include enough extra observations on each end, until further augmentation of the subchain will not significantly change the local beliefs. In other words, the local beliefs are within an ϵ-ball of the optimal $q^*(s_t)$ resulting from the entire chain. Foti *et al.* [17] has proven that such stochastic natural gradient ascent method converges to a local maximum as long as step sizes ρ_κ satisfy $\sum_\kappa \rho_\kappa^2 < \infty$ and $\sum_\kappa \rho_\kappa = \infty$. We adopt the

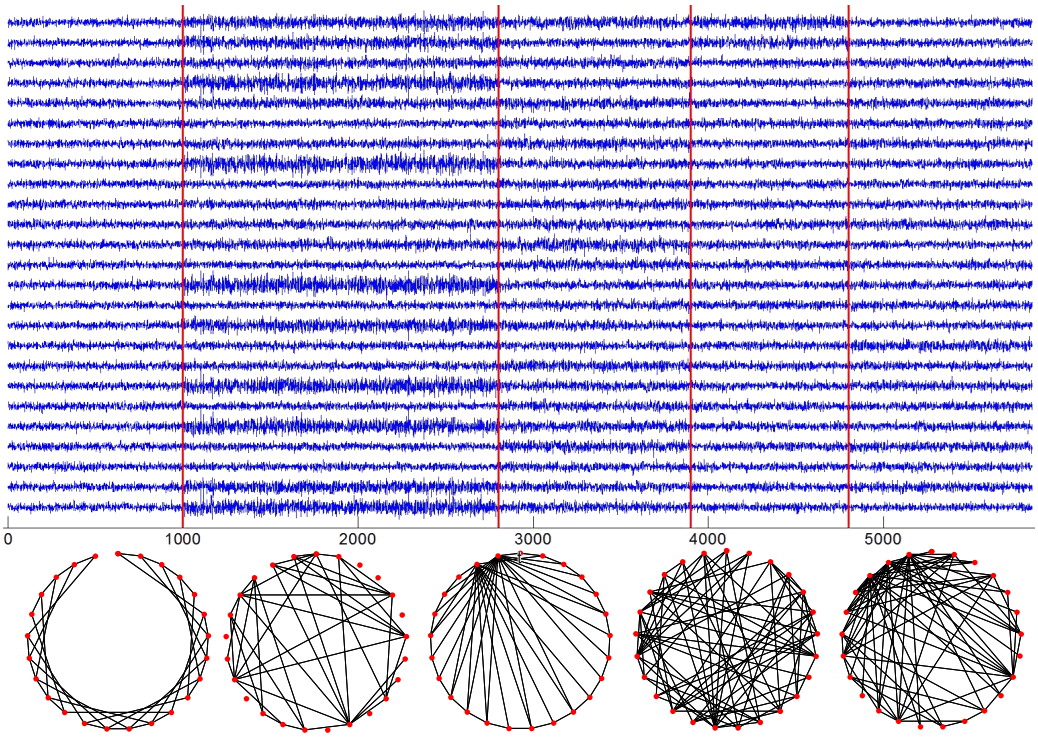

Fig. 1: 25-dimensional synthetic data, change points (red lines), and the true graphical models of all the segments.

automatic methods proposed in [26] to tune step sizes. The entire algorithm is summarized in Table I.

IV. Numerical Results

In this section, we present our results on synthetic data; we benchmark the proposed Bayesian model with the optimization method-based model [14] ,and the graphical lasso method [3] that is used to infer GGMs from stationary data. The penalty parameters in the second and third model are chosen by adaptive methods, cf [14].

The synthetic dataset has 25 variables and 5850 samples. The true value of change points are $\{1000, 2800, 3900, 4800\}$. The signals and the graphical models of the five time segments are shown in Fig. 1. We then test the three models using this dataset. More specifically, we compare the accuracy of change point detection, the accuracy of graphical model inference, and the running time. For graphical model inference, we consider three criteria, that is, precision, recall, and F_1-score. Precision is defined as the proportion of correctly estimated edges to all the edges in the estimated graph; recall is defined as the proportion of successfully estimated edges to all the edges in the true graph; F_1-score is defined as 2·precision·recall/(precision+recall), which is a weighted average of the precision and recall. We set $K = 10$, $L_s = 2$, and $M = 98$ when running the proposed stochastic variational inference algorithm. Due to the stochastic natural of the algorithm, the running time may vary depending on the variance of the stochastic gradient in each iteration. Therefore, we average the value of the running time as well as other criteria over 100 trials. All the simulations are running on a 20-core 3GHz CPU. Parallel computing is implemented for the penalty parameter selection procedure of the optimization model and the graphical lasso, as well as the forward-backward algorithm in the M subchains in the proposed model. The results are summarized in Table II.

TABLE II: Quantitative comparison of different models

Models	Bayesian Model	Optimization Model	Graphical Lasso
Change Points	$\{995.44, 2800, 3900, 4800\}$	$\{995, 2800, 3900, 4800\}$	N.A.
Precision	0.8962	0.9474	0.2855
Recall	0.9708	0.8417	0.6007
F_1-score	0.9316	0.8914	0.3870
Running Time	947.75	6763.51	23.69

We can find from the table that both the proposed Bayesian model and the optimization-based model performs well in terms of the accuracy of change point detection. The Bayesian model yields slightly better estimates of the first change point in some trials. The estimated position is 997, which is more close to the ground truth 1000. For graphical model inference, the recall of the Bayesian model is much larger than that of the optimization model, indicating that the proposed Bayesian model can reliably recover the true graph. On the other hand, the precision of the Bayesian model is slightly lower than the optimization, implying that the Bayesian model introduces few more false positives. In summary, compared with the optimization model, the proposed Bayesian model yields a relatively dense graph, successfully recovering the true graph

at the cost of including a few extra edges. Such slightly dense graphs are often favored in practice, as false positives can be identified in further analysis whereas false negatives are buried by the other absent edges. Another obvious advantage of the proposed model is that it is much faster than the optimization model. Finally, we notice that graphical lasso gives biased estimation to the graphical models of all segments, because of the wrong assumption that the data is stationary. It is therefore necessary to develop specific methods for non-stationary time series.

V. Conclusion and Future Work

In this paper, we focus on change point detection and graphical model inference for piecewise-stationary time series. We formulate the problem as inferring the hidden states and the emission distributions of a HMM. We further develop a low-complexity stochastic variational inference algorithm to learn the model. Numerical results show that the proposed model can automatically estimate the number and position of change points as well as the sparse graphical models without tuning any parameters.

One of our future work is to apply the proposed model to real data, such as multi-electrode brain recordings and financial time series. It is also interesting to investigate an online version of the proposed algorithm, since it is straightforward to incorporate new information in the stochastic optimization framework.

References

[1] A. T. Ihler, S. Krishner, M. Ghil, A. W. Robertson, and P. Smyth, "Graphical Models for Statistical Inference and Data Assimilation," *Physica D* vol. 230, pp. 72–87, 2007.

[2] H.-A. Loeliger, J. Dauwels, J. Hu, S. Korl, P. Li, and F. Kschischang, "The factor graph approach to model-based signal processing," *Proceedings of the IEEE* 95(6), pp. 1295–1322, 2007.

[3] J. Friedman, T. Hastie, and R. Tibshirani, "Sparse inverse covariance estimation with the graphical lasso," *Biostatistics*, vol. 9, no. 3, pp. 432-441, 2008.

[4] V. Chandrasekaran, P. A. Parrilo, and A. S. Willsky, "Latent Variable Graphical Model Selection via Convex Optimization," *The Annals of Statistics*, vol. 40, no. 4, pp. 1935–1967, 2012.

[5] H. Liu, J. Lafferty, and L. Wasserman, "The Nonparanormal: Semiparametric Estimation of High Dimensional Undirected Graphs," *Journal of Machine Learning Research,* pp. 2295–2328, 2010.

[6] H. Yu, J. Dauwels, and X. O. Wang, "Copula Gaussian Graphical Models with Hidden Variables," *Proceedings of ICASSP 2012*, pp. 2177–2180, 2012.

[7] H. Yu, J. Dauwels, X. Zhang, S. Y. Xu, and W. I. T. Uy, "Copula Gaussian Multiscale Graphical Models with Application to Geophysical Modeling," *Proceedings of 15th International Conference on Information Fusion*, pp. 1741–1748, 2012.

[8] J. Dauwels, H. Yu, S. Y. Xu, and X. O. Wang, "Copula Gaussian Graphical Model for Discrete Data", *Proceedings of ICASSP 2013*, pp. 2177–2180, 2013.

[9] 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", *The Journal of Neuroscience*, pp. 10076–10085, 2010.

[10] X. Xuan, and K. Murphy, "Modeling Changing Dependency Structure in Multivariate Time Series", *Proceedings of the 24th ICML*, 2007.

[11] I. Cribben, R. Haraldsdottir, L. Y. Atlas, T. D. Wager, and M. A. Lindquist, "Dynamic Connectivity Regression: Determining State-related Changes in Brain Conectivity", *NeuroImage* 61, pp. 907–920, 2012.

[12] M. Lavielle, and G. Teyssière, "Adaptive Detection of Multiple Change-Points in Asset Price Volatility", in: G. Teyssière and A. Kirman (Eds), *Long-Memory in Economics*, Springer, pp. 129–156, 2005.

[13] D. Angelosante, and G. B. Giannakis, "Sparse Graphical Modeling of Piecewise-Sationary Time Series", *Proceedings of ICASSP 2011*, pp. 1960–1963, 2011.

[14] H. Yu, C. Li, and J. Dauwels, "Network Inference and Change Point Detection for Piecewise-Stationary Time Series," *Proceedings of ICASSP 2014*, pp. 4498-4502, 2014.

[15] R. Killick, P. Fearnhead, I. A. Eckley, "Optimal detection of change-points with a linear computational cost," *Journal of the American Statistical Association* vol. 107, no. 500, pp. 1590–1598, 2012.

[16] M. D. Hoffman, D. M. Blei, C. Wang, and J. W. Paisley, "Stochastic variational inference," *Journal of Machine Learning Research*, vol. 14, no. 1, pp. 1303-1347, 2013.

[17] N. J. Foti, J. Xu, D. Laird, and E. B. Fox, "Stochastic Variational Inference for Hidden Markov Models," *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, 2014.

[18] D. Wulsin, E.B. Fox, and B. Litt, "Modeling the Complex Dynamics and Changing Correlations of Epileptic Events," *Artificial Intelligence*, vol. 216, pp. 55-75, 2014.

[19] S. I. M. Ko, T. T. L. Chong, and P. Ghosh, "Dirichlet Process Hidden Markov Multiple Change-point Model," *Bayesian Analysis*, pp. 1-22, 2015.

[20] J. Paisley, and L. Carin, "Hidden Markov Models with Stick-Breaking Priors," *IEEE transactions on Signal Processing*, vol. 57, no. 10, pp. 3905-3917, 2009.

[21] E. B. Fox, E. B. Sudderth, M. I. Jordan, A. S. Willsky, "A Sticky HDP-HMM with Application to Speaker Diarization," *Annals of Applied Statistics*, vol. 5, no. 2A, pp. 1020-1056, 2011.

[22] S. D. Babacan, M. Luessi, R. Molina, and A. K. Katsaggelos, "Sparse Bayesian Methods for Low-Rank Matrix Estimation," *IEEE trans. Signal Process.*, vol. 60, no. 8, pp. 3964-3977, 2012.

[23] H. Yu, and J. Dauwels, "Variational Bayes Learning of Multiscale Graphical Models," to be appear in *Proceedings of ICASSP 2015*, 2015.

[24] M. Chen, H. Wang, X. Liao, and L. Carin, "Bayesian Learning of Sparse Gaussian Graphical Models", *Technical report*, 2012.

[25] M. J. Beal, *Variational algorithms for approximate Bayesian inference*, PhD thesis, University of London, 2003.

[26] R. Ranganath, C. Wang, D. Blei, and E. Xing, "An adaptive learning rate for stochastic variational inference," *Proceedings of the 30th International Conference on Machine Learning (ICML)*, pp. 298-306, 2013.