A Non-parametric Approach to Pair-wise Dynamic Topic Correlation Detection

Yang Song[1], Lu Zhang[2], C. Lee Giles[3,1]
[1]Department of Computer Science and Engineering,
[2]Department of Statistics,
[3]College of Information Sciences and Technology,
The Pennsylvania State University,
University Park, PA 16802, U.S.A.

Abstract

We introduce dynamic correlated topic models (DCTM) for analyzing discrete data over time. This model is inspired by the hierarchical Gaussian process latent variable models (GP-LVM). DCTM is essentially a non-linear dimension reduction technique which is capable of (1) detecting topic evolution within a document corpus, (2) discovering topic correlations between document corpora, (3) monitoring topic and correlation trends dynamically. Unlike generative aspect models such like LDA, DCTM demonstrates a much faster converging rate with better model fitting to the data. We empirically assess our approach using 268,231 scientific documents, from the year 1988 to 2005. Posterior inferences suggest that DCTM is useful for capturing topic and correlation dynamics, as well as predicting their trends.

1 Introduction

Topic models have been powerful tools for statistical analysis of text document. As an example, the latent Dirichlet allocation (LDA) model [3] assumes that documents are mixtures of topics, and topics are probability distribution of words, where topics are shared by all documents. The LDA model further assumes the *exchangeability* of words, i.e., words from each document are drawn independently from a mixture of multinomials. The model uses a Dirichlet prior to draw the topic proportions, so that each document may exhibit different topic distributions. LDA is capable of modeling the semantic relations between words and topics, and using multiple topics to describe document collections.

Since most topic models are *generative models*, scalability is always an issue. With a large number of model parameters, the time for the models to converge is prohibitively long. As one example, we applied LDA to over 700,000 full-text scientific documents. The program took more than one week to finish for a 200-component model. Additionally, these models inevitably suffer from the problem of overfitting. As stated in [6], the variational inference for parameter estimation in LDA is problematic, which failed to achieve accurate inference for large data sets.

Moreover, since the LDA model treats words exchangeably, it is not suitable to capture the evolution of documents over time. LDA is also unable to model the topic correlations since it assumes topics are drawn from unique priors. These two issues have been addressed by two extensions of LDA, the dynamic topic models [1] and the correlated topic models (CTM) [2], respectively. Nevertheless, neither of these two models is immune to the aforementioned issues.

In this paper, we present the *dynamic correlated topic models* (DCTM) for analyzing document topics over time. Our model is inspired by the hierarchical Gaussian process latent variable model (HGP-LVM) [5] which has been used for human motion capture. DCTM maps the high-dimensional observed space (words) into low-dimensional latent space (topics), which models the dynamic topic evolution within a corpus. A document corpus considered here is either a conference proceedings or a collection of journal articles. Furthermore, the topic latent space is mapped into a lower-dimensional space which captures the correlations between document corpora. The dynamics of the topics and correlations are captured by a temporal prior, which constructs a hierarchy over the correlation latent space. Unlike generative models, DCTM makes no assumption on word exchangebility. All variables (words, topics and correlations) exhibit dynamics at different time point. By marginalizing out the parameters rather than the latent variables, DCTM becomes a *non-parametric* model with a much faster model convergency rate than the generative processes. The posterior inference of topic and correlation distributions in DCTM is helpful for discovering the dynamic changes of topic-specific word probabilities, and predicting the evolutions of topics and correlations.

2 Related Work

(Correlated Topic Models) An evident limitation of the LDA model attributes to the fact that the topics generated by the multinomial distribution are mutually exclusive. This assumption can be seriously violated in practice. To address this issue, Blei proposed a correlated topic model (CTM) [2], in which the topic proportions are correlated through logistic normal distribution. Mean-field variational methods were employed for parameter estimation. The model was empirically studied by 16,351 *Science* documents over 10 years. A 100-topic CTM shows superior over the traditional LDA model in terms of the predictive performance.

(Dynamic Topic Models) For modeling topic trends over time, Blei developed a time series model, or the dynamic topic models [1], to capture the time evolution of topics in document collections. Rather than using a Dirichlet prior, the dynamic topic model uses a more reasonable Gaussian prior for the topic parameters β, which can capture the evolutions of the topics over the time slices. The topic proportions are drawn from a logistic normal distribution α whose mean values also follow a Gaussian distribution. Two approximate inference methods are developed, namely variational Kalman filtering and wavelet regression. Experiments were performed on a large collection of 30,000 *Science* documents, ranging from 1881 to 1999.

3 Gaussian Process Latent Variable Models

For an introduction to Gaussian Processes, interested readers are suggested to review [7]. In Gaussian process latent variable models (GP-LVM), given a set of n observations $\mathbf{Y} \in \mathbb{R}^{n\times d}$, it seeks a probabilistic approach to non-linear dimension reduction by introducing the latent variables $\mathbf{X} \in \mathbb{R}^{n\times q}$, where $q \ll d$, via a parameterized function

$$Y_{ij} = f(\mathbf{X}_i; \mathbf{W}) + \epsilon_i, \tag{1}$$

where Y_{ij} corresponds to the entry from the i^{th} row and j^{th} column of the matrix $\mathbf{Y}$, $\mathbf{X}_i$ is the i^{th} row of $\mathbf{X}$ with the noise ϵ_i, and $\mathbf{W}$ is the matrix of parameters to be estimated. Traditional non-linear probabilistic approach seeks to maximize the likelihood of the model w.r.t. $\mathbf{W}$ by placing prior distribution $p(\mathbf{X})$ over the latent variables $\mathbf{X}$. Nevertheless, from the Bayesian perspective of view, the parameters $\mathbf{W}$ are trivial and should be marginalized out. Therefore, in GP-LVM, a Gaussian prior is placed on the parameters, i.e., $p(\mathbf{W}) = \prod_{ij} p(w_{ij}) = \prod_{ij} N(w_{ij}|0, 1)$. The marginal likelihood can then be optimized w.r.t. the latent variables ($\mathbf{f}$ being the latent functions)

$$p(\mathbf{Y}|\mathbf{X}) = \int p(\mathbf{Y}|\mathbf{f})p(\mathbf{f}|\mathbf{X})d\mathbf{f}. \tag{2}$$

It has been shown [4] that this model leads to principal component analysis (PCA) given a *linear* covariance function, or a probabilistic non-linear latent variable model given a *non-linear* covariance function. Consequently, the optimized latent variables $\mathbf{X}$ are capable of reducing the original data into a much lower representation.

4 Dynamic Correlated Topic Models

Assume that a set of n document corpora is given, i.e., $\mathbf{D} = \{\mathbf{D}_1, ...\mathbf{D}_n\}$, in which each corpus $\mathbf{D}_i$ contains documents divided into several sets by their timestamps, e.g., the year of publication for scientific documents. We assume that all corpora in our setting share the same timescale, denoted as $[1, ..., T]$, so that each $\mathbf{D}_i = \{\mathbf{D}_{i,1}, ...\mathbf{D}_{i,T}\}$, where $\mathbf{D}_{i,t}$ denotes the set of documents appeared in corpus $\mathbf{D}_i$ at time t. We further assume that a controlled vocabulary with size d is shared across all $\mathbf{D}_i$ over time, so that each $\mathbf{D}_{i,t}$ can be represented into a matrix, $\mathbf{D}_{i,t} \in \mathbb{R}^{N_{i,t}\times d}$, with $N_{i,t}$ denoting the number of document in $\mathbf{D}_i$ at time t. Note that the value of $N_{i,t}$ may vary for different i and t.

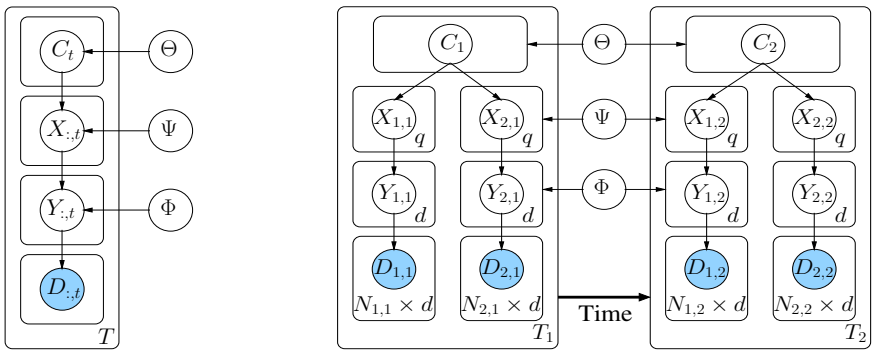
As in most topic models, we also assume that a set of q underlying latent topics exist for each $\mathbf{D}_i$, where the number of topics remain the same over time. In order to model the correlations of those topics over time, we need to first discover the latent topics at time t for each corpus $\mathbf{D}_i$, and specify a proper function for calculating the correlations between topics and corpora. Furthermore, we wish to capture the dynamics of the latent spaces. In what follows, we extend the hierarchical Gaussian process latent variable model (HGP-LVM) [5] for dynamic topic correlation detection.

We first represent each $\mathbf{D}_{i,t}$ into a vector form $\mathbf{Y}_{i,t} \in \mathbb{R}^d$ by aggregating the corresponding features in all instances

$$Y_{i,t}^k = \frac{\sum_{j=1}^{N_{i,t}} (D_{i,t}^{j,k} - \overline{\mathbf{D}_{i,t}^{:,k}})}{\text{var}(\mathbf{D}_{i,t}^{:,k})}, \text{for } k = 1, ..., d, \tag{3}$$

where $Y_{i,t}^k$ is the summarized value of feature k in $\mathbf{D}_{i,t}$, $D_{i,t}^{j,k}$ is the number of times feature k occurred in the j's document of $\mathbf{D}_{i,t}$, $\overline{\mathbf{D}_{i,t}^{:,k}}$ denotes the mean value of feature k and the denominator computes the variance of feature k. In this way we summarize the contributions of individual documents at a certain time and leave only the relationship between words and time.

In the context of textual documents, each $\mathbf{Y}_i = \{\mathbf{Y}_{i,1}, ..., \mathbf{Y}_{i,T}\}$ has the dimensionality of $T \times d$, with each $Y_{i,t}^k$ corresponding to the latent position of word k at time t in $\mathbf{D}_i$, i.e., the position that k appears most probably according to the maximum likelihood estimation. To find q latent topics given $\mathbf{Y} = \{\mathbf{Y}_1, ..., \mathbf{Y}_n\}$, we define n sets of q-dimensional latent variables, with $\mathbf{X}_i = \{\mathbf{X}_{i,1}, ..., \mathbf{X}_{i,T}\} \in \mathbb{R}^{T\times q}, i = 1, ..., n$. We use GP-LVM to model the relations



(a) The DCTM model. (b) An example of DCTM with two corpora over two time frames.

Figure 1. Graphical representation of the DCTM model. Shaded nodes represent observed values.

between each pair of $\mathbf{Y}_i$ and $\mathbf{X}_i$,

$$P(\mathbf{Y}_i|\mathbf{X}_i) = \prod_{j=1}^{d} N(\mathbf{Y}_{i,:}^j|\mathbf{0}, \mathbf{K}_x^{(i)}). \tag{4}$$

Each $\mathbf{Y}_{i,:}^j$ is a size T column vector of $\mathbf{Y}_i$, with each element representing the latent position of word j at different time point. $\mathbf{K}_x^{(i)}$ is a kernel covariance matrix of size $T \times T$, where each element is defined by a kernel function, $[\mathbf{K}_x^{(i)}]_{m,n} = k_x(\mathbf{X}_{i,m}, \mathbf{X}_{i,n})$. In this paper, we use the radial basis function (RBF) kernel

$$k_x(\mathbf{X}_{i,m}, \mathbf{X}_{i,n}) = \phi_1 \exp\left(-\frac{\|\mathbf{X}_{i,m} - \mathbf{X}_{i,n}\|^2}{2\phi_2}\right) + \phi_3\delta_{mn}, \tag{5}$$

with $\Phi = \{\phi_1, \phi_2, \phi_3\}$ being the kernel parameters, where δ_{mn} is the delta function that has the value 1 if $m = n$ and 0 otherwise. Our assumption is that given a topic, words follow a zero-mean Gaussian distribution, where the highest probability occurs when a word appears most in a topic. Note that this zero-mean assumption is valid here since the mean values of word frequency have been extracted from $\mathbf{D}$ during the initialization in eq.(3). To ensure a well-defined probability distribution of topics at each t, we seek to transform the original $\mathbf{X}_i$ using the multiple logistic function

$$P(\tilde{\mathbf{X}}_i|\mathbf{X}_i) = \frac{\exp(\mathbf{X}_{i,:}^j)}{\sum_{j'} \exp(\mathbf{X}_{i,:}^{j'})}, \text{ so that } \sum_j P(\tilde{\mathbf{X}}_{i,t}^j) = 1. \tag{6}$$

In this way the relations between $\mathbf{Y}_i$ and $\mathbf{X}_i$ are rewritten as $P(\mathbf{Y}_i|\mathbf{X}_i) = P(\mathbf{Y}_i|\tilde{\mathbf{X}}_i)P(\tilde{\mathbf{X}}_i|\mathbf{X}_i)$, with $P(\mathbf{Y}_i|\tilde{\mathbf{X}}_i)$ computed using eq.(4).

We then construct a hierarchy by placing a latent variable $\mathbf{C}$ over $\mathbf{X}$, which captures the correlation between each pair of topic sets $\mathbf{X}_i$ and $\mathbf{X}_j$. A proper approach is the Gaussian process where topics that are highly correlated are also close in geometrical interpretation. One approach used in [5] is

Algorithm 1 Parameter Optimization for DCTM

Input: a set of document corpora $\mathbf{D} = \{\mathbf{D}_1, ..., \mathbf{D}_n\}$, number of estimated topics q, number of time frames T, the size of the vocabulary d, initial kernel parameters $\{\Phi, \Theta, \Psi\}$, number of iterations I.
Initialize each $\mathbf{Y}_i \in \mathbb{R}^{T \times d}$ for each $\mathbf{D}_i$ by eq.(3),
Initialize each $\mathbf{X}_i \in \mathbb{R}^{T \times q}$ through SVD from each $\mathbf{Y}_i$,
Initialize each latent correlation variable set $\mathbf{C}$ through SVD for each pair of $[\mathbf{X}_i\ \mathbf{X}_j]$.
for $i = 1$ to I
  for $j = 1$ to n
    optimize each $\{\mathbf{X}_i, \Phi, \Theta\}$ using gradient method
  end for
  for $j = 1$ to n
    optimize $\{\mathbf{C}, \Psi\}$ using the optimized $\mathbf{X}$
  end for
end for

to construct the concatenation of latent variables $[\mathbf{X}_i\ \mathbf{X}_j]$ and find $\mathbf{C}$ by principal component analysis (PCA). This method works well for high-dimensional problems such as video tracking. An alternative is to use singular value decomposition (SVD) where features (words) are usually of equal importance such as in text analysis.

Furthermore, to capture the correlation dynamically, we place a temporal prior over the element of $\mathbf{C}$,

$$P(\mathbf{C}|\mathbf{t}) = \prod_{i=1}^{n} N(\mathbf{c}_{:,i}|\mathbf{0}, \mathbf{K}_t), \tag{7}$$

where $\mathbf{K}_t$ is the covariance matrix for $\mathbf{t} = \{1, ..., T\}$, which takes the exact form as eq.(5) except for the input of $\mathbf{t}$ with a different parameter set $\Theta = \{\theta_1, \theta_2, \theta_3\}$. Fig.1 shows the graphical representation of the general DCTM model.

The temporal prior can be combined with equations to marginalize out latent variables $\mathbf{Y}, \mathbf{X}$ and $\mathbf{C}$. The joint

probability distribution of the hierarchy can be written as

$$\begin{aligned} P(\mathbf{D}_1, ..., \mathbf{D}_n|\mathbf{t}) &= \int P(\mathbf{D}_1|\mathbf{Y}_1)P(\mathbf{Y}_1|\tilde{\mathbf{X}}_1)P(\tilde{\mathbf{X}}_1|\mathbf{X}_1)\cdots \\ &\times \int P(\mathbf{D}_n|\mathbf{Y}_n)P(\mathbf{Y}_n|\tilde{\mathbf{X}}_n)P(\tilde{\mathbf{X}}_n|\mathbf{X}_n)\cdots \\ &\times \int P(\mathbf{X}_1, ..., \mathbf{X}_n|\mathbf{C})P(\mathbf{C}|\mathbf{t}) \\ & \quad \mathrm{d}\mathbf{C}\mathrm{d}\mathbf{X}_1\cdots\mathrm{d}\mathbf{X}_n\mathrm{d}\tilde{\mathbf{X}}_1\cdots\mathrm{d}\tilde{\mathbf{X}}_n. \end{aligned}$$

However, this marginalization is intractable so that we instead attempt to use a maximum a posterior (MAP) approach to approximating the integration, i.e., to maximize the aggregated Gaussian process log likelihoods [5]

$$\begin{aligned} & \mathcal{L}(\mathbf{D}) \\ \triangleq\ & \log P(\mathbf{D}_1, ..., \mathbf{D}_n|\mathbf{t}) \\ =\ & \sum_{m=1}^{n}(\log P(\mathbf{D}_m|\mathbf{Y}_m) + \log P(\mathbf{Y}_m|\tilde{\mathbf{X}}_m) + \\ & \log P(\tilde{\mathbf{X}}_m|\mathbf{X}_m)) + \log P(\mathbf{X}_1, ..., \mathbf{X}_n|\mathbf{C}) + \log P(\mathbf{C}|\mathbf{t}) \end{aligned} \tag{8}$$

w.r.t. each $\mathbf{X}_m$ and $\mathbf{C}$. The solution of eq.(8) can be easily found by gradient search methods.

Practically, when optimizing the latent variables and parameters, we seek a fast converging algorithm which also avoids local minimum. To this point, we initialize each latent variable $\mathbf{X}_i$ and $\mathbf{C}$ by using SVD as described in Alg.1. We then minimize $\mathcal{L}$ by optimizing each set of latent variables and their correlations alternatively.

4.1 Smoothing

In eq.(8), $\mathcal{L}_1$ corresponds to the estimation of the learned latent positions, while all terms in $\mathcal{L}_2$ sum up to the MAP estimation of the dynamic correlations. It can be observed that unsmooth correlations usually result in high values which are not desirable. However, due to the effect of summation of $\mathcal{L}_1$ which involves a large number of instances, the value of $\mathcal{L}_2$ is usually underestimated in practice.

Therefore, to encourage smoothness of $\mathcal{L}(\mathbf{D})$ by penalizing the correlations and the positions on the same granularity, we seek to balance the contribution of both terms by raising the dynamics density function to the ratio of their dimensions, $i.e., \pi = d/q$. Thus the terms corresponding to the dynamics are rescaled in eq.(8) [8]:

$$\pi\left(\frac{q}{2}\log|\mathbf{K}_c| + \log|\mathbf{K}_t| - \frac{1}{2}\sum_{i=1}^{q}\mathbf{X}_{:,i}^T\mathbf{K}_c\mathbf{X}_{:,i} - \frac{1}{2}\mathbf{C}_{:,i}^T\mathbf{K}_t\mathbf{C}_{:,i}\right), \tag{9}$$

which leads to a simple and balanced learning function for the model. Empirically, this has shown to be effective for Gaussian process-based 3D people tracking [8].

4.2 Inference and Predictions

(Posterior Inference) Since we made an assumption on the conditional distribution of $P(\mathbf{Y}_i|\mathbf{X}_i)$ by eq.(4), the topic-specific word distributions $P(\mathbf{Y}_i|\mathbf{X}_i)$ can not be straightforwardly inferred from the model. Instead, we can make inference on the word-specific topic probabilities, to monitor the change of words over time. First, inference can be made for $P(\mathbf{X}_i|\mathbf{Y}_i)$ by using the Bayes rule,

$$P(\mathbf{X}_i|\mathbf{Y}_i) \propto P(\mathbf{Y}_i|\mathbf{X}_i)P(\mathbf{X}_i), \tag{10}$$

so that we can get the word-specific topic probabilities at a certain time t, $\mathbf{X}_{i,t}$, by marginalizing out all latent variables $\mathbf{X}_i$ except for $\mathbf{X}_{i,t}$ (denoted as $\mathbf{X}_{-i,t}$):

$$\begin{aligned} P(\mathbf{X}_{i,t}|\mathbf{Y}_{i,t}) &= \int P(\mathbf{X}_i|\mathbf{Y}_{i,t})d\mathbf{X}_{-i,t} \\ &\propto \int P(\mathbf{Y}_{i,t}|\mathbf{X}_i)P(\mathbf{X}_i)d\mathbf{X}_{-i,t}, \end{aligned} \tag{11}$$

We use importance sampling to estimate the integral.

(Topic & Correlation Predictions) We show the predictive power of DCTM by proposing two prediction methods, using regression analysis and Gaussian processes.

Besides between-topic correlations, the autocorrelations (AC) within each topic can also be computed. Specifically, we can model the autocorrelations of a set of topic distributions over time $\mathbf{X}_i = \{\mathbf{X}_{i,1}, ..., \mathbf{X}_{i,T}\}$ by

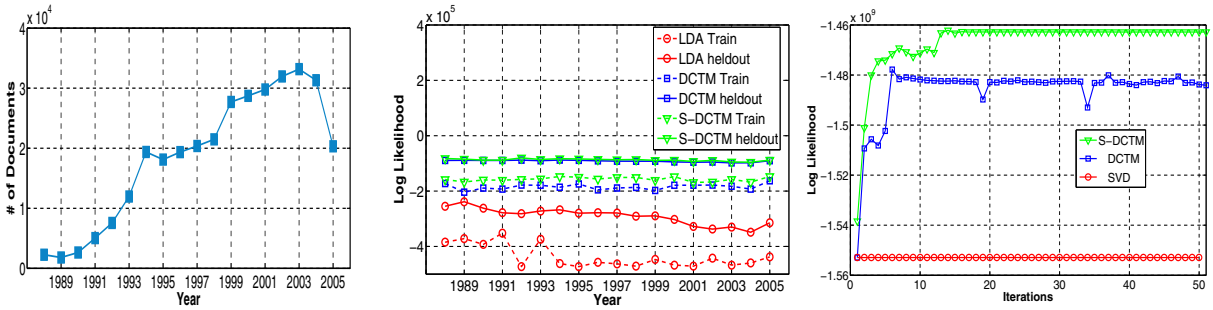
$$\begin{aligned} P(AC(l)|\mathbf{X}_i) &= \frac{\sum_{j=1}^{T-l}(\mathbf{X}_{i,j} - \bar{\mathbf{X}}_i) * (\mathbf{X}_{i,j+l} - \bar{\mathbf{X}}_i)}{\sum_{j=1}^{T}(\mathbf{X}_{i,j} - \bar{\mathbf{X}}_i)^2}, \\ \text{for } l &= 1, ..., T-1, \end{aligned} \tag{12}$$

where $AC(l)$ corresponds to the lag-l autocorrelation function and $\bar{\mathbf{X}}_i$ takes the mean value of $\mathbf{X}_i$. A typical autocorrelation generally decreases with the increase of lag, indicating that only the first few lags demonstrate significantly non-zero. The values of the lags are often used to discover repeating patterns in the data such as the topic distributions during a certain period of time. Mathematically, the values can be used as the coefficient for the regression function.

Meanwhile, due to conjugation, the posterior probabilities of topics and correlations are also Gaussian. We thus propose a simple Gaussian dynamic prediction model [1, 8] for the next time point $t+1$:

$$\begin{aligned} \mathbf{X}_{i,t+1}|\mathbf{X}_{i,t} &\sim N(\mu(\mathbf{X}_{i,t}), \sigma^2(\mathbf{X}_{i,t})\mathbf{I}), \text{ where} \\ \mu(\mathbf{X}_{i,t}) &= \mathbf{K}(\mathbf{X}_i, \mathbf{X}_{i,t})^T\mathbf{K}_X^{-1}\mathbf{X}_i, \\ \sigma^2(\mathbf{X}_{i,t}) &= \mathbf{K}_X(\mathbf{X}_i, \mathbf{X}_i) - \mathbf{K}(\mathbf{X}_i, \mathbf{X}_{i,t})^T \\ & \quad \mathbf{K}_X^{-1}\mathbf{K}(\mathbf{X}_i, \mathbf{X}_{i,t}). \end{aligned} \tag{13}$$

From a standard Gaussian process perspective, making predictions require averaging all parameter values, with



(a) Number of documents per year.(b) Training and held-out log likelihood. (c) Convergency of the log likelihood.

Figure 2. Results of log likelihood on the CiteSeer data set.

their associated posterior weights. However, this approach is computationally demanding which involves expensive Monte Carlo sampling methods. Thus, what we suggested here can be considered as a shortcut of achieving roughly the same predictive power, with less computational cost.

5 Experiments

We analyzed a subset of 268,231 scientific documents (from the year 1988 to 2005) from the CiteSeer[1] digital library. We applied information gain to reduce the dimensionality and resulted in top 24,351 words. We ran a series of experiments on different numbers of topics from 10 to 200. Due to space consideration, we only show the result with 25 topics. To investigate the change of the log likelihood in eq.(8), we split the data into 90% for modeling (training), and use the rest 10% for testing the model with optimized parameters. Figure 2 (b) demonstrates the log likelihood of these two data sets. It is clearly that DCTM shows better fit than LDA for documents across all years. Meanwhile, the smoothing method we used for DCTM (S-DCTM) does show a positive effect on refining the model, by showing higher likelihood than DCTM. It can also be observed that with the increased number of documents by year (Figure 2 (a)), LDA generally shows worse performance with lower likelihood. However, this has minor effect on our models, which supports our argument that DCTM does not suffer from overfitting of large data sets. It can also been seen from Figure 2 (c) that the convergency of DCTM is fast. The log likelihood converges after merely 10 iterations.

Figure 3 presents some results for the SIGMOD corpus. The top figure shows the top 6 venues which have the highest correlations with SIGMOD for each year. It can be observed from the list that most top-ranked venues from the posterior inference are database-related venues. The research trends of SIGMOD can also be observed. While maintaining a steady and strong correlations with traditional database-related venues like ICDE, PODS and VLDB, the correlations of SIGMOD with application-oriented venues are decreasing gradually, e.g., DEXA. Instead, SIGMOD correlates more with data-mining and information-retrieval venues like WWW, AAAI and ICDM (cf middle figure).

The bottom figure depicts two highly-correlated topics in SIGMOD at three different years. The words are sampled from the distribution with probabilities directly computed from the prior. Based on our knowledge, the first topic focuses on *algebra* and *association rules*, with *mining* gradually gets more attention. The second topic addresses *users* and *programming*, which later shifted to *web* applications.

5.1 Prediction Performance

We assessed the predictive powerful of our model. The objective is to predict the correlations between SIGMOD and 5 other venues. We trained our model using data containing the first 16 years (1988–2003), and tested on the year of 2004 and 2005. Both autocorrelation regression (ACR) and mean prediction (MP) are tested. Least square error is applied to measure the performance of the prediction. We also made a simple comparison to the dynamic LDA models [1] by using the variational wavelet regression (VWR). Table 1 lists the results. Both of our methods outperform VWR on all venues.

5.2 Discussion

The comparison of DCTM and LDA did not go through *perplexity* as well as other metrics. This is because these two models differ from each other fundamentally. As explained, our model is able to make inferences on corpus-level correlations, which is a clear advantage over LDA.

The inferences of our model and LDA are also quite different. In LDA, top-ranked words for each topic can be

[1] http://citeseer.ist.psu.edu

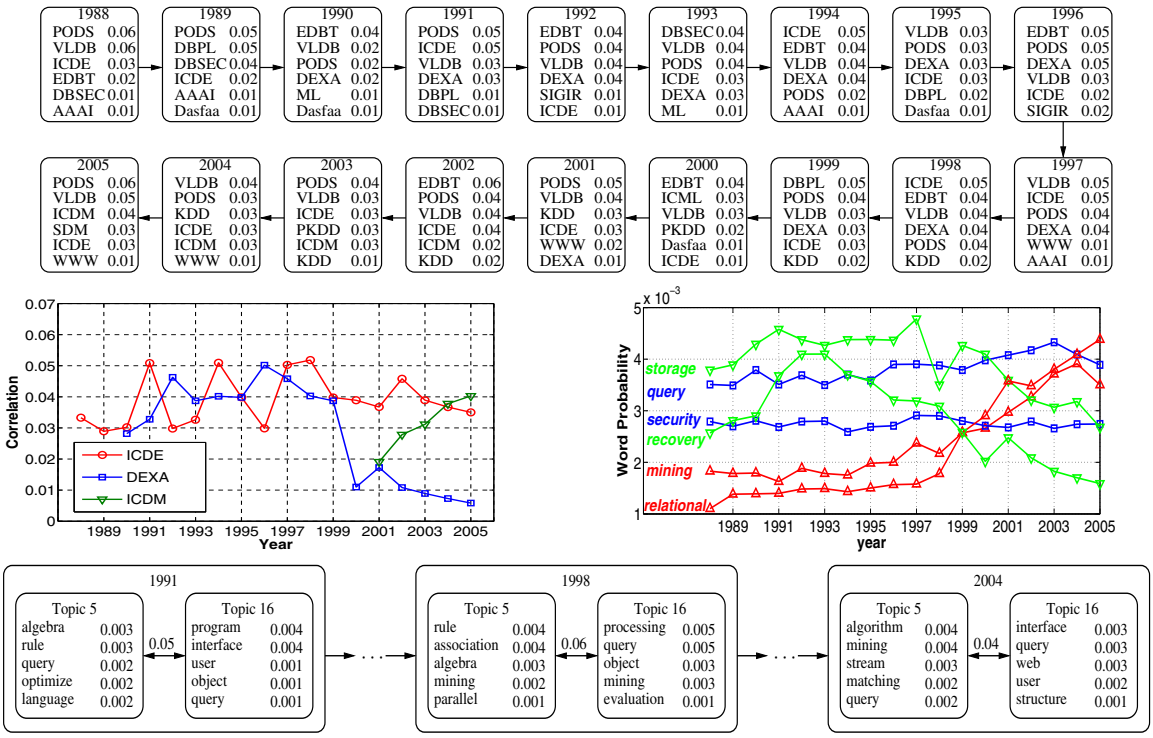


Figure 3. Results of SIGMOD.

Table 1. Correlation results of SIGMOD.

Venue Name	ACR	MP	VWR
AAAI	13.203	10.557	15.625
ICML	45.209	45.317	47.194
KDD	33.004	27.508	34.175
PODS	27.854	24.692	34.215
VLDB	37.225	36.901	45.229
mean	**27.203**	**24.572**	**31.254**

discovered by the posterior inference of topic-specific word probabilities. This is usually used for *naming* topics. However, this approach is very subjective and often requires a good domain knowledge for judgment. Comparatively, our model monitors topic probabilities given a specific word, by marginalizing out the topics at the same time, we can directly observe the popularity of that word at a certain time.

The most controversial part of our model is the initialization step. To minimize the computational cost, we initialized our model by SVD, which is a linear dimensionality reduction method. This method is known to have issues when applied to LSA, though it seems to work well for Gaussian-based models when applied to human motion caption [5]. Besides, due to the restriction of matrix decomposition in SVD, monitoring a large number of topics in a fixed timescale becomes unachievable. The model needs to be re-trained once we change the number of topics.

References

[1] D. M. Blei and J. D. Lafferty. Dynamic topic models. In *ICML '06: Proceedings of the 23rd international conference on Machine learning*, pages 113–120. ACM, 2006.

[2] D. M. Blei and J. D. Lafferty. A correlated topic model of science. NIPS 2007.

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, 2003.

[4] N. D. Lawrence. Gaussian process latent variable models for visualisation of high dimensional data. In S. Thrun, L. Saul, and B. Schölkopf, editors, *Advances in Neural Information Processing Systems 16*. MIT Press, Cambridge, MA, 2004.

[5] N. D. Lawrence and A. J. Moore. Hierarchical gaussian process latent variable models. In *ICML '07*, pages 481–488.

[6] T. Minka and J. Lafferty. Expectation-propagation for the generative aspect model. In *the 18th Conference on Uncertainty in Artificial Intelligence*, pages 352–359, 2002.

[7] C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, 2006.

[8] J. M. Wang, D. J. Fleet, and A. Hertzmann. Gaussian process dynamical models for human motion. *Transactions on Pattern Analysis and Machine Intelligence*, 30(2):283–298.