Stochastic Variational Inference for Dynamic Correlated Topic Models

Federico Tomasi
federicot@spotify.com
Spotify

Praveen Ravichandran
praveenr@spotify.com
Spotify

Gal Levy-Fix*
gal.levy-fix@columbia.edu
Columbia University

Mounia Lalmas
mounial@spotify.com
Spotify

Zhenwen Dai
zhenwend@spotify.com
Spotify

Abstract

Correlated topic models (CTM) are useful tools for statistical analysis of documents. They explicitly capture the correlation between topics associated with each document. We propose an extension to CTM that models the evolution of both topic correlation and word co-occurrence over time. This allows us to identify the changes of topic correlations over time, *e.g.*, in the machine learning literature, the correlation between the topics "stochastic gradient descent" and "variational inference" increased in the last few years due to advances in stochastic variational inference methods. Our temporal dynamic priors are based on Gaussian processes (GPs), allowing us to capture diverse temporal behaviours such as smooth, with long-term memory, temporally concentrated, and periodic. The evolution of topic correlations is modeled through generalised Wishart processes (GWPs). We develop a stochastic variational inference method, which enables us to handle large sets of continuous temporal data. Our experiments applied to real world data demonstrate that our model can be used to effectively discover temporal patterns of topic distributions, words associated to topics and topic relationships.

1 INTRODUCTION

Topic models (Blei et al., 2003) are a popular class of tools to automatically analyse large sets of categorical data, including text documents or other data that can be represented as bag-of-words, such as images. Topic models have been widely used in various domains, *e.g.*, information retrieval (Blei et al., 2007; Mehrotra et al., 2013; Balikas et al., 2016), computational biology (Zhao et al., 2014; Gopalan et al., 2016; Kho et al., 2017), recommender systems (Liang et al., 2017) and computer vision (Fei-Fei & Perona, 2005; Kivinen et al., 2007; Chong et al., 2009). In the original topic model by Blei et al. (2003), which is also known as Latent Dirichlet Allocation (LDA), the words in a document come from a mixture of topics, where each topic is defined as a distribution over a vocabulary. The variations in the mixtures of topics across documents are captured by a Dirichlet distribution. However, a limitation is that it does not model the correlation in the co-occurrence of topics. To overcome this limitation, Blei & Lafferty (2006a) proposed the correlated topic models (CTM), which extends LDA with a correlated prior distribution for mixtures of topics.

An important piece of information associated with a textual document is when the document has been written. For human writings, both the meanings of topics, popularity and correlations among topics evolve over time. Modeling such evolution is very important for understanding the topics in a collection of documents across a period of time. For example, consider the topic *machine learning*. The distribution of the words associated with it has been gradually changing over the past few years, revolving around *neural networks*, shifting towards *support vector machines*, *kernel methods*, and finally again on *neural networks* and *deep learning*. In addition, due to the evolution of meaning, the topic *machine learning* probably increasingly correlates with *high performance computing* and *GPU* following the emerging of deep learning.

In this paper, we propose the dynamic correlated topic model (DCTM), which allows us to learn the temporal dynamics of all the relevant components in CTM. To model the evolution of the meanings of topics, we construct a temporal prior distribution for topic representation, which is derived from a set of Gaussian processes (GP). This enables us to handle documents in continuous time and to interpolate and extrapolate the topic representations at unseen time points. In CTM, the prior distribution for

*The work was part of internship at Spotify.

Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI), PMLR volume 124, 2020.

mixtures of topics is derived from a multivariate normal distribution, in which the mean encodes the popularity of individual topics while the covariance matrix encodes the co-occurrence of topics. We extend the prior for mixtures of topics into a dynamic distribution by providing a set of GPs as the prior for the mean, and a generalised Wishart Process (GWP) as the prior for the covariance matrices. With DCTM, apart from assuming the individual documents at a given time points are independently sampled, we can jointly model the evolution of the representations of topics, the popularity of topics and their correlations.

A major challenge applying topic models to real world applications is the scalability of the inference methods. A large group of topic models come with the inference methods based on Markov chain Monte Carlo (often Gibbs sampling in particular), which are hard to apply to corpora of millions of documents. To allow the model to deal with large datasets, we develop a stochastic variational inference method for DCTM. To enable mini-batch training, we use a deep neural network to encode the variational posterior of the mixtures of topics for individual documents. For the GPs and the generalised Wishart Process, we augment the model with auxiliary variables like in the stochastic variational GP (Hensman et al., 2013) to derive a scalable variational lower bound. As the final lower bound is intractable, we marginalise the discrete latent variables and apply a Monte Carlo sampling approximation with the reparameterisation trick, which allows us to have a low-variance estimate for the gradients.

The main contributions of this paper are as follows:

- We propose a full dynamic version of CTM, which allows us to model the evolution of the representations of topics, topic popularity and their correlations.
- We derive a stochastic variational inference method for DCTM, which enables mini-batch training and is scalable to millions of documents.

Outline. Section 2 discusses related work. Section 3 presents our novel contribution and the generalised dynamic correlated topic model. Section 4 describes an efficient variational inference procedure for our model, built on top of sparse Gaussian processes. Section 5 presents our experiments and validation of the model on real data. Section 6 concludes with a discussion and future research directions.

2 RELATED WORK

Static Topic Models. LDA was proposed by Blei et al. (2003) as a technique to infer a mixture of topics starting from a collection of documents. Each topic is a probability distribution over a vocabulary, and each topic is assumed to be independent from one another. However, such independent assumption usually does not hold in real world scenarios, in particular when the number of topics is large. The CTM (Blei & Lafferty, 2006a) relaxes this assumption, allowing us to infer correlated topics through the use of a logistic normal distribution. Similar models have been proposed with modifications to the prior distribution of the topics, in particular using a Gaussian process to model topic proportions while keeping topics static (Agovic & Banerjee, 2010; Hennig et al., 2012). However, the static nature of such models makes them unsuitable to model topics in a set of documents ordered by an evolving index, such as time.

Dynamic Topic Models. Topic models have been extended to allow for topics and words to change over time (Blei & Lafferty, 2006b; Wang et al., 2008b), making use of the inherent structure between documents appearing at different indices. These models considered latent Wiener processes, using a forward-backward learning algorithm, which requires a full pass through the data at every iterations if the number of time stamps is comparable with the total number of documents. A similar approach was proposed by Wang & McCallum (2006), with time being an observed variable. Such approach allowed for scalability, while losing the smoothness of inferred topics. Another scalable approach was proposed by Bhadury et al. (2016), to model large topic dynamics by relying on stochastic gradient MCMC sampling. However, such approach is still restricted to Wiener processes. Finally, Jähnichen et al. (2018) recently proposed a model that allows for scalability under a general framework to model time dependency, overcoming the limitation of Wiener processes. An attempt to model a latent correlation between topics in discrete time stamps has been shown in (Song et al., 2008), where topic correlation is computed using principal component analysis based on their closeness in the latent space. However, to the best of our knowledge, no general procedure has been proposed to explicitly model dynamic topic models with evolving correlations over continuous time.

Stochastic Variational Inference. We develop a scalable inference method for our model based on stochastic variational inference (SVI) (Hoffman et al., 2013), which combines variational inference with stochastic gradient estimation. Two key ingredients of our inference method are amortised inference and the reparameterisation trick (Kingma & Welling, 2014). Amortised inference has been widely used for enabling mini-batch training in the models with local latent variables such as variational autoencoder (Kingma & Welling, 2014) and

deep Gaussian processes (Dai et al., 2015). The reparamterisation trick allows us to obtain low-variance gradient estimates with Monte Carlo sampling for intractable variational lower bounds. Note that SVI is usually applied to the models, where the data points are i.i.d. given the global parameters such as Bayesian neural networks, which does apply to GP and GWP. Although the log marginal likelihood of GP and GWP cannot be easily approximated with data sub-sampling, we use the stochastic variational sparse GP formulation (Hensman et al., 2013), where an unbiased estimate of the variational lower bound could be derived from data sub-sampling, which is essential for mini-batch training. Recently, Jähnichen et al. (2018) developed a stochastic variational inference for DTM, which is a dynamic version of LDA. This is different from our approach, which is a dynamic version of CTM, where the correlations in the mixture of topics are modelled dynamically.

3 DYNAMIC CORRELATED TOPIC MODEL

DCTM is a correlated topic model in which the temporal dynamics are governed by GPs and GWPs. Consider a corpus W of documents associated with an index (for example a time stamp). We denote the index of a document as d and its time stamp as t_d. While taking into account the dynamics underlying the documents, our goal is twofold: *(i)* infer the vocabulary distributions for the topics, and *(ii)* infer the distribution of the mixture of topics. We use continuous processes to model the dynamics of words and topics, namely the Gaussian process. These incorporate temporal dynamics into the model, and capture diverse evolution patterns, maybe in the forms of smooth, with long-term memory or periodic.

Following the notation of the CTM (Blei & Lafferty, 2006a), we denote the probability of word w to be assigned to topic k as β_{wk}, and the probability of topic k for the document d as η_{dk}. DCTM assumes that a N_d-word document d at the time t_d is generated according to the following generative process:

1. Draw a mixture of topics $\boldsymbol{\eta}_d \sim \mathcal{N}(\boldsymbol{\mu}_{t_d}, \Sigma_{t_d})$;
2. For each word $n = 1, \ldots, N_d$:
   (a) Draw a topic assignment $z_n|\boldsymbol{\eta}_d$ from a multinomial distribution with the parameter $\sigma(\boldsymbol{\eta}_d)$;
   (b) Draw a word $w_n|z_n, \boldsymbol{\beta}$ from a multinomial distribution with the parameter $\sigma(\boldsymbol{\beta}_{z_n})$,

where σ represents the softmax function, *i.e.*, $\sigma(\boldsymbol{z})_i = e^{z_i}/\sum_{j=1}^{K} e^{z_j}$. Note that the softmax transformation is required for both $\boldsymbol{\eta}_d$ and $\boldsymbol{\beta}_{z_n}$, as they are assumed to be defined in an unconstrained space. The softmax transformation converts the parameters to probabilities, to encode the proportion of topics for document d and the distribution of the words for a topic $\boldsymbol{\beta}_{z_n}$, respectively.

Under this generative process, the marginal likelihood for corpus W becomes:

$$p(W|\boldsymbol{\mu}, \Sigma, \boldsymbol{\beta}) = \prod_{d=1}^{D} \int \left(\sum_{z_n=1}^{k} p(W_d|z_n, \boldsymbol{\beta}_{t_d}) p(z_n|\boldsymbol{\eta}_d) \right) p(\boldsymbol{\eta}_d|\boldsymbol{\mu}_{t_d}, \Sigma_{t_d}) \mathrm{d}\boldsymbol{\eta}_d. \tag{1}$$

The individual documents are assumed to be i.i.d. given the document-topic proportion and topic-word distribution.

The key idea of CTM is to relax the parameterisation of $\boldsymbol{\eta}$ by allowing topics to be correlated with each other, *i.e.*, by allowing a non-diagonal Σ_{t_d}. We follow the same intuition as in (Blei & Lafferty, 2006a), using a logistic normal distribution to model $\boldsymbol{\eta}$. This allows the probability of the topics to be correlated with each other. However, especially in the presence of a long period of time, we argue that it is unlikely that the correlations among topics remain constant. Intuitively, the degree of correlations among topics changes over time, as they simply reflect the co-occurrence of the concepts appearing in documents. Consider the correlation between the topics "stochastic gradient descent" and "variational inference", which increased in recent years due to advances in stochastic variational inference methods. We propose to model the dynamics of the covariance matrix of the topics, as well as the document-topic distribution and the topic-word distribution.

Dynamics of $\boldsymbol{\mu}$, $\boldsymbol{\beta}$ and Σ. First, we model the topic probability $(\boldsymbol{\mu}_{t_d})_{d=1}^{D}$ and the distribution of words for topics $(\boldsymbol{\beta}_{t_d})_{d=1}^{D}$ as zero-mean Gaussian processes, *i.e.*, $p(\boldsymbol{\mu}) = \mathcal{GP}(\mathbf{0}, \kappa_\mu)$ and $p(\boldsymbol{\beta}) = \mathcal{GP}(\mathbf{0}, \kappa_\beta)$. We model the series of covariance matrices $(\Sigma_{t_d})_{d=1}^{D}$ using generalised Wishart processes, a generalisation of Gaussian processes to positive semi-definite matrices (Wilson & Ghahramani, 2011; Heaukulani & van der Wilk, 2019). Wishart process are constructed from i.i.d. collections of Gaussian processes as follows. Let $\boldsymbol{f}$ be $D \times \nu$ i.i.d. Gaussian processes with zero mean function, so that

$$f_{di} \sim \mathcal{GP}(\mathbf{0}, \kappa_\theta), d \leq D, i \leq \nu \tag{2}$$

and (shared) kernel function κ_θ, where θ denotes any parameters of the kernel function. For example, in the case

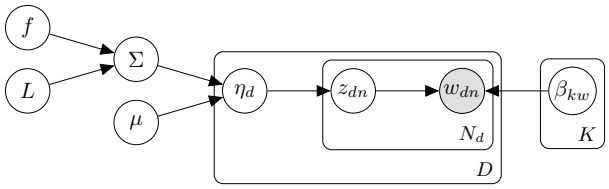

Figure 1: The graphical model for DCTM.

of $\kappa_\theta = \theta_1^2 * \exp(-||x-y||^2/(2*\theta_2^2))$, $\theta = (\theta_1, \theta_2)$ corresponds to the amplitude and length scale of the kernel (assumed to be independent from one another).

The positive integer-valued $\nu \geq D$ is denoted as the *degrees of freedom* parameter. Let $F_{ndk} := f_{dk}(\boldsymbol{x}_n)$, and let $F_n := (F_{ndk}, d \leq D, k \leq \nu)$ denote the $D \times \nu$ matrix of collected function values, for every $n \geq 1$. Then, consider

$$\Sigma_n = LF_nF_n^\top L^\top, n \geq 1, \tag{3}$$

where $L \in \mathbb{R}^{D\times D}$ satisfies the condition that the symmetric matrix $LL^\top$ is positive definite. With such construction, Σ_n is (marginally) Wishart distributed, and Σ is correspondingly called a Wishart process with degrees of freedom ν and scale matrix $V = LL^\top$. We denote $\Sigma_n \sim \mathcal{GWP}(V, \nu, \kappa_\theta)$ to indicate that Σ_n is drawn from a Wishart process. The dynamics of the process of covariance matrices Σ are inherited by the Gaussian processes, controlled by the kernel function κ_θ. With this formulation, the dependency between D Gaussian processes is static over time, and regulated by the matrix V.

We consider L to be a triangular Cholesky factor of the positive definite matrix V, with $M = D(D+1)/2$ free elements. We vectorise all the free elements into a vector $\boldsymbol{\ell} = (\ell_1, \ldots, \ell_M)$ and assign a spherical normal distribution $p(\ell_m) = \mathcal{N}(0, 1)$ to each of them. Note that the diagonal elements of L need to be positive. To ensure that, we apply *change of variable* to the prior distribution of the diagonal elements by applying a soft-plus transformation $\ell_i = \log(1 + \exp(\hat{\ell}_i)), \hat{\ell}_i \sim \mathcal{N}(0, 1)$. Hence, $p(L)$ is a set of independent normal distributions with diagonal entries constrained to be positive by a change of variable transformation.

Figure 1 shows the graphical model of DCTM.

Collapsing z's. Stochastic gradient estimation with discrete latent variables is difficult, often results into significantly higher variance in gradient estimation even with state-of-the-art variance reduction techniques. Fortunately, the discrete latent variables z in DCTM can be marginalised out in closed form. The resulting marginalised distribution $p(W_d|\boldsymbol{\eta}_d, \boldsymbol{\beta}_{t_d})$ becomes a multinomial distribution over the word-count in each document,

$$W_d \sim \prod_{n=1}^{N_d} \text{Multinomial}(1, \sigma(\boldsymbol{\beta}_{t_d}\boldsymbol{\eta}_d)). \tag{4}$$

This trick has also been used by Srivastava & Sutton (2017) to derive a variational lower bound for LDA.

4 VARIATIONAL INFERENCE

Given a collections of documents covering a period of time, we are interested in analysing the evolution of not only the word distributions of individual topics but also the evolution of the popularity of individual topics in the corpora and the correlations among topics. With the aim of handling millions of documents, we develop a stochastic variational inference method to perform mini-batch training with stochastic gradient descent methods.

4.1 AMORTISED INFERENCE FOR DOCUMENT-TOPIC PROPORTION

An essential component of the SVI method for DCTM is to enable mini-batch training over documents. After defining a variational posterior $q(\boldsymbol{\eta}_d)$ for each document, a variational lower bound of the log probability over the documents can be derived as follows,

$$\begin{aligned}
&\log p(W|\boldsymbol{\mu}, \Sigma, \boldsymbol{\beta}) \\
&\geq \sum_{d=1}^{D} \int q(\boldsymbol{\eta}_d) \log \frac{p(W_d|\boldsymbol{\eta}_d, \boldsymbol{\beta}_{t_d}) p(\boldsymbol{\eta}_d|\boldsymbol{\mu}_{t_d}, \Sigma_{t_d})}{q(\boldsymbol{\eta}_d)} \mathrm{d}\boldsymbol{\eta}_d \\
&= \sum_{d=1}^{D} \Big(\mathbb{E}_{q(\boldsymbol{\eta}_d)} \left[\log p(W_d|\boldsymbol{\eta}_d, \boldsymbol{\beta}_{t_d})\right] \\
&\qquad - \mathrm{KL}\left(q(\boldsymbol{\eta}_d)||p(\boldsymbol{\eta}_d|\boldsymbol{\mu}_{t_d}, \Sigma_{t_d})\right) \Big).
\end{aligned} \tag{5}$$

Denote the above lower bound as $\mathcal{L}_W$. As the lower bound is a summation over individual documents, it is straight-forward to derive a stochastic approximation of the summation by sub-sampling the documents,

$$\begin{aligned}
\mathcal{L}_W \approx \frac{D}{B} \sum_{i \in \mathcal{D}_\mathcal{B}} \Big(&\mathbb{E}_{q(\boldsymbol{\eta}_d)} \left[\log p(W_d|\boldsymbol{\eta}_d, \boldsymbol{\beta}_{t_d})\right] \\
&- \mathrm{KL}\left(q(\boldsymbol{\eta}_d)||p(\boldsymbol{\eta}_d|\boldsymbol{\mu}_{t_d}, \Sigma_{t_d})\right) \Big),
\end{aligned} \tag{6}$$

where $\mathcal{D}_\mathcal{B}$ is a random sub-sampling of the document indices with the size B. The above data sub-sampling allows us to perform mini-batch training, where the gradients of the variational parameters are stochastically approximated from a mini-batch. An issue with the above data sub-sampling is that only the variational parameters associated with the mini-batch get updated, which causes synchronisation issues when running stochastic

gradient descent. To avoid this, we assume the variational posteriors $q(\boldsymbol{\eta}_d)$ for individual documents are generated according to parametric functions,

$$q(\boldsymbol{\eta}_d) = \mathcal{N}(\phi_m(W_d), \phi_S(W_d)), \tag{7}$$

where ϕ_m and ϕ_S are the parametric functions that generate the mean and variance of $q(\boldsymbol{\eta}_d)$, respectively. This is known as amortised inference. With this parameterisation of the variational posteriors, a common set of parameters are always updated no matter which documents are sampled into the mini-batch, thus overcoming the synchronisation issue.

The lower bound $\mathcal{L}_W$ cannot be computed analytically. Instead, we compute an unbiased estimate of $\mathcal{L}_W$ via Monte Carlo sampling. As $q(\boldsymbol{\eta}_d)$ are normal distributions, we can easily obtain a low-variance estimate of the gradients of the variational parameters via the reparameterisation strategy (Kingma & Welling, 2014).

4.2 VARIATIONAL INFERENCE FOR GAUSSIAN PROCESSES

In DCTM, both the word distributions of topics $\boldsymbol{\beta}$ and the mean of the prior distribution of the document-topic proportion $\boldsymbol{\mu}$ follow Gaussian processes that take the time stamps of individual documents as inputs, i.e., $p(\boldsymbol{\beta}|\boldsymbol{t})$ and $p(\boldsymbol{\mu}|\boldsymbol{t})$. The inference of these Gaussian processes are due to the cubic computational complexity with respect to the number of documents. To scale the inference for real-world problems, we take a stochastic variational Gaussian process (Hensman et al., 2013, SVGP) approach to construct the variational lower bound of our model. We first augment each Gaussian process with a set of auxiliary variables with a set of corresponding time stamps, i.e.,

$$p(\boldsymbol{\beta}|\boldsymbol{t}) = \int p(\boldsymbol{\beta}|U_\beta, \boldsymbol{t}, \boldsymbol{z}_\beta) p(U_\beta|\boldsymbol{z}_\beta) \mathrm{d}U_\beta, \tag{8}$$

$$p(\boldsymbol{\mu}|\boldsymbol{t}) = \int p(\boldsymbol{\mu}|U_\mu, \boldsymbol{t}, \boldsymbol{z}_\mu) p(U_\mu|\boldsymbol{z}_\mu) \mathrm{d}U_\mu, \tag{9}$$

where U_β and U_μ are the auxiliary variables for $\boldsymbol{\beta}$ and $\boldsymbol{\mu}$ respectively and $\boldsymbol{z}_\beta$ and $\boldsymbol{z}_\mu$ are the corresponding time stamps. Both $p(\boldsymbol{\beta}|U_\beta, \boldsymbol{t}, \boldsymbol{z}_\beta)$ and $p(U_\beta|\boldsymbol{z}_\beta)$ follow the same Gaussian processes as the one for $p(\boldsymbol{\beta}|\boldsymbol{t})$, *i.e.*, these Gaussian processes have the mean and kernel functions. The same also applies to $p(\boldsymbol{\mu}|U_\mu, \boldsymbol{t}, \boldsymbol{z}_\mu)$ and $p(U_\mu|\boldsymbol{z}_\mu)$. Note that, as shown in Equations (8) and (9), the above augmentation does not change the prior distributions for $\boldsymbol{\beta}$ and $\boldsymbol{\mu}$.

The variational posteriors of $\boldsymbol{\beta}$ and $\boldsymbol{\mu}$ are constructed in a special form to enable efficient inference (Titsias, 2009): $q(\boldsymbol{\beta}, U_\beta) = p(\boldsymbol{\beta}|U_\beta)q(U_\beta)$ and $q(\boldsymbol{\mu}, U_\mu) = p(\boldsymbol{\mu}|U_\mu)q(U_\mu)$. Both $q(U_\beta)$ and $q(U_\mu)$ are multivariate normal distributions, in which the mean and covariance are variational parameters, $q(U_\beta) = \mathcal{N}(M_\beta, S_\beta)$, $q(U_\mu) = \mathcal{N}(M_\mu, S_\mu)$. $p(\boldsymbol{\beta}|U_\beta)$ and $p(\boldsymbol{\mu}|U_\mu)$ are conditional Gaussian processes, as defined in (Hensman et al., 2013). When $\boldsymbol{\beta}$ and $\boldsymbol{\mu}$ are used in the down-stream distributions, a lower bound can be derived,

$$\log p(\cdot|\boldsymbol{\beta}) \geq \mathbb{E}_{q(\boldsymbol{\beta})}[p(\cdot|\boldsymbol{\beta})] - \mathrm{KL}\left(q(U_\beta)||p(U_\beta)\right), \tag{10}$$

$$\log p(\cdot|\boldsymbol{\mu}) \geq \mathbb{E}_{q(\boldsymbol{\mu})}[p(\cdot|\boldsymbol{\mu})] - \mathrm{KL}\left(q(U_\mu)||p(U_\mu)\right), \tag{11}$$

where $q(\boldsymbol{\beta}) = \int p(\boldsymbol{\beta}|U_\beta)q(U_\beta)\mathrm{d}U_\beta$ and $q(\boldsymbol{\mu}) = \int p(\boldsymbol{\mu}|U_\mu)q(U_\mu)\mathrm{d}U_\mu$.

4.3 VARIATIONAL INFERENCE FOR WISHART PROCESSES

The generalised Wishart process for Σ is derived from a set of GPs. At each time point, the covariance matrix is defined as $\Sigma_t = LF_tF_t^\top L^\top$. The vector stacking each entry of the matrix F_t across all the time points, $\boldsymbol{f}_{ij} = ((F_1)_{ij}, \ldots, (F_T)_{ij}))$, follows a Gaussian process $p(\boldsymbol{f}_{ij}|\boldsymbol{t}) = \mathcal{GP}(0, \kappa)$. A stochastic variational inference method for the used Wishart Process could be derived similar to the GPs in the previous section. For each GP $p(\boldsymbol{f}_{ij}|\boldsymbol{t})$ in the Wishart process, we augment it with a set of auxiliary variables with a set of the corresponding time stamps,

$$p(\boldsymbol{f}_{ij}|\boldsymbol{t}) = \int p(\boldsymbol{f}_{ij}|\boldsymbol{u}_{ij}, \boldsymbol{t}, \boldsymbol{z}_{ij}) p(\boldsymbol{u}_{ij}|\boldsymbol{z}_{ij}) \mathrm{d}\boldsymbol{u}_{ij}, \tag{12}$$

where $\boldsymbol{u}_{ij}$ is the auxiliary variable, $\boldsymbol{z}_{ij}$ is the corresponding time stamps and $p(\boldsymbol{f}_{ij}|\boldsymbol{u}_{ij})$ is a conditional Gaussian process, as defined in (Hensman et al., 2013). We define the variational posterior of $\boldsymbol{f}_{ij}$ to be $q(\boldsymbol{f}_{ij}, \boldsymbol{u}_{ij}) = p(\boldsymbol{f}_{ij}|\boldsymbol{u}_{ij})q(\boldsymbol{u}_{ij})$, where $q(\boldsymbol{u}_{ij}) = \mathcal{N}(\boldsymbol{m}_{ij}, \boldsymbol{s}_{ij})$. We also define the variational posterior of $\boldsymbol{\ell}$ to be $q(\boldsymbol{\ell}) = \mathcal{N}(\boldsymbol{m}_\ell, S_\ell)$, where S_ℓ is a diagonal matrix. As the diagonal elements of L needs to be positive, we also apply *change of variable* to the variational posterior of the diagonal elements, *i.e.*, $\ell_m = \log(1 + \exp(\hat{\ell}_m))$, $q(\hat{\ell}_m) = \mathcal{N}(\boldsymbol{m}_{\ell_m}, S_{\ell_m})$.

Note that $\boldsymbol{z}_\beta$, $\boldsymbol{z}_\mu$ and $\boldsymbol{z}_{ij}$ are variational parameters and not random variables. For this reason, we will omit them from the notation for convenience.

With such a set of variational posterior for all the entries $\{\boldsymbol{f}_{ij}\}$ and $\boldsymbol{\ell}$, a variational lower bound could be derived, when Σ is used for some down-stream distributions,

$$\begin{aligned}\log p(\cdot|\Sigma) \geq \\ \mathbb{E}_{q(F)q(\boldsymbol{\ell})}[p(\cdot|\Sigma)] - \sum_{i,j} \mathrm{KL}\left(q(\boldsymbol{u}_{ij})||p(\boldsymbol{u}_{ij})\right) \\ - \mathrm{KL}\left(q(\boldsymbol{\ell})||p(\boldsymbol{\ell})\right),\end{aligned} \tag{13}$$

where $q(F) = \prod_{ij} q(\boldsymbol{f}_{ij})$ with $q(\boldsymbol{f}_{ij}) = \int p(\boldsymbol{f}_{ij}|\boldsymbol{u}_{ij})q(\boldsymbol{u}_{ij})\mathrm{d}\boldsymbol{u}_{ij}$.

4.4 INFERENCE FOR DCTM

After deriving the variational lower bound for all the components, we will show how the lower bounds of the individual components can be assembled together for a stochastic variational inference for DCTM. The document-topic proportion for each document d follows a prior distribution $p(\boldsymbol{\eta}_d|\boldsymbol{\mu}_{t_d}, \Sigma_{t_d})$, where the GP of $\boldsymbol{\mu}$ provides the mean and the generalised Wishart process for Σ provides the covariance matrix at the time stamp t_d. The word distributions for individual topics are used in defining the distribution of individual words for each document d, $p(W_d|\boldsymbol{\eta}_d, \boldsymbol{\beta}_{t_d})$. Combining the lower bounds (10), (11) and (13), we can derive the complete variational lower bound $\mathcal{L}$ of DCTM,

$$\begin{aligned}\log p(W) \geq & \\ & \mathbb{E}_{q(\boldsymbol{\mu})q(F)q(L)q(\boldsymbol{\beta})}\left[\mathcal{L}_W\right] - \mathrm{KL}\left(q(U_\beta)||p(U_\beta)\right) \\ & - \mathrm{KL}\left(q(U_\mu)||p(U_\mu)\right) - \sum_{i,j}\mathrm{KL}\left(q(\boldsymbol{u}_{ij})||p(\boldsymbol{u}_{ij})\right) \\ & - \mathrm{KL}\left(q(\boldsymbol{\ell})||p(\boldsymbol{\ell})\right) \\ & = \mathcal{L}.\end{aligned}$$

The first term of $\mathcal{L}$ can be further decomposed by plugging in Equation (5),

$$\begin{aligned}&\mathbb{E}_{q(\boldsymbol{\mu})q(F)q(L)q(\boldsymbol{\beta})}\left[\mathcal{L}_W\right] = \\ &\sum_{d=1}^{D}\Big(\mathbb{E}_{q(\boldsymbol{\eta}_d)q(\boldsymbol{\beta}_{t_d})}\left[\log p(W_d|\boldsymbol{\eta}_d, \boldsymbol{\beta}_{t_d})\right] \\ &- \mathbb{E}_{q(\boldsymbol{\eta}_d)q(\boldsymbol{\mu}_{t_d})q(F_{t_d})q(L)}\left[\mathrm{KL}\left(q(\boldsymbol{\eta}_d)||p(\boldsymbol{\eta}_d|\boldsymbol{\mu}_{t_d}, \Sigma_{t_d})\right)\right]\Big).\end{aligned} \tag{14}$$

This formulation allows us to easily perform mini-batch training by data sub-sampling. For each mini-batch, we randomly sub-sampling the data set and re-weight the term $\mathbb{E}_{q(\boldsymbol{\mu})q(F)q(L)q(\boldsymbol{\beta})}\left[\mathcal{L}_W\right]$ according to the ratio between the size of dataset and the size of the mini-batch as shown in Equation (6). Note that all variational parameters of $q(\boldsymbol{\mu})$, $q(F)$, $q(L)$, $q(\boldsymbol{\beta})$ are optimised.

5 EXPERIMENTS

We validated the DCTM on real datasets in which the documents are created over the course of a period of time. We demonstrate the ability of DCTM in terms of capturing not only the changes within a topic but also how the topic changes themselves affect their relationships. Below we present the details of the datasets used in this study to compare DCTM against the state-of-art topic models.

5.1 DATASETS

We used three time-stamped text corpora with different characteristics. Firstly, we used the *State of the Union* (SotU) corpus, a long-term dataset with a small number of documents for each time step (only one document per time step). Secondly, we analysed the *Department of Justice* (DoJ) dataset, which has a short short-span but includes an high number of documents. Finally, we validated our model on the *NeurIPS* dataset, that features a medium-length time span and includes more documents at every time point. The first and last dataset were also used by Jähnichen et al. (2018) to validate their dynamic topic models. We applied simple preprocessing techniques used in prior works to all three datasets: text tokenisation, punctuations removal, and filtering the stop words from a standard list of English stop words.[1] Further details about each dataset are provided below.

State of the Union corpus (1790-2018). The corpus represents an annual address by the President of the United States before a joint session of congress. The dataset includes one document per year, from 1790 to 2018 (229 years). After our preprocessing, our vocabulary includes 1112 words. Finally, we split the data into 170 documents as training and 57 documents as test data.

Department of justice press releases (2009-2018). The dataset includes 13087 press releases from the Department of Justice from 2009 to 2018, for a total of 115 unique time points. It includes information such as criminal cases, actions taken against felons or general current administration updates. Our preprocessing leads to 1801 unique tokens. Finally, the documents were split into 9674 for training and for 3413 testing.

NeurIPS conference papers (1987-2015). The dataset contains 5804 conference papers from 1987 to 2015 (29 years), with an average of 34 papers per year. Our preprocessing produces 1047 tokens. We used 4237 documents as training data and 1567 as test data.

5.2 MODEL COMPARISON

We compared our approach with the state-of-the-art topic models and evaluate the generalisation on unseen documents. We compare with the following models, including both static and dynamic models: *(i)* LDA model with standard online mean-field variational inference[1] (Hoffman et al., 2010); *(ii)* ProdLDA, an autoencoding inference method to learn an LDA model (Srivastava & Sutton, 2017). Similar to correlated topic models, ProdLDA approximates the $\boldsymbol{\eta}$ distribution using a logistic normal;

[1] Available in `scikit-learn` library.

(iii) CTM using variational inference (Blei & Lafferty, 2006a), where we also add the word-topic assignment (the parameter $\boldsymbol{\beta}$) as a variational parameter, using the same base framework as our model (but without the temporal component); *(iv)* FastDTM, an implementation of MCMC-based dynamic topic model that relies on latent Wiener processes (Bhadury et al., 2016); *(v)* gDTM, a continuous dynamic topic model that generalises the DTM, and includes a temporal dynamics to the word-topic distribution parameter $\boldsymbol{\beta}$ (Jähnichen et al., 2018).

5.3 IMPLEMENTATION DETAILS

All datasets were divided into two disjoint parts, namely training and test set, using 75% of documents for training and 25% for testing. To model a real world scenario, documents in the training data are associated with a temporal index disjoint to the index of the test documents.

In our experimental setting, we optimised the static models (LDA, ProdLDA, and CTM) for documents belonging to our datasets disregarding the actual temporal index that each document is associated with. Since FastDTM does not handle continuous dynamics, we discretised the index points and assigned each index point to the closest bin, *i.e.*, we split the datasets into 10 bins and grouped together adjacent time points. As the model does not allow us to predict during unseen time steps, we compute our metrics by simply matching the discretisation of the index points for the test set.

Model Parameters. In our experiments we used a Matérn kernel (with parameter $1/2$) for $\boldsymbol{\beta}$, as we allow topic to quickly incorporate new words (especially useful with neologisms). To model $\boldsymbol{\mu}$ and $\boldsymbol{f}$ we instead used a squared exponential kernel, as we allow more freedom to topic probability and their correlation to change rapidly.

We initialised amplitude and length scale of kernels as 1 and 0.1 respectively, which are then learnt in our model using the approximate empirical Bayes approach (Maritz, 2018).

Hyperparameters. Our models are implemented using TensorFlow (Dillon et al., 2017). Experiments were conducted using Adam optimiser with learning rate 0.001 and 2000 iterations. To validate the output of topic models based on variational inference we compute the perplexity using the exponential the average negative predictive log-likelihood for each word (Blei et al., 2003), where the ELBO for a test document d^* is computed using Equation (14). We experimented with different number of topics, and selected the best ones for each dataset (30 for NeurIPS and *DoJ*, 20 for *SotU*). We also experimented with a different number of inducing points for the three components $\boldsymbol{\beta}$, $\boldsymbol{\mu}$ and $\boldsymbol{f}$, which control the complexity of the variational posterior. We used 15/20/15 (NeurIPS), 10/10/10 (DoJ), 6/6/6 (SotU).

Table 1: Average per-word perplexity on test data of three real world datasets (the lower the better).

Dataset	LDA	ProdLDA	CTM	FastDTM	gDTM	DCTM (ours)
SotU	1340.82	896.15	1012.79	1317.72	1263.41	**790.12**
DoJ	663.70	873.62	1340.73	890.65	2248.93	**440.41**
NeurIPS	1033.18	1081.6	1028.54	874.06	3581.72	**505.88**

5.4 RESULTS

We report both quantitative and qualitative results on the three datasets previously described.

Quantitative Analysis. We report the per-word perplexity computed on the held-out test set of different models described in Section 5.2. The per-word perplexity is used as a measure of best fit to compare models and is computed as an exponent of the average negative ELBO per word (Wang et al., 2008a).

Table 1 shows the average per-word perplexity of our proposed model and various baselines on unseen test documents. We see that by considering temporal dynamics, our proposed model generalises better and captures the presence of correlated topics. In general, our proposed DCTM model outperforms all baselines on the three datasets considered in our evaluation, by capturing the change in topics over time. Amongst the three datasets considered in this study, *SotU* dataset covers the longest time span and includes only one document per time point. In such case we argue it is of fundamental importance for topic models to capture time dynamics. However, we note how a static topic model (ProdLDA) performs already quite well in this case by considering a perplexity score. This is related to the fact that, by considering a single document at each time point, ProdLDA is able to fix different topics and to correctly assign documents to the relevant topics (however without information on the temporal aspect of the topics). In such case extreme case of a single document per time point, we note how DCTM can choose to optimise for different topics (of which their proportion change over time but the words describing them stay constant) or to consider fixed topic proportions over time with words that change.

Next, the performance of our model on the *DoJ* corpus demonstrates the robustness of our model. While the performance of our proposed DCTM is better than the baselines, the perplexity values are also quite close to a simple LDA. This is somewhat expected because the dataset spans over a relatively short time period and we

Table 2: Top 30 most probable words associated to Topic 2 (*neural networks*), Topic 15 & 19 (*neuroscience*). These topics are highly correlated (as shown in Figure 3). This is also reflected by the common words shared among them (such as those highlighted).

Topic num	Words
2	layer unit hidden architectur input deep hinton output network weight connect recept net pool convolut modul activ represent propag epoch competit learn train gate unsupervis neural sigmoid code recurr sensori
15	voltag channel signal spike auditori circuit frequenc filter sound chip nois movement veloc record sourc decod motor analog gain delay current modul power conduct motion adapt respons tempor period neuron
19	neuron synapt fire cortic cortex stimulus cell synaps spike stimuli popul neurosci respons tune recept sensori orient biolog oscil correl mechan visual brain activ potenti connect spatial domin modul pattern

suspect that topics do not change much, while having a low correlation as documents are relatively short and only focused on a particular topic. The poor performance of CTM and gDTM models on this dataset confirms our expectations. Our results demonstrates how DCTM adapts well to datasets including mostly static topics.

We observe one of the biggest performance gain of using our model on the NeurIPS dataset, which spans quite a long-time and has are many documents associated with every time point. DCTM is able to correctly handle datasets when topics share similarities, as it is the case for the NeurIPS dataset, but the topics themselves may be identified by different words. Indeed, modeling the covariance structure between the topics through the use of a Wishart process allows not only to model topics correlation, but also their evolution over time. Particularly for the NeurIPS dataset this has shown to be strongly effective (as the dataset spans over almost 30 years), and topics are associated to possible different words (for example based on neologisms, or such as names of novel models).

Qualitative Analysis on NeurIPS dataset. Figure 2 shows a sample of the most relevant topics over time, computed based on distribution of the topics for each document. The topic distribution shows a decreasing trend in *neural networks* popularity after early 1990. Unfortunately, our dataset does not include documents after 2015, when the topic began popular again. Also, consistently with our prior knowledge, there is a decreasing trend for topics associated with 15 and 19. Both of these topics show a similar temporal pattern. By inspecting the words associated with them (Table 2), we can indeed interpret such topics as both related to the area of *neuroscience*, which indeed began have less presence in the conference especially after the year 2000.

[2]This is calculated as $\langle \sigma(\boldsymbol{\eta}_{t_*}) \rangle_{p(\boldsymbol{\eta}_{t_*}|D)}$, where $p(\boldsymbol{\eta}_{t_*}|D) = \int p(\boldsymbol{\eta}_{t_*}|\boldsymbol{\mu_{t_*}}, \Sigma_{t_*}) p(\boldsymbol{\mu_{t_*}}|D) p(\Sigma_{t_*}|D) \mathrm{d}\boldsymbol{\mu_{t_*}} \mathrm{d}\Sigma_{t_*}$.

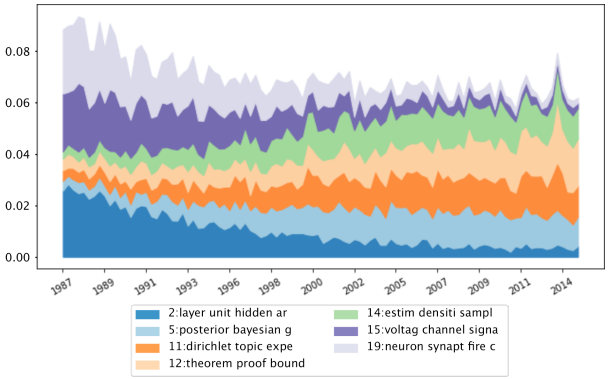

Figure 2: Evolution of the mean of the mixture of topics $\sigma(\boldsymbol{\eta})$ for a few selected topics across the time span predicted from the trained DCTM.[2] Topics 15 and 19 are both associated with *neuroscience*, which exhibit a decreasing trend over time. Similarly there is a decreasing trend for *neural networks* after 1990.

To understand the behaviour of topics over time we can use topic correlations as learnt by our model (Figure 3). The plot shows the temporal correlation between topic 19 (associated with *neuroscience*) and other topics, such as topic 15 (again, mostly associated with *neuroscience*) and topic 2 (*neural networks*). Consistently with their interpretation, topic 15 and 19 exhibit high level of correlation across time. Also, we note the increasing correlation with topic 2, associated with *neural networks*.

To better analyse this behaviour, Table 2 includes a list of the most relevant words for those topics. We highlighted the words shared among at least two of those topics. The table shows two fundamental features of the our model. First, our model is robust enough to discriminate between different topics that share the vocabulary (such as topics 2 with respect to both 15 and 19). Indeed, we point out how the literature on *neural networks* has been using words traditionally connected to the field of *neuroscience*. At the same time, DCTM is able to consider multiple topics with similar interpretation (such as topics 15 and 19), hence splitting facets of a single topic (which may be the case when the number of topics specified as a hyperparameter is too high).

Figure 4 shows another topic, which is related to the field of *topic modeling*. While the probability associated with this topic is stable over time (topic 11 in Figure 2), we note how the words associated with this topic change drastically over time. In particular, after their introduction in 2003, *topic modeling* has been referred to through the use of words such as *lda*, *dirichlet*, and *topic*. Before this date the model identified other words associated with this topic (or a closely related one), such as *mixture* and *expert*. Indeed, there is a strong connection between *topic*

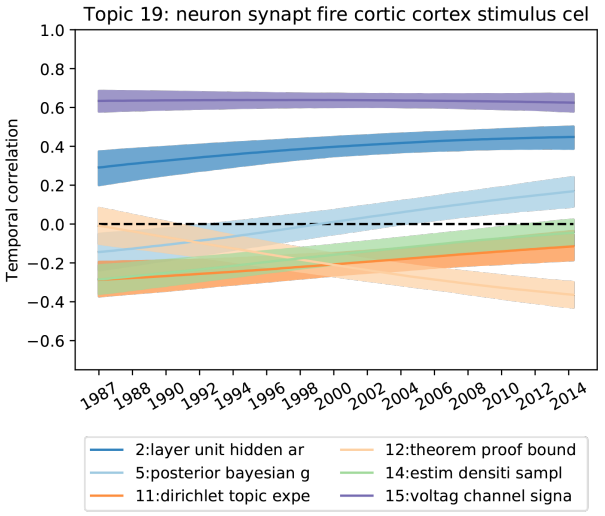

Figure 3: Evolution of the correlations between Topic 19 (*neuroscience*) and a few selected topics from DCTM. The solid line shows the 50th percentile and shaded area show the credible interval between 5th and 95th percentile. Topic 15 is consistently positively correlated with topic 19, which shares lost of similar words. The correlation with Topic 2 slowly increases over time. Instead, correlations with other topics are small and close to zero (dashed line), as they are mostly identified by disjoint sets of words.

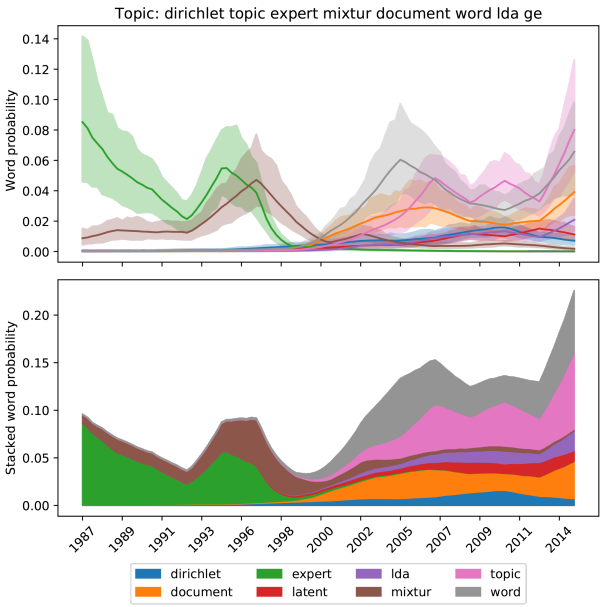

Figure 4: Evolution of word-topic probabilities over time for a topic. This is computed from the posterior predictive probability β. The top figure shows the 50th percentile in solid line and the credible interval between 5th and 95th percentile as shaded areas. The figure below shows the stacked word mean probability for a few words. While the word *topic* appears only after 2000, the model associate the words *mixture* and *expert* to this topic. These words have decreasing trends, balanced with the increasing trend of *lda* (after its introduction in 2003), *latent* and *dirichlet*. Such models have been mostly applied to textual data (reflected in the high probability of *document* and *word*).

modeling and *mixture models*, as the former can be seen as a particular case of the latter.

Modeling dynamics. We remark that we did not experiment with the choice of the kernel functions and their parameters. However, we argue that in real use cases it is better and beneficial to tune the model based on the data at hand to account for different temporal dynamics, possibly at the word level. Indeed, while in our experiment we included a single copy of the kernel for each word, it would be beneficial to allow for each topic to evolve according to a particular behaviour. We note that this is readily available in our framework, as it is general in the choice of the kernel function.

6 CONCLUSION

In this paper, we developed a novel dynamic correlated topic model that incorporates a time dynamics for each of the word-topic distribution, topic proportions, and topic correlations. Our model incorporates dynamics through the use of continuous processes, namely Gaussian and Wishart processes. We developed a stochastic variational inference for DCTM, which enables us to scale for large collections of documents. The inference of the covariance structure between the topics over time is beneficial to understand the similarity between topics. Modeling topic correlation is also fundamental in real use cases, as shown in our experiments, where topics are expected to be related. By considering such structure between topics, the model benefits as it improves statistical robustness and performance.

A current limitation of this approach is the long-term forecasting, due to the limitation of the stationary kernels used in DCTM. A solution would be to extend the model to use more sophisticated time-series models, such as forecasting and state-space models, which provide a general framework to analyse deterministic and stochastic dynamical systems (Hyndman & Athanasopoulos, 2018).

Finally, our proposed approach is able to scale to large datasets for two reasons. First, the use of sparse Gaussian process decreases the computational cost during inference. Secondly, using the stochastic variational inference method introduced, we are able to perform mini-batches on large datasets in a parallel.

References

Agovic, A. and Banerjee, A. Gaussian process topic models. In *UAI*, pp. 10–19, 2010.

Balikas, G., Amini, M.-R., and Clausel, M. On a topic model for sentences. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pp. 921–924, 2016.

Bhadury, A., Chen, J., Zhu, J., and Liu, S. Scaling up dynamic topic models. In *WWW*, pp. 381–390, 2016.

Blei, D. and Lafferty, J. Correlated topic models. *NIPS*, 18:147, 2006a.

Blei, D. M. and Lafferty, J. D. Dynamic topic models. In *ICML*, 2006b.

Blei, D. M., Ng, A., and Jordan, M. Latent dirichlet allocation. *JMLR*, 3:993–1022, 2003.

Blei, D. M., Lafferty, J. D., et al. A correlated topic model of science. *The Annals of Applied Statistics*, 1 (1):17–35, 2007.

Chong, W., Blei, D., and Li, F.-F. Simultaneous image classification and annotation. In *CVPR*, pp. 1903–1910. IEEE, 2009.

Dai, Z., Damianou, A., González, J., and Lawrence, N. Variational auto-encoded deep gaussian processes. In *ICLR*, 2015.

Dillon, J. V., Langmore, I., Tran, D., Brevdo, E., Vasudevan, S., Moore, D., Patton, B., Alemi, A., Hoffman, M., and Saurous, R. A. Tensorflow distributions. *arXiv preprint arXiv:1711.10604*, 2017.

Fei-Fei, L. and Perona, P. A bayesian hierarchical model for learning natural scene categories. In *CVPR*, volume 2, pp. 524–531. IEEE, 2005.

Gopalan, P., Hao, W., Blei, D. M., and Storey, J. D. Scaling probabilistic models of genetic variation to millions of humans. *Nature genetics*, 48(12):1587, 2016.

Heaukulani, C. and van der Wilk, M. Scalable bayesian dynamic covariance modeling with variational wishart and inverse wishart processes. In *NeurIPS 32*, pp. 4584–4594. Curran Associates, Inc., 2019.

Hennig, P., Stern, D., Herbrich, R., and Graepel, T. Kernel topic models. In *Artificial Intelligence and Statistics*, pp. 511–519, 2012.

Hensman, J., Fusi, N., and Lawrence, N. D. Gaussian processes for big data. In *UAI*, pp. 282. Citeseer, 2013.

Hoffman, M., Bach, F. R., and Blei, D. M. Online learning for latent dirichlet allocation. In *NIPS*, pp. 856–864, 2010.

Hoffman, M. D., Blei, D. M., Wang, C., and Paisley, J. Stochastic variational inference. *The Journal of Machine Learning Research*, 14(1):1303–1347, 2013.

Hyndman, R. J. and Athanasopoulos, G. *Forecasting: principles and practice*. OTexts, 2018.

Jähnichen, P., Wenzel, F., Kloft, M., and Mandt, S. Scalable generalized dynamic topic models. In *AISTATS*, pp. 1427–1435, 2018.

Kho, S. J., Yalamanchili, H. B., Raymer, M. L., and Sheth, A. P. A novel approach for classifying gene expression data using topic modeling. In *Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, pp. 388–393, 2017.

Kingma, D. and Welling, M. Auto-encoding variational bayes. In *ICLR*, 2014.

Kivinen, J. J., Sudderth, E. B., and Jordan, M. I. Learning multiscale representations of natural scenes using dirichlet processes. In *ICCV*, 2007.

Liang, Q., Zheng, X., Wang, M., Chen, H., and Lu, P. Optimize recommendation system with topic modeling and clustering. In *ICEBE*, pp. 15–22. IEEE, 2017.

Maritz, J. S. *Empirical Bayes methods with applications*. Chapman and Hall/CRC, 2018.

Mehrotra, R., Sanner, S., Buntine, W., and Xie, L. Improving lda topic models for microblogs via tweet pooling and automatic labeling. In *SIGIR*, 2013.

Song, Y., Zhang, L., and Giles, C. L. A non-parametric approach to pair-wise dynamic topic correlation detection. In *2008 Eighth IEEE International Conference on Data Mining*, pp. 1031–1036. IEEE, 2008.

Srivastava, A. and Sutton, C. Autoencoding variational inference for topic models. In *ICLR*, 2017.

Titsias, M. Variational learning of inducing variables in sparse gaussian processes. In *AISTATS*, pp. 567–574, 2009.

Wang, C., Blei, D., and Heckerman, D. Continuous time dynamic topic models. In *UAI*, pp. 579–586, Arlington, Virginia, USA, 2008a. AUAI Press. ISBN 0974903949.

Wang, C., Blei, D. M., and Heckerman, D. Continuous time dynamic topic models. In *UAI*, 2008b.

Wang, X. and McCallum, A. Topics over time: a non-markov continuous-time model of topical trends. In *KDD*, pp. 424–433, 2006.

Wilson, A. G. and Ghahramani, Z. Generalised wishart processes. In *UAI*, pp. 736–744. AUAI Press, 2011.

Zhao, W., Zou, W., and Chen, J. J. Topic modeling for cluster analysis of large biological and medical datasets. In *BMC bioinformatics*, volume 15, pp. S11. BioMed Central, 2014.