Robust Estimation for Mixture of Probability Tables based on β-likelihood

Yu Fujimoto* Noboru Murata*

Abstract

Modeling of a large joint probability table is problematic when its variables have a large number of categories. In such a case, a mixture of simpler probability tables could be a good model. And the estimation of such a large probability table frequently has another problem of data sparseness. When constructing mixture models with sparse data, EM estimators based on the β-likelihood are expected more appropriate than those based on the log likelihood. Experimental results show that a mixture model estimated by the β-likelihood approximates a large joint probability table with sparse data more appropriately than EM estimators.

Keywords: EM estimation, mixture model, sparse data, robustness, β-likelihood.

1 Introduction.

In categorical data analysis, sample-based parametric models are often used to describe data relationships [1]. While modeling large complex data, however, we are bothered with following two problems. The first problem is model fitness: models might be "too fitted" (over-fitted) or "unfitted" (under-fitted) and none of them are "appropriate". This is caused by adopting models with not appropriate complexity. The second problem is data sparseness. In a practical situation, the size of samples tends to be insufficient, especially when the model possesses a quite large number of parameters.

Let X and Y be two categorical variables, X with I categories, and Y with J categories. The log-linear model is a typical model to analyze the relationship between X and Y. An independency between X and Y is denoted by

$$\mathrm{P}(x_i, y_j) = \mathrm{P}(x_i)\mathrm{P}(y_j) \ , \tag{1.1}$$

where $\mathrm{P}(x_i, y_j)$ is the joint probability of the event $(X, Y) = (x_i, y_j)$ on the contingency table. The independent log-linear model is derived from the logarithm of (1.1). Let N_{ind} be the number of parameters in the independent model, and $N_{\mathrm{ind}} = I + J - 2$. When X and Y are not independent, (1.1) does not describe their relationship sufficiently, then the saturated log-linear model should be considered. Let N_{sat} be the number of parameters to express all the cells in the joint probability table, and $N_{\mathrm{sat}} = I \times J - 1$. Difference between N_{sat} and N_{ind} becomes quite large as I and J increase. If the independent model does not fit data, we hesitantly have to select the saturated model with an extremely large number of parameters. However, the saturated model with small samples may not be appropriate because the number of estimating parameters is larger than sample size. Very large contingency tables often have this sparseness. If variables are ordinal, This problem can be mitigated by adding the natural ordering term to the independent model, which is called linear-by-linear association model [1]. However, we cannot construct such a model for nominal variables.

On the other hand, Gaussian mixture models are frequently adopted to approximate probability distributions of continuous variables. The advantage of Gaussian mixture models is their flexibility beyond the class of Gaussian distribution. In this paper, we adopt mixture models which reduce the number of parameters when the probability table is large. We also show experimentally that the estimation of mixture models based on sparse data works appropriately with a robust method based on a modified likelihood.

In Section 2, mixture models of probability tables are presented. Section 3 shows a problem of estimating joint probability tables with small sample data sets and describes an estimation method to avoid the above-mentioned sparseness problem. Section 4 shows some experimental results. Finally, concluding remarks are given in Section 5.

2 Mixture of Probability Tables.

A mixture of simple probability tables is constructed with various components. In this section, we introduce two type of components: an independent table and a category integration table.

2.1 Examples of components. Independencies between discrete variables given by (1.1) can be simple components of mixture models. Let K be the number of prepared independent tables to express the joint probability distribution between X and Y. The joint

*Waseda University.

probability $\mathrm{P}(x_i, y_j)$ is described as

$$\mathrm{P}(x_i, y_j) = \sum_{k=1}^{K} \pi_k \mathrm{P}_k(x_i)\mathrm{P}_k(y_j) \quad , \tag{2.2}$$

where $\mathrm{P}_k(x_i)$ and $\mathrm{P}_k(y_j)$ are marginal probability distributions contained in the k-th independent table, and π_k is a weight parameter of the k-th component which satisfies $\sum_{k=1}^{K} \pi_k = 1$, and $\pi_k > 0$ for all k. Equation (2.2) has $N_{\mathrm{mix}}(K) = (I + J - 1)K - 1$ parameters. For fixed K, $N_{\mathrm{mix}}(K)$ is much smaller than N_{sat} for large I and J. Therefore mixture models whose K satisfies

$$N_{\mathrm{ind}} < N_{\mathrm{mix}}(K) < N_{\mathrm{sat}}, \tag{2.3}$$

have intermediate fitness and flexibility between the independent model and the saturated model. If I and J have 100 categories, the independent model and the saturated model have $N_{\mathrm{ind}} = 198$ and $N_{\mathrm{sat}} = 9,999$ parameters: N_{sat} is extremely larger than N_{ind}. In this case, the mixture of independent tables with $K \leq 526$ satisfies (2.3).

The mixture of independent tables in (2.2) is equivalent to the latent class model [1]. From the viewpoint of introducing a latent variable, it is often called the aspect model [5].

A probability table is also simplified by integrating some cells into one category based on similarity. Category integration tables could be components of mixture models. The concept of CMAC [2] is similar to the mixture of category integration tables. Let M_k be the number of cells in the k-th category integration table, that is $M_k < I \times J$. The joint probability on the k-th table is given by $\mathrm{P}_k(x_i, y_j) = \frac{\mathrm{P}_k(m_k)}{c(m_k)}$ where m_k is a category on the k-th table to which (x_i, y_j) belongs, and $c(m_k)$ is the number of cells in the integrated category m_k. The joint probability $\mathrm{P}(x_i, y_j)$ is described as

$$\mathrm{P}(x_i, y_j) = \sum_{k=1}^{K} \pi_k \frac{\mathrm{P}_k(m_k)}{c(m_k)} \quad . \tag{2.4}$$

In this model, joint probability in some cells are grouped into exhaustive partitions on each component. Figure 1 shows an example of the mixture model whose components have uniform grids. The number of cells in each component is small while that in the mixed table is large. In general, components in mixture models have arbitrary cell groupings and need not have uniform grids. Equation (2.4) has $N_{\mathrm{mix}}(K, \mathbf{M}) = \sum_k^K (M_k - 1) + K - 1$ parameters, where $\mathbf{M} = \{M_1, \ldots, M_K\}$. Then, mixture models with moderate $N_{\mathrm{mix}}(K, \mathbf{M})$ also have appropriate fitness and flexibility in the same way as (2.3).

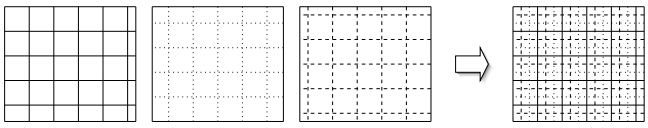


Figure 1: Mixture of category integration tables.

For example of Fig. 1, three components have $\mathbf{M} = \{30, 30, 36\}$ cells. Then the number of parameters in the mixture model is $\sum_{k=1}^{3}(M_k - 1) + 3 - 1 = 95$, and it is smaller than those of the original one $N_{\mathrm{sat}} = 223$.

2.2 Estimation of Mixture of Probability Tables. In general, parameters of the mixture model can be estimated by the EM algorithm [6]. Let $\mathrm{P}(\mathbf{x})$ be the joint probability in the cell $\mathbf{x} \in \mathbf{X}$ that

$$\mathrm{P}(\mathbf{x}; \theta) = \sum_{k=1}^{K} \pi_k \mathrm{P}_k(\mathbf{x}; \phi_k) \quad , \tag{2.5}$$

where ϕ_k is a parameter set in the k-th component and θ is a parameter set in the mixture model, that is, $\theta = (\pi_1, \ldots, \pi_K, \phi_1, \ldots, \phi_K)$. The maximum likelihood (ML) estimation of the mixture model is accomplished by maximizing the log likelihood, and the mean value of log likelihoods is given as

$$\begin{aligned} l(\mathbf{X}; \theta) &= \frac{1}{n} \sum_{\mathbf{x}} n_{\mathbf{x}} \log \mathrm{P}(\mathbf{x}; \theta) \\ &= \frac{1}{n} \sum_{\mathbf{x}} n_{\mathbf{x}} \log \sum_{k=1}^{K} \pi_k \mathrm{P}_k(\mathbf{x}; \phi_k) \quad , \end{aligned} \tag{2.6}$$

where $n_{\mathbf{x}}$ is a count of cell $\mathbf{x}$ in the contingency table and n is a total size of samples, that is, $n = \sum_{\mathbf{x}} n_{\mathbf{x}}$. Since direct maximization of the log likelihood for the mixture model is generally difficult, the EM algorithm is applied. In the EM algorithm, the Q-function is prepared instead of (2.6), and maximized iteratively to obtain estimators.

Let $z_{k\mathbf{x}}$ be unobserved value which describes the posterior probability that a sample $\mathbf{x}$ belongs to the k-th component of the mixture. Then the EM algorithm is constructed with the following Q-function,

$$\begin{aligned} Q(\theta; \theta^{(t)}) &= \mathrm{E}_{\theta^{(t)}}[\log \mathrm{P}(\mathbf{X}, \mathbf{z}; \theta) | \mathbf{X}] \quad . \\ &= \sum_{\mathbf{x}} \sum_{k=1}^{K} n_{\mathbf{x}} \log(\pi_k \mathrm{P}_k(\mathbf{x}; \phi_k))^{z_{k\mathbf{x}}^{(t)}} \quad , \end{aligned} \tag{2.7}$$

where

$$z_{k\mathbf{x}}^{(t)} = \frac{\pi_k^{(t)} \mathrm{P}_k(\mathbf{x}; \phi_k^{(t)})}{\sum_{k=1}^{K} \pi_k^{(t)} \mathrm{P}_k(\mathbf{x}; \phi_k^{(t)})} \quad . \tag{2.8}$$

The EM algorithm in mixture models is shown as follows.

E-step On the $(t+1)$-th iteration, calculate $Q(\theta;\theta^{(t)})$, that is given by (2.7).

M-step Choose $\theta^{(t+1)}$ that is,

$$\theta^{(t+1)} = \arg\max_{\theta} Q(\theta;\theta^{(t)}) \ .$$

The EM algorithm proceeds iteratively until the Q-function converges to the maximum depending on a given initial parameter $\theta^{(0)}$.

When one of the distributions in ϕ_k has probability zero, we add small counts c to it for smoothing the parameter set, and it ensures that the Q-function is bounded. For instance, if $\mathrm{P}_k(X)^{(t+1)}$ has probability zero on the $(t+1)$-th E-step, we manipulate it as $\mathrm{P}_k(x_i)^{(t+1)} = \frac{\mathrm{P}_k(x_i)+c}{\sum_l(\mathrm{P}_k(x_l)+c)}$ for all i. In this paper, the Q-function was calculated with the manipulation $(c = 10^{-6})$ in such cases.

3 Difficulty of the Model Estimation.

The parameter estimation of the probability model often accompanies some difficulties. In this section, we discuss the over-fitting problem on mixture models.

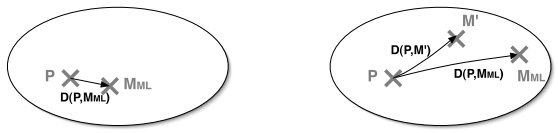
3.1 Sparseness of Data. Mixtures of probability tables show their effectiveness when the number of mixtures satisfies (2.3). In a practical scene, the sample size tends to be insufficient as cells in the estimated table increase, and the probability table has to be estimated with sparse data. In such a case, the model tends to deviate from the true distribution.

Let us measure the distance between the true probability distribution P and the estimated model M with the Kullback-Leibler (KL) divergence, which is

$$\mathrm{D}(P,M) = \sum_{i=1}^{I}\sum_{j=1}^{J} p_{ij}\log\frac{p_{ij}}{m_{ij}} \ ,$$

where $P = \{p_{ij}\}$ and $M = \{m_{ij}\}$ are probability tables, and each has $I \times J$ cells, respectively.

Figure 2 shows the geometrical interpretation of $\mathrm{D}(P,M)$ corresponding to the size of sample data, i.e. how the ML estimators deviate from the true distribution depending on the sample size. With a sufficiently large data set, they are closely located like Fig. 2(a): M_{ML}, which is obtained with the ML estimation, approaches P. With a small data set, P and M_{ML} are located like Fig. 2(b): M_{ML} is far from P because of small samples. The EM estimation, which is based on "maximizing log likelihood", have the same problem. Then, it is preferable to obtain M' by maximizing another type of likelihood, which satisfies $\mathrm{D}(P,M') < \mathrm{D}(P,M_{\mathrm{ML}})$, for small samples.



(a) With large samples. (b) With small samples.

Figure 2: The geometrical interpretation of probability distribution space.

3.2 Robust Estimation with β-likelihood. Fujisawa and Eguchi have proposed an EM-like iterative algorithm in the view of maximizing the β-likelihood instead of the mean log likelihood [3]. The β-likelihood $l_\beta(\mathbf{X};\theta)$ of distribution $\mathrm{P}(\mathbf{X};\theta)$ is given by

$$\begin{aligned} l_\beta(\mathbf{X};\theta) = &\frac{1}{n\beta}\sum_{\mathbf{x}} n_{\mathbf{x}}\mathrm{P}(\mathbf{x};\theta)^\beta \\ &-\frac{1}{1+\beta}\sum_{\mathbf{x}}\mathrm{P}(\mathbf{x};\theta)^{1+\beta} \ , \end{aligned} \tag{3.9}$$

where β is a tuning parameter, that is $0 \le \beta \le 1$. The β-likelihood possesses robustness to outliers depending on the β value. The EM-like algorithm prepares a Q-function based on (3.9) instead of (2.6), and iteratively processes EM steps until it converges. To distinguish from the original EM algorithm based on the log likelihood, we call this EM-like procedure a "β-EM" algorithm.

The Q-function in the β-EM algorithm is modified from (2.7) as

$$\begin{aligned} Q_\beta(\theta;\theta^{(t)}) = &\sum_{\mathbf{x}}\frac{n_{\mathbf{x}}}{n\beta}\left(\prod_{k=1}^{K}(\pi_k\mathrm{P}_k(\mathbf{x};\phi_k))^{z_{k\mathbf{x}}^{(t)}}\right)^\beta \\ &-\frac{1}{1+\beta}\sum_{\mathbf{x}}\left(\prod_{k=1}^{K}(\pi_k\mathrm{P}_k(\mathbf{x};\phi_k))^{z_{k\mathbf{x}}^{(t)}}\right)^{1+\beta} \ , \end{aligned} \tag{3.10}$$

where $z_{k\mathbf{x}}^{(t)}$ is described in (2.8).

Then E-step and M-step in the β-EM algorithm are modified from the original EM as follows.

E-step On the $(t+1)$-th iteration, calculate $Q_\beta(\theta;\theta^{(t)})$, that is given by (3.10).

M-step Choose $\theta^{(t+1)}$ that is,

$$\theta^{(t+1)} = \arg\max_{\theta} Q_\beta(\theta;\theta^{(t)}) \ .$$

Robustness of the β-likelihood can be intuitively explained as follows. Let $g(z)$ be a function which scores fitness of a sample $\mathbf{x}$, where its probability is given by $z = \mathrm{P}(\mathbf{x})$. The difference of the likelihood type is the

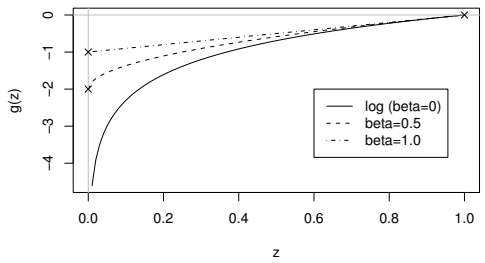


Figure 3: Shapes of $g(z)$.

difference of function $g(\cdot)$ [7]. In the log likelihood, $g(\cdot)$ is given by

$$g(z) = \log(z) \ ,$$

and in the β-likelihood, it is given by

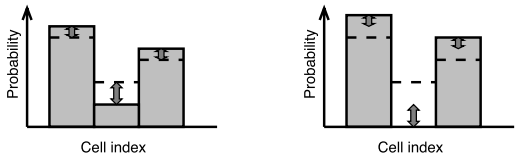
$$g(z) = \frac{z^{\beta} - 1}{\beta} \ .$$

Figure 3 shows shapes of $g(z)$. The log likelihood shows that $g(0) = -\infty$ while β-likelihoods show that $g(0) = -\frac{1}{\beta}$. For a small sample set, the contingency table often has empty cells with $n_{\mathbf{x}} = 0$ where the true probability $\mathrm{P}(\mathbf{x})$ is quite small, and these empty cells are regarded as outliers. Therefore behavior of $g(z)$ where $z \to 0$ is very important for estimation, and robustness of the β-likelihood mainly comes from this feature because the effect of the score with small z is restricted as $g(z) \geq -\frac{1}{\beta}$. Due to robustness of the β-likelihood, the β-EM estimators are expected to be more accurate than the original EM estimators, especially with sparse data.

4 Experimental Results.

In this section, the mixture of independent tables estimated by the β-EM algorithm is evaluated experimentally. At first, robustness of β-EM estimators is analyzed. Secondly, models estimated with the EM method and the β-EM method are compared.

4.1 Robustness of β-EM Estimation. It is difficult to discuss robustness of the probability table estimation generally. We analyzed it experimentally under the special situation. Figure 4 shows an image of the difference between the true distribution and the empirical distribution. With a sufficiently large number of samples, the empirical distribution does not differ much from the true one as shown in Fig. 4(a). However, with a small number of samples, some cells in the empirical probability distribution might become zero as shown in Fig. 4(b). Two typical situations are shown as arrows in Fig. 4(a) and (b), and these two cases were intensively analyzed in our experiment.



(a) With large samples. (b) With small samples.

Figure 4: Image of true distribution (dotted line) and empirical distribution.

Table 1: F.

	x_1	x_2	x_3	x_4
y_1	0.061	0.053	0.058	0.053
y_2	0.093	0.078	0.089	0.080
y_3	0.005	0.001	0.005	0.004
y_4	0.117	0.091	0.113	0.100

Table 2: Example of δ.

	x_1	x_2	x_3	x_4
y_1	0.097	0.085	0.093	0.084
y_2	0.148	0	0	0.128
y_3	0.008	0.002	0.008	0.006
y_4	0	0	0.180	0.159

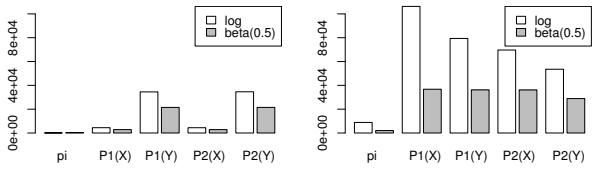
Joint probability distribution F of two variables shown in Table 1 was expressed by a mixture of two independent tables ($K = 2$). Let δ be a zero-contained distribution, which is obtained by modifying 1-4 cells in F with $\mathrm{P}(x_i, y_j) = 0$ and being normalized. We defined two indices to evaluate robustness of estimators based on gross-error-sensitivity (GES) [4], which are given by

$$\gamma_t = \sup_{\delta} \frac{||\theta((1-\varepsilon)F + \varepsilon\delta) - \theta(F)||}{\varepsilon} \ ,$$
$$\gamma_c = \sup_{\delta} \frac{||\theta((1-\varepsilon)\delta + \varepsilon F) - \theta(\delta)||}{\varepsilon}$$

with $\varepsilon = 10^{-5}$, where $\theta(F)$ is an estimator obtained under the distribution F. For a large sample set, the empirical distribution typically becomes like Fig. 4(a), and the largest difference between parameters estimated from F and $((1-\varepsilon)F + \varepsilon\delta)$ is evaluated by γ_t. On the other hand, for a small sample set, the empirical distribution becomes like Fig. 4(b), and the largest difference is evaluated by γ_c. In this experiment, $\theta(F) = \{\pi, \mathrm{P}_1(X), \mathrm{P}_2(X), \mathrm{P}_1(Y), \mathrm{P}_2(Y)\}$, and the difference of two estimators $||\theta(F) - \theta(F')||$ under distributions F and F' is measured based on L_1 norm.

Figure 5 shows γ_t and γ_c values of EM estimators and β-EM estimators with $\beta = 0.5$. In Fig. 5(a), γ_t values of β-EM estimators are slightly better than those of original EM estimators. However, in Fig. 5(b), γ_c values on original EM estimators are extremely larger than those of β-EM. This shows the size of sample sets does not affect β-EM estimators very much, while that would make original EM estimators deviate seriously from those of the true distribution.

From the results, β-EM estimators of mixture models are robust to a small sample set on Table 1 in the sense of GES. Although shown examples are considered under typical situations, β-EM estimators have the robust feature in general.



(a) Empirical GES γ_t. (b) Empirical GES γ_c.

Figure 5: Empirical GES of estimators.

Table 3: Components in P_1.

π:	$\{0.2, 0.4, 0.4\}$
$\mathrm{P}_1(X)$:	$\{1/55, 2/55, \ldots, 10/55\}$
$\mathrm{P}_1(Y)$:	$\{1/55, 2/55, \ldots, 10/55\}$
$\mathrm{P}_2(X)$:	$\{5/60, 4/60, \ldots, 1/60, 1/60, 2/60, \ldots, 5/60\}$
$\mathrm{P}_2(Y)$:	$\{1/60, 2/60, \ldots, 5/60, 5/60, 4/60, \ldots, 1/60\}$
$\mathrm{P}_3(X)$:	$\{2/15, 1/15, 2/15, 1/15, \ldots, 1/15\}$
$\mathrm{P}_3(Y)$:	$\{3/25, 2/25, 3/25, 2/25, \ldots, 2/25\}$.

4.2 Estimation with the β-EM Method. In this experiment, the distance between the mixture model M and the true probability distribution P was measured based on the KL divergence $\mathrm{D}(P, M)$ to evaluate the estimating methods.

At first, 10×10 contingency tables were generated subject to P_1, which was equivalent to the mixture model with three independent tables, that is given as Table 3. Twenty contingency tables were generated for the total sample size of 50, 100, 500 and $1,000$ for each. Then, M, the mixture model with three independent tables, was estimated with each contingency table. In the ideal situation, $\mathrm{D}(P_1, M)$ converges to zero with sufficiently large sample data set. Estimators were given by the EM and the β-EM methods with $\beta = 0.5$. To avoid the local maxima problem, models were estimated five times from randomly generated initial values, and the one which achieves the highest likelihood was adopted for both methods.

Figure 6 shows the relationship between $\mathrm{D}(P_1, M)$ and the size of the sample data set. From the figure, $\mathrm{D}(P_1, M)$ estimated with the EM algorithm is converged to zero as the sample size increases. And results of the β-EM estimation shows similar tendency. It is expected that $\mathrm{D}(P_1, M)$ converges to zero with large sample data sets. However, $\mathrm{D}(P_1, M)$ of the original EM estimation does not show the remarkable improvement where the sample size is 100 or less. This is a problem of the estimation based on a sparse data set as described in Section 3. On the other hand, $\mathrm{D}(P_1, M)$ of the β-EM estimation shows the improvement from those of the EM estimation where the sample size is 50-100 especially. This is due to robustness of the β-likelihood which is maximized in the β-EM algorithm.

Therefore, it is effective to estimate mixture models with the β-EM method when the contingency table does not have a sufficient number of samples. Moreover, $\mathrm{D}(P, M)$ of the β-EM algorithm converges to zero as

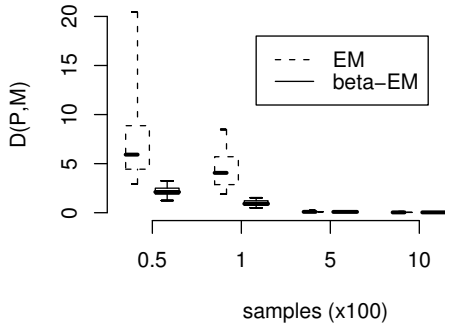


Figure 6: The KL divergence and the sample size.

well as that of the EM algorithm even if its sample size is very large. From the result, the mixture model estimation with the β-EM method is expected to work appropriately in wide situation.

5 Conclusion.

In the paper, we have investigated the robustness of estimators for mixture models based on the β-likelihood for a small sample set. In our experiments, the fixed β value was used though it must have an optimal β with not only robustness but efficiency. In practical situations, we should select a tuning parameter β by cross-validation or something, because well-known information criteria, such as AIC or BIC, or model testing statistics, such as Pearson's X^2 or likelihood-ratio G^2, do not always evaluate models correctly with a small sample set.

To discuss the robustness analytically in general situation, we should derive influence function of the β-EM estimator. We expect that the β-EM method for a small sample set works differently from the other method of avoiding the over-fitting problem, and their behavior should be compared as a further work.

References

[1] A. Agresti. *Categorical Data Analysis*. Wiley Inc., 2 edition, 2002.

[2] J. S. Albus. *Brains, Behavior and Robotics*. BYTE Books, Nov 1981.

[3] H. Fujisawa and S. Eguchi. Robust estimation in the normal mixture model. Research memorandum, ISM, Japan, Feb. 2003.

[4] F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel. *Robust Statistics*. Wiley, 1986.

[5] T. Hofmann. Probabilistic latent semantic analysis. In *Proc. of UAI'99*, 1999.

[6] G. J. McLachlan and T. Krishnan. *The EM Algorithm and Extensions*. Wiley-Interscience, 1996.

[7] N. Murata, T. Takenouchi, T. Kanamori, and S. Eguchi. Information geometry of U-boost and Bregman divergence. *Neural Comp.*, 16:1437–1481, 2004.