2009－CVIM－166 (12)
2009／3／13

Semi-supervised learning scheme using Dirichlet process EM-algorithm

Tomoaki KIMURA†, Yohei NAKADA††, Arnaud DOUCET†††, and Takashi MATSUMOTO††

† Graduate School of Advanced Science and Engineering, Waseda University
3-4-1, Ohkubo, Shinjuku-ku, Tokyo, 169-8555, Japan.
†† Faculty of Science and Engineering, Waseda University
3-4-1, Ohkubo, Shinjuku-ku, Tokyo, 169-8555, Japan.
††† The Institute of Statistical Mathematics
4-6-7 Minami-Azabu, Minato-ku, Tokyo 106-8569, Japan.
E-mail: †{kimura06,yohei,takashi}@matsumoto.elec.waseda.ac.jp, †††arnaud@ism.ac.jp

Abstract Learning with a dataset that contains both labeled data and unlabeled data is often called a semi-supervised learning problem. In the last decade, the semi-supervised learning problem has become an important research problem in many fields. This article presents a novel semi-supervised learning scheme using a Bayesian Maximum A Posteriori Expectation Maximization (MAP-EM) algorithm with a Dirichlet process prior (stick-breaking representation). The proposed scheme enables us to estimate a mixture model under an unknown number of components and provides a simpler implementation than other implementations such as Markov Chain Monte Carlo (MCMC) implementations. Several examples of a Gaussian mixture are examined to validate the proposed scheme.

Key words learning system, semi-supervised learning, Dirichlet process, stick-breaking representation, EM algorithm, Gaussian mixture

1. Introduction

1.1 Motivation

Learning from a dataset containing labeled data and unlabeled data is known as the semi-supervised learning problem [1] [2] [3]. These problems can be found in many fields, including signal processing, image processing, pattern recognition, and machine learning.

In this article, we propose a novel semi-supervised learning scheme with a stick-breaking process prior [4], which is a representation of the Dirichlet process prior. The proposed scheme is based on a Bayesian Maximum A Posteriori (MAP) approach using an Expectation Maximization (EM) algorithm [5] for a stick-breaking process prior, which enables us to estimate a mixture model under an unknown number of components. It also provides a simpler implementation for a mixture model with a stick-breaking process prior than other implementations such as Markov Chain Monte Carlo (MCMC). Several examples with Gaussian mixture models are examined to evaluate the proposed scheme.

1.2 Related work

There are many implementations of the Diriclet process prior [6] [7]. Several MCMC implementations for this prior can be found in [8] [9] [10] [11]. There are also several alternative implementations for this prior using the Variational Bayesian approach [12] or the Sequential Monte Carlo approach [13]. Although these implementations can be extended to semi-supervised learning (e.g., [14]), the MAP-EM implementation based on

our previous study [15] [16] [17] enabled us to realize a simpler and faster procedure.

2. Dirichlet process EM-algorithm for semi-supervised learning

2.1 Semi-supervised learning

Let $Y := (y_1, \cdots, y_T)$ be independent and identically distributed d-dimensional random variables that can be grouped into N_C groups. The dataset Y consists of both data y_i with group label $c_i (\in \{1, \cdots, N_C\})$ (labeled data), and data y_i without group label c_i (unlabeled data). Under such conditions, learning the dataset is often called the semi-supervised learning problem.

2.2 Bayesian mixture model

The model for such a dataset can be described as

$$p(Y, C_{obs}|\Theta) := \prod_{i=1}^{T} f(y_i, c_i; \Theta), \tag{1}$$

where

$$f(y, c; \Theta) := \begin{cases} p(c|\Theta)p(y|c, \Theta) & (\text{If } c \text{ is given}) \\ p(y|\Theta) & (\text{otherwise}) \end{cases} \tag{2}$$

C_{obs} stands for observed label variables, Θ stands for all parameters of this model, and $p(y|\Theta)$ is the marginal distribution of y given by

$$p(y|\Theta) = \sum_{c} p(c|\Theta)p(y|c, \Theta). \tag{3}$$

Obviously, this model can be considered as a mixture model with group component distributions $p(y|c, \Theta)$. Typically, the probability of belonging to the group $p(c|\Theta)$ is defined as

$$p(c|\Theta) := \mathcal{M}ulti(c; \rho), \tag{4}$$

where $\mathcal{M}ulti(\cdot; \rho)$ is a multinomial distribution with parameter vector $\rho := (\rho_1, \cdots, \rho_{N_C})$ under $\rho_i \geqq 0$ and $\sum_{i=1}^{N_C} \rho_i = 1$. In many Bayesian approaches, the (prior) distribution of parameter vector ρ is set as a natural conjugate Dirichlet prior, i.e.,

$$p(\rho) := \mathcal{D}ir(\rho; \gamma), \tag{5}$$

where $\mathcal{D}ir(\cdot; \gamma)$ is the Dirichlet distribution with the parameter γ, which is set as $\gamma := (1, \cdots, 1)$ in many cases. This setting is also used in the experiments described later.

When the group component distributions are expected to have simple shapes, the standard probability distribution families (such as normal distribution) are used for group component distribution $p(y|c, \Theta)$ in many cases. In contrast, the group component distributions should be more flexible when the group component distributions are expected to have complicated shapes, as in many real data cases. To realize such more flexible settings, one way that has been considered is to also define the group component distributions as mixture models:

$$p(y|c, \Theta) = \sum_{k=1}^{n_c} \pi_{c,k} h(y; \theta_{c,k}). \tag{6}$$

Here, $h(y; \theta_{c,k})$ is the k-th component distribution of the group component distributions $p(y|c, \Theta)$; let us call it the lower component distribution. The variable $\theta_{c,k}$ is its parameter, and $\pi_{c,k} (\in \mathbb{R})$ represents its mixing ratio.

In this paper, we propose a flexible Bayesian approach for this model with a stick-breaking process prior based on our group's previous study [15], which enables us to avoid deciding the number of components n_c. More details are given below.

2.2.1 Group component distribution $p(y|c, \Theta)$

For a flexible Bayesian approach, we use a stick-breaking process prior (a stick-breaking representation of the Dirichlet process prior), as mentioned in our previous study [15]. By using a stick-breaking process prior, the group component distribution $p(y|c, \Theta)$ can be written by using an infinite mixture model:

$$p(y|c, \Theta) := \sum_{k=1}^{\infty} \pi_{c,k} h(y; \theta_{c,k}), \tag{7}$$

where the mixing ratio $\pi_{c,k}$ is described by using another variable $v_{c,k} (\in \mathbb{R})$ as

$$\pi_{c,k} := \begin{cases} v_{c,k} \prod_{l=1}^{k-1} (1 - v_{c,l}) & (k \geqq 2) \\ v_{c,k} & (k = 1) \end{cases}, \tag{8}$$

the (prior) distributions of $v_{c,k}$ are defined as

$$p(v_{c,k}) := \mathcal{B}e(v_{c,k}; 1, \alpha), \tag{9}$$

Here, $\mathcal{B}e(\cdot)$ is a beta distribution, and $\alpha (\in \mathbb{R})$ is a hyperparameter representing the scale parameter of the Dirichlet process. Instinctively, the hyperparameter α corresponds to the "shrink" level of the mixing ratios of the redundant lower component distributions [15].

To implement such a model in the following section, considering that the expectation of the mixing ratio $\pi_{c,k}$ rapidly decreases corresponding to the lower component index k, it can be assumed that lower compo-

nents are truncated at sufficiently large index.

Note that, by using a latent variable z corresponding to the index of the lower component distribution $h(y;\theta_{c,k})$, the group component distribution (7) can also be represented as

$$p(y|c,\Theta) := \sum_z p(z|c,\Theta)p(y|z,c,\Theta), \tag{10}$$

where

$$p(y|z,c,\Theta) = h(y;\theta_{c,k}),$$
$$p(z|c,\Theta) = \mathcal{M}ulti(z;\pi_c),$$

and $\pi_c := (\pi_{c,1}, \pi_{c,2}, \cdots, \pi_{c,\infty})$.

2.2.2 Lower component distribution $h(y;\theta_{c,k})$

In the experiments described in this paper, we defined the component distribution $h(y;\theta_{c,k})$ as a Gaussian distribution, i.e.,

$$h(y;\theta_{c,k}) := \mathcal{N}(y;\theta_{c,k}). \tag{11}$$

Here $\mathcal{N}(\cdot)$ represents the Gaussian distribution. The parameter $\theta_{c,k}$ can be represented by

$$\theta_{c,k} = (m_{c,k}, \Sigma_{c,k}), \tag{12}$$

where $m_{c,k}(\in \mathbb{R}^d)$ is a mean vector of the lower component, and $\Sigma_{c,k}(\in \mathbb{R}^{d\times d})$ is the covariance matrix.

In this paper, for a convenient Bayesian implementation described later, the prior distribution of the parameter $\theta_{c,k}$ is defined as a natural conjugate prior given by

$$p(\theta_{c,k}) := p(m_{c,k}|\Sigma_{c,k})p(\Sigma_{c,k}), \tag{13}$$
$$p(m_{c,k}|\Sigma_{c,k}) := \mathcal{N}(m_{c,k};\mu,\lambda^{-1}\Sigma_{c,k}), \tag{14}$$
$$p(\Sigma_{c,k}) := \mathcal{IW}(\Sigma_{c,k};n_0,R_0), \tag{15}$$

where $\mu(\in \mathbb{R}^d)$, $\lambda(\in \mathbb{R})$, $n_0(\in \mathbb{R})$, and $R_0(\in \mathbb{R}^{d\times d})$ are hyperparameters of the natural conjugate prior, and $\mathcal{IW}(\cdot)$ denotes the inverted Wishart distribution.

2.3 Maximum A Posteriori approach and its implementation

Under such settings of the Bayesian model, the posterior distribution can be described as

$$p(\Theta|Y,C_{obs}) = \frac{p(Y,C_{obs}|\Theta)p(\Theta)}{\int p(Y,C_{obs}|\Theta)p(\Theta)d\Theta}. \tag{16}$$

where $\Theta = (\rho, \{v_{c,k}\}, \{\theta_{c,k}\})$, and the prior $p(\Theta)$ is

$$p(\Theta) = p(\rho)\prod_{c=1}^{N_c}\prod_{k=1}^{\infty} p(v_{c,k})p(\theta_{c,k}), \tag{17}$$

For a simple and fast Bayesian implementation for semi-supervised learning modelling, we consider the Maximum A Posteriori (MAP) approach, i.e., only the most probable parameter,

$$\Theta_{\mathrm{MAP}} := \arg\max_{\Theta} p(Y,C_{obs}|\Theta)p(\Theta). \tag{18}$$

is used for the data modelling.

To estimate the most probable parameter Θ_{MAP}, we consider an Expectation Maximization (EM) implementation based on the algorithm in [15]. As mentioned in Section 2.2.1, it is assumed that lower components are truncated at sufficiently large index K. To describe details of this MAP-EM implementation, let us define a MAP-EM Q function as

$$Q(\Theta;\Theta') := \sum_{i=1} q_{y_i}(\Theta;\Theta'), \tag{19}$$
$$q_y(\Theta;\Theta') := \sum_c\sum_z r(c;y,\Theta')p(z|y,c,\Theta') \times \log p(y|z,c,\Theta)p(z|c,\Theta)p(y|c,\Theta)p(c|\Theta)p(\Theta), \tag{20}$$

where

$$p(z|y,c,\Theta) = \frac{p(y|z,c,\Theta)p(z|c,\Theta)}{\sum_z p(y|z,c,\Theta)p(z|c,\Theta)}, \tag{21}$$

$$r(c;y,\Theta) := \begin{cases} I(c=c') & (\text{If } c \text{ is given}) \\ \frac{p(y|c,\Theta)p(c|\Theta)}{\sum_c p(y|c,\Theta)p(c|\Theta)} & (\text{otherwise}) \end{cases} \tag{22}$$

Here $I(\cdot)$ is an indicator function, c' is the observed group index and Θ' denotes the current parameter of the model. By using this MAP-EM Q function (19), our MAP-EM implementation can be summarized as follows.

Procedure of the MAP-EM algorithm

(1) Initialize Θ^{new}.

(2) Repeat the following 2 steps until the Q function converges.

E-step: Evaluate the Q function: $Q(\Theta;\Theta')$ by using Eqn. (19) after Θ' is set to $\Theta' \leftarrow \Theta^{\mathrm{new}}$.

M-step: Maximize the Q function: $Q(\Theta;\Theta')$ to update the parameter $\Theta^{\mathrm{new}} \leftarrow \arg\max_\Theta Q(\Theta;\Theta')$.

Specifically, the procedure in the E-step and M-step are described as follows.

Procedure in E-step

To evaluate the Q function, compute following equations.

$$O_{c,k} := p(z_n = k | c_n = c, y_n, \Theta') \tag{23}$$

$$O_{T,c,k} := \sum_{n=1}^{T} O_{c,k} \tag{24}$$

$$M_{T,c,k} := \sum_{n=1}^{T} O_{c,k}\, y_n \tag{25}$$

$$S_{T,c,k} := \sum_{n=1}^{T} O_{c,k}\, y_n y_n^T \tag{26}$$

Procedure in M-step

To maximize the Q function, update parameters by following equations.

$$v_{c,k}^{new} = \frac{O_{T,c,k}}{O_{T,c,k} + \sum_{l=k+1}^{K} O_{T,c,l} + \alpha - 1} \tag{27}$$

$$m_{c,k}^{new} = \frac{M_{T,c,k} + \lambda\mu}{O_{T,c,k} + \lambda} \tag{28}$$

$$\Sigma_{c,k}^{new} = \frac{\hat{S}_{c,k}}{O_{T,c,k} + n_0 + d + 2} \tag{29}$$

$$\pi_{c,k}^{new} = \begin{cases} v'_{c,k} \prod_{l=1}^{k-1} (1 - v'_{c,l}) & (k \geqq 2) \\ v'_{c,1} & (k = 1) \end{cases} \tag{30}$$

where

$$\hat{S}_{c,k} = \begin{array}{l} S_{T,c,k} - m_{c,k}^{new} M_{T,c,k}^T \\ -M_{T,c,k} m_{c,k}^{newT} \\ +m_{c,k}^{new} m_{c,k}^{newT} O_{T,c,k} \\ +R_0^{-1} + \lambda (m_{c,k}^{new} - \mu)(m_{c,k}^{new} - \mu)^T \end{array} \tag{31}$$

2.4 Hyperparameter settings

In this subsection, we describe the settings of hyperparameters that are used in the experiments shown in the next section. In the stick-breaking process prior (9), the scale parameter α is set to 2.0. The hyperparameters in the natural conjugate priors for the lower components (13)–(15) are set to $\lambda = 0.5$, $\mu = 0_d$, $n_0 = d$, and $R_0 = I_d$. Here 0_d is a d dimensional zero vector, and I_d represents a unit matrix of size $d \times d$. For the natural conjugate Dirichlet prior (4) for the mixing ratio ρ, the hyperparameter γ is set to $\gamma := (1, \cdots, 1)$, as mentioned before.

3. Experimental results

To evaluate the proposed scheme, we performed several numerical experiments based on two group component distributions.

3.1 Example 1: Two-dimensional example

First, we consider a two-dimensional synthetic example. The dataset is generated by the following equations:

$$y_t = ((a + r_t)\cos(s_t), (a + r_t)\sin(s_t))^T, \tag{32}$$

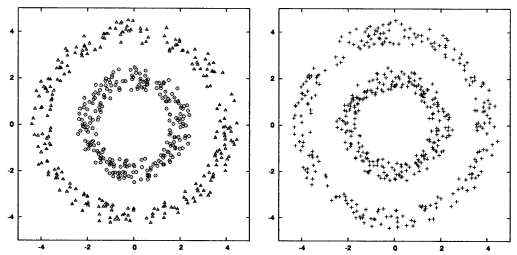
$$r_t \sim i.i.d.U(-0.5, 0.5), \quad s_t \sim i.i.d.U(0, 2\pi), \tag{33}$$

where

$$a := \begin{cases} 2.0 & \text{(group 1)} \\ 4.0 & \text{(group 2)} \end{cases}, \tag{34}$$

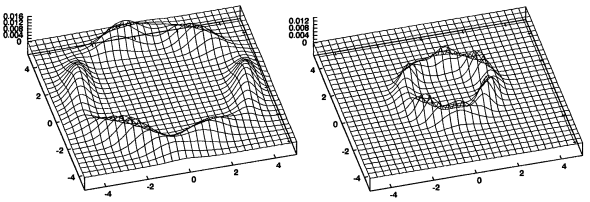
and $U(a, b)$ denotes a continuous uniform distribution with range (a, b). By using these equations, 250 labeled data items and 250 unlabeled data items are generated from both groups (a total of 1000 data items). The generated dataset is shown in Fig. 1.

Fig. 2 shows the group component distribution estimated by the proposed scheme with a sufficient number of iterations. As seen, this result indicates the reasonable capability of the proposed scheme.



(a) Labeled dataset. (b) Unlabeled dataset.

Fig. 1 Dataset for modelling in Example 1. In (a), circles indicate labeled dataset of the 1st group, and triangles the labeled dataset of the 2nd group. (b) The unlabeled dataset from both groups.



(a) 1st group. (b) 2nd group.

Fig. 2 Group component distributions obtained by the proposed scheme in Example 1.

3.2 Example 2: Dual shrinking spirals

In this example, we considered dual shrinking spirals data based on the shrinking spirals data in [18]. The dataset is generated by

$$y_t = u_t + \nu_t, \tag{35}$$

$$u_t = \begin{cases} (r_t \cos s_t, -r_t \sin s_t, 8\pi s_t)^T & \text{(group 1)} \\ (-r_t \cos s_t, r_t \sin s_t, 8\pi s_t)^T & \text{(group 2)} \end{cases} \tag{36}$$

where

$$s_t \sim i.i.d.U(0, 2\pi), \quad \nu_t \sim i.i.d.N(0_3, I_3), \tag{37}$$

and $r_t = 13 - s_t$. By using these equations, 250 labeled data items and 250 unlabeled data items are generated from both groups, the same as in Example 1. The generated dataset is shown in Fig. 3.

Fig. 4 plots the estimated results obtained with the proposed scheme. This figure indicates that the shapes of the dual spiral are clearly obtained by the proposed scheme under the semi-supervised learning settings.

3.3 Example 3: Effect of unlabeled data

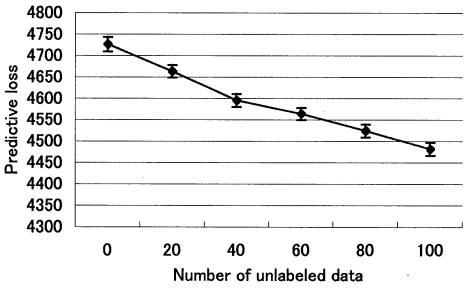
To validate the effect of using of unlabeled data, we considered plural size of two-dimensional dataset generated from the same equations as Section 3.1. The size of datasets is to be set as described in table 1.

Fig. 3.3 shows the predictive loss and accuracy rate obtained by a hundred of times of the experiment for each datasize. In this paper, predictive loss is defined as negative logarithm of likelihood function for the unlabeled test dataset, which is generated by the same equations as the dataset for learning. And accuracy rate is the percentage of correctly classified the test dataset to the group components.

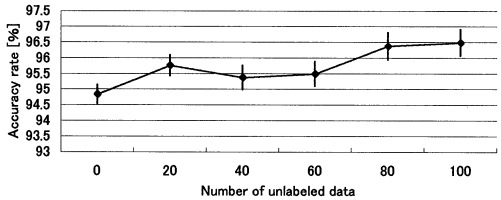
This figure indicates that the predictive loss decreases and the accuracy rate is better as the number of unlabeled data increase.

Table 1 The size of the dataset used in example 3

1st group		2nd group		Total
Labeled	Unlabeled	Labeled	Unlabeled	
50	0	50	0	100
50	10	50	10	120
50	20	50	20	140
50	30	50	30	160
50	40	50	40	180
50	50	50	50	200



(a) Predictive loss.



(b) Accuracy ratio.

Fig. 5 Predictive loss and accuracy rate in example 3.

4. Conclusion

In this article, we proposed a novel semi-supervised learning scheme using a Bayesian Maximum A Posteriori Expectation Maximization (MAP-EM) algorithm with a Dirichlet process prior (stick-breaking representation). Several examples of Gaussian mixtures are examined to evaluate the proposed scheme. The results indicated the capability of the proposed scheme. Especially, in case of the number of labeled data is small, the proposed scheme can be seen advantageous.

REFERENCES

[1] V. K. Mansinghka, D. M. Roy, R. Rifkin and J. Tenenbaum: "Aclass: A simple, online, parallelizable algorithm for probabilistic classification", AISTATS (2007).

[2] B. Krishnapuram, D. Williams, Y. Xue, A. J. Hartemink, L. Carin and M. A. T. Figueiredo: "On semi-supervised classification", NIPS (2004).

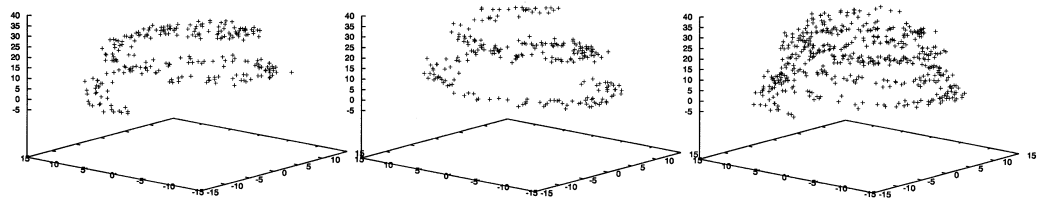
[3] M. Seeger: "Learning with labeled and unlabeled data", Technical report (2001).

[4] J. Sethuraman: "A constructive definition of dirichlet priors", Stat. Sin., **4**, pp. 639–650 (1994).

[5] A. P. Dempster, N. M. Laird and R. D. B.: "Maximum likelihood from incomplete data via the EM algorithm", J. R. Stat. Soc. Ser. B Stat. Methodol., **39**, 1, pp. 1–38 (1977).

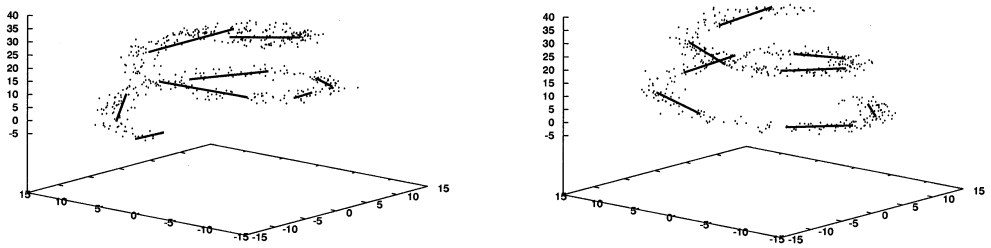
[6] T. S. Ferguson: "A Bayesian analysis of some nonparametric problems", Ann. of Stat., **1**, 2, pp. 209–230 (1973).

[7] C. E. Antoniak: "Mixtures of Dirichlet processes with



(a) Labeled dataset (1st group). (b) Labeled dataset (2nd group). (c) Unlabeled dataset.

Fig. 3 Dual shrinking spirals dataset. (a) and (b) show the labeled dataset from both groups. (c) shows the unlabeled data from both groups.



(a) 1st group. (b) 2nd group.

Fig. 4 Estimated results in Example 2. Lines show estimated eigenvectors of the maximum eigenvalues. Plots show data of both groups.

applications to Bayesian nonparametric problems", Ann. Stat., **2**, 6, pp. 1152–1174 (1974).

[8] L. F. J. H. Ishwaran: "Gibbs sampling methods for stick-breaking priors", J. Am. Stat. Assoc., **96**, pp. 161–173 (2001).

[9] C. E. Rasmussen: "The infinite Gaussian mixture model", Advances in information processing systems 12 (Ed. by S. e. a. Solla), MIT Press, pp. 554–560 (2000).

[10] S. Jain and R. M. Neal: "A split-merge markov chain monte carlo procedure for the dirichlet process mixture model", J. Comput. Graph. Stat., **13**, pp. 158–182 (2000).

[11] Y. Teh, M. Jordan, M. Beal and D. Blei: "Hierarchical Dirichlet processes" (2003).

[12] D. Blei and M. Jordan: "Variational methods for the Dirichlet process" (2004).

[13] P. Fearnhead: "Particle filters for mixture models with an unknown number of components", J. Stat. Comput., **14**, pp. 11–21 (2004).

[14] N. Ueda, T. Yamada and S. Kuwata: "Semi-supervised learning based on Dirichlet process mixture models", Tech. Rep. IEICE. PRMU, **107**, 115, pp. 87–92 (20070621).

[15] T. Kimura, Y. Nakada, T. Matsumoto and A. Doucet: "Recursive em algorithm for parameter estimation in mixture models", to appear.

[16] T. Tokuda, T. Kimura, Y. Nakada and T. Matsumoto: "Maximum A posteriori Estimation For Language Models", IEICE Technical Report Pattern recognition and media understanding (2009).

[17] Y. Goto, T. Kimura, A. Matsui, Y. Nakada and T. Matsumoto: "Face Detection with Dirichlet Process EM", IEICE Technical Report Pattern recognition and media understanding (2009).

[18] N. Ueda, R. Nakano, Z. Ghahramani and G. E. Hinton: "SMEM algorithm for mixture models", Neural Comput., **12**, 9, pp. 2109–2128 (2000).