On Topic Evolution

Eric P. Xing
School of Computer Science
Carnegie Mellon University
epxing@cs.cmu.edu

Technical Report: CMU-CALD-05-115

December 2005

Abstract

I introduce topic evolution models for longitudinal epochs of word documents. The models employ marginally dependent latent state-space models for evolving topic proportion distributions and topic-specific word distributions; and either a logistic-normal-multinomial or a logistic-normal-Poisson model for document likelihood. These models allow posterior inference of latent topic themes over time, and topical clustering of longitudinal document epochs. I derive a variational inference algorithm for non-conjugate generalized linear models based on truncated Taylor approximation, and I also outline formulae for parameter estimation based on variational EM principle.

1 Introduction

Text information, such as media documents, journal articles and emails, often come as temporal streams. Current information retrieval systems working on corpora collected over time make little use of the time stamps associated with the documents. They often merely pool all the documents into a single collection, in which each document is treated as an iid sample from some topical distribution [Hofmann, 1999; Blei *et al.*, 2003; Griffiths and Steyvers, 2004]; or model the topics of each time-specific epoch separately and then examine relationships among the independently inferred time-specific topics [Steyvers *et al.*, 2004]. In practice, topic themes that generate the documents can evolve over time, and there exist dependencies among documents over time. In this report, I develop a principled statistical framework for modeling topic evolution and extracting high-level insights of the topic history based on latent space dynamic processes, and I derive the formulae for posterior inference and parameter estimation.

2 Topic Evolution

Let $\{\mathcal{D}_1, \ldots, \mathcal{D}_T\}$ represent a temporal series of corpus, where $\mathcal{D}_t \equiv \{\mathbf{x}_d^{(t)}\}_{d=1}^{N_t}$ denotes the set of N_t documents available at time t, $\mathbf{x}_d$ denotes a document consisting of word sequence $(x_{d,1}, \ldots, x_{d,N_d})$, and $\vec{n}_d = (n_{d,1}, \ldots, n_{d,M})$ denotes an M-dimensional count vector corresponding to the frequencies of M words (defined by a fixed vocabulary) in document d. We assume that every document $\mathbf{x}_d$ can express multiple topics coming from a predefined topic space, and the weights of every topic can be represented by a normalized vector θ_d of fixed dimension. Furthermore, we assume that each topic can be represented by a set of parameters that determine how words from a fixed vocabulary can be drawn in a topic-specific manner to compose the document (for simplicity, here we assume a bag-of-word model for the word-to-document relationship, so that topic-specific semantics only translate to measures on word rates, but not to non-trivial syntactic grammars).

Under a topic evolution model, the prior distributions of topic proportions of every document, and the representations of each of the topics themselves, are evolving over time. In the following, I present two topic evolution models defined on two different kinds of topic representations, and derive the variational inference formulas in each case.

2.1 A Dynamic Logistic-Normal-Multinomial Model

In this model we assume that each document is an *admixture* of topics resulting from a bag of topic-specific instances of words, each of which is marginally a mixture of topics. Each topic, say topic k, is represented by an M-dimensional normalized word frequency vector $\vec{\beta}_k$, which parameterizes a topic-specific multinomial distribution of word. Here is an outline of a generative process under such a model (a graphical model representation of this model is illustrated in Figure 1):

- We assume that the topic proportion vector $\vec{\theta}$ for each document follows a time-specific logistic normal prior $\mathcal{LN}(\vec{\mu}_t, \Sigma_t)$, whose mean $\vec{\mu}_t$ is evolving over time according to a linear Gaussian model: (For simplicity, we assume that the Σ_t's capturing time-specific topic correlations are independent across time.)
  - $\vec{\mu}_1 \sim \text{Normal}(\nu, \Phi)$, sample the mean of the topic mixing prior at time 1.
  - $\vec{\mu}_t = \text{Normal}(\mathbf{A}\vec{\mu}_{t-1}, \Phi)$, sample the means of the topic mixing priors over time.
  - $\vec{\theta}_d^{(t)} \sim \text{LogisticNormal}(\vec{\mu}_t, \Sigma_t)$, for each document, sample a topic proportion vector (for simplicity, in the sequel we will omit the time index t and/or document index d when describing a general law that applies to all time points and/or all documents).

  Notice that the last step above can be broken into two sub-steps:
  - $\vec{\gamma}_d^{(t)} \sim \text{Normal}(\vec{\mu}_t, \Sigma_t)$,

* $\theta_{d,k}^{(t)} = \frac{\exp(\gamma_{d,k}^{(t)})}{\sum_{k'} \exp(\gamma_{d,k'}^{(t)})}, \quad \forall k = 1, \ldots, K.$

  Furthermore, due to the normalizability constrain of the multinomial parameters, $\vec{\theta}$ only has $K-1$ degree of freedom. Thus, as described in detail in the sequel, we only need to draw the first $K-1$ components of $\vec{\gamma}$ from a $(K-1)$-dimensional multivariate Gaussian, and leave $\gamma_K = 0$. But for simplicity, we omit this technicality in the forth coming general description of our model.

- We further assume that the representation of each topic, in this case a topic-specific multinomial vector $\vec{\beta}$ of word frequencies, is also evolving over time. By defining $\vec{\beta}$ as a logistic transformation of a multivariate normal random vector $\vec{\eta}$, we can model the temporal evolution of $\vec{\beta}$ in a simplex via a linear Gaussian dynamics model:
  - $\vec{\eta}_k^{(1)} \sim \text{Normal}(\iota_k, \Psi_k)$, sample the topic k at time 1.
  - $\vec{\eta}_k^{(t)} \sim \text{Normal}(\mathbf{B}_k \vec{\eta}_k^{(t-1)}, \Psi_k)$, sample topic k over subsequent time points.
  - $\beta_{k,w}^{(t)} = \frac{\exp(\eta_{k,w}^{(t)})}{\sum_{w'} \exp(\eta_{k,w'}^{(t)})}, \ \forall w = 1, \ldots, M$, compute word probabilities via logistic transformation.
- Now we assume that each occurrence of word, e.g., the nth word in document d at time t, $x_{d,n}^{(t)}$, is drawn from a topic-specific word distribution $\vec{\beta}_k^{(t)}$, specified by a latent topic indicator $z_{d,n}^{(t)} = k$.
  - $z_{d,n}^{(t)} \sim \text{Multinomial}(\theta_d^{(t)})$ sample the latent topic indicator (again, for simplicity, indices t and d will be omitted in the sequel where no confusion arises.)
  - $x_{d,n}^{(t)} | z_{d,n}^{(t)} = k \sim \text{Multinomial}(\vec{\beta}_k^{(t)})$, sample the word from a topic-specific word distribution.

In principle, we can use the above topic evolution model to capture not only topic correlation among documents at a specific time (as did in [Blei and Lafferty, 2006]), but also dynamic coupling (i.e., co-evolution) of topics via covariance matrix Φ, and topic-specific word-coupling via covariance matrices $\{\Psi_k\}$. In the simplest scenario, when $\mathbf{A} = I$, $\mathbf{B} = I$, $\Phi = \sigma I$, and $\Psi_k = \rho I$, this model reduces to random walk in both the topic spaces, and the topic-mixing space. Since in most realistic temporal series of corpus, both the proportions of topics, and the semantic representations of topics are unlikely to be invariant over time, we expect that even a random walk topic evolution model can provide a better fit of the data than a static model that ignores the time stamps of all documents.

2.2 A Dynamic Log-Normal-Poisson Model

The above topic evolution process assumes an admixture likelihood model for documents belonging to a specific time interval, and the admixing is realized at the word level, i.e., the marginal probability of each word in the document is defined by a mixture of topic-specific word distributions.

Now we present another text likelihood model employing a different topic mixing mechanism, which can be also plugged into the topic evolution model. Note that in a bag-of-word model all we observed are counts of words in the documents. Instead of assuming each occurrence of a word is sampled from the topic-specific word distribution, we can directly assume that the total count n_w of word w is made up of fractions each contributed by a specific topic according to a topic-specific Poisson distribution $\text{Poisson}(\omega \theta_k \tau_{w,k})$, where ω denotes the length of the document, θ_k denotes the proportion of topic k in the document as defined before, and $\tau_{w,k}$ is a rate measure for word w associated with topic k. Specifically, $n_w = \sum_k n_{w,k}$, $n_{w,k} \sim \text{Poisson}(\omega \theta_k \tau_{w,k})$. It can be shown that under this model we have:

$$n_w \sim \text{Poisson}(\omega \sum_k \theta_k \tau_{w,k}) = \exp\{n_w \log(\omega \sum_k \theta_k \tau_{w,k}) - \omega \sum_k \theta_k \tau_{w,k} - \Gamma(n_w + 1)\}.$$

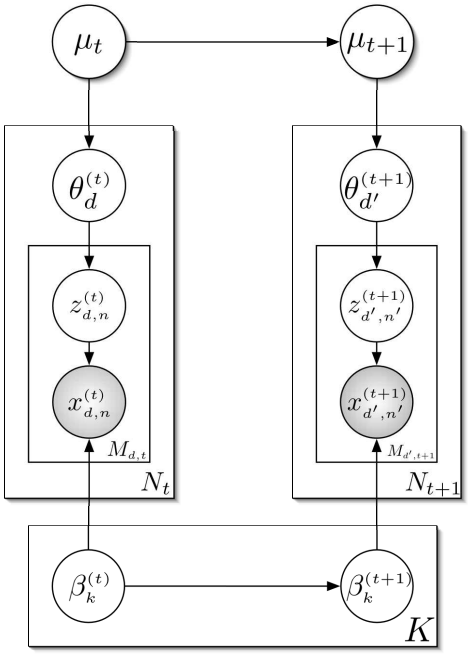

Figure 1: A graphical model representation of the dynamic logistic-normal-multinomial model for topic evolution.

Note that in the above setting for each word w we have a row vector of rates each associated with a specific topic: $\vec{\tau}_{w,\cdot} = (\tau_{w,1}, \ldots, \tau_{w,K})$. For each topic k, we have a column vector of rates each associated with a specific word, $\vec{\tau}_{\cdot,k} = (\tau_{1,k}, \ldots, \tau_{M,k})'$. Unlike the multinomial topic model parameterized by column-normalized topic matrix $\boldsymbol{\beta} = [\vec{\beta}_1, \ldots, \vec{\beta}_K]$, the Poisson topic model is parameterized by matrix $\boldsymbol{\tau} = [\vec{\tau}_{\cdot,1}, \ldots, \vec{\tau}_{\cdot,K}]$ that does not have to be column- or row-normalized. Thus we can directly use a Log-Normal distribution to model $\vec{\tau}$, which is simpler than the logistic-normal distribution. This leads to the following generative model for topic evolution (assuming we are interested in modeling cross-topic coupling of word rates):

- $\vec{\mu}_1 \sim \text{Normal}(\nu, \Phi)$, sample the mean of the topic mixing prior at time 1.
  - $\vec{\mu}_t = \text{Normal}(\mathbf{A}\vec{\mu}_{t-1}, \Phi)$, sample the means of the topic mixing priors over time.
  - $\vec{\theta}_d^{(t)} \sim \text{LogisticNormal}(\vec{\mu}_t, \Sigma_t)$, for each document, sample a topic proportion vector.
- $\vec{\zeta}_w^{(1)} \sim \text{Normal}(0, \Psi_w)$, sample rates for word w at time 1.
  - $\vec{\zeta}_w^{(t)} \sim \text{Normal}(\mathbf{B}_w\vec{\zeta}_w^{(t-1)}, \Psi_w)$, sample rates for word w over subsequent time points.
  - $\tau_{w,k}^{(t)} = \exp(\zeta_{w,k}^{(t)}), \quad \forall w = 1, \ldots, M,$ compute word rates.
- $n_{d,w}^{(t)} \sim \text{Poisson}(\omega \sum_k \theta_k \tau_{w,k})$ sample the word counts.

Figure 2 illustrates a graphical model representation of such a dynamic log-normal-Poisson model.

3 Variational Inference

3.1 Variational Inference for the Logistic-Normal-Multinomial Model

Under the Logistic-Normal-Multinomial topic evolution model, the complete likelihood function can be written as follows:

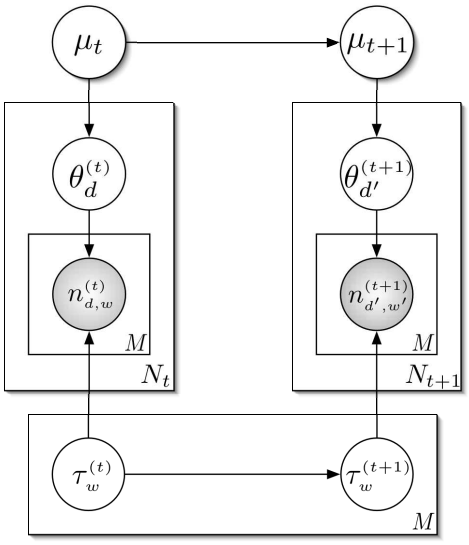

Figure 2: A graphical model representation of the dynamic log-normal-Poisson model for topic evolution. Note that the topic representations evolve as M independent word-rate vectors, each of which defines the rates of a word of a fixed set of topics.

$$
\begin{aligned}
& p(\mathcal{D}, \{\vec{\mu}_t\}, \{\vec{\theta}_d^{(t)}\}, \{\vec{\eta}_k^{(t)}\}, \{z_{d,n}^{(t)},\}) \\
= \;& p(\{\vec{\mu}_t\}) p(\{\vec{\theta}_d^{(t)}\}|\{\vec{\mu}_t\}) p(\{\vec{\eta}_k^{(t)}\}) p(\{z_{d,n}^{(t)}\}|\{\vec{\theta}_d^{(t)}\}) p(\{x_{d,n}^{(t)}\}|\{z_{d,n}^{(t)}\}, \{\vec{\eta}_k^{(t)}\}) \\
= \;& \Big(p(\vec{\mu}_1|\nu, \Phi) \prod_{t=2}^{T} p(\vec{\mu}_t|\vec{\mu}_{t-1}) \Big) \times \Big(\prod_{t=1}^{T} \prod_{d} p(\vec{\theta}_d^{(t)}|\mu_t) \Big) \times \Big(\prod_{k} \Big(p(\vec{\eta}_k^{(1)}|\iota_k, \Psi_k) \prod_{t=2}^{T} p(\vec{\eta}_k^{(t)}|\vec{\eta}_k^{(t-1)}) \Big) \Big) \\
& \times \Big(\prod_{t=1}^{T} \prod_{d=1}^{N_t} \prod_{n=1}^{N_d} p(z_{d,n}^{(t)}|\vec{\theta}_d^{(t)}) p(x_{d,n}^{(t)}|z_{d,n}^{(t)}, \{\vec{\eta}_k^{(t)}\}) \Big) \\
= \;& \Big(\mathcal{N}(\vec{\mu}_1|\nu, \Phi) \prod_{t=2}^{T} \mathcal{N}(\vec{\mu}_t|\mathbf{A}\vec{\mu}_{t-1}, \Phi) \Big) \times \Big(\prod_{t=1}^{T} \prod_{d} \mathcal{LN}(\vec{\theta}_d^{(t)}|\vec{\mu}_t, \Sigma_t) \Big) \times \Big(\prod_{k} \Big(\mathcal{N}(\vec{\eta}_k^{(1)}|\iota_k, \Psi_k) \prod_{t=2}^{T} \mathcal{N}(\vec{\eta}_k^{(t)}|\mathbf{B}\vec{\eta}_k^{(t-1)}, \Psi) \Big) \Big) \\
& \times \Big(\prod_{t=1}^{T} \prod_{d=1}^{N_t} \prod_{n=1}^{N_d} \text{Multinomial}(z_{d,n}^{(t)}|\vec{\theta}_d^{(t)}) \text{Multinomial}(x_{d,n}^{(t)}|z_{d,n}^{(t)}, \{\text{Logistic}(\vec{\eta}_k^{(t)})\}) \Big). \qquad (1)
\end{aligned}
$$

The posterior of $\big\{\{\vec{\mu}_t\}, \{\vec{\theta}_d^{(t)}\}, \{\vec{\eta}_k^{(t)}\}, \{z_{d,n}^{(t)}\}\big\}$ under the above model is intractable, therefore we approximate $p(\{\vec{\mu}_t\}, \{\vec{\theta}_d^{(t)}\}, \{\vec{\eta}_k^{(t)}\}, \{z_{d,n}^{(t)}\})$ with a product of simpler marginals, each on a cluster of latent variables: $q = q_\mu(\{\mu_t\}) q_\theta(\{\vec{\theta}_d^{(t)}\}) q_\eta(\{\vec{\eta}_k^{(t)}\}) q_z(\{z_{d,n}^{(t)}\})$. Based on the generalized mean field theorem [Xing *et al.*, 2003], the optimal parameterization of each marginal, $q(\cdot|\Theta)$, can be derived by plugging the generalized mean field (GMF) messages received by the cluster of variables (say, $\mathbf{X}_C$) under each marginal to the original conditional distribution of each variable cluster given its *Markov blanket* (MB); the GMF messages can be thought of as surrogates of the dependent variables $\mathbf{X}_{\text{MB}}$ in the Markov blanket of the cluster of variables under the marginal (e.g., $\mathbf{X}_C$), and they will be used to replace the original values of the dependent variables in the MB, e.g., $p(\mathbf{X}_C|\mathbf{X}_{\text{MB}}) \Rightarrow p\big(\mathbf{X}_C|\text{GMF}(\mathbf{X}_{\text{MB}})\big)$. [Xing *et al.*, 2003] showed that in case of generalized linear model, the generalized mean field message corresponds to an expectation of the *sufficient statistics* of the relevant Markov blanket variables under its associated GMF cluster marginal. In the sequel, we use $\langle S_{\mathbf{x}} \rangle_{q_{\mathbf{x}}}$ to denote the GMF message due to latent variable $\mathbf{x}$; and the optimal GMF approximation to $p(\mathbf{X}_C)$

is:

$$q^*(\mathbf{X}_C) = p\big(\mathbf{X}_C | \langle S_{\mathbf{y}} \rangle_{q_{\mathbf{y}}} : \forall \mathbf{y} \in \mathbf{X}_{\text{MB}}\big) \tag{2}$$

As a prelude for detailed derivations, we first rearrange some relevant local conditional distributions in our model into the canonical form of generalized linear models. As mentioned before, the multinomial parameters θ_k are logistic transformations of elements of a multivariate normal vector $\vec{\gamma}$: $\theta_t = e^{\gamma_k} / \sum_l e^{\gamma_l}$. In fact, since $\vec{\theta}$ is a multinomial parameter vector, it has only $K-1$ degree of freedom. Therefore, we only need to model a $K-1$ dimensional normal vector, and pad it with an vacuous element $\gamma_K = 0$. Under this parameterization, the logistic transformation from $\vec{\gamma}$ to $\vec{\theta}$ remains the same, but the inverse of this transformation takes a simple form: $\gamma_k = \ln \frac{\theta_k}{1-\sum_{i=1}^{K-1}\theta_i} = \ln \frac{\theta_k}{\theta_K}$. Assuming that z is a normalized K-dimensional random binary vector, that is, when z indicate the kth event, we have $z_k = 1$, $z_{i\neq k} = 0$, and $\sum_i z_i = 1$; the exponential family representation of a multinomial distribution for a topic indicator z is:

$$\begin{aligned} p(z|\vec{\theta}) = \exp\{\sum_{k=1}^{K} z_k \ln\theta_k\} &= \exp\Big\{ \sum_{k=1}^{K-1} z_k\big(\gamma_k - \ln(1+\sum_{k=1}^{K-1} e^{\gamma_k})\big) - z_K \ln(1+\sum_{k=1}^{K-1} e^{\gamma_k})\Big\} \\ &= \exp\Big\{ \sum_{k=1}^{K-1} z_k\gamma_k - \sum_{k=1}^{K} z_k \ln\big(1+\sum_{k=1}^{K-1} e^{\gamma_k}\big)\Big\}. \end{aligned} \tag{3}$$

For a collection of topic indicators $\{z_{d,n}^{(t)}\}$, we have the following conditional likelihood at time t:

$$\begin{aligned} p(\{z_{d,n}^{(t)} : \forall n\}|\vec{\theta}_d^{(t)}) &= \exp\Big\{ \sum_{k=1}^{K-1}\sum_n z_{d,n,k}^{(t)}\gamma_{d,k}^{(t)} - \sum_{k=1}^{K}\sum_n z_{d,n,k}^{(t)} \ln\big(1+\sum_{k=1}^{K-1}\exp\{\gamma_{d,k}^{(t)}\}\big)\Big\} \\ &= \exp\Big\{ \sum_{k=1}^{K-1} n_{d,k}^{(t)}\gamma_{d,k}^{(t)} - \sum_{k=1}^{K} n_{d,k}^{(t)} \ln\big(1+\sum_{k=1}^{K-1}\exp\{\gamma_{d,k}^{(t)}\}\big)\Big\} \\ &= \exp\Big\{\vec{m}_d^{(t)}\vec{\gamma}_d^{(t)} - \vec{n}_d^{(t)}\mathbf{1}\times c(\vec{\gamma}_d^{(t)}) \\ &= \exp\Big\{\vec{m}_d^{(t)}\vec{\gamma}_d^{(t)} - N_d^t c(\vec{\gamma}_d^{(t)})\Big\} \end{aligned} \tag{4}$$

where $n_{d,k}^{(t)} = \sum_n z_{d,n,k}^{(t)}$ is the number of words from topic k in document d at time t, $\vec{m}_d^{(t)} = (n_{d,1}^{(t)}, \ldots, n_{d,K-1}^{(t)})$ denotes the row vector of total word-counts from topics 1 to $K-1$ at time t; $\vec{n}_d^{(t)} = (\vec{m}_d^{(t)}, n_{d,K}^{(t)})$ denotes the row vector of total word-counts in document d from all topics; $\mathbf{1}$ is a column vector of all ones; and $c(\vec{\gamma}_d^{(t)}) = \ln\big(1+\sum_{k=1}^{K-1}\exp\{\gamma_k^{(t)}\}\big)$ is a scalar determined by $\vec{\gamma}_d^{(t)}$.

Similarly, the local conditional probability of the data $\{x_{d,n}^{(t)}\}$, where x is also defined as a M-dimensional norm-1 binary indicator vector, can be written as:

$$\begin{aligned} & p(\{x_{d,n}^{(t)} : \forall n,d\}|\{z_{d,n}^{(t)} : \forall n,d\},\{\vec{\eta}^{(t)}\}) \\ = \ & \exp\Big\{ \sum_{k=1}^{K}\sum_{w=1}^{M-1}\sum_{d,n} x_{d,n,w}^{(t)} z_{d,n,k}^{(t)} \eta_{k,w}^{(t)} - \sum_{k=1}^{K}\sum_{w=1}^{M}\sum_{d,n} x_{d,n,w}^{(t)} z_{d,n,k}^{(t)} \ln\big(1+\sum_{w=1}^{M-1}\exp\{\eta_{k,w}^{(t)}\}\big)\Big\} \\ = \ & \exp\Big\{ \sum_{k=1}^{K}\sum_{w=1}^{M-1} n_{k,w}^{(t)}\eta_{k,w}^{(t)} - \sum_{k=1,w=1}^{K,M} n_{k,w}^{(t)} \ln\big(1+\sum_{w=1}^{M-1}\exp\{\eta_{k,w}^{(t)}\}\big)\Big\} \\ = \ & \exp\Big\{ \sum_{k=1}^{K}\vec{m}_k^{(t)}\vec{\eta}_k^{(t)} - \sum_{k=1}^{K}\vec{n}_k^{(t)}\mathbf{1}\times c(\vec{\eta}_k^{(t)})\Big\} \\ = \ & \exp\Big\{ \sum_{k=1}^{K}\vec{m}_k^{(t)}\vec{\eta}_k^{(t)} - \sum_{k=1}^{K} N_k^t \times c(\vec{\eta}_k^{(t)})\Big\}, \end{aligned} \tag{5}$$

where $n^{(t)}_{k,w} = \sum_{d,n} x^{(t)}_{d,n,w} z^{(t)}_{d,n,k}$ is the count for word w from topic k at time t; $\vec{m}^{(t)}_k = (n^{(t)}_{k,1}, \ldots, n^{(t)}_{k,M-1})$ denotes the row vector of total word-counts of all but the last word of topic k at time t; $\vec{n}^{(t)}_k = (\vec{m}^{(t)}_k, n^{(t)}_{k,M})$ denotes the row vector of counts of every word generated from topic k at time t; and $c(\vec{\eta}^{(t)}_k)$ is a scalar determined by $\vec{\eta}^{(t)}_k$. Note that we have the following identity: $\vec{n}_t = \sum_k \vec{n}^{(t)}_k = \sum_d \vec{n}^{(t)}_d$.

With the above specifications of local conditional probability distributions, in the following we can write down one by one the GMF approximations to marginal posteriors of subsets of latent variables.

3.1.1

We first show that the marginal posterior of $\{\vec{\mu}_t\}$ can be approximated by a re-parameterized state-space model.

$$
\begin{aligned}
& q_\mu(\{\vec{\mu}_t\}) \\
= \;& p(\{\vec{\mu}_t\}|\langle S_{\boldsymbol{\theta}}\rangle_{q_\theta}) = p(\{\vec{\mu}_t\}|\langle S_{\boldsymbol{\gamma}}\rangle_{q_\gamma}) \\
\propto \;& \frac{1}{(2\pi)^{K-1/2}|\Phi|^{1/2}} \exp\{-\frac{1}{2}(\vec{\mu}_1 - \nu)'\Phi^{-1}(\vec{\mu}_1 - \nu)\} \times \big(\frac{1}{(2\pi)^{K-1/2}|\Phi|^{1/2}}\big)^T \exp\{-\frac{1}{2}\sum_{t=2}^{T}(\vec{\mu}_t - \mathbf{A}\vec{\mu}_{t-1})'\Phi^{-1}(\vec{\mu}_t - \mathbf{A}\vec{\mu}_{t-1})\} \\
& \times \prod_{t=1}^{T}\prod_{d=1}^{N_t} \frac{1}{(2\pi)^{K-1/2}|\Sigma_t|^{1/2}} \exp\{-\frac{1}{2}(\langle\vec{\gamma}^{(t)}_d\rangle - \vec{\mu}_t)'\Sigma_t^{-1}(\langle\vec{\gamma}^{(t)}_d\rangle - \vec{\mu}_t)\} \qquad (6)
\end{aligned}
$$

where $\langle\vec{\gamma}^{(t)}_d\rangle = (\langle\gamma^{(t)}_{d,1}\rangle, \ldots, \langle\gamma^{(t)}_{d,K-1}\rangle)$ is the expected topic vector of document d at time t, in which $\langle\gamma^{(t)}_{d,k}\rangle$ denotes the expectation of $\gamma^{(t)}_{d,k} = \ln\frac{\theta_{d,k}}{\theta_{d,K}}$ under variational marginal $q_\gamma(\{\vec{\gamma}^{(t)}_d\})$. For simplicity, we define $\vec{y}^t_d = \langle\vec{\gamma}^{(t)}_d\rangle$ as a short hand for the expected topic vector, and $\mathbf{Y}_t = \{\langle\vec{\gamma}^{(t)}_d\rangle\}_{d=1}^{N_t}$ as a short hand for all such vectors at time t.

Note that the above Eq. 6 is a linear Gaussian SSM, except that at each time the output is not a single observation $\langle\vec{\gamma}^{(t)}_d\rangle$, but a set of observations $\{\langle\vec{\gamma}^{(t)}_d\rangle\}_{d=1}^{N_t}$. It is well known that under a standard SSM, the posterior distribution of the centroid $\vec{\mu}_t$ given the entire observation sequence is still a normal distribution, of which the mean and covariance matrix can be readily estimated using the Kalman filtering (KF) and Rauch-Tung-Striebel (RTS) smoothing algorithms. Here we give the modified Kalman filter "measurement-update" equations that take into account multiple rather than single output data points [1]. The RTS equations and the "time-update" equations of KF is identical to the standard case for single output. Let $\hat{\mu}_{t|t}$ denote the mean of $\vec{\mu}_t$ conditioned on partial sequence $\mathbf{Y}_1, \ldots, \mathbf{Y}_t$. The convariance matrix of $\vec{\mu}_t$ conditioned on partial sequence $\mathbf{Y}_1, \ldots, \mathbf{Y}_t$ is denoted $P_{t|t}$; that is:

$$
\begin{aligned}
\hat{\mu}_{t|t} &\equiv E[\vec{\mu}_t|\mathbf{Y}_1, \ldots, \mathbf{Y}_t] \\
P_{t|t} &\equiv E[(\vec{\mu}_t - \hat{\mu}_{t|t}) \times (\vec{\mu}_t - \hat{\mu}_{t|t})'|\mathbf{Y}_1, \ldots, \mathbf{Y}_t].
\end{aligned}
$$

Similarly, we let $\hat{\mu}_{t+1|t}$ denotes the mean of μ_{t+1} conditioned on the partial sequence $\mathbf{Y}_1, \ldots, \mathbf{Y}_t$; $P_{t+1|t}$ denotes the covariance matrices of $\mu_{t+1|t}$ conditioned of partial sequences $\mathbf{Y}_1, \ldots, \mathbf{Y}_t$; and so on. Thus, the SSM inference formulae are as follows:

- Time update:

$$
\begin{aligned}
\hat{\mu}_{t+1|t} &= \mathbf{A}\hat{\mu}_{t|t} \qquad (8) \\
P_{t+1|t} &= \mathbf{A}P_{t|t} + \Phi \qquad (9)
\end{aligned}
$$

[1] This can be derived using the fact that the posterior mean and covariance matrix of the mean of a normal distribution $\mathcal{N}(\mu, \Sigma)$ given data $\mathbf{Y}$ and prior of the mean $\mathcal{N}(\mu_0, \Sigma_0)$ is:

$$
\Sigma_p = (n\Sigma^{-1} + \Sigma_0^{-1})^{-1}, \qquad \mu_p = (n\Sigma^{-1} + \Sigma_0^{-1})^{-1}(n\Sigma^{-1}\tilde{y} + \Sigma_0^{-1}\mu_0) \qquad (7)
$$

- Measurement update:

$$\begin{aligned}\hat{\mu}_{t+1|t+1} &= \hat{\mu}_{t+1|t} + P_{t+1|t}(P_{t+1|t} + \Sigma_t/N_t)^{-1}(\tilde{\gamma}_{t+1} - \hat{\mu}_{t+1|t}) \quad (10)\\ P_{t+1|t+1} &= P_{t+1|t} - P_{t+1|t}(P_{t+1|t} + \Sigma_t/N_t)^{-1}P_{t+1|t}, \quad (11)\end{aligned}$$

- RTS smoothing:

$$\begin{aligned}L_t &\equiv P_{t|t}\mathbf{A}'P_{t+1|t} \quad (12)\\ \hat{\mu}_{t|T} &= \hat{\mu}_{t|t} + L_t(\hat{\mu}_{t+1|T} - \hat{\mu}_{t+1|t}) \quad (13)\\ P_{t|T} &= P_{t|t} + L_t\big(P_{t+1|T} - P_{t+1|t}\big)L_t' \quad (14)\end{aligned}$$

where $\tilde{\gamma}_{t+1}$ denote the sample mean of observations at time $t+1$: $\tilde{\gamma}_{t+1} = \frac{1}{N_{t+1}}\sum_d \vec{y}_d^{t+1}$.

To estimate the state dynamic matrix $\mathbf{A}$, we also need to compute the cross-time convariance matrix about $\vec{\mu}_t$ and $\vec{\mu}_{t-1}$ conditioned on complete sequence $\mathbf{Y}_1, \ldots, \mathbf{Y}_T$: $P_{t,t-1|T} \equiv E[(\vec{\mu}_t - \hat{\mu}_{t|T}) \times (\vec{\mu}_{t-1} - \hat{\mu}_{t-1|T})'|\mathbf{Y}_1, \ldots, \mathbf{Y}_t]$. It can be shown that $P_{t-1\to t|T}$ satisfies the following backward recursion:

$$P_{t-1\to t|T} = P_{t|t}L_{t-1}' + L_t(P_{t\to t+1|T} - \mathbf{A}P_{t|t})L_{t-1}', \quad (15)$$

which is initialized by $P_{T-1\to T|T} = (I - K_T)\mathbf{A}P_{T-1|T-1}$, where $K_T \equiv P_{t+1|t}(P_{t+1|t} + \Sigma_t/N_t)^{-1}$ is the *Kalman gain matrix*.

3.1.2

Now we move on to the variational marginal $q_\gamma(\{\vec{\gamma}_d^{(t)}\})$.

$$\begin{aligned}& q_\gamma(\{\vec{\gamma}_d^{(t)}\})\\ =\ & p(\{\vec{\gamma}_d^{(t)}\}|\langle S_\mu\rangle_{q_\mu}, \langle S_{\mathbf{z}}\rangle_{q_z})\\ =\ & \prod_{t=1}^{T}\prod_{d=1}^{N_t} p(\vec{\gamma}_d^{(t)}|\langle S_{\mu_t}\rangle_{q_\mu}, \langle S_{\mathbf{z}_{t,d}}\rangle_{q_z})\\ \propto\ & \prod_{t=1}^{T}\prod_{d=1}^{N_t} \frac{1}{(2\pi)^{K-1/2}|\Sigma_t|^{1/2}} \exp\{-\frac{1}{2}(\vec{\gamma}_d^{(t)} - \langle\vec{\mu}_t\rangle)'\Sigma_t^{-1}(\vec{\gamma}_d^{(t)} - \langle\vec{\mu}_t\rangle)\} \exp\Big\{\langle\vec{m}_d^{(t)}\rangle\vec{\gamma}_d^{(t)} - \langle\vec{n}_d^{(t)}\rangle\mathbf{1}\times c(\vec{\gamma}_d^{(t)})\Big\}, \quad (16)\end{aligned}$$

where $\langle\vec{n}_d^{(t)}\rangle = (\langle n_{d,1}^{(t)}\rangle, \ldots, \langle n_{d,K}^{(t)}\rangle)$ (c.f. $\langle\vec{m}_d^{(t)}\rangle$), and $\langle n_{d,k}^{(t)}\rangle = \sum_n \langle z_{d,n,k}^{(t)}\rangle$ denotes the sum of expected topic-specific counts for each word in document d under $q_z(\{z_{d,n}^{(t)}\})$ (which will be specified in the sequel).

Due to the complexity of $c(\vec{\gamma}_d^{(t)}) = \ln\big(1 + \sum_{k=1}^{K-1}\exp\{\gamma_{d,k}^{(t)}\}\big)$, the $q(\vec{\gamma}_d^{(t)})$ defined above is not integratable during inference (e.g., for computing an expectation of $\vec{\gamma}$). In [Blei and Lafferty, 2006], an variational approximation based on optimizing a relaxed bound of the KL-divergence between $q(\cdot)$ and $p(\cdot)$ is used to approximate $q(\vec{\gamma}_d^{(t)})$. In the following, we present a different approach that overcome the non-conjugacy between the multinomial likelihood and the logistic-normal prior, and make the joint tractable. We seek a normal approximation to $q(\vec{\gamma}_d^{(t)})$ using Taylor expansion technique. Let's take a second-order Taylor expansion of $c(\vec{\gamma})$ with respect to $\vec{\gamma}$:

$$\begin{aligned}\frac{\partial c}{\partial\gamma_i} &= \frac{e^{\gamma_i}}{1 + \sum_{k=1}^{K-1} e^{\gamma_k}} \equiv g_i\\ \frac{\partial^2 c}{\partial\gamma_i\partial\gamma_i} &= \frac{\partial}{\partial\gamma_i}\Big(\frac{e^{\gamma_i}}{1 + \sum_{k=1}^{K-1} e^{\gamma_k}}\Big) = \frac{-e^{\gamma_i}e^{\gamma_i} + \big(1 + \sum_{k=1}^{K-1} e^{\gamma_k}\big)e^{\gamma_i}}{\big(1 + \sum_{k=1}^{K-1} e^{\gamma_k}\big)^2}\\ &= \frac{\big(1 + \sum_{k=1,k\neq i}^{K-1} e^{\gamma_k}\big)e^{\gamma_i}}{\big(1 + \sum_{k=1}^{K-1} e^{\gamma_k}\big)^2} \equiv h_{ii}\\ \frac{\partial^2 c}{\partial\gamma_i\partial\gamma_j} &= \frac{\partial}{\partial\gamma_j}\Big(\frac{e^{\gamma_i}}{1 + \sum_{k=1}^{K-1} e^{\gamma_k}}\Big) = \frac{-e^{\gamma_j}e^{\gamma_i}}{\big(1 + \sum_{k=1}^{K-1} e^{\gamma_k}\big)^2} \equiv h_{ij}. \quad (17)\end{aligned}$$

Therefore, the 2nd-order Taylor series of $c(\vec{\gamma}) = \ln\big(1 + \sum_{k=1}^{K-1} \exp\{\gamma_k\}\big)$ with respect to some $\hat{\gamma}$ is:

$$c(\vec{\gamma}) = c(\hat{\gamma}) + \vec{g}'_\gamma(\vec{\gamma} - \hat{\gamma}) + \frac{1}{2}(\vec{\gamma} - \hat{\gamma})' H_\gamma (\vec{\gamma} - \hat{\gamma}) + R_2, \tag{18}$$

where $\vec{g}'_\gamma = \nabla_\gamma c(\vec{\gamma}) = (g_1, \ldots, g_{K-1})$ denotes the gradient vector of $c(\vec{\gamma})$, $H_\gamma = \{h_{ij}\}$ denotes the Hessian matrix, and R_2 is the Lagrange remainder. Assuming that $\hat{\gamma}$ is close enough to the true γ for each document at each time (e.g., the posterior mean of all γ), we have:

$$c(\vec{\gamma}) \approx c(\hat{\gamma}) + \vec{g}'_\gamma(\hat{\gamma})(\vec{\gamma} - \hat{\gamma}) + \frac{1}{2}(\vec{\gamma} - \hat{\gamma})' H_\gamma(\hat{\gamma})(\vec{\gamma} - \hat{\gamma}). \tag{19}$$

It can be shown that since $c(\vec{\gamma})$ is convex w.r.t. $\vec{\gamma}$, the above approximation is a 2nd-order polymornial lower bound of $c(\vec{\gamma})$ [Jordan *et al.*, 1999].

Now we have:

$$\begin{aligned}
& p(\vec{\gamma}_d^{(t)} | \langle S_{\mu_t} \rangle_{q_\mu}, \langle S_{\mathbf{z}_{t,d}} \rangle_{q_z}) \\
\propto\ & \frac{1}{(2\pi)^{K-1/2} |\Sigma_t|^{1/2}} \exp\{-\frac{1}{2}(\vec{\gamma}_d^{(t)} - \langle \vec{\mu}_t \rangle)' \Sigma_t^{-1} (\vec{\gamma}_d^{(t)} - \langle \vec{\mu}_t \rangle)\} \exp\Big\{ \langle \vec{m}_d^{(t)} \rangle' \vec{\gamma}_d^{(t)} - \langle \vec{n}_d^{(t)} \rangle \mathbf{1} \times c(\vec{\gamma}_d^{(t)}) \Big\} \\
\approx\ & \frac{1}{(2\pi)^{K-1/2} |\Sigma_t|^{1/2}} \exp\Big\{ -\frac{1}{2} \vec{\gamma}'^{(t)}_d \Sigma_t^{-1} \vec{\gamma}_d^{(t)} + \vec{\gamma}'^{(t)}_d \Sigma_t \langle \vec{\mu}_t \rangle - \frac{1}{2} \langle \vec{\mu}_t \rangle' \Sigma_t \langle \vec{\mu}_t \rangle + \vec{\gamma}'^{(t)}_d \langle \vec{m}_d^{(t)} \rangle \\
& - N_d^t \big(c(\hat{\gamma}) + \vec{g}'_\gamma(\hat{\gamma})(\vec{\gamma}_d^{(t)} - \hat{\gamma}) + \frac{1}{2}(\vec{\gamma}_d^{(t)} - \hat{\gamma})' H_\gamma(\hat{\gamma})(\vec{\gamma}_d^{(t)} - \hat{\gamma}) \big) \Big\} \\
\propto\ & \exp\Big\{ -\frac{1}{2} \vec{\gamma}'^{(t)}_d \big(\Sigma_t^{-1} + N_d^t H_\gamma(\hat{\gamma}) \big) \vec{\gamma}_d^{(t)} + \vec{\gamma}'^{(t)}_d \big(\Sigma_t^{-1} \langle \vec{\mu}_t \rangle + \langle \vec{m}_d^{(t)} \rangle - N_d^t \vec{g}_\gamma(\hat{\gamma}) + N_d^t H_\gamma(\hat{\gamma}) \hat{\gamma} \big) \Big\}
\end{aligned} \tag{20}$$

Rearranging the terms, and setting $\hat{\gamma} = \langle \vec{\mu}_t \rangle = \hat{\mu}_{t|T}$ (from §2.1.1), we have the following multivariate-normal approximation:

$$p(\vec{\gamma}_d^{(t)} | \langle S_{\mu_t} \rangle_{q_\mu}, \langle S_{\mathbf{z}_{t,d}} \rangle_{q_z}) \approx \mathcal{N}(\vec{\gamma}_d^{(t)} | \tilde{\mu}_d^t, \tilde{\Sigma}_d^t), \tag{21}$$

where

$$\tilde{\Sigma}_d^t = \mathrm{inv}\big(\Sigma_t^{-1} + N_d^t H_\gamma(\hat{\mu}_{t|T}) \big), \tag{22}$$

$$\begin{aligned}
\tilde{\mu}_d^t &= (\tilde{\Sigma}_d^t)^{-1} \big((\Sigma_t^{-1} \hat{\mu}_{t|T} + N_d^t H_\gamma(\hat{\mu}_{t|T}) \hat{\mu}_{t|T} + \langle \vec{m}_d^{(t)} \rangle - N_d^t \vec{g}_\gamma(\hat{\mu}_{t|T}) \big) \\
&= \hat{\mu}_{t|T} + (\tilde{\Sigma}_d^t)^{-1} \big(\langle \vec{m}_d^{(t)} \rangle - N_d^t \vec{g}_\gamma(\hat{\mu}_{t|T}) \big).
\end{aligned} \tag{23}$$

3.1.3

Now we compute variational marginal $q_\beta(\{\vec{\beta}_k^{(t)}\})$.

$$q_\beta(\{\vec{\beta}_k^{(t)}\}) = \prod_{k=1}^{K} p(\vec{\beta}_k^{(1)}, \ldots, \vec{\beta}_k^{(T)} | \langle S_{\mathbf{z}} \rangle_{q_z}) \tag{24}$$

This is a product of conditionally independent SSMs (given sufficient statistics $\langle S_\gamma \rangle_{q_\gamma}$, $\langle S_{\mathbf{z}} \rangle_{q_z}$, model parameters $\{\iota_k, \Psi_k, \mathbf{B}_k\}$, and data $\mathcal{D}$). The variational marginal of a single chain of a evolving topic represented in pre-transformed normal vector $\vec{\eta}_k^{(1)}, \ldots, \vec{\eta}_k^{(T)}$ is:

$$\begin{aligned}
& p(\vec{\eta}_k^{(1)}, \ldots, \vec{\eta}_k^{(T)} | \langle S_{\mathbf{z}} \rangle_{q_z}) \\
\propto\ & \Big(\frac{1}{(2\pi_k)^{(M-1)/2} |\Phi_k|^{1/2}} \Big)^T \exp\Big\{ -\frac{1}{2}(\vec{\eta}_k^{(1)} - \iota_k)' \Psi_k^{-1} (\vec{\eta}_k^{(1)} - \iota_k) - \frac{1}{2} \sum_{t=2}^{T} (\vec{\eta}_k^{(t)} - \mathbf{B}_k \vec{\eta}_k^{(t-1)})' \Psi^{-1} (\vec{\eta}_k^{(t)} - \mathbf{B}_k \vec{\eta}_k^{(t-1)}) \Big\} \\
& \times \prod_{t=1}^{T} \exp\Big\{ \langle \vec{m}_k^{(t)} \rangle \vec{\eta}_k^{(t)} - \langle \vec{n}_k^{(t)} \rangle \mathbf{1} \times c(\vec{\eta}_k^{(t)}) \Big\}.
\end{aligned} \tag{25}$$

Recall that we can approximate $c(\vec{\eta}_k^{(t)})$ with its second-order truncated Taylor series with respect to an estimate of $\vec{\eta}_k^{(t)}$, say :

$$c(\vec{\eta}_k^{(t)}) \approx c(\hat{\vec{\eta}}_k^{(t)}) + g(\hat{\vec{\eta}}_k^{(t)})(\vec{\eta}_k^{(t)} - \hat{\vec{\eta}}_k^{(t)}) + \frac{1}{2}(\vec{\eta}_k^{(t)} - \hat{\vec{\eta}}_k^{(t)})'\mathbf{H}(\hat{\vec{\eta}}_k^{(t)})(\vec{\eta}_k^{(t)} - \hat{\vec{\eta}}_k^{(t)}). \tag{26}$$

In the following we first outline a normal approximation to a multinomial distribution of count vector n. In particular, we assume that the multinomial parameters are logistic transformations of a real vector η:

$$\begin{aligned}
p(n|\eta) &= \exp\Big\{n'\eta + N\ln(1 - \sum_{w=1}^{M-1}\eta_w)\Big\} \\
&\approx \exp\Big\{(n' - Ng' + \hat{\eta}'N\mathbf{H})\eta - \frac{1}{2}\eta'N\mathbf{H}\eta - \frac{1}{2}\hat{\eta}'N\mathbf{H}\hat{\eta}' - N\ln(1 + \sum_{w=1}^{M-1}\hat{\eta}_w) + N\hat{\eta}'g\Big\} \\
&= \exp\Big\{ -\frac{1}{2}\big((N\mathbf{H})^{-1}(n - Ng + N\mathbf{H}\hat{\eta}) - \eta\big)'N\mathbf{H}\big((N\mathbf{H})^{-1}(n - Ng + N\mathbf{H}\hat{\eta}) - \eta\big) - \\
&\quad N\ln(1 + \sum_{w=1}^{M-1}\hat{\eta}_w) + \frac{1}{2}(n - Ng)'(N\mathbf{H})^{-1}(n - Ng) + n'\hat{\eta}\Big\} \\
&\approx \mathcal{N}(v|\eta, (N\mathbf{H})^{-1})
\end{aligned} \tag{27}$$

where $g = \nabla_\eta c|_{\eta=\hat{\eta}}$, $\mathbf{H} = \text{Hessian}_\eta c|_{\eta=\hat{\eta}}$, $v = (N\mathbf{H})^{-1}(n - Ng + N\mathbf{H}\hat{\eta}) = \hat{\eta} + (N\mathbf{H})^{-1}(n - Ng)$; and the Taylor expansion point $\hat{\eta}$ can be set at the empirical estimate or just a guess of η.

With this approximation, we can approximate Eq. 25 by an SSM with linear Gaussian emission models:

$$\begin{aligned}
& p(\vec{\eta}_k^{(1)}, \ldots, \vec{\eta}_k^{(T)}|\langle S_{\mathbf{z}}\rangle_{q_z}) \\
\approx\ & \Big(\frac{1}{(2\pi_k)^{(M-1)/2}|\Phi_k|^{1/2}}\Big)^T \exp\Big\{ -\frac{1}{2}(\vec{\eta}_k^{(1)} - \iota_k)'\Psi_k^{-1}(\vec{\eta}_k^{(1)} - \iota_k) - \frac{1}{2}\sum_{t=2}^{T}(\vec{\eta}_k^{(t)} - \mathbf{B}_k\vec{\eta}_k^{(t-1)})'\Psi^{-1}(\vec{\eta}_k^{(t)} - \mathbf{B}_k\vec{\eta}_k^{(t-1)})\Big\} \\
& \times \prod_{t=1}^{T}\exp\Big\{\langle\vec{m}_k^{(t)}\rangle\vec{\eta}_k^{(t)} - \langle N_k\rangle c(\hat{\vec{\eta}}_k^{(t)}) - \langle N_k\rangle g(\hat{\vec{\eta}}_k^{(t)})(\vec{\eta}_k^{(t)} - \hat{\vec{\eta}}_k^{(t)}) - \frac{\langle N_k\rangle}{2}(\vec{\eta}_k^{(t)} - \hat{\vec{\eta}}_k^{(t)})'H(\hat{\vec{\eta}}_k^{(t)})(\vec{\eta}_k^{(t)} - \hat{\vec{\eta}}_k^{(t)}))\Big\} \\
\approx\ & \mathcal{N}(\vec{\eta}_k^{(1)}|\iota_k, \Phi_k)\prod_{t=2}^{T}\mathcal{N}(\vec{\eta}_k^{(t)}|\mathbf{B}_k\vec{\eta}_k^{(t-1)}, \Phi_k) \times \prod_{t=1}^{T}\mathcal{N}(\vec{v}_k^{(t)}|\vec{\eta}_k^{(t)}, (N\mathbf{H})^{-1}),
\end{aligned} \tag{28}$$

where the "observation" $\vec{v}_k^{(t)} = \hat{\eta}_k^{(t)} + \big(\langle N_k^{(t)}\rangle\mathbf{H}(\hat{\vec{\eta}}_k^{(t)})\big)^{-1}\Big(\langle\vec{m}_k^{(t)}\rangle - \langle N_k^{(t)}\rangle g(\hat{\vec{\eta}}_k^{(t)})\Big)$, $\hat{\eta}_k^{(t)}$ can be set to be its estimate in the previous round of GMF iteration (see §2.1.5); and the expectation of the count vector $\langle\vec{m}_k^{(t)}\rangle$ and total word count $\langle N_k^{(t)}\rangle$ associated with topic k at time t can be computed using variation marginal $q_{\mathbf{z}}(\cdot)$, specifically, $\langle n_{k,w}^{(t)}\rangle = \sum_{d,n} x_{d,n,w}^{(t)}\langle z_{d,n,k}^{(t)}\rangle$ is the expected count for word w from topic k at time t; $\langle\vec{m}_k^{(t)}\rangle = (\langle n_{k,1}^{(t)}\rangle, \ldots, \langle n_{k,M-1}^{(t)}\rangle)$ denotes the expected row vector of total word-counts of all but the last word of topic k at time t; $\langle\vec{n}_k^{(t)}\rangle = (\langle\vec{m}_k^{(t)}\rangle, \langle n_{k,M}^{(t)}\rangle)$ denotes the row vector of counts of every word generated from topic k at time t; $\langle N_k^{(t)}\rangle = |\langle\vec{n}_k^{(t)}\rangle|$.

Now the posterior of $\vec{\eta}_k^{(t)}$ can be approximate by a multivariate Gaussian $\mathcal{N}(\cdot|\hat{\eta}_{k,t|T}, P_{k,t|T})$, here we give the formula for the KF time/measurement update, and the RTS smoothing of the topics distribution parameters at time t:

- Time update:

$$\begin{aligned}
\hat{\eta}_{k,t+1|t} &= \mathbf{B}_k\hat{\eta}_{k,t|t} \\
P_{k,t+1|t} &= \mathbf{B}_kP_{k,t|t} + \Psi_k
\end{aligned} \tag{29}$$

- Measurement update:

$$\begin{aligned}\hat{\eta}_{k,t+1|t+1} &= \hat{\eta}_{k,t+1|t} + P_{k,t+1|t}\big(P_{k,t+1|t} + (N\mathbf{H})^{-1}\big)^{-1}(\vec{v}_k^{(t)} - \hat{\eta}_{k,t+1|t}) \\ P_{k,t+1|t+1} &= P_{k,t+1|t} - P_{k,t+1|t}\big(P_{k,t+1|t} + (N\mathbf{H})^{-1}\big)^{-1}P_{k,t+1|t} \end{aligned} \tag{30}$$

- RTS smoothing:

$$\begin{aligned} L_{k,t} &\equiv P_{k,t|t}\mathbf{B}_k' P_{k,t+1|t} \\ \hat{\eta}_{k,t|T} &= \hat{\eta}_{k,t|t} + L_{k,t}(\hat{\eta}_{k,t+1|T} - \hat{\eta}_{k,t+1|t}) \\ P_{k,t|T} &= P_{k,t|t} + L_{k,t}\big(P_{k,t+1|T} - P_{k,t+1|t}\big)L_{k,t}' \\ P_{k,t-1\to t|T} &= P_{k,t|t}L_{k,t-1}' + L_{k,t}(P_{k,t\to t+1|T} - \mathbf{B}_k P_{k,t|t})L_{k,t-1}', \end{aligned} \tag{31}$$

We can estimate the parameters ι_k, Ψ_k and $\mathbf{B}_k$ using an EM algorithm.

3.1.4

Now we compute variational marginal $q_{\mathbf{z}}(\{z_{d,n}^{(t)}\})$.

$$\begin{aligned} & p(\mathbf{z}|\mathcal{D}, \langle S_\gamma\rangle_{q_\gamma}, \langle S_\eta\rangle_{q_\eta}) \\ = & \prod_{d,n,t} p(z_{d,n}^{(t)}|x_{d,n}^{(t)}, \langle S(\gamma_d^{(t)})\rangle_{q_\gamma}, \{\langle S(\eta_k^{(t)})\rangle_{q_\eta}\}) \end{aligned} \tag{32}$$

For notational simplicity, we omit indices, and give bellow a generic formula for the variational approximation for singleton marginal: (Recall that z is a unit base vector, thus $|z| = \sum_k z_k = 1$. Similar definition applies to x.)

$$\begin{aligned} p(z|x, \langle S_\gamma\rangle, \langle S_\eta\rangle) &\propto p(z|\langle S_\gamma\rangle)p(x|z, \langle S_\eta\rangle) \\ &= \exp\big\{z'\langle\vec{\gamma}\rangle - \langle c(\vec{\gamma})\rangle + x'\langle\mathbf{\Xi}\rangle z - \sum_k z_k\langle c(\vec{\eta}_k)\rangle\big\}, \end{aligned} \tag{33}$$

where γ follows a Gaussian distribution, and $\mathbf{\Xi}$ is an $(M-1)\times K$ matrix whose column vectors $\vec{\eta}_k$ also follows a Gaussian distribution.

The close form solutions of $\langle c(\vec{\gamma})\rangle$ and $\langle c(\vec{\eta}_k)\rangle$ under normal distribution is not available. Note that the multinomial parameter vector $\theta = \frac{1}{\sum_k \pi_k}(\pi_1, \ldots, \pi_{K-1}, \pi_K)$, where vector $\vec{\pi} = (\pi_1, \ldots, \pi_{K-1})$ follows a multivariate log-normal distribution, and $\pi_K = 1$. To better approximate $\langle c(\vec{\gamma})\rangle$ (and similarly also $\langle c(\vec{\eta}_k)\rangle$), we rewrite $c(\vec{\gamma}) = \ln(1 + \sum_{k=1}^{K-1}\exp\gamma_k) = c(\vec{\pi}) = \ln(|\vec{\pi}|)$, where $\vec{\pi}$ is the unnormalized version of multinomial parameter vector $\vec{\theta}$. Now we expand $c(\vec{\pi})$ w.r.t. to $\vec{\pi}$ around the mean of $\vec{\pi}$ up to the second order.

The gradient of $c(\vec{\pi})$ w.r.t. to $\vec{\pi}$ is:

$$\begin{aligned} \frac{\partial}{\partial\pi_i}\ln(|\vec{\pi}|) &= \frac{1}{|\vec{\pi}|} \\ \Rightarrow \quad \nabla_\pi \ln(|\vec{\pi}|) &= \frac{1}{|\vec{\pi}|}\mathbf{1} \equiv \vec{g}_\pi. \end{aligned} \tag{34}$$

The Hessian of $c(\vec{\pi})$ w.r.t. to $\vec{\pi}$ is:

$$\begin{aligned} \frac{\partial^2}{\partial\pi_i\partial\pi_i}\ln(|\vec{\pi}|) &= \frac{\partial}{\partial\pi_i}\Big(\frac{1}{|\vec{\pi}|}\Big) = -\frac{1}{|\vec{\pi}|^2} \\ \frac{\partial^2}{\partial\pi_i\partial\pi_j}\ln(|\vec{\pi}|) &= -\frac{1}{|\vec{\pi}|^2} \\ \Rightarrow \quad \mathbf{H}_\pi \ln(|\vec{\pi}|) &= -\frac{1}{|\vec{\pi}|^2}\begin{vmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 1 \end{vmatrix} = -\frac{1}{|\vec{\pi}|^2}\mathbf{1}\times\mathbf{1}', \end{aligned} \tag{35}$$

where $\mathbf{1} \times \mathbf{1}'$ represents an outer product of the two one-vectors.

Therefore,

$$c(\vec{\pi}) \approx \ln(|\hat{\pi}|) + \vec{g}'_{\pi}(\pi - \hat{\pi}) + \frac{1}{2}(\pi - \hat{\pi})'\mathbf{H}_{\pi}(\pi - \hat{\pi}). \tag{36}$$

Let $\hat{\pi} = E[\vec{\pi}]$ under $\vec{\pi} \sim \mathrm{LN}_{K-1}(\vec{\mu}, \Sigma)$, then we have:

$$\begin{aligned} \langle c(\vec{\pi}) \rangle &\approx \ln(|E[\vec{\pi}]|) + \frac{1}{2}\mathrm{Tr}\big(\mathbf{H}_{\pi}(E[\vec{\pi}]) \cdot E[(\vec{\pi} - E[\vec{\pi}])(\vec{\pi} - E[\vec{\pi}])']\big) \\ &= \ln(|E[\vec{\pi}]|) + \frac{1}{2}\mathrm{Tr}\big(\mathbf{H}_{\pi}(E[\vec{\pi}]) \cdot \hat{\Sigma}_{\vec{\pi}}\big), \end{aligned} \tag{37}$$

where $\hat{\Sigma}_{\vec{\pi}}$ is the covariance of $\vec{\pi}$ under a multivariate log-normal distribution. It can be shown that [Kleiber and Kotz, 2003]:

$$cov(\pi_i, \pi_j) = \exp\big\{\mu_i + \mu_j + \frac{1}{2}(\sigma_{ii} + \sigma_{jj})\big\}\big(\exp(\sigma_{ij}) - 1\big) \tag{38}$$

$$E(\pi_i) = \exp\big\{\mu_i + \frac{1}{2}\sigma_{ii}\big\} \tag{39}$$

$$E[\vec{\pi}] = \exp\{\vec{\mu} + \frac{1}{2}\mathrm{Diag}(\Sigma)\}. \tag{40}$$

This computation can be applied to the expectation of both the pre-normalized topic proportion vector $\vec{\pi}$ and the pre-normalized topic-specific word frequency vector $\vec{\xi}_k$ corresponding to $\vec{\beta}_k$. So, we have

$$\begin{aligned} p(z|x, \langle S_{\gamma} \rangle, \langle S_{\eta} \rangle) &\propto \exp\big\{z'\langle\vec{\gamma}\rangle - \langle c(\vec{\pi})\rangle + x'\langle\mathbf{\Xi}\rangle z - \sum_k z_k \langle c(\vec{\xi}_k)\rangle\big\} \\ &= \exp\Big\{z'E[\vec{\gamma}] - \ln(|E[\vec{\pi}]|) - \frac{1}{2}\mathrm{Tr}\big(\mathbf{H}_{\pi}(E[\vec{\pi}]) \cdot \hat{\Sigma}_{\vec{\pi}}\big) \\ &\quad + x'E[\mathbf{\Xi}]z - \sum_k z_k\Big(\ln(|E[\vec{\xi}_k]|) - \frac{1}{2}\mathrm{Tr}\big(\mathbf{H}_{\xi}(E[\vec{\xi}_k]) \cdot \hat{\Sigma}_{\vec{\xi}_k}\big)\Big)\Big\}, \end{aligned} \tag{41}$$

where $\mathbf{H}_{\xi}(E[\vec{\xi}_k]) = \frac{1}{|E[\vec{\xi}_k]|^2}\mathbf{1} \times \mathbf{1}'$, $E[\vec{\xi}_k]$ and $\hat{\Sigma}_{\vec{\xi}_k}$ are the mean and covariance of $\vec{\xi}_k$ under a log-normal distribution as in Eqs. (38)-(40), $E[\mathbf{\Xi}]$ consists of column-by-column expectation, $E[\vec{\eta}_k]$, under a normal distribution of $\vec{\eta}_k$. Note that in the above computation, one must be careful about appropriately recovering the K-dimensional multinomial distribution of z from the $(K-1)$-dimensional pre-tranformed natural parameter vector $\vec{\gamma}$, and the $(M-1) \times K$ dimensional pre-tranformed natural parameter matrix $\mathbf{\Xi} = (\vec{\eta}_1, \ldots, \vec{\eta}_K)$. I omit details of such manipulations.

We need to compute the above singleton marginal for each $z_{d,n}^{(t)}$ given $\langle\vec{\gamma}_d^{(t)}\rangle, \langle\vec{\pi}_d^{(t)}\rangle, \{\langle\vec{\eta}_k^{(t)}\rangle\}$, and $\{\langle\vec{\xi}_k^{(t)}\rangle\}$; where $\vec{\pi}_d^{(t)} \sim \mathrm{LN}(\tilde{\mu}_d^{(t)}, \tilde{\Sigma}_d^{(t)})$, $\vec{\gamma}_d^{(t)} \sim \mathrm{N}(\tilde{\mu}_d^{(t)}, \tilde{\Sigma}_d^{(t)})$, $\vec{\xi}_k^{(t)} \sim \mathrm{LN}(\hat{\eta}_{k,t|T}, P_{k,t|T})$, and $\vec{\eta}_k^{(t)} \sim \mathrm{N}(\hat{\eta}_{k,t|T}, P_{k,t|T})$.

3.1.5 Summary

The above 4 variational marginals are coupled and thus constitute a set of fixed-point equations, computing the GMF message for one marginal require the marginals of other sets of variables. Thus, we can iteratively update each marginal until convergence (i.e., all the GMF messages stop changing). This approximation scheme can be shown to minimize the KL divergence between the variational posterior and the true posterior of latent variables.

We can use a variational EM scheme to estimate the parameters in our model, which are essentially the SSM parameters. Operationally, VEM is no different from a standard EM for SSM—we have observation sequence $\tilde{\gamma}_t = \frac{1}{N_t}\sum_d \langle\vec{\gamma}_d^{(t)}\rangle$ for the topic mixing SSM (as defined by $q_{\mu}(\cdot)$), and observation sequence $\vec{u}_k^{(k)}$ for each of the K topic representation (i.e., word-frequency) SSMs (as defined by $q_{\eta_k}(\cdot)$); and we can use the standard learning learning rules for SSM parameter estimation.

In the E step, we use Eqs.(13)-(14), and Eqs.(31)-(31) to estimate the expected sufficient statistics of the latent states; In the M step, we update the parameters Φ, $\mathbf{A}$, Σ_t of the topic mixing SSM, and Ψ_k, $\mathbf{B}_k$ of the topic representation SSM (see [Ghahramani and Hinton, 1996; Ghahramani and Hinton, 1998] for details).

3.2 Variational Inference for Log-Normal-Poisson model

Now we need to approximate

$$n_w \sim \text{Poisson}(N \sum_k \theta_k \tau_{w,k}) = \exp\{n_w \ln(N \sum_k \theta_k \tau_{w,k}) - N \sum_k \theta_k \tau_{w,k} - \Gamma(n_w + 1)\}.$$

Again, let $\vec{\pi}$ denote the unnormalized version of the multinomial parameter vector $\vec{\theta}$. Note that the Jacobian of vector $\vec{\theta}$ with respect to $\vec{\pi}$ is:

$$\begin{aligned}
\frac{\partial \theta_i}{\partial \pi_i} &= \frac{\sum_k \pi_k - \pi_i}{(\sum_k \pi_k)^2} = \frac{\sum_{k \neq i} \pi_k}{(\sum_k \pi_k)^2} \\
\frac{\partial \theta_i}{\partial \pi_j} &= -\frac{\pi_i}{(\sum_k \pi_k)^2} \\
J(\theta) &= \frac{1}{(\sum_k \pi_k)^2} \begin{vmatrix} \sum_{k \neq 1} \pi_k & -\pi_1 & \cdots & -\pi_1 \\ -\pi_1 & \sum_{k \neq 2} \pi_k & \cdots & -\pi_1 \\ \vdots & \vdots & \ddots & \vdots \\ -\pi_K & -\pi_K & \cdots & \sum_{k \neq K} \pi_k \end{vmatrix}
\end{aligned} \tag{42}$$

From this derivation, we know that

$$\begin{aligned}
\frac{\partial \vec{\theta}}{\partial \pi_i} &= \frac{-1}{(\sum_k \pi_k)^2} (\pi_1, \ldots, \sum_{k \neq i} \pi_k, \ldots, \pi_K)' \\
&= \frac{-1}{(\sum_k \pi_k)^2} \big(|\vec{\pi}| e_i - \vec{\pi}\big) \\
\Rightarrow \quad \frac{\partial \vec{\theta}' \vec{\tau}_w}{\partial \pi_i} &= \frac{-1}{(\sum_k \pi_k)^2} \big(|\vec{\pi}| e_i - \vec{\pi}\big)' \vec{\tau}_w = \frac{-1}{|\vec{\pi}|^2} \big(|\vec{\pi}| \tau_{w,i} - \vec{\pi}' \vec{\tau}_w\big) \\
\Rightarrow \quad \frac{\partial \ln\big(N \vec{\theta}' \vec{\tau}_w\big)}{\partial \pi_i} &= \frac{N}{N \vec{\theta}' \vec{\tau}_w} \frac{\partial \vec{\theta}' \vec{\tau}_w}{\partial \pi_i} = \frac{1}{\vec{\theta}' \vec{\tau}_w} \times \frac{-1}{|\vec{\pi}|^2} \big(|\vec{\pi}| \tau_{w,i} - \vec{\pi}' \vec{\tau}_w\big) \\
&= \frac{1}{\vec{\theta}' \vec{\tau}_w} \times \frac{-1}{|\vec{\pi}|} \big(\tau_{w,i} - \vec{\theta}' \vec{\tau}_w\big) = \frac{1}{|\vec{\pi}|} \big(1 - \tau_{w,i} / \vec{\theta}' \vec{\tau}_w\big) \\
&= \frac{1}{|\vec{\pi}|} - \frac{\tau_{w,i}}{\vec{\pi}' \vec{\tau}_w}.
\end{aligned} \tag{43}$$

similarly,

$$\begin{aligned}
\frac{\partial^2 \ln\big(N \vec{\theta}' \vec{\tau}_w\big)}{\partial \pi_i \partial \pi_i} &= \frac{\partial}{\partial \pi_i} \Big(\frac{1}{|\vec{\pi}|} - \frac{\tau_{w,i}}{\vec{\pi}' \vec{\tau}_w}\Big) = -\frac{1}{|\vec{\pi}|^2} + \frac{\tau_{w,i}^2}{(\vec{\pi}' \vec{\tau}_w)^2} \\
\frac{\partial^2 \ln\big(N \vec{\theta}' \vec{\tau}_w\big)}{\partial \pi_i \partial \pi_j} &= \frac{\partial}{\partial \pi_j} \Big(\frac{1}{|\vec{\pi}|} - \frac{\tau_{w,i}}{\vec{\pi}' \vec{\tau}_w}\Big) = -\frac{1}{|\vec{\pi}|^2} + \frac{\tau_{w,i} \tau_{w,j}}{(\vec{\pi}' \vec{\tau}_w)^2}.
\end{aligned} \tag{44}$$

Therefore, here is the matrix form of the gradient and Hessian of the Poisson log-likelihood with respect

to θ:

$$\begin{aligned}
\nabla_\pi \ln\left(N\vec{\theta}'\vec{\tau}_w\right) &= \frac{1}{|\vec{\pi}|}\mathbf{1} - \frac{1}{\vec{\pi}'\vec{\tau}_w}\vec{\tau}_w \\
H_\pi \ln\left(N\vec{\theta}'\vec{\tau}_w\right) &= \frac{1}{(\vec{\pi}'\vec{\tau}_w)^2}\begin{vmatrix} \tau_{w,1}^2 & \tau_{w,1}\tau_{w,2} & \cdots & \tau_{w,1}\tau_{w,K} \\ \tau_{w,1}\tau_{w,2} & \tau_{w,2}^2 & \cdots & \tau_{w,2}\tau_{w,K} \\ \vdots & \vdots & \ddots & \vdots \\ \tau_{w,1}\tau_{w,K} & \tau_{w,K-1}\tau_{w,K} & \cdots & \tau_{w,K}^2 \end{vmatrix} - \frac{1}{|\vec{\pi}|^2}\begin{vmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 1 \end{vmatrix} \\
&= \frac{1}{(\vec{\pi}'\vec{\tau}_w)^2}\vec{\tau}_w \times \vec{\tau'}_w - \frac{1}{|\vec{\pi}|^2}\mathbf{1}\times\mathbf{1}', \qquad (45)
\end{aligned}$$

where $\vec{\tau}_w \times \vec{\tau'}_w$ represents an outer product of the two vectors.

The gradient and Hessian with respect to τ is:

$$\begin{aligned}
\frac{\partial \ln\left(N\vec{\theta}'\vec{\tau}_w\right)}{\partial \tau_i} &= \frac{\theta_i}{\vec{\theta}'\vec{\tau}_w} \\
\frac{\partial^2 \ln\left(N\vec{\theta}'\vec{\tau}_w\right)}{\partial \tau_i \partial \tau_i} &= \frac{\partial}{\partial \tau_i}\Big(\frac{\theta_i}{\vec{\theta}'\vec{\tau}_w}\Big) = -\frac{\theta_i^2}{(\vec{\theta}'\vec{\tau}_w)^2} \\
\frac{\partial^2 \ln\left(N\vec{\theta}'\vec{\tau}_w\right)}{\partial \tau_i \partial \tau_j} &= \frac{\partial}{\partial \tau_j}\Big(\frac{\theta_i}{\vec{\theta}'\vec{\tau}_w}\Big) = -\frac{\theta_i\theta_j}{(\vec{\theta}'\vec{\tau}_w)^2}. \qquad (46)
\end{aligned}$$

In matrix form:

$$\begin{aligned}
\nabla_\tau \ln\left(N\vec{\theta}'\vec{\tau}_w\right) &= \frac{1}{\vec{\pi}'\vec{\tau}_w}\vec{\theta} \\
H_\tau \ln\left(N\vec{\theta}'\vec{\tau}_w\right) &= -\frac{1}{(\vec{\pi}'\vec{\tau}_w)^2}\vec{\theta}\times\vec{\theta}'. \qquad (47)
\end{aligned}$$

Assume that θ and τ are independent, i.e., $cov(\tau,\theta) = 0$, we have the following approximation of $\ln\left(N\vec{\theta}'\vec{\tau}_w\right)$:

$$\begin{aligned}
& \ln\left(N\vec{\theta}'\vec{\tau}_w\right) \\
\approx\ & \ln\left(N\hat{\theta}'\hat{\tau}_w\right) + \nabla_\tau[\hat{\tau}_w](\vec{\tau}_w - \hat{\tau}_w) + \nabla_\pi[\hat{\pi}](\vec{\pi} - \hat{\pi}) \\
& + \frac{1}{2}(\vec{\tau}_w - \hat{\tau}_w)' H_\tau[\hat{\tau}_w](\vec{\tau}_w - \hat{\tau}_w) + \frac{1}{2}(\vec{\pi} - \hat{\pi})' H_\pi[\hat{\pi}](\vec{\pi} - \hat{\pi})
\end{aligned} \qquad (48)$$

where $\hat{\pi}$ and $\hat{\tau}_w$ are some estimates of the true π and τ. Note that under this approximation, computing the expectation $\Big\langle \ln\left(N\vec{\theta}'\vec{\tau}_w\right)\Big\rangle$ under $q(\vec{\theta})$ and $q(\vec{\tau}_w)$ can be done (approximately) in close-from by using the variational marginals of $\vec{\gamma}$ and $\vec{\zeta}_w$.

Now, the variational marginal for $\{\vec{\zeta}_w^{(t)}\}$ (the inverse logistic-transformation of $\{\vec{\tau}_w^{(t)}\}$, also known as the *natural parameter* of the multinomial) and $\{\vec{\gamma}_d^{(t)}\}$ (the inverse logistic-transformation of $\{\vec{\theta}_d^{(t)}\}$) can be derived from the following GMF approximations to the marginal posterior of $\{\vec{\zeta}_w^{(t)}\}$ and $\{\vec{\tau}_w^{(t)}\}$, respectively.

$$\begin{aligned}
& p(\vec{\zeta}_w^{(1)}, \ldots, \vec{\zeta}_w^{(T)} | \langle S_{\boldsymbol{\theta}} \rangle_{q_\theta}) \\
\propto\ & \Big(\frac{1}{(2\pi)^{K/2}|\Psi_w|^{1/2}}\Big)^T \exp\Big\{ -\frac{1}{2}\vec{\zeta}'^{(1)}_w \Psi_w^{-1}\vec{\zeta}_w^{(1)} - \frac{1}{2}\sum_{t=2}^{T}(\vec{\zeta}_w^{(t)} - \mathbf{B}_w\vec{\zeta}_w^{(t-1)})'\Psi_w^{-1}(\vec{\zeta}_w^{(t)} - \mathbf{B}_w\vec{\zeta}_w^{(t-1)})\Big\} \\
& \times \prod_{t=1}^{T} \exp\Big\{ n_w^{(t)}\Big\langle \log(\omega_t \sum_k \theta_k \tau_{w,k})\Big\rangle_{q_\theta} - \omega_t \sum_k \langle\theta_k\rangle\tau_{w,k} - \Gamma(n_w^{(t)} + 1)\Big\}; \qquad (49)
\end{aligned}$$

and

$$
\begin{aligned}
& q_\gamma(\{\vec{\gamma}_d^{(t)}\}) \\
= \; & p(\{\vec{\gamma}_d^{(t)}\}|\langle S_\mu\rangle_{q_\mu}, \langle S_\tau\rangle_{q_\tau}) \\
= \; & \prod_{t=1}^{T}\prod_{d=1}^{N_t} p(\vec{\gamma}_d^{(t)}|\langle S_{\mu_t}\rangle_{q_\mu}, \langle S_{\mathbf{z}_{t,d}}\rangle_{q_z}) \\
\propto \; & \prod_{t=1}^{T}\prod_{d=1}^{N_t} \frac{1}{(2\pi)^{K-1/2}|\Sigma_t|^{1/2}} \exp\{-\frac{1}{2}(\vec{\gamma}_d^{(t)} - \langle\vec{\mu}_t\rangle)'\Sigma_t^{-1}(\vec{\gamma}_d^{(t)} - \langle\vec{\mu}_t\rangle)\} \\
& \times \exp\Big\{ n_w^{(t)} \Big\langle \log(\omega_t \sum_k \theta_k \tau_{w,k}) \Big\rangle_{q_\tau} - \omega_t \sum_k \theta_k \langle\tau_{w,k}\rangle - \Gamma(n_w^{(t)}+1) \Big\}. \qquad (50)
\end{aligned}
$$

Note that by introducing the taylor approximation to $\ln\big(N\vec{\theta}'\vec{\tau}_w\big)$, and using the laws for computing the means and covariance of $\vec{\tau}_w$ and $\vec{\pi}$ under multivariate log-normal distribution (i.e., Eqs. (38)-(40), the expectation terms in the above equations can be approximately solved. Using similar techniques employed in §2.1, they can be approximated by standard SSMs with Gaussian emissions.

4 Parameter Estimation

As mentioned before, we can use a variational EM scheme to estimate the parameters in our model, which are essentially the SSM parameters. Operationally, VEM is no different from a standard EM for SSM—we have observation sequence $\tilde{\gamma}_t = \frac{1}{N_t}\sum_d \langle\vec{\gamma}_d^{(t)}\rangle$ for the topic mixing SSM (as defined by $q_\mu(\cdot)$), and observation sequence $\vec{u}_k^{(k)}$ for each of the K topic representation (i.e., word-frequency) SSMs (as defined by $q_{\eta_k}(\cdot)$); and we can use the standard learning learning rules for SSM parameter estimation.

In the E step, we use Eqs.(13)-(14), and Eqs.(31)-(31) to estimate the expected sufficient statistics of the latent states; In the M step, we update the parameters Φ, $\mathbf{A}$, Σ_t of the topic mixing SSM, and Ψ_k, $\mathbf{B}_k$ of the topic representation SSM. Following [Ghahramani and Hinton, 1996; Ghahramani and Hinton, 1998], which gave detailed derivations for the MLE standard SSM and switching SSM, bellow we give the relevant formulae of MLE for the parameters in our model. Each of the estimate can be derived by taken the corresponding partial derivative of the expected loglikelohood under the our variational approximation to the true posterior, setting to zero and solving.

- Topic mixing dynamic matrix:

$$
\mathbf{A}^* = \big(\sum_2^T V_{t|t-1}(\mu)\big)\big(\sum_2^T V_{t-1}(\mu)^{-1}\big), \qquad (51)
$$

  where $V_t(\mu) = E[\vec{\mu}_t\vec{\mu}_t'|\mathbf{Y}_1,\ldots,\mathbf{Y}_t]$ (not to be confused with the center moment P_t in Eq. (**??**)), and $V_{t|t-1}(\mu) = E[\vec{\mu}_t\vec{\mu}_{t-1}'|\mathbf{Y}_1,\ldots,\mathbf{Y}_t]$. From RTS smoother, it is easy to see:

$$
\begin{aligned}
V_t(\mu) &= P_{t|T} + \hat{\mu}_{t|T}\hat{\mu}_{t|T}' \\
V_{t|t-1}(\mu) &= P_{t\leftarrow t-1|T} + \hat{\mu}_{t|T}\hat{\mu}_{t-1|T}' \qquad (52)
\end{aligned}
$$

  where the posterior estimate of the self and cross-time covariance matrices $P_{t|T}$ and $P_{t,t-1|T}$ can be computed from Eqs. (13)-(15).

- Noise covariance matrix for topic mixing state:

$$
\mathbf{\Phi}^* = \frac{1}{T}\big(\sum_2^T V_t(\mu) - \mathbf{A}^*\sum_2^T V_{t|t-1}(\mu)\big). \qquad (53)
$$

- Output covariance matrix for topic mixing vectors:

$$\Sigma_t^* = \frac{1}{N_t}\sum_{d=1}^{N_t}(\langle\vec{\gamma}_d^{(t)}\rangle - \hat{\mu}_{t|T})(\langle\vec{\gamma}_d^{(t)}\rangle - \hat{\mu}_{t|T})'. \quad (54)$$

- Topic representation (i.e., topic-specific word frequency vector) dynamic matrix:

$$\mathbf{B}_k^* = \big(\sum_2^T V_{k,t|t-1}(\eta)\big)\big(\sum_2^T V_{k,t-1}(\eta)^{-1}\big), \quad (55)$$

where

$$\begin{aligned} V_{k,t}(\eta) &= P_{k,t|T} + \hat{\eta}_{k,t|T}\hat{\eta}'_{k,t|T} \\ V_{k,t|t-1}(\eta) &= P_{k,t\leftarrow t-1|T} + \hat{\eta}_{k,t|T}\hat{\eta}'_{k,t-1|T} \end{aligned} \quad (56)$$

- Noise covariance matrix for topic representation vector:

$$\Psi_k^* = \frac{1}{T}\big(\sum_2^T V_{k,t}(\eta) - \mathbf{B}^*\sum_2^T V_{k,t|t-1}(\eta)\big). \quad (57)$$

We set the initial vectors ι_k and ν to be zero vectors instead of estimating them from the data. Finally, note that what are given above are the most general form of transition and correlations of topics and words. In practice, to avoid over-parameterization, we can choose to reduce, for example, the transition matrices $\mathbf{B}_k$'s and the covariance matrices Ψ_k's of the topic representations to be sparse or diagonal matrix to model only random walk effects.

5 Conclusion

In this report I introduce topic evolution models for longitudinal epochs of word documents. The models employ marginally dependent latent state-space models for evolving topic proportion distributions and topic-specific word distributions; and both a logistic-normal-multinomial and a logistic-normal-Poisson model for document likelihood. These models allow posterior inference of latent topic themes over time, and topical clustering of longitudinal document epochs. I derive a variational inference algorithm for non-conjugate generalized linear models based on truncated Taylor approximation, and I also outline formulae for parameter estimation based on variational EM principle. In the current model, I assume all topics coexist over time, and no new topic will emerge over time. In a companion report, I present a birth-death process model that capture more complicated and realistic behaviors of topic evolution, such as aggregation, emergence, extinction, and split of topics over time.

6 BIBLIOGRAPHY

References

[Blei and Lafferty, 2006] D. Blei and J. Lafferty. Correlated topic models. In *Advances in Neural Information Processing Systems 18*, 2006.

[Blei *et al.*, 2003] D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3, 2003.

[Ghahramani and Hinton, 1996] Z. Ghahramani and G. E. Hinton. Parameter estimation for linear dynamical systems. University of Toronto Technical Report CRG-TR-96-2, 1996.

[Ghahramani and Hinton, 1998] Z. Ghahramani and G. E. Hinton. Variational learning for switching state-space models. *Neural Computation*, 12(4):963–296, 1998.

[Griffiths and Steyvers, 2004] T. Griffiths and M. Steyvers. Finding scientific topics. *Proc Natl Acad Sci U S A*, 101 (Suppl 1):5228–5235, 2004.

[Hofmann, 1999] Thomas Hofmann. Probabilistic latent semantic indexing. In *Proc. of the 22nd Intl. ACM SIGIR conference*, pages 50–57, 1999.

[Jordan *et al.*, 1999] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul. An introduction to variational methods for graphical models. In M. I. Jordan, editor, *Learning in Graphical Models*, pages 105–161. Kluwer Academic Publishers, 1999.

[Kleiber and Kotz, 2003] C. Kleiber and S. Kotz. *Statistical Size Distributions in Economics and Actuarial Sciences*. Wiley InterScience, 2003.

[Steyvers *et al.*, 2004] M. Steyvers, P. Smyth, M. Rosen-Zvi, and T. Griffiths. Probabilistic author-topic models for information discovery. In *The Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2004.

[Xing *et al.*, 2003] E. P. Xing, M. I. Jordan, and S. Russell. A generalized mean field algorithm for variational inference in exponential families. In *Proceedings of the 19th Annual Conference on Uncertainty in AI*, 2003.