

---

# Bayesian Latent Semantic Analysis

---

Nando de Freitas

Kobus Barnard

Computer Science Division, UC Berkeley  
387 Soda Hall, Berkeley CA 94720-1776, USA  
{jfgf,kobus}@cs.berkeley.edu

## Abstract

We extend recently proposed models for probabilistic latent semantic analysis using a hierarchical Bayesian framework. This approach enables us to carry out automatic regularisation of large, complex probabilistic models for multimedia databases. Moreover, it allows us to introduce *a priori* knowledge into the modelling process using specific word and image preferences, as well as, semantic hierarchies obtained using WordNet.

## 1 Introduction

We introduce a Bayesian treatment for a probabilistic latent semantic model for discrete data (Hofmann 1999), recently extended for discrete and continuous data (Barnard and Forsyth 2001) in the context of images with associated text. Learning the joint probability distribution of image segment features and associated text has applications to computer vision, image understanding, multimedia database browsing, and document retrieval. The model supports browsing by exposing the semantics of the data. Furthermore, the model can be queried for documents using keywords and/or image features (search), images using text blocks (auto-illustrate), and words using images (auto-annotate), the latter process having clear ties to object recognition.

These exciting applications are characterised by complex models which are large relative to the size of the data; thereby motivating the Bayesian approach. By incorporating priors into the model, we can regularise and take advantage of prior knowledge. For example we can assign an advantageous prior probability to (application specific) special words. Natural language processing can also provide useful priors. In this work, since the model generates document items from a hierarchical set of nodes with more general “concepts” being emitted from higher levels, we use information from WordNet (Miller, Beckwith, Fellbaum, Gross and Miller 1998) to encourage the model to emit words at a level corresponding to their level of semantic abstraction.

## 2 Probabilistic model

Following (Hofmann and Puzicha 1998), we treat the documents,  $d \in \mathcal{D}$ , and their attributes,  $a \in \mathcal{A}$ , as dyadic observations  $\mathcal{O} \triangleq (d_i, a_i)_{i=1}^N$ . Document attributes can be categorical, such as words, or continuous, such as Gaussian feature vectors extracted from image segments. The observations are modelled using a hierarchical mixture model that contains classes,  $c$ , in the, say, horizontal direction and levels,  $l$ , in the vertical direction. By grouping identical dyads together, this model can be written as follows

$$p(\mathcal{O}) = \prod_{d \in \mathcal{D}} \sum_{c=1}^{n_c} p(c) \prod_{a \in \mathcal{A}} \left[ \sum_{l=1}^{n_l} p(a|l, c) p(l|d, c) p(d) \right]^{N(d, a)}, \quad (1)$$

where  $N(d, a) \triangleq |\{(d_i, a_i) : d_i = d, a_i = a\}|$  denotes the empirical co-occurrence frequencies,  $\mathcal{D}$  the set of existing documents,  $\mathcal{A}$  the set of different attributes,  $n_l$  the number of levels and  $n_c$  the number of clusters. The probabilities  $p(c)$ ,  $p(a|l, c)$  and  $p(l|d, c)$  are unknown parameters that must be estimated.  $p(d)$  is chosen independently to ensure proper document-length normalisation (Hofmann and Puzicha 1998). The cardinalities of the sets  $\mathcal{D}$  and  $\mathcal{A}$  are  $n_d$  and  $n_a$  respectively.

It is trivial to modify the model to obtain clustering, hierarchical clustering and aspect models (Hofmann and Puzicha 1998). The choice of model is typically application dependent. With our browsing applications in mind, we impose a tree structure on the model by tying some of the parameters. In this setting, the terminal nodes represent clusters and the inner and terminal nodes represent aspects, see (Hofmann and Puzicha 1998) for details.

Words  $w$  are assumed to be distributed according to a multinomial model with parameters  $p(a|l, c) = p(w|l, c)$ . That is, we are adopting the standard “bag of words” model where, for computational simplicity, the words are assumed to be independent given the indicator variables ( $l$  and  $c$  account for the correlations in the word model). Information from images is extracted by first segmenting the images and then grouping the features of each segment into a segment vector  $s$  (Belongie, Carson, Greenspan and Malik 1997). The segment vectors are assumed to be distributed according to  $p(a|l, c) = p(s|l, c) = \mathcal{N}(\boldsymbol{\mu}_{l,c}, \boldsymbol{\Sigma}_{l,c})$ . For practical reasons, the Gaussian covariance is block-diagonal, and we tie the parameters by allowing only one Gaussian per node.

### 2.1 Priors

In addition to modelling the data, we need to introduce prior distributions. This Bayesian extension of the maximum likelihood (ML) approach is justified by the following points:

–**Ill-conditioning**: In the ML framework, the likelihood is often unbounded. For example, when dealing with mixtures of Gaussians, nothing prevents a mixture component density from being assigned to a single observation. When this happens, the variance goes to zero and the likelihood goes to infinity, thus causing serious ill-conditioning problems. To circumvent this common problem, people either prune components by hand or add extra tuning parameters as in ridge regression (Marquardt and Snee 1975). From a Bayesian perspective, the introduction of a prior reduces this problem.

–**A priori knowledge**: One can use the prior to specify domain-specific knowledge (some rules derived from an expert) or subjective preferences (favouring simpler

models). For example, we have implemented a prior that encourages concept levels in our hierarchical mixture model to reflect semantic levels obtained with WordNet.

–**Regularisation:** Since the data set is finite and noisy, one needs to take care of not overfitting the data. The prior distribution can be used to favour simpler (smooth) models that avoid fitting the noise and, therefore, extrapolate reasonable well.

–**Multiple overlapping copies of clusters:** ML estimation often splits an underlying category into several components with identical parameters whose component weights  $\lambda_c$  add up to the correct one. Bayesian estimation avoids this problem by specifying priors that favour sparse models.

–**Starting point for more sophisticated modelling:** The Bayesian perspective lays the groundwork for more sophisticated models that enable us to, for example, achieve robustness with respect to the specification of the prior distributions (no parameter tuning), perform model selection, extend point estimators to average estimators and consider different loss functions in a principled way: see for example (Andrieu, de Freitas and Doucet 2000, Bernardo and Smith 1994).

We follow a hierarchical Bayesian strategy, where the unknown parameters  $\varphi \triangleq \{p(w|l, c), p(c), p(a|d, c), \boldsymbol{\mu}, \boldsymbol{\Sigma}\}$  are regarded as being drawn from appropriate prior distributions. We acknowledge our uncertainty about the exact form of the prior by specifying it in terms of some unknown parameters (hyperparameters). We also need to introduce the latent allocation variables  $x_i \in \{1, \dots, n_c\}$  and  $y_i \in \{1, \dots, n_l\}$  to indicate that the  $i$ -th dyad belongs to a specific group  $c$  and level  $l$ . We model these variables with multinomial distributions

$$p(x_i|p(c)) = \prod_{c=1}^{n_c} p(c)^{\mathbb{I}_c(x_i)},$$

$$p(y_i|p(l|d, c)) = \prod_d \prod_{c=1}^{n_c} \prod_{l=1}^{n_l} p(l|d, c)^{\mathbb{I}_l(y_i)}$$

where  $\mathbb{I}_E(z)$  denotes the indicator function of the set  $E$  (1 if  $z \in E$ , 0 otherwise). The categorical parameters are modelled using parameterised Dirichlet distributions

$$p(p(c)|\boldsymbol{\alpha}) = \frac{\Gamma(\sum_c \alpha_c)}{\Gamma(\alpha_1) \dots \Gamma(\alpha_{n_c})} \prod_{c=1}^{n_c} p(c)^{\alpha_c - 1}$$

$$p(p(l|d, c)|\boldsymbol{\gamma}) = \prod_d \prod_{c=1}^{n_c} \frac{\Gamma(\sum_l \gamma_{d,l,c})}{\Gamma(\gamma_{d,1,c}) \dots \Gamma(\gamma_{d,n_l,c})} \prod_{l=1}^{n_l} p(l|d, c)^{\gamma_{d,l,c} - 1}$$

$$p(p(w|l, c)|\boldsymbol{\beta}) = \prod_{l=1}^{n_l} \prod_{c=1}^{n_c} \frac{\Gamma(\sum_w \beta_{w,l,c})}{\Gamma(\beta_{1,l,c}) \dots \Gamma(\beta_{n_w,l,c})} \prod_w p(w|l, c)^{\beta_{w,l,c} - 1}$$

where  $\boldsymbol{\alpha}$ ,  $\boldsymbol{\beta}$  and  $\boldsymbol{\gamma}$  denote the hyperparameters. For the multivariate Gaussian image segment vectors of length  $n_g$ , we adopt the following normal-inverse Wishart prior (McLachlan and Peel 2000)

$$\boldsymbol{\mu}_{l,c} \sim \mathcal{N}_{n_g}(\boldsymbol{\omega}_{l,c}, \boldsymbol{\Sigma}_{l,c}/\kappa_{l,c})$$

$$\boldsymbol{\Sigma}_{l,c}^{-1} \sim \mathcal{W}_{n_g}(r_{l,c}, \boldsymbol{\Delta}_{l,c})$$

where  $\mathcal{W}_{n_g}(r_{l,c}, \boldsymbol{\Delta}_{l,c})$  denotes a Wishart distribution. In the above expressions,  $\boldsymbol{\Delta}_{l,c}$  is a symmetric, positive definite,  $n_g \times n_g$  matrix, while  $\kappa_{l,c}$  and  $r_{l,c}$  are regularisation parameters, with  $r_{l,c} > n_g$ . For exposition purposes, we introduce  $\boldsymbol{\eta}$  to denote all the model hyperparameters.

A rigorous Bayesian analysis would involve specifying priors on the hyperparameters. This would require that we develop computationally demanding approximate algorithms based, for example, on variational bounds or Markov chain Monte Carlo. Here, we opt for more pragmatic solutions. We either set the hyper-parameters using *a priori* knowledge or use the data to compute point estimates of the hyper-parameters. In the latter case, we aim to find an  $\boldsymbol{\eta}^*$  that maximises the marginal likelihood  $p(\mathcal{O}|\boldsymbol{\eta})$ . That is, we are trying to find the most likely model hypothesis (Bernardo and Smith 1994). This approach of estimating the priors from the data is an empirical Bayes method known as maximum likelihood type II (Good 1983). It is based on the assumption that  $p(\boldsymbol{\eta}|\mathcal{O})$  is fairly sharply peaked around its mode  $\boldsymbol{\eta}^*$  and, consequently, approximations such as  $p(\boldsymbol{\varphi}, \mathbf{x}, \mathbf{y}|\mathcal{O}) \approx p(\boldsymbol{\varphi}, \mathbf{x}, \mathbf{y}|\boldsymbol{\eta}^*, \mathcal{O})$  are valid.

### 3 Computation

We develop an EM algorithm to compute the ML and MAP point estimates of the model parameters. Our EM algorithm can be derived either by formulating and differentiating the expected complete log-posterior subject to probabilistic constraints (using Lagrange multipliers) or by deriving expressions for the posterior modes and replacing the cluster and level indicator variables with their *a posteriori* expectations. The resulting algorithm involves iterating between the following standard steps:

**E step:** Compute the expectations of the indicator variables.

$$p(x_d = c|\mathcal{O}, \boldsymbol{\varphi}) \propto p(c) \left[ \prod_{a \in \mathcal{A}} \sum_{l=1}^{n_l} p(a|l, c) p(l|d, c) \right]^{N(d, a)}$$

$$p(y_i = l|a, d, c, \boldsymbol{\varphi}) \propto p(a|l, c) p(l|d, c)$$

**M step:** Update the model parameters.

$$p(w|l, c) \propto \sum_{i=1}^N \mathbb{I}_w(w_i) \xi_{i, l, c} + \beta_{w, l, c} - 1$$

$$p(l|d, c) \propto \sum_{i=1}^N \mathbb{I}_d(d_i) p(y_i = l|a, d, c, \boldsymbol{\varphi}) + \gamma_{d, l, c} - 1$$

$$p(c) \propto \sum_{d \in \mathcal{D}} p(x_d = c|\mathcal{O}, \boldsymbol{\varphi}) + \alpha_c - 1$$

$$\boldsymbol{\mu}_{l, c} = \left( \sum_{i=1}^N \xi_{i, l, c} s_i + \kappa_{l, c} \boldsymbol{\omega}_{l, c} \right) \left( \sum_{i=1}^N \xi_{i, l, c} + \kappa_{l, c} \right)^{-1}$$

$$\boldsymbol{\Sigma}_{l, c} = \left( \boldsymbol{\Delta}_{l, c}^{-1} + \widehat{\mathbf{V}}_{l, c} + \kappa_{l, c} (\boldsymbol{\omega}_{l, c} - \boldsymbol{\mu}_{l, c})(\boldsymbol{\omega}_{l, c} - \boldsymbol{\mu}_{l, c})' \right) \left( \sum_{i=1}^N \xi_{i, l, c} + r_{l, c} - n_g \right)^{-1}$$

where  $\xi_{i, l, c} \triangleq p(x_i = c, y_i = l|a, d, \boldsymbol{\varphi})$  and  $\widehat{\mathbf{V}}_{l, c} \triangleq \sum_{i=1}^N \xi_{i, l, c} (s_i - \boldsymbol{\mu}_{l, c})(s_i - \boldsymbol{\mu}_{l, c})'$ . The ML updates follow by setting the hyperparameters to obtain uninformative priors. That is,  $\boldsymbol{\alpha} \rightarrow \mathbf{1}, \boldsymbol{\beta} \rightarrow \mathbf{1}, \boldsymbol{\gamma} \rightarrow \mathbf{1}, \boldsymbol{\kappa} \rightarrow \mathbf{0}, r_{l, c} \rightarrow n_{g_a}$  and  $\boldsymbol{\Delta}^{-1} \rightarrow \mathbf{0}$ . To improve the mode search, stochastic versions of the algorithm (Monte Carlo EM) are easily obtained by sampling from the discrete posterior distribution of the cluster and level indicator variables and then computing the empirical expectation of the samples in the E step. The algorithms can also be easily annealed.

### 3.1 Estimating the hyperparameters

As mentioned earlier, we obtain the hyperparameters by maximising the expected marginal log-likelihood given by

$$\mathbb{E}_{p(\mathbf{x}, \mathbf{y} | \mathcal{O}, \boldsymbol{\eta})} [\log p(\mathbf{x}, \mathbf{y}, \mathcal{O} | \boldsymbol{\eta})] = \mathbb{E}_{p(\mathbf{x}, \mathbf{y} | \mathcal{O}, \boldsymbol{\eta})} \left[ \log \int p(\mathbf{x}, \mathbf{y}, \mathcal{O} | \boldsymbol{\varphi}) p(\boldsymbol{\varphi} | \boldsymbol{\eta}) d\boldsymbol{\varphi} \right]$$

This results on a set of equations that needs to be iterated in order to increase the lower bound on the marginal likelihood

$$\alpha_c^{(\text{new})} = \alpha_c \frac{\Psi \left( \sum_{i=1}^N p(x_i = c | a, d, \boldsymbol{\varphi}) + \alpha_c \right) - \Psi(\alpha_c)}{\Psi \left( n_d + \sum_{c=1}^{n_c} \alpha_c \right) - \Psi \left( \sum_{c=1}^{n_c} \alpha_c \right)}$$

$$\beta_{w,l,c}^{(\text{new})} = \beta_{w,l,c} \frac{\Psi \left( \sum_{i=1}^N \mathbb{I}_w(w_i) \xi_{i,l,c} + \beta_{w,l,c} \right) - \Psi(\beta_{w,l,c})}{\Psi \left( \sum_w \left( \sum_{i=1}^N \mathbb{I}_w(w_i) \xi_{i,l,c} + \beta_{w,l,c} \right) \right) - \Psi \left( \sum_w \beta_{w,l,c} \right)}$$

$$\gamma_{d,l,c}^{(\text{new})} = \gamma_{d,l,c} \frac{\Psi \left( \sum_{i=1}^N \mathbb{I}_d(d_i) p(y_i = l | a, d, c, \boldsymbol{\varphi}) + \gamma_{d,l,c} \right) - \Psi(\gamma_{d,l,c})}{\Psi \left( \sum_{l=1}^{n_c} \left( \sum_{i=1}^N \mathbb{I}_d(d_i) p(y_i = l | a, d, c, \boldsymbol{\varphi}) + \gamma_{d,l,c} \right) \right) - \Psi \left( \sum_{l=1}^{n_c} \gamma_{d,l,c} \right)}$$

where  $\Psi(\alpha_c) \triangleq \frac{\partial}{\partial \alpha_c} \log \Gamma(\alpha_c)$  is the digamma function. It is also possible to derive faster Newton-Raphson schemes as shown in (Narayanan 1991). However, one has to take care that this algorithm does not become numerically unstable. Lastly, estimators to compute the hyperparameters of the Gaussian components are discussed in (Gelman, Carlin, Stern and Rubin 1995).

## 4 Experiments

We first verified that the Bayesian model outperformed the ML model substantially when recovering the parameters of a second model used to generate the data. These experiments and some retrieval results are described in detail in our technical report (de Freitas and Barnard 2001). Subsequently, we performed several experiments on the Corel image database. This database contains images annotated with approximately 3 to 5 keywords each. The images on each of the CDs provided by Corel are samples from a particular theme. This makes this database very suitable for testing our algorithms. For example, when applying the various algorithms (ML, MAP with  $\alpha = 1$  and  $\beta = 2$  and MAP with empirical Bayes (EB)) to cluster a dataset consisting of 10 CDs (10 themes - 1000 documents), and assuming 20 clusters initially, we obtained the cluster probabilities shown in Figure 1. Clearly, the Bayesian strategies allow us to obtain a number of clusters that is in more agreement with the Corel human choice. Moreover this improves the coherence of the clusters <sup>1</sup>.

Figure 2 shows a section of the results obtained with a hierarchical quad-tree (with more than *30,000 parameters*), when trained with 1000 documents using both image segment features and associated words. In this case, we applied WordNet to the Corel annotations to generate more words in a semantic hierarchy. The more general words were assigned a higher prior probability at the top level of our hierarchical mixture model and the less general words were assigned a higher prior probability

<sup>1</sup>Since it is difficult to present these results in this paper format, we have made them available at <http://elib.CS.Berkeley.EDU/papers/clustering/bayesian/index.html>.

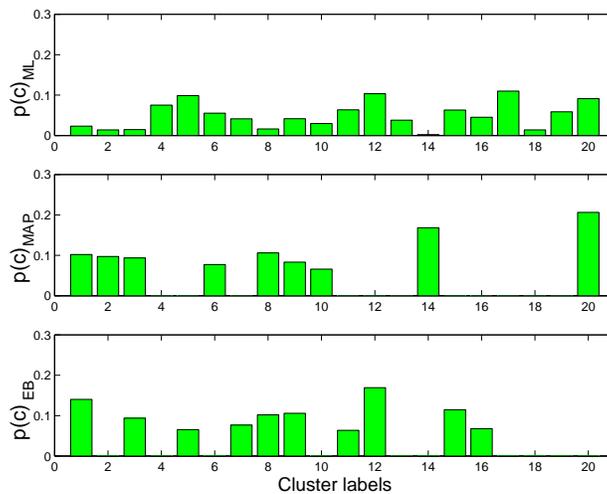


Figure 1: Cluster probabilities in the Corel example. Maximum likelihood (top) maximum *a posteriori* (middle) and empirical Bayes (bottom). The lack of a bar indicates that the respective cluster has been automatically pruned by the regulariser.

at the bottom levels. In other words, we used the prior to encode the WordNet semantic hierarchy in our model. The results indicate that higher levels are indeed related to more general concepts. The blank boxes are branches that have been pruned automatically by the prior.

## 5 Conclusions

In this paper, we have taken an important step towards improving models for Probabilistic latent semantic analysis. In particular, by adopting the Bayesian approach, we showed that it becomes possible to regularise and to introduce a priori knowledge in the training of large, complex probabilistic models for multimedia databases.

## Acknowledgments

We would like to thank Arnaud Doucet, Pinar Duygulu, David Forsyth, Thomas Hofman, Jan Puzicha and Stuart Russell for their help.

## References

- Andrieu, C., de Freitas, N. and Doucet, A. (2000). Robust full Bayesian learning for radial basis networks, to appear in *Neural Computation*.
- Barnard, K. and Forsyth, D. (2001). Learning the semantics of words and pictures, to appear in *ICCV2001*.
- Belongie, S., Carson, C., Greenspan, H. and Malik, J. (1997). Color and texture-based image segmentation using EM and its application to content-based image retrieval, *International Conference on Computer Vision*, pp. 675–682.
- Bernardo, J. M. and Smith, A. F. M. (1994). *Bayesian Theory*, Wiley Series in Applied Probability and Statistics.

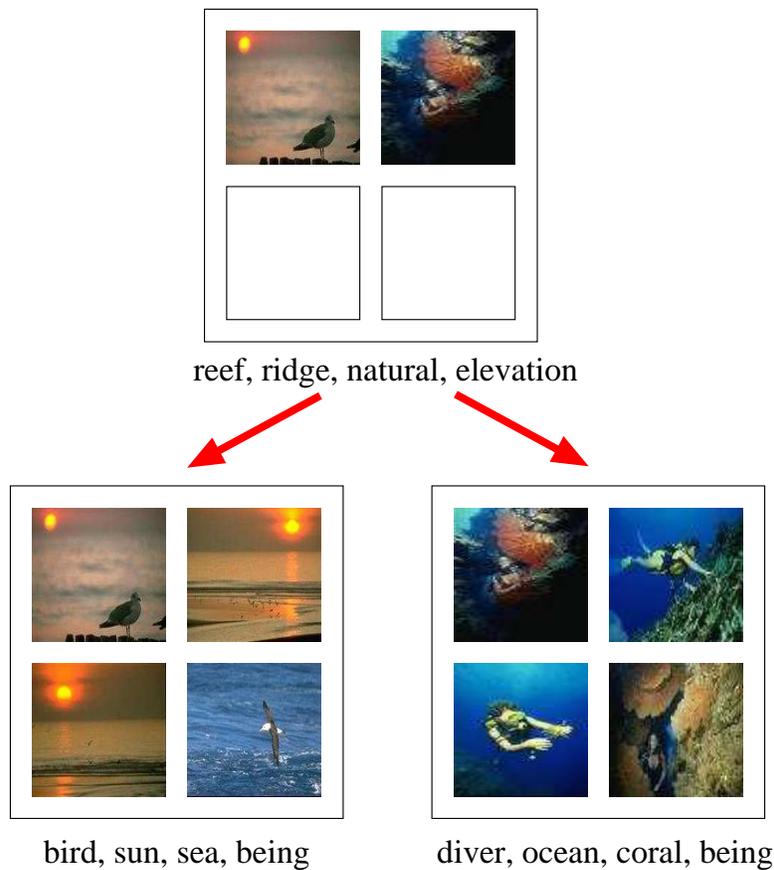


Figure 2: Section of the results generated by the hierarchical model.

- de Freitas, N. and Barnard, K. (2001). Bayesian modelling of documents with images and text, Computer Science Division, UC Berkeley.
- Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (1995). *Bayesian Data Analysis*, Chapman and Hall.
- Good, I. J. (1983). *Good Thinking: The Foundations of Probability and its Applications*, Minnesota Press, Minneapolis.
- Hofmann, T. (1999). Probabilistic latent semantic analysis, *Uncertainty in Artificial Intelligence*.
- Hofmann, T. and Puzicha, J. (1998). Unsupervised learning from dyadic data, *Technical Report TR-98-042*, International Computer Science Institute.
- Marquardt, D. W. and Snee, R. D. (1975). Ridge regression in practice, *American Statistician* **29**(1): 3–20.
- McLachlan, G. and Peel, D. (2000). *Finite Mixture Models*, Wiley Series in Probability and Statistics, New York.
- Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D. and Miller, K. J. (1998). Introduction to WordNet: An on-line lexical database, *International Journal of Lexicography* **3**: 235–244.
- Narayanan, A. (1991). Maximum likelihood estimation of the parameters of the Dirichlet distribution, *Applied Statistics* **40**(2): 365–374.