Representation method for a set of documents from the viewpoint of Bayesian statistics

Masayuki GOTO,
Fac. of Environmental and Information Studies
Musashi Institute of Technology, Japan
goto@yc.musashi-tech.ac.jp

Takashi ISHIDA, and Shigeichi HIRASAWA
Dep. of Science and Engineering,
Waseda University, Japan
ishida@hirasa.mgmt.waseda.ac.jp

Abstract – *In this paper, we consider the Bayesian approach for representation of a set of documents. In the field of representation of a set of documents, many previous models, such as the latent semantic analysis (LSA), the probabilistic latent semantic analysis (PLSA), the Semantic Aggregate Model (SAM), the Bayesian Latent Semantic Analysis (BLSA), and so on, were proposed. In this paper, we formulate the Bayes optimal solutions for estimation of parameters and selection of the dimension of the hidden latent class in these models and analyze it's asymptotic properties.*

Keywords: Probabilistic Latent Semantic Indexing, Automated Document Indexing, Information Retrieval, Bayesian Statistics

1 Introduction

Recently, huge repositories of textual data are available to use. In the field of information retrieval, Latent Semantic Indexing (LSI)[1] was proposed. Although the typical information retrieval systems match the keywords in a user's query to the index words for all documents in the database, LSI computes a smaller semantic subspace from the original word-document matrix. On the other hand, Probabilistic Latent Semantic Indexing (PLSI)[3], which was proposed by T.Hofmann, is an interesting approach to automated document indexing and information retrieval which is based on a statistical latent class model for factor analysis of counted data. From the probabilistic representation, we can treat the latent semantic indexing problem based on the probability theory. However, the probabilistic model is estimated based on the likelihood estimation in PLSI. In the problems of the automated document indexing and the information retrieval, the number of parameters which must be estimated is huge. Therefore, the sample size may not be sufficient to estimate the parameters in these problems even if we can get huge repositories of textual data. This is because we cannot calculate the EM algorithm for huge size data from the viewpoint of computational complexity. Although the Semantic Aggregate Model (SAM) proposed by D.Mochihashi and U.Matsumoto is a model representing the meaning of words, the SAM has also same properties as PLSI. In these models, there arises a problem: "How can we select the dimension of the model?".

On the other hand, we can find the results of studies for estimating the order of a hidden Markov model(HMM)[11]. Because HMM model does not satisfy the smoothness assumptions for probabilistic model class under which the AIC, the BIC, and the MDL were derived, we cannot apply these model selection criteria to select the order of HMM as they are. The PLSI model is similar with the HMM model. Then, the similar model selection criteria in [11] may be applied to make a decision of the number of the unobserved class.

In this paper, we propose the Bayesian representation of the probability model for latent semantic analysis. The Bayes method is well known as a useful method for small size of data in statistics[7]. Moreover, Bayesian statistics is congenial to the model selection problems. However, any criterion from the viewpoints of Bayesian statistics has not been proposed to select the number of the unobserved class in the latent semantic analysis. In the field of latent semantic analysis, N.Freitas and K.Barnard propose a general Bayesian treatment of the latent semantic analysis problem. However, in the BLSA, the algorithm to estimate the parameters is based on maximum likelihood method. Although the prior distribution with hyper-parameters is assumed in this model, the estimation of parameters or hyper-parameters was focused. These parameters are estimated by maximizing of the log-likelihood or marginal log-likelihood.

We give a new probabilistic model for automated document indexing and information retrieval based on Bayesian statistics, derive the asymptotic properties, and then propose the criterion to select the number of class (model selection) based on Bayes theory. From this formulation, we can derive the Bayesian criterion to estimate the probabilistic inference in automated document indexing and information retrieval. At first, we formulate the latent semantic analysis based on Bayesian decision theory for the case that the number of the hid-

0-7803-7952-7/03/$17.00 © 2003 IEEE.

den latent states is given. Moreover, we consider the problem how we should select the number of the hidden latent states (dimension of a model) from the viewpoint of Bayes decision theory. By the asymptotic analysis and simulation experiments, we show the effectiveness of our Bayesian method.

2 Preliminaries

Here, we show the conventional models for the information retrieval and modeling of word meaning.

Through the paper, $P(\cdot)$ is used for a probability and $f(\cdot)$ is used for a probability density. When we don't specify the probability or probability density, then we use the notation $p(\cdot)$ for a meaning of the probability distribution.

2.1 Latent Semantic Indexing (LSI) [1]

Let each document be represented by a vector x_i ($i \in \{1, 2, \cdots, n\}$) containing the frequencies of d index words.

Let X be a word-document matrix given by

$$X = (x_1, x_2, \cdots, x_n) \tag{1}$$

where n is the number of documents. Given a query q by a user, the retrieval system computes a list of scores $s_i = q^T x_i$ and put it out as a result. LSI represents the word-document matrix in a much smaller k-dimensional subspace and this is done by the truncated singular value decomposition (SVD). From SVD, we can decompose X as

$$X = \sum_{i=1}^{r} t_i \sigma_i d_i = TSD, \tag{2}$$

where $T = (t_1, t_2, \cdots, t_r)$ and $D = (d_1, d_2, \cdots, d_r)$ are left and right singular vectors, $S = diag(\sigma_1, \sigma_2, \cdots, \sigma_r)$, and $\sigma_1, \sigma_2, \cdots, \sigma_r$ are the singular values. Selecting k big singular values, we can re-represent X as

$$\hat{X}_k = T_k S_k D_k^T, \tag{3}$$

where $k < r$. The query is transformed to $q^t T_k$, the documents are represented as $S_k D_k^T$. The relevance score is computed as $s = (q^t T_k)(S_k D_k^T)$.

Here a problem arises. The truncated SVD is the best approximation of X in the reduced k-dimensional space from the viewpoint of square error. However, we must reasonably determine the dimension k. A range of k from 100 to 500 or more have been suggested based on empirical evidences. In [4], MDL criterion to determine the dimension k is proposed and the effectiveness of the proposal is shown from the simulation experiments.

2.2 Probabilistic Latent Semantic Indexing (PLSI) [3]

The PLSI is a latent variable model for general co-occurrence data which associates an unobserved class variable (the hidden latent class) $c \in \mathcal{C} = \{c_1, c_2, \cdots, c_k\}$ with each observation, i.e., with each occurrence of a word $w \in \mathcal{W} = \{w_1, w_2, \cdots, w_d\}$ in a document $x \in \mathcal{X} = \{x_1, x_2, \cdots, x_n\}$. In terms of a generative model it can be defined in the following way: 1) select a document x with probability $P(x)$, 2) pick a latent class c with probability $P(c|x)$, 3) generate a word w with probability $P(w|c)$. As a result one obtains an observed pair (x, w), which leads the latent class in the expression

$$P(x, w) = \sum_{c \in \mathcal{C}} P(w|c)P(c|x)P(x). \tag{4}$$

Following the likelihood principle, we can estimate $P(c)$, $P(x|c)$, and $P(w|c)$ by maximization of the log-likelihood function

$$\mathcal{L} = \sum_{x \in \mathcal{X}} \sum_{w \in \mathcal{W}} n(x, w) \log P(x, w), \tag{5}$$

where $n(x, w)$ is the term frequency, i.e., the number of times w occurred in x.

In fact, the maximum likelihood estimator can be calculated by EM algorithm.

1) E-step:

$$P(c|x, w) = \frac{P(x|c)P(w|c)P(c)}{\sum_{c \in \mathcal{C}} P(x|c)P(w|c)P(c)}$$

2) M-step:

$$P(x|c) = \frac{\sum_w n(x, w)P(c|x, w)}{\sum_{x,w} n(x, w)P(c|x, w)},$$

$$P(w|c) = \frac{\sum_x n(x, w)P(c|x, w)}{\sum_{x,w} n(x, w)P(c|x, w)},$$

$$P(c) \propto \sum_{x,w} n(x, w)P(c|x, w).$$

We can rewrite (4) as follows:

$$P(x, w) = \sum_{c \in \mathcal{C}} P(w|c)P(x|c)P(c). \tag{6}$$

2.3 Semantic Aggregate Model (SAM) [5]

Mochihashi and Matsumoto proposed a Semantic Aggregate Model (SAM) on word meanings by extending the PLSI. By this representation, the semantic distance and semantic weights of words can be reformulated mathematically.

In SAM, the probability model of co-occurrence of words w and w' is introduced.

$$P(w,w') = \sum_{c\in\mathcal{C}} P(w|c)P(w'|c)P(c). \quad (7)$$

This model is extended version of the Aggregate Markov Model $P(w'|w) = \sum_{c\in\mathcal{C}} P(w'|c)P(c|w)$ proposed by F.Saul and F.Pereira[6].

2.4 Bayesian Latent Semantic Analysis (BLSA) [8]

N.Freitas and K.Barnard propose a general Bayesian treatment of the latent semantic analysis problem. They constructed a model of the problem as a general formulation. For X given by (1), each documents x_i is assumed to be drawn from the following mixture model

$$\begin{aligned} P(x_i) &= \sum_{c\in\mathcal{C}} P(x_i|\theta_c,c)P(c) \\ &= \sum_{c\in\mathcal{C}}\prod_{j=1}^{d} P(x_{i,j}|\theta_{c,j},c)P(c) \end{aligned} \quad (8)$$

where $x_{i,j}$ is j-th attribute of x_i related to j-th word and $\theta_c = (\theta_{c,1},\theta_{c,2},\cdots,\theta_{c,d})$ is a parameter vector. If $x_{i,j}\in\{0,1\}$, then $P(x_i|\theta_c,c)$ is given by the form

$$P(x_i|\theta_c,c) = \prod_{j=1}^{d}(\theta_{c,j})^{x_{i,j}}(1-\theta_{c,j})^{1-x_{i,j}}. \quad (9)$$

Rewriting $P(c)$ using the parameters as $P(c_l)=\lambda_l$, we have

$$P(x_i) = \sum_{l=1}^{k}\lambda_l\prod_{j=1}^{d}P(x_{i,j}|\theta_{l,j},c_l) \quad (10)$$

where $\sum_{l=1}^{k}\lambda_l = 1$.

If $\mathcal{C}$ is given, the problem is reduced to estimate θ_c and $P(c)$. They consider three types of Bayesian approaches, namely simple Bayes, empirical Bayes and hierarchical Bayes.

2.4.1 The simple Bayes method

The simple Bayes compute the maximum a posteriori (MAP) estimator instead of the maximum likelihood estimator. The EM algorithm is easily modified to produce the MAP estimator[9],p.30.

2.4.2 The hierarchical and empirical Bayes methods

The hierarchical Bayes model proposed in [8] has the following three levels:

$$P(X|\varphi,z) = \prod_{i=1}^{n}P(x_i|\varphi,z_i) \quad (11)$$

$$p(\varphi,z|\eta) = \prod_{i=1}^{n}p(\varphi,z_i|\eta) \quad (12)$$

$$p(\eta) \quad (13)$$

where $\varphi=(\lambda,\theta)$ and $\lambda=(\lambda_1,\lambda_2,\cdots,\lambda_k)$. η is the hyper-parameters which specify the probability distribution of the parameter $\varphi=(\lambda,\theta)$. If the posterior probability distribution $f(\eta|X)$ is fairly sharply peaked around its mode $\tilde{\eta}$, then we can approximately use $p(\varphi,z_i|\tilde{\eta},X)$ instead of the marginal posterior probability

$$p(\varphi,z_i|X) = \int p(\varphi,z_i|\eta,X)f(\eta|X)d\eta. \quad (14)$$

This approach is the empirical Bayes method.

3 Latent Semantic Analysis based on Bayesian Decision Theory

The Bayesian Latent Semantic Analysis proposed in [8] is very interesting from the viewpoint of accuracy of estimation from the insufficient size of document data. However, in the BLSA, the algorithm to estimate the parameters is based on EM algorithm. Although the prior distribution with hyper-parameters is assumed in this model, the estimation of parameters or hyper-parameters was focused. These parameters are estimated by maximizing of the log-likelihood or marginal log-likelihood.

In this section, we formulate the latent semantic analysis based on Bayesian decision theory for the case that the number of the hidden latent states, k, is given.

3.1 Basic Probability Model

For X given by (1), each documents x_i is assumed to be drawn from the following mixture model

$$P(x_i|\theta,\lambda) = \sum_{l=1}^{k}\lambda_l\prod_{j=1}^{d}P(x_{i,j}|\theta_{l,j},c_l^k), \quad (15)$$

for $i\in\{1,2,\cdots,n\}$. If $x_{i,j}\in\{0,1\}$, then $P(x_i|\theta,\lambda)$ is given by the form

$$P(x_i|\theta,\lambda) = \sum_{l=1}^{k}\lambda_l\prod_{j=1}^{d}(\theta_{l,j})^{x_{i,j}}(1-\theta_{l,j})^{1-x_{i,j}}, \quad (16)$$

where $0<\theta_{l,j}<1$ and $\sum_l\lambda_l = 1$. Here, $\theta = (\theta_1,\theta_2,\cdots,\theta_k)$, where $\theta_l = (\theta_{l,1},\theta_{l,2},\cdots,\theta_{l,d})$, and $\lambda=(\lambda_1,\lambda_2,\cdots,\lambda_k)$ are previously unknown parameters. $P(X|\theta,\lambda)$ is given by

$$\begin{aligned} P(X|\theta,\lambda) &= \prod_{i=1}^{n}\left\{\sum_{l=1}^{k}\lambda_l\prod_{j=1}^{d}(\theta_{l,j})^{x_{i,j}}(1-\theta_{l,j})^{1-x_{i,j}}\right\} \\ &= \sum_{l_1=1}^{k}\cdots\sum_{l_n=1}^{k}\left\{\prod_{i=1}^{n}\lambda_{l_i}\prod_{j=1}^{d}(\theta_{l_i,j})^{x_{i,j}}(1-\theta_{l_i,j})^{1-x_{i,j}}\right\}. \end{aligned} \quad (17)$$

3.2 Method for Calculation of Bayes Optimal Solutions

We assume that the prior distributions $f(\theta_{l,j}|c_l^k)$ and $f(\lambda)$. Then the posterior distribution of (θ, λ) given X is given by

$$P(\theta, \lambda|X) = \frac{P(X|\theta, \lambda)f(\theta)f(\lambda)}{\int_\lambda \int_\theta P(X|\theta, \lambda)f(\theta)f(\lambda)d\theta d\lambda}, \quad (18)$$

When the square error loss function is considered, the Bayes optimal estimator of (θ, λ) is given by

$$\left(\tilde{\theta}, \tilde{\lambda}\right) = \int_\theta \int_\lambda (\theta, \lambda)P(\theta, \lambda|X)d\theta d\lambda. \quad (19)$$

In this case that we consider the prediction of future observations, the loss function can be defined by $L(Ax, x)$, the loss between the prediction Ax and observation value x, or $L(AP, P)$, the loss between the probability distribution of future observation and the true distribution. Typically we can assume the square loss for $L(Ax, x)$ or the logarithmic loss for $L(AP, P)$. From the Bayes decision theory, the optimal prediction is given by calculation of the Bayes predictive distribution

$$P(x|X) = \int_{\theta,\lambda} P(x|\theta, \lambda)P(\theta, \lambda|X)d\theta d\lambda. \quad (20)$$

3.3 Method for Calculation of Bayes Optimal Solutions

To calculate the posterior probability $P(\theta, \lambda|X)$ or the Bayes predictive distribution $P(x|X)$, we give the form of the useful prior distribution. Usually, we can assume the Dirichlet prior distribution on the mixing coefficients λ. The Dirichlet prior distribution is given by

$$f(\lambda) = \frac{\Gamma(\alpha_1 + \alpha_2 + \cdots + \alpha_k)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\cdots\Gamma(\alpha_k)}\lambda_1^{\alpha_1 - 1}\lambda_2^{\alpha_2 - 1}\cdots\lambda_k^{\alpha_k - 1}, \quad (21)$$

where $\Gamma(\cdot)$ is the gamma function. For the prior of θ, we can assume the beta prior distribution:

$$f(\theta_{l,j}|c_l^k) = \frac{\Gamma(\beta_1^{l,j} + \beta_2^{l,j})}{\Gamma(\beta_1^{l,j})\Gamma(\beta_2^{l,j})}(\theta_{l,j})^{\beta_1^{l,j} - 1}(1 - \theta_{l,j})^{\beta_2^{l,j} - 1}, \quad (22)$$

for all $l \in \{1, 2, \cdots, k\}$ and $j \in \{1, 2, \cdots, d\}$. If we know nothing about parameters previously, we can set $\alpha_l = 1$ and $\beta_1^{l,j} = \beta_2^{l,j} = 1$.

For convenience, we denote $l = (l_1, l_2, \cdots, l_n)$ and $\mathcal{L} = \{(l_1, l_2, \cdots, l_n)|l_i \in \{1, 2, \cdots, k\}\}$. $l = l_i$ means that the i-th document is emitted from the l-th hidden latent state c_l^k. Define $z(l|l)$ as

$$z(l|l) = \sum_{i=1}^{n} \mathbf{1}\{l = l_i\}, \quad (23)$$

where $\mathbf{1}\{\cdot\}$ is an indicator function. $z(l|l)$ means the number of l-th hidden latent state c_l^k in l. Define $x_i(l|l)$ as

$$x(j|l, l) = \sum_{i:l=l_i} x_{i,j}. \quad (24)$$

$x(j|l, l)$ means the number of 1 of j-th element of x emitted from the l-th hidden latent state c_l^k in X on the setting l is given.

On the above setting, the likelihood function (17) can be rewritten as

$$P(X|\theta, \lambda) = \sum_{l\in\mathcal{L}}\prod_{l=1}^{k}\left\{(\lambda_l)^{z(l|l)}\prod_{j=1}^{d}(\theta_{l,j})^{x(j|l,l)}(1 - \theta_{l,j})^{1 - x(j|l,l)}\right\}. \quad (25)$$

The posterior distribution of $\theta_{l,j}$ conditioned by X and λ is given by

$$f(\theta_{l,j}|X, \lambda) \propto \sum_{l\in\mathcal{L}}\left\{\prod_{i=1}^{n}\lambda_{l_i}\frac{\Gamma(\beta_1^{l,j} + \beta_2^{l,j})}{\Gamma(\beta_1^{l,j})\Gamma(\beta_2^{l,j})} \cdot (\theta_{l,j})^{x(j|l,l) + \beta_1^{l,j} - 1}(1 - \theta_{l,j})^{n - x(j|l,l) + \beta_2^{l,j} - 1}\right\}. \quad (26)$$

The posterior distribution of $(\theta_{l,j}, \lambda)$ is given by

$$f(\theta_{l,j}, \lambda|X) = \frac{1}{K_{l,j}}\sum_{l\in\mathcal{L}}\left\{\left(\prod_{l'=1}^{k}(\lambda_{l'})^{z(l'|l) + \alpha_{l'} - 1}\right)\frac{\Gamma(\beta_1^{l,j} + \beta_2^{l,j})}{\Gamma(\beta_1^{l,j})\Gamma(\beta_2^{l,j})} \cdot (\theta_{l,j})^{x(j|l,l) + \beta_1^{l,j} - 1}(1 - \theta_{l,j})^{n - x(j|l,l) + \beta_2^{l,j} - 1}\right\}. \quad (27)$$

The standardization constant $K_{l,j}$ is

$$K_{l,j} = \sum_{l\in\mathcal{L}}\left\{\frac{\prod_{l'=1}^{k}\Gamma(z(l'|l) + \alpha_{l'} - 1)}{\Gamma(n + \alpha_{l'})} \cdot \frac{\prod_{i=0}^{x(j|l,l)-1}\left(i + \beta_1^{l,j}\right)\prod_{i=0}^{n - x(j|l,l)-1}\left(i + \beta_2^{l,j}\right)}{\prod_{i=0}^{n-1}\left(i + \beta_1^{l,j} + \beta_1^{l,j}\right)}\right\}. \quad (28)$$

Letting $K_\lambda(l)$ and $K_{\theta_{l,j}}(l)$ be

$$K_\lambda(l) = \frac{\prod_{l'=1}^{k}\Gamma(z(l'|l) + \alpha_{l'} - 1)}{\Gamma(n + \alpha_{l'})}$$

$$K_{\theta_{l,j}}(l) = \frac{\prod_{i=0}^{x(j|l,l)-1}\left(i + \beta_1^{l,j}\right)\prod_{i=0}^{n - x(j|l,l)-1}\left(i + \beta_2^{l,j}\right)}{\prod_{i=0}^{n-1}\left(i + \beta_1^{l,j} + \beta_1^{l,j}\right)},$$

K can be rewritten as

$$K_{l,j} = \sum_{l\in\mathcal{L}} K_\lambda(l)K_{\theta_{l,j}}(l). \quad (29)$$

Therefore, the Bayes optimal estimators $\tilde{\theta}_{l,j}$ of $\theta_{l,j}$ are given by

$$\tilde{\theta}_{l,j} = \frac{1}{K}\sum_{l\in\mathcal{L}}\left\{K_\lambda(l)K_{\theta_{l,j}}(l)\left(\frac{x(j|l,l)+\beta_1^{l,j}}{n+\beta_1^{l,j}+\beta_2^{l,j}}\right)\right\}. \tag{30}$$

Then the Bayes optimal estimators of $\theta_{l,j}$ is given by the form of weighted sum of Laplace estimators calculated for the all case of l.

On the other hand, the posteriot distribution of λ is given by

$$f(\lambda|X) = \int_\theta \Big(\prod_{l,j} f(\theta_{l,j},\lambda|X)\Big)d\theta$$

$$= \frac{1}{K_\lambda}\sum_{l\in\mathcal{L}}\prod_{l=1}^{k}\left\{(\lambda_l)^{z(l|l)+\alpha_l-1}\prod_{j=1}^{d}K_{\theta_{l,j}}(l)\right\}, \tag{31}$$

where K_λ is given by

$$K_\lambda = \sum_{l\in\mathcal{L}}\frac{\prod_{l=1}^{k}\Gamma(z(l|l)+\alpha_l)K_{\theta_{l,j}}(l)}{\Gamma(n+\alpha_1+\alpha_2+\cdots+\alpha_k)}. \tag{32}$$

The Bayes optimal estimator $\tilde{\lambda}_l$ of λ is given by

$$\tilde{\lambda}_l =$$

$$\frac{1}{K_\lambda}\sum_{l\in\mathcal{L}}\left\{K_\lambda(l)\prod_{j=1}^{d}K_{\theta_{l,j}}(l)\left(\frac{z(l|l)+\alpha_l}{n+\alpha_1+\alpha_2+\cdots+\alpha_k}\right)\right\}. \tag{33}$$

We can see that the Bayes optimal estimators of λ_l is also given by the form of weighted sum of Laplace estimators calculated for the all case of l.

Therefore, we can construct the algorithm to calculate the Bayes optimal estimator from the Laplace estimators for all l. Although the hidden latent state cannot be observed from x_i, we can get the optimal estimator from the weighted mixture of all cases of l.

4 Model Specification and Model Mixture Method from the Viewpoint of Bayesian Theory

The Bayesian Latent Semantic Analysis proposed in [8] is very interesting from the viewpoint of accuracy of estimation from the insufficient size of document data. However, in the BLSA, the number of class c, k, that is the size of $\mathcal{C}$, is given. In the settings of LSI, PLSI, SAM, and BLSA, the selection of the number of hidden class c is important. This is the problem to select the order of a hidden class and essentially equivalent to the model selection problem in documents and words model. For LSI model, a method using the MDL criterion was proposed [4]. However, the Bayes method is useful to the model selection problems. In this paper, we formulate the selection of the order k based on the Bayesian statistics, and analyze the performance from the viewpoints of the asymptotic properties.

4.1 Basic Probability Model

Let m be specifying the order of hidden class $\mathcal{C}_m$, where $\mathcal{C}_m = \{c_1, c_2, \cdots, c_{k_m}\}$. We call m *a model*. k_m is the order of the model m.

For X given by (1), each documents x_i is assumed to be drawn from the following mixture model

$$P(x_i|m) = \int_\lambda\int_\theta\Big\{\sum_{l=1}^{k_m}\prod_{j=1}^{d}P(x_{i,j}|m,\theta_{l,j}^{k_m},c_l^{k_m})\lambda_l^{k_m}\Big\}$$

$$f(\theta_{l,j}^{k_m}|m,c_l^{k_m})f(\lambda_l^{k_m}|m)d\theta_{l,j}^{k_m}d\lambda_l^{k_m}, \tag{34}$$

where $\sum_{i=1}^{k_m}\lambda_i^{k_m} = 1$. Similary $P(X|m)$ is given by

$$P(X|m) = \int_\lambda\int_\theta\Big\{\sum_{l=1}^{k_m}\prod_{i=1}^{n}\prod_{j=1}^{d}P(x_{i,j}|m,\theta_{l,j}^{k_m},c_l^{k_m})\lambda_l^{k_m}\Big\}$$

$$f(\theta_{l,j}^{k_m}|m,c_l^{k_m})f(\lambda_l^{k_m}|m)d\theta_{l,j}^{k_m}d\lambda_l^{k_m}, \tag{35}$$

Let $\mathcal{M}$ be a finite model class and the cardinality of $\mathcal{M}$ be M. That is, $\mathcal{M} = \{m_1, m_2, \cdots, m_M\}$ and $M = |\mathcal{M}|$. We assume that $k_m \neq k_{m'}$ for $m \neq m'$, $m, m' \in \mathcal{M}$. Although the true model m^* omitting the data X exists, we cannot know it previously. Therefore, we must select m for modeling the probability structure of X using (34). It is equivalent to select the order of dimension of the hidden class.

We assume that the true model m^* emitting documents and words data X exists in $\mathcal{M}$. The true model m^* is defined by

$$m^* =$$

$$\arg_{m\in\mathcal{M}}\min\Big\{k_m\Big|\exists\theta^{k_m},\lambda^{k_m},P(x|m,\theta^{k_m},\lambda^{k_m}) = P^*(x)\Big\} \tag{36}$$

4.2 Bayesian Formulation of Model Selection

We assume the prior probability over the model class, $P(m)$. Let $L(Am,m)$ be a loss function between the decision $Am \in \mathcal{M}$ (selected model) and each model $m \in \mathcal{M}$. Then the Bayes risk is given by

$$BR(Am) = \sum_{m\in\mathcal{M}}L(Am,m)P(m|X), \tag{37}$$

where $P(m|X)$ is given by

$$P(m|X) = \frac{P(X|m)P(m)}{\sum_{m\in\mathcal{M}}P(X|m)P(m)}$$

$$= \frac{\prod_{i=1}^{n}P(x_i|m)P(m)}{\sum_{m\in\mathcal{M}}\prod_{i=1}^{n}P(x_i|m)P(m)}. \tag{38}$$

If we set a 0-1 loss function such as

$$L(Am, m) = \begin{cases} 0 & \text{if } Am = m, \\ 1 & \text{if } Am \neq m, \end{cases} \quad (39)$$

then the Bayes optimal model selection $\hat{m}_{BD} = Am^*$ is given by

$$\hat{m}_{BD} = Am^* = \arg_m \max P(m|X). \quad (40)$$

4.3 Method for Calculation of Bayes Optimal Solutions

To calculate the posterior probability of m, $P(m|X)$, we must calculate $P(X|m)$ given by (35).

$$P(X|m) =$$

$$\int_\lambda \int_\theta \left\{ \sum_{l=1}^{k_m} \lambda_l^{k_m} \prod_{i=1}^{n} \prod_{j=1}^{d} \left(\theta_{l,j}^{k_m}\right)^{x_{i,j}} \left(1 - \theta_{l,j}^{k_m}\right)^{1-x_{i,j}} \right\}$$

$$\cdot f(\theta_{l,j}^{k_m}|m, c_l^{k_m}) f(\lambda_l^{k_m}|m) d\theta_{l,j}^{k_m} d\lambda_l^{k_m}, \quad (41)$$

We can assume the Dirichlet prior distribution on the mixing coefficients λ and the beta prior distribution on $f(\theta_{l,j}^{k_m}|m, c_l^{k_m})$. The Dirichlet prior distribution as the prior of λ^{k_m} of m is given by

$$f(\lambda^{k_m}|m) =$$

$$\frac{\Gamma(\alpha_1 + \alpha_2 + \cdots + \alpha_{k_m})}{\Gamma(\alpha_1)\Gamma(\alpha_2)\cdots\Gamma(\alpha_{k_m})} \lambda_1^{\alpha_1 - 1} \lambda_2^{\alpha_2 - 1} \cdots \lambda_{k_m}^{\alpha_{k_m} - 1}, \quad (42)$$

and the bata distribution $f(\theta_{l,j}^{k_m}|m, c_l^{k_m})$ as the prior of $\theta_{l,j}^{k_m}$ of model m is given by

$$f(\theta_{l,j}^{k_m}|m, c_l^{k_m}) = \frac{\Gamma(\beta_1 + \beta_2)}{\Gamma(\beta_1)\Gamma(\beta_2)} \left(\theta_{l,j}^{k_m}\right)^{\beta_1 - 1} \left(1 - \theta_{l,j}^{k_m}\right)^{\beta_2 - 1}, \quad (43)$$

Using (42) and (43), we can calculate $P(X|m)$ using the techniques shown in Section 3. This calculation is not based on the EM algorithm or other iteration algorithm.

5 Asymptotic Analysis

We can show the following theorem using the asymptotic normality of the posterior density [10].

Theorem:

Assuming that the true model m^* and true parameters $\theta_{m^*}^{k_{m^*}}$ and λ^* emitting documents and words data X exists in $\mathcal{M}$, the model selection by (40) satisfy

$$\hat{m}_{BD} \to m^*, \quad a.s. \quad (44)$$

That is, we can asymptotically find the optimal number of the hidden latent states, k_{m^*}.

6 Conclusions

In this paper, we propose the Bayesian representation of the probability model for latent semantic analysis and propose the criterion to select the number of class (model selection) based on Bayes theory. Moreover we analyze the asymptotic properties. We may use the results for the hidden Markov model class [11],[12] to analyze the properties of proposal. This is a future work.

Because the proposed method is not based on estimation of parameters using maximum likelihood, the iterated algorithm, such as the EM algorithm, don't need. Computational complexity should be considered as the future work.

References

[1] S, Deerwester, S. T. Dumais, G.W. Furnas, T. K. Landauer, and R. Harshman: "Indexing by Latent Semantic Analysis", *Journal of the American Society of Information Science*, Vol.41, No.6, pp.391-407, (1990)

[2] C. H. Q. Ding: "A Probabilistic Model for Latent Semantic Indexing in Information Retrival and Filtering", in *Computational Information Retrival*, pp.65-73, (2000)

[3] T. Hofmann: "Probabilistic Latent Semantic Indexing", In *Proc. of the 22nd International Conference on Research and Development in Information Retrieval(SIGIR '99)*, pp50-57, (1999)

[4] H.Zha: "A Subspace-based Model for Information Retrieval with Applications in Latent Semantic Indexing",

[5] D. Mochihashi and U. Matsumoto: "Probabilistic Representation of Meanings", IPSJ SIG Notes on Natural Language(NL), 2002-NL-147-12, pp.77-84, in Japanese, (2002)

[6] L.Saul and F.Pereira: "Aggregate and mixed-order Markov Models for Statistical Natural Language Processing", In *Proc. of the Second Conference on Empirical Method in Natural Language Processing.* pp.81-89, (1997)

[7] J.O. Berger: *Statistical Decision Theory and Bayesian Analysis (Second Edition)*, Springer - Verlag, (1985)

[8] Nando de Freitas and Kobus Barnard: "Bayesian Latent Semantic Analysis ", http://elib.cs.berkeley.edu/papers/clustering/ bayesian/ ,(2000)

[9] G. J. McLachlan and T. Krishnan: *The EM Algorithm and Extensions*, John Wiley and Sons, Inc., (1997)

[10] B.S.Clarke : "Asymptotic Normality of the Posterior in Relative Entropy", *IEEE Trans. Information Theory*, Vol.45, No.1, pp.165-176, (1999)

[11] R.J. MacKay: "Estimating the order of a hidden Markov model", *The Canadian Journal of Statistics*, Vol. 30, No 4, pp. 573-589,(2002)

[12] L.Mevel and L.Linesso: "Bayesian estimation of hidden Markov models", Proc. Mathematical Theory of Networks and Systems MTNS-2000, Perpignan, (2002)