Chapter 4
Topic Modeling for Speech and Language Processing

Jen-Tzung Chien

Abstract In this chapter, we present state-of-art machine learning approaches for speech and language processing with highlight on topic models for structural learning and temporal modeling from unlabeled sequential patterns. In general, speech and language processing involves extensive knowledge of statistical models. We require designing a flexible, scalable, and robust system to meet heterogeneous and nonstationary environments in the era of big data. This chapter starts from an introduction of unsupervised speech and language processing based on factor analysis and independent component analysis. Unsupervised learning is then generalized to a latent variable model which is known as the topic model. The evolution of topic models from latent semantic analysis to hierarchical Dirichlet process, from non-Bayesian parametric models to Bayesian nonparametric models, and from single-layer model to hierarchical tree model is investigated in an organized fashion. The inference approaches based on variational Bayesian and Gibbs sampling are introduced. We present several case studies on topic modeling for speech and language applications including language model, document model, segmentation model, and summarization model.

4.1 Unsupervised Learning in General

Machine learning is generally categorized into supervised learning and unsupervised learning. Supervised learning aims to find a function mapping from observations to their classes, while the unsupervised learning has a broad goal of extracting salient features and discovering structural information from the given data. In the era of big data, an enormous amount of multimedia data is available in Internet which contains speech, text, image, music, video, social network, and many other specialized technical data. It is challenging to extract reliable features and explore latent structure

J.-T. Chien (✉)
Department of Electrical and Computer Engineering, National Chiao Tung University,
Hsinchu, Taiwan
e-mail: jtchien@nctu.edu.tw

G.W. Peters and T. Matsui (eds.), *Modern Methodology and Applications in Spatial-Temporal Modeling*, JSS Research Series in Statistics,
DOI 10.1007/978-4-431-55339-7_4

from these abundant heterogeneous data which are prone to be noisy, mismatched, mislabeled, misaligned, and ill-posed. In addition, the probabilistic learning models may be improperly assumed, overestimated, or underestimated. The issue of model regularization plays an important role in machine learning.

In general, we need some statistical models or tools for modeling, analyzing, searching, recognizing, and understanding real-world data. Such modeling should faithfully represent the uncertainty in model structure and parameters. The noise condition in observation data should be sufficiently reflected. The learning method should be automatic and adaptive to unknown environments and scalable for large amount of data. The uncertainty in heterogeneous data may be expressed by a prior distribution or even a prior process. We aim to construct a learning machine which provides the ways to organize, understand, search, and summarize a large amount of electronic archives automatically. It is attractive to learn such a model in an unsupervised manner which discovers the hidden themes or topics that pervade data collection. This model can be used to annotate any kinds of documents according to their latent themes. With these annotations, we can organize, summarize, search, and predict for future data.

In this chapter, we first survey a series of unsupervised models in Sect. 4.1.1 and address the history and the evolution of different topic models in Sect. 4.1.2. We then focus on topic model based on the latent Dirichlet allocation (LDA) [7] in Sect. 4.1.3. We introduce the inference procedures of LDA including the approximate inference based on variational inference and Gibbs sampling. Section 4.2 addresses the issue of model selection and its solution based on Bayesian nonparametrics (BNP). We briefly survey BNP approaches to topic models including hierarchical Dirichlet process, the nested Dirichlet process and hierarchical Pitman–Yor process in Sect. 4.2. Section 4.3 presents some advances in topic models especially for the applications of speech and language processing including language model in Sect. 4.3.1, document model in Sect. 4.3.2, segmentation model in Sect. 4.3.3, and summarization model in Sect. 4.3.4. Finally, the summary and future direction are provided in Sect. 4.4.

4.1.1 Unsupervised Models

There are many unsupervised learning approaches in the literature which are available to explore latent features of observation data. Principal component analysis (PCA) [30] is known as a statistical procedure that uses an orthogonal transformation to project a set of possibly correlated observation variables $\mathbf{x} \in \mathcal{R}^D$ into a set of linearly uncorrelated variables $\mathbf{z} \in \mathcal{R}^K$ where $K \ll D$. The projected variables are treated as a kind of latent variables which are also called the principal components. The projection is obtained by finding the eigenvalues and the corresponding eigenvectors of the covariance matrix of observation data. The maximal amount of variance is achieved by this linear projection.

Factor analysis (FA) [1] is closely related to PCA but with more domain-specific constraints on the underlying structure. FA uses the regression model for the error terms, while PCA is a descriptive statistical method for the variance. FA incorporates the common factors $\mathbf{z} \in \mathcal{R}^K$ with a factor loading matrix $\mathbf{W} \in \mathcal{R}^{D \times K}$ and a specific factor vector $\boldsymbol{\varepsilon}$ in order to represent the observed data via $\mathbf{x} = \mathbf{W}\mathbf{z} + \boldsymbol{\varepsilon}$. FA is seen as a latent variable model owing to the common factors $\mathbf{z}$ which are unseen in unsupervised learning procedure. FA model is constructed by imposing the following conditions. The common factors and specific factors are distributed by the zero-mean Gaussians with $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_K)$ and $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Psi})$, respectively, where $\mathbf{I}_K$ is an $K \times K$ identity matrix and $\boldsymbol{\Psi}$ is an $D \times D$ diagonal matrix. And, two sets of factors are uncorrelated by $\mathbb{E}[\mathbf{z}\boldsymbol{\varepsilon}^T] = 0$. The latent factors account for common variance in the data. Basically, PCA and FA are solved by eigen-analyzing different covariance matrices and accordingly correspond to the second-order approaches where the principal components in PCA and common factors in FA are Gaussian distributed.

Independent component analysis (ICA) [21] and blind source separation find a set of latent components that are non-Gaussian and mutually independent, i.e., a much stronger assumption. ICA assumes that the observation vector $\mathbf{x}$ is mixed from a set of independent components $\mathbf{z}$ by $\mathbf{x} = \mathbf{W}\mathbf{z}$ where $\mathbf{W}$ is an $D \times K$ mixing matrix. ICA discovers the independent components or latent sources by maximizing the statistical independence or non-Gaussianity of the estimated components which can be measured based on the information-theoretic criterion using mutual information [2] and the higher order statistics using kurtosis [28]. The demixing matrix is estimated by optimizing such a contrast function. The iterative learning solution to ICA is obtained accordingly. In general, ICA is known as a higher order approach to explore independent components for unsupervised learning which produces a tighter or stronger clustering than the uncorrelated components in PCA and the uncorrelated factors in FA.

PCA, FA, and ICA have been successfully developed as the unsupervised approaches to explore latent variables for a number of applications in speech and language processing. For example, PCA was employed in the technique called eigenvoice [33] which assumed that the supervector of acoustic parameters lay in a subspace spanned by a few eigenvectors or latent components. Speaker adapted acoustic model was obtained by estimating the coefficients of a linear expansion over the eigenvectors. FA was adopted to explore the common factors from acoustic features and apply them to build the streamed hidden Markov model [17] where the streaming regularity was governed by the correlation between speech features which was inherent in common factors. FA was also applied for subspace-based speech enhancement [16] where the principal subspace and minor subspace were constructed from common factors and partitioned according to the values of eigenvalues so that the representation of noisy speech was improved for estimation of clean speech. In addition, ICA was exploited for speech recognition where an unsupervised learning was performed to compensate the pronunciation variations in acoustic model via an ICA algorithm [12]. More recently, a convex divergence [15] was designed as a contrast function for ICA algorithm which improved the convergence speed for blind source separation of speech and music signals. In general, the unsupervised learning algorithms using PCA, FA, and ICA are useful to identify salient features or mixture sources $\mathbf{z}$ from continuous observations $\mathbf{x}$ based on a whole collection of observation vectors $\mathcal{D} = \{\mathbf{x}_1, \ldots, \mathbf{x}_T\}$.

4.1.2 Evolution of Topic Models

Latent variable model based on a whole set of continuous observation vectors could be extended to the one based on the *groups* of *discrete* observation data. This extension was originally developed to conduct a latent semantic analysis [22] and build a latent topic model using a set of grouped words from different documents $\mathcal{D} = \{\mathbf{w}_1, \ldots, \mathbf{w}_M\}$ where each document $\mathbf{w}_m = \{w_{mn}\}$ is composed of N_m words and each word is from a dictionary of $\mathcal{V}$ words. Topic model is developed as an unsupervised learning approach to discover latent features or semantic topics which are used to index or annotate the observed text documents. The annotations could be applied for information retrieval and many other applications. Beyond text annotations, the acoustic topic model was proposed for audio tag classification where the acoustic characteristics were represented by discrete symbols for estimation of latent acoustic topics [31]. In [35], topic model was developed to conduct audio mixture analysis where the acoustic data in time–frequency domain were treated as a bag of frequencies to find acoustic topics. A bag of spectrograms was created to build the convolutive topic model with shift-invariance property in both time and frequency. In the fields of computer vision [24], topic model was established as a Bayesian hierarchical model for scene classification where the image of a scene was seen as a collection of local regions or a bag of image features. Each image was automatically annotated with the themes determined by using topic model.

Topic models have been widely developed as a powerful tool for data analysis, annotation, regression, and classification. Figure 4.1 briefly illustrates the evolution and history of topic models. The earliest topic model called latent semantic analysis (LSA) was proposed by Deerwester et al. [22] in 1990. LSA was invented for automatic indexing and retrieval through a singular value decomposition (SVD) over a word-by-document matrix. The latent structure of words and documents was explored

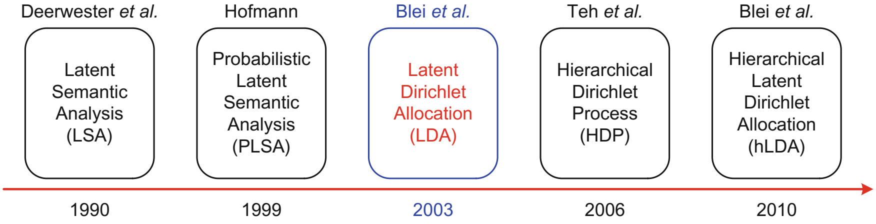
Deerwester et al.
Latent Semantic Analysis (LSA)
1990
Hofmann
Probabilistic Latent Semantic Analysis (PLSA)
1999
Blei et al.
Latent Dirichlet Allocation (LDA)
2003
Teh et al.
Hierarchical Dirichlet Process (HDP)
2006
Blei et al.
Hierarchical Latent Dirichlet Allocation (hLDA)
2010

Fig. 4.1 Evolution and history of topic models

from the decomposed matrices. The next milestone of topic model was achieved by the method called probabilistic latent semantic analysis (PLSA) proposed by Hofmann [27] in 1999. PLSA is a probabilistic framework of LSA where the parameters given latent semantic topics were estimated by maximum likelihood theory using the expectation maximization (EM) algorithm [23]. In 2003, Blei et al. proposed the latent Dirichlet allocation (LDA) [7] for text modeling, document classification, and collaborative filtering. LDA is known as the most popular topic model with the largest citations in the literature. LDA is an extended paradigm from PLSA by introducing a Dirichlet prior to represent the topic probabilities or topic proportions so that the unseen documents could be generalized from Bayesian perspective without greatly increasing the number of parameters. LDA parameters are inferred by maximizing the marginal likelihood over latent topics and topic proportions according to the variational Bayesian (VB) inference [7] and the Gibbs sampling inference [26].

In 2006, Teh et al. proposed the hierarchical Dirichlet process (HDP) [39] which relaxes the constraint of LDA that the number of topics should be known and fixed in topic model. A Bayesian nonparametric (BNP) approach was developed as an expressive probabilistic representation with less assumption-laden approach to inference. The prior process is introduced to conduct a flexible Bayesian learning with infinite topic representation. HDP was implemented by the stick-breaking process and inferred by using the Gibbs sampling procedure. However, topic models based on LDA and HDP assume that topics are independent. To incorporate the topic correlation or even the topic hierarchy into topic model, Blei et al. proposed the nested Chinese restaurant process (nCRP) and built the hierarchical LDA (hLDA) for document representation [3, 4] in 2010. Gibbs sampling was applied to sample a tree path and then sample a tree layer to represent a word w_{mn} in a target document $\mathbf{w}_m$. The tree layers in a tree path reflect different degrees of sharing in the estimated topic parameters. In this chapter, we focus on the topic model based on LDA and its inference procedures using VB-EM algorithm and Gibbs sampling in Sect. 4.1.3. The extensions to HDP and nCRP will be addressed in Sect. 4.2. Some advances in topic model for speech and language processing are described in Sect. 4.3. First of all, we address the early works on topic model based on LSA and PLSA.

Latent Semantic Analysis

Latent semantic analysis (LSA) [22] goes beyond the lexical level from a collection of text documents $\mathcal{D}$ and aims to reveal the latent semantic structure in low-dimensional data space. This algorithm first constructs a word-by-document matrix $\mathbf{W}$ with the element ω_{vm} representing the number of times of a word v occurring in document m. This $\mathcal{V} \times M$ matrix is then decomposed and approximated using the SVD method to produce $\mathbf{W} \approx \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\top$ where $\boldsymbol{\Sigma}$ is an $K \times K$ diagonal matrix with a reduced dimension $K < \min(\mathcal{V}, M)$, $\mathbf{U}$ is an $\mathcal{V} \times K$ matrix whose columns are the first K eigenvectors derived from word-by-word correlation matrix $\mathbf{W}\mathbf{W}^\top$, and $\mathbf{V}$ is an $M \times K$ matrix whose columns are the first K eigenvectors derived from the document-by-document correlation matrix $\mathbf{W}^\top\mathbf{W}$. Each column of $\boldsymbol{\Sigma}\mathbf{V}^\top$ characterizes the location of a particular document in the reduced K-dimensional semantic topic space. Based on this property, we measure the similarity between two documents m and m' by projecting the corresponding document vectors $\mathbf{v}_m$ and $\mathbf{v}_{m'}$ into the semantic topic space as $\boldsymbol{\Sigma}\mathbf{v}_m$ and $\boldsymbol{\Sigma}\mathbf{v}_{m'}$ and then calculating the cosine similarity between two K-dimensional vectors $\cos(\boldsymbol{\Sigma}\mathbf{v}_m, \boldsymbol{\Sigma}\mathbf{v}_{m'})$. Using this similarity, we accordingly conduct the information retrieval by finding the similarity between a query q and a reference document d_m based on $\cos(\boldsymbol{\Sigma}\mathbf{v}_q, \boldsymbol{\Sigma}\mathbf{v}_m)$ where the query vector in semantic topic space is calculated by $\mathbf{v}_q = \boldsymbol{\Sigma}^{-1}\mathbf{U}^\top\boldsymbol{\omega}_q$ using the vector $\boldsymbol{\omega}_q$ consisting of the number of occurrences of different words in query q.

Probabilistic Latent Semantic Analysis

LSA model was established by applying the SVD method which minimizes the approximation error by using the decomposed matrices. LSA is seen as a nonparametric method where there is no probabilistic distribution assumed in this topic model. The system performance and model generalization are constrained. Hofmann [27] introduced a probabilistic solution to LSA based on maximum likelihood (ML) theory. Figure 4.2 shows the graphical representation of the probabilistic LSA (PLSA). PLSA is seen as an aspect model which represents the co-occurrence data of words (denoted by w_n) and documents (denoted by d_m) associated with a topic or latent variable $z_n = k$. The generative model for co-occurrence w_n and d_m is expressed by the joint probability $p(w_n, d_m)$. Under this latent variable model, the joint likelihood function of training data $\mathcal{D} = \{w_n, d_m\}$ is formed by

$$p(\mathcal{D}|\Theta) = \prod_{m=1}^{M}\prod_{n=1}^{N_m}\sum_{k=1}^{K} p(w_n|z_n = k)p(z_n = k|d_m)p(d_m) \tag{4.1}$$

where PLSA parameters $\Theta = \{p(w_n = v|z_n = k), p(z_n = k|d_m)\}$ consist of two sets of topic-based multinomials with the number of parameters given by $\mathcal{V}K + KM$. ML estimation of PLSA parameters is performed by maximizing Eq. (4.1) with respect to Θ. However, such ML estimation suffers from the incomplete data problem due to the missing variable $z_n = k$ or simply z_k. EM algorithm is applied to resolve this problem by alternatively and iteratively performing the E step which calculates the auxiliary function $Q(\Theta'|\Theta) = \mathbb{E}_{(Z)}[\log p(\mathcal{D}, Z|\Theta')|\mathcal{D}, \Theta]$ and then the M step which maximizes $Q(\Theta'|\Theta)$ with respect to Θ'. Here, the auxiliary function $Q(\Theta'|\Theta)$ is calculated as an expectation of log likelihood function using new parameter estimate Θ' given the current estimate Θ. The expectation is performed over latent variables $Z = \{z_k\}$. After EM iterations, ML PLSA parameters are converged at the mode $\hat{\Theta}$.

By expanding the joint probability $p(w_v, d_m)$ where w_v implies $w_n = v$, we may bridge the connection between PLSA and LSA by defining $\mathbf{U} = \{p(w_v|z_k)\}_{v,k}$, $\mathbf{V} = \{p(d_m|z_k)\}_{m,k}$ and $\boldsymbol{\Sigma} = \text{diag}\{p(z_k)\}_k$. And, a matrix with likelihood entries is formed by $\mathbf{P} = \{p(w_v, d_m)\}_{v,m} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\top$. Basically, PLSA assumes that the estimated parameters for different topics are nonnegative, while the elements of the decomposed matrices in LSA, estimated from the eigen-analysis, are not guaranteed to be nonnegative. LSA may violate the nonnegative nature of word count. In addition, the Dirichlet priors for multinomial parameters $\{p(w_v|z_k)\}$ and $\{p(z_k|d_m)\}$ were introduced to conduct the maximum a posteriori (MAP) estimation with constraints $\sum_v p(w_v|z_k) = 1$ and $\sum_k p(z_k|d_m) = 1$. MAP PLSA model was developed for an adaptive topic model which adapted the PLSA parameters to fit the topic-changing domains [18].

4.1.3 Latent Dirichlet Allocation

There are three issues in PLSA topic model. First, the PLSA parameters estimated by ML theory are prone to be overtrained. Model generalization is not assured. Second, PLSA could not model the unseen documents. Third, the number of parameters is proportionally increased by the number of topics K and the number of documents M. To overcome these issues, latent Dirichlet allocation (LDA) [7] was proposed by introducing a Dirichlet prior with hyperparameters $\boldsymbol{\alpha}$ for document-dependent topic proportions $\boldsymbol{\theta}_m = \{p(z_k|d_m)\}$ over K topics as seen in the graphical representation in Fig. 4.2b. Each document is treated as a “random mixture” over latent topics. Topic model is generalized to unseen data through the shared prior distribution $p(\boldsymbol{\theta}_m|\boldsymbol{\alpha})$ with a common hyperparameter $\boldsymbol{\alpha} = \{\alpha_k\}$ where $\alpha_k > 0$. Model construction using LDA is described as follows:

1. For each document $\mathbf{w}_m = \{w_{mn}|n = 1, \ldots, N_m\}$
   a. Draw topic proportions $\boldsymbol{\theta}_m \sim \text{Dir}(\boldsymbol{\alpha})$
   b. For each word w_{mn}
      i. Choose a topic by $z_{mn} = k \sim \text{Mult}(\boldsymbol{\theta}_m)$
      ii. Choose a word by $w_{mn} = v|z_{mn} = k, \boldsymbol{\beta} \sim \text{Mult}(\beta_{vk})$

Here, $\boldsymbol{\beta} = \{\beta_{vk}\} = \{p(w_v|z_k)\}$ denotes the $\mathcal{V} \times K$ multinomial matrix consisting of conditional multinomials β_{vk} for different words under different topics. There are two latent variables in LDA including topic proportions $\boldsymbol{\theta} = \{\theta_{mk}\}$ and topic assignments $\mathbf{z} = \{z_{mn}\}$. LDA parameters $\{\boldsymbol{\alpha}, \boldsymbol{\beta}\}$ are estimated by maximizing the marginal likelihood over two latent variables

$$p(\mathcal{D}|\boldsymbol{\alpha}, \boldsymbol{\beta}) = \prod_{m=1}^{M} \int p(\boldsymbol{\theta}_m|\boldsymbol{\alpha}) \prod_{n=1}^{N_m} \sum_{k=1}^{K} p(z_{mn} = k|\boldsymbol{\theta}_m) p(w_{mn}|z_{mn} = k, \boldsymbol{\beta}) d\boldsymbol{\theta}_m. \tag{4.2}$$

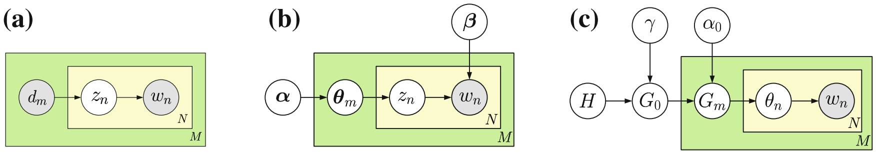

Fig. 4.2 Graphical representation for **a** PLSA, **b** LDA and **c** HDP

We can see that the number of parameters in LDA is $\mathcal{V}K + K$ which is much smaller than $\mathcal{V}K + KM$ for PLSA. A shared $\boldsymbol{\alpha}$ for all documents in LDA can be used to generalized to unseen data and keep a compact model complexity.

However, the exact solution to model inference based on Eq. (4.2) does not exist due to the coupling of multiple latent variables $\boldsymbol{\theta}$ and $\mathbf{z}$ in posterior distribution $p(\boldsymbol{\theta}, \mathbf{z}|\mathcal{D}, \boldsymbol{\alpha}, \boldsymbol{\beta})$. In what follows, we introduce the approximate inference procedures based on variational Bayesian and Gibbs sampling.

Inference by Variational Bayesian

Variational Bayesian (VB) inference is known as the deterministic approach to infer model parameters through a convexity-based variational procedure which is implemented by using the Jensen's inequality. VB aims to resolve the intractable posterior distribution $p(\boldsymbol{\theta}, \mathbf{z}|\mathcal{D}, \boldsymbol{\alpha}, \boldsymbol{\beta})$ by using a factorizable variational distribution

$$q(\boldsymbol{\theta}, \mathbf{z}|\boldsymbol{\gamma}, \boldsymbol{\phi}) = \prod_{m=1}^{M} q(\boldsymbol{\theta}_m|\boldsymbol{\gamma}_m) \prod_{n=1}^{N_m} q(z_{mn}|\phi_{mn}) \tag{4.3}$$

through maximizing a lower bound of the logarithm of marginal likelihood $\mathcal{L}(\boldsymbol{\gamma}, \boldsymbol{\phi}; \boldsymbol{\alpha}, \boldsymbol{\beta})$ where $\boldsymbol{\gamma}$ and $\boldsymbol{\phi}$ denote the variational Dirichlet and multinomial parameters, respectively. We have the relation

$$\log p(\mathcal{D}|\boldsymbol{\alpha}, \boldsymbol{\beta}) = \mathcal{L}(\boldsymbol{\gamma}, \boldsymbol{\phi}; \boldsymbol{\alpha}, \boldsymbol{\beta}) + \mathrm{KL}(q(\boldsymbol{\theta}, \mathbf{z}|\boldsymbol{\gamma}, \boldsymbol{\phi})\|p(\boldsymbol{\theta}, \mathbf{z}|\mathcal{D}, \boldsymbol{\alpha}, \boldsymbol{\beta})). \tag{4.4}$$

Therefore, maximizing the lower bound $\mathcal{L}(\boldsymbol{\gamma}, \boldsymbol{\phi}; \boldsymbol{\alpha}, \boldsymbol{\beta})$ with respect to variational parameters $\{\boldsymbol{\gamma}, \boldsymbol{\phi}\}$ is equivalent to estimating the new variational distribution $q(\boldsymbol{\theta}, \mathbf{z}|\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\phi}})$ which is closest to the true posterior $p(\boldsymbol{\theta}, \mathbf{z}|\mathcal{D}, \boldsymbol{\alpha}, \boldsymbol{\beta})$ with the smallest Kullback–Leibler divergence $\mathrm{KL}(\cdot\|\cdot)$. Basically, finding the approximate posterior distribution $q(\boldsymbol{\theta}, \mathbf{z}|\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\phi}})$ is seen as an expectation step (also called VB-E step) in VB-EM algorithm. Then, in VB-M step, we upgrade the lower bound using the new variational parameters $\mathcal{L}(\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\phi}}; \boldsymbol{\alpha}, \boldsymbol{\beta})$ and maximize the updated lower bound with respect to the model parameters $\{\boldsymbol{\alpha}, \boldsymbol{\beta}\}$ so as to estimate the new LDA parameters $\{\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}\}$. VB-EM algorithm is run to upgrade the variational distribution and increase the lower bound, and accordingly improve the marginal likelihood using the continuously updated model parameters. LDA parameters are finally estimated with convergence after VB-EM iterations as detailed in [7]. Notably, since the Dirichlet distribution in LDA is seen as the conjugate prior for the multinomial likelihood of the observed words, the solutions to variational Dirichlet parameter vector $\hat{\boldsymbol{\gamma}} = \{\hat{\gamma}_k\}$, variational multinomial parameters $\hat{\boldsymbol{\phi}} = \{\hat{\phi}_{nk}\}$, and conditional multinomial distributions $\hat{\boldsymbol{\beta}} = \{\hat{p}(w_v|z_k)\}$ are derived in the closed form. Only the solution to Dirichlet model parameters $\hat{\boldsymbol{\alpha}}$ is calculated by the Newton–Raphson algorithm. Importantly, the variational Dirichlet parameters $\hat{\boldsymbol{\gamma}}$ are seen as the surrogate of the Dirichlet model parameters $\hat{\boldsymbol{\alpha}}$ which sufficiently reflect the topic proportions $\boldsymbol{\theta}$. The variational lower bound $\mathcal{L}(\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\phi}}; \boldsymbol{\alpha}, \boldsymbol{\beta})$ is treated as a tractable surrogate for the intractable log marginal likelihood $\log p(\mathcal{D}|\boldsymbol{\alpha}, \boldsymbol{\beta})$.

Inference by Gibbs Sampling

Griffiths and Steyvers [26] presented a Markov chain Monte Carlo (MCMC) inference solution to LDA topic model. MCMC provides another realization of approximate inference which fulfills the full Bayesian perspective. Different from the deterministic approximation based on VB, MCMC is known as a stochastic approximation. MCMC uses the numerical sampling computation rather than solving the integral and expectation analytically. MCMC provides highly flexible models without limitation of any specific distribution and can be used to infer the infinite-dimensional topic models based on HDP and nCRP which will be addressed in Sect. 4.2. MCMC is computationally expensive without convergence guaranteed. The asymptotically exact solution can be found. However, VB never generates the exact solution but guarantees convergence and fast implementation. The strengths and weaknesses using VB and MCMC are complementary.

Gibbs sampling is a simple and widely applicable realization of MCMC algorithm. Every single state of a Markov chain is seen as an outcome of a latent variable in a variable sequence $\mathbf{z} = \{z_1, \ldots, z_K\}$. Each step of the Gibbs sampling procedure replaces the value for one of the variables z_k by a value drawn from the distribution of that variable conditioned on the values of the remaining states $\mathbf{z}_{-k}$ (i.e., $\mathbf{z} = \{z_k, \mathbf{z}_{-k}\}$) including the preceding states $z_{1:(k-1)}^{(\tau+1)}$ in new iteration $\tau + 1$ and the succeeding states $z_{k+1:K}^{(\tau)}$ in current iteration τ

$$z_k^{(\tau+1)} \sim p\left(z_k \middle| z_{1:(k-1)}^{(\tau+1)}, z_{(k+1):K}^{(\tau)}\right). \tag{4.5}$$

The sampling procedure is repeated with T iterations by cycling through the variables in a particular order or in a random order with some distribution.

Using Gibbs sampling procedure for LDA, we sample the topic assignment z_k according to the predictive posterior distribution $p(z_{mn} = k|\mathbf{z}_{-(mn)}, \mathcal{D})$ given by

$$\begin{aligned} p(w_{mn} &= v|z_{mn} = k, \mathbf{z}_{-(mn)}, \mathbf{w}_{-(mn)})p(z_{mn} = k|\mathbf{z}_{-(mn)}) \\ &= \mathbb{E}[\beta_{vk}|\mathbf{z}_{-(mn)}, \mathbf{w}_{-(mn)}]\, \mathbb{E}[\theta_{mk}|\mathbf{z}_{-(mn)}] \\ &= \frac{\eta+\sum_{m=1}^{M}\sum_{i=1,i\neq n}^{N_m} z_{mi}^k w_{mi}^v}{\mathcal{V}\eta+\sum_{m=1}^{M} N_m-1}\ \frac{\alpha+\sum_{i=1,i\neq n}^{N_m} z_{mi}^k}{K\alpha+N_m-1} \end{aligned} \tag{4.6}$$

where η is the Dirichlet parameter of β_{vk}, $w_{mn} = v$ is expressed by $w_{mn}^v = 1$ and $z_{mn} = k$ is written by $z_{mn}^k = 1$. Here, we use the property of predictive multinomial

$$p(z_k|\mathbf{z}_{-k}) = \int p(z_k|\theta)p(\theta|\mathbf{z}_{-k})d\theta = \mathbb{E}[\theta|\mathbf{z}_{-k}]. \tag{4.7}$$

With a set of samples of topic assignments for different words and documents $\mathbf{z} = \{z_{mn}\}$, we can estimate the multinomial parameters for topics $\hat{\boldsymbol{\theta}} = \{\hat{\theta}_{mk}\}$ and for words under different topics $\hat{\boldsymbol{\beta}} = \{\hat{\beta}_{vk}\}$ by using the expected value of multinomials as given in Eq. (4.7).

4.2 Bayesian Nonparametric Learning

Topic models based on LSA, PLSA, and LDA are constructed as a finite-dimensional mixture representation which assumes that (1) the number of topics is fixed and (2) different topics are independent. These assumptions constrain the flexibility and performance of topic model in presence of scalable data under heterogeneous condition. The topic models based on HDP [39] and nCRP [3, 4] were accordingly developed to resolve these two assumptions through Bayesian nonparametric (BNP) learning. In general, BNPs are used to characterize a big parameter space and construct the probability measure over this space. We setup a stochastic prior process on probability distributions which is a measure on function space. A Bayesian model on an infinite-dimensional parameter space is established. BNPs allow data representation to grow structurally when more data are collected. Number of clusters or topics (or model structure) is unknown a priori. In what follows, we describe BNP learning based on the Dirichlet process and the Pitman–Yor (PY) process. We then introduce the topic models produced by HDP and nCRP and the language model drawn from the hierarchical PY (HPY) process [38].

Dirichlet Process

Dirichlet process (DP) is realized to find the flexible data partitions and provide the nonparametric prior over the number of topics K via a distribution over probability measures $G \sim \text{DP}(\alpha_0, G_0)$ where $\alpha_0 > 0$ is a strength parameter and G_0 is a base measure over a probability space Ω with any partitions $A_1, \ldots, A_k \in \Omega$ as

$$(G(A_1), \ldots, G(A_k)) \sim \text{Dir}(\alpha_0 G_0(A_1), \ldots, \alpha_0 G_0(A_k)) \tag{4.8}$$

which is an infinite-dimensional generalization of Dirichlet distribution. The topic-based representation of a single document $\mathbf{w}$ is formed by drawing the probability measure θ_n for each word w_n using an DP G. The predictive multinomial for new parameter θ_{n+1} in partition A given the previous ones $\theta_{1:n}$ is obtained by Eq. (4.7) as

$$\begin{aligned} p(\theta_{n+1} \in A | \theta_{1:n}, \alpha_0, G_0) = \mathbb{E}[G(A)|\theta_{1:n}] &= \sum_{i=1}^{n} \frac{1}{\alpha_0 + n} \delta_{\theta_i}(A) + \frac{\alpha_0}{\alpha_0 + n} G_0(A) \\ &= \sum_{k=1}^{K} \frac{n_k}{\alpha_0 + n} \delta_{\phi_k}(A) + \frac{\alpha_0}{\alpha_0 + n} G_0(A) \end{aligned} \tag{4.9}$$

where $\phi_1, \ldots, \phi_K$ denote the distinct values from $\theta_{1:n}$. DP can be realized by using the stick-breaking process (SBP) and the Chinese restaurant process (CRP). Equation (4.9) can be explained as a metaphor of CRP with the existing K tables (or clusters). New customer θ_{n+1} enters a restaurant and chooses an occupied table k with probability $\frac{n_k}{\alpha_0 + n}$ or a new table with probability $\frac{\alpha_0}{\alpha_0 + n}$ where n_k denotes the number of customers who have seated in table ϕ_k. On the other hand, using the

SBP, we randomly break a unit-length stick into two segments and find the proportions $\boldsymbol{\pi} = \{\pi_k\} \sim \text{GEM}(\alpha_0)$ with constraint $\sum_k \pi_k = 1$ using the GEM distribution through a process of drawing beta variables $\{\pi'_k\}$. An DP, $G \sim \text{DP}(\alpha_0, G_0)$, is implemented by

$$\phi_k \sim G_0, \quad \pi'_k|\alpha_0 \sim \text{Beta}(1, \alpha_0), \quad \pi_k = \pi'_k \prod_{j=1}^{k-1}(1 - \pi'_j), \quad G = \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k}. \tag{4.10}$$

Pitman–Yor Process

Pitman–Yor (PY) process [34], $\text{PY}(d_0, \alpha_0, G_0)$, is expressed as a three-parameter distribution over distributions where $0 \leq d_0 < 1$ is a discount parameter which characterizes the power-law distribution in natural language, namely *many unique words are observed and most of them rarely*. Basically, d_0 controls the asymptotic growth of the number of unique words, while α_0 controls the overall number of unique words. When $d_0 = 0$, this PY process reverts to $\text{DP}(\alpha_0, G_0)$. When $d_0 > 0$, PY process draws a longer tail probability measure than the DP. Let $G_\emptyset = [G_\emptyset(w)]_{w \in \Omega_v}$ represent the vector of unigrams with empty context $\emptyset$ and $G_0(w) = \frac{1}{\mathcal{V}}$. The predictive unigram probability of a new word w is calculated by

$$\begin{aligned} p(w|\mathcal{D}, d_0, \alpha_0) &= \sum_{k=1}^{m.} \frac{n_k - d_0}{\alpha_0 + n.} \delta_{\phi_k}(w) + \frac{\alpha_0 + d_0 m.}{\alpha_0 + n.} G_0(w) \\ &= \frac{n_w - d_0 m_w}{\alpha_0 + n.} + \frac{\alpha_0 + d_0 m.}{\alpha_0 + n.} \frac{1}{|\mathcal{V}|} \end{aligned} \tag{4.11}$$

where $n. = \sum_k n_k$ is the total number of customers in different tables, m_w is the number of occupied tables labeled by word w, and $m. = \sum_w m_w$ is calculated over different words. Physical meaning of discounting scheme using d_0 is obvious in both terms of right-hand side of Eq. (4.11). The number of occurrences of the seen words is discounted and distributed for those of the unseen words in case of $n_w = m_w = 0$.

Hierarchical Dirichlet Process

HDP deals with the mixed membership representation for multiple documents or grouped data where each document $\mathbf{w}_m$ is associated with a mixture model which is drawn from an DP by $G_m \sim \text{DP}(\alpha_0, G_0)$. Data in different documents share a global mixture model drawn from a global DP by $G_0 \sim \text{DP}(\gamma, H)$ as seen in Fig. 4.2c. HDP can be expressed by the mixture models with the shared atoms $\{\phi_k\}_{k=1}^{\infty}$ but different weights or proportions $\boldsymbol{\beta} = \{\beta_k\}_{k=1}^{\infty}$ and $\boldsymbol{\pi}_m = \{\pi_{mk}\}_{k=1}^{\infty}$ so that we have $G_0 = \sum_k \beta_k \delta_{\phi_k}$ and $G_m = \sum_k \pi_{mk} \delta_{\phi_k}$ with constraints $\sum_k \beta_k = \sum_k \pi_{mk} = 1$. The HDP topic model is accordingly established through a stick-breaking process based on an GEM distribution

$$\begin{aligned} \boldsymbol{\beta}|\gamma \sim \text{GEM}(\gamma), \quad \boldsymbol{\pi}_m|\alpha_0, \boldsymbol{\beta} \sim \text{DP}(\alpha_0, \boldsymbol{\beta}), \quad z_{mn}|\boldsymbol{\pi}_m \sim \text{Mult}(\boldsymbol{\pi}_m) \\ \phi_k|H \sim H, \qquad w_{mn}|z_{mn}, \{\phi_k\}_{k=1}^{\infty} \sim \text{Mult}(\phi_{z_{mn}}) \end{aligned} \tag{4.12}$$

where the infinite-dimensional topic multinomials $\{\phi_k\}_{k=1}^{\infty}$ are incorporated. Importantly, a two-stage SBP was implemented to connect the relation between the topic proportions for words in corpus level $\boldsymbol{\beta}$ and in document level $\boldsymbol{\pi}_m$ [39].

The Nested Chinese Restaurant Process

The topic model based on LDA assumes that different topics are independent. To relax this restriction, the correlated topic model (CTM) [6] was proposed by introducing a multivariate logistic Gaussian distribution as a prior distribution to replace the Dirichlet prior distribution for topic proportions $\boldsymbol{\theta}$ in Sect. 4.1.3. Logistic Gaussian adopts a softmax transformation to impose the condition of summing the proportions to be one. The non-diagonal elements of the corresponding covariance matrix induce the dependencies between the transformed topic multinomials. However, CTM fixed the number of topics and did not consider the topic hierarchy.

Blei et al. proposed the nested Chinese restaurant process (nCRP) [4] and built the hierarchical LDA [3] to explore different levels of aspects for topic modeling without fixing the model structure. Figure 4.3a depicts an infinitely branching tree structure for nCRP representation of words (denoted by blue circles) and document (denote by yellow rectangle). Thick arrows denote a tree path $\mathbf{c}_m$ drawn from nine words of a document $\mathbf{w}_m$ or d_m. Each word w_{mn} is assigned by a topic parameter ϕ_k at a tree node along $\mathbf{c}_m$ using topic proportions $\boldsymbol{\pi}_m$.

1. For each node k in the infinite tree
   a. Draw a topic parameter $\phi_k | H \sim H$
2. For each document $\mathbf{w}_m = \{w_{mn} | n = 1, \ldots, N_m\}$
   a. Draw a tree path by $\mathbf{c}_m \sim \text{nCRP}(\alpha_0)$
   b. Draw topic proportions over layers of $\mathbf{c}_m$ by a stick-breaking process $\boldsymbol{\pi}_m \sim \text{GEM}(\gamma)$
   c. For each word w_{mn}
      i. Choose a layer or a topic by $z_{mn} = k \sim \boldsymbol{\pi}_m$

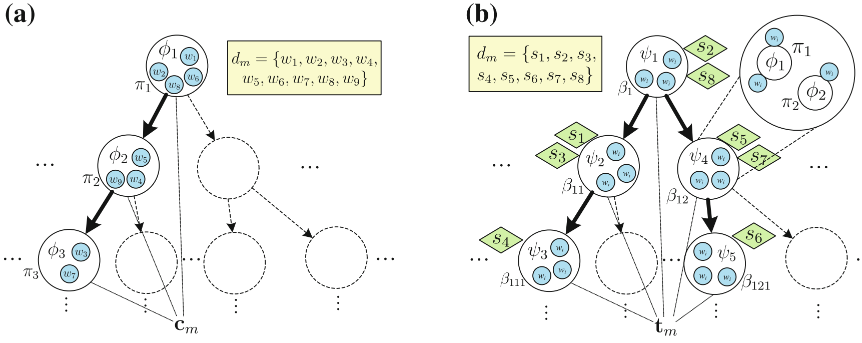

Fig. 4.3 Graphical representation for **a** nCRP and **b** sentence-based nCRP

ii. Choose a word based on topic $z_{mn} = k$ by
$w_{mn}|z_{mn}, \mathbf{c}_m, \{\phi_k\}_{k=1}^{\infty} \sim \text{Mult}\left(\phi_{\mathbf{c}_m(z_{mn})}\right)$

In implementation of nCRP, Gibbs sampling is applied to sample the posterior tree path and word topic $\{\mathbf{c}_m, z_{mn}\}$ for M documents in $\mathcal{D} = \{\mathbf{w}_m\}$ with N_m words in each document according to the individual posterior probabilities of $\mathbf{c}_m$ and z_{mn} given $\mathcal{D}$ and the current values of all the other latent variables, i.e., $p(\mathbf{c}_m|\mathbf{c}_{-m}, \mathcal{D}, \mathbf{z}, \alpha_0, H)$ and $p(z_{mn}|\mathcal{D}, \mathbf{z}_{-(mn)}, \mathbf{c}_m, \gamma, H)$. Again, "−" denotes the self-exception. The tree path $\mathbf{c}_m$ is selected for each customer or document $\mathbf{w}_m$. The tree nodes along $\mathbf{c}_m$ imply a series of visits of this customer to different restaurants in different days. A hierarchical topic model is constructed with different degrees of sharing from root node (broad topic) to leaf nodes (specific topics).

Hierarchical Pitman–Yor Process

Teh [38] presented an BNP learning for language model (LM) to deal with the issue of data sparseness in higher order n-gram model. To cope with this issue, conventional method using the Kneser-Ney (KN) LM smoothing [32] was empirically developed by discounting the number of occurrences for seen n-gram events and distributing these occurrences to unseen n-gram events. Such discounting mechanism reflects the power-law property of natural language and does improve n-gram modeling. Interestingly, KN-LM can be interpreted as a hierarchical Bayesian framework according to the hierarchical Pitman–Yor (HPY) process. Similar to the style of hierarchical generative process based on HDP, HPY process conducts a hierarchical generation of PY processes to draw the discounted n-gram probabilities $p(w_i|w_{i-n+1}^{i-1})$ where the predictive probability of next word $w = w_i$ is based on a history or a context vector consisting of previous $n-1$ words $\mathbf{u} = \{w_{i-n+1}, \ldots, w_{i-1}\} \triangleq w_{i-n+1}^{i-1}$. The HPY process is expressed by a recursive formula where the PY process $G_{\mathbf{u}}$ is formed with a nested base measure $G_{\pi(\mathbf{u})}$ of backoff context $\pi(\mathbf{u})$, which is also an PY process given by a base measure of doubly backoff context $\pi(\pi(\mathbf{u}))$ in a much lower order model. We have

$$G_{\mathbf{u}} \sim \text{PY}(d_{|\mathbf{u}|}, \alpha_{|\mathbf{u}|}, G_{\pi(\mathbf{u})}), \qquad G_{\pi(\mathbf{u})} \sim \text{PY}(d_{|\pi(\mathbf{u})|}, \alpha_{|\pi(\mathbf{u})|}, G_{\pi(\pi(\mathbf{u}))}) \tag{4.13}$$

where the parameters $d_{|\mathbf{u}|}$ and $\alpha_{|\mathbf{u}|}$ depend on the length of context $|\mathbf{u}|$. This is repeated until we reach to PY process for unigram model with empty context $\emptyset$, $G_{\emptyset} \sim \text{PY}(d_0, \alpha_0, G_0)$, as implemented in Eq. (4.11). A kind of linearly interpolated LM (called HPY-LM) is accordingly produced by using the HPY process which combines the mixture probability measures from the higher order statistics in the nth-order model from $\mathbf{u}$ and the lower order LM in the $(n-1)$th-order base measure from backoff context $G_{\pi(\mathbf{u})}$. The combination weights are formed from an PY process mixture model. In Sect. 4.3.1, we will present a new BNP inference procedure for topic-based LM.

4.3 Advanced Topic Models and Their Applications

We have surveyed the fundamental topic models based on the non-Bayesian parametric methods using LSA and PLSA, the Bayesian parametric method using LDA, and the Bayesian nonparametric methods using HDP and nCRP. Model structure has been extended from single-layer model (LSA, PLSA, LDA, HDP) to multiple-layer model (nCRP). Approximate inference algorithms using VB for LDA and Gibbs sampling for LDA, HDP, and nCRP have been addressed. In this section, we will present a series of advanced topic models for different applications including speech recognition, information retrieval, document classification, text segmentation, and document summarization. Here, we categorize these advanced topic models into different information models ranging from language model, document model, segmentation model to summarization model. Going beyond LDA topic model, some other issues are concerned and tackled to achieve flexible, scalable, and robust information systems for real-world applications.

4.3.1 Language Model

Speech recognition system is constructed with two essential models: acoustic model and language model (LM) which considerably affect the system performance. LM provides a prior word probability which characterizes the regularities in natural language. LM is not only useful for speech recognition but also for many other information systems including optical character recognition, spell correction, question answering, automatic summarization, information retrieval, etc. Basically, LM based on n-gram probability $p(w_i|w_{i-n+1}^{i-1})$ is constrained with two weaknesses: (1) lack of training data for higher order LM with large n and (2) lack of long-distance information due to the limitation of n-gram window. To deal with the sparseness of training data, HPY process [38] in Sect. 4.2 was presented to draw the smoothed LM with discounting scheme which was seen as Bayesian interpretation for the heuristic solution based on KN-LM [32]. Considering the issue of long-distance information, the topic-based LMs were proposed by merging the latent semantic information which relaxes the constraint of using short-term lexical information. In [25], PLSA topic model was incorporated into the construction of n-gram model. In addition, the LDA-LM was constructed by employing LDA-based topic information into LM training where the topic prediction was based on the hypothesis of either history words [36] or the words in a whole sentence [37]. In what follows, we introduce the extension of PLSA-LM and LDA-LM to the Dirichlet class LM [13] and the generalization of HPY-LM to the hierarchical Pitman–Yor-Dirichlet LM [9] where the topic models are taken into account.

Dirichlet Class Language Model

The key issue in LDA-LM [36, 37] is that topic information for word prediction is estimated from a set of training documents $\mathcal{D}$ which is treated as a bag of words.

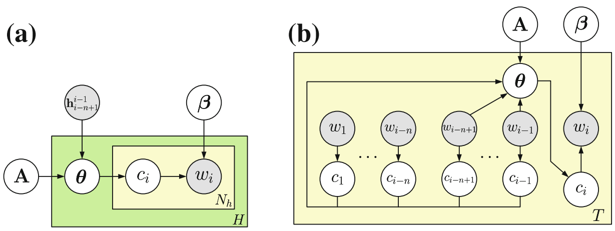

Fig. 4.4 Graphical representation for **a** DC-LM and **b** cache DC-LM

Such estimation did not consider the latent variables based on the *sequential order* of $n-1$ history words $\{w_{i-n+1}, \ldots, w_{i-1}\}$. Such ordering information is crucial for word prediction in natural language. Dirichlet class LM (DC-LM) [13] was proposed to deal with this issue through the representation of history words w_{i-n+1}^{i-1} by concatenating a sequence of $n-1$ history word vectors which are encoded by 1-of-$\mathcal{V}$ coding scheme. An $(n-1)\mathcal{V} \times 1$ supervector $\mathbf{h}_{i-n+1}^{i-1}$ is formed as the surrogate of w_{i-n+1}^{i-1} and then projected into an C-dimensional class space or topic space so that the class proportions are drawn from a Dirichlet prior $\boldsymbol{\theta} \sim \text{Dir}(\mathbf{A}^\top \mathbf{h}_{i-n+1}^{i-1})$. Graphical representation is shown in Fig. 4.4a. Here, the parameter $\mathbf{A} = \{\mathbf{a}_1, \ldots, \mathbf{a}_C\}$ in DC-LM plays a similar role to $\boldsymbol{\alpha}$ in LDA. The other parameters $\boldsymbol{\beta} = \{\beta_{vc}\}$ are seen as the class conditional multinomials for $\mathcal{V}$ words. In a corpus $\mathcal{D}$, there are H histories with N_h words predicted by each history. As a result, DC-LM is calculated by integrating over different classes c_i and proportions $\boldsymbol{\theta}$

$$\begin{aligned} p(w_i|\mathbf{h}_{i-n+1}^{i-1}, \mathbf{A}, \boldsymbol{\beta}) &= \sum_{c_i=1}^{C} p(w_i|c_i, \boldsymbol{\beta}) \int p(\boldsymbol{\theta}|\mathbf{h}_{i-n+1}^{i-1}, \mathbf{A}) p(c_i|\boldsymbol{\theta}) d\boldsymbol{\theta} \\ &= \sum_{c=1}^{C} \beta_{ic} \frac{\mathbf{a}_c^\top \mathbf{h}_{i-n+1}^{i-1}}{\sum_{j=1}^{C} \mathbf{a}_j^\top \mathbf{h}_{i-n+1}^{i-1}}. \end{aligned} \tag{4.14}$$

DC-LM parameters $\{\mathbf{A}, \boldsymbol{\beta}\}$ are estimated according to an VB-EM procedure [13]. DC-LM acts as a new Bayesian class LM which is a smoothed LM over the classes of histories. However, the long-distance information outside n-gram window was not characterized. For this concern, a cache DC-LM was proposed by incorporating the cache memory from all history words w_1^{i-1} into DC-LM as illustrated in Fig. 4.4. We can see that cache DC-LM is calculated through choosing the best class sequence $\hat{c}^{i-1}$ associated with each history word sequence $\hat{w}^{i-1}$. However, DC-LM is constructed with the fixed number of classes or topics C without considering power-law property.

Hierarchical Pitman–Yor-Dirichlet Language Model

Using HPY-LM, the predictive n-gram from $G_{\mathbf{u}}$ is inferred by marginalizing out the prior measure of backoff context $G_{\pi(\mathbf{u})}$. HPY-LM copes with the issue of data sparseness and holds the power-law property of natural language. But, topic information was not captured and accordingly the long-distance information was missed in HPY-LM. In [9], a hierarchical Pitman–Yor-Dirichlet LM (HPYD-LM) was proposed to achieve an BNP learning for the discounted topic-based LM which is seen as a flexible LM with power-law distributions and latent topics where the number of topics is unbounded. An HPYD process is constructed to draw the HPYD-LM. Different from the parametric topic mixture model

$$p(w_i|w_{i-n+1}^{i-1}) = \sum_{k=1}^{K} p(z_i = k|w_{i-n+1}^{i-1}) p(w_i|w_{i-n+1}^{i-1}, z_i = k) \tag{4.15}$$

HPYD process combines a prior process for drawing the topic-dependent smoothed n-gram $p(w_i|w_{i-n+1}^{i-1}, z_i = k)$ from an PY process, and a prior process for topic mixture probability $p(z_i = k|w_{i-n+1}^{i-1})$ from an DP. Starting from the uniform seed measure $H_0(w) = 1/\mathcal{V}$ for all words $w \in \Omega_v$, we draw a word measure from a global topic by $G_0 \sim \text{DP}(\gamma_0, H_0)$. The distribution of topic-dependent unigram $G_{\emptyset z_i}$ with empty context $\emptyset$ and topic assignment z_i is sampled by an PY process $G_{\emptyset z_i} \sim \text{PY}(d_1, \alpha_1, G_0)$ where G_0 is acted as a prior base measure. Next, $G_{\emptyset z_i}$ serves as a base measure for an DP to draw a distribution of unigrams $G_{w_i} \sim \text{DP}(\gamma_1, G_{\emptyset z_i})$. Using G_{w_i} as a prior measure, we draw the distribution of topic-dependent bigrams by using PY process $G_{w_{i-1}z_i} \sim \text{PY}(d_2, \alpha_2, G_{w_i})$. This measure is again acted as a prior basis for an DP to draw the distribution of bigrams $G_{w_{i-1}w_i} \sim \text{DP}(\gamma_2, G_{w_{i-1}z_i})$. Therefore, HPYD process is recursively realized by sampling the distribution of topic-dependent n-grams $p(w_i|w_{i-n+1}^{i-1}, z_i)$ from $G_{w_{i-n+1}^{i-1}z_i}$ and then that of n-grams $p(w_i|w_{i-n+1}^{i-1})$ from $G_{w_{i-n+1}^{i}}$ by

$$G_{w_{i-n+1}^{i-1}z_i} \sim \text{PY}\left(d_n, \alpha_n, G_{w_{i-n+1}^{i-1}}\right), \qquad G_{w_{i-n+1}^{i}} \sim \text{DP}\left(\gamma_n, G_{w_{i-n+1}^{i-1}z_i}\right). \tag{4.16}$$

A hierarchical Chinese restaurant process (HCRP) [9] was designed to implement the HPYD process and infer the HPYD-LM. Imagine that there are Chinese restaurants serving customers with infinite tables, infinite menus, and infinite dishes. For each restaurant with context $\mathbf{u}$, the first customer or word with parameter θ_1 enters the restaurant and chooses the first table in restaurant $\mathbf{u}$. He or she draws a shared menu for all customers seating with the same table and then orders a dish which is labeled by a distinct word $w_{\mathbf{u}1}$. Each customer θ_i only chooses one table and one dish from the single menu corresponding to that table. Each table has its own menu. Following this way, each customer chooses a table with a distinct menu and then draws a dish from that menu. Note that the menus in this HCRP are associated with the topics in HPYD-LM. The menus in restaurant $\mathbf{u}$ are obtained from two information sources: (1) the corresponding menus from the lower order or back off restaurant $\pi(\mathbf{u})$ and (2) the clustering information from the customers in higher order restaurant $\mathbf{u}$. The HPYD n-gram is determined by calculating the predictive or marginal probability of a test word w appearing after a context $\mathbf{u}$ given by a set of training data $\mathcal{D}$. The marginalization is performed over the arrangements of tables $\mathbf{t} = \{t_i, \mathbf{t}_{-i}\}$, menus $\mathbf{z} = \{z_i, \mathbf{z}_{-i}\}$, dishes $\mathbf{l} = \{l_i, \mathbf{l}_{-i}\}$ of all training words $\mathbf{w} = \{w_i, \mathbf{w}_{-i}\}$, and the hyperparameters $\boldsymbol{\lambda} = \{d_m, \alpha_m, \gamma_m | 1 \leq m \leq n\}$. A Gibbs sampling procedure was developed to draw the tables, the menus, and the dishes according to the corresponding posterior probabilities $p(t_i = t|\mathbf{t}_{-i}, \mathbf{z}, \boldsymbol{\lambda}, \mathbf{w}, \mathbf{u})$, $p(z_i = k|\mathbf{z}_{-i}, \mathbf{t}, \boldsymbol{\lambda}, \mathbf{w}, \mathbf{u})$, and $p(l_i = w|\mathbf{l}_{-i}, z_i = k, \mathbf{z}_{-i}, \boldsymbol{\lambda}, \mathbf{w}_{-i}, \mathbf{u})$, respectively [9]. At last, we realize the HPYD process and obtain the HPYD n-gram $p(w_i = w|\mathbf{w}_{-i}, \mathbf{z}, \boldsymbol{\lambda}, \mathbf{u})$.

4.3.2 Document Model

Some other advanced topic models are developed for robust document modeling by compensating the nonstationary condition or conducting the sparse representation.

Dynamic Topic Model

Blei and Lafferty [5] proposed a dynamic topic model (DTM) to analyze the time evolution of topics in a large document collection. The state space models using natural parameters of LDA topic model were implemented to provide a qualitative window over the content of a large data collection. In particular, the topics associated with time slice t evolve from the topics associated with slice $t-1$. Accordingly, the conditional multinomials $\boldsymbol{\beta} = \{\boldsymbol{\beta}_k\}$ and the Dirichlet parameters $\boldsymbol{\alpha}$ are represented by the state space model with the evolution using Gaussians given by the isotropic covariance parameters σ^2 and δ^2

$$\boldsymbol{\beta}_{t,k}|\boldsymbol{\beta}_{t-1,k} \sim \mathcal{N}(\boldsymbol{\beta}_{t-1,k}, \sigma^2 I), \qquad \boldsymbol{\alpha}_t|\boldsymbol{\alpha}_{t-1} \sim \mathcal{N}(\boldsymbol{\alpha}_{t-1}, \delta^2 I). \tag{4.17}$$

Such time-dependent continuous variables are converted into the proportion variables to draw topics $\{z_{mnt}\}$ using $\boldsymbol{\alpha}_t$ and choose the corresponding words $\{w_{mnt}\}$ using $\boldsymbol{\beta}_{t,z}$ for each time slice t. DTM is an extension of LDA to meet nonstationary condition and has been successfully applied to analyze the evolution of topic words in the journal *Science* over 120 years [5].

Sparse Topic Model

The real-world text documents are usually contaminated with noises and redundancies. Sparse representation is helpful to establish a compact model which is robust to adverse conditions. Recently, a sparse Bayesian learning was introduced to perform sparse document representation using the sparse LDA (sLDA) [11]. Previous topic model based on LDA assumes that all of K topics are fully connected to each word w_{mn} in a document. The sLDA topic model aims to select salient features in LDA network by incorporating the spike-and-slab priors [29] into a Bayesian framework. The spike distribution is used to select salient features, while the slab distribution is applied to establish topic model based on the selected relevant topics. As addressed in Sect. 4.1.3, the connections between topics and words in LDA network are sufficiently reflected by the variational multinomial parameters $\{\hat{\phi}_{nk}\}$ which are introduced as the hyperparameters of the variational distributions of latent variables $\{z_{mn} = k\}$.

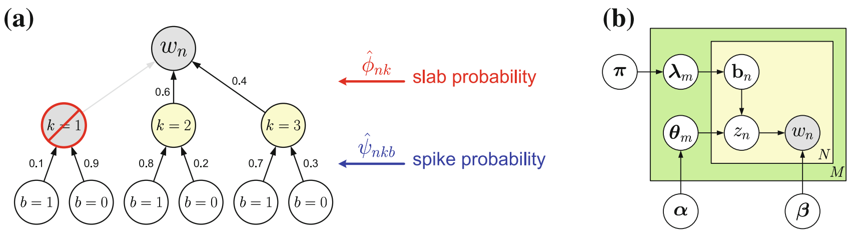

Fig. 4.5 **a** Illustration for feature selection using spike-and-slab priors. **b** Graphical representation for sparse LDA

Such connection is used to select salient features or topics for document representation. Figure 4.5a illustrates the feature selection using spike-and-slab priors. The variational parameter $\hat{\phi}_{nk}$ is treated as a slab probability which connects the representation of a target word w_{mn} using the relevant topics (here $k = 2$ and $k = 3$). This judgment is made from an indicator $b_{nmk} \sim \text{Bern}(\lambda_{mk})$ using a Bernoulli parameter drawn from a beta distribution $\lambda_{mk} \sim \text{Beta}(\boldsymbol{\pi})$. A word w_{mn} is chosen using the conditional multinomial where only the relevant topic k with $b_{nmk} = 1$ is merged, namely

$$w_{mn} = v | b_{nmk} = 1, z_{mn} = k \sim \text{Mult}(\beta_{vk}). \tag{4.18}$$

Graphical representation of sLDA is shown in Fig. 4.5b. An VB-EM procedure was developed to infer the sLDA parameters $\{\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\pi}\}$ by maximizing the marginal likelihood $p(\mathcal{D}|\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\pi})$ over four latent variables $\{\mathbf{z}, \boldsymbol{\theta}, \mathbf{b}, \boldsymbol{\lambda}\}$. Notably, marginal likelihood is only accumulated for all training samples $\{w_{mn\}}$ connected with their associated topics $z_{nm} = k$ with condition $b_{nmk} = 1$. The variational distributions with parameters $\{\boldsymbol{\phi}, \boldsymbol{\gamma}, \boldsymbol{\psi}, \boldsymbol{\eta}\}$ corresponding to latent variables $\{\mathbf{z}, \boldsymbol{\theta}, \mathbf{b}, \boldsymbol{\lambda}\}$ are estimated by maximizing the variational lower bound. Importantly, the variational binomial parameters $\hat{\boldsymbol{\psi}} = \{\hat{\psi}_{nkb}\}$ for binomial indicators $\mathbf{b} = \{b_{nmk}\}$ are estimated as the spike probabilities for feature selection, while the variational multinomial parameters $\hat{\boldsymbol{\phi}} = \{\hat{\phi}_{nk}\}$ for multinomial topics $\mathbf{z} = \{z_{nm}\}$ are calculated as the slab probabilities to model those selected features. In this illustration, the spike probability for topic $k = 1$ under $b_{nmk} = 1$ is too small to contribute the generation of a target word w_{mn}.

4.3.3 Segmentation Model

Sequential patterns in natural language usually appear without explicit boundaries but with the variations of temporal topics. Text segmentation aims to partition the text data into homogeneous processing units or semantically coherent chunks. This research horizon is crucial for many applications including language modeling, speech recognition, text categorization, retrieval and summarization, and also topic detection and tracking. However, in real world, the observed text stream is constructed by a set of heterogeneous documents, making it difficult to extract homogeneous topics. In what follows, we introduce how LDA topic model is extended to cope with the stream-level segmentation and the document-level segmentation [14]. In stream-level segmentation, the text stream is partitioned into topic-coherent documents. In document-level segmentation, the pseudodocument is further segmented into word-coherent paragraphs. Such a hierarchical segmentation makes it feasible to build a precise topic model to compensate the varying distributions of topics and words in nonstationary conditions. This idea can be applied to conduct automatic transcription for lecture speech where the discussion topics are changed by time. This is similar to the situation that the topics are moving between two concatenated documents.

Topic-Based Stream-Level Segmentation

Segmentation of a text stream can be treated as a task of detecting the boundary of documents according to the similarity between sentences $\mathbf{w}_{t-1}$ and $\mathbf{w}_t$ at each sentence time t which is measured by calculating the cosine distance between the corresponding topic proportions $s(\boldsymbol{\theta}_{t-1}, \boldsymbol{\theta}_t)$. The sentence-dependent topic proportions $\boldsymbol{\theta}_t = \{\theta_{tk}\}$ are determined by using the MAP estimate of variational posteriors $\mathbb{E}[\theta_{tk}|\hat{\gamma}_{tk}]$. We draw a segmentation probability based on the beta distribution using this one-sided contextual similarity, i.e., $\omega \sim \text{Beta}(1 - \varepsilon_t, \varepsilon_t)$ where $\varepsilon_t = s(\boldsymbol{\theta}_{t-1}, \boldsymbol{\theta}_t)$. The segmentation label c for each pair of sentences is then chosen by a binomial distribution $c \sim \text{Bin}(\boldsymbol{\omega})$. The segmentation boundary is detected when $c = 1$, otherwise this sentence is grouped into the previous segment. The number of segments is determined automatically. In this study, contextual topic information plays an important role for stream-level segmentation. In [14], the one-sided contextual similarity $s(\boldsymbol{\theta}_{t-1}, \boldsymbol{\theta}_t)$ was improved by using the two-sided contextual similarity for beta parameter via $\varepsilon_t = \max\{s(\boldsymbol{\theta}_{t-1}, \boldsymbol{\theta}_t), s(\boldsymbol{\theta}_{t-1}, \boldsymbol{\theta}_{t+1})\}$. A smoothed boundary detection was performed. The segmentation error due to the non-topic sentences was alleviated. This stream-level segmentation is performed to compensate the variations of topic distributions $\boldsymbol{\theta}$ in a text stream.

Topic-Based Document-Level Segmentation

Furthermore, the variations of word distributions within a pseudodocument are treated in the document-level segmentation. It is because that the usage of the same words in a natural language system is gradually varied over different paragraphs or segments due to the composition style and document structure. Accordingly, we merge a Markov chain to characterize the dynamics of word distributions in LDA topic model. Figure 4.6a shows graphical representation of the resulting nonstationary LDA. Here, each word w_{mn} or simply w_n is generated due to both topic z_n and segment or state s_n. A left-to-right hidden Markov model topology without state skipping is implemented for document-level segmentation. The model parameters consist of $\{\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{A}\}$ where $\mathbf{A} = \{a_{s_{n-1}s_n}\}$ denotes the state transition probabilities.

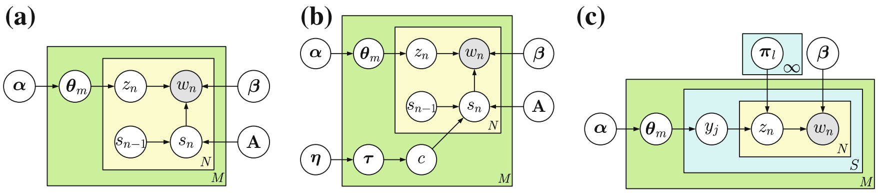

Fig. 4.6 Graphical representation for **a** nonstationary LDA, **b** adaptive and nonstationary LDA, and **c** sentenced-based LDA

Again, the VB-EM algorithm is applied to estimate model parameters $\{\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{A}\}$ by maximizing the marginal likelihood

$$\begin{aligned} p(\mathcal{D}|\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{A}) = \prod_{m=1}^{M} \int p(\boldsymbol{\theta}_m|\boldsymbol{\alpha}) \sum_{s} \prod_{n=1}^{N_m} \sum_{k=1}^{K} p(z_{mn} = k|\boldsymbol{\theta}_m) \\ \times p(w_{mn}|z_{mn} = k, s_{mn}, \boldsymbol{\beta}) p(s_{mn} = s|s_{m,n-1}, \mathbf{A}) d\boldsymbol{\theta}_m. \end{aligned} \tag{4.19}$$

This nonstationary LDA was constructed from spoken documents and merged into n-gram language model. The speech recognition results were rescored for spoken documents [19]. In [20], an adaptive segmentation model was proposed by introducing a style variable c which indicated the number of stylistic changes in a document as depicted in Fig. 4.6b. Style variable is modeled by a multinomial distribution $c \sim \text{Mult}(\boldsymbol{\tau})$ with the style proportions drawn from a Dirichlet prior $\boldsymbol{\tau} \sim \text{Dir}(\boldsymbol{\eta})$. The hybrid stream-level and document-level segmentation was successfully applied for topic detection and tracking in [14].

4.3.4 Summarization Model

Automatic summarization aims to extract the thematic contents or sentences from a large set of documents. A good summary is helpful for browsers to capture the themes and concepts from multiple documents in a very short time. Beyond document representation in word level and document level using LDA, the key issue in a summarization system is to conduct a hierarchical modeling over words, sentences, and documents in a corpus. Given the trained parameters, we can measure the similarity between a document and individual sentence and select the top-ranked sentences according to the Kullback–Leibler (KL) divergence. In a practical system, we usually observe heterogeneous documents where the topics are ambiguous, inconsistent, and diverse. A good summary should reflect the diversity of topics in documents and keep the redundancy to be minimum. In what follows, we survey two advanced topic models for document summarization. One is the parametric model based on the sentence-based LDA [8] and the other one is the nonparametric model based on the sentenced-based nested Chinese restaurant process [10].

Sentence-Based Latent Dirichlet Allocation

A simple extension to allow sentence modeling in LDA topic model is to introduce the sentence-level latent variable $y_j = l$ for each sentence s_j and connect it with the word-level latent variable $z_n = k$ for document representation. Different from the latent topics in word-level representation, we use another related concept called "themes" as the latent variables for sentence-level representation. A sentence-based LDA is constructed as depicted in Fig. 4.6c. Each word $w_n = v$ in sentence s_j $(1 \leq j \leq S)$ and document d_m is drawn by using a word-level multinomial parameter β_{vk} where the latent topic $z_n = k$ is determined by using a theme-dependent topic proportion π_{lk} with latent theme $y_j = l$ $(1 \leq l \leq L)$. This theme is drawn from a document-dependent theme proportion θ_{ml} which is governed by a Dirichlet prior with hyperparameters $\boldsymbol{\alpha} = \{\alpha_l\}$. Notably, each sentence is associated with a latent theme $y_j = l$. Each theme is used to draw the corresponding latent topic $z_n = k$ for representation of a target word $w_n = v$. As a result, document summarization is performed by calculating the KL divergence using the sentence-based unigram $p(w_n|s_j)$ and document-based unigram $p(w_n|d_m)$. The estimated model parameters $\{\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\pi}}\}$ and their variational parameters via VB-EM algorithm are used to calculate these two unigram probabilities $p(w_n|s_j)$ and $p(w_n|d_m)$.

Sentence-Based Nested Chinese Restaurant Process

Similar to what we have discussed in Sect. 4.2 for standard LDA, there are two limitations in the sentence-based LDA which constrain the performance of document representation and summarization. First, the number of themes L and the number of topics K are fixed in advance. Second, different themes l are assumed to be independent while different topics k are independent as well. A sentence-based nCRP was proposed to relax these two assumptions and apply for Bayesian nonparametric document summarization [10]. A metaphor for sentence-based nCRP (snCRP) is displayed in Fig. 4.3b. An infinitely branching tree structure is built for representation of words, sentences, and documents based on an nCRP compound HDP by a two-stage procedure. In the first stage, each sentence s_j of a document d_m is drawn from a document-dependent theme mixture model $G_{s,m}$ via an nCRP. In the second stage, each word w_n of a sentence s_j under a tree node is drawn from a theme-dependent topic mixture model $G_{w,l}$ via an HDP. The probability measures of two models and the relation between the measures of theme ψ_l and topic ϕ_k are expressed by

$$G_{s,m} = \sum_{l=1}^{\infty} \theta_{ml} \delta_{\psi_l}, \qquad G_{w,l} = \sum_{k=1}^{\infty} \pi_{lk} \delta_{\phi_k}, \qquad \psi_l \sim \sum_{k} \pi_{lk} \phi_k. \tag{4.20}$$

Here, the theme proportions $\boldsymbol{\theta}_m = \{\theta_{ml}\}$ and the topic proportions $\boldsymbol{\pi}_l = \{\pi_{lk}\}$ in sentence-based nCRP are similar to those in sentenced-based LDA.

Using this approach, the document-dependent theme mixture model $G_{s,m}$ is established under a sentence-based tree model with atoms $\{\psi_l\}_{l=1}^{\infty}$. Different from the word-based nCRP in Fig. 4.3a using a single tree path $\mathbf{c}_m$ for representation of words $\{w_n\}$ in a document d_m, the sentence-based nCRP in Fig. 4.3b represents the sentences

$\{s_j\}$ of a document d_m based on the theme parameters $\{\psi_l\}$ along the subtree path $\mathbf{t}_m \sim \text{snCRP}(\alpha_0)$. A wide coverage of thematic information in $\mathbf{t}_m$ is beneficial to compensate the thematic uncertainties or variations in the sentences from heterogeneous documents $\mathcal{D}$. Furthermore, the theme-dependent topic mixture model $G_{w,l}$ is constructed by treating the words of the sentences in a tree node l as the grouped data and modeling those grouped data in different tree nodes according to an HDP. The shared atoms $\{\phi_k\}_{k=1}^{\infty}$ are involved. Each word w_n in sentence s_j and document d_m is chosen by a multinomial distribution with parameter $\phi_{\mathbf{t}_m(y_j,z_n)}$ which is selected from the parameter of topic $z_n = k$ under a tree node of theme $y_j = l$ from a subtree path $\mathbf{t}_m$. The topic k and theme l are drawn from the topic proportions $\boldsymbol{\pi}$ and theme proportions $\boldsymbol{\theta}$, respectively. Importantly, the theme-dependent topic proportions are drawn by an GEM distribution $\boldsymbol{\pi}_l|\gamma_w \sim \text{GEM}(\gamma_w)$ using a word-level strength parameter γ_w through a stick-breaking processing (SBP). The document-dependent theme proportions are chosen by a treeGEM distribution $\boldsymbol{\theta}_m|\gamma_s \sim tree\text{GEM}(\gamma_s)$ using a sentence-level parameter γ_s through a tree SBP. In [10], a Gibbs sampling was developed to sample a document-dependent subtree branches $\mathbf{t}_m = \{t_{mj}\}$, document-dependent theme labels $\mathbf{y} = \{y_j\}$ and theme-dependent topic labels $\mathbf{z} = \{z_n\}$ according to the posterior probabilities $p(t_{mj} = t|\mathbf{t}_{m(-j)}, \mathcal{D}, \mathbf{y}, \alpha_0)$, $p(y_j = l|d_m, \mathbf{y}_{-j}, \mathbf{t}_m, \gamma_s)$ and $p(z_n = k|\mathcal{D}, \mathbf{z}_{-n}, y_j = l, \gamma_w)$, respectively. A document summarization system was established through a sentence selection procedure over the inferred tree model for sentences.

4.4 Summary and Future Direction

We have presented the theoretical background and surveyed some advances in topic models for speech and language processing. In theoretical background, we started from the general unsupervised learning methods using latent variable models based on FA and ICA and then moved to general topic models for natural language applications. We systematically addressed the evolution of topic models from the parametric models using LSA, PLSA, and LDA to the Bayesian nonparametric models using HDP and nCRP. The inference solutions to LDA based on VB and Gibbs sampling procedures were investigated. The Bayesian nonparametric learning methods via DP, PY process, HDP, and HPY process were introduced. From these theoretical surveys, we would like to move beyond baseline topic model using LDA toward building a flexible, hierarchical, adaptive, and scalable topic model to meet a variety of heterogeneous conditions in real-world information systems.

In the advanced studies, we presented a series of extended topic models which were developed and applied for speech recognition, document retrieval, text segmentation, and document summarization. We discussed different issues in LDA topic model including topic correlation, model complexity, topic structure, model smoothing, power-law property, temporal modeling, overtrained problem, sparse representation, nonstationary condition, and ill-posed condition. A variety of solutions were proposed to achieve finite-dimensional and infinite-dimensional topic-based language models, dynamic and sparse topic-based document models, topic-based stream-level and document-level segmentation, and sentence-based LDA and nCRP summarization models. The HPY compound HDP was developed for topic-based language model, while the nCRP compound HDP was exploited for sentence clustering and hierarchical modeling of words, sentences, and documents.

Some suggestions are provided for future direction. In the era of big data, we build an infinite model from heterogeneous data. We should think more seriously about the problems at hand, systematically extract the latent information, and carefully represent the model variations. We need to take care of some challenging issues including parallel processing in algorithm level as well as in system level, rapid inference algorithm and sequential MCMC algorithm and work on big learning for topic model. It is interesting to discover ubiquitous extensions and connections to the nonnegative matrix factorization, tensor decomposition and deep neural network and apply them to speech recognition, speaker recognition, speech synthesis, music classification, source separation, etc.

References

1. Basilevsky, A.: Statistical Factor Analysis and Related Methods—Theory and Applications. Wiley, New York (1994)
2. Bell, A.J., Sejnowski, T.J.: An information-maximization approach to blind separation and blind deconvolution. Neural Comput. **7**, 1129–1159 (1995)
3. Blei, D., Griffiths, T., Jordan, M., Tenenbaum, J.: Hierarchical topic models and the nested Chinese restaurant process. Adv. Neural Inf. Proc. Syst. **16**, 17–24 (2004)
4. Blei, D.M., Griffiths, T.L., Jordan, M.I.: The nested Chinese restaurant process and Bayesian nonparametric inference of topic hierarchies. J. ACM **57**(2) (2010). Article 7
5. Blei, D.M., Lafferty, J.D.: Dynamic topic models. In: Proceedings of International Conference on Machine Learning, pp. 113–120 (2006)
6. Blei, D.M., Lafferty, J.D.: A correlated topic model of science. Ann. Appl. Stat. **1**(1), 17–35 (2007)
7. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)
8. Chang, Y.L., Chien, J.T.: Latent Dirichlet learning for document summarization. In: Proceedings of International Conference on Acoustics, Speech, and Signal Processing, pp. 1689–1692 (2009)
9. Chien, J.T., Chang, Y.L.: Hierarchical Pitman-Yor and Dirichlet process for language model. In: Proceedings of Annual Conference of International Speech Communication Association, pp. 2212–2216 (2013)
10. Chien, J.T., Chang, Y.L.: Hierarchical theme and topic model for summarization. In: Proceedings of IEEE International Workshop on Machine Learning for Signal Processing, pp. 1–6 (2013)
11. Chien, J.T., Chang, Y.L.: Bayesian sparse topic model. J. Signal Proc. Syst. **74**(3), 375–389 (2014)
12. Chien, J.T., Chen, B.C.: A new independent component analysis for speech recognition and separation. IEEE Trans. Audio Speech Lang. Process. **14**(4), 1245–1254 (2006)
13. Chien, J.T., Chueh, C.H.: Dirichlet class language models for speech recognition. IEEE Trans. Audio, Speech, Lang. Process. **19**(3), 482–495 (2011)

14. Chien, J.T., Chueh, C.H.: Topic-based hierarchical segmentation. IEEE Trans. Audio Speech Lang. Process. **20**(1), 55–66 (2012)
15. Chien, J.T., Hsieh, H.L.: Convex divergence ICA for blind source separation. IEEE Trans. Audio Speech Lang. Process. **20**(1), 290–301 (2012)
16. Chien, J.T., Ting, C.W.: Factor analyzed subspace modeling and selection. IEEE Trans. Audio Speech Lang. Process. **16**(1), 239–248 (2008)
17. Chien, J.T., Ting, C.W.: Acoustic factor analysis for streamed hidden Markov model. IEEE Trans. Audio Speech Lang. Process. **17**(7), 1279–1291 (2009)
18. Chien, J.T., Wu, M.S.: Adaptive Bayesian latent semantic analysis. IEEE Trans. Audio Speech Lang. Process. **16**(1), 198–207 (2008)
19. Chueh, C.H., Chien, J.T.: Nonstationary latent Dirichlet allocation for speech recognition. In: Proceedings of Annual Conference of International Speech Communication Association, pp. 372–375 (2009)
20. Chueh, C.H., Chien, J.T.: Adaptive segment model for spoken document retrieval. In: Proceedings of International Symposium on Chinese Spoken Language Processing, pp. 261–264 (2010)
21. Comon, P.: Independent component analysis, a new concept? Signal Process. **36**(3), 287–314 (1994)
22. Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R.: Indexing by latent semantic analysis. J. Am. Soc. Inf. Sci. **41**(6), 391–407 (1990)
23. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. B **39**(1), 1–38 (1977)
24. Fei-Fei, L., Perona, P.: A Bayesian hierarchical model for learning natural scene categories. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 524–531 (2005)
25. Gildea, D., Hofmann, T.: Topic-based language models using EM. In: Proceedings of European Conference on Speech Communication and Technology, pp. 2167–2170 (1999)
26. Griffiths, T.L., Steyvers, M.: Finding scientific topics. Proc. Nat. Acad. Sci. U.S.A. **101**(1), 5228–5235 (2004)
27. Hofmann, T.: Probabilistic latent semantic indexing. In: Proceedings of Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 50–57 (1999)
28. Hyvarinen, A.: Fast and robust fixed-point algorithms for independent component analysis. IEEE Trans. Neural Netw. **10**(3), 626–634 (1999)
29. Ishwaran, H., Rao, J.S.: Spike and slab variable selection: frequentist and Bayesian strategies. Ann. Stat. **33**(2), 730–773 (2005)
30. Jolliffe, I.T.: Principal Component Analysis. Springer, New York (1986)
31. Kim, S., Georgiou, P., Narayanan, S.: Latent acoustic topic models for unstructured audio classification. APSIPA Trans. Signal Inf. Process. **1** (2012). doi:10.1017/ATSIP.2012.7
32. Kneser, R., Ney, H.: Improved backing-off for *m*-gram language modeling. In: Proceedings of International Conference on Acoustics, Speech, and Signal Processing, pp. 181–184 (1995)
33. Kuhn, R., Junqua, J.C., Nguyen, P., Niedzielski, N.: Rapid speaker adaptation in eigenvoice space. IEEE Trans. Audio Speech Lang. Process. **8**(4), 695–707 (2000)
34. Pitman, J., Yor, M.: The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. Ann. Probab. **25**, 855–900 (1997)
35. Smaragdis, P., Shashanka, M., Raj, B.: Topic models for audio mixture analysis. In: Proceedings of NIPS Workshop on Applications for Topic Models: Text and Beyond (2009)
36. Tam, Y.C., Schultz, T.: Dynamic language model adaptation using variational Bayes inference. In: Proceedings of Annual Conference of International Speech Communication Association, pp. 5–8 (2005)
37. Tam, Y.C., Schultz, T.: Unsupervised language model adaptation using latent semantic marginals. In: Proceedings of Annual Conference of International Speech Communication Association (2006)

38. Teh, Y.W.: A hierarchical Bayesian language model based on Pitman-Yor processes. In: Proceedings of International Conference on Computational Linguistics and Annual Meeting of the Association for Computational Linguistics, pp. 985–992 (2006)
39. Teh, Y.W., Jordan, M.I., Beal, M.J., Blei, D.M.: Hierarchical Dirichlet processes. J. Am. Stat. Assoc. **101**(476), 1566–1581 (2006)