Geometric Clustering for Multiplicative Mixtures of Distributions in Exponential Families

(Extended Abstract)

Mary Inaba and Hiroshi Imai
Department of Information Science, University of Tokyo
Hongo, Bunkyo-ku, Tokyo, 113-0033 Japan.
E-mail: {mary,imai}@is.s.u-tokyo.ac.jp

1 Introduction

Estimating unknown parameters from observed data generated by a mixture of distributions in the exponential family is a useful and well-studied problem in statistics [1, 2, 7, 8]. This paper investigates this problem from the viewpoint of computational geometry. We define an analogue to the problem in a setting of geometric clustering, specifically introduce a multiplicative version of the likelihood function for the mixture, and analyze its complexity via proximity relations induced from statistical structures. This extends work on geometric clustering in [4, 5, 6].

2 Exponential Families

A probability distribution determined by a parameter (vector) $\theta = [\theta^i]$ belongs to the exponential family [1, 2, 8] if its probability density function $f(x;\theta)$ with probability variable (vector) x is expressed as

$$f(x;\theta) = \exp[C(x) + \sum_i \theta^i F_i(x) - \psi(\theta)].$$

Example 1 (Normal distribution) *The probability density function of a normal distribution with mean μ and standard deviation σ is given by*

$$\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$
$$= \exp\left(\frac{\mu}{\sigma^2}x - \frac{1}{2\sigma^2}x^2 - \frac{\mu^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)\right)$$

Hence, by setting parameters in the following way,

$$\theta^1 = \frac{\mu}{\sigma^2}, \quad \theta^2 = -\frac{1}{2\sigma^2},$$
$$F_1(x) = x, \quad F_2(x) = x^2, \quad C(x) = 0$$
$$\psi(\theta) = \frac{\mu^2}{2\sigma^2} + \frac{1}{2}\log(2\pi\sigma^2)$$

and this belongs to the exponential family.

Example 2 (Finite discrete distribution) *This is a discrete distribution over a finite set $\{0, 1, \ldots, d\}$ such that the probability for a probabilistic variable x to be 0 is $1 - \sum_{i=1}^{d} \xi_i$, and to be i is ξ_i where*

$$\xi_i > 0 \ (i = 1, \ldots, d), \quad 1 - \sum_{j=1}^{d} \xi_j > 0$$

Setting

$$F_i(x) = \begin{cases} 1 & \text{when } x = i \ (\in \{1, \ldots, d\}), \\ 0 & \text{otherwise} \end{cases}$$
$$\theta^i = \log \frac{\xi_i}{1 - \sum_{j=1}^{d} \xi_j},$$
$$\psi(\theta) = -\log(1 - \textstyle\sum_{j=1}^{d} \xi_j), \quad C(x) = 0$$

this is in the exponential family.

Other examples are Poisson distribution, exponential distribution, etc.

3 Mixture of Exponential Families

With k parameters $\theta^{(j)}$ $(j = 1, \ldots, k)$ and mixing parameters $p_1, \ldots, p_k$ with $\sum_{j=1}^{k} p_j = 1$, $p_j \geq 0$, the mixture of these k distributions is determined by the following probability density function:

$$\sum_{j=1}^{k} p_j f(x; \theta^{(j)})$$

Estimating unknown parameters $\theta^{(j)}$ and p_j from n identically and independently observed data $x^{(l)}$ $(l = 1, \ldots, n)$ generated by this mixture distribution is well-studied problem in statistics. A typical method is to find parameters that maximizes the following likelihood

$$\prod_{l=1}^{n}\left(\sum_{j=1}^{k} p_j f(x^{(l)}; \theta^{(j)})\right)$$

by steepest ascent algorithms, EM algorithm, etc. To determine which among the k distributions generates each observed data, the posterior probability is used, and each observed data is regarded as being generated by the distribution with the maximum posterior probability.

4 Multiplicative Mixture and Classification Likelihood

In this paper, we investigate a discrete version of this problem, which focus on clustering, and its computational complexity. In the above statistically sound model, the mixture is defined by the weighted sum. Instead, we may consider a multiplicative version for mixing the distributions as follows:

$$\prod_{l=1}^{n}\left(\max_{s_{j,l}\geq 0,\ \sum_j s_{j,l}=1}\prod_{j=1}^{k}(p_j f(x^{(l)};\theta^{(j)}))^{s_{j,l}}\right)$$

where $p_j = \frac{1}{n}\sum_{l=1}^{n} s_{j,l}$. $s_{j,l}$ denotes a ratio for data $x^{(l)}$ to belong to cluster j, and this multiplicative version of the likelihood function is maximized over all possible $s_{j,l}$.

This multiplicative version might look artificial, but, restricting $s_{j,l}$ to be 0 or 1, this reduces to the classification likelihood function [3, 7] defined by

$$\prod_{l=1}^{n}\left(\max_{j=1}^{k}(p_j f(x^{(l)};\theta^{(j)}))\right)$$

The classification likelihood function has some computational merits over the additive mixture model, and, the above-defined multiplicative mixture likelihood function shares similar properties.

5 Our Results

We now state our main theorem for this problem.

Theorem 1 *For an optimal $\hat{s}_{j,l}$, $\hat{p}_j$ and $\hat{\theta}^{(j)}$ that maximizes the multiplicative mixture likelihood function, define $J(l)$ $(l = 1, \ldots, n)$ by*

$$J(l) = \{j' \mid \hat{p}_{j'} f(x^{(l)};\hat{\theta}^{(j')}) = \max p_j f(x^{(l)};\theta^{(j)})\}.$$

Then, $\hat{s}(j'', l) = 0$ $(j'' \notin J(l))$.

Theorem 2 *When an optimal $\hat{s}_{j,l}$ that maximizes the multiplicative mixture likelihood function takes values of 0 or 1, the problem of maximizing the multiplicative likelihood function can be solved in $O(n^{d+1})$ time.*

These theorems have strong connection with the Voronoi diagram determined by the divergence in information geometry [5, 6, 9].

References

[1] S. Amari: *Differential Geometrical Methods in Statistics.* Lecture Notes in Statistics, Vol.28, Springer-Verlag, New York, 1985.

[2] O. Barndorff-Nielsen: *Information and Exponential Families in Statistical Theory.* John Wiley & Sons, Inc., 1978.

[3] G. Celeux and G. Govaert: A Classification EM Algorithm for Clustering and Two Stochastic Versions. *Computational Statistics & Data Analysis*, Vol.14 (1992), pp.315–332.

[4] M. Inaba and H. Imai: Geometric Clustering Models for Multimedia Databases. *Proceedings of the 10th Canadian Conference on Computational Geometry*, 1998, pp. 110–111.

[5] M. Inaba, H. Imai and K. Sadakane: Geometric Clustering on Feature Manifold. Preprint, 1998.

[6] M. Inaba, N. Katoh and H. Imai: Applications of Weighted Voronoi Diagrams and Randomization to Variance-Based k-Clustering. *Proceedings of the 10th ACM Symposium on Computational Geometry*, 1994, pp.332–339.

[7] B. Mirkin: *Mathematical Classification and Clustering.* Nonconvex Optimization and Its Applications, Vol.11, Kluwer Academic Publishers, Boston, 1996.

[8] M. K. Murray and J. W. Rice: *Differential Geometry and Statistics.* Chapman & Hall, 1993.

[9] K. Onishi and H. Imai: Voronoi Diagrams for an Exponential Family of Probability Distributions in Information Geometry. *Japan-Korea Joint Workshop on Algorithms and Computation*, Fukuoka, 1997, pp.1–8.