TWO-DIMENSIONAL CORRELATED TOPIC MODELS

Suwon Suh and Seungjin Choi

Department of Computer Science and Engineering, POSTECH, Korea
{caster,seungjin}@postech.ac.kr

ABSTRACT

Latent Dirichlet allocation (LDA) is a widely-used topic model where a set of hidden topics is learned to model a collection of data in the form of bag-of-words. Correlated topic model (CTM) extends LDA, modelling topic correlation by using logistic normal prior for topic proportion vectors, instead of Dirichlet prior. However, in the case of bag-of-words from multiple data sources (for instance, streams of time-stamped measurements in sensor networks), where each word in a document is labeled with one of data sources from which the data comes, it is desirable to consider correlations across topics as well as across data sources. In this paper we present *two-dimensional correlated topic model* (2D-CTM) where we use a *matrix-variate* logistic normal distribution for a topic proportion matrix, in which correlations across topics as well as across data sources are captured by two covariance matrices of the matrix-variate normal distribution. We develop a mean-field variational inference algorithm for approximate posterior inference in our model 2D-CTM. We apply 2D-CTM to the problem of human activity recognition and sport tactic analysis, using data collected on multiple on-body sensors, with comparison to existing topic models.

Index Terms— Latent Dirichlet allocation, sports tactic analysis, topic models.

1. INTRODUCTION

Latent Dirichlet allocation (LDA) is a widely-used topic model, which was originally developed to model text corpora [3]. LDA is a hierarchical Bayesian model in which each observed token is modelled as a finite mixture over an underlying set of topics and each topic is characterized by a distribution over words. Correlated topic model (CTM) extends LDA to overcome the inability of LDA to model topic correlations, replacing Dirichlet distribution by logistic normal distribution for the topic proportions [2].

Various extensions of LDA or CTM have been developed to handle more complex data or to incorporate meta data. Pachinko allocation uses a directed acyclic graph to capture arbitrary correlations between topics [10]. The Dirichlet-multinomial regression topic model [11] is an exemplary model where meta data is incorporated into topic models. One interesting model, which is closely related to our work, is multi-field CTM (mf-CTM) [15] which allows multiple sets of topics to be used for different fields or for different data sources. Topic proportion vectors are drawn from logistic normal distributions, as in CTM, however, they are partitioned into several vectors, each of which is associated with corresponding field. The size of covariance matrix in mf-CTM is increasing in accordance with the number of fields as well as with the number of topics, requiring a large number of parameters, compared to CTM.

In this paper we consider a case where bag-of-words data come from multiple data sources (for instance, streams of time-stamped measurements in sensor networks), where each word in a document is labeled with one of data sources from which the data comes. In such a case, it is desirable to consider both correlations across topics and across data sources. To this end, we present *two-dimensional correlated topic model* (2D-CTM) where we use a *matrix-variate* logistic normal distribution for a topic proportion matrix, in which both correlations across topics and across data sources are captured by two covariance matrices of the matrix-variate normal distribution. This provides a more compact parameterization, in comparison to mf-CTM. We develop a mean-field variational inference algorithm for approximate posterior inference in our model 2D-CTM. We validate the performance of 2D-CTM, applying it to the on-body sensor-based activity recognition, where data collected from multiple on-body sensors is classified into one of predefined activity categories.

A two-dimensional topic-aspect model [14] or factorial LDA [13] is also an extension of CTM using a multi-dimensional technique. However, these models focus on multi-faceted data, which is different from our goal. In fact, our model 2D-CTM shares a similar underlying spirit to 2D-LDA [17] or 2D-CCA [9] where subspace methods are extended to handle matrix data directly rather than multivariate data.

2. RELATED WORK: CORRELATED TOPIC MODEL

We briefly give an overview of correlated topic model (CTM) [2]. Each document, denoted by $\boldsymbol{w}_{d,1:N}$, is a sequence of N words, for $d = 1, \ldots, D$ (D is the size of a corpus) and each word $\boldsymbol{w}_{d,n} \in \mathbb{R}^V$ (V is the size of vocabulary) is a unit vector that has a single entry equal to one and all other entries equal zeros. For instance, if $\boldsymbol{w}_{d,n}$ is the vth word in the vocabulary, then $w_{d,n,v} = 1$ and $w_{d,n,j} = 0$ for $j \neq v$. The graphical model for LDA is shown in Fig. 1a, where each word $\boldsymbol{w}_{d,n}$ for $n = 1, \ldots, N$ in document d is assumed to be generated as follows:

- Draw a vector $\boldsymbol{\theta}_d \in \mathbb{R}^K$, $\boldsymbol{\theta}_d \,|\, \{\boldsymbol{\mu}, \boldsymbol{\Sigma}\} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$.
- Compute a vector of topic proportions, $\widetilde{\boldsymbol{\theta}}_d \in \mathbb{R}^K$, whose entry $\widetilde{\theta}_{d,k}$ is the soft-max transform of $\theta_{d,k}$, a mapping of real values drawn from the Gaussian to the probability simplex, $\widetilde{\theta}_{d,k} = \exp(\theta_{d,k}) / \sum_{j=1}^{K} \exp(\theta_{d,j})$.
- For each word n,
  1) Draw a topic assignment $\boldsymbol{z}_{d,n} \in \mathbb{R}^K$ from multinomial distribution: $\boldsymbol{z}_{d,n} \,|\, \boldsymbol{\theta}_d \sim \text{Mult}\left(\widetilde{\boldsymbol{\theta}}_d\right) = \prod_{k=1}^{K} \left(\widetilde{\theta}_{d,k}\right)^{z_{d,n,k}}$.
  2) Draw a word $\boldsymbol{w}_{d,n} \in \mathbb{R}^V$:
  $$\boldsymbol{w}_{d,n} | \boldsymbol{z}_{d,n}, \boldsymbol{\phi}_{1:K} \sim p(\boldsymbol{w}_{d,n} | \boldsymbol{z}_{d,n}, \boldsymbol{\phi}_{1:K}),$$
  where $\boldsymbol{\phi}_{1:K}$ is a shorthand notation for $\{\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_K\}$.

Variational inference was applied in [2] to calculate approximate posterior distributions over hidden variables, $\{\boldsymbol{\theta}_d, \boldsymbol{z}_{d,n}\}$, by maximizing the variational lower-bound on the log marginal likelihood $p(\boldsymbol{w}_{1:D,1:N}|\boldsymbol{\alpha}, \boldsymbol{\phi}_{1:K})$.

3. TWO-DIMENSIONAL CORRELATED TOPIC MODEL

In this section, we present our main contribution, 2D-CTM, whose graphical representation is shown in Fig. 1b. In contrast to CTM or LDA, we consider the case where data come from multiple data sources. Each word $\boldsymbol{w}_{d,n}$ is given with label variable $\boldsymbol{l}_{d,n}$, for data source, which is a I-dimensional unit vector (I is the number of data sources) used to represent which data source yields the word $\boldsymbol{w}_{d,n}$. For instance, if word j in document d corresponds to the data sample n measured at sensor i, then $w_{d,n,j} = 1$ and $l_{d,n,i} = 1$. A distinct property of 2D-CTM is to use a matrix of topic proportions, $\widetilde{\boldsymbol{\Theta}}_d \in \mathbb{R}^{I\times K}$, whose entries $\widetilde{\Theta}_{d,i,k}$ are soft-max transform of $\Theta_{d,i,k}$ which are elements of a *random matrix* drawn from *matrix-variate normal* distribution, in order to capture correlations both across topics and across data sources. A brief review of matrix-variate normal distribution is provided below, followed by describing 2D-CTM model and presenting corresponding mean-field variational inference and parameter estimation.

3.1. Matrix-Variate Normal Distribution

The random matrix $\boldsymbol{X} \in \mathbb{R}^{R\times C}$ is said to have a *matrix-variate normal distribution* [6] with mean matrix $\boldsymbol{M} \in \mathbb{R}^{R\times C}$ and covariance matrices $\boldsymbol{\Sigma} \in \mathbb{R}^{R\times R}$ and $\boldsymbol{\Psi} \in \mathbb{R}^{C\times C}$,

$$\boldsymbol{X} \sim \mathcal{MN}_{R\times C}(\boldsymbol{M}, \boldsymbol{\Sigma}, \boldsymbol{\Psi}),$$

if $\mathrm{vec}(\boldsymbol{X})$ obeys RC-dimensional multivariate normal distribution, i.e., $\mathrm{vec}(\boldsymbol{X}) \sim \mathcal{N}\big(\mathrm{vec}(\boldsymbol{M}), \boldsymbol{\Psi}\otimes\boldsymbol{\Sigma}\big)$, where $\mathrm{vec}(\cdot)$ is the vectorization of a matrix which converts a matrix into a column vector obtained by stacking the columns of a matrix on top of another, and $\otimes$ represents Kronecker product. The probability density function of $\boldsymbol{X}$ is given by

$$p(\boldsymbol{X}) = \frac{\exp\Big\{-\frac{1}{2}\mathrm{tr}\Big(\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}-\boldsymbol{M})\boldsymbol{\Psi}^{-1}(\boldsymbol{X}-\boldsymbol{M})^\top\Big)\Big\}}{(2\pi)^{\frac{RC}{2}}|\boldsymbol{\Sigma}|^{\frac{C}{2}}|\boldsymbol{\Psi}|^{\frac{R}{2}}},$$

where $\mathrm{tr}(\cdot)$ denotes the trace operator which computes the sum of all diagonal entries of a matrix.

3.2. Model

The generation process for each trajectory word $\{\boldsymbol{w}_{d,n}\}$ for $n = 1,\ldots,N$ in data d is as follows.

- Draw a matrix $\boldsymbol{\Theta}_d \in \mathbb{R}^{I\times K}$, from matrix-variate normal distribution: $\boldsymbol{\Theta}_d \sim \mathcal{MN}_{I\times K}(\boldsymbol{M}, \boldsymbol{\Sigma}, \boldsymbol{\Psi})$.
- Compute a matrix of topic proportions, $\widetilde{\boldsymbol{\Theta}}_d \in \mathbb{R}^{I\times K}$, whose entry $\widetilde{\Theta}_{d,i,k}$ is the soft-max transform of $\Theta_{d,i,k}$, $\widetilde{\Theta}_{d,i,k} = \exp(\Theta_{d,i,k})/\sum_{k'=1}^{K}\exp(\Theta_{d,i,k'})$.
- For each word n,
  1) Draw a topic assignment $\boldsymbol{z}_{d,n} \in \mathbb{R}^K$ from multinomial distribution: $\boldsymbol{z}_{d,n} \,|\, \boldsymbol{\Theta}_d, \boldsymbol{l}_{d,n} \sim p(z_{d,n}|\widetilde{\Theta}_d, l_{d,n})$.
  2)Draw a word $\boldsymbol{w}_{d,n} \in \mathbb{R}^V$:
  $\boldsymbol{w}_{d,n} \,|\, \boldsymbol{z}_{d,n}, \boldsymbol{\phi}_{1:K} \sim p(\boldsymbol{w}_{d,n}|\boldsymbol{z}_{d,n}, \boldsymbol{\phi}_{1:K})$.

3.3. Variational Inference

We present variational inference [8] for 2D-CTM, where a variational lower-bound on marginal likelihood is maximized to approximately compute posterior distributions over hidden variables and to determine the most probable values of parameters. We define sets of variables $\mathcal{W} = \{\boldsymbol{w}_{d,n}\}$, $\mathcal{Z} = \{\boldsymbol{z}_{d,n}\}$, $\mathcal{L} = \{\boldsymbol{l}_{d,n}\}, \boldsymbol{\Theta} = \{\boldsymbol{\Theta}_d\}$ and a set of parameters $\{\boldsymbol{M}, \boldsymbol{\Sigma}, \boldsymbol{\Psi}, \boldsymbol{\phi}_{1:K}\}$. Then the joint distribution over these variables obeys the following factorization:

$$\begin{aligned} &p(\boldsymbol{\Theta}, \mathcal{Z}, \mathcal{W}, \mathcal{L}|\boldsymbol{M}, \boldsymbol{\Sigma}, \boldsymbol{\Psi}, \boldsymbol{\phi}_{1:K}) \\ &\quad = \quad p(\boldsymbol{\Theta}|\boldsymbol{M}, \boldsymbol{\Sigma}, \boldsymbol{\Psi})p(\mathcal{Z}|\boldsymbol{\Theta}, \mathcal{L})p(\mathcal{W}|\mathcal{Z}, \mathcal{L}, \boldsymbol{\phi}_{1:K})p(\mathcal{L}), \end{aligned}$$

where the first distribution is parameterized by a product of matrix-variate normal distributions $p(\boldsymbol{\Theta}|\boldsymbol{M}, \boldsymbol{\Sigma}, \boldsymbol{\Psi}) = \prod_{d=1}^{D} p(\boldsymbol{\Theta}_d|\boldsymbol{M}, \boldsymbol{\Sigma}, \boldsymbol{\Psi})$, where

$$\begin{aligned} &p(\boldsymbol{\Theta}_d|\boldsymbol{M}, \boldsymbol{\Sigma}, \boldsymbol{\Psi}) \\ &\quad = \quad \frac{\exp\Big\{-\frac{1}{2}\mathrm{tr}\Big(\boldsymbol{\Sigma}^{-1}(\boldsymbol{\Theta}_d - \boldsymbol{M})\boldsymbol{\Psi}^{-1}(\boldsymbol{\Theta}_d - \boldsymbol{M})^\top\Big)\Big\}}{(2\pi)^{\frac{IK}{2}}|\boldsymbol{\Sigma}|^{\frac{K}{2}}|\boldsymbol{\Psi}|^{\frac{I}{2}}}, \end{aligned}$$

and the second and third distributions are parameterized by products of multinomial distributions

$p(\mathcal{Z}|\boldsymbol{\Theta}, \mathcal{L}) = \prod_{d=1}^{D}\prod_{n=1}^{N}\prod_{i=1}^{I}\prod_{k=1}^{K}\big(\widetilde{\Theta}_{d,i,k}\big)^{l_{d,n,i}\, z_{d,n,k}}$,
$p(\mathcal{W}|\mathcal{Z}, \mathcal{L}, \boldsymbol{\phi}_{1:K}) = \prod_{d=1}^{D}\prod_{n=1}^{N}\prod_{k=1}^{K}\prod_{j=1}^{V}\big(\phi_{k,j}\big)^{z_{d,n,k}\, w_{d,n,j}}$,
and $p(\mathcal{L})$ is a constant, which will be left out in subsequent calculations, since $\mathcal{L}$ is a set of observed variables.

Define a set of parameters as $\mathcal{M} = \{\boldsymbol{M}, \boldsymbol{\Sigma}, \boldsymbol{\Psi}, \boldsymbol{\phi}_{1:K}\}$. Marginalizing hidden variables out yields the log marginal likelihood that is of the form

$$\begin{aligned} &\log p(\mathcal{W}, \mathcal{L}|\mathcal{M}) \\ &\quad \geq \int_{\boldsymbol{\Theta}}\sum_{\mathcal{Z}} q(\boldsymbol{\Theta}, \mathcal{Z})\log\left(\frac{p(\boldsymbol{\Theta}, \mathcal{Z}, \mathcal{W}, \mathcal{L}|\mathcal{M})}{q(\boldsymbol{\Theta}, \mathcal{Z})}\right)d\boldsymbol{\Theta} = \mathcal{F}(q), \end{aligned}$$

where $q(\boldsymbol{\Theta}, \mathcal{Z})$ denotes the *variational distribution* and Jensen's inequality is to used to reach the *variational lower-bound* $\mathcal{F}(q)$. We assume that the variational distribution factorizes as $q(\boldsymbol{\Theta}, \mathcal{Z}) = q(\boldsymbol{\Theta})q(\mathcal{Z})$, where each distribution is assumed to be of the form of Table .1. Variational parameters, $\Big\{\{\overline{M}_{d,i,k}\}, \{\Gamma_{d,i,k}\}, \{\rho_{d,n,k}\}\Big\}$, are determined by maximizing the variational lower-bound

$$\begin{aligned} \mathcal{F}(q) \quad &= \quad \mathbb{E}_q\Big[\log p(\boldsymbol{\Theta}|\boldsymbol{M}, \boldsymbol{\Sigma}, \boldsymbol{\Psi}) + \log p(\mathcal{Z}|\boldsymbol{\Theta}, \mathcal{L}) \\ &\quad + \quad \log p(\mathcal{W}|\mathcal{Z}, \mathcal{L}, \boldsymbol{\phi}_{1:K})\Big] - \mathbb{E}_q\Big[\log q(\boldsymbol{\Theta}) + \log q(\mathcal{Z})\Big], \end{aligned}$$

where $\mathbb{E}_q[\cdot]$ denotes the statistical expectation with respect to the variational distribution $q(\cdot)$.

One thing must be noted in this calculation is that unlike CTM, we consider a local quadratic bound for the log-sum of exponentials [4] where the standard quadratic bound [7] for $\log(1+e^x)$,

$\log(1+e^x) \leq \lambda(\xi)(x^2-\xi^2) + \frac{x-\xi}{2} + \log(1+e^\xi)$,

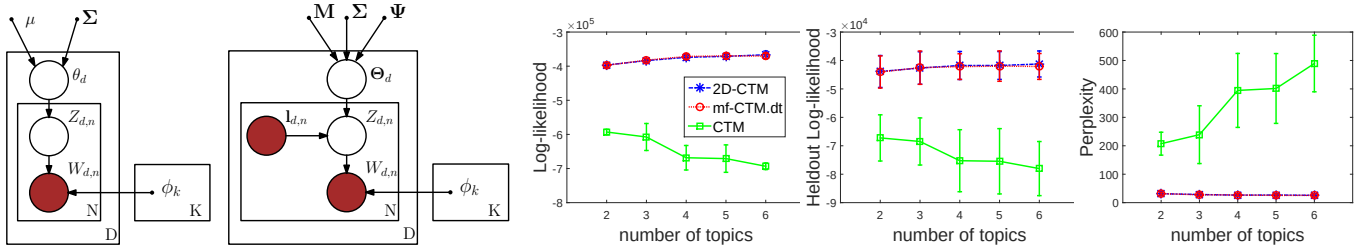
follows the fact that the sum of exponentials is upper-bounded by a product of sigmoids, $\sum_i e^{x_i} \leq \prod_i(1+e^{x_i})$, where $\lambda(\xi) = \frac{1}{4\xi}\tanh(\frac{\xi}{2})$. As a result, this local quadratic bound enable us to have closed form solutions for all the parameter updates unlike CTM. Closed-form updates for variational parameters are obtained by solving a corresponding stationary point equation of $\mathcal{F}_q$ for a parameter of interest, which is summarized in Table. 1. Variational parameters $\xi_{d,i,j}$ for document d, which appear in the local quadratic approximation for log-sum of exponentials, are updated by

$$\xi_{d,i,j} = \Big(\Gamma_{d,i,j} + \overline{M}^2_{d,i,j}\Big)^{\frac{1}{2}},$$

for $i = 1,\ldots,I$ and $j = 1,\ldots,K$.

Table 1. Updating equations for variational parameters of 2D-CTM.

Variational posterior distributions	Updating equations for variational parameters
$q(\boldsymbol{\Theta}) = \prod_{d=1}^{D}\prod_{i=1}^{I}\prod_{k=1}^{K}\mathcal{N}(\Theta_{d,i,k}\|\overline{M}_{d,i,k},\Gamma_{d,i,k})$	$\overline{M}_{d,i,k} = B_{d,i,k}/(\Sigma_{i,i}^{-1}\Psi_{k,k}^{-1} + 2\sum_{n=1}^{N} l_{d,n,i}\lambda(\xi_{d,i,k}))$, $\Gamma_{d,i,k} = (\Sigma_{i,i}^{-1}\Psi_{k,k}^{-1} + 2\sum_{n=1}^{N} l_{d,n,i}\lambda(\xi_{d,i,k}))^{-1}$, where $B_{d,i,k} = -\sum_{j\neq i}\sum_{l\neq k}\Sigma_{i,j}^{-1}(\overline{M}_{d,j,l} - M_{j,l})\Psi_{l,k}^{-1} - \sum_{j\neq i}\Sigma_{i,j}^{-1}(\overline{M}_{d,j,k} - M_{j,k})\Psi_{k,k}^{-1}$ $-\sum_{l\neq k}\Sigma_{i,i}^{-1}(\overline{M}_{d,i,l} - M_{i,l})\Psi_{l,k}^{-1} + \Sigma_{i,i}^{-1}M_{i,k}\Psi_{k,k}^{-1} + \sum_{n=1}^{N} l_{d,n,i}\rho_{d,n,k} - \frac{1}{2}\sum_{n=1}^{N} l_{d,n,i}$
$q(\mathcal{Z}) = \prod_{d=1}^{D}\prod_{n=1}^{N}\prod_{k=1}^{K}(\rho_{d,n,k})^{z_{d,n,k}}$	$\log\rho_{d,n,k} \propto \sum_{j=1}^{V} w_{d,n,j}\log\phi_{k,j} + \sum_{i=1}^{I} l_{d,n,i}\Big[\overline{M}_{d,i,k}$ $-\sum_{j=1}^{K}\Big\{\lambda(\xi_{d,i,j})(\Gamma_{d,i,j} + \overline{M}_{d,i,j}^2 - \xi_{d,i,j}^2) + \frac{(\overline{M}_{d,i,j} - \xi_{d,i,j})}{2} + \log(1 + e^{\xi_{d,i,j}})\Big\}\Big]$



(a) CTM. (b) 2D-CTM. (c) ELOB (d) Held-out ELOB (e) Partially observed

Fig. 1. Graphical representation of (a)CTM and (b)2D-CTM with performance comparison in terms of log-likelihood and perplexity: (c) log-likelihood of training data in 'walk' class; (d) log-likelihood of held-out data in 'walk' class using 10-fold cross validation; (e) perplexity when half of words in each document are observed;

Table 2. Performance comparison in terms of F-measure.

CTM	mf-CTM.dt	2D-CTM	NN	3-NN	PCA-NN
0.8885	0.7450	0.7700	0.4701	0.4250	0.3568

3.4. Parameter Estimation

Maximizing the variational lower-bound $\mathcal{F}_q$ also yields updates for parameters. Multinomial parameters $\{\phi_{k,j}\}$ are updated by

$$\phi_{k,j} = \frac{\sum_{d=1}^{D}\sum_{n=1}^{N}\rho_{d,n,k}w_{d,n,j}}{\sum_{j=1}^{V}\sum_{d=1}^{D}\sum_{n=1}^{N}\rho_{d,n,k}w_{d,n,j}}.$$

Parameters $\{\boldsymbol{M},\boldsymbol{\Sigma},\boldsymbol{\Psi}\}$ involving matrix-variate normal distribution are also updated by

$$\boldsymbol{M} = \frac{1}{D}\sum_{d=1}^{D}\overline{\boldsymbol{M}}_d,$$

$$\Sigma_{i,j} = \frac{1}{KD}\sum_{d=1}^{D}\Big[(\widetilde{\boldsymbol{M}}_d)\boldsymbol{\Psi}^{-1}(\widetilde{\boldsymbol{M}}_d)^{\top}\Big]_{i,j} + \frac{\delta_{i,j}}{KD}\sum_{d=1}^{D}\sum_{k=1}^{K}\Gamma_{d,i,k}\Psi_{k,k}^{-1},$$

$$\Psi_{i,j} = \frac{1}{ID}\sum_{d=1}^{D}\Big[(\widetilde{\boldsymbol{M}}_d)^{\top}\boldsymbol{\Sigma}^{-1}(\widetilde{\boldsymbol{M}}_d)\Big]_{i,j} + \frac{\delta_{i,j}}{ID}\sum_{d=1}^{D}\sum_{i=1}^{I}\Gamma_{d,i,k}\Sigma_{i,i}^{-1},$$

where $\widetilde{\boldsymbol{M}}_d = (\overline{\boldsymbol{M}}_d - \boldsymbol{M})$, $\delta_{i,j}$ denotes Kronecker delta which equals one for $i = j$ and zero otherwise. We further regularize two covariance submatrices $\boldsymbol{\Sigma},\boldsymbol{\Phi}$ with sparse covariance estimation as follows [1]:

$$\boldsymbol{\Sigma} \leftarrow \mathcal{S}(\boldsymbol{\Sigma},\eta\boldsymbol{P}_K), \boldsymbol{\Psi} \leftarrow \mathcal{S}(\boldsymbol{\Psi},\eta\boldsymbol{P}_I),$$

where $\mathcal{S}(\boldsymbol{C},\boldsymbol{P})$ is the elementwise soft-thresholding operator $\mathcal{S}(\boldsymbol{C},\boldsymbol{P})_{i,j} = \text{sgn}(C_{i,j})\max(C_{i,j} - P_{i,j}, 0)$, $\boldsymbol{P}_K \in \mathbb{R}^{K\times K}$ is a matrix filled with ones but zeros on diagonal entries, and η is a sparsity parameter. If η has a larger value, the covariance matrix will be more sparse.

4. EXPERIMENTS

We evaluate the performance of our 2D-CTM, applying it to the problem of human activity recognition (HAR), where CTM was successfully used [16]. We compare 2D-CTM with CTM [2] and mf-CTM [15], carrying out experiments on the on-body multiple sensor-based HAR dataset [5]. We evaluate our model in unsupervised task as well as supervised task.

In an unsupervised task, we evaluate the log-likelihood of training set for each of 5 classes, but here we report the results for class 'walk' only. The log-likelihood of training set for each model (2D-CTM, CTM, and mf-CTM) is shown in Fig. 1c, where 2D-CTM and mf-CTM.dt outperforms CTM. The difference of 2D-CTM and mf-CTM.dt is almost negligible. The average log-likelihood of held-out data using 10-fold cross-validation for each model is shown in Fig. 1d, where 2D-CTM and mf-CTM provide a better fit than CTM and the difference of 2D-CTM and mf-CTM.dt is almost negligible as before. We evaluate the perplexity of each model which measures how well the model predicts the rest of words when a fraction of words in a document is observed. The perplexity of unobserved words in held-out documents is defined as the reciprocal geometric

mean of the likelihood of unobserved words $\boldsymbol{w}_{d,n}$ in document d given observed words $\boldsymbol{w}_{d,1:B_d}$ and trained parameters $\mathcal{M}$:

$$\text{Perp} = \left(\prod_{d=1}^{D} \prod_{n=B_d+1}^{N_d} p(\boldsymbol{w}_{d,n} | \boldsymbol{w}_{d,1:B_d}; \mathcal{M}) \right)^{\frac{-1}{\sum_{d=1}^{D} N_d - B_d}},$$

where B_d is the number of observed words in held-out document d. We consider the first half of words along time stamp in each held-out document is observed and the rest is to be predicted (see Fig. 1e). In this cases, both 2D-CTM and mf-CTM outperform CTM, while K varies from 2 to 6. And 2D-CTM and mf-CTM shows similar result as before.

Both models considering correlation over sources and topics beat the model only considering correlation over topics, which justifies why we consider correlation over sources in addition. Moreover, 2D-CTM requires only $I(I+1)/2 + K(K+1)/2$ parameters for covariance matrices, while mf-CTM.dt needs $(IK)(IK+1)/2$ parameters for the big covariance matrix. Even tough the compact parameterization of 2D-CTM compared to mf-CTM.dt, the performance in unsupervised task do not degrade, which implies 2D-CTM may scale better than mf-CTM.dt without degrading performance.

In a supervised task, we compare topic models with baseline methods in [5], t o evaluate how well a locomotion mode is predicted given an unseen document. In Opportunity challenge [5], nearest neighbor (NN) methods were successfully used, including NN, 3-NN, PCA-NN (PCA is followed by NN). In our experiments, we first learn parameters of 5 topic models $\mathcal{M}_{stand}, \mathcal{M}_{walk}, \mathcal{M}_{none}, \mathcal{M}_{sit}, \mathcal{M}_{lie}$, each of which corresponds to Stand, Walk, None, Sit and Lie documents. Then, given an unseen document d, we choose a model that maximizes posterior probability $p(\mathcal{M}_{class}|d) \propto p(d|\mathcal{M}_{class})p(\mathcal{M}_{class})$ with uniform distribution over the model parameters. In this way, we can fully exploit the influence of the covariance matrix of topic models. Results are summarized in Table 2[1] where events with missing values are discarded. 2D-CTM outperforms NN methods in terms of F-measure as well as mf-CTM. This result shows the robustness of 2D-CTM in the presence of missing values, which frequently occur in sensor networks. However, CTM shows better performance than 2D-CTM and mf-CTM.dt, implying that the supervised task favors a model that considers correlations over topics only.

In addition, we apply 2D-CTM to DEBS 2013 dataset [12] for sport tactic analysis and define *tactic* as two correlated movements(topics) are carried out by two correlated players(sources). DEBS 2013 consist of real time trajectories of 8 vs 8 soccer game. We first make input stream data quantization with location, acceleration and direction with sliding time window and take it as a word with player identity as label only if its acceleration exceeds certain threshold. Then, we label 29 attacking scenes in the first half as documents. We apply two types of 2D-CTM (*fixed* means $\boldsymbol{\Sigma}$ is fixed as empirical covariance across sources; *init* means $\boldsymbol{\Sigma}$ is just initialized with it) and mf-CTM.dt to this dataset. As shown in Fig. 3, 2D-CTM has the highest average held-out log-likelihood when K=4 and both 2D-CTM variants works robustly as the number of topic increase due to the compact parameterization than mf-CTM.dt, which is prone to over-fitting. The result of sports tactics analysis is demonstrated in Fig. 2 [2], where two orthogonal tactics are mined;

[1]We select the model with highest F-measure: CTM(K=3), mf-CTM.dt(K=5), 2D-CTM(K=5).

[2]Even though it is less predictive than K=4, we use 15 topics (K=15) because the mined tactics are more recognizable by human.

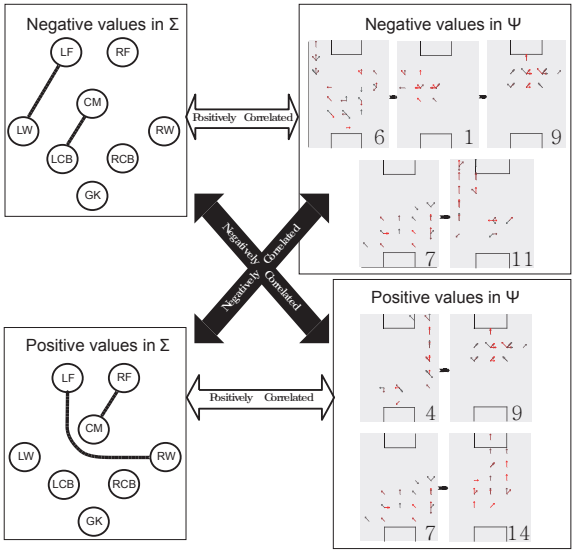


Fig. 2. Mined football tactics with DEBS2013 dataset; two left panels side show correlations between two players(source), while two right panels show correlation across movements(topics). Each row stands for the such tactics, which is orthogonal to each other.

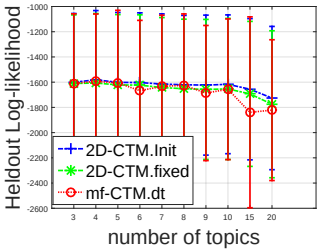


Fig. 3. Held-out ELOB

left and right side attack with corresponding pair of movements and pair of players.

5. CONCLUSIONS

In this paper, we have presented 2D-CTM in order to model discrete data measured from multiple data sources, incorporating correlations between data sources into the mode. Matrix-variate logistic normal distribution was introduced as an prior for the topic proportion matrix. In contrast to the CTM, 2D-CTM captured correlations across topics as well as correlations across data sources, which allowed us to find topics shared over all the data sources. We have evaluated the performance of 2D-CTM, applying it the on-body sensor-based human activity recognition task dataset as well as DEBS2013 dataset for sport tactic analysis. Our model 2D-CTM yielded more compact parameterization, compared to mf-CTM.dt, while achieving the compatible (or slightly better) performance and resistant to over-fitting.

Acknowledgments: This work was supported by the IT R&D Program of MSIP/IITP (B0101-15-0307, Machine Learning Center) and National Research Foundation (NRF) of Korea (NRF-2013R1A2A2A01067464). I have been grateful for discussion with the late Jaedeug Choi and Hyohyeong Kang.

6. REFERENCES

[1] J. Bien and R. Tibshirani, "Sparse estimation of a covariance matrix," *Biometrika*, vol. 98, no. 4, pp. 807–820, 2011.

[2] D. M. Blei and J. D. Lafferty, "Correlated topic models," in *Advances in Neural Information Processing Systems (NIPS)*, vol. 18, 2006.

[3] D. M. Blei, A. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[4] G. Bouchard, "Efficient bounds for the softmax function, applications to inference in hybrid models," in *Proceedings of the NIPS Workshop on Approximate Bayesian Inference in Continuous/Hybrid Systems*, 2007.

[5] R. Chavarriaga, H. Sagha, A. Calatroni, S. T. Digumarti, G. Tröster, J. del. R. Millán, and D. Roggen, "The opportunity challenge: A benchmark database for on-body sensor-based activity recognition," *Pattern Recognition Letters*, vol. 34, pp. 2033–2042, 2013.

[6] A. K. Gupta and D. K. Nagar, *Matrix Variate Distributions*. Chapman & Hall/CRC, 1999.

[7] T. S. Jaakkola and M. I. Jordan, "A variational approach to Bayesian logistic regression problems and their extensions," in *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 1996.

[8] ——, "Bayesian parameter estimation via variational methods," *Statistics and Computing*, vol. 10, pp. 25–37, 2000.

[9] S. H. Lee and S. Choi, "Two-dimensional canonical correlation analysis," *IEEE Signal Processing Letters*, vol. 14, no. 10, pp. 753–758, 2007.

[10] W. Li and A. McCallum, "Pachinko allocation: DAG-structured mixture models of topic correlations," in *Proceedings of the International Conference on Machine Learning (ICML)*, Pittsburgh, PA, USA, 2006.

[11] D. Mimno and A. McCallum, "Topic models conditioned on arbitrary features with Dirichlet-Multinomial regression," in *Proceedings of the Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, Helsinki, Finland, 2008.

[12] C. Mutschler, H. Ziekow, and Z. Jerzak, "The DEBS 2013 grand challenge," in *Proceedings of The 7th ACM International Conference on Distributed Event-Based Systems (DEBS)*, Arlington, Texas, USA, 2013.

[13] M. Paul and M. Dredze, "Factorial LDA: Sparse multi-dimensional text models," in *Advances in Neural Information Processing Systems (NIPS)*, vol. 25, 2012.

[14] M. Paul and R. Girju, "A two-dimensional topic-aspect model for discovering multi-faceted topics," in *Proceedings of the AAAI National Conference on Artificial Intelligence (AAAI)*, Atlanta, Georgia, USA, 2010.

[15] K. Salomatin, Y. Yang, and A. Lad, "Multi-field correlated topic modeling," in *Proceedings of the SIAM International Conference on Data Mining (SDM)*, Sparks, Nevada, USA, 2009.

[16] J. Seiter, O. Amft, M. Rossi, and G. Tr"oster, "Discovery of activity composites using topic models: An analysis of unsupervised methods," *Pervasive and Mobile Computing*, 2014, in press.

[17] J. Ye, R. Janardan, and Q. Li, "Two-dimensional linear discriminant analysis," in *Advances in Neural Information Processing Systems (NIPS)*, 2005, pp. 1569–1576.