Multi-field Correlated Topic Modeling

Konstantin Salomatin
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213
ksalomat@cs.cmu.edu

Yiming Yang
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213
yiming@cs.cmu.edu

Abhimanyu Lad
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213
alad@cs.cmu.edu

Abstract

Popular methods for probabilistic topic modeling like the Latent Dirichlet Allocation (LDA, [1]) and Correlated Topic Models (CTM, [2]) share an important property, i.e., using a common set of topics to model all the data. This property can be too restrictive for modeling complex data entries where multiple fields of heterogeneous data jointly provide rich information about each object or event. We propose a new extension of the CTM method to enable modeling with multi-field topics in a global graphical structure, and a mean-field variational algorithm to allow joint learning of multinomial topic models from discrete data and Gaussian-style topic models for real-valued data. We conducted experiments with both simulated and real data, and observed that the multi-field CTM outperforms a conventional CTM in both likelihood maximization and perplexity reduction. A deeper analysis on the simulated data reveals that the superior performance is the result of successful discovery of the mapping among field-specific topics and observed data.

1 Introduction

There is a great need in practical applications for analyzing and maintaining data collections where each entity (object or event) consists of multiple fields with different but interrelated contents. Consider a computer hardware support department that keeps all the trouble reports for past malfunctions and repairs. Each record may contain several free-text fields, such as a brief problem description by a user, an initial analysis of the problem by a technical specialist, and a detailed technical description by the expert(s) who fixed the problem. Other fields in the record may include related information in the forms of nominal, categorical, ordinal and numerical attributes, such as by whom the problem was reported, what level of urgency was specified, which expert(s) was assigned, what categories of domain expertise were required, how long it took to solve the problem, and so on. When a new case is reported, the engineer assigned to the task would like to find similar cases in the past for reference. Obviously, if our system could make a good use of the multi-field heterogeneous data in combination, then the relevance assessments over past cases would be more accurate than the result of using a standard search engine for retrieval based on a single amorphous text field, e.g., the problem description alone or the content of all fields viewed as a merged one. It is not obvious, however, how to model the dependencies among multiple fields so that the rich connections among tasks can be effectively leveraged. Addressing this open challenge is our primary interest in this paper. Specifically, we propose a new extension of the Correlated Topic Model (CTM) by Blei and Lafferty [2] to allow multiple sets of topics to be used for different fields instead of forcing all the fields to share a common set of topics, and to better capture cross-field dependencies at proper levels of granularity.

Topic modeling methods have received considerable attention in recent machine learning research, among which the LDA (Latent Dirichlet Allocation, [1]), its modifications [4, 5, 6, 8, 9] and CTM [2] are representative of Bayesian graphical models. These methods describe the probabilistic generation process of data. Both LDA and CTM focus on the modeling of data collections with single-field entities, e.g., documents, with a set of hidden variables as the "topics" that explain observed data. LDA has the limitation of being incapable of modeling correlated topics, and CTM addresses that limitation by introducing a logistic normal prior of topics instead of the Dirichlet prior and by using the covariance matrix of the variables in the logistic normal model to capture correlations among topics. While CTM has been successfully applied to single-field data, showing the importance of modeling correlated topics, it is not clear how to apply CTM to multi-field data. A straightforward application is to merge all the contents in different fields of each entity to obtain a synthetic "bag of feature" representation of the entity, and to train CTM in the conventional way on a collection of such field-merged data. We will refer to this way of using CTM as conventional CTM or "single-field CTM" in the rest of the paper, in distinction from the multi-field CTM we develop as an alternative.

We argue that single-field CTM is insufficient for best modeling of multi-field data. Its limitation comes from the design choice, i.e., using the common set of topics to model different data in all the fields. To see why, let us revisit our example of the hardware troubleshooting scenario, and focus on two fields: the problem summaries and the technical resolutions in trouble reports. Both fields contain free-text descriptions; however, the number of topics required to model the latent concepts in user descriptions can greatly differ from the number of topics required to model the latent concepts in expert descriptions (e.g. the former could be much smaller). Moreover, the latent concepts underlying the two fields are related but not necessarily identical. Hence, model the two fields using a common set of topics is an over-simplification of the assumed data generation process. A natural alternative is to allow multiple sets of topics in a unified graphical model where each set of the topics corresponds to a particular data field or a subset of the data fields. In this way of modeling, the generation of multi-field data is broken down into conditionally independent processes, each of which has its own set of topics. Of course the multiple sets of topics can be correlated, and we develop a new extension of the conventional CTM to enable the learning and inference based on the correlated multi-field topics. The conditionally independent nature of data generation for different fields is the main distinction at the concept level of our multi-field CTM from the conventional CTM (and LDA). The letter does not enforce this property; instead, each field would be modeled as if it were generated from all the latent factors, not just from the corresponding set of specific latent factors. This leads to an overly general model that fails to best leverage the rich information in multi-field data, i.e., the tight correspondences between subsets of the topics and the fields.

The main contribution in this paper is a novel extension of the conventional CTM, namely, the multi-field CTM (mf-CTM). It allows multiple sets of topics to be used for different fields. We use logistic-normal distribution to model the correlations among topics, within and across the topic sets. We also develop a variant of the mean-field variational algorithm as the approximation procedure to perform inference and parameter estimation. This procedure is generic and does not restrict the model types for specific fields: these can be Multinomial for textual data, Gaussian for real-valued data or other types of latent-factor models. The effectiveness of the proposed method is evident in our evaluation of this method in comparison with the single-field CTM, on both real and simulated data.

The rest of the paper is organized as the following: Section 2 outlines the conventional CTM as related background. Section 3 proposes the mf-CTM approach as our new extension. Section 4 develops a variant of a mean-field variational algorithm as the approximation procedure for learning mf-CTM and making inferences with the model. Section 5 describes the datasets we prepared for evaluation, including both real and simulated collections of multi-field entities. Section 6 reports our controlled experiments with mf-CTM and the original CTM. Section 7 concludes our findings.

2 Related Background

Let us briefly outline the conventional CTM (Correlated Topic Model) and related notation which is necessary for later description of our new extensions of CTM. Figure 1 illustrates CTM using a standard graphical structure. The circles are random variables or model parameters, and the edges specify probabilistic dependencies (or the conditional independencies) among them; a box is a compact notation of multiple (N, D or K) instances of the variables or parameters. Using a document collection as a concrete example, the symbols are defined as:

- D, the number of documents in the collection
- $d = 1, 2, \ldots, D$, the index of an individual document in the collection
- N, the total count of word occurrences in document d
- $n = 1, 2, \ldots, N$, the index of a word occurrence in document d
- K, the number of hidden variables ("topics") in the model
- $k = 1, 2, \ldots, K$, the index of a hidden variable ("topic")
- $w_{d,n}$ (or w_n), the random variable whose value is the observed word ("feature") at position n in document d, we will sometimes omit index d and use w_n when it is not important to distinguish between documents
- $z_{d,n}$ (or z_d), the random variable whose value is the hidden topic behind $w_{d,n}$
- β_k, the topic-specific word distribution, defining the word emission probabilities for documents in topic k
- $\eta_d \sim N(\mu, \Sigma)$ (or η), a K-dimensional vector, specifying topic priors for each document
- μ and Σ, the parameters (mean and covariance matrix) of a multivariate Gaussian process

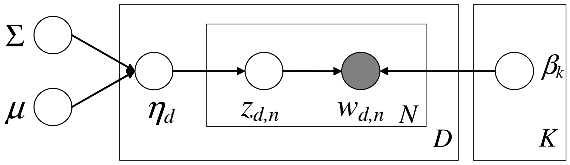

Figure 1: Graphical model representation for the CTM.

The generative process of the CTM is defined as:

A. Draw $\eta|\{\mu, \Sigma\} \sim N(\mu, \Sigma)$

B. Compute the document-specific topic proportions θ using the logistic normal transformation as

$$\theta_i = \exp \eta_i / \sum_j \exp \eta_j \tag{2.1}$$

C. For $n \in \{1, \ldots, N\}$

  a) Draw topic assignment $z_n|\eta$ from $\mathrm{Mult}(\theta)$

  b) Draw feature $w_n|\{z_n, \beta_{1:K}\}$ according to $p(w_n|z_n, \beta)$

According to this model the probability of the document with words w, topic variables η and individual topic assignments z is:

$$p(\eta, z, w|\mu, \Sigma, \beta) = p(\eta|\mu, \Sigma) \prod_{i=1}^{N} p(z_i|\eta) p(w_i|z_i, \beta) \tag{2.2}$$

Notice that only the word-level representation is observed and the topical-level information (η and z) is hidden. To estimate the likelihood of observed words (w) we need to sum out η and z as:

$$p(w|\mu, \Sigma, \beta) = \int_\eta \sum_z p(\eta, z, w|\mu, \Sigma, \beta) d\eta \tag{2.3}$$

As for the whole corpus $\mathbf{w_1}, \ldots \mathbf{w_D}$ where $\mathbf{w_d}$ denotes the word sequence in a document, the likelihood of observing the entire data is:

$$p(Corpus|\mu, \Sigma, \beta) = \prod_{d=1}^{D} p(\mathbf{w_d}|\mu, \Sigma, \beta) \tag{2.4}$$

The expression in (2.3) is intractable due to integration and summation over hidden variables, as shown in [2]. Approximate method (variational approximation, [3], [7]) has been used to estimate the likelihood to perform training and to estimate most likely topic proportions η and topic assignments z (details can be found in [2]).

In the above example we focused on using CTM for modeling text and therefore $\beta_{1:K}$ define K multinomial distributions for modeling word conditioned on topics. Generally speaking, variable $w_{n,d}$ is not restricted to be a word, and the emission probabilities do not need to follow a multinomial process. For example, $w_{n,d}$ can be defined as a Gaussian random variable to model real-valued data.

3 Multi-field Correlated Topic Modeling

We propose two ways to extend the conventional CTM and compare them. One way is to use a common set of topics for all the fields, but allow each field (or each subset of fields) to have its own feature set and feature emission probabilities conditioned on each common topic. Another way is to allow each field (or each subset of fields) to have its own topic set, feature set and feature emission probabilities conditioned on field-specific topics. Briefly, we refer to the former as mf-CTM.ct (for *multi-field CTM with a common topic set*) and the latter as mf-CTM.dt (for *multi-field CTM with different topic sets*)

3.1 Multi-field CTM with a common topic set.

Let us introduce additional notation to support multi-field modeling:

- S, the number of fields in the data entry
- $s = 1, 2, \ldots, S$, the index of an individual field
- β_k^s, feature distribution in the field s conditioned on topic k, defining the feature emission probabilities

In the rest of the paper, we will use the upper index to denote a field. For example, instead of using w_n for the word at position n in document d, we will use w_n^s to specify the word at position n of field s in document d.

Figure 2 illustrates the graphical structure of mf-CTM.ct. Comparing it to the generative process in Figure 1, the only difference is that the word (feature) emission is not only conditioned on a topic, but also conditioned on the field. That is, we use $w_n^s \sim p(w_n^s|z_n^s, \beta^s)$ to replace topic-conditioned word emission probabilities in step C-b) of Section 2.

3.2 Multi-field CTM with different topic sets.

Now we modify the previous model to support multiple sets of topics modeling different fields. This allows us to address the granularity issue and to model the inter-field relations explicitly. To achieve this we divide topic

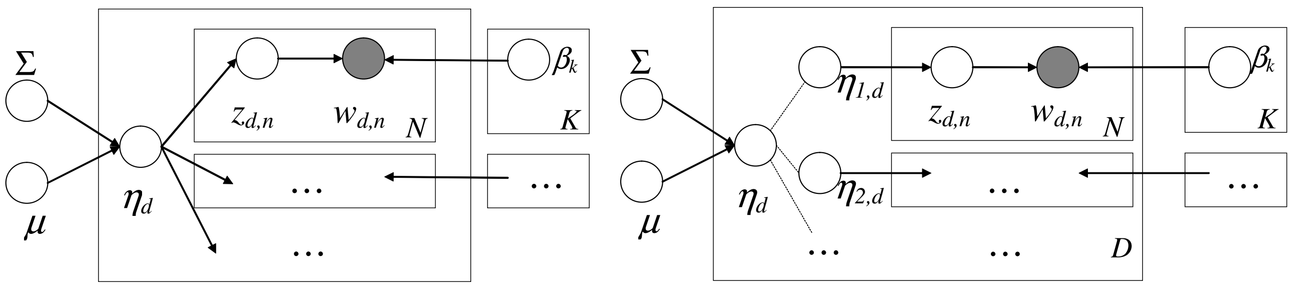

Figure 2: Graphical representation of multi-field CTM with common topic set (on the left) and multi-field CTM with different topic sets (on the right).

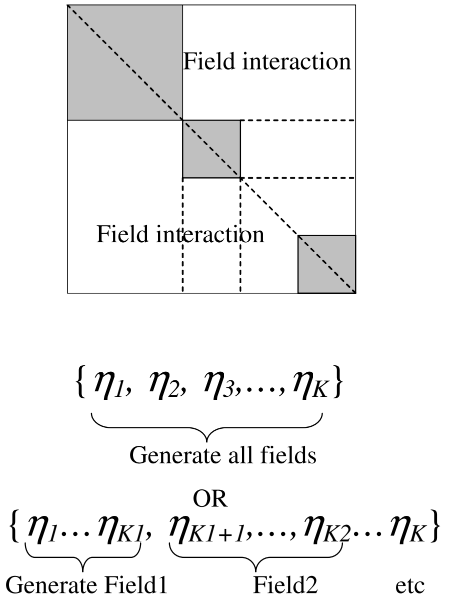

Figure 3: On the top is the topic covariance matrix in the mf-CTM.dt model; on the bottom we compare the vector representations of topic prior in mf-CTM.ct (top) and mf-CTM.dt (bottom).

prior η (a vector) into S disjoint sub-vectors as shown in Figure 3. Each sub-vector corresponds to a field (or a group of fields). The field-joint topic prior is denoted as:

$$\eta = \{\eta^1, \dots \eta^S\} \quad \text{where} \quad \eta^s = \{\eta_{k_s+1}, \dots \eta_{k_{s+1}}\}$$

The mapping from sub-vector η^s to the topic priors θ^s in the corresponding field is defined as:

$$\theta_i = \frac{\exp \eta_i}{\sum_{j=k_s+1}^{k_{s+1}} \exp \eta_j} \tag{3.5}$$

where $i \in \{k_s + 1, \dots k_{s+1}\}$ is the index of a topic in the field. Each sub-vector θ^s lies on a simplex, i.e., its elements are real numbers between zero and one, and these elements sum to one.

The entire data generation process is defined as:

A. Draw $\eta | \{\mu, \Sigma\} \sim N(\mu, \Sigma)$

B. For $s \in \{1, \dots, S\}$ (enumerating the fields)

   a) compute θ^s using transformation (3.5)

   b) for all tokens in the field: $n \in \{1, \dots, N^s\}$

      i. Draw topic assignment $z_n | \eta$ from Mult(θ^s)

      ii. Draw feature $w_n | \{z_n, \beta^s\}$ according to $p(w_n | z_n, \beta^s)$

This gives us an important conditional independence property: the words $\mathbf{w^s}$ of any field s are independent of all the rest words $\mathbf{w^r}$ $(r \neq s)$ given the topical variables η^s. This means that we move the word-level interactions among fields to the topic-level interactions in our model. Denote Σ^i a square diagonal sub-matrix of the covariance matrix Σ (Figure 3) that corresponds to sub-vector η^i, we calculate the joint probability of both hidden topics and observed variables in a multi-field document as:

$$\begin{aligned} p(\eta, z, w | \mu, \Sigma, \beta) &= p(\eta | \mu, \Sigma) \prod_{s=1}^{S} p(z^s, w^s | \eta, \beta) \\ &= p(\eta | \mu, \Sigma) \prod_{s=1}^{S} p(z^s, w^s | \eta^s, \beta^s) \end{aligned} \tag{3.6}$$

Here z^s is a vector of topic assignments of the field s and w^s is a vector of its observable features. Each probability $p(z^s, w^s | \eta^s, \beta^s)$ is the product over all observables as in (2.2) and β^s is a field-specific set of parameters. The posterior likelihood of observing the data given the model is calculated in the same way as before, i.e., using formula (2.3).

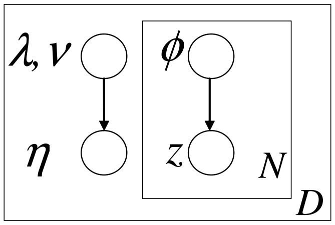

Figure 4: The variational approximation of mf-CTM.dt

4 Mean-field Variational Procedure

For learning the extended CTM models we need the corresponding modification of the mean-field variational approximation method [3] in the conventional CTM. Readers who do not want the computational and mathematical details of the learning and inference procedures can skip this section without loss of context in general. Figure 4 shows a graphical representation of the mean-field approximation for the mf-CTM.dt model. We use the Jensen's inequality to define the lower bound of the likelihood of data in our models as:

$$\begin{aligned}\ln p(w) &= \ln \int_\eta \sum_z p(\eta, z, w) d\eta \\ &= \ln \int_\eta \sum_z \frac{p(\eta, z, w)}{q(\eta, z)} q(\eta, z) d\eta \\ &\geq E_q[\ln p(\eta, z, w)] - E_q[\ln q(\eta, z)] \\ &= E_q[\ln p(\eta, z, w)] + H(q)\end{aligned} \tag{4.7}$$

Here variational distribution q is the approximation of p, and E_q is its expectation and H is its entropy. The mean-field factorization of q is defined as:

$$q(\eta, z|\lambda, \nu^2, \phi) = \prod_{i=1}^{K} q(\eta_i|\lambda_i, \nu_i^2) \prod_{n=1}^{N} q(z_n|\phi_n) \tag{4.8}$$

where $\eta_{1:K}$ are distributed as independent univariate Gaussians (parameterized by $\{\lambda_i, \nu_i^2\}$) and $z_{1:N}$ is distributed according to the K-dimensional vectors $\phi_{1:N}$ that specify the multinomial parameters for each document. Notice that variables η and z are decoupled to make the approximation tractable, and we no longer concern the observable word variable as another part of the simplification. We also no longer consider the non-diagonal covariance elements and the variational parameters are optimized for each field independently. We can rewrite q as $q(\eta, z) = \prod_s q^s(\eta^s, z^s)$ (where q^s is defined on the the s-th field of the model) and substitute q in the right-hand side of (4.7) that yields:

(4.9)

$$\begin{aligned}E_q[\ln p(\eta, z, w)] &= E_q\left[\ln\left(p(\eta)\prod_{s=1}^{S} p(z^s, w^s|\eta^s)\right)\right] \\ &= E_q[\ln p(\eta)] + \sum_{s=1}^{S} E_q[\ln p(z^s, w^s|\eta^s)] \\ &= E_q[\ln p(\eta)] + \sum_{s=1}^{S} E_{q^s}[\ln p(z^s, w^s|\eta^s)] \\ &= E_q[\ln p(\eta)] + \\ \sum_{s=1}^{S} \{E_{q^s}[\ln p(\eta^s)] &- E_q[\ln p(\eta^s)] + E_{q^s}[\ln p(z^s, w^s|\eta^s)]\}\end{aligned}$$

In the last line we add and subtract equal terms. To complete our derivation of the likelihood lower bound we use the fact that the entropy is additive, i.e.

$$H(q) = \sum_{s=1}^{S} H(q^s) \tag{4.10}$$

Substituting the above in formula (4.7) we get the lower bound as:

(4.11)

$$\begin{aligned}L(w) \geq E_q[\ln p(\eta)] &- \sum_{s=1}^{S} E_q[\ln p(\eta^s)] \\ &+ \sum_{s=1}^{S} \Big\{ E_{q^s}[\ln p(\eta^s)] + E_{q^s}[\ln p(z^s, w^s|\eta^s)] \\ &+ H(q^s) \Big\} \\ = \Big\{ E_q[\ln p(\eta)] - \sum_{s=1}^{S} E_q[\ln p(\eta^s)] \Big\} &+ \sum_{s=1}^{S} L(\text{s-th field})\end{aligned}$$

The first part (in the curly parentheses) of the sum represents the cross-field interactions; the second part is the sum of the log-likelihood bounds of individual field-specific models. So far we have not restricted ourselves to any specific form of these field-specific models; in fact, any model with a normal prior over the latent factors such as CTM can be used.

For inference the expression in (4.11) should be optimized with respect to the variational parameters:

$$\begin{aligned}E_q[\log p(\eta|\mu, \Sigma)] &= (1/2)\log|\Sigma^{-1}| \\ -(K/2)\log 2\pi &- (1/2)E_q[(\eta-\mu)^T\Sigma^{-1}(\eta-\mu)]\end{aligned} \tag{4.12}$$

where

$$\begin{aligned}E_q[(\eta-\mu)^T\Sigma^{-1}(\eta-\mu)] &= Tr(diag(\nu^2)\Sigma^{-1}) \\ &+ (\lambda-\mu)^T\Sigma^{-1}(\lambda-\mu)\end{aligned} \tag{4.13}$$

The optimization of field-specific parameters can be done independently for each field using existing procedure for particular models. Estimation of λ and ν involves all the summands simultaneously. However, the update of the procedure is straightforward for gradient methods because the derivatives can be computed independently for all the components (see [2] for the details of optimizing the lower bound in case of CTM).

5 Datasets

We prepared two datasets for evaluation: a real dataset and a simulated data set.

5.1 The collection of aircraft troubleshooting reports. The real dataset consists of roughly 12,000 trouble reports we obtained from Boeing. Those reports describe software and hardware failures and the corresponding troubleshooting events for F/A-18 fielded aircraft maintenance. Each report consists of several free text fields (namely, "summary", "title", "description" and "resolution") and 18 categorical or nominal entries (e.g., those of hardware/software categories, priority levels, locations, engineers, etc.). To test our two-field topic model, we used the "problem summary" of each report as field 1, and we merged the categorical/nominal entries in the report to obtain a synthetic field 2. Each field contains a bag of features: with words as the features in field 1, and with categorical or nominal IDs as the features in field 2. The feature-set size of field 1 is 2841 and the feature-set size of field 2 is 702.

5.2 The simulated dataset. In our simulated data set we control the process of data generation, and use the true model for evaluating automatically learned models and comparing their strengths and weaknesses. We are particularly interested in model comparison regarding the ability to learn the inter-group relationship among topics, or the topics representing different levels of granularity for the underlying concepts of data. More specifically, we generated 2-field simulated data with the following properties:

- Topics of field 1 and topics of field 2 are correlated, each topic of field 1 is likely to co-occur with one or several topics of field 2. Let i be a topic in field 1 and j be a topic in field 2, we have $\sum_{j:i\Rightarrow j} \Pr(j|i) = 1$. Notice that multiple topics in field 1 may imply the same topic in field 2.

- Simulated documents in each field are generated independently and randomly, conditioned on the topic-specific distribution of features which exhibits the power law. Each topic has a randomly assigned power-law slope. We reinforce the power law in feature generation to mimic the importance property of words and class labels distributed over documents in realistic applications.

The procedure for generating the simulated dataset consists of the following steps.

Algorithm 1 Generate topics

```
Specify topic counts $k^1$, $k^2$ and feature-set sizes $N^1$, $N^2$ for fields 1 and 2, respectively
for $i \in \{1, \ldots, k^1\}$ do
    Draw $m_i$ uniformly from $\{2, 3, 4\}$
    Draw $m_i$ topics uniformly from $\{1, \ldots, k^2\}$ as the implied topics: $l_{i,1}, \ldots, l_{i,m_i}$
end for
for $s \in \{1, 2\}$ do
    for $i \in \{1, \ldots, k^s\}$ do
        Draw the power-law slope $\gamma_i^s$ uniformly from interval $[1, 2]$
        Choose $\sigma$, random permutation of $\{1 \ldots N^s\}$; set $\beta^s_{i,\sigma(j)} = -\gamma \log j \Big|_{j=1\ldots N^s}$
    end for
end for
```

Algorithm 2 Generate collection

```
Repeat M times
    Draw $i$ uniformly from $\{1, \ldots, k_1\}$, set $\theta^1_i = 1$
    Draw $\{\theta^2_{l_{i,j}}\}_{j=1}^{m_i}$ from $Dirichlet(\alpha)$
    Draw tokens from the mixture of multinomials defined by matrix $\beta$ and weights $\theta$
```

6 Experiments

We evaluated the three models on the held-out datasets, i.e., the conventional CTM model, and mf-CTM.ct and mf-CTM.dt, respectively. We use the posterior likelihood (defined in (4.7) and calculated using the variational procedure) of each model to measure the performance. We also evaluated the perplexity of each model, measuring how well the model predicts the rest of a multi-field document after observing a part of the document. In our experimental settings the words in field 1 are treated as fully observed; as for field 2, only a fraction of the tokens is observed and the rest is to be predicted. Let X be the observed part of the document and Y be the part to be predicted, the average perplexity over the test set is defined as:

$$\text{Perp} = \left(\prod_{d=1}^{D} \prod_{y \in Y_d} \mathrm{p}(y|X_d) \right)^{\frac{-1}{\sum_{d=1}^{D} |Y_d|}} \tag{6.14}$$

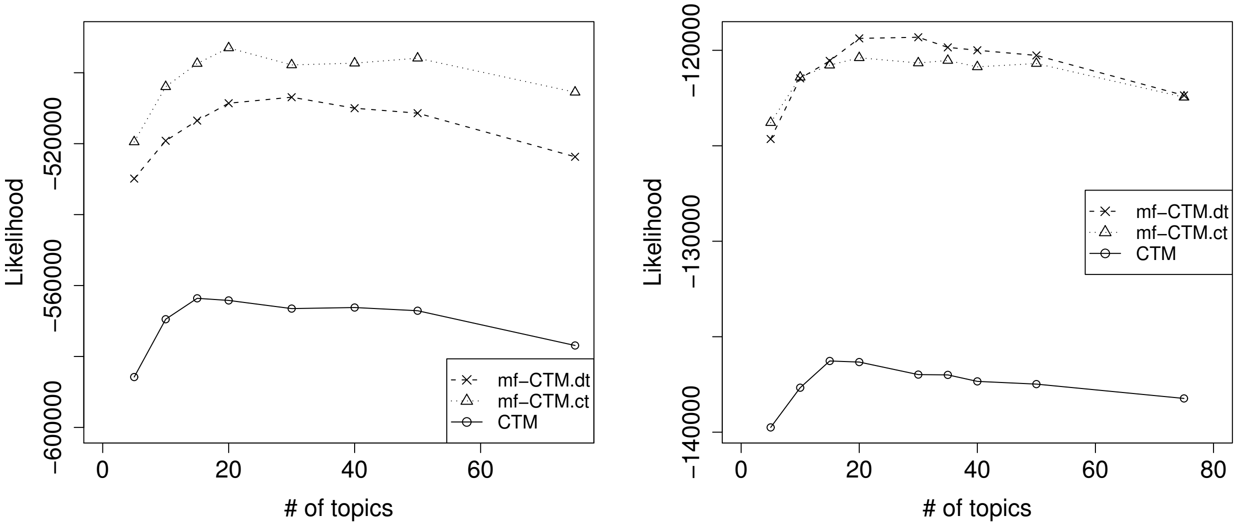

Figure 5: Likelihood of the trouble reports dataset: the training set likelihood on the left and the held-out set likelihood on the right.

The perplexity is essentially the average inverse-likelihood of observing the next token given the partially observed data entry. It differs from the likelihood metric as it emphasizes the prediction power of the model over the output variables instead of focusing on maximizing the likelihood for both the input part and the output part of the data. From a task-oriented evaluation point of view, perplexity is more informative or pertinent than likelihood-based comparison. For completeness we provide our results with both measures.

In addition, we also used the simulated data to analyze the ability of mf-CTM in discovering the structure in data generation, in comparison with the conventional CTM.

6.1 Main Results. Figure 5 shows the performance curves of the three methods measured using the training-set likelihood (in the graph on the left) and the test-set likelihood (in the graph on the right) in response to tuning the number of topics as a parameter in those models. Both the mf-CTM models have significantly better performance than the baseline CTM. This can be explained by the fact that both of the mf-CTM models utilize the multi-filed information in data such as field-specific feature set and field-specific emission probabilities, while the baseline (conventional) CTM ignores such information. As for the performance difference in mf-CTM.ct and mf-CTM.dt, we notice that the former has higher likelihood than the latter on the training data, but we observe the opposite on the held out test-set. This indicates that mf-CTM.ct has an overfitting issue. In other words, mf-CTM.dt is proven to be more robust than mf-CTM.ct because its use of field-specific topic sets to explicitly model inter-field relationship, and with less parameters in the model. The number of parameters is reduced in mf-CTM.dt because we only allow the local features to interact with local topics while in mf-CTM.ct, we allow all the features to interact with all the (common) topics.

Figure 6 shows the perplexity curves (the lower the better) of the three models in response to varying percentage of data being predicted in field 1 and field 2: the larger the proportion, the harder the prediction problem. The number of topics was optimized through cross-validation (Figure 5, Right) using a part of the training data, and for simplicity we chose to use the same number of topics for both field 1 and field 2.[1] The mf-CTM.dt method significantly outperformed mf-CTM.ct and the conventional CTM when a larger portion of field 2 is not observed, i.e., when the prediction problem is harder. It is also consistently better in predicting field 1. It is evident in these results that proper modeling of the multiple fields gives more predictive power to our approach than that of the baseline CTM.

[1] This means that there is a room for improvement if we allow the fields to have different numbers of topics. We leave this kind of fine turning to future exploration.

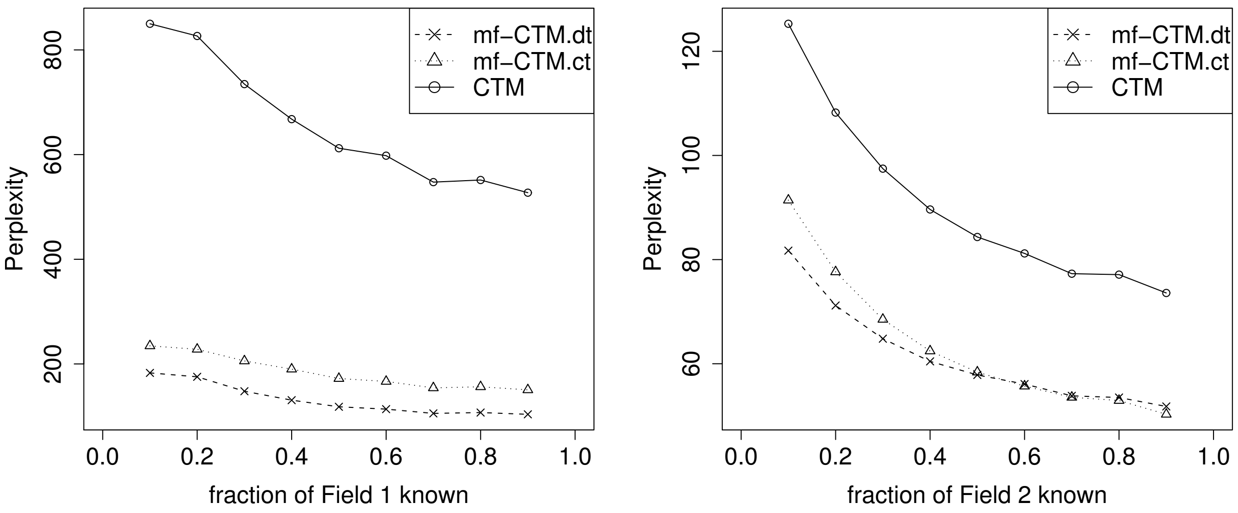

Figure 6: Predictive perplexity for the trouble reports dataset. Prediction of partially observed Field 1 given Field 2 is fully observed on the left and prediction of partially observed Field 2 given Field 1 is fully observed on the right.

In our experiments with the simulated dataset we only focus on the comparison of the conventional CTM and our stronger method, i.e., mf-CMT.dt. In these experiments we have the ground truth of topic numbers, i.e., $k_1 = k_2 = 15$. We trained the baseline CTM with $k = 30$ and the mf-CTM.dt with $k_1 = k_2 = 15$, that is, we purposely made the number of topics to be identical in the two models. Figure 7 shows their perplexity curves; again multi-field CTM significantly outperformed the standard CTM on this dataset.

6.2 Analysis with respect to structure discovery. To understand why and how the conventional CTM and our multi-field CTM (mf-CTM.dt) differ from each other, we analyzed these models regarding their ability to rediscover the true underlying structure in the simulated dataset. Using the held-out data we analyzed the co-occurrence patterns between the true topics and the system-learned topics as the following. Let $\theta_{i,d}$ denote the true weight of topic i given multi-field document d, and $\psi_{i,d}$ denote the learned weight of topic i given document d. For each true topic i and each predicted topic j, we compute a score for the implication from i to j as:

$$\text{Score}(i \to j) = \frac{1}{\sum_{d \in D} \theta_{i,d}} \sum_{d \in D} (\theta_{i,d} \cdot \psi_{j,d}) \tag{6.15}$$

For each node the sum of all outgoing scores is 1:

$$\forall i : \quad \sum_j \text{Score}(i \to j) = 1 \tag{6.16}$$

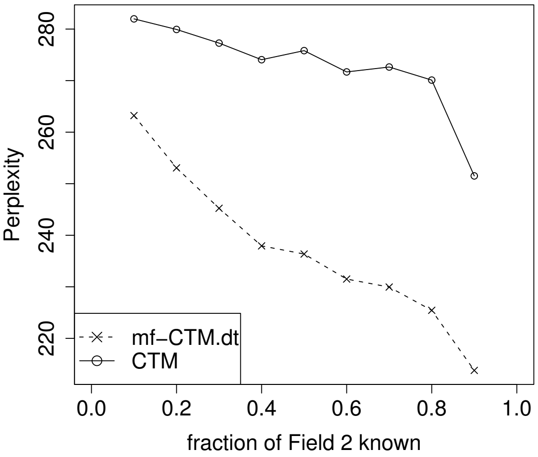

Figure 7: Predictive perplexity curves of different methods on the simulated data.

The co-occurrence based scores define a directed graph with weighed edges that point from true topics to system-predicted topics. To visualize the dominant correspondences in the graph, we need to apply a threshold to the edge weights: the edges with a weight below the threshold will be ignored for visualization; similarly, system-predicted topics without any edge from a true topic after thresholding will be removed from the graph. Figure 8 shows the result after applying of threshold of 0.2. Of course the choice of the threshold and the choice of the heuristic scoring function are subjective; different graphs will correspond to different thresholds. However, our purpose here is to demonstrate qualitatively the behavioral difference between the two methods, by focusing on the dominant connections and ignoring the details.

In this graph, each gray node is a true topic and each white node is a system-predicted topic. The left penal shows the graph structure recovered by the conventional CTM approach and the right penal shows the graph structure recovered by our mf-CTM method. The left-most column of gray nodes in the left/right panel are the true topics in field 1, and the right most columns of gray nodes in the left/right panel are the true topics in field 2. In an ideal situation, i.e., if the true topic structure is fully recovered, we should see each gray node is linked to one and only one white node in the adjacent column, and vice versa. In Figure 8, however, none of the system-predicted topic sets is perfect: two of the true topics in field 1 are not linked to any of the system-predicted topics by the conventional CTM, and one of the true topics in field 2 is not linked to any system-predicted topic by mf-CTM. Nevertheless, we see a more "clean" mapping in the graph structure recovered by mf-CTM for the topics in field 2, compared to that in the graph structure recovered by the conventional CTM; as for the topics in field 1, the two methods are comparable in the structure recovery. This suggests that mf-CTM successfully captured the differences among topics for the data in different fields and at different granularity levels. In other words, multi-field CTM is more powerful in discovering the underlying topical structures in different fields.

7 Concluding Remarks

In this paper we present a new principled solution for a challenging problem in graph structure learning, that is, modeling complex multi-field data and leveraging structural relationships among topics. With novel extensions of CTM and a new variant of the mean-field variation algorithm, our approach enables the joint learning of multi-field topics from both discrete data fields and numerical fields, and the combined use of multinomial models and Gaussian models in a unified framework.

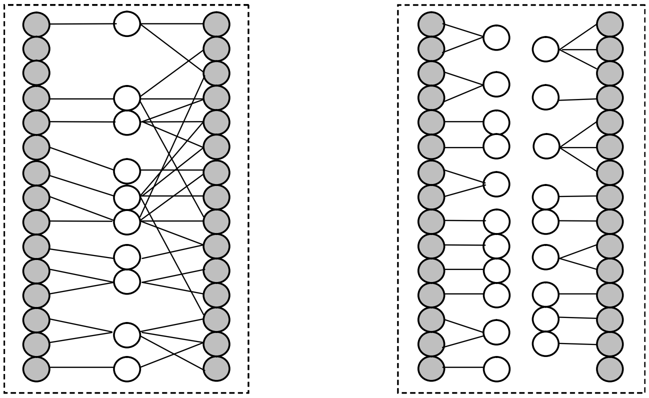

Figure 8: On the left is the graph structure recovered by the conventional CTM and on the right is the graph structure recovered by mf-CTM. Grey circles denote true topics for field 1 and field 2, respectively; white circles denote the predicted topics. The edge directions are omitted for the clarity of the graphs.

The effectiveness of the proposed approach is evident in our experiments on both real data and simulated data. The significant performance improvements of the multi-field CTM over the conventional CTM show the benefits of modeling multi-field topics explicitly, to support topic modeling at various granularity levels and to effectively leverage cross-field dependencies through field-specific topics.

For future work we would like to broaden the scope of our experimentation with various heterogeneous datasets and theoretically and empirically compare a broad set of approaches, including both generative and discriminative methods for multi-field topic modeling.

8 Acknowledgments

This work is supported in parts by the National Science Foundation (NSF) under grant IIS_0704689, and the Defense Advanced Research Project Agency (DARPA) under contract NBCHD030010. Any opinions, findings, conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the sponsors.

References

[1] D. Blei, A. Ng, M. Jordan, *Latent Dirichlet Allocation*, Journal of Machine Learning Research 3 993-1022, 2003.

[2] D. M. Blei and J. D. Lafferty, *Correlated topic models*, in Advances in Neural Information Processing Systems

18 (Y.Weiss, B. Scholkopf and J. Platt, eds.). MIT Press, Cambridge, MA, 2006.

[3] M. Jordan, Z. Ghahramani, T. Jaakkola and L. Saul, *An Introduction to Variational Methods for Graphical Models*, Machine Learning, 37, 183233, 1999.

[4] D. Blei, M.Jordan, *Modeling Annotated Data*, in Proceedings of the 26th annual International ACM SIGIR Conference on Research and Development in Information Retrieval 127134. ACM Press, New York, NY, 2003.

[5] D. Blei, T. Griffiths, M. Jordan, and J. Tenenbaum, *Hierarchical topic models and the nested Chinese restaurant process*, in S. Thrun, L. Saul, and B. Scholkopf, editors, Advances in Neural Information Processing Systems (NIPS) 16. MIT Press, Cambridge, MA, 2004.

[6] A. McCallum, A. Corrada-Emmanuel and X. Wang, *The authorrecipienttopic model for topic and role discovery in social networks: Experiments with Enron and academic email*, Tech. Report, Univ. Massachusetts, Amherst, 2004.

[7] M. Wainwright and M.Jordan, *A variational principle for graphical models*, in New Directions in Statistical Signal Processing, chapter 11. MIT Press, 2005.

[8] M. Steyvers, et al., *Probabilistic author-topic models for information discovery*, in 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 306-315, Seattle, WA, 2004.

[9] X. Wang and A. McCallum., *Topics over time: A non-Markov continuous-time model of topical trends*, in SIGKDD, 2006.