

Critical Lines in Symmetry of Mixture Models and its Application to Component Splitting

Kenji Fukumizu*

Institute of Statistical Mathematics

fukumizu@ism.ac.jp

Shotaro Akaho

National Institute of Advanced Industrial Science and Technology

s.akaho@aist.go.jp

Shun-ichi Amari

RIKEN Brain Science Institute

amari@brain.riken.go.jp

July 9, 2002

Abstract

We show the existence of critical points as lines for the likelihood function of mixture-type models. They are given by embedding a critical point for models with less components. A sufficient condition that the critical line gives local maxima or saddle points is also derived. Based on this fact, a component-split method is proposed for mixture of Gaussian components, and its effectiveness is verified through experiments.

1 Introduction

The likelihood function of a mixture model often has a complex shape so that calculation of an estimator can be difficult, whether maximum likelihood or Bayesian approach is used. In maximum likelihood estimation, convergence of EM algorithm to the global maximum is not guaranteed in many cases, while it is a standard method. Study of likelihood function for mixture models is very important to develop effective methods for learning.

This paper discusses critical points of the likelihood function for mixture-type models by analyzing their hierarchical symmetric structure. As generalization of [1], we show that, given a critical point of the likelihood for the model

*URL <http://www.ism.ac.jp/~fukumizu/>. Currently visiting Department of Statistics, University of California, Berkeley. *fukumizu@stat.berkeley.edu*

with $(H - 1)$ components, duplication of a component gives critical points as lines for the model with H components. We call them *critical lines* of mixture models. We derive also a sufficient condition that the critical lines gives maxima or saddle points of the larger model, and show that in mixture of Gaussian components, given a maximum of likelihood function, split of a component can give an ascending direction of the likelihood. Based on this theory, we propose a stable method of splitting a component, which works effectively with EM optimization for avoiding the dependency on the initial condition and improving the optimization. The usefulness of the algorithm is verified through experiments.

2 Hierarchical Symmetry and Critical Lines of Mixture Models

2.1 Symmetry of Mixture models

Suppose $f_H(x | \boldsymbol{\theta}^{(H)})$ is a mixture model with H components, defined by

$$f_H(x | \boldsymbol{\theta}^{(H)}) = \sum_{j=1}^H c_j p(x | \beta_j), \quad c_j = \frac{\alpha_j}{\alpha_1 + \dots + \alpha_H}, \quad (1)$$

where $p(x | \beta)$ is a probability density function with parameter β . We write for simplicity $\boldsymbol{\alpha}^{(H)} = (\alpha_1, \dots, \alpha_H)$, $\boldsymbol{\beta}^{(H)} = (\beta_1, \dots, \beta_H)$, and $\boldsymbol{\theta}^{(H)} = (\boldsymbol{\alpha}^{(H)}; \boldsymbol{\beta}^{(H)})$.

The key of our discussion is the following two properties;

(S-1) $f_H(x | \boldsymbol{\alpha}^{(H)}; \boldsymbol{\beta}^{(H-2)}, \beta_{H-1}, \beta_{H-1}) = f_{H-1}(x | \boldsymbol{\alpha}^{(H-2)}, \alpha_{H-1} + \alpha_H; \boldsymbol{\beta}^{(H-1)})$.

(S-2) There exists a function $A(\boldsymbol{\alpha})$ such that for $j = H - 1$ and H ,

$$\frac{\partial f_H}{\partial \beta_j}(x | \boldsymbol{\alpha}^{(H)}; \boldsymbol{\beta}^{(H-2)}, \beta_{H-1}, \beta_{H-1}) = \frac{\alpha_j}{A(\boldsymbol{\alpha})} \frac{\partial f_{H-1}}{\partial \beta_{H-1}}(x | \boldsymbol{\alpha}^{(H-2)}, \alpha_{H-1} + \alpha_H; \boldsymbol{\beta}^{(H-1)}).$$

In mixture models, the function $A(\boldsymbol{\alpha})$ is simply given by $\alpha_1 + \dots + \alpha_H$.

Hereafter, we discuss in general a model with the assumptions (S-1) and (S-2). The results in Section 2.1 and 2.2 depend only on these assumptions¹. While in mixture models similar conditions are satisfied with any choices of two components, we describe only the case of $H - 1$ and H just for simplicity. We write Θ_H for the space of parameter $\boldsymbol{\theta}^{(H)}$.

Another example which satisfies (S-1) and (S-2) is Latent Dirichlet Allocation (LDA, [2]), which models data with group structure (e.g. document as a set of words). For $x = (x_1, \dots, x_M)$, LDA with H components is defined by

$$f_H(x | \boldsymbol{\theta}^{(H)}) = \int_{\Delta_{H-1}} \mathcal{D}_H(\mathbf{u}^{(H)} | \boldsymbol{\alpha}^{(H)}) \prod_{\nu=1}^M \left(\sum_{j=1}^H u_j p(x_\nu | \beta_j) \right) d\mathbf{u}^{(H)}, \quad (2)$$

¹The results do not require $p(x | \beta)$ is a density function. Thus, they are generalization of the results on multilayer neural networks in [1].

where

$$\mathcal{D}_H(\mathbf{u}^{(H)}|\boldsymbol{\alpha}^{(H)}) = \frac{\Gamma(\sum_j \alpha_j)}{\prod_j \Gamma(\alpha_j)} \prod_{j=1}^H u_j^{\alpha_j - 1}$$

is the Dirichlet distribution over the $(H - 1)$ -dimensional simplex Δ_{H-1} . It is easy to see (S-1) and (S-2) for LDA by using Lemma 6 in Appendix. LDA includes mixture models eq.(1) as the special case of $M = 1$.

It is straightforward from (S-1) that, given a parameter $\boldsymbol{\theta}^{(H-1)} = (\boldsymbol{\gamma}^{(H-1)}; \boldsymbol{\eta}^{(H-1)})$ of the model with $(H - 1)$ components and a scalar λ , the parameter $\boldsymbol{\theta}_\lambda \in \Theta_H$ defined by

$$\begin{aligned} \alpha_j &= \gamma_j, & \beta_j &= \eta_j & (1 \leq j \leq H - 2) \\ \alpha_{H-1} &= \lambda\gamma_{H-1}, & \alpha_H &= (1 - \lambda)\gamma_{H-1}, & \beta_{H-1} &= \beta_H = \eta_{H-1} \end{aligned} \quad (3)$$

gives the same function as $f_{H-1}(x|\boldsymbol{\theta}^{(H-1)})$. In mixture models/LDA, this corresponds to duplication of the $(H - 1)$ -th component with partitioning the mixing/Dirichlet parameter in the ratio $\lambda : (1 - \lambda)$. Since λ is arbitrary, a point in the smaller model is embedded into the larger model as a line in the parameter space Θ_H . This implies that the parameter to realize $f_{H-1}(x|\boldsymbol{\theta}^{(H-1)})$ lacks identifiability in Θ_H . Such singular structure of a model causes various interesting phenomena in estimation, learning, and generalization ([3],[4]).

2.2 Critical Lines – Embedding of a Critical Point

Given sample $X^{(1)}, \dots, X^{(N)}$, we consider an objective function for learning:

$$L_H(\boldsymbol{\theta}^{(H)}) = \sum_{n=1}^N \Omega_n(f_H(X^{(n)}; \boldsymbol{\theta}^{(H)})), \quad (4)$$

where $\Omega_n(f)$ are functions, which may depend on n . The aim of learning is to maximize L_H . If $\Omega_n(f) = \log f$ for all n , this is equal to maximum likelihood estimation.

Suppose $\boldsymbol{\theta}_*^{(H-1)} = (\gamma_1^*, \dots, \gamma_{H-1}^*; \eta_1^*, \dots, \eta_{H-1}^*)$ is a critical point of $L_{H-1}(\boldsymbol{\theta}^{(H-1)})$, that is, $\frac{\partial L_{H-1}}{\partial \boldsymbol{\theta}^{(H-1)}}(\boldsymbol{\theta}_*^{(H-1)}) = 0$. Embedding of this point into Θ_H gives a critical line;

Theorem 1 (Critical Line). *Suppose that a model satisfies (S-1) and (S-2). Let $\boldsymbol{\theta}_*^{(H-1)}$ be a critical point of L_{H-1} with $\gamma_{H-1}^* \neq 0$, and $\boldsymbol{\theta}_\lambda$ be a parameter given by eq.(3) for $\boldsymbol{\theta}_*^{(H-1)}$. Then, $\boldsymbol{\theta}_\lambda$ is a critical point of $L_H(\boldsymbol{\theta}^{(H)})$ for all λ .*

Proof. Although this is essentially the same as Theorem 1 in [1], the following proof gives better intuition. Let $(s, t; \zeta, \xi)$ be reparametrization of $(\alpha_{H-1}, \alpha_H; \beta_{H-1}, \beta_H)$, defined by

$$s = \alpha_{H-1} + \alpha_H, \quad t = \alpha_{H-1} - \alpha_H, \quad \beta_{H-1} = \zeta + \alpha_H \xi, \quad \beta_H = \zeta - \alpha_{H-1} \xi. \quad (5)$$

This is a one-to-one correspondence, if $\alpha_{H-1} + \alpha_H \neq 0$. Note that $\xi = 0$ is equivalent to the condition $\beta_{H-1} = \beta_H$. Let $\boldsymbol{\omega} = (\boldsymbol{\alpha}^{(H-2)}, s, t; \boldsymbol{\beta}^{(H-2)}, \zeta, \xi)$ be the new coordinate, $\ell_H(\boldsymbol{\omega})$ be the objective function eq.(4) under this parametrization, and $\boldsymbol{\omega}_\lambda$ be the parameter corresponding to $\boldsymbol{\theta}_\lambda$. Since we have by definition $\ell_H(\boldsymbol{\omega}) = L_H(\boldsymbol{\alpha}^{(H-2)}, \frac{s+t}{2}, \frac{s-t}{2}; \boldsymbol{\beta}^{(H-2)}, \zeta + \frac{s-t}{2}\xi, \zeta - \frac{s+t}{2}\xi)$, the condition (S-1) means

$$\ell_H(\boldsymbol{\alpha}^{(H-2)}, s, t; \boldsymbol{\beta}^{(H-2)}, \zeta, 0) = L_{H-1}(\boldsymbol{\alpha}^{(H-2)}, s; \boldsymbol{\beta}^{(H-2)}, \zeta). \quad (6)$$

Then, it is clear that the first derivatives of ℓ_H at $\boldsymbol{\omega}_\lambda$ with respect to $(\boldsymbol{\alpha}^{(H-2)}, s, \boldsymbol{\beta}^{(H-2)}, \zeta)$ are equal to those of $L_{H-1}(\boldsymbol{\theta}^{(H-1)})$ at $\boldsymbol{\theta}_*^{(H-1)}$, and they are zero. The derivative $\partial \ell_H(\boldsymbol{\omega}_\lambda) / \partial t$ vanishes from eq.(6), and $\partial \ell_H(\boldsymbol{\omega}_\lambda) / \partial \xi = 0$ from following Lemma 2. \square

Lemma 2. *Let \mathcal{H} be a hyperplane given by $\{\boldsymbol{\omega} \mid \xi = 0\}$. Then, for all $\boldsymbol{\omega}_o \in \mathcal{H}$, we have*

$$\frac{\partial f_H}{\partial \xi}(x; \boldsymbol{\omega}_o) = 0. \quad (7)$$

Proof. Straightforward from the assumption (S-2) and $\frac{\partial}{\partial \xi} = \alpha_H \frac{\partial}{\partial \beta_{H-1}} - \alpha_{H-1} \frac{\partial}{\partial \beta_H}$. \square

Given that a maximum of L_H is larger than that of L_{H-1} , Theorem 1 implies that the function L_H always has critical points which are not global maximum. Those points lie on lines in the parameter space. Further embedding of the critical lines into larger models gives high-dimensional critical planes in the parameter space. This property is very general, and in LDA and mixture models we do not need any assumptions on $p(x \mid \beta)$. In these models, by the permutation symmetry of components, there are many choices for embedding, which provides many critical lines and planes for L_H .

2.3 Embedding of a Maximum Point in LDA and Mixture Models

A next interesting question is whether the critical lines formed by a maximum of L_{H-1} gives maxima of L_H . The answer requires information on second derivatives and depends on models. We show a general result on LDA, and that on mixture models as its corollary.

Theorem 3. *Suppose that the model is LDA defined by eq.(2). Let $\boldsymbol{\theta}_*^{(H-1)}$ be an isolated maximum point of L_{H-1} , and $\boldsymbol{\theta}_\lambda$ its embedding given by eq.(3). Define a symmetric matrix R of the size $\dim \beta$ by*

$$R = \sum_{n=1}^N \Omega'_n (f_{H-1}(X^{(n)} \mid \boldsymbol{\theta}_*^{(H-1)})) \left\{ \sum_{\mu=1}^M I_\mu^{(n)} \frac{\partial^2 p(X_\mu^{(n)} \mid \eta_{H-1}^*)}{\partial \beta \partial \beta} \right. \\ \left. + \frac{1}{\sum_{j=1}^{H-1} \gamma_j^* + 1} \sum_{\mu=1}^M \sum_{\tau \neq \mu}^M J_{\mu, \tau}^{(n)} \frac{\partial p(X_\mu^{(n)} \mid \eta_{H-1}^*)}{\partial \beta} \frac{\partial p(X_\tau^{(n)} \mid \eta_{H-1}^*)}{\partial \beta} \right\},$$

where

$$I_\mu^{(n)} = \int_{\Delta_{H-2}} \mathcal{D}_{H-1}(\mathbf{u} | \gamma_1^*, \dots, \gamma_{H-2}^*, \gamma_{H-1}^* + 1) \prod_{\nu \neq \mu} (\sum_{j=1}^{H-1} u_j p(X_\nu^{(n)} | \beta_j)) d\mathbf{u}^{(H-1)},$$

$$J_{\mu,\tau}^{(n)} = \int_{\Delta_{H-2}} \mathcal{D}_{H-1}(\mathbf{u} | \gamma_1^*, \dots, \gamma_{H-2}^*, \gamma_{H-1}^* + 2) \prod_{\nu \neq \mu, \tau} (\sum_{j=1}^{H-1} u_j p(X_\nu^{(n)} | \beta_j)) d\mathbf{u}^{(H-1)}.$$

Then, we have

- (i) If R is negative definite, the parameter θ_λ is a maximum of L_H for all $\lambda \in (0, 1)$.
- (ii) If R has a positive eigenvalue, the parameter θ_λ is a saddle point for all $\lambda \in (0, 1)$.

Remark: The conditions depend only on the parameter $\theta_*^{(H-1)}$.

Proof. We use the parametrization ω defined by eq.(5). For each t , let \mathcal{H}_t be a hyperplane with t fixed, and $\tilde{L}_{H,t}$ be the function L_H restricted on \mathcal{H}_t . The hyperplane \mathcal{H}_t is a slice transversal to the critical line, along which L_H has the same value. Therefore, if the Hessian matrix of $\tilde{L}_{H,t}$ on \mathcal{H}_t is negative definite at the intersection ω_λ ($\lambda = (t+1)/2$), the point is a maximum of L_H , and if the Hessian has a positive eigenvalue, ω_λ is a saddle point.

Since in ω coordinate we have $\tilde{L}_{H,t}(\alpha_1, \dots, \alpha_{H-2}, s; \beta_1, \dots, \beta_{H-2}, \zeta, 0) = L_{H-1}(\alpha_1, \dots, \alpha_{H-2}, s; \beta_1, \dots, \beta_{H-2}, \zeta)$, the Hessian of $\tilde{L}_{H,t}$ at ω_λ is given by

$$\text{Hess} \tilde{L}_{H,t}(\omega_\lambda) = \begin{pmatrix} \text{Hess} L_{H-1}(\theta_*^{(H-1)}) & O \\ O & \frac{\partial^2 \tilde{L}_{H,t}(\omega_\lambda)}{\partial \xi \partial \xi} \end{pmatrix}. \quad (8)$$

The off-diagonal blocks are zero, because we have $\frac{\partial^2 \tilde{L}_{H,t}(\omega_\lambda)}{\partial \xi \partial \omega_a} = 0$ for $\omega_a \neq \xi$ from Lemma 2. By assumption, $\text{Hess} L_{H-1}(\theta_*^{(H-1)})$ is negative definite. Noting that the terms including $\partial f_H(X^{(n)}; \theta_\lambda) / \partial \xi$ vanish from Lemma 2, it is easy to obtain

$$\frac{\partial^2 \tilde{L}_{H,t}(\omega_\lambda)}{\partial \xi \partial \xi} = \frac{\lambda(1-\lambda)(\gamma_{H-1}^*)^3}{\sum_{j=1}^{H-1} \gamma_j^*} R$$

by using Lemma 6 and the definition of ξ . \square

By setting $M = 1$ in LDA model, we have the sufficient conditions for mixture models.

Corollary 4. For a mixture model, the same assertions as Theorem 3 holds for

$$\tilde{R} = \sum_{n=1}^N \Omega'_n(f_{H-1}(X^{(n)} | \theta_*^{(H-1)})) \frac{\partial^2 p(X^{(n)} | \eta_{H-1}^*)}{\partial \beta \partial \beta}. \quad (9)$$

Proof. For $M = 1$, $J_{\mu,\tau}^{(m)} = 0$ and $I^{(n)} = \gamma_{H-1}^* / \sum_{j=1}^{H-1} \gamma_j^*$. The assertion is obvious. \square

In mixture models, given a maximum point $\theta_*^{(H-1)}$, the calculation of \tilde{R} is straightforward. Corollary 4 tells that we can easily see whether the critical line is local maxima or saddle points.

2.4 Critical Lines in Various Models

We further investigate the critical lines for specific models. Hereafter, we assume that the objective function is defined by $\Omega_n(f) = \log f$ for all n .

2.4.1 Gaussian Mixture, Mixture of Factor Analyzers, and Mixture of PCA

Assume that each component is $\phi(\mathbf{x}; \boldsymbol{\mu}, V)$, D -dimensional Gaussian density with mean $\boldsymbol{\mu}$ and variance-covariance matrix V as parameters. The matrix \tilde{R} in eq.(9) has a form

$$\tilde{R} = \begin{pmatrix} S_2 & S_3 \\ S_3^T & S_4 \end{pmatrix}, \quad (10)$$

where S_2 , S_3 , and S_4 correspond to the second derivatives with respect to $(\boldsymbol{\mu}, \boldsymbol{\mu})$, $(\boldsymbol{\mu}, V)$, and (V, V) , respectively. It is well known that for Gaussian density the second derivative $\partial^2 \phi / \partial \boldsymbol{\mu} \partial \boldsymbol{\mu}$ is equal to the first derivative $\partial \phi / \partial V$. Then, S_2 is equal to zero by the critical point condition. If the data is randomly generated, S_3 and S_4 are of full rank almost surely. This type of matrix necessarily has a positive eigenvalue. It is not difficult to extend this discussion to models with scalar or diagonal covariance matrices for variable parameters.

Similar arguments hold for mixture of factor analyzers (MFA, [5]) and mixture of probabilistic PCA (MPCA, [6]). In factor analyzers or probabilistic PCA, the variance-covariance matrix is restricted to the form

$$V = FF^T + S,$$

where F is a factor loading of rank k and S is a diagonal or scalar matrix. Because the first derivative of $\phi(\mathbf{x}; \boldsymbol{\mu}, FF^T + S)$ with respect to F is $\frac{\partial \phi(\mathbf{x}; \boldsymbol{\mu}, FF^T + S)}{\partial V} F$, the block in \tilde{R} corresponding to the second derivatives on μ is not of full rank. In a similar manner to Gaussian mixtures, \tilde{R} has a positive eigenvalue. In summary, we have the following

Theorem 5. *Suppose the model is Gaussian mixture, MFA, or MPCA. If \tilde{R} is of full rank, every point θ_λ on the critical line is a saddle point of L_H .*

This theorem means that if we have the maximum likelihood estimator for $H - 1$ components, we can find an ascending direction of likelihood by splitting a component and modifying their means and variance-covariance matrices in the direction of the positive eigenvector. This leads a component splitting method, which will be shown in Section 3.1.

2.4.2 Latent Dirichlet Allocation

We investigate LDA with multinomial components. Using the D -dimensional random vector $x = (x_a) \in \{(1, 0, \dots, 0)^T, \dots, (0, \dots, 0, 1)^T\}$, which indicates a chosen element, the multinomial distribution over D elements is expressed as an exponential family by

$$p(x | \beta) = \prod_{a=1}^D (p_a)^{x_a} = \exp \left\{ \sum_{a=1}^{D-1} \beta^a x_a - \log \left(1 + \sum_{a=1}^{D-1} e^{\beta^a} \right) \right\},$$

where p_a is the expectation of x_a , and $\beta \in \mathbb{R}^{D-1}$ is a natural parameter given by $\beta^a = \log(p_a/p_D)$. It is easy to obtain

$$R = \sum_{n=1}^N \Omega'(f_{H-1}(X^{(n)} | \theta_*^{(H-1)})) \sum_{\mu=1}^M \sum_{\tau \neq \mu} J_{\mu\tau}^{(n)} p(X_\mu^{(n)} | \gamma_{H-1}^*) p(X_\tau^{(n)} | \gamma_{H-1}^*) \\ \times (\tilde{X}_\mu^{(n)} - p_{(H-1)}^*) (\tilde{X}_\tau^{(n)} - p_{(H-1)}^*)^T, \quad (11)$$

where $\tilde{X}_\nu^{(n)}$ is the truncated $(D-1)$ -dimensional vector, and $p_{(H-1)}^* \in (0, 1)^{D-1}$ is the expectation parameter for $(H-1)$ -th component of $\theta_*^{(H-1)}$.

In general, $J_{\mu\tau}^{(n)}$ are intractable in large problems. We explain a simple case of $H = 2$ and $M = D$. Let \hat{p} be the frequency vector of the elements, which is the maximum likelihood estimator for the one multinomial model. In this case, we have $J_{\mu\tau}^{(n)} = 1$ and

$$R = \sum_{n=1}^N \left\{ \sum_{\mu, \tau=1}^M (\tilde{X}_\mu^{(n)} - \hat{p})(\tilde{X}_\tau^{(n)} - \hat{p})^T - \sum_{\mu=1}^M (\tilde{X}_\mu^{(n)} - \hat{p})(\tilde{X}_\mu^{(n)} - \hat{p})^T \right\}.$$

First, suppose we have a data set $X_\nu^{(n)} = e_\nu$ for all n and $1 \leq \nu \leq D = M$, where e_j is the D -dimensional vector with the j -th component 1 and others zero. Then, we have $\hat{p} = (1/D, \dots, 1/D)$ and $\sum_{\mu=1}^D (\tilde{X}_\mu^{(n)} - \hat{p}) = 0$, which means $R < 0$. The critical line gives maxima for LDA with $H = 2$. Next, suppose the data consists of D groups, and every data in the j -th group is given by $X_\nu^{(n)} = e_j$. While we have again $\hat{p} = (1/D, \dots, 1/D)$, the matrix R is $\sum_{j=1}^D (N/D) \times D(D-1)(e_j - \hat{p})(e_j - \hat{p})^T > 0$. Thus, all the points on the critical lines are saddle points. These two examples explain two extreme cases; in the former we have no advantage in using two components because all the data $X^{(n)}$ are the same, while in the latter multiple components fits better to the variety of $X^{(n)}$.

Algorithm 1 : EM with component splitting for Gaussian mixture

1. Initialization: Calculate the sample mean μ_1 and variance-covariance matrix V_1 .
2. $H := 1$.
3. For all $1 \leq h \leq H$, diagonalize V_{h*} as $V_{h*} = U_h \Lambda_h U_h^T$, and calculate \tilde{R}_h according to eq.(13) in Appendix.
4. For $1 \leq h \leq H$, calculate the eigenvector (r_h, W_h) of \tilde{R}_h corresponding to the largest eigenvalue.
5. For $1 \leq h \leq H$, optimize β by line search to maximize the likelihood in

$$\begin{aligned} c_h &= \frac{1}{2}c_{h*}, & \mu_h &= \mu_{h*} - \rho r_h, & V_h &= U_h e^{-\rho W_h} \Lambda_h e^{-\rho W_h} U_h^T, \\ c_{H+1} &= \frac{1}{2}c_{h*}, & \mu_{H+1} &= \mu_{h*} + \rho r_h, & V_{H+1} &= U_h e^{\rho W_h} \Lambda_h e^{\rho W_h} U_h^T. \end{aligned} \quad (12)$$

Let ρ_h^o be the optimizer and L_h be the likelihood.

6. For h_* to give the largest L_h , split h_* -th component according to eq.(12) with $\rho_{h_*}^o$.
 7. Optimize the parameter $\theta^{(H+1)}$ using EM algorithm.
 8. If $H + 1 = \text{MAX_H}$, then END. Otherwise, $H := H + 1$ and go to 3.
-

3 Component Splitting Method in Mixture of Gaussian Components

3.1 EM with Component Splitting

It is well known that EM algorithm suffers from strong dependency on initialization. Moreover, in mixture of Gaussian components, the likelihood is not upper bounded if infinitely small variance is used. We have to use an optimization technique which can find an appropriate maximum. Theorem 5 enables us to derive a stable and effective way of splitting a Gaussian component to increase the likelihood. We propose EM with component splitting, which adds a component one by one after maximizing the likelihood at each size. Ueda et al ([7]) proposes Split and Merge EM algorithm, in which the components repeats split and merge in a triplet, keeping the total number fixed. While their method works well, it requires a large number of trials of EM for candidate triplets, and the splitting method is heuristic. Our splitting method is well based on theory, and EM with splitting gives a series of estimators for all model sizes in a single run.

It is noteworthy that in splitting a component, not only the means but also

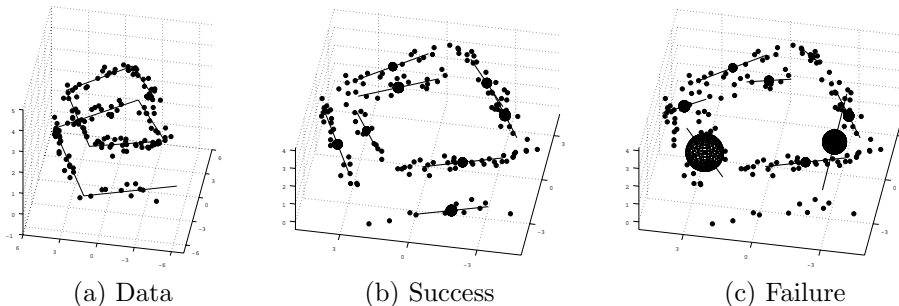


Figure 1: Spiral data. In (b) and (c), the lines represent the vectors F_h and $-F_h$ at the mean values, and the radius of a sphere is the scalar part of the variance.

	EM	EMCS
Best	-534.9 (6 times)	-534.9 (26 times)
Worst	-648.1	-587.9
Av.	-583.9	-541.3

Table 1: Likelihood for spiral data (30 runs). EM is the conventional EM with random initialization, and EMCS is the proposed EM with component splitting.

the variance-covariance matrices must be modified. The simple additive rule $V_{new} = V_{old} + \Delta V$ tends to fail, because it may make the matrix non-positive definite. To solve this problem, we use Lie algebra expression to add a vector of ascending direction. Let $V = U\Lambda U^T$ be the diagonalization of V , and consider $V(W) = Ue^W \Lambda e^W U^T$ for a symmetric matrix W . This gives a local coordinate of the positive definite matrices around $V = V(0)$. Modification of V through W gives a stable way of updating variance-covariance matrices.

Algorithm 1 is the procedure of learning. I show only the case of mixture of Gaussian. The exact algorithms for other models will be shown in a forthcoming paper.

3.2 Experimental results

We show through experiments how the proposed EM with component splitting effectively maximizing the likelihood. In the first experiment, the mixture of PCA with 8 components of rank 1 is employed to fit the synthesized 150 data generated along a piecewise linear spiral (Fig.1). Table 1 shows the results over 30 trials with different random numbers. We use on-line EM algorithm ([8]), presenting data one-by-one in a random order. The EM with random initialization reaches the best state (Fig.1-(b)) only 6 times, while EM with component splitting achieves it 26 times. Fig.1-(c) shows an example of failure.

Next experiment is an image compression problem, in which the standard

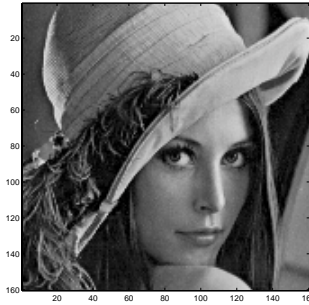


Figure 2: Image data "Lenna".

$\times 10^4$	EM	EMCS
Best	5.94	5.38
Worst	6.40	6.12
Av.	6.15	5.78

Table 2: Residual square errors for "Lenna" (10 runs). EM is the conventional EM with random initialization, and EMCS is the proposed EM with component splitting.

image "Lenna" of 160×160 pixels with 8 bit gray scale (Fig.2) is used. The image is partitioned into 20×20 blocks of 8×8 pixels, and these blocks are regarded as 400 data of 64 dimension. We use the mixture of PCA with 10 components of rank 4 to fit the data, and obtain a compressed image by $\hat{X} = F_h(F_h^T F_h)^{-1} F_h^T X$, where X is a 64 dimensional block and h indicates the component of the shortest Euclidean distance $\|X - \mu_h\|$. Table 2 shows the residual square error (RSE), $\sum_{j=1}^{400} \|X_j - \hat{X}_j\|^2$, which shows the quality of the compression. In both experiments, we can see the better optimization performance of the proposed algorithm.

4 Discussions

In EM with component splitting, we obtain the estimators for up to the specified number of components. We need some model selection technique to choose the best one, which is another important problem. We do not discuss it in this paper because our method can be combined with any techniques, which select a model after obtaining the estimators. However, we should note that famous methods such as AIC and MDL, which are based on statistical asymptotic theory, cannot be applied to mixture models because of the unidentifiability of the parameter. Further study is necessary from both of theory and practice on model selection for mixture models.

In Algorithm 1, in selecting a component to split, we try line search and

choose the one giving the largest likelihood. While this works well in our experiments, the proposed method of component splitting can be combined with any criterions to select a component, and there are other possibilities. One of them is to select the component giving the largest eigenvalue of R_h . In Gaussian mixture models, this method is very natural; the block of second derivatives on V in \tilde{R} is equal to the weighted fourth cumulant, and a component with a large cumulant should be split. However, in mixture of FA and PCA, this does not necessarily work well, because the decomposition $V = FF^T + S$ does not give a natural parametrization. Although this paper discusses only local properties, a method using global information might be more desirable. This is left as a future work.

This paper mainly focuses on mixture models. However, the origin of conditions (S-1) and (S-2) is the singular structure induced by latent variables. There are many other models, such as hidden Markov models, which have latent and symmetric structure. Further study is needed to investigate possible extensions of the results in this paper to other models.

Appendix

A Dirichlet integral

Lemma 6. *Suppose $\varphi_H(\mathbf{u}^{(H)}; \boldsymbol{\beta}^{(H)})$ satisfies the assumption (S-1). Define $I_H(\boldsymbol{\alpha}^{(H)}; \boldsymbol{\beta}^{(H)}) = \int_{\Delta_{H-1}} \varphi(\mathbf{u}^{(H)}; \boldsymbol{\beta}^{(H)}) \mathcal{D}_H(\mathbf{u}^{(H)} | \boldsymbol{\alpha}^{(H)}) d\mathbf{u}^{(H)}$. Then, we obtain*

$$I_H(\boldsymbol{\alpha}^{(H)}; \boldsymbol{\beta}^{(H-2)}, \beta_{H-1}, \beta_{H-1}) = I_{H-1}(\boldsymbol{\alpha}^{(H-2)}, \alpha_{H-1} + \alpha_H; \boldsymbol{\beta}^{(H-1)}).$$

Proof. Direct calculation. \square

B Matrix \tilde{R}_h for Gaussian mixture

We omit the index h for simplicity, and use Einstein's convention. Let $U = (u_1, \dots, u_D)$ and $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_D)$. For $V(W) = Ue^W \Lambda e^W U^T$, we have $\partial V(O)/W_{ab} = (\lambda_a + (1 - \delta_{ab})\lambda_b)(u_a u_b^T + u_b u_a^T)$, where δ_{ab} is Kronecker's delta. Let $T^{(3)}$ and $T^{(4)}$ be the weighted third and fourth sample moments, respectively, with weights $\frac{\phi(x^{(n)}; \mu_*, V_*)}{f^{(H-1)}(x^{(n)}; \boldsymbol{\theta}_*^{(H-1)})}$. $\tilde{T}_{(3)}$ and $\tilde{T}_{(4)}$ are defined by $\tilde{T}_{(3)}^{abc} = V^{ap} V^{bq} V^{cr} T_{pqr}^{(3)}$ and $\tilde{T}_{(4)}^{abcd} = V^{ap} V^{bq} V^{cr} V^{ds} T_{pqrs}^{(4)}$, respectively, where V^{ap} is the (ap) -component of V^{-1} . Then, direct calculation leads that the matrix $\tilde{R} = \begin{pmatrix} O & B \\ B^T & C \end{pmatrix}$, where the decomposition corresponds to $\beta = (\mu, W)$, is given by

$$\begin{aligned} B_{\mu_a, W_{bc}} &= (\lambda_b + (1 - \delta_{bc})\lambda_c) u_b^T \tilde{T}_{(3)}^{..a} u_c \\ C_{W_{ab} W_{cd}} &= (\lambda_a u_b u_a^T + (1 - \delta_{ab})\lambda_b u_a u_b^T)_{pq} (\lambda_c u_d u_c^T + (1 - \delta_{cd})\lambda_d u_c u_d^T)_{rs} \\ &\quad \times \{ \tilde{T}_{(4)}^{pqrs} - (V^{pq} V^{rs} + V^{pr} V^{qs} + V^{ps} V^{qr}) \}. \end{aligned} \quad (13)$$

In the above equation, $\tilde{T}_{(3)}^{\cdot a}$ is the $D \times D$ matrix with fixed a for $\tilde{T}_{(3)}^{bca}$.

References

- [1] K. Fukumizu and S. Amari. Local minima and plateaus in hierarchical structures of multilayer perceptrons. *Neural Networks*, 13(3):317–327, 2000.
- [2] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *Advances in Neural Information Processing Systems*, 14, 2002. MIT Press.
- [3] S. Amari, H. Park, and T. Ozeki. Geometrical singularities in the neuromanifold of multilayer perceptrons. *Advances in Neural Information Processing Systems*, 14, 2002. MIT Press.
- [4] K. Fukumizu. Likelihood Ratio of Unidentifiable Models and Multilayer Neural Networks. *The Annals of Statistics*. To be published. <http://www.stat.washington.edu/annstat/RFI/FuturePapers.html>
- [5] Z. Ghahramani and G. Hinton. The EM algorithm for mixtures of factor analyzers. Technical Report CRG-TR-96-1, University of Toronto, Department of Computer Science, 1997.
- [6] M. Tipping and C. Bishop. Mixtures of probabilistic principal component analysers. *Neural Computation*, 11:443–482, 1999.
- [7] N. Ueda, R. Nakano, Z. Ghahramani, and G. Hinton. SMEM algorithm for mixture models. *Neural Computation*, 12(9):2109–2128, 2000.
- [8] M. Sato and S. Ishii. On-line EM algorithm for the normalized Gaussian network. *Neural Computation*, 12(2):2209–2225, 2000.