Generalizing and Scaling up Dynamic Topic Models via Inducing Point Variational Inference

Patrick Jähnichen, Florian Wenzel
Department of Computer Science
Humboldt-Universität zu Berlin, Germany
{jaehnicp, wenzelfl}@hu-berlin.de

Marius Kloft
Department of Computer Science
TU Kaiserslautern, Germany
kloft@cs.uni-kl.de

Stephan Mandt
Data Science Institute
Columbia University, NY, USA
stephan.mandt@gmail.com

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 topics change continuously over time and therefore impose continuous stochastic process priors on their model parameters. In this paper, we extend the class of tractable priors from Wiener processes to the generic class of Gaussian processes (GPs). Second, we show how to perform scalable approximate inference in these models based on ideas around stochastic variational inference and Gaussian processes with inducing points. Our experiments show that our generalized model allows us to find interesting patterns that were not accessible by previous approaches.

1 Introduction

Topic models belong to the standard machine learning toolbox and 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).

Dynamic topic models take into account that topics may change their word frequencies over long periods of time (Blei and Lafferty, 2006; Wang and McCallum, 2006; Wang et al., 2008a). Consider the example of the topic *technology* when training topic models on historical articles [1]. Around the year 1900, we find words such as *engine*, *electricity*, and *wire* whereas for more recent articles, we may find *devices*, *gates*, and *silicon* among the top words. 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 limited: first, the latent topics' dynamics always follow a Wiener process. Second, they lack scalability, relying on batch algorithms for approximate inference. In this paper, we generalize dynamic topic models in two ways: we extend the class of tractable priors from Wiener processes to the more general class of Gaussian processes and we derive a scalable approximate Bayesian inference algorithm based on inducing points. This allows us to

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

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

apply our model to contemporary large text collections. In our experiments, we use different kernels to filter time-localized topics in a set of presidents' speeches on the State of the Union.

2 Related Work

Our work connects to dynamic and correlated topic models, sparse GPs and stochastic variational inference. **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. (2008b) extended the approach to continuous time. Both rely on a latent Wiener process and use the forward-backward algorithm for learning. Additionally, Bhadury et al. (2016) proposed a new approach for learning in classical DTMs based on stochastic gradient MCMC (Welling and Teh, 2011; Mandt et al., 2016), again restricted to latent Wiener processes. **Correlated and GP Topic Models** modify static topic models to break 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). Both approaches assume that the topics themselves are static and only the topic proportions change. In contrast, we adhere to the notion of DTMs and treat the proportions as independent and identically distributed (iid) and impose dynamics on the topics themselves. Our algorithm builds on **Stochastic Variational Inference (SVI)** (Hoffman et al., 2013) and **sparse GPs**. SVI can normally only be 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. (2012, 2013) have shown that one can derive a tractable lower bound to the marginal likelihood of the data that allows for data subsampling. This 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 and Matthews, 2015; Wenzel et al., 2017). None of this work has been applied in the context of topic models.

3 Scalable Dynamic Topic Models Using Inducing Points

Assume that we observe a corpus of D documents, each of which is associated with a time stamp τ_{t_d} with index $t_d \in \{1, \dots, T\}$. For a simpler notation we denote the number of words in a document as N. For a given document d with time index t_d, 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 distribution over the vocabulary at time t. Our model exhibits the following joint distribution:

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

The function $\pi(\cdot)$ is the softmax function which normalizes the topic $\beta_{k\cdot t}$ over the vocabulary indices so that the second part (everything apart from $p(\beta)$) of (1) is just a regular LDA model (at time t). 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:

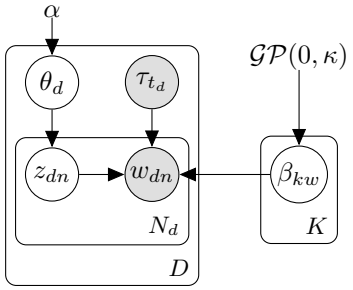

Figure 1: The model in plate notation.

$$\beta_{kw\cdot} \sim \mathrm{GP}(0, \kappa) \Leftrightarrow \beta_{kw,1:T} \sim \mathcal{N}_T(0, K_{TT}), \tag{2}$$

$$K_{\tau,\tau'} = \kappa(\tau, \tau'), \quad \tau, \tau' \in \{\tau_1, \cdots, \tau_T\}. \tag{3}$$

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. In more detail, we consider several different kernels commonly used for time-series modeling (Roberts et al., 2012): *Wiener kernels* with covariance function $\kappa_{\text{Wie}}(\tau, \tau') = \sigma^2 \min(\tau, \tau')$, recovering the original DTM setup of Wang et al. (2008b) and serving

as a baseline, *Ornstein-Uhlenbeck (OU) kernels* with $\kappa_{\text{OU}}(\tau, \tau') = \sigma^2 \exp\left(-\frac{|\tau-\tau'|}{l}\right)$, leading to temporally localized changes in topics and *Cauchy kernels* with $\kappa_{\text{Cau}}(\tau, \tau') = \sigma^2 \left(1 + \frac{(\tau-\tau')^2}{l^2}\right)^{-1}$, leading to temporal correlations that 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.

Sparse DTMs. The bottleneck of inference in the model so far is the inversion of the $T \times T$ kernel matrix K_{TT}. 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 $\hat{T} \ll T$. For this, 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. Defining $p(u_{kw}) = \mathcal{N}(0, K_{\hat{T}\hat{T}})$, we obtain $p(\beta_{kw}|u_{kw}) = \mathcal{N}(K_{T\hat{T}}K_{\hat{T}\hat{T}}^{-1}u_{kw}, \tilde{K})$, and perform approximate inference over u (with $\tilde{K} = K_{TT} - K_{T\hat{T}}K_{\hat{T}\hat{T}}^{-1}K_{\hat{T}T}$). 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

$$p(\beta, w, z, \theta, u) = p(w|\beta, z)p(z|\theta)p(\beta|u)p(u), \tag{4}$$

which summarizes our model.

Approximate marginalization. We first marginalize over β in the augmented joint distribution (4) and obtain

$$\log p(w_{dn}|z_{dn} = k, u, t_d) = K_{t_d\hat{T}}K_{\hat{T}\hat{T}}^{-1}u_k w_{dn} - \mathbb{E}_{p(\beta_{k\cdot t_d}|u)}\left[\log \sum_w \exp(\beta_{kwt_d})\right], \tag{5}$$

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 (5) is still intractable due to the sum inside of the logarithm. Following Blei and Lafferty (2006), we can lower bound this quantity by

$$\log \tilde{p}(w_{dn}|z_{dn} = k, u, t_d),$$

introducing additional free variational parameters (see appendix). Next, we use this lower bounded log-likelihood to derive a tractable variational objective which we can optimize.

Stochastic Variational Inference. Our main contribution is the derivation of the global updates for the variational parameters of u based on natural gradients. We consider Gaussian distributions $q(u_{kw}) = \mathcal{N}_{\hat{T}}(m_{kw}, \Sigma_{kw})$ in *natural parameterization*, i.e. using the parameters $\eta_{kw}^{(1)} = \Sigma_{kw}^{-1}m_{kw}$ and $\eta_{kw}^{(2)} = -\frac{1}{2}\Sigma_{kw}^{-1}$, where m_{kw} are the variational Gaussian means and Σ_{kw} the covariances.

In general it holds for objectives $\mathcal{F}$ that depend on a Gaussian distribution that

$$\hat{\nabla}_{(\eta_1,\eta_2)}\mathcal{F}(\eta) = \left(\nabla_\mu \mathcal{F}(\eta) - 2\nabla_\Sigma \mathcal{F}(\eta)\mu,\ \nabla_\Sigma \mathcal{F}(\eta)\right)^\top, \tag{6}$$

where $\hat{\nabla}$ denotes the natural gradient and ∇ the Euclidean gradient. Applying (6) to the variational objective, we obtain

$$\hat{\nabla}_{\eta_{kw}}\mathcal{L} = \left(\Xi_{kw} + B_{kw} \circ (m_{kw} - 1) - \eta_{kw}^{(1)}, -\frac{1}{2}K_{\hat{T}\hat{T}}^{-1} - \frac{1}{2}C_{kw} - \eta_{kw}^{(2)}\right)^\top \tag{7}$$

We used the following abbreviations:

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

$$\Xi_{kw} = K_{\hat{T}\hat{T}}^{-1} \sum_t \sum_{d:t_d=t} K_{\hat{T}t} n_{dw}\phi_{dwk},\ \ C_{kw} = B_{kw}K_{t\hat{T}}K_{\hat{T}\hat{T}}^{-1}.$$

Above, $\circ$ denotes the Hadamard product and n_{dw} the number of occurrence of term w in document d. Details are provided the appendix. Iterating through those updates completes the algorithm.

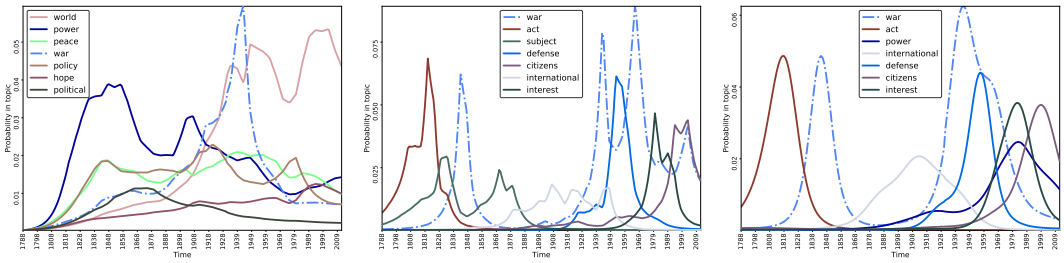

Figure 2: SoU: Learned word trajectories of the "war" topic using the Wiener kernel (left), OU kernel (middle) and Cauchy kernel (right). The Cauchy kernel provides smoother trajectories yet the OU kernel is able to provide a better resolution in time.

4 Experiments

We use the "State of the Union" addresses of U.S. presidents, which span more than two centuries, resulting in $T = 224$ different time stamps [2]. 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 imported 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 (see appendix) is less than a certain threshold, resulting in dictionaries of size $V = 3127$. We consider a topic of war and peace. Figure 2 shows the word probabilities within this topic over time for all three considered kernels. Kernel hyper-parameters were determined via grid search to obtain interpretable models with low test set perplexity. 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 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. Note that 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, or Cauchy kernel. Ultimately, many other kernels may be designed for different purposes.

5 Conclusion and Future Work

We presented a 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 Wiener process/Brownian motion priors to arbitrary Gaussian process priors. We showed in our experiments that our approach leads to interesting new qualitative findings, such as temporally localized topics, and topics that display long-range temporal dependencies. In the future, 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.

[2] http://www.presidency.ucsb.edu/sou.php

References

Agovic, A. and Banerjee, A. (2012). Gaussian process topic models. *arXiv preprint arXiv:1203.3462*.

Bhadury, A., Chen, J., Zhu, J., and Liu, S. (2016). Scaling up dynamic topic models. In *Proceedings of the 25th International Conference on World Wide Web*, pages 381–390. International World Wide Web Conferences Steering Committee.

Blei, D. M. and Lafferty, J. D. (2006). Dynamic topic models. *Proceedings of the 23rd International Conference on Machine Learning*.

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

Blei, D. M. and Lafferty, J. D. (2009). Visualizing topics with multi-word expressions. *Arxiv preprint arXiv:0907.1013*.

Charlin, L. and Zemel, R. S. (2013). The toronto paper matching system: an automated paper-reviewer assignment system. In *International Conference on Machine Learning (ICML)*, volume 10.

Chong, W., Blei, D., and Li, F.-F. (2009). Simultaneous image classification and annotation. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1903–1910. IEEE.

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 *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 2, pages 524–531. IEEE.

Gopalan, P., Hao, W., Blei, D. M., and Storey, J. D. (2016). Scaling probabilistic models of genetic variation to millions of humans. Technical report, Nature Research.

Hennig, P., Stern, D. H., Herbrich, R., and Graepel, T. (2012). Kernel topic models. In *AISTATS*, pages 511–519.

Hensman, J., Fusi, N., and Lawrence, N. D. (2013). Gaussian Processes for Big Data. In *Conference on Uncertainty in Artifical Intelligence*.

Hensman, J. and Matthews, A. (2015). Scalable Variational Gaussian Process Classification. In *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics,*.

Hensman, J., Rattray, M., and Lawrence, N. D. (2012). Fast variational inference in the conjugate exponential family. In *Advances in Neural Information Processing Systems*.

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

Mandt, S., Hoffman, M. D., and Blei, D. M. (2016). A Variational Analysis of Stochastic Gradient Algorithms. In *International Conference on Machine Learning*, pages 354–363.

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. *NIPS'04 Workshop on Structured Data and Representations in Probabilistic Models for Categorization*.

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.

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 the International Conference on Artificial Intelligence and Statistics*, pages 567–574.

Wang, C., Blei, D., and Heckerman, D. (2008a). Continuous time dynamic topic models. In *Proceedings of the Twenty-Fourth Conference on Uncertainty in Artificial Intelligence*, UAI'08, pages 579–586, Arlington, Virginia, United States. AUAI Press.

Wang, C. and Blei, D. M. (2011). Collaborative topic modeling for recommending scientific articles. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 448–456. ACM.

Wang, C., Blei, D. M., and Heckerman, D. (2008b). Continuous Time Dynamic Topic Models. In *Conference on Uncertainty in Artifical Intelligence*.

Wang, X. and McCallum, A. (2006). Topics over time: a non-Markov continuous-time model of topical trends. *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 424–433.

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 the 28th International Conference on Machine Learning*, 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 the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*.

A Appendix

A.1 Approximate marginalization

Following (Blei and Lafferty, 2009), we lower bound the intractable expectation in (5) by computing the first order Taylor approximation of the logarithm around an arbitrary location parameter $\zeta_{kt} > 0$,

$$\mathbb{E}_{p(\beta_t|u)}\left[\log \sum_v \exp(\beta_{kvt})\right] \leq \zeta_{kt}^{-1} \sum_v \exp\left(K_{t\hat{T}} K_{\hat{T}\hat{T}}^{-1} u_{kv\cdot} + \frac{\tilde{K}_{tt}}{2}\right) + \log(\zeta_{kt}) - 1.$$

Updating the Taylor expansion location parameters. In each iteration in our inference algorithm we optimize the location parameter of the Taylor expansion to achieve the tightest possible bound on true marginal likelihood (c.f. equation 3). Setting the derivative of $\mathcal{L}$ w.r.t. ζ_{kt} to zero and solving for ζ_{kt} gives the update

$$\zeta_{kt} = \sum_v \exp\left(m_{kvt} + \frac{1}{2}(\Lambda_{kvt} + \tilde{K}_{tt})\right).$$

A.2 Derivation of the Variational Objective

Recall the variational objective

$$\mathcal{L}(\lambda, \phi, \mu, \Sigma) = \mathbb{E}_q[\log \tilde{p}(w|u, z)p(z|\theta)p(\theta)p(u)] - \mathbb{E}_q[\log q(\theta)q(z)q(u)].$$

The first term is

$$\begin{aligned}
&\mathbb{E}_q\left[\log \tilde{p}(w|z, u)\right] \\
&= \sum_{t,n,k} \mathbb{E}_q\left[z_{tnk} \log \tilde{p}(w_{tn}|z_{tn} = k, u)\right] \\
&= \sum_{t,n,k} \mathbb{E}_q[z_{tnk}] \left\{ K_{t\hat{T}} K_{\hat{T}\hat{T}}^{-1} \mathbb{E}_q[u_{k\cdot\cdot}] w_{tn} - \zeta_{kt}^{-1} \sum_v \mathbb{E}_q\left[\exp\left(K_{t\hat{T}} K_{\hat{T}\hat{T}}^{-1} u_{kv\cdot} + \frac{\tilde{K}_{tt}}{2}\right)\right] - \log(\zeta_{kt}) + 1 \right\} \\
&= \sum_{t,n,k} \phi_{tnk} \Big\{ K_{t\hat{T}} K_{\hat{T}\hat{T}}^{-1} \mu_{k\cdot\cdot} w_{tn} - \zeta_{kt}^{-1} \sum_v \exp\left(K_{t\hat{T}} K_{\hat{T}\hat{T}}^{-1} \mu_{kv} + \frac{1}{2}(K_{t\hat{T}} K_{\hat{T}\hat{T}}^{-1} \Sigma_{kv} K_{\hat{T}\hat{T}}^{-1} K_{\hat{T}t} + \tilde{K}_{tt})\right) \\
&\quad - \log(\zeta_{kt}) + 1 \Big\} \\
&= \sum_{t,n,k} \phi_{tnk} \Big\{ w_{tn\cdot}^\top m_{k\cdot t} - \zeta_{kt}^{-1} \sum_v \exp\left(m_{kvt} + \frac{1}{2}(\Lambda_{kvt} + \tilde{K}_{tt})\right) - \log(\zeta_{kt}) + 1 \Big\},
\end{aligned}$$

where $m_{kvt} = K_{t\hat{T}} K_{\hat{T}\hat{T}}^{-1} \mu_{kv}$ and $\Lambda_{kvt} = K_{t\hat{T}} K_{\hat{T}\hat{T}}^{-1} \Sigma_{kv} K_{\hat{T}\hat{T}}^{-1} K_{\hat{T}t}$.

The second term is

$$\begin{aligned}
\mathbb{E}_q[\log p(z|\theta)] &= \sum_{t,n} \mathbb{E}_q[\log p(z_{tn}|\theta_t)] = \sum_{t,n,k} \phi_{tnk} \mathbb{E}_q[\log \theta_{tk}] \\
&= \sum_{t,n,k} \phi_{tnk} \left(\psi(\lambda_{tk}) - \psi\left(\lambda_{t0}\right)\right),
\end{aligned}$$

where $\lambda_{t0} = \sum_k \lambda_{dk}$.

The negative KL terms are

$$\mathbb{E}_q[\log p(u) - \log q(u)] = -\sum_{k,v} \mathrm{KL}(q(u_{kv})||p(u_{kv})) \overset{\mathsf{c}}{=} -\frac{1}{2}\sum_{k,v}\left(\mu_{kv}K^{-1}_{\hat{T}\hat{T}}\mu_{kv} + \mathrm{tr}(\Sigma_{kv}K^{-1}_{\hat{T}\hat{T}}) - \log|\Sigma_{kv}|\right).$$

and

$$\begin{aligned}\mathbb{E}_q[\log p(\theta) - \log q(\theta)] &= -\sum_t \mathrm{KL}(q(\theta_t)||p(\theta_t)) \\ &\overset{\mathsf{c}}{=} \sum_{t,k}\left((\alpha_k - \lambda_{tk})(\psi(\lambda_{tk}) - \psi(\lambda_{t0})) + \log\Gamma(\lambda_{tk})\right) - \Gamma(\lambda_{t0}).\end{aligned}$$

The entropy of $q(z)$ is

$$-\mathbb{E}_q[q(z)] = -\sum_{t,n,k}\phi_{tnk}\log\phi_{tnk}.$$

Finally, summing all terms gives the variational objective

$$\begin{aligned}\mathcal{L}(\lambda,\phi,\mu,\Sigma) = \sum_{t,n,k}\phi_{dnk}\Big\{&w_{tn}^\top m_{k.t} - \zeta_{kt}^{-1}\sum_v \exp\left(m_{kvt} + \frac{1}{2}(\Lambda_{kvt} + \tilde{K}_{tt})\right) - \log(\zeta_{kt}) + 1 \\ &+ \psi(\lambda_{tk}) - \psi(\lambda_{t0}) - \log\phi_{tnk}\Big\} - \frac{1}{2}\sum_{k,v}\left(\mu_{kv}K^{-1}_{\hat{T}\hat{T}}\mu_{kv} + \mathrm{tr}(\Sigma_{kv}K^{-1}_{\hat{T}\hat{T}}) - \log|\Sigma_{kv}|\right) \\ &+ \sum_{t,k}\left((\alpha_k - \lambda_{tk})(\psi(\lambda_{tk}) - \psi(\lambda_{t0})) + \log\Gamma(\lambda_{tk})\right) - \Gamma(\lambda_{t0}) + \mathrm{const.}\end{aligned}$$

A.3 SVI Updates

In the following we provide more details on how the the parameter updates are derived.

Updating the Taylor expansion location parameter The derivative of the variational objective with respect to the location parameter of the Taylor expansion is

$$\frac{\partial\mathcal{L}}{\partial\zeta_{kt}} = \sum_n \phi_{tnk}\zeta_{kt}^{-1}\left(\zeta_{kt}^{-1}\sum_v \exp\left(m_{kvt} + \frac{1}{2}(\Lambda_{kvt} + \tilde{K}_{tt})\right) - 1\right).$$

Stetting the derivative zero and solving for ζ_{kt} gives the update

$$\zeta_{kt} = \sum_v \exp\left(m_{kvt} + \frac{1}{2}(\Lambda_{kvt} + \tilde{K}_{tt})\right).$$

Updating the local variables of $q(z)$ The derivative of $\mathcal{L}$ w.r.t. ϕ_{tnk} is

$$\frac{\partial\mathcal{L}}{\partial\phi_{tnk}} = w_{tn}^\top m_{k.t} - \zeta_{kt}^{-1}\sum_v \exp\left(m_{kvt} + \frac{1}{2}(\Lambda_{kvt} + \tilde{K}_{tt})\right) - \log(\zeta_{kt}) + \psi(\lambda_{tk}) - \psi(\lambda_{t0}) - \log\phi_{tnk}$$

Setting the derivative zero leads to

$$\phi_{tnk} = \exp\left\{w_{tn}^\top m_{k.t} - \zeta_{kt}^{-1}\sum_v \exp\left(m_{kvt} + \frac{1}{2}(\Lambda_{kvt} + \tilde{K}_{tt})\right) - \log(\zeta_{kt}) + \psi(\lambda_{tk}) - \psi(\lambda_{t0})\right\}.$$

Inserting the update of the previous update of ζ_{kt} this simplifies to

$$\begin{aligned}\phi_{tnk} &= \exp\left\{w_{tn}^\top m_{k.t} - 1 - \log(\zeta_{kt}) + \psi(\lambda_{tk}) - \psi\left(\lambda_{t0}\right)\right\} \\ &\propto \exp\left\{w_{tn}^\top m_{k.t} - \log(\zeta_{kt}) + \psi(\lambda_{tk}) - \psi\left(\lambda_{t0}\right)\right\}.\end{aligned}$$

The update for the parameter vector $\phi_{tn.}$ is obtained by renormalizing (such that $||\phi_{tn.}||_1 = 1$).

Updating the global variables The standard Euclidean gradient of $\mathcal{L}$ with respect to the mean and covariance parameters of $q(u_{kv})$ is

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial \mu_{kv}} &= \Xi_{kv} - B_{kv} - K_{\hat{T}\hat{T}}^{-1}\mu_{kv}, \\ \frac{\partial \mathcal{L}}{\partial \Sigma_{kv}} &= -\frac{1}{2}C_{kv} + \frac{1}{2}\Sigma_{kv}^{-1} - \frac{1}{2}K_{\hat{T}\hat{T}}^{-1},\end{aligned}$$

where $\Xi_{kv} = \sum_{t,n}\phi_{tnk}w_{tnv}K_{\hat{T}\hat{T}}^{-1}K_{\hat{T}t}$, $B_{kv} = \sum_{t,n}\zeta_{kt}^{-1}\phi_{tnk}\exp\left(m_{kvt} + \frac{\Lambda_{kvt}+\tilde{K}_{tt}}{2}\right)K_{\hat{T}\hat{T}}^{-1}K_{\hat{T}t}$ and $C_{kv} = \sum_{t,n}\zeta_{kt}^{-1}\phi_{tnk}\exp\left(m_{kvt} + \frac{\Lambda_{kvt}+\tilde{K}_{tt}}{2}\right)K_{\hat{T}\hat{T}}^{-1}K_{\hat{T}t}K_{t\hat{T}}K_{\hat{T}\hat{T}}^{-1}$.

We now consider the Gaussian distributions $q(u_{kv})$ in natural parametrization using $\eta_{kv}^{(1)} = S_{kv}^{-1}\mu_{kv}$ and $\eta_{kv}^{(2)} = -\frac{1}{2}S_{kv}^{-1}$. Applying formula 6 we obtain the natural gradient w.r.t. natural parameters,

$$\begin{aligned}\hat{\nabla}_{\eta_{kv}^{(1)}}\mathcal{L} &= \Xi_{kv} - B_{kv} - K_{\hat{T}\hat{T}}^{-1}\mu_{kv} - 2(-\frac{1}{2}C_{kv} + \frac{1}{2}\Sigma_{kv}^{-1} - \frac{1}{2}K_{\hat{T}\hat{T}}^{-1})\mu_{kv} \\ &= \Xi_{kv} + B_{kv}\circ(m_{kv} - 1) - \eta_{kv}^{(1)}\end{aligned}$$

and

$$\hat{\nabla}_{\eta_{kv}^{(2)}}\mathcal{L} = -\frac{1}{2}C_{kv} - \frac{1}{2}K_{\hat{T}\hat{T}}^{-1} - \eta_{kv}^{(2)}.$$

Note that m_{kv} as function of the natural parameters is

$$m_{kv} = K_{T\hat{T}}K_{\hat{T}\hat{T}}^{-1}\mu_{kv} = -\frac{1}{2}K_{T\hat{T}}K_{\hat{T}\hat{T}}^{-1}\left(\eta_{kv}^{(2)}\right)^{-1}\eta_{kv}^{(1)}.$$

A.4 Global td-idf score

To determine important words, we use an extension to the classic tf-idf scoring scheme. The score of a word

$$\text{score}(w) = \frac{n_w}{M}\ln\left(\frac{D}{n_{dw}}\right)$$

where M is the total amount of terms in the corpus, D is the number of documents, n_{dw} is the frequency of word w in document d and $n_w = \sum_d n_{dw}$.