Scalable Generalized Dynamic Topic Models

Patrick Jähnichen [* 1]   **Florian Wenzel** [* 1 2]   **Marius Kloft** [1 2]   **Stephan Mandt** [3]

[1]Humboldt-Universität zu Berlin, Germany   [2]TU Kaiserslautern, Germany   [3]Columbia University, New York, USA

Abstract

Dynamic topic models (DTMs) model the evolution of prevalent themes in literature, online media, and other forms of text over time. DTMs assume that word co-occurrence statistics change continuously and therefore impose continuous stochastic process priors on their model parameters. These dynamical priors make inference much harder than in regular topic models, and also limit scalability. In this paper, we present several new results around DTMs. First, we extend the class of tractable priors from Wiener processes to the generic class of Gaussian processes (GPs). This allows us to explore topics that develop smoothly over time, that have a long-term memory or are temporally concentrated (for event detection). Second, we show how to perform scalable approximate inference in these models based on ideas around stochastic variational inference and sparse Gaussian processes. This way we can train a rich family of DTMs to massive data. Our experiments on several large-scale datasets show that our generalized model allows us to find interesting patterns that were not accessible by previous approaches.

1   Introduction

Probabilistic topic models help us to organize and browse large collections of documents (Blei, 2012). Topic models have been successfully applied in information retrieval (McCallum et al., 2004; Wang et al., 2007; Charlin and Zemel, 2013), computational biology (Pritchard et al., 2000; Gopalan et al., 2016), recommendation systems (Wang and Blei, 2011), and computer vision (Fei-Fei and Perona, 2005; Chong et al., 2009). Topic models assume that all words in a document were independently drawn from a finite set of probability distributions over words, termed the 'topics'. This way, every document is a mixture of topics. The limitation is that this approach assumes that topics are static.

Topics change over time. To provide some intuition, consider the example of the topic *technology* when training topic models on historical articles [1]. Restricting the corpus to articles around 1900, we find words such as *engine*, *electricity*, and *wire* to be mainly associated with this topic. For modern articles, we may find *devices*, *gates*, and *silicon* among the top words. In applications as this, we want to be able to associate documents with similar topic proportions with each other over large time spans. But at the same time, we want to allow topics to 'modernize', meaning to dynamically adjust their vocabulary. This is achieved in dynamic topic models (DTMs) (Blei and Lafferty, 2006; Wang and McCallum, 2006; Wang et al., 2008). DTMs model the evolution of topics as a continuous Wiener process. This dynamic prior determines how strongly topics may change their vocabulary. This way, DTMs share statistical strengths over all times, while giving the topics enough flexibility to change.

Current formulations of dynamic topic models are subject to the major limitation that they are restricted to a particular type of stochastic process for the latent topical dynamics, namely Wiener processes. This formulation does not allow us to analyze long-term effects, events, or other more complicated temporal dependencies. Second, relying on the forward-backward algorithm, they lack scalability. If the data are distributed across many different time-stamps, they require a full pass through the data in every iteration. This lack of scalability may be the reason why DTMs have been much less used in large-scale scientific or industrial applications than their static counterparts. In this paper, we generalize dynamic topic models in two ways: first we extend the class of tractable priors from Wiener processes to the more general class of Gaussian processes. Second, we derive a scalable approximate Bayesian inference algorithm based on inducing points. This allows us to apply our model to contemporary large text collections. In more detail, our main contributions are as follows:

- We formulate DTMs in terms of latent Gaussian process priors on topic evolution. This opens a wealth of possibilities for new models in which the topics display different types of temporal (or even spatial) correlations.

---

[1] Example from David Blei's tutorial slides on topic modeling, http://www.cs.columbia.edu/~blei/talks/Blei_ICML_2012.pdf

Going beyond the typical Wiener processes, we analyze Ornstein-Uhlenbeck processes for event detection, Gaussian processes with Cauchy kernels (for long-term memory effects) and squared exponential kernels (for rather short-term memory effects).

- We derive a scalable variational inference algorithm for this new model class. Our approach relies on inducing points for Gaussian process inference (Snelson and Ghahramani, 2006; Titsias, 2009; Hensman et al., 2013). All natural gradients are given in closed-form and do not rely on numerical optimization or sampling approaches. Natural gradients have the advantage that they are invariant to reparameterization of the variational family (Amari and Nagaoka, 2007; Martens, 2017) and provide effective second-order optimization updates (Hoffman et al., 2013; Wenzel et al., 2018). While a naive implementation would scale cubically in the number of time stamps, our approach scales cubically in the number of inducing points, which is typically much smaller.

- In our experiments, we investigate dynamic topics using different kernels. These new priors allow us to find patterns which were not accessible before. For instance, we filter time-localized topics in a set of speeches on the State of the Union and in news articles as published in the New York Times.

This paper is organized as follows. In section 2 we discuss related work. We describe the novel generalized dynamic topic model in section 3 and present an efficient variational inference algorithm for our model in section 4. Section 5 concludes with experiments. For implementation details, we refer readers to the website of the first author of this paper[2].

2   Background and Related Work

We connect to dynamic and correlated topic models, sparse GPs and stochastic variational inference (SVI).

Dynamic Topic Models.   DTMs form the basis of our approach. While Blei and Lafferty (2006) originally proposed a model with equidistant time slices, Wang et al. (2008) extended the approach to continuous time. Both rely on a latent Wiener process and use the forward-backward algorithm for learning, which requires full passes through the data in every iteration if the number of time stamps is comparable with the total number of documents. Wang and McCallum (2006) proposed a different approach where time is an observed variable with some prior over a finite time interval. While in principle being scalable, the resulting topics are non-smooth. Finally, Bhadury et al. (2016) proposed a new approach for learning in topic models based on

---

[2]https://patrickjae.github.io

stochastic gradient MCMC (Welling and Teh, 2011; Mandt et al., 2016). Their approach similarly is restricted to latent Wiener processes.

Correlated and GP Topic Models.   This class of modified static topic models breaks the independence assumptions of the per-document topic proportions. Instead, the topic proportions are jointly drawn from some prior which induces correlations (Blei and Lafferty, 2007). If this prior is a Gaussian process, this leads to the kernel topic model (Hennig et al., 2012) or Gaussian process topic model (Agovic and Banerjee, 2012). Note that both approaches assume that the topics themselves are static and only the topic proportions change. In contrast, we treat the proportions as independent and identically distributed (iid) and impose dynamics on the topics themselves. None of these models have been formulated in a scalable manner.

Stochastic Variational Inference and sparse GPs. Our algorithm builds on stochastic variational inference (SVI) (Hoffman et al., 2013), which combines variational inference with stochastic optimization. SVI can normally only applied if the data are iid conditioned on a global set of paramaters, which is an assumption that is typically broken in Gaussian process modelling setups. Hensman et al. (2013, 2012) have shown that one can derive a tractable lower bound to the marginal likelihood of the data that allows for data subsampling. This so-called inducing point or sparse approach dates back to earlier work by Titsias (2009); Snelson and Ghahramani (2006) and Csató and Opper (2002) and has been successfully applied to a variaty of GP models (e.g. Hensman et al., 2015; Wenzel et al., 2017). None of this work has been applied in the context of topic models.

3   Generalized Dynamic Topic Models

Dynamic topic models are mixed-membership bag-of-words models which allow their mixture components—the topics—to drift over time. This allows to dynamically fade-in new words, and fade-out old words which loose their semantic significance in a topic. In the classic DTM this continuity is achieved by imposing a Wiener process prior on the topic matrices (Blei and Lafferty, 2006; Wang et al., 2008) (see also (Bamler and Mandt, 2017) for a related approach for word embeddings).

In this paper, we propose Gaussian processes as priors on the topic matrices. Since the Wiener process is a specific type of GP, our approach is a strict generalization of dynamic topic models but covers a much richer class of dynamics. We introduce the generalized dynamic topic model in section 3.1 and present a scalable version in section 3.2.

3.1 Generalized DTMs

For what follows, we borrow notation from the topic modeling literature (Blei et al., 2003). We assume that we observe a corpus of D documents, each of which is associated with a time stamp τ_{t_d} with index $t_d \in \{1, \ldots, T\}$. For a simpler notation we denote the number of words in a document as N. For a given document d with time index t, let $w_{d1}, \cdots, w_{dN}$ be the words it contains, θ_d be a K-vector of topic proportions and z_{dn} the assignment of word w_{dn} to a topic. The model consists of K time dynamic topics whereby $\beta_{k \cdot t}$ denotes a topic's V-dimensional distribution over the vocabulary at time t.

Our model exhibits the following joint distribution:

$$p(w, z, \theta, \beta) = p(\beta) \prod_{t=1}^{T} \prod_{k=1}^{K} p(w_t, z_t, \theta | \pi(\beta_{k \cdot t})). \quad (1)$$

The function $\pi(\cdot)$ is the softmax function which normalizes the topic $\beta_{k \cdot t}$ over the vocabulary indices. The remaining likelihood,

$$p(w_t, z_t, \theta | \pi(\beta_{\cdot \cdot t})) =$$
$$\prod_{d:t_d=t} p(\theta_d) \prod_n p(w_{dn} | \pi(\beta_{z_{dn} \cdot t})) p(z_{dn} | \theta_d),$$

is just a regular LDA model (at time t), where $p(w_{dn} | \pi(\beta_{z_{dn} \cdot t})) = \text{Mult}(\pi(\beta_{z_{dn} \cdot t}))$, $p(z_{dn} | \theta_d) = \text{Mult}(\theta_d)$, $p(\theta_d) = \text{Dir}(\alpha)$. The graphical model is shown in Figure 1.

The distinctive feature of dynamic topic models is their dynamic prior $p(\beta)$. In our model each of the V words out of K topics is a latent function over time, drawn from a GP with kernel function κ. This GP is observed at times $\tau_1, \cdots, \tau_T$ and can thus be described as a T-dimensional multivariate normal distribution[3]:

$$\beta_{kw \cdot} \sim \text{GP}(0, \kappa) \Leftrightarrow \beta_{kw, 1:T} \sim \mathcal{N}_T(0, K_{TT}), \quad (2)$$
$$K_{\tau, \tau'} = \kappa(\tau, \tau'), \quad \tau, \tau' \in \{\tau_1, \cdots, \tau_T\}. \quad (3)$$

Using a Wiener kernel function in our model results in the classic DTM of Wang et al. (2008). However, due to the model's flexibility we can model *any* stochastic process that falls into the class of GPs by simply altering the covariance function κ. As an aside, this setup not only covers the dynamic setup, but also allows for incorporating other types of meta data as e.g. spatial modeling if the text documents are associated with location coordinates.

In this paper, we focus on the time-specific setup. In more detail, we consider several different kernels commonly used for time-series modeling (Roberts et al., 2012).

---

[3]We call attention to the slight overloading of notation: a plain K always is the number of topics, using subscripts or a tilde it denotes a kernel/covariance matrix.
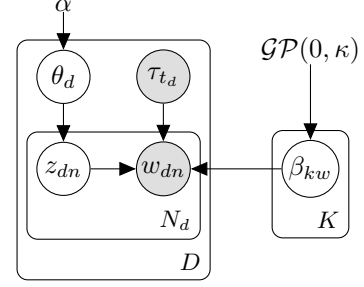

Figure 1: The generalized dynamic topic model.

- **Wiener kernels**, $\kappa_{\text{Wie}}(\tau, \tau') = \sigma^2 \min(\tau, \tau')$. Using a Wiener kernel (Brownian motion kernel) in our model recovers the typical DTM setup. This serves as our baseline.

- **Ornstein-Uhlenbeck kernels**, $\kappa_{\text{OU}}(\tau, \tau') = \sigma^2 \exp\left(-\frac{|\tau - \tau'|}{l}\right)$. The Ornstein-Uhlenbeck (OU) process is essentially a Wiener process in the presence of a mean-reverting force which pulls the process state back to its mean and thus acts like a regularizer. An effect of this is that topics may die-off and other topics may dynamically emerge (using a zero-mean process). As we show in our experiments, this leads to temporally localized changes in topics.

- **Squared exponential kernels**, $\kappa_{\text{SE}}(\tau, \tau') = \sigma^2 \exp\left(-\frac{(\tau - \tau')^2}{2l^2}\right)$. Squared exponential (SE) kernels have the property that the resulting trajectories are smoother compared to Wiener kernels. The resulting prior functions are infinitely often differentiable. The exponential decay of the temporal correlations leads to memory effects that can be parameterized by the kernel's length scale l. With a suitable chosen l this allows for temporally localized topics.

- **Cauchy Kernels**, $\kappa_{\text{Cau}}(\tau, \tau') = \sigma^2 \left(1 + \frac{(\tau - \tau')^2}{l^2}\right)^{-1}$. Cauchy kernels are constructed similarly as SE kernels, but instead of using the Gaussian density one uses a Cauchy density. This kernel has long-range memory, which means that temporal correlations decay not exponentially but polynomially, which in some cases is more realistic.

Note that *any* additive or multiplicative combination of covariance functions again results in a valid covariance function again and so can similarly be used. This adds considerable to the flexibility of the proposed prior.

We again stress that all these kernels use the same inference algorithm. The problem is that a naive implementation would scale cubically in the number of time stamps. We therefore propose a more efficient version based on the concept of sparse GPs.

3.2 Sparse DTMs

The bottleneck of inference in the model introduced in section 3.1 is the inversion of the $T \times T$ kernel matrix K_{TT}. One solution is to bin time stamps into groups, thus artificially reducing T. But this way we loose valuable information, especially when the number of distinct time stamps is comparable to the number of observed documents themselves, i.e. $T \approx D$.

Instead, we present a scalable version of the generalized dynamic topic model based on inducing points (Hensman et al., 2013). This is a low-rank approximation to the T-dimensional GPs based on $\hat{T}$ artificial time stamps (inducing points) where $\hat{T} \ll T$. The inversion necessary for inference is only based on the $\hat{T} \times \hat{T}$ covariance matrix of the approximating GP and can therefore be computed efficiently.

Following Hensman et al. (2013), let K_{TT} be the kernel evaluated at all training points (i.e. the full rank kernel as in (2)), $K_{\hat{T}\hat{T}}$ the kernel evaluated at inducing points, and $K_{T\hat{T}}$ and $K_{\hat{T}T}$ be kernels evaluated in-between these sets of points. Furthermore, let u be a $\hat{T}$-dimensional variable. We make use of the following Gaussian integral:

$$\mathcal{N}(0, K_{TT}) = \int \mathcal{N}(K_{T\hat{T}}K_{\hat{T}\hat{T}}^{-1}u, \tilde{K})\mathcal{N}(u; 0, K_{\hat{T}\hat{T}})\mathrm{d}u, \tag{4}$$

where $\tilde{K} = K_{TT} - K_{T\hat{T}}K_{\hat{T}\hat{T}}^{-1}K_{\hat{T}T}$. Thus, we introduce latent auxiliary variables u_{kw} for every β_{kw} such that the resulting marginal distribution of β_{kw} does not change (when integrating over u_{kw}). Defining $p(u_{kw}) = \mathcal{N}(0, K_{\hat{T}\hat{T}})$, we obtain

$$p(\beta_{kw}|u_{kw}) = \mathcal{N}(K_{T\hat{T}}K_{\hat{T}\hat{T}}^{-1}u_{kw}, \tilde{K}), \tag{5}$$

and perform approximate inference over u. Also note that conditioning of GPs involves inversion of the kernel matrix. In our approach, inverting a $T \times T$ matrix is now replaced by inverting one of size $\hat{T} \times \hat{T}$.

The augmented joint distribution is

$$p(\beta, w, z, \theta, u) = p(w|\beta, z)p(z|\theta)p(\beta|u)p(u). \tag{6}$$

This summarizes our model (the discussion of marginalizing over β is deferred to the next section). Next, we present details about the inference procedure. Readers primarily interested in experimental results may therefore skip section 4 and continue with section 5.

4 Inference

In Bayesian latent variable models such as DTMs, our goal is to compute the posterior distribution over the latent variables. This quantity is intractable and we have to resort to approximate methods. We use variational inference, which

maps the inference problem to an optimization problem, minimizing Kullback-Leibler divergence between a simple proxy distribution and the posterior. This is equivalent to optimizing a lower bound to the marginal likelihood of the model, termed evidence lower bound (ELBO) (Jordan et al., 1999). In particular, we use stochastic variational inference (SVI) (Hoffman et al., 2013), which optimizes the ELBO using stochastic gradient descent.

We first carry-out the approximate marginalization over β, lower-bounding the likelihood term. We then show how we can decompose the ELBO into a part which is equivalent to LDA, and into another part which contains the GP prior and therefore is more complex. We list all modified updates on the local and global parameters, with detailed calculations given in the appendix.

Approximate marginalization. We first marginalize over β in the augmented joint distribution (6). Unfortunately, the marginal likelihood term cannot be computed in closed-form. We use Jensen's inequality to obtain a lower bound on the log likelihood,

$$\log p(w_{dn}|z_{dn} = k, u, t_d) \tag{7}$$
$$= \log \mathbb{E}_{p(\beta_{k \cdot t_d}|u)}\left[p(w_{dn}|z_{dn} = k, \beta_{k \cdot t_d}]\right.$$
$$\geq \mathbb{E}_{p((\beta_{k \cdot t_d}|u)}\left[\log \mathrm{Mult}\left(w_{dn}|\pi(\beta_{k \cdot t_d})\right)\right]$$
$$= K_{t_d\hat{T}}K_{\hat{T}\hat{T}}^{-1}u_k w_{dn} - \mathbb{E}_{p((\beta_{k \cdot t_d}|u)}\left[\log \sum_w \exp(\beta_{kwt})\right], \tag{8}$$

where u_k is a $\hat{T} \times V$ matrix and $K_{t_d\hat{T}}$ is the t_d-th row of $K_{T\hat{T}}$. The remaining expectation in (8) is still intractable due to the sum inside of the logarithm. Following Blei and Lafferty (2006), we introduce additional free variational parameters ζ_{kt} (see appendix). This results in a lower bound to $\log p(w_{dn}|z_{dn} = k, u, t_d)$:

$$\log\tilde{p}(w_{dn}|z_{dn} = k, u, t_d) \tag{9}$$
$$= K_{t_d\hat{T}}K_{\hat{T}\hat{T}}^{-1}u_k w_{dn}$$
$$- \zeta_{kt_d}^{-1}\sum_w \exp\left(K_{t_d\hat{T}}K_{\hat{T}\hat{T}}^{-1}u_{kwt_d} + \frac{\tilde{K}_{t_d t_d}}{2}\right)$$
$$- \log(\zeta_{kt_d}) + 1.$$

Next, we use this lower bounded log-likelihood to derive a tractable variational objective which we can optimize.

Stochastic Variational Inference. We follow a variational structured mean-field approach (Wainwright and Jordan, 2007) and impose the following variational distributions on the latent variables, $q(\theta_d|\lambda_d) = \mathrm{Dir}(\lambda_d)$, $q(z_{dn}|\phi_{dn}) = \mathrm{Mult}(\phi_{dn})$ and $q(u_{kw}|\mu_{kw}, \Sigma_{kw}) = \mathcal{N}_T(\mu_{kw}, \Sigma_{kw})$. Eq. (9) gives rise to the following tractable lower bound of the

marginal likelihood:

$$\mathcal{L}(\lambda, \phi, \mu, \Sigma, \zeta) = \underbrace{\mathbb{E}_q[\log \tilde{p}(w|u,z)]}_{\mathcal{L}_1} \qquad (10)$$
$$+ \underbrace{\mathbb{E}_q[\log(p(z|\theta)p(\theta)p(u))]}_{\mathcal{L}_2} + \underbrace{H(q)}_{\text{Entropy}}.$$

The entropy term and $\mathcal{L}_2$ consist of standard results and a part that can be computed similarly as in standard LDA (see appendix). We also compute $\mathcal{L}_1$ in closed form:

$$\mathcal{L}_1 = \sum_{t,k,w} \sum_{d:t_d=t} n_{dw}\phi_{dwk} \left\{ m_{kwt} - \log(\zeta_{kt}) + 1 \right.$$
$$\left. - \zeta_{kt}^{-1} \sum_{w'} \exp\left(m_{kw't} + \frac{1}{2}(\Lambda_{kw't} + \tilde{K}_{tt}) \right) \right\},$$

with

$$m_{kwt} = K_{t\hat{T}} K_{\hat{T}\hat{T}}^{-1} \mu_{kw}$$
$$\Lambda_{kwt} = K_{t\hat{T}} K_{\hat{T}\hat{T}}^{-1} \Sigma_{kw} K_{\hat{T}\hat{T}}^{-1} K_{\hat{T}t}.$$

Objective $\mathcal{L}$ is optimized using SVI (Hoffman et al., 2013), i.e. for global variational parameters, we follow noisy natural gradients based on minibatches. Local variational parameter updates are similar to those in (Wang et al., 2008) and we do not replicate them here. Further details are provided in the appendix.

Global updates. We consider the Gaussian distributions $q(u_{kw})$ in *natural parameterization*, i.e. using the parameters $\eta_{kw}^{(1)} = \Sigma_{kw}^{-1}\mu_{kw}$ and $\eta_{kw}^{(2)} = -\frac{1}{2}\Sigma_{kw}^{-1}$, where μ_{kw} are the Gaussian means and Σ_{kw} the covariances. In SVI, we update these global parameters using stochastic estimates of the natural gradient and it turns out that in this case natural parameters result in simpler and more effective updates.

More specifically, for a Gaussian distribution, properties of the Fisher information matrix expose the simplification that the *natural gradient* w.r.t. the natural parameters can be expressed in terms of the *Euclidean gradient* w.r.t. the canonical parameters (i.e. mean and covariance). Namely, in general it holds for objectives $\mathcal{F}$ that depend on a Gaussian distribution that

$$\hat{\nabla}_{(\eta_1,\eta_2)}\mathcal{F}(\eta) = \left(\nabla_\mu \mathcal{F}(\eta) - 2\nabla_\Sigma \mathcal{F}(\eta)\mu, \ \nabla_\Sigma \mathcal{F}(\eta) \right), \qquad (11)$$

where $\hat{\nabla}$ denotes the natural gradient and ∇ the Euclidean gradient. Applying (11) to the variational objective (10), we obtain

$$\hat{\nabla}_{\eta_{kw}^{(1)}}\mathcal{L} = \Xi_{kw} + B_{kw} \circ (m_{kw} - 1) - \eta_{kw}^{(1)},$$
$$\hat{\nabla}_{\eta_{kw}^{(2)}}\mathcal{L} = -\frac{1}{2}K_{\hat{T}\hat{T}}^{-1} - \frac{1}{2}C_{kw} - \eta_{kw}^{(2)}. \qquad (12)$$

We used the following abbreviations:

$$\Xi_{kw} = K_{\hat{T}\hat{T}}^{-1} \sum_t \sum_{d:t_d=t} K_{\hat{T}t} n_{dw} \phi_{dwk},$$
$$B_{kw} = \sum_t \sum_{d:t_d=t} \zeta_{kt}^{-1} n_{dw} \phi_{dwk}$$
$$\times \exp\left(m_{kwt} + \frac{\Lambda_{kwt} + \tilde{K}_{tt}}{2} \right) K_{\hat{T}\hat{T}}^{-1} K_{\hat{T}t},$$
$$C_{kw} = B_{kw} K_{t\hat{T}} K_{\hat{T}\hat{T}}^{-1}.$$

Above, $\circ$ denotes the Hadamard product. Details are provided the appendix. Iterating through those updates completes the algorithm.

5 Experiments

We evaluate our method on three time-stamped text corpora. Compared to standard DTMs with Wiener kernels, we find that incorporating other dynamic priors may lead to improved predictive likelihoods and perplexity on held-out data. Using different kernel functions within our framework, we find new insights in the data that could not be found using the classic DTM of Wang et al. (2008), which uses the Wiener kernel and thus results in an unbounded variance over the time span. This promotes topics that are consistent, albeit relatively static.

By making use of the greater flexibility that comes with general GPs, we show how to extend and enhance an analysis. For instance, by using the *OU kernel*, we introduce a mean reverting force that quickly "draws" word probabilities towards zero, resulting in topics that are consistent and constrained in time and more sensitive to changes. Further, in situations in which the classic approach collapses most of the probability mass to single words per time stamp, we compensate by using *Cauchy kernels*, which place a smooth filter on word probabilities onto the topic over time. On the other hand, more fine-grained temporal dynamics can be captured by *RBF kernels*, due to their short-range memory.

Data and preprocessing.
1. We use the "The New York Times Annotated Corpus" (**NYT**) (Sandhaus, 2008), which consists of over 1.8 million articles published between 1987 and 2007 with $T = 7475$ unique time stamps. We subsample 100000 documents.

2. We use the **NIPS** dataset that contains 2711 papers from the NIPS conferences between 1987 and 2006[4] resulting in $T = 19$ time stamps.

3. We use the "State of the Union" (**SoU**) addresses of U.S. presidents, which span more than two centuries, resulting in

---

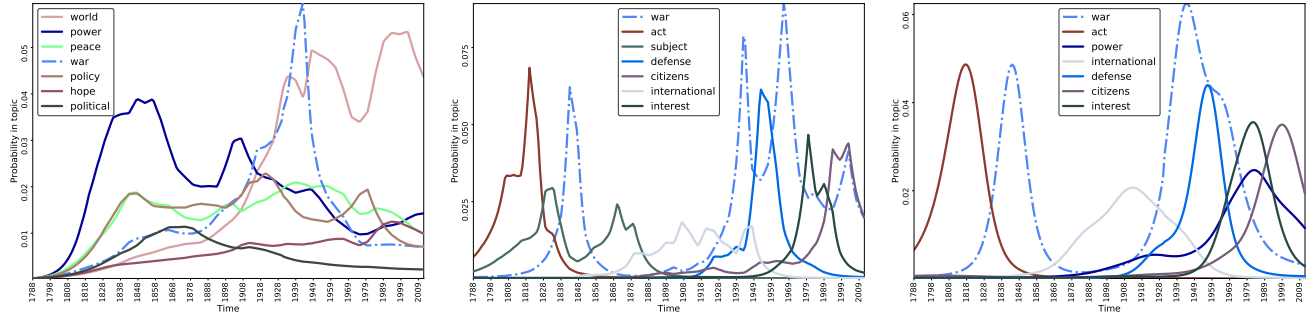[4] http://www.datalab.uci.edu/author-topic/NIPs.htm

Figure 2: SoU: Learned word trajectories of the "war" topic using the Wiener kernel (left), OU kernel (middle) and Cauchy kernel (right). The Cauchy kernel provides smoother trajectories yet the OU kernel is able to provide a better resolution in time. Both outperform the baseline in terms of perplexity.
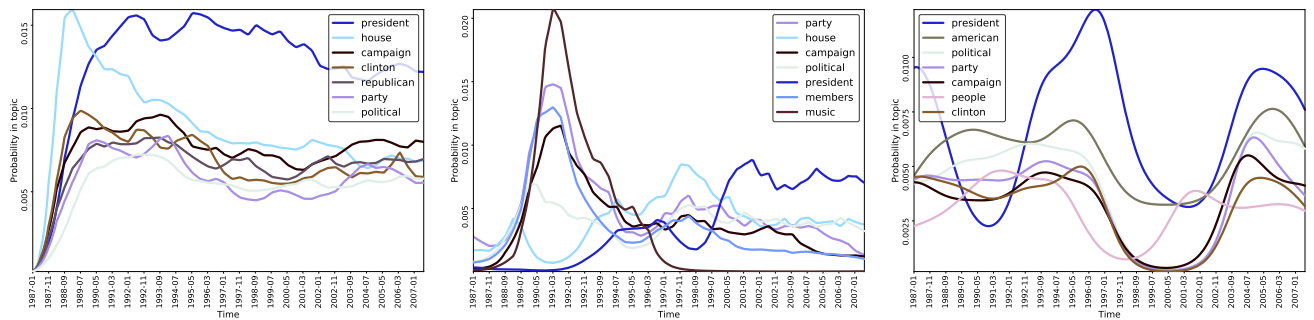

Figure 3: NYT: Learned word trajectories of the "election campaign" topic using the Wiener kernel (left), OU kernel (middle) and Cauchy kernel (right), which results in the smoothest curves.

$T = 224$ different time stamps [5]. We increase the number of documents to 4428 by treating every chunk of ten paragraphs in a speech as a separate document.

For preprocessing, we filtered the raw data using a standard stop word list. After collecting word statistics, we remove words that appear less than 25 times across a whole corpus. We further shrink the vocabulary by removing words whose score is less than a certain threshold, resulting in dictionaries of reasonable size (see appendix). After this step, we remove documents with effective lengths less than ten word occurrences. We initialize our models by randomly selecting K documents for any given time stamp and setting probabilities in topic k of occurring words proportional to their frequencies in the k-th document.

Hyperparameters. In our experiments we select the hyperparameters via grid search but they could also be directly learned in our inference scheme using the approximate empirical Bayes approach (Maritz and Lwin, 1989).

Qualitative Results. We now discuss the qualitative results obtained from applying our model on all three corpora. For certain example topics, we plot and discuss the

---

probabilities of the most important words in this topic over time. As a general tendency, we find that our proposed Ornstein-Uhlenbeck and Cauchy kernels outperform the standard Wiener kernel in terms of interpretability and in terms of usefulness for detecting events.

SoU. We consider a topic of with war and peace, found when fitting our generalized DTM to the state-of-the-union corpus. Figure 2 shows the word probabilities within this topic over time for all three considered kernels. The Wiener kernel is able to find a semantically coherent word distribution for this topic. We observe a relatively high probability of the term *"war"* over the whole time span with a sharp peak around 1939 (World War II). Using the Cauchy kernel, we are able to gain a better resolution of the dynamics for the importance of this term. We observe two separate high-probability periods of the word "war". One is matching the time of the American-Mexican war 1846-1848, the other one the World Wars and Vietnam war. We attribute this finding to the fact that the Cauchy kernel shares more statistical strength over time due to its long-term memory property.

While this model already provides a better insight into active time periods of the topic, additionally introducing a mean-reverting force via the OU kernel provides a mean to "super-resolve" topic activity quite accurately to certain
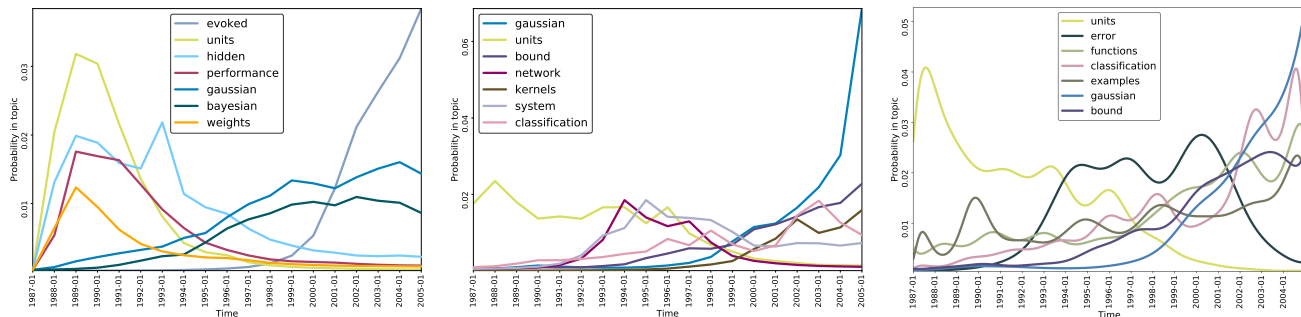
Figure 4: NIPS: Learned word trajectories of the "function approximation" topic using the Wiener kernel (left), OU kernel (middle) and Cauchy kernel (right). All three approaches identify terms that gain or loose importance within the topic over time. Since the Cauchy kernel shares statistical strength over a broader time horizon, its word trajectories are smoother.

events. We observe high probability for the term *"war"* again around 1848, a small plateau in the 1910s (World War I) rising to a high value in 1939 (World War II) and the 1960s (Vietnam war). We even observe a small bump in the beginning and through the 1980s (possibly the war in Afghanistan) and another peak in the mid 2000s (second Afghanistan war). Additionally, when looking at the words with highest probability at these times, we observe that the model is able to place probability mass on terms relating to the different wars, e.g. *"texas"* for the American-Mexican war (which was fought over Texas) or *"attack"* and *"japanese"* in 1942 (where the attack on Pearl Harbor took place). Based on these findings, the Ornstein-Uhlenbeck kernel seemed most appropriate for this task.

NYT. Another interesting use case scenario is the analysis of news texts. One of the topics identified when analyzing the New York Times corpus deals with presidential election campaigns. Figure 3 shows probability trajectories of terms in that topic for the Wiener (left), Ornstein-Uhlenbeck (middle), and Cauchy (right) kernels, respectively. We observe, that the baseline model (Wiener kernel) is able to capture meaningful terms. Going beyond this, the OU kernel arguably reflects the election campaigns in 1992 and 1996. The Cauchy kernel results in even smoother trajectories. These findings, however, deserve a more thorough investigation and interpretation. Nevertheless, this example shows that different kernels reveal qualitatively different phenomena.

NIPS. Applying generalized DTMs on the NIPS corpus allows us to track trends in machine learning over the last 20 years. We present probability trajectories of a topic related to classification and function approximation. Again, we show results for Wiener, OU, and Cauchy kernels (Figure 4, left to right). We observe from the baseline model that neural networks gained attention in the late 1980s and early 1990s. However, excitement subsided in the late 1990s and Bayesian methods were on the rise (our data set is not recent enough to detect the uptrend of neural networks in the last 5-10 years).

While the Wiener kernel models overall development, the Ornstein-Uhlenbeck process is able to better react to small scale changes, resulting in a more realistic representation of term development. Additionally, it finds more general terms, such as *"network"*, *"classification"* and *"system"*. Using the Cauchy kernel with its long-term memory prevents from placing large probability mass on the rapidly rising term *"gaussian"*. The Cauchy kernel is also able to identify more general terms.

Quantitative Results. We show that using our approach not only leads to interesting dynamic topics but also generalizes better to unseen data. We use all documents associated with time stamps T_{train} as training set and analyze the predictive held-out likelihoods on remaining documents (associated with the time stamps $T_{\text{test}} = T \setminus T_{\text{train}}$). We experiment on the NYT dataset and randomly select T_{train} to hold 85% of the unique time stamps.
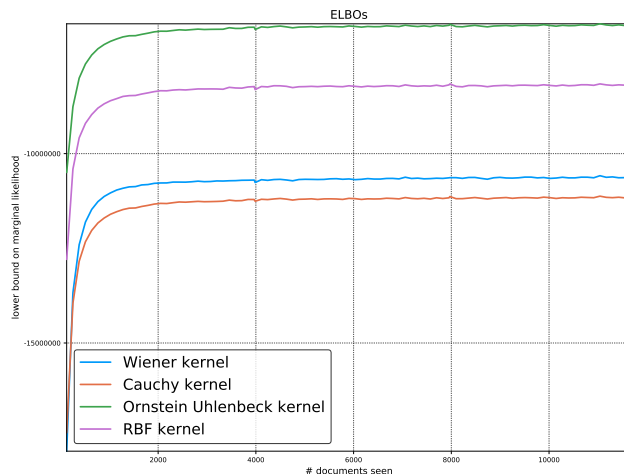

Figure 5: SoU: Evidence lower bound against the number of documents seen. On all used kernels, the objective function converges to an optimum.

Data	cDTM (baseline)	gDTM OU	gDTM Cauchy	gDTM RBF
NYT	1.42323	**1.42073**	1.42129	1.42374
NIPS	1.4931	1.48149	**1.48105**	1.4821
SoU	1.46854	1.45594	**1.45575**	1.46023

Table 1: Per-word predictive perplexities (lower numbers are better). We constantly outperform the baseline on all data sets.

Table 1 shows that our method outperforms the baseline in terms of per-word predictive perplexity (e.g. Blei and Lafferty, 2007). We observe that the perplexity on both the NIPS and SoU dataset is best when the dynamics are modeled by a GP with Cauchy kernel while the NYT dataset is best captured by a OU kernel. This shows again the advantage of using our approach over the state-of-the-art. Having the flexibility of modeling the dynamics by a GP we can account for the different dynamics that may underlie different datasets. Additionally, Figure 5 shows the ELBO objective function when fitting a model to the SoU data set, eventually reaching an optimum. Results on the different data sets were similar.

Remarks. We argue that as always in probabilistic modeling, the prior should not be chosen based on predictive likelihood alone. Instead, a prior is a modeling choice that helps reveal the effects that one searches for. Depending on the problem at hand, a practitioner would choose the suitable kernel, be it the Wiener kernel, Ornstein-Uhenbeck kernel, RBF kernel or Cauchy kernel. The Ornstein-Uhlenbeck kernel has the favorable property of localizing topics in time, which may be a promising tool for event detection. However, if the length scale is too small, topics change their word distributions at a frequency which is too high, in which case the results are less interpretable. On the other hand, the RBF kernel (and even more so the Cauchy kernel) has long-time memory and is generally more data efficient, which has advantages if the data set is small. Ultimately, many other kernels may be designed for different purposes.

6 Conclusion and Future Work

We presented the generalized dynamic topic model, which allows for dynamic topic modeling with a broader class of dynamic priors, and which easily scales up to very large text collections. In particular, we generalized dynamic topic models from Brownian motion priors to arbitrary Gaussian process priors. We showed in our experiments that our approach leads to better predictive likelihoods on held-out documents, and to interesting new qualitative findings, such as temporally localized topics, and topics that display long-range temporal dependencies. As a possible future extensions, we plan to consider periodic kernels for repeating events, and to extend dynamic topic modeling from the time domain to the geo-spatial domain, such as text equipped with location information.

References

Agovic, A. and Banerjee, A. (2012). Gaussian Process Topic Models. *arXiv preprint arXiv:1203.3462*.

Amari, S. and Nagaoka, H. (2007). *Methods of Information Geometry*. American Mathematical Society.

Bamler, R. and Mandt, S. (2017). Dynamic word embeddings via skip-gram filtering. *arXiv preprint arXiv:1702.08359*.

Bhadury, A., Chen, J., Zhu, J., and Liu, S. (2016). Scaling up dynamic topic models. In *Proceedings of WWW*.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, 55(4):77–84.

Blei, D. M. and Lafferty, J. D. (2006). Dynamic topic models. *Proceedings of ICML*.

Blei, D. M. and Lafferty, J. D. (2007). A Correlated Topic Model of Science. *The Annals of Applied Statistics*.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022.

Charlin, L. and Zemel, R. S. (2013). The toronto paper matching system: an automated paper-reviewer assignment system. In *Proceedings of ICML*.

Chong, W., Blei, D., and Li, F.-F. (2009). Simultaneous image classification and annotation. In *Proceedings of CVPR*.

Csató, L. and Opper, M. (2002). Sparse on-line gaussian processes. *Neural computation*, 14(3):641–668.

Fei-Fei, L. and Perona, P. (2005). A bayesian hierarchical model for learning natural scene categories. In *Proceedings of CVPR*.

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

Hennig, P., Stern, D. H., Herbrich, R., and Graepel, T. (2012). Kernel topic models. In *Proceedings of AISTATS*.

Hensman, J., Fusi, N., and Lawrence, N. D. (2013). Gaussian Processes for Big Data. In *Proceedings of UAI*.

Hensman, J., Matthews, A., and Ghahramani, Z. (2015). Scalable Variational Gaussian Process Classification. *Proceedings of AISTATS*.

Hensman, J., Rattray, M., and Lawrence, N. D. (2012). Fast variational inference in the conjugate exponential family. In *Proceedings of NIPS*.

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

Jordan, M. I., Ghahramani, Z., Jaakkola, T., and Saul, L. (1999). An introduction to variational methods for graphical models. *Machine Learning*, 37(2):183–233.

Mandt, S., Hoffman, M. D., and Blei, D. M. (2016). A Variational Analysis of Stochastic Gradient Algorithms. *arXiv.org*.

Maritz, J. and Lwin, T. (1989). Empirical Bayes Methods with Applications. *Monographs on Statistics and Applied Probability*.

Martens, J. (2017). New insights and perspectives on the natural gradient method. *Arxiv Preprint*.

McCallum, A., Corrada-Emmanuel, A., and Wang, X. (2004). The Author-Recipient-Topic Model for Topic and Role Discovery in Social Networks: Experiments with Enron and Academic Email. *Workshop on Structured Data and Representations in Probabilistic Models for Categorization at NIPS*.

Pritchard, J. K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959.

Roberts, S., Osborne, M., Ebden, M., Reece, S., Gibson, N., and Aigrain, S. (2012). Gaussian Processes for Timeseries Modelling. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 371.

Sandhaus, E. (2008). The New York Times Annotated Corpus. Linguistic Data Consortium.

Snelson, E. and Ghahramani, Z. (2006). Sparse Gaussian processes using pseudo-inputs.

Titsias, M. K. (2009). Variational learning of inducing variables in sparse Gaussian processes. In *Proceedings of AISTATS*.

Wainwright, M. J. and Jordan, M. I. (2007). Graphical Models, Exponential Families, and Variational Inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305.

Wang, C., Blei, D., and Heckerman, D. (2008). Continuous time dynamic topic models. In *Proceedings of UAI*.

Wang, C. and Blei, D. M. (2011). Collaborative topic modeling for recommending scientific articles. In *Proceedings of ACM SIGKDD*.

Wang, X. and McCallum, A. (2006). Topics over time: a non-Markov continuous-time model of topical trends. *Proceedings of ACM SIGKDD*.

Wang, X., McCallum, A., and Wei, X. (2007). Topical N-Grams: Phrase and Topic Discovery, with an Application to Information Retrieval. *Data Mining*.

Welling, M. and Teh, Y.-W. (2011). Bayesian learning via stochastic gradient Langevin dynamics. In *Proceedings of ICML*, pages 681–688.

Wenzel, F., Galy-Fajou, T., Deutsch, M., and Kloft, M. (2017). Bayesian Nonlinear Support Vector Machines for Big Data. In *Proceedings of ECML*.

Wenzel, F., Galy-Fajou, T., Donner, C., Kloft, M., and Opper, M. (2018). Efficient Gaussian Process Classification Using Pólya-Gamma Data Augmentation. *arXiv preprint*.