Probabilistic Dynamic Non-negative Group Factor Model for Multi-source Text Mining

Chien Lu
Tampere University
Tampere, Finland
chien.lu@tuni.fi

Jaakko Peltonen
Tampere University
Tampere, Finland
jaakko.peltonen@tuni.fi

Jyrki Nummenmaa
Tampere University
Tampere, Finland
jyrki.nummenmaa@tuni.fi

Kalervo Järvelin
Tampere University
Tampere, Finland
kalervo.jarvelin@tuni.fi

ABSTRACT

Nonnegative matrix factorization (NMF) is a popular approach to model data, however, most models are unable to flexibly take into account multiple matrices across sources and time or apply only to integer-valued data. We introduce a probabilistic, Gaussian Process based, more inclusive NMF-based model which jointly analyzes nonnegative data such as text data word content from multiple sources in a temporal dynamic manner. The model collectively models observed matrix data, source-wise latent variables and their dependencies and temporal evolution with a full-fledged hierarchical approach including flexible nonparametric temporal dynamics. Experiments on simulated data and real data show the model outperforms comparable models. A case study on social media and news demonstrates the model discovers semantically meaningful topical factors and their evolution.

CCS CONCEPTS

• **Computing methodologies** → **Non-negative matrix factorization**; **Natural language processing**.

KEYWORDS

Nonnegative Matrix Factorization; Gaussian Process; Multiple Sources

ACM Reference Format:
Chien Lu, Jaakko Peltonen, Jyrki Nummenmaa, and Kalervo Järvelin. 2020. Probabilistic Dynamic Non-negative Group Factor Model for Multi-source Text Mining. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management (CIKM '20), October 19–23, 2020, Virtual Event, Ireland.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3340531.3411956

1 INTRODUCTION

Factor analysis is a popular approach to extract latent components describing variable relationships within data sets, and non-negative matrix factorization (NMF) [14, 18] in particular has become a prominent solution for data sets in matrix form, applicable in numerous settings where measurements and their latent factors are expected to be nonnegative, such as in several text analytics and bioinformatics settings. However, much factor analysis work has focused on factorization of individual matrices.

Analyzing data from multiple sources has attracted increasing attention in the machine learning community [11]. For instance text data such as online discussions or news articles from a single source may not provide a sufficiently thorough understanding of the underlying phenomena. Analyzing the factors underlying data matrices from multiple sources jointly is a promising approach to infer improved models that better represent the phenomena, have better predictive performance, and allow discovery of relations and interactions between different sources.

In addition to multiple sources, modeling temporal variation of the phenomena from data collected over time is also often desired. Models including flexible generative approaches such as Gaussian Processes (GPs) [19] and their extensions have been proposed to model temporal dynamics. Temporal analysis should ideally reveal both variation of the underlying factor prevalences and variation of the factors' contents over time.

Although NMF has been widely accepted as a classical approach when analyzing text data, to our knowledge there are only few probabilistic matrix-factorization models that address the multiple sources aspect or the temporal aspect, and none that address both.

We introduce a novel probabilistic non-negative matrix factorization model, suitable for analysis of multiple data matrices across sources and time, applicable to any series of nonnegative real-valued matrices. The proposed method models the matrix data, the underlying source-wise parameters of factor prevalence and content, and inter-source parameters of factor relationships across sources. Temporal dynamics of topic prevalence, topic content and source-source interaction are modeled with a flexible (Hierarchical) Gaussian Process Latent Variable Model (GPLVM) [12, 13, 15] based approach. Modeling temporal dynamics with GP priors can model smooth temporal changes without fixing a rigid parametric form [8]. We carry out variational inference for the model.

The model has superior performance in experiments in predicting held-out data. We demonstrate the model both on simulated data and a case study on news and social media. We use a text analytics case for simplicity of illustrating results, but the model

CIKM '20, October 19–23, 2020, Virtual Event, Ireland

ACM ISBN 978-1-4503-6859-9/20/10.
https://doi.org/10.1145/3340531.3411956

applies to all similar domains with non-negative data and is not restricted e.g. to integer-valued count data, unlike some text analysis solutions.

The rest of the paper is organized as follows. Next, Section 2 describes preliminaries and related work. Sections 3 and 4 present the basic structure of the proposed model and its variational inference. Sections 5 and 6 describe the experiments with simulated and real data. Section 7 provides conclusions and discussion.

2 RELATED BACKGROUND

Non-negative Matrix Factorization. NMF is a widely used data analysis approach in domains such as bioinformatics [23], image processing [14], and text mining [16]. In short, NMF finds an approximate decomposition of a $N \times D$ matrix $\mathbf{X}$ containing only nonnegative element values into a product of two lower-rank matrices $\mathbf{X} \approx \mathbf{Z}\mathbf{W}^\top$ where $\mathbf{Z}$ is a $N \times K$ matrix, $\mathbf{W}$ is a $D \times K$ matrix, and K is the number of latent factors, where $\mathbf{Z}$ and $\mathbf{W}$ also contain only nonnegative values. For example in text analytics $\mathbf{X}$ may be a term-document matrix of N terms and D documents, $\mathbf{W}$ can be interpreted as a topic loading matrix of K topics of D documents, so that each row $\mathbf{w}_d$ contains the topic loadings for document d, and $\mathbf{Z}$ can be interpreted as a topic content matrix of N terms across the K topics, each column $\mathbf{z}_{\cdot k}$ is a discrete distribution over terms for topic k. Different NMF variants use different divergences to measure difference between $\mathbf{X}$ and its approximation $\mathbf{Z}\mathbf{W}^\top$ and regularize $\mathbf{Z}$ and $\mathbf{W}$ by different penalties. We adopt the form where the model is specified by a particular noise model between $\mathbf{Z}\mathbf{W}^\top$ and the observed $\mathbf{X}$ and particular priors for $\mathbf{Z}$ and $\mathbf{W}$; the latter incorporate a hierarchical model for cross-sources and temporal dynamics.

Related Work. Some NMF based methods have been proposed to model temporal dynamics [21, 24] of text data or data from multiple sources [4, 7, 22]; most of these are not hierarchical approaches or deal with only one of the two aspects (multi-source or temporal). For example, in the Joint Past-Present Decomposition Model (JPP; [24]) at each time slice the term-document matrix is explained by both current topics and topics at the previous time slice.

A noteworthy example of Matrix factorization approaches is Bayesian Group Factor Analysis (GFA) [10, 25] which analyzes data from multiple sources (groups). GFA considers the joint data set $\mathbf{Y} = \{\mathbf{X}_1, \ldots, \mathbf{X}_M\}$ of matrices $\mathbf{X}_1 \in \mathbb{R}^{N \times D_1}, \ldots, \mathbf{X}_M \in \mathbb{R}^{N \times D_M}$. GFA factorizes $\mathbf{Y}$ into matrices $\mathbf{Z}$ and $\mathbf{W}$ as $\mathbf{Y} \approx \mathbf{Z}\mathbf{W}^\top$ where $\mathbf{W} = [\mathbf{W}_1^\top \ldots \mathbf{W}_M^\top]^\top$, $\mathbf{W}_m \in \mathbb{R}^{D_m \times K}$ and each element $w_{m,k}(d)$ in $\mathbf{W}_m$ is normally distributed with zero mean and a group-wise precision parameter $\alpha_{m,k}$ as $w_{m,k}(d) \sim N(0, \alpha_{m,k}^{-1})$. The precision parameter $\alpha_{m,k}$ enables GFA to model shared underlying features between groups. However, GFA has no model for temporal dynamics. Moreover, GFA is not designed to model non-negative factorization and hence it can yield negative-valued factors even for nonnegative-valued data, making it unsuitable to be directly applied in cases when factors are required to be nonnegative e.g. for interpretability, such as loadings and contents of topics in text data. We use GFA as a comparison both as is and with a simple correction for nonnegativity.

One similar work [9] tries to model the temporal dynamics but only takes the dynamics of the left-hand side matrix $\boldsymbol{Z}$ into account, the loading matrix $\boldsymbol{W}$ is considered static.

Another group of approaches are the models based on Poisson factor analysis (PFA) [1, 6, 17, 28]. However, since the Poisson distribution only models positive integers, the approaches only model positive-integer-valued matrices but not positive real-valued matrices; the latter occur in many domains including text mining, e.g. real-valued term weighting such as TF-IDF is often crucial for document representation. This paper focuses on methods applicable to positive real-valued matrices.

3 PROPOSED MODEL

We now present the proposed dynamic non-negative Bayesian group factor (DNBGFA) model. For clarity we use text data terminology (documents, terms, topics) but the model is general. DNBGFA considers a temporal sequence of T term-document matrices $\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \ldots, \mathbf{X}^{(T)}$, sharing the same vocabulary of N terms (words). For each time slice t, $\mathbf{X}^{(t)} = [\mathbf{X}_1^{(t)}, \ldots \mathbf{X}_M^{(t)}]$ is a combined matrix of M text sources, each $\mathbf{X}_m^{(t)}$ contains N terms and $D_m^{(t)}$ documents, and the total document count at time t is $D^{(t)} = \sum_m D_m^{(t)}$.

For each time slice t, the task is to approximately factorize the $N \times D^{(t)}$ term-document matrix $\mathbf{X}^{(t)}$ as

$$\mathbf{X}^{(t)} \approx \mathbf{Z}^{(t)} \mathbf{W}^{(t)\top} \tag{1}$$

where $\mathbf{Z}^{(t)}$ is a $N \times K$ matrix which represents the topic content and $\mathbf{W}^{(t)}$, a $D^{(t)} \times K$ matrix, represents the topic prevalence, and both matrices are nonnegative. The setup is illustrated in Figure 1. We infer the factorization as part of a hierarchical generative model for the data.

The graphical plate model representation of the model is shown in Figure 2. We assume a truncated-Gaussian likelihood where each Gaussian is truncated from below at 0 as is appropriate for nonnegative data, so that

$$p(\mathbf{X}^{(t)} | \mathbf{Z}^{(t)}, \mathbf{W}^{(t)}) = \prod_{n,d} N^+ \left(x_{n,d}^{(t)} | \mathbf{z}_n^{(t)\top} \mathbf{w}_d^{(t)}, \sigma^2 \right) \tag{2}$$

where $\mathbf{w}_d^{(t)}$ denotes the dth column of $\mathbf{W}^{(t)}$ representing the topic prevalence in document d, $\mathbf{z}_n^{(t)}$ denotes the nth row of $\mathbf{Z}^{(t)}$ representing the weight of the nth vocabulary word across the topics The σ^2 controls the noisiness of the observations. We set it equal for every document in the following implementations but one can also take the advantage of the flexibility to make a more sophisticated model if needed. For example, a more detailed document-specific variance σ_d^2 representing a source-specific noise parameter can be assigned as $\sigma_d = \sigma_{m(d)}$ where $m(d)$ denotes the group that document d belongs to.

The key idea is to generate the factor matrices in a way that flexibly ties them over time, topics, and sources, without restricting the time dependency to a pre-given form; we generate the dependencies (covariance matrices) as functions of latent variables that are draws from flexible nonparametric time series models, as detailed in Sections 3.1 and 3.2.

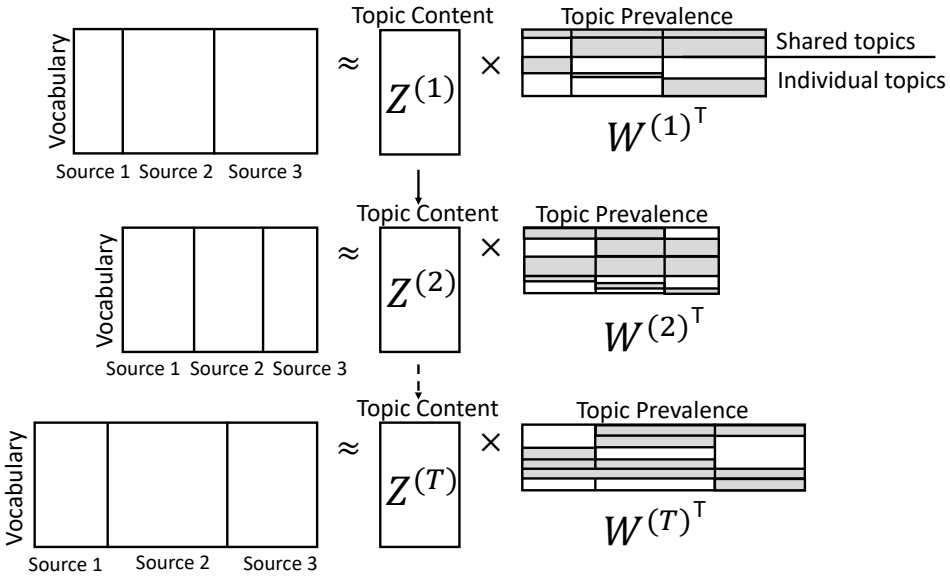

Figure 1: Illustration of the DNBGFA model. A sequence of non-negative matrices $\mathbf{X}^{(1)}, ..., \mathbf{X}^{(T)}$ is factorized into $\mathbf{Z}^{(1)}, ..., \mathbf{Z}^{(T)}$ and $\mathbf{W}^{(1)}, ..., \mathbf{W}^{(T)}$ while modeling temporal dependencies of factors.

3.1 Topic Content

To enforce non-negativity of the topic content matrix, each element $z_{k,n}^{(t)}$ of $\mathbf{Z}^{(t)}$ is parameterized by a softmax transformation

$$z_{k,n}^{(t)} = \frac{\exp(\eta_{k,n}^{(t)})}{\sum_{n'=1}^{N} \exp(\eta_{k,n'}^{(t)})} \tag{3}$$

which ensures the summation of word proportions of each topic $\sum_{n'=1}^{N} z_{k,n'}^{(t)}$ is equal to 1. Note that we will model magnitude of numbers in observed matrices by the loading matrices, hence we can without loss of generality fix the sums as above. Similar transformations are often used in text mining models [3, 20].

GPLVM based model. For each term n, the variable $\boldsymbol{\eta}_n = [\eta_{1,n}^{(1)} \dots \eta_{K,n}^{(1)} \dots \eta_{1,n}^{(T)} \dots \eta_{K,n}^{(T)}]^\top$ controls topic content and the dependencies between its elements represent dependencies across sources and time. We model them in a nonparametric approach by a Gaussian process latent variable model (GPLVM) which lets us model temporal dynamics in a flexible way. In a GPLVM, the parameters of a Gaussian distribution are constructed by a draw from another GP :

$$\left[\eta_{1,n}^{(1)}\eta_{1,n}^{(2)} \dots \eta_{1,n}^{(T)} \dots \eta_{K,n}^{(1)} \dots \eta_{K,n}^{(T)}\right]^\top \sim \mathbf{N}\left(\mathbf{0}, \Sigma_{\boldsymbol{\eta}}\right) \tag{4}$$

where

$$\Sigma_{\boldsymbol{\eta}} = \mathcal{K}_{\boldsymbol{\eta}} + \epsilon_{\boldsymbol{\eta}}\mathbf{I} \tag{5}$$

and $\mathcal{K}_{\boldsymbol{\eta}}$ consists of elements

$$\mathcal{K}_{k,l}^{(\eta)}(t_i, t_j) = k_0^{(\eta)}(t_i, t_j)\delta_{k,l} + k_{k,l}^{(\eta)}(t_i, t_j) \tag{6}$$

where k, l are topic indices and $k_0^{(\eta)}(t_i, t_j)\delta_{k,l}$ is a kernel function which governs the within topic consistency over time, $k_{k,l}^{(\eta)}(t_i, t_j)$ governs the topic-topic interaction, and $\epsilon_{\boldsymbol{\eta}}$ controls noisiness.

The kernel $k_0^{(\eta)}$ can be formed by an arbitrary kernel function of time slices t_i, $i = 1, \dots, T$. In this paper, we use RBF kernel which is defined as

$$rbf_{(\xi,\iota)}(t_i, t_j) = \iota^2 \times e^{\frac{-||t_i - t_j||^2}{\xi^2}} \tag{7}$$

where hyperparameters ξ and ι control dependencies over time. This is a nonparametric time series model for the changing of the term n over time in topic k. As in GPs, no specific functional form is assumed for behavior over time, only that values at similar time points are correlated as described.

$k_{k,l}^{(\eta)}(t_i, t_j)$ controls topic-topic interactions not only within a time-slice but also across two different time-slices. We construct it as

$$k_{k,l}^{(\eta)}(t_i, t_j) = e^{-\lambda_\eta |t_i - t_j|} r_k^{(t_i)} r_l^{(t_j)} \tag{8}$$

which consists of an exponential time decay term $\lambda_\eta \sim Gamma(a, b)$ and products $r_k^{(t_i)} r_l^{(t_j)}$ that control topic-topic interactions of topics k and l across time in a more flexible way: for each topic the vector $\mathbf{r}_k = [r_k^{(t_1)}, \dots, r_k^{(t_T)}]^\top$ is drawn as a realization of a GP as

$$\mathbf{r}_k \sim GP(\mathbf{0}, \Sigma_{\mathbf{r}}) \quad , \quad \Sigma_{\mathbf{r}} = \mathcal{K}_{\mathbf{r}} + \epsilon_r \mathbf{I} \tag{9}$$

where $\mathcal{K}_{\mathbf{r}}$ consists of elements

$$\mathcal{K}^{(r)}(t_i, t_j) = k_0^{(r)}(t_i, t_j) \tag{10}$$

and ϵ_r controls noisiness. Large values of the product $r_k^{(t_i)} r_l^{(t_j)}$ strengthen the dependency $k_{k,l}^{(\eta)}(t_i, t_j)$ between two time slices

whereas small values of the product decrease the dependency, allowing new topic content to emerge.

Like $k_0^{(\eta)}$, the kernel $k_0^{(r)}$ can be computed given time slices t_i, $i = 1, \ldots, T$ with an RBF kernel shown in equation (7). Noisiness variables ϵ_η and ϵ_r could be given priors or be used as hyperparameters, we did the latter for simplicity.

The kernel $k_0^{(\eta)}$ is identical in different topics, hence it acts as a regularizing term controlling word (dis)similarities within topics over time whereas $k_{k,l}^{(\eta)}$ models flexibility of topic-topic interactions.

3.2 Topic Prevalence

Similar to the topic content model, to enforce non-negativity, each $w(d)_{m,k}^{(t)}$ of $\mathbf{W}^{(t)}$ is sampled from a truncated normal distribution with mean 0 and a source-wise variance $e^{\alpha_{m,k}^{(t)}}$:

$$w(d)_{m,k}^{(t)} \sim N^{+}(0, e^{\alpha_{m,k}^{(t)}}) \ . \tag{11}$$

The source-wise latent variables $\alpha_{m,k}^{(t)}$ which control the sparsity of topic in data sources m and time slices t are again a realization of a GPLVM

$$\left[\alpha_{1,k}^{(1)} \ldots \alpha_{1,k}^{(T)}, \ldots, \alpha_{M,k}^{(1)} \ldots \alpha_{M,k}^{(T)}\right]^{\top} \sim \mathbf{N}(\mathbf{0}, \Sigma_{\boldsymbol{\alpha}}) \tag{12}$$

where

$$\Sigma_{\boldsymbol{\alpha}} = \mathcal{K}_{\boldsymbol{\alpha}} + \epsilon_{\boldsymbol{\alpha}}\mathbf{I} \tag{13}$$

and $\mathcal{K}_{\boldsymbol{\alpha}}$ consists of elements

$$\mathcal{K}_{m,n}^{(\alpha)}(t_i, t_j) = k_0^{(\alpha)}(t_i, t_j)\delta_{m,n} + k_{m,n}^{(\alpha)}(t_i, t_j) \tag{14}$$

where $k_0^{(\alpha)}(t_i, t_j)\delta_{m,n}$ is a kernel function governs the within source consistency of topic prevalence over time and $k_{m,n}^{(\alpha)}(t_i, t_j)$ governs the cross-source interactions.

The cross-source interactions $k_{m,n}^{(\alpha)}(t_i, t_j)$ are constructed as

$$k_{m,n}^{(\alpha)}(t_i, t_j) = e^{-\lambda_\alpha |t_i - t_j|} s_m^{(t_i)} s_n^{(t_j)} \tag{15}$$

where ϵ_α controls noisiness and the matrix is otherwise again composed of products of two terms, an exponential time decay term with decay variable $\lambda_\alpha \sim Gamma(c, g)$ and the products $s_m^{(t_i)} s_n^{(t_j)}$ that control correlation sources across time in a flexible manner, by generating for each source m the vector $\mathbf{s}_m = [s_m^{(1)}, \ldots, s_m^{(T)}]^{\top}$ from an independent GP as

$$\mathbf{s}_m \sim \mathbf{N}(\mathbf{0}, \Sigma_{\mathbf{s}}) \ , \ \ \Sigma_{\mathbf{s}} = \mathcal{K}_s + \epsilon_s\mathbf{I} \tag{16}$$

where $\mathcal{K}_{\mathbf{s}}$ consists of elements

$$\mathcal{K}^{(s)}(t_i, t_j) = k_0^{(s)}(t_i, t_j) \tag{17}$$

and ϵ_s controls noisiness.

As before, covariances $k_0^{(\alpha)}$ and $k_0^{(s)}$ are obtained by RBF kernel, whose hyperparameters control time depencency; we used RBF. For the noisiness parameters ϵ_α, and ϵ_s could again be given their own priors but for simplicity we kept them as hyperparameters.

The models of topic content and prevalence in the previous section and this section are highly analogous just like matrices $\mathbf{Z}^{(t)}$ and $\mathbf{W}^{(t)}$ have highly analogous roles. The differences are the different way to enforce non-negativity, and the different role of topics and sources: in the previous section correlations were modeled by GPLVMs for each term across topics and time slices, here correlations are modeled by GPLVMs for each topic across sources and time slices. This establishes a flexible framework for factorization of matrices related across sources and time. The factorizations at each time slice (i.e., the parameter posteriors) are learned based on both the hierarchical prior and the likelihood.

The hierarchical prior lets the model handle cases where at some time slices no documents belonging to a source exist; we test this in an experiment in Section 5. If a source is known to be inactive (not just missing) at some time slices, such as birth/death of sources, it can be specified into the priors e.g. by larger ϵ_α , if such expert knowledge is available.

Algorithm 1 Variational EM Procedure

Require:
$\mathbf{X}^{(1)} \ldots \mathbf{X}^{(T)}$: Observed matrices
K: number of topics
σ_d: Hyper-parameters (likelihood)
$k_0^{(\eta)}, k_0^{(r)}, \epsilon_r, \epsilon_\eta, a, b$: Hyper-parameters (content)
$k_0^{(\alpha)}, k_0^{(s)}, \epsilon_s, \epsilon_\alpha, c, g$: Hyper-parameters (prevalence)
Ensure:
1: **for** iter ← 1 to maxit **do**
2: E-step: update $\boldsymbol{\eta}, \boldsymbol{\alpha}, \mathbf{W}$
3: M-step: update r, s, $\lambda_\eta, \lambda_\alpha$
4: **end for**
5: **return** $\boldsymbol{\eta}, \boldsymbol{\alpha}, \mathbf{W}, \mathbf{r}, \mathbf{s}, \lambda_\alpha$ =0

4 VARIATIONAL INFERENCE

To deliver time-efficient inference, we derive variational inference algorithms. Approaches such as Gibbs sampling are possible, here we focus on the variational approach. The inference constructs a variational posterior distribution q for each parameter of interest; update rules for parameters of the q distributions are given below. We update the parameters in an EM manner, as shown in Algorithm 1. Inference algorithms are further described.

4.1 Topic Content Variables $\boldsymbol{\eta}$ and Topic Sparsity Sariables $\boldsymbol{\alpha}$

A Laplace's method based inference [26] is used. The variational distribution $q(\boldsymbol{\eta}_n^{(t)}) = \mathbf{N}(\boldsymbol{\eta}_n^{(t)} | \mathbf{m}_{\boldsymbol{\eta}_n^{(t)}}, -\nabla^2 f(\mathbf{m}_{\boldsymbol{\eta}_n^{(t)}})^{-1})$ where the mean $\mathbf{m}_{\boldsymbol{\eta}_n^{(t)}}$ is set to the value of the MAP solution which maximizes the joint log-probability f defined as

$$f(\boldsymbol{\eta}_n) = \sum_{(t)} E_{q(\mathbf{w})}\left[\log p(\mathbf{x}_n^{(t)} | \mathbf{z}_n^{(t)}, \mathbf{w}^{(t)})\right] + E_{q(\mathbf{r})}\left[\log p(\boldsymbol{\eta}_n^{(t)} | \mathbf{r})\right] \ . \tag{18}$$

In this work, we obtained the $\mathbf{m}_{\boldsymbol{\eta}_n^{(t)}} = \arg\max_{\boldsymbol{\eta}_n^{(t)}} f(\boldsymbol{\eta}_n^{(t)})$ using an optimizer called simulated annealing (SANN) [2]. The covariance matrix $\nabla^2 f(\mathbf{m}_{\boldsymbol{\eta}_n^{(t)}})$ is the Hessian matrix of f evaluated at the point $\mathbf{m}_{\boldsymbol{\eta}_n^{(t)}}$.

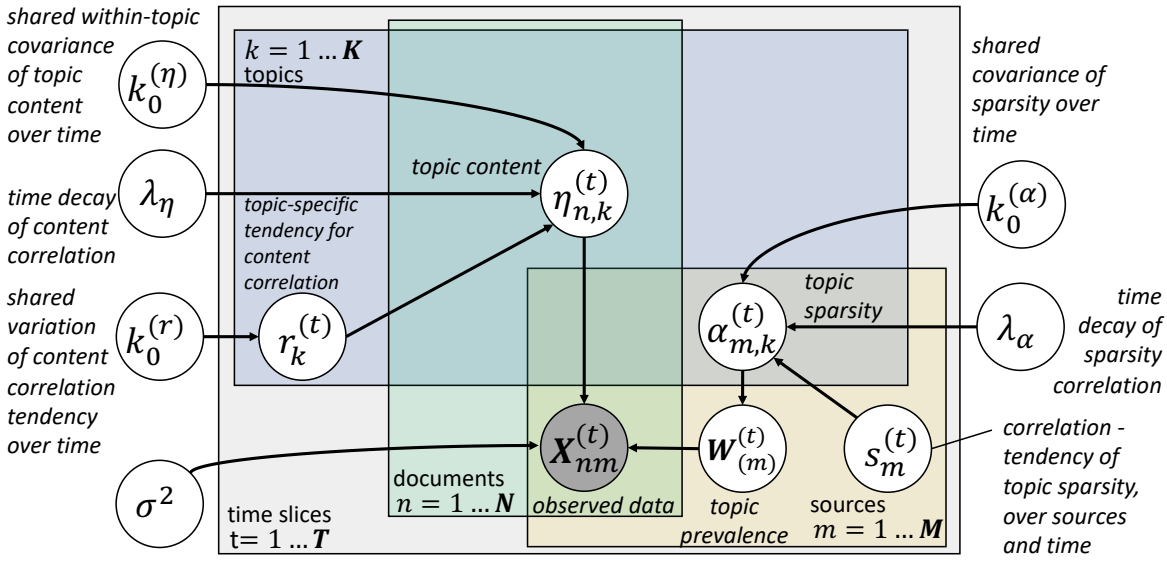

Figure 2: Graphical representation of the DNBGFA model. Noisiness parameters ϵ_η, ϵ_r, ϵ_s, ϵ_α not shown for clarity.

Inference of $\boldsymbol{\alpha}$ is similar, the variational distribution $q(\boldsymbol{\alpha}_k)$ is $\mathrm{N}(\boldsymbol{\alpha}_k|\mathbf{m}_{\boldsymbol{\alpha}_k}, -\nabla^2 f(\mathbf{m}_{\boldsymbol{\alpha}_k})^{-1})$ and the corresponding objective function $f(\boldsymbol{\alpha}_k)$ is

$$f(\boldsymbol{\alpha}_k) = \sum_{(t)} E_{q\left(\mathrm{w}_k^{(t)}\right)}\left[\log p(\mathbf{w}_k^{(t)}|\boldsymbol{\alpha}_k)\right] + E_{q(\mathrm{s})}\left[\log p(\boldsymbol{\alpha}_n^{(t)}|\mathbf{s})\right] \quad (19)$$

4.2 Topic Content Correlation r and Sparsity Correlation Tendencies s

To carry out the posterior inference of the variables $\mathbf{r}$ and $\mathbf{s}$ describing the topic-specific content and source-wise sparsity correlation tendencies, we adapted a recently developed framework proposed by Damianou et al. [5] which is able to capture the complexity of the interactions between latent variables. In the framework, auxiliary variables $\mathbf{u}^{(r)}$ and $\mathbf{r}_u$ are induced. The joint probability related to $\mathbf{r}$ is then expanded, written as

$$\prod_{n=1}^{N} p(\boldsymbol{\eta}_n|\mathbf{u}_n^{(r)}, \mathbf{r}, \mathbf{r}_u) p(\mathbf{u}_n^{(r)}|\mathbf{r}_u) p(\mathbf{r}) \quad (20)$$

where $p(\boldsymbol{\eta}_n|\mathbf{u}_n^{(r)}, \mathbf{r}, \mathbf{r}_u) = \mathrm{N}(\boldsymbol{\eta}_n|\mathbf{a}_\mathbf{n}, \Sigma_\eta^*)$ with $\mathbf{a}_\mathbf{n} = \mathcal{K}_{\eta u}{\mathcal{K}_u^{(r)}}^{-1}\mathbf{u}_n$ and $\Sigma_\eta^* = \Sigma_\eta - \mathcal{K}_{\eta u}{\mathcal{K}_u^{(r)}}^{-1}\mathcal{K}_{u\eta}$. The pseudo-inputs $\mathbf{r}_u = [r_u^{(1)}, \dots, r_u^{(T)}]^\top$ are the constructing variables of $\mathbf{u}^{(r)}$, that is,

$$p(\mathbf{u}_n^{(r)}|\mathbf{r}_u) = \mathrm{N}(\mathbf{u}_n^{(r)}|\mathbf{0}, \mathcal{K}_u^{(r)}), \quad (21)$$

where $\mathcal{K}_u^{(r)}$ consists of elements

$$\mathcal{K}_u^{(r)}(t_i, t_j) = e^{-\lambda_\eta|t_i - t_j|} r_u^{(t_i)} r_u^{(t_j)} + k_0^{(\eta)}(t_i, t_j)\,. \quad (22)$$

The posterior is then approximated with

$$\prod_{n=1}^{N} p(\boldsymbol{\eta}_n|\mathbf{u}_n^{(r)}, \mathbf{r}', \mathbf{r}_u) q(\mathbf{u}_n^{(r)}) q(\mathbf{r}') \quad (23)$$

where $\mathbf{r}' = [r_1^{(1)} \dots r_k^{(1)} \dots r_1^{(T)} \dots r_k^{(T)}]^\top$; $q(\mathbf{r}')$ is a Gaussian distribution $q(\mathbf{r}') = \mathrm{N}(\mathbf{r}'|\mathbf{m}_{\mathbf{r}'}, \mathrm{S}_{\mathbf{r}'})$ where the variational mean vector $\mathbf{m}_{\mathbf{r}'}$ and covariance matrix $\mathrm{S}_{\mathbf{r}'}$ are obtained via maximizing an objective function $\hat{\mathcal{F}}(\mathbf{r}') - KL\left(q(\mathbf{r}')||p(\mathbf{r}')\right)$ which is a Jensen's lower bound of the marginal likelihood, with respect to $\mathbf{m}_{\mathbf{r}'}$ and $\mathrm{S}_{\mathbf{r}'}$ together with $\mathbf{r}_u$.

We have

$$\hat{\mathcal{F}}(\mathbf{r}') = \frac{\sum_{n=1}^{N} \boldsymbol{\eta}_n^\top \mathbf{W}^{(r)} \boldsymbol{\eta}_n}{-2} + N \log\left(\frac{\epsilon_\eta^{-(K\times T)} |\mathcal{K}_u^{(r)}|^{\frac{1}{2}}}{(2\pi)^{\frac{(K\times T)}{2}} |\epsilon_\eta^{-2}\Psi_2^{(r)} + \mathcal{K}_u^{(r)}|^{\frac{1}{2}}}\right) + \frac{tr\left({\mathcal{K}_u^{(r)}}^{-1}\Psi_2^{(r)}\right) - \psi_0^{(r)}}{2\epsilon_\eta^2/N} \quad (24)$$

where the matrices involved are computed as

$$\mathbf{W}^{(r)} = \epsilon_\eta^{-2}\mathbf{I}_{(K\times T)} - \epsilon_\eta^{-4}\Psi_1^{(r)}\left(\epsilon_\eta^{-2}\Psi_2^{(r)} + \mathcal{K}_u^{(r)}\right)^{-1}{\Psi_1^{(r)}}^\top, \quad (25)$$

$$\psi_0^{(r)} = \mathbf{m}_{\mathbf{r}'}^\top \mathbf{m}_{\mathbf{r}'} + tr(\mathrm{S}_{\mathbf{r}'}), \quad (26)$$

$$\Psi_1^{(r)} = \mathbf{r}_u \mathbf{m}_{\mathbf{r}'}^\top \circ \mathbf{D}^{(\eta u)}, \quad (27)$$

$$\Psi_2^{(r)} = \mathbf{D}^{(u\eta)} \circ \mathbf{r}_u \left(\mathbf{m}_{\mathbf{r}'}\mathbf{m}_{\mathbf{r}'}^\top + Tr(\mathrm{S}_{\mathbf{r}'})\right)\mathbf{r}_u^\top \circ \mathbf{D}^{(\eta u)}, \quad (28)$$

$$\mathrm{S}_{\mathbf{r}'} = \left({\Sigma_{\mathbf{r}'}}^{-1} + diag(\xi_{\mathbf{r}'})\right)^{-1}, \quad (29)$$

where

$$\mathbf{D}^{(\eta u)} = \mathbf{1}_K \otimes \begin{bmatrix} 1 & \dots & e^{-|1-T|\lambda_\eta} \\ & \ddots & \\ e^{-|T-1|\lambda_\eta} & \dots & 1 \end{bmatrix} \tag{30}$$

and $\mathbf{D}^{(u\eta)} = \mathbf{D}^{(\eta u)\top}$. Note that $\circ$ denotes Hadamard product and $\otimes$ denotes Kronecker product.

For the parameters $\mathbf{s}$ which define the tendency of the topics' sparsity to correlate, the inference is done in a similar manner by imposing $\mathbf{u}^{(s)}$ and $\mathbf{s}_u$. The variational distribution $q(\mathbf{s}')$ related parameters $\{\mathbf{m}_{\mathbf{s}'}, \boldsymbol{\xi}_{\mathbf{s}'}, \mathbf{s}_u\}$ are obtained via optimizing the objective function $\hat{\mathcal{F}}(\mathbf{s}') - KL\left(q(\mathbf{s}')||p(\mathbf{s}')\right)$. The computation of $\hat{\mathcal{F}}(\mathbf{s}')$ is similar to the computation of (r') via replacing corresponding variables.

4.3 Time Decay Parameters λ_η and λ_α

Here we obtain the point estimates of λ_η and λ_α by optimizing the following objective functions:

$$f(\lambda_\eta) = E_{q(\boldsymbol{\eta})q(\mathbf{r})}\left[\sum_n \log p(\boldsymbol{\eta}_n|\mathbf{r}, \lambda_\eta)\right] + \log p(\lambda_\eta|a, b) \tag{31}$$

and

$$f(\lambda_\alpha) = E_{q(\boldsymbol{\alpha})q(\mathbf{s})}\left[\sum_k \log p(\boldsymbol{\alpha}_k|\mathbf{s}, \lambda_\alpha)\right] + \log p(\lambda_\alpha|c, g) \tag{32}$$

which can be done by standard optimizers, here we again use the SANN optimizer.

4.4 Topic Prevalence W

The truncated normal distribution preserves the Gaussian-Gaussian conjugacy, therefore, the variational distribution can can be obtained analytically:

$$q\left(\mathbf{w}_d^{(t)}\right) = \mathbf{N}^+(\mathbf{w}_d^{(t)}|\mathbf{m}_{\mathbf{w}_d}, \sigma^2 \mathbf{S}_{\mathbf{w}_d}) \tag{33}$$

where we have

$$\mathbf{m}_{\mathbf{w}_d} = \mathbf{S}_{\mathbf{w}_d} E_q[\mathbf{Z}^{(t)\top}]\mathbf{x}_d^{(t)} \tag{34}$$

and

$$\mathbf{S}_{\mathbf{w}_d} = \left(E_q\left[\mathbf{Z}^{(t)\top}\mathbf{Z}^{(t)}\right]^{-1} + \Sigma^{-1}_{\alpha_d^{(t)}}\right)^{-1}. \tag{35}$$

5 SIMULATION EXPERIMENTS

We evaluate the proposed model both on simulated and on real data. We focus on cases where individual matrices are relatively small, so that good modeling assumptions become crucial for strong predictive performance. In this section we first compare the model with other approaches using artificial data in the same range as our collected data, simulated from an underlying DNBGFA model with $t = 1, ..., 10$, $N = 200$, each $D_m^{(t)} = 20$ and hyper-parameters: $k_0^{(\eta)} = k_0^{(\alpha)} = rbf_{(0.1,100)}$, $k_0^{(r)} = k_0^{(s)} = rbf_{(1,0.1)}$, $\epsilon_r = \epsilon_2 = 1$, $\epsilon_\eta = \epsilon_\alpha = 0.1$, $a = c = 1$, $b = g = 10$, $\sigma = 0.01$. The above RBF kernel parameters emphasize time dependency in the simulated data.

We compare the proposed method DNBGFA to six other methods: NMF, GFA and its variant denoted NGFA, JPP, and an integer-based method denoted DTM, as described below.

In these experiments as well as the case studies, the data are real-valued and we focus on comparing methods that are applicable to such real-valued data; therefore, NMF, GFA and JPP are selected as comparison methods designed for real-valued data. In contrast to the above methods, methods that are restricted to integers [6, 27, 28] are not readily applicable to real-valued data. We will compare to one such method, Dynamic Topic Model (DTM) [3] as a prominent example of integer-restricted dynamic methods; due to its restriction to integer data, DTM's model building is here based on integer-rounded observations. We compare performance of the methods in two scenarios below.

Partial Article. In this scenario, we simulate a situation where only partial content of articles are observed and we aim to predict the rest. A model built from the observed document parts is used to predict left-out content of the same documents. This scenario corresponds e.g. to using news RSS feed snippets to predict the news content, or using abstracts to predict the content of full-text research articles.

We simulate the scenario by leaving a randomly selected 10% of the content of each document vector $\mathbf{x}_d^{(t)}$ in the training data set. In detail, each column of the training term-document matrix $\mathbf{X}_{train}^{(t)}$ is generated by a multinomial draw. For each document vector $\mathbf{x}_d^{(t)}$ column (document) $\mathbf{x} = [x_1, \dots, x_N]^\top$ of the original matrix $\mathbf{X}^{(t)}$, denote the total term occurrence by $||\mathbf{x}||_1$ and the vector of term occurrence proportions by $\mathbf{x}/||\mathbf{x}||_1$; we fill the corresponding column of $X_{train}^{(t)}$ as the count vector of $0.10 \cdot ||\mathbf{x}||_1$ trials from the distribution $\mathbf{x}/||\mathbf{x}||_1$. The resulting training matrix contains 10% as many term occurrences as the original.

After training a model (DNBGFA, NMF, GFA, NGFA, DTM and JPP) to obtain the underlying topic content and topic prevalence matrices, the left-out term-document matrices of complete articles $X^{(t)}$ are then estimated by $\tilde{X^{(t)}} \approx Z_{train}^{(t)} W_{train}^{(t)\top} \times 10$, where the multiplier scales the prediction to the size of left-out data.

Interpolating Missing Data. In this scenario, we leave out the entire term-document matrix out of the 10 time slices, and we repeat the scenario 10 times leaving out a different time slice each time. The task is to estimate the missing slice given its time index and number of documents. For NMF, GFA and NGFA, the missing matrix is estimated using the result of the previous time slice, where topic loadings of an unseen document are estimated by average loadings. We have also tried to use the result from the next time slice and the performance is very similar. As DTM does not directly allow missing time slices we train it with the missing slice omitted and predict using the result from the previous time slice.

For DNBGFA, the matrices of the left-out time slice $Z^{(t)}$ and $W^{(t)}$ are directly estimated from the hierarchical model based on the time index of the held-out slice, thus the missing matrix $X^{(t)}$ can be directly estimated.

Results. For both scenarios, we repeat the process 20 times to account for stochasticity in data generation and in training methods, the root mean square error (RMSE) between predicted matrix content and true left-out content is employed as the performance

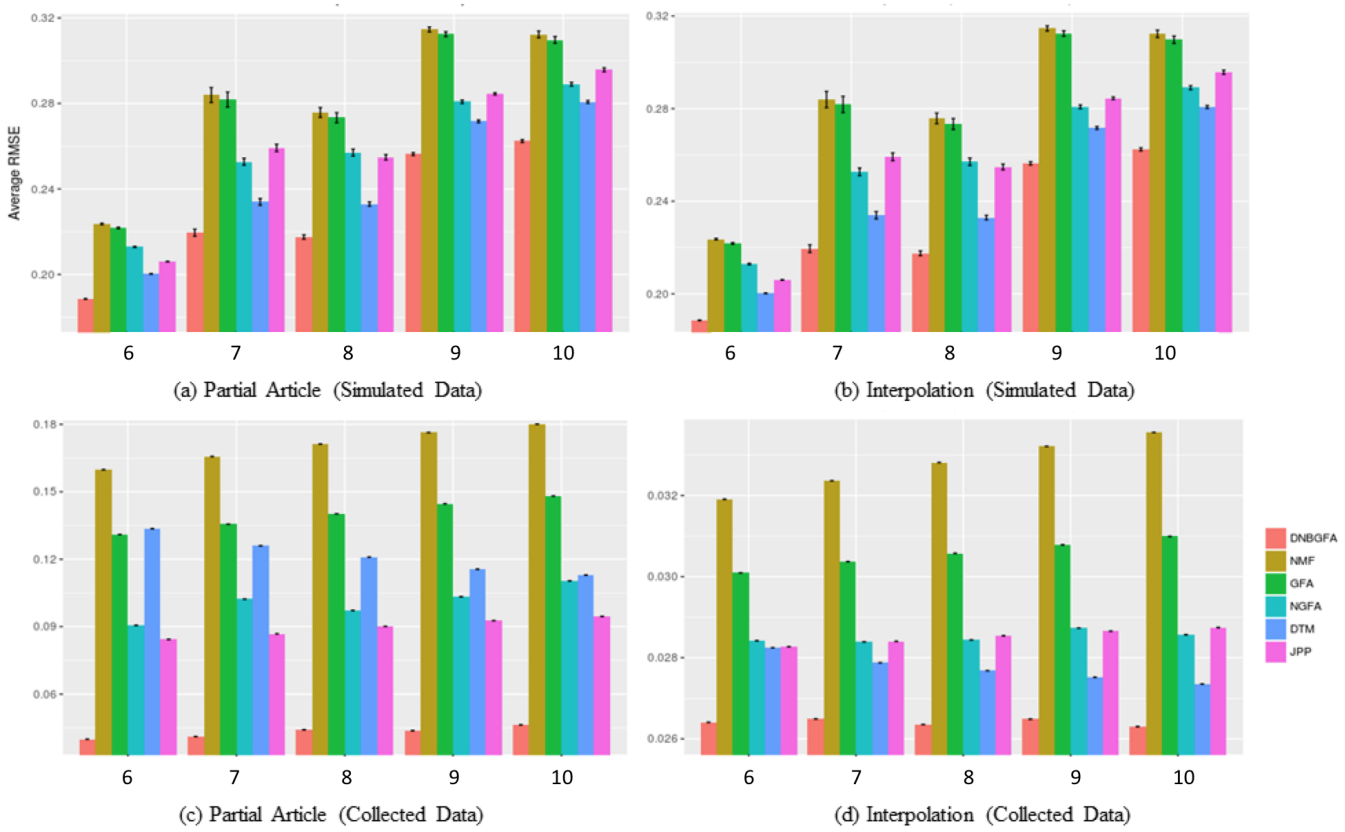

Figure 3: Performances are compared with averaged RMSE. Error bars are the variances of the mean value. DNBGFA attains the lowest average RMSE and outperforms other approaches in all four experiments (a)-(d), and for all cases over K (number of topic on the horizontal axis).

measure and pairwise t-tests between DNBGFA and other methods are then conducted to verify if the differences are statistically significant. Results can be found in Figure 3. In all cases DNBGFA achieves clearly smaller prediction error than other methods, and the differences between DNBGFA and other methods are statistically significant ($p < 0.01$).

6 CASE STUDY: FINNISH NEWS AND SOCIAL MEDIA

We apply the model to data from three text sources in 12 time slices (months) from September 2011 to August 2012, including *Helsingin Sanomat* (a Finnish newspaper), *Finnish Twitter Census* (www.finnishtwitter.com) and *Suomi24* (Finnish online forum; we take text from sections Talous (Economics) and Yhteiskunta (Society)). We remove stop-words and rare terms, lemmatize the text, then form TF-IDF weighted term-document matrices from the processed text.

Comparative Study. A comparison study is presented here, analogous to the two scenarios in Section 5 but with the above-mentioned data. For each experiment, we randomly sample 20 documents from each source and each time slice. The hyper-parameters are set as in the section 5 Results are shown in Figure 3. DNBGFA again outperforms other methods and differences are statistically significant.

Case Study: Exploratory Analysis of Topic Evolution. We further apply the proposed model to a subset of the above-mentioned dataset which contains the 150 longest documents from each time slice and each text source, yielding 5400 documents in total and 1286 terms after removing rare words and stop words. Figure 4 displays two example topics of the posterior analysis, showing their prevalence and topic content across time slices. The topic content evolution is extracted from the posterior of η (terms with highest loadings for each time slice are shown) and the prevalence is extracted from posterior of α (controlling ability of the topic to appear in documents; higher value yields higher chance to appear). Both of these topics start from low prevalences in September and October 2011, rise rapidly in November 2011, and continue with greater prevalences thereafter. Both topics have roughly equal prevalence across the sources (Suomi24 social media, Helsingin Sanomat news and Twitter), but the prevalences have differing time behavior. Prevalence in Twitter attains a peak fastest for both topics; for the Media topic Twitter prevalence has only one broad peak whereas for the Education topic there are three peaks. Prevalence in Helsingin sanomat shows two peaks for the Media topic,

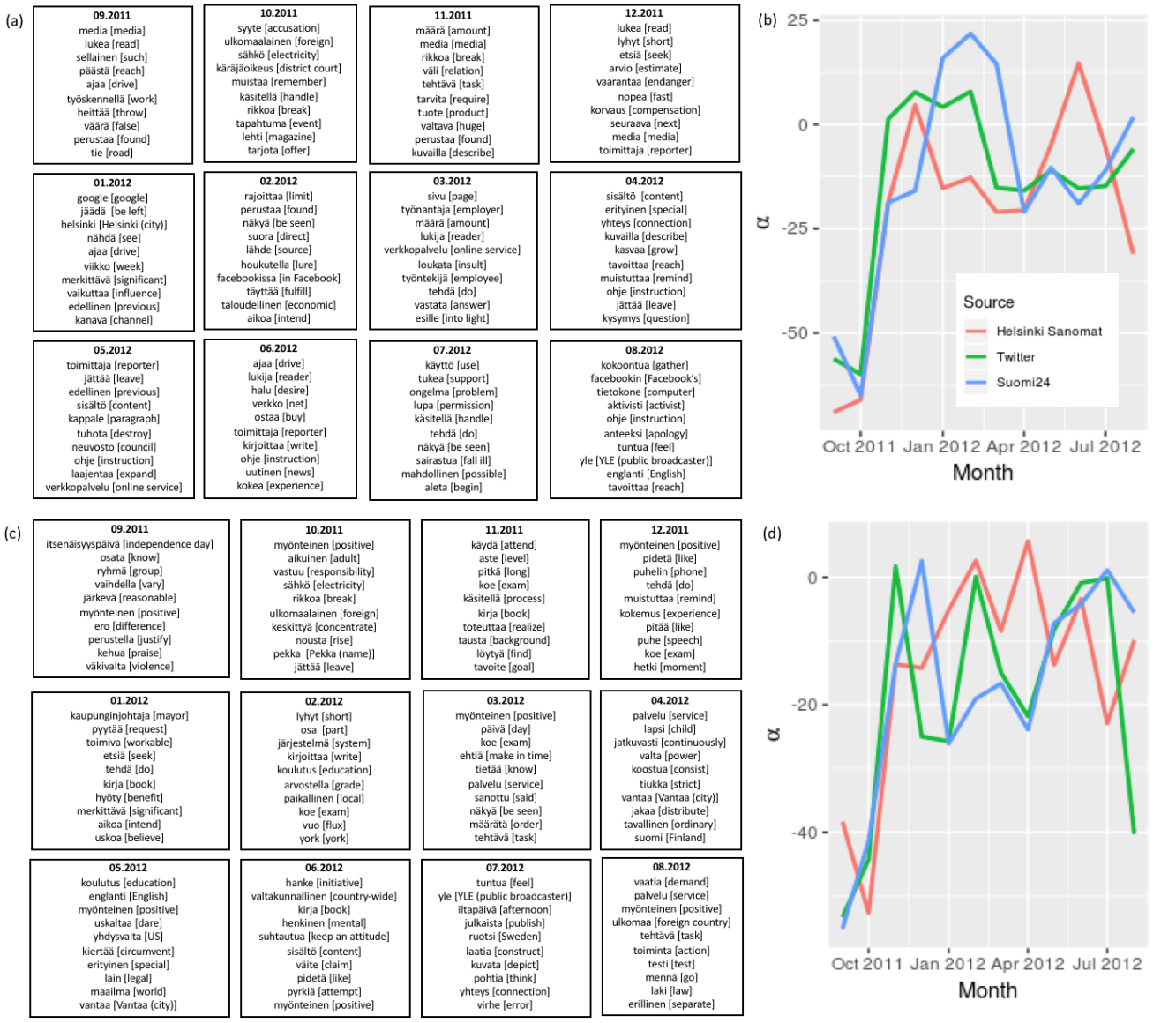

Figure 4: Evolution over time of topic content and topic sparsity (prevalence) in different sources: (a) evolution of the content of topic 'Media' in Finnish news and social media, (b) evolution of topic sparsity for the topic 'Media', (c) evolution of the content of topic 'Education' in Finnish news and social media, (d) evolution of topic sparsity for the topic 'Education'.

and noisy behavior for the Education topic. Prevalence in Suomi24 has a single peak for the Media topic in February 2012, and two peaks in December 2011 and July 2012 for the Education topic. Both topics are sensible in terms of their content and experience reasonable variation of prevalence and content over time. For example, in Figure 4 (a) the top words are all relevant to media but each time slice emphasizes a different aspect of media. It seems that time slices 09.2011, 21.2011, and 05.2012 focus more on news (contain words 'read', 'reporter' and 'paragraph') and time slices 02.2012 and 08.2012 focus more on social media (containing words 'Facebook', 'source' and 'computer'). Similarly, in Figure 4 (c) the top words refer to education with different time slices emphasizing different aspects, for example the time slice 03.2012 04.2012 emphasizes performance evaluation (containing words 'positive', 'exam' and 'task') whereas 04.2012 focuses more on education as a public service (containing words 'service', 'child' and 'city'). Our approach allows smooth changing of topic content, for example in Figure 4 (a) the word 'media' appears in adjacent time slices 11.2011 and 12.2011 of the Media topic, but with less prevalence in the latter.

7 CONCLUSIONS AND DISCUSSION

We introduced DNBGFA, a probabilistic NMF-based model that enables flexible modeling of temporal dynamics using multiple sources of data across data sources (domains) and time slices. Novelties include a Softmax+GP prior and overall structure of the hierarchical model; the model is a novel solution to address temporal dynamics and multiple sources at the same time. The hierarchical structure lets the model incorporate prior knowledge, especially underlying structure of source-source interactions and temporal dynamics, to inference, in addition to the data. The model achieved better generalization ability (ability to predict left-out data) than comparable models in realistic scenarios.The case study showed the model enables discovery of topic evolution and interactions. The model is applicable beyond text data to nonnegative matrices with multiple sources and temporal dynamics.

Our contributions are 1. Hierarchical modeling of topics shared across sources and time and topics unique to sources or time slices; 2. Discovering temporal dynamics of both topic content and prevalence; 3. Comparative studies using both simulated data and real-world data; 4. A real-world demonstration using data from three Finnish text sources.

ACKNOWLEDGEMENTS

This work was supported by the Academy of Finland, decision numbers 312395, 313748, 295694, and 327352.

REFERENCES

[1] Ayan Acharya, Joydeep Ghosh, and Mingyuan Zhou. 2015. Nonparametric Bayesian factor analysis for dynamic count matrices. *arXiv preprint arXiv:1512.08996* (2015).
[2] Claude JP Bélisle. 1992. Convergence theorems for a class of simulated annealing algorithms on d. *Journal of Applied Probability* 29, 4 (1992), 885–895.
[3] David M Blei and John D Lafferty. 2006. Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*. ACM, 113–120.
[4] Yong Chen, Hui Zhang, Junjie Wu, Xingguang Wang, Rui Liu, and Mengxiang Lin. 2015. Modeling emerging, evolving and fading topics using dynamic soft orthogonal nmf with sparse representation. In *Data Mining (ICDM), 2015 IEEE International Conference on*. IEEE, 61–70.
[5] Andreas C Damianou, Michalis K Titsias, and Neil D Lawrence. 2016. Variational inference for latent variables and uncertain inputs in Gaussian processes. *The Journal of Machine Learning Research* 17, 1 (2016), 1425–1486.
[6] Zhe Gan, Changyou Chen, Ricardo Henao, David Carlson, and Lawrence Carin. 2015. Scalable deep Poisson factor analysis for topic modeling. In *International Conference on Machine Learning*. 1823–1832.
[7] Sunil Gupta, Dinh Phung, Brett Adams, and Svetha Venkatesh. 2011. A matrix factorization framework for jointly analyzing multiple nonnegative data. In *Proceedings of the Ninth Workshop on Text Mining-Eleventh SIAM International Conference on Data Mining*. Omnipress.
[8] Patrick Jähnichen, Florian Wenzel, Marius Kloft, and Stephan Mandt. 2018. Scalable generalized dynamic topic models. *arXiv preprint arXiv:1803.07868* (2018).
[9] Bin Ju, Yuntao Qian, Minchao Ye, Rong Ni, and Chenxi Zhu. 2015. Using dynamic multi-task non-negative matrix factorization to detect the evolution of user preferences in collaborative filtering. *PloS one* 10, 8 (2015).
[10] Arto Klami, Seppo Virtanen, Eemeli Leppäaho, and Samuel Kaski. 2015. Group factor analysis. *IEEE transactions on neural networks and learning systems* 26, 9 (2015), 2136–2147.
[11] Dana Lahat, Tülay Adali, and Christian Jutten. 2015. Multimodal data fusion: an overview of methods, challenges, and prospects. *Proc. IEEE* 103, 9 (2015), 1449–1477.
[12] Neil D Lawrence. 2004. Gaussian process latent variable models for visualisation of high dimensional data. In *Advances in neural information processing systems*. 329–336.
[13] Neil D Lawrence and Andrew J Moore. 2007. Hierarchical Gaussian process latent variable models. In *Proceedings of the 24th international conference on Machine learning*. ACM, 481–488.
[14] Daniel D Lee and H Sebastian Seung. 1999. Learning the parts of objects by non-negative matrix factorization. *Nature* 401, 6755 (1999), 788.
[15] Ping Li and Songcan Chen. 2016. A review on gaussian process latent variable models. *CAAI Transactions on Intelligence Technology* 1, 4 (2016), 366–376.
[16] Minnan Luo, Feiping Nie, Xiaojun Chang, Yi Yang, Alexander Hauptmann, and Qinghua Zheng. 2017. Probabilistic non-negative matrix factorization and its robust extensions for topic modeling. In *Thirty-first AAAI conference on artificial intelligence*.
[17] John W Paisley, David M Blei, and Michael I Jordan. 2014. Bayesian Nonnegative Matrix Factorization with Stochastic Variational Inference.
[18] V Paul Pauca, Farial Shahnaz, Michael W Berry, and Robert J Plemmons. 2004. Text mining using non-negative matrix factorizations. In *Proceedings of the 2004 SIAM International Conference on Data Mining*. SIAM, 452–456.
[19] Carl Edward Rasmussen and Christopher KI Williams. 2006. *Gaussian process for machine learning*. MIT press.
[20] Margaret E Roberts, Brandon M Stewart, and Edoardo M Airoldi. 2016. A model of text for experimentation in the social sciences. *J. Amer. Statist. Assoc.* 111, 515 (2016), 988–1003.
[21] Ankan Saha and Vikas Sindhwani. 2010. Dynamic nmfs with temporal regularization for online analysis of streaming text. In *NIPS Workshop on Machine Learning for Social Computing, pp. 1C8*.
[22] Ankan Saha and Vikas Sindhwani. 2012. Learning evolving and emerging topics in social media: a dynamic nmf approach with temporal regularization. In *Proceedings of the fifth ACM international conference on Web search and data mining*. ACM, 693–702.
[23] Leo Taslaman and Björn Nilsson. 2012. A framework for regularized non-negative matrix factorization, with application to the analysis of gene expression data. *PloS one* 7, 11 (2012), e46331.
[24] Carmen K Vaca, Amin Mantrach, Alejandro Jaimes, and Marco Saerens. 2014. A time-based collective factorization for topic discovery and monitoring in news. In *Proceedings of the 23rd international conference on World wide web*. ACM, 527–538.
[25] Seppo Virtanen, Arto Klami, Suleiman Khan, and Samuel Kaski. 2012. Bayesian group factor analysis. In *Artificial Intelligence and Statistics*. 1269–1277.
[26] Chong Wang and David M Blei. 2013. Variational inference in nonconjugate models. *Journal of Machine Learning Research* 14, Apr (2013), 1005–1031.
[27] Xuerui Wang and Andrew McCallum. 2006. Topics over time: a non-Markov continuous-time model of topical trends. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 424–433.
[28] Mingyuan Zhou, Lauren A Hannah, David B Dunson, and Lawrence Carin. 2012. Beta-negative binomial process and Poisson factor analysis. *Journal of Machine Learning Research* (2012).