Density Estimation by Mixture Models with Smoothing Priors

Akio Utsugi
National Institute of Bioscience and Human-Technology,
1-1 Higashi Tsukuba Ibaraki 305, Japan
April 19, 1998

Abstract
In the statistical approach for self-organizing maps (SOMs), learning is regarded as an estimation algorithm for a Gaussian mixture model with a Gaussian smoothing prior on the centroid parameters. The values of the hyperparameters and the topological structure are selected on the basis of a statistical principle. However, since the component selection probabilities are fixed to a common value, the centroids concentrate on areas with high data density. This deforms a coordinate system on an extracted manifold and makes smoothness evaluation for the manifold inaccurate. In this paper, we study an extended SOM model whose component selection probabilities are variable. To stabilize the estimation, a smoothing prior on the component selection probabilities is introduced. An estimation algorithm for the parameters and the hyperparameters based on empirical Bayesian inference is obtained. The performance of density estimation by the new model and the SOM model is compared via simulation experiments.

1 Introduction

Mixture of probability distributions is an elemental method for constructing a complex stochastic model from simple stochastic components (McLachlan and Basford, 1988). Although the density of a mixture model is only a weighted sum of its component densities, estimation for the model is difficult in comparison with that for the associated single-component model. However, once a standard estimation method by an expectation-maximization (EM) algorithm was established, mixture models have been applied to many problems. In particular, Gaussian mixture
models are used for discovering clusters from data. Moreover, their extension has been attempted in various manners. For example, prior probabilities on the parameters are introduced to stabilize the estimation (Hanson et al., 1991; Ormoneit and Tresp, 1996).

Furthermore, we can consider a topological structure for the assembly of the components, independently of the data space. That is, the components are assumed to lie on a hidden space, which is called the inner space of the model. When the dimension of the inner space is smaller than the data space, this model represents stochastic data generation from a manifold embedded in the data space. In addition, when the manifold is smooth, we can use the smoothness for the regularization of the parameters. In this case, the manifold is modeled as a sample of a smooth stochastic process, which is represented by a smoothing prior probability distribution on the centroid parameters of the components. When this model is used for the extraction of the hidden manifold from data, rather than discovering clusters, many components are required for the detailed representation of curvature structure on the manifold. The regularization of the centroid parameters enables stable estimation on the many components.

In particular, the simplest kind of Gaussian mixture model with a Gaussian smoothing prior on the centroid parameters can be regarded as a stochastic model for self-organizing maps (SOMs). In other words, a learning algorithm for SOMs is derived as an estimation algorithm for this stochastic model (Utsugi, 1996, 1997). Although SOMs are well-known as unique neural networks to extract hidden topological structure from data, they had no suitable statistical model for a long time (Kohonen, 1995). The statistical treatment of SOMs has many advantages. For example, the values of the hyperparameters and the topology can be selected on the basis of a statistical criterion.

However, experiments in the earlier studies were confined to the SOM model with one-dimensional inner spaces. In the current paper, we attempt an experiment for the model with a multi-dimensional inner space. Then, problems proper to the multi-dimensional case are observed. For example, the centroids of the components tend to concentrate on areas with high data density, since the component selection probabilities are fixed to a common value. This property causes the deformation of a coordinate system on an extracted manifold. Especially in the multi-dimensional case, this deformation is large. When the smoothness of the manifold is evaluated for the hyperparameter selection, the deformation of the coordinate system makes this evaluation inaccurate.

This difficulty is partially due to the fixation of the component selection probabilities. If we use the component selection probabilities as variable parameters, the components can spread more uniformly on the extracted manifold. This suppresses the deformation of the coordinate system. However, if the component selection
probabilities are completely free, the model has too many free parameters to obtain stable estimates. Thus, we need the regularization of the component selection probabilities.

In the remainder of this paper, we study a mixture model with a smoothing prior on the component selection probabilities in addition to that on the centroids. First, a maximum a posteriori (MAP) estimation algorithm for the parameters is obtained using a generalized EM algorithm. Next, we obtain a hyperparameter estimation algorithm by an empirical Bayesian method. Finally, the performance of density estimation based on the new model and the SOM model is compared via simulation experiments.

2 SOM with Multi-Dimensional Inner Space

2.1 Stochastic Model for SOM

The probability density function of a mixture model for a datapoint $x$ is defined as

$$f(x|w, v, \beta) = \sum_{s=1}^{r} v_s g(x|w_s, \beta)$$

using the component selection probabilities $v_s$ and the component densities $g(x|w_s, \beta)$ with the component parameters $w_s$ and the common parameter $\beta$, where $v = (v_1, \ldots, v_r)'$ and $w = (w_1', \ldots, w_r')'$. In this paper, we focus on a special kind of mixture model for a data point $x = (x_1, \ldots, x_m)' \in \mathbb{R}^m$, whose components have spherical Gaussian densities with a common variance

$$g(x|w_s, \beta) = \left(\frac{\beta}{2\pi}\right)^{m/2} \exp\left(-\frac{\beta}{2}\|x - w_s\|^2\right)$$

where $w_s = (w_{s1}, \ldots, w_{sm})' \in \mathbb{R}^m$ are the centroids of the components. The likelihood function for a data set $X = (x_1, \ldots, x_n)'$ is given as

$$f(X|w, v, \beta) = \prod_{i=1}^{n} f(x_i|w, v, \beta)$$

$$= \left(\frac{\beta}{2\pi}\right)^{mn/2} \prod_{i=1}^{n} \sum_{s=1}^{r} v_s \exp\left(-\frac{\beta}{2}\|x - w_s\|^2\right).$$

In addition to this likelihood, a smoothing prior on the centroids along an inner space is introduced. Using a discretized differential operator $D$ on the inner space, which is considered in the next section, we define the smoothness of a
centroid configuration along the inner space by $-\sum_{j=1}^{r} \|Dw_{(j)}\|^2$, where $w_{(j)} = (w_{1j}, \ldots, w_{rj})'$. Then, we assume a Gaussian smoothing prior

$$f(w|\alpha, D) = \prod_{j=1}^{r} \left( \frac{\alpha}{2\pi} \right)^{l/2} (\det + M)^{1/2} \exp \left( -\frac{\alpha}{2} \|Dw_{(j)}\|^2 \right)$$  \hspace{1cm} (2.4)$$

where $M = D'D$, $l = \text{rank} M$ and $\det + M$ denotes the product of positive eigenvalues of $M$. This prior gives larger probability density to smoother centroid configurations. Since $M$ is generally singular, this prior density function is partially improper. That is, it specifies density only on a subspace of the parameter space. This subspace is spanned by the eigenvectors of $M$ with positive eigenvalues.

From the likelihood and the prior, a posterior on the centroids is calculated by Bayes’ rule

$$f(w|X, v, \alpha, \beta, D) \propto f(X, w|v, \alpha, \beta, D)$$

$$= f(X|v, \beta) f(w|\alpha, D).$$  \hspace{1cm} (2.5)$$

The MAP estimates of the centroids are given as the maximizer of this posterior. In particular, if $v_s$ are fixed to $1/r$ for every $s$, a learning algorithm of SOMs is derived from the posterior as an approximate MAP estimation algorithm (Utsugi, 1997).

Next, we consider the estimation of $\alpha$ and $\beta$. Since $\beta$ is the inverse of the squared variance of the Gaussian components, it can be estimated through an EM algorithm together with the centroids. On the other hand, $\alpha$ is a parameter in the prior, that is, a hyperparameter. Although there are several methods for the determination of the hyperparameter, we adopt the empirical Bayesian method because of its stability and usability (Utsugi, 1996, 1997). In the empirical Bayesian method, $\beta$ is treated as a hyperparameter and the estimates of the hyperparameters are given by the maximizer of their evidence

$$f(X|\alpha, \beta, D) = \int f(X, w|\alpha, \beta, D) dw.$$  \hspace{1cm} (2.6)$$

A Gaussian approximation for the evidence was presented (Utsugi, 1997). In addition, an efficient search algorithm for the maximizer of the evidence using the derivatives was presented (Utsugi, 1996).

### 2.2 Discretized Differential Operator on Multi-Dimensional Inner Space

On a one-dimensional inner space with a simple line-segment topology, the discretized differential operator $D$ is given by a discretized Laplacian $D_r^{(2)}$, whose
entries are
\[
d_{ij} = \begin{cases} 
  -2 & |i - j + 1| = 0 \\
  1 & |i - j + 1| = 1 \quad i = 1, \ldots, r-2; \quad j = 1, \ldots, r. \\
  0 & \text{otherwise}
\end{cases}
\] (2.7)

Using this operator, we can regard each term \( \|Dw_j\|_2^2 \) in the Gaussian smoothing prior as a discretized spline regularizer. A spline regularizer for a real function \( w_j(x) \) on the range \([0, 1]\) is given by
\[
S(w_j) = \int_0^1 \left( \frac{\partial^2 w_j}{\partial x^2} \right)^2 \, dx.
\] (2.8)

We regard \( w_{(j)} \) as a discretized vector of \( w_j(x) \), that is, \( w_{sj} = w_j((s-1)\delta), s = 1, \ldots, r \), where \( \delta = 1/(r-1) \). The discretized spline regularizer is now given by
\[
S(w_{(j)}) = \sum_{s=2}^{r-1} \left( \frac{w_{s-1,j} + w_{s+1,j} - 2w_{sj}}{\delta} \right)^2 = \frac{1}{\delta^2} \|Dw_{(j)}\|^2.
\] (2.9)

We ignore the constant \( 1/\delta^2 \), since it can be absorbed into \( \alpha \).

A natural extension of the spline regularizer for functions on multi-dimensional spaces is the thin-plate spline regularizer (Wahba, 1990). In the two-dimensional case, this is given as
\[
S(w_j) = \int \int_\Omega \left\{ \left( \frac{\partial^2 w_j}{\partial x^2} \right)^2 + \left( \frac{\partial^2 w_j}{\partial y^2} \right)^2 + 2 \left( \frac{\partial^2 w_j}{\partial x \partial y} \right)^2 \right\} \, dx \, dy
\] (2.10)
for a function \( w_j(x,y) \) on a region \( \Omega \subset \mathbb{R}^2 \). Here, we consider that \( w_{(j)} \) is a discretized vector of the function \( w_j(x,y) \) at the vertices of a regular-interval lattice on \( \Omega \). When the lattice is rectangular and its vertices are indexed in order by \((k,t), k = 1, \ldots, p, t = 1, \ldots, q\), the value of the function at a vertex \((k,t)\) gives the \( st \)th entry of \( w_{(j)} \) at \( s = q(k-1) + t \). The discretization of the thin-plate spline regularizer is now given as
\[
S(w_{(j)}) = \|D_{xx}w_{(j)}\|^2 + \|D_{yy}w_{(j)}\|^2 + 2\|D_{xy}w_{(j)}\|^2
\] (2.11)
using
\[
D_{xx} = D_p^{(2)} \otimes I_q
\] (2.12)
\[
D_{yy} = I_p \otimes D_q^{(2)}
\] (2.13)
\[
D_{xy} = D_p^{(1)} \otimes D_q^{(1)}
\] (2.14)
where \( I_p \) and \( I_q \) denote identity matrices with sizes \( p \) and \( q \), and \( \otimes \) denotes the Kronecker product. The matrices \( D_p^{(2)} \) and \( D_q^{(2)} \) are the one-dimensional discretized Laplacians with sizes \( p \) and \( q \). The matrices \( D_p^{(1)} \) and \( D_q^{(1)} \) are the one-dimensional first-order differential operators. For example, \( D_p^{(1)} \) is defined as a matrix whose entries are

\[
    d_{ij} = \begin{cases} 
        1 & i = j \\
        -1 & i + 1 = j \quad i = 1, \ldots, p - 1; j = 1, \ldots, p \\\n        0 & \text{otherwise}
    \end{cases}
\]  

(2.15)

The discretized thin-plate spline regularizer is also expressed as \( S(w_{(j)}) = \|Dw_{(j)}\|_2 \) using

\[
    D = [D'_{xx}, D'_{yy}, D'_{xy}, D'_{xy}]'.
\]

(2.16)

The SOM model with a rectangular inner space uses this \( D \) in its smoothing prior. We can also make other boundary shapes of the inner space than rectangles by manipulating \( D \). In the next, we perform a simulation experiment for this model.

### 2.3 Simulation Experiment

In the simulation experiment, we use artificial data generated from two types of quadric surfaces in \( \mathbb{R}^3 \): an elliptic paraboloid and a hyperbolic paraboloid, which are represented by equations \( x_3 = ax_1^2 + bx_2^2 \) and \( x_3 = ax_1^2 - bx_2^2 \) respectively. Initially, we obtain points on a plane spanned by the \( x_1 \) and \( x_2 \) axes using a two-dimensional Gaussian random generator with zero mean and a covariance matrix \( \text{diag}(c^2, d^2) \), where \( \text{diag}(c^2, d^2) \) is a diagonal matrix with the diagonal \( (c^2, d^2) \). Then, we obtain points on the quadric surfaces using the above equations. Finally, by adding three-dimensional random vectors to the points on the surfaces, we obtain the data points. The random vectors are generated from a Gaussian random generator with zero mean and covariance \( \sigma^2 I_3 \). We use \( a = 0.25, b = 1, c = 0.7, d = 0.35, \sigma = 0.1 \) and \( n = 300 \).

To the data sets, we apply the SOM model whose inner space is specified by a rectangular lattice with \( p = 10 \) and \( q = 10 \). The centroids and the hyperparameters are estimated by the fast search algorithm presented earlier (Utsugi, 1996). This algorithm is the mixture of the centroid estimation algorithm based on posterior maximization and the hyperparameter search algorithm based on evidence maximization. The initial values of the centroids are made from a Gaussian random generator with zero mean and covariance \( I_3 \). The initial values of the hyperparameters are set to \( \alpha = 1000 \) and \( \beta = 25 \). We attempt 50 learning sessions using different initial centroids.

Figure 1 shows the configurations of the centroids giving maximum evidence out of the 50 learning sessions. We use the Gaussian approximation for the calculation...
Figure 1: Centroid configurations of the SOM model giving maximum evidence. The meshes and their nodes represent the extracted surfaces and the centroids respectively. The dots represent the data points generated from two sorts of quadric surfaces. The graphs viewed from two angles are displayed.
of the evidence (2.6). The nodes of the meshes in the figure represent the centroids. The meshes are made by linking the nearest components in the inner space. From this figure, we observe that the centroids concentrate on a center area with high data density and this deforms the coordinate system on the surfaces. In particular, the data from the elliptic paraboloid yield large deformation. Although centroid configurations giving maximum evidence have different patterns for other data sets, they have also large deformation.

If a hidden manifold has small curvature and data density on the manifold is uniform, \( \alpha \) becomes large. In this case, the strong regularization produces a regular centroid configuration. In the simulations, \( \alpha \) decreases because the data density on the manifold is not uniform. However, if the component selection probabilities can be adapted to the data density on the manifold, \( \alpha \) may maintain large value, and thus a regular centroid configuration may be obtained. In the following section, such a model is constructed.

3 Mixture Model with Smoothing Prior on Component Selection Probabilities

3.1 Smoothing Prior on Component Selection Probabilities

In this section, we study a mixture model with variable component selection probabilities. The simplest method for estimating the component selection probabilities is the maximum likelihood (ML) method. In this case, these parameters are regarded as completely free and estimated using an EM algorithm. However, this method produces unstable estimates, because we use many components for the detailed representation of curvature structure on the hidden manifold, unlike the original mixture model. In particular, multi-dimensional manifolds require many components for the interpolation of sparse data points. Some components may have too few data points to estimate their parameters properly.

To stabilize the estimation, we need a regularization prior on the component selection probabilities \( \mathbf{v} \). A candidate for such a prior is a Dirichlet prior, which is conjugate for the multinomial distribution on \( \mathbf{v} \) (Ormoneit and Tresp, 1996). However, the Dirichlet prior has a bias towards even component selection, and thus, it has difficulty in capturing large and smooth variation of the component selection probabilities. Sufficient regularization by this prior often produces the same result as the SOM model. For the stable estimation of the largely variable component selection probabilities, we use a smoothing prior on \( \mathbf{v} \) along the inner space.
Since $\mathbf{v}$ is a probability vector, that is, its entries are restricted to the range $[0,1]$ and sum into one, the direct use of a Gaussian prior on $\mathbf{v}$ is not natural. Instead, we consider a Gaussian prior on another variables $\mathbf{\mu} = (\mu_1, \ldots, \mu_r)' \in \mathbb{R}^r$ that produce $\mathbf{v}$ via the soft-max function

$$v_s = \frac{\exp \mu_s}{\sum_{t=1}^r \exp \mu_t}. \tag{3.1}$$

A similar parameter conversion is used in the mixture-of-experts model (Jordan and Jacobs, 1994; Jordan and Xu, 1995). We can now make a Gaussian smoothing prior on $\mathbf{\mu}$ as

$$f(\mathbf{\mu}|\gamma, \mathbf{D}) = \left( \frac{\gamma}{2\pi} \right)^{1/2} (\det^+ \mathbf{M})^{1/2} \exp \left( -\frac{\gamma}{2} \| \mathbf{D} \mathbf{\mu} \|^2 \right). \tag{3.2}$$

However, when the second-order differential operator is used as $\mathbf{D}$, the estimation of $\mathbf{\mu}$ is very unstable. A reason for this instability is explained as follows. For the second-order differential operator, $\mathbf{M} = \mathbf{D}' \mathbf{D}$ has two sorts of eigenvectors with zero eigenvalues, which are a constant vector and linear-variation vectors. This means that the prior does not constrain the bias and the linear trend of $\mathbf{\mu}$. For the centroid parameters, their bias and linear trend can be determined properly from data without constraint. On the other hand, the linear trend of $\mathbf{\mu}$ is difficult to determine, because it interferes with the linear trend of component intervals. We need not determine the bias of $\mathbf{\mu}$.

Instead of this problematic prior, we use another smoothing prior with constraint on the linear trend. The new prior is based on both the first-order and second-order differential operators. The second-order differential operator $\mathbf{D}_2$ has been already given by (2.16). The first-order differential operator $\mathbf{D}_1$ is defined such that $\| \mathbf{D}_1 \mathbf{\mu} \|^2$ is the discretization of the first-order regularizer for a function $\mu(x,y)$

$$S(\mu) = \int \int_\Omega \left\{ \left( \frac{\partial \mu}{\partial x} \right)^2 + \left( \frac{\partial \mu}{\partial y} \right)^2 \right\}. \tag{3.3}$$

Such an operator is given by

$$\mathbf{D}_1 = [\mathbf{D}_x', \mathbf{D}_y']' \tag{3.4}$$

where

$$\mathbf{D}_x = \mathbf{D}^{(1)}_p \otimes \mathbf{I}_q \tag{3.5}$$
$$\mathbf{D}_y = \mathbf{I}_p \otimes \mathbf{D}^{(1)}_q. \tag{3.6}$$

The matrices $\mathbf{D}^{(1)}_p$ and $\mathbf{D}^{(1)}_q$ are the one-dimensional first-order differential operators, defined in section 2.2.
The new smoothing prior on $\mu$ is now given as
\[
f(\mu|\gamma_1, \gamma_2, D_1, D_2) = \frac{1}{(2\pi)^{l/2}}(\det + H_{PM})^{1/2} \exp \left[-\frac{1}{2}(\gamma_1 \| D_1 \mu \|^2 + \gamma_2 \| D_2 \mu \|^2) \right]
\]
where
\[
H_{PM} = \gamma_1 M_1 + \gamma_2 M_2
\]
(3.7)

$M_1 = D_1' D_1$, $M_2 = D_2' D_2$ and $l = \text{rank } H_{PM}$. The first and second terms in the exponential function are the first-order and second-order regularizers, respectively. The new hyperparameters $\gamma_1$ and $\gamma_2$ mean the strength of the regularization.

In addition, we use a similar prior on the centroids
\[
f(w|\alpha_1, \alpha_2, D_1, D_2) = \prod_{j=1}^{m} \frac{1}{(2\pi)^{l/2}}(\det + H_{PW})^{1/2} \exp \left[-\frac{1}{2}(\alpha_1 \| D_1 w(j) \|^2 + \alpha_2 \| D_2 w(j) \|^2) \right]
\]
where
\[
H_{PW} = \alpha_1 M_1 + \alpha_2 M_2.
\]
(3.9)

If $\alpha_1 = 0$, this prior is identical to (2.4). Here, we introduce the first-order regularization to observe the effect of it.

Here, we refer to the parameters of the model as $\theta = (w', \mu')'$, the hyperparameters as $h = (\alpha_1, \alpha_2, \beta, \gamma_1, \gamma_2)$ and the topological structure as $T = \{D_1, D_2\}$, respectively. From the mixture likelihood (2.3) and the two smoothing priors (3.7) and (3.9), a posterior on $\theta$ is calculated by
\[
f(\theta|X, h, T) \propto f(X, \theta|h, T).
\]
\[
= f(X|\theta, \beta)f(w|\alpha_1, \alpha_2, T)f(\mu|\gamma_1, \gamma_2, T).
\]
(3.11)

In the next section, we construct an MAP estimation algorithm for this posterior by a generalized EM algorithm.

### 3.2 MAP Estimation Algorithm for Parameters

To make the EM algorithm, we need a likelihood function for the complete data set $\{X, Y\}$, where $Y$ is missing data consisting of the binary memberships of the data points for the components. This likelihood is given as
\[
f(X, Y|\theta, \beta) = \prod_{i=1}^{n} \prod_{s=1}^{r} (v_s g(x_i|w_s, \beta))^{y_{si}}.
\]
(3.12)
Using this, a conditional expectation of the log posterior is given as

\[ Q(\theta|\theta^*) = E_Y \{ \log f(X, Y|\theta, \beta)|X, \theta^* \} + \log f(w|\alpha_1, \alpha_2, T) + \log f(\mu|\gamma_1, \gamma_2, T) \]

\[ = \sum_{i=1}^{n} \sum_{s=1}^{r} p_{si} \log v_s + \frac{nm}{2} \log \frac{\beta}{2\pi} - \frac{\beta}{2} \sum_{i=1}^{n} \sum_{s=1}^{r} p_{si} \|x_i - w_s\|^2 \]

\[ - \frac{ml}{2} \log 2\pi + \frac{m}{2} \log \det + H_{PW} - \frac{1}{2} \sum_{j=1}^{m} \left( w_j^t H_{PW} w_j \right) \]

\[ - \frac{l}{2} \log 2\pi + \frac{1}{2} \log \det + H_{PM} - \frac{1}{2} \mu' H_{PM} \mu \]  

(3.13)

where \( \theta^* \) is a temporary estimate and

\[ p_{si} = p(y_{si} = 1|x_i, \theta^*, \beta) = \frac{v_s^* g(x_i|w^*_s, \beta)}{\sum_t v_t^* g(x_i|w^*_t, \beta)} \]  

(3.14)

are posterior component selection probabilities. In the EM algorithm, the maximizer of this function is used as the next temporary estimate and this updating is iterated until convergence.

The centroids maximizing \( Q \) are easily obtained by solving linear equations

\[ \frac{\partial Q}{\partial w_{(j)}} = \beta \sum_{i=1}^{n} x_{ij} p_i - \beta N w_{(j)} - H_{PW} w_{(j)} = 0 \]  

(3.15)

where \( p_i = (p_{i1}, \ldots, p_{ri})' \) and \( N = \sum_i^{n} \text{diag} \ p_i \). Thus, the next temporary estimates of the centroids are given as

\[ \hat{w}_{(j)} = \beta H_{QW}^{-1} \sum_{i=1}^{n} x_{ij} p_i \]  

(3.16)

where

\[ H_{QW} = -\frac{\partial^2 Q}{\partial w_{(j)} \partial w'_{(j)}} = \beta N + \alpha_1 M_1 + \alpha_2 M_2. \]  

(3.17)

This updating rule has the same form as the EM algorithm for the SOM model.

On the other hand, to obtain \( \mu \) maximizing \( Q \), we need to solve a nonlinear equation

\[ \frac{\partial Q}{\partial \mu} = \sum_{i=1}^{n} (p_i - v) - H_{PM} \mu = 0 \]  

(3.18)

which requires an iterative method, such as a Newton-Raphson algorithm. Instead, we consider to use one step of the Newton-Raphson algorithm as an updating rule for \( \mu \), which gives a generalized EM algorithm (Jordan and Xu, 1995). However,
this manner fails for our model, since the Hesse matrix of \(-Q\) is singular. The Hesse matrix of \(-Q\) is given as

\[
H_{QM} = -\frac{\partial^2 Q}{\partial \mathbf{\mu} \partial \mathbf{\mu}'} = nV + \gamma_1 M_1 + \gamma_2 M_2
\]  

(3.19)

where \(V = \text{diag} v - vv'\). We can show that \(H_{QM}\) is non-negative definite and has \(1_r\), an \(r\)-dimensional vector with ones, as an eigenvector with zero eigenvalue. This singularity is due to the non-identifiability of the model along the direction of \(1_r\). That is, the posterior is invariable by adding any constant vector to \(\mu\). To eliminate this non-identifiability, we fix the projection of \(\mu\) onto \(1_r\). When this projection is zero, \(\mu\) is restricted to a linear manifold given by \(\sum_s \mu_s = 0\). We now maximize a Lagrange function

\[
L(\mu, \lambda) = Q(\mu) + \lambda \sum_{s=1}^r \mu_s
\]  

(3.20)

rather than \(Q\), where \(\lambda\) is a Lagrange multiplier. An increment vector \(\Delta \mu\) by a Newton-Raphson algorithm for this Lagrange function is obtained by solving an equation

\[
\begin{bmatrix}
H_{QM} & 1_r \\
-1_r' & 0
\end{bmatrix}
\begin{bmatrix}
\Delta \mu \\
\lambda
\end{bmatrix}
= -
\begin{bmatrix}
d_{QM} \\
0
\end{bmatrix}
\]  

(3.21)

where \(d_{QM} = -\partial Q/\partial \mu\) is given in (3.18).

In the result, the updating rule for \(\theta\) is given by (3.16) and

\[
\hat{\mu} = \mu + \Delta \mu.
\]  

(3.22)

### 3.3 Hyperparameter Selection

Evidence for \(h = (\alpha_1, \alpha_2, \beta, \gamma_1, \gamma_2)\) is defined as

\[
f(X|h, T) = \int_{\Theta} f(X, \theta|h, T) d\theta
\]  

(3.23)

where \(\Theta = \{\theta : w \in R^m, \sum_{s=1}^r \mu_s = 0, \mu \in R^r\}\). Here, we obtain a calculable formula for the evidence using a Gaussian approximation (MacKay, 1992). For the Gaussian approximation, we need the Hessian of the negative integrand at the maximizer of the integrand. In the present case, this maximizer is given by the MAP estimate \(\hat{\theta}\). The Hessian must be evaluated on the restricted parameter space \(\Theta\), rather than on the unrestricted parameter space \(R^{m+r}\). We can obtain the Hessian on \(\Theta\) from the Hesse matrix on \(R^{m+r}\):

\[
H_{\theta \theta} = -\frac{\partial^2}{\partial \theta \partial \theta'} \log f(X, \hat{\theta}|h, T).
\]  

(3.24)
The Hessian on $\Theta$ is given by $\text{det}^+ H_{\theta\theta}$. This is because $H_{\theta\theta}$ is non-negative definite and its only eigenvector with zero eigenvalue is orthogonal to $\Theta$. The local evidence at $\hat{\theta}$ is defined as

$$f(X, S_{\hat{\theta}}|h, T) = \int_{S_{\hat{\theta}}} f(X, \theta|h, T)d\theta \approx (2\pi)^{(rm+r-1)/2} f(X, \hat{\theta}|h, T)(\text{det}^+ H_{\theta\theta})^{-1/2}$$

(3.25)

where $S_{\hat{\theta}} \subset \Theta$ is a region dominated by $\hat{\theta}$. Although the Hesse matrix is obtained exactly in appendix A, it is complicated. Rather, we approximate it as

$$\log \text{det}^+ H_{\theta\theta} \simeq m \log \text{det} H_{QW} + \log \text{det}^+ H_{QM}. \quad (3.26)$$

The log evidence is now given as

$$e(h) = \frac{nm}{2} \log \frac{\beta}{2\pi} + \sum_{i=1}^{n} \log \sum_{s=1}^{r} \hat{v}_s \exp \left( -\frac{\beta}{2} \| x_i - \hat{w}_s \|^2 \right)$$

$$-\frac{ml}{2} \log 2\pi + \frac{m}{2} \log \text{det}^+ H_{PW} - \frac{1}{2} \sum_{j=1}^{m} \hat{w}'(j) H_{PW} \hat{w}(j)$$

$$-\frac{l}{2} \log 2\pi + \frac{1}{2} \log \text{det}^+ H_{PM} - \frac{1}{2} \hat{\mu}' H_{PM} \hat{\mu}$$

$$-\frac{1}{2}(m \log \text{det} H_{QW} + \log \text{det}^+ H_{QM}) + \frac{rm + r - 1}{2} \log 2\pi. \quad (3.27)$$

An approximate Newton-Raphson algorithm for the maximizer of this evidence is proposed in appendix B.

### 3.4 Simulation Experiment

We apply the new model to the data used in section 2.3. The results of the previous simulations are employed as the initial values of the centroids and the hyperparameters $\alpha_1$ and $\beta$ in the current simulations. The initial values of the other hyperparameters are set to $\alpha_2 = 100$, $\gamma_1 = 100$ and $\gamma_2 = 1000$. The initial values of the component selection probabilities are set to a constant. Figure 2 shows the configurations of the centroids with maximum evidence. By comparing them with the configurations in Figure 1, we find that the deformation of the coordinate system on the surfaces relaxes. Figure 3 shows the distributions of the estimated component selection probabilities. These distributions are similar to a Gaussian distribution used in the data generation.
Figure 2: Centroid configurations of the new model giving maximum evidence.
To compare the performance of density estimation between the new model and the SOM model, we calculate an expected log likelihood (ELL) for each model. ELL is a measure of compatibility between a data distribution and a model distribution. Here, we approximate ELL by log likelihood for 1000 newly generated data points. On both data sets, the new model shows larger ELL.

Next, we apply the model to data under severer conditions.

First, the noise level is increased from 0.1 to 0.2. Under this condition, \( \gamma_1 \) and \( \gamma_2 \) grow infinitely on all learning sessions. In this case, the component selection probabilities approach constant. We set the upper limit of \( \gamma_1 \) and \( \gamma_2 \) to \( 10^5 \). The centroid configurations with maximum evidence are shown in Figure 4. In this case, the new model yields the same solution as the SOM model.

Second, the data size is reduced from 300 to 100. Under this condition, the component selection probabilities become constant again.

Finally, we increase the curvature of the hidden surfaces by doubling the values of \( a \) and \( b \). The centroids and the component selection probabilities with maximum evidence are shown in the Figure 5 and 6, respectively. For the hyperbolic-paraboloid data, the solution has a similar pattern to the solution under the low curvature condition. On the other hand, for the elliptic-paraboloid data, the centroid configuration is largely deformed and the component selection probabilities become monotonic by infinite growth of \( \gamma_2 \).
Figure 4: Centroid configurations giving maximum evidence in high-noise condition.
Figure 5: Centroid configurations giving maximum evidence in high-curvature condition.
4 Discussion

4.1 Condition for Convergence into Correct Structure

In the proposed method, topology extraction is formalized as an inverse problem for a stochastic generative model. Thus, if the generative model is valid, the correct structure is given by the optimal estimate for the model. Since we use the generalized EM algorithm for the parameter estimation, convergence into a local optimal estimate is guaranteed if the hyperparameters are fixed (Dempster et al., 1977).

On the other hand, the hyperparameter estimation fails if the data size is too small, as shown in the simulations. This is because empirical Bayesian inference has a bias toward simple structure. Without sufficient data for structure determination, it selects the simplest structure (Utsugi, 1997). In the current case, the simplest structure has a regular centroid configuration on a flat plane and constant component selection probabilities. In particular, the component selection probabilities easily become constant or monotonic if the flexibility of the manifold is large, because data density on the manifold can be also represented by component density. Furthermore, it is essentially difficult to fit a flat coordinate system to a high-curvature manifold globally. In conclusion, we need many data and small curvature of the manifold for exact hyperparameter estimation.
4.2 Relation to Other Related Methods

4.2.1 Principal curves

Here, we consider the relationship of our model to principal curves and principal manifolds (Hastie and Stuetzle, 1989; Tibshirani, 1992). A principal curve is a generalization of a linear principal component. This is originally defined for a probability distribution. However, several algorithms to estimate it from data were also presented. In particular, Tibshirani’s algorithm is derived from a generative model, which is based on a Gaussian mixture model with a constraint on its centroids. Although this generative principal-curve model is similar to our model, they are different in the manner of parameterization.

In the principal-curve model, each data point is assumed to be generated from an individual component on a smooth manifold. That is, the components have a one-to-one correspondence with data points. We need to estimate the positions of the components on the manifold in addition to the estimation of the manifold itself. Since these position parameters are not regularized and their number increases in proportion to the data size, the estimates maintain large variance even if the data size increases. This property degrades the performance of the density estimation.

On the other hand, our model estimates the distribution of the data sources on the manifold, instead of estimating the positions of the data sources on the manifold. The distribution are represented by the component selection probabilities, whose number is fixed independently of the data size. In addition, these parameters are regularized by the smoothing prior. Thus, the effective number of the parameters is considerably reduced in comparison with the principal-curve model. We can say that our model is a regularized version of the principal-curve model. Thus, our model is expected to make better estimation than the principal-curve model, if the assumption of the regularization is valid. Moreover, our model is more suitable for Bayesian inference, including the determination of the hyperparameters, because it is governed by a few free hyperparameters.

4.2.2 Autoassociative bottleneck networks

It is known that three-layered autoassociative neural networks with backpropagation learning are related to the principal component analysis (PCA), if their hidden units are less than their input-output units. Furthermore, five-layered autoassociative networks can perform nonlinear PCA, if they have more units in the second and fourth layers and less units in the third layer than the input-output layers (Diamantaras and Kung, 1996). This nonlinear PCA net produces fully distributed representation in the third layer, while the SOM model produce a localized activity blob in the inner unit layer. The nonlinear PCA net represents
smooth mapping between the inner space and the data space by a pair of three-layered neural networks. Although the three-layered neural networks can emulate any smooth function, they require many inner units for the emulation of nonlinear mapping. Thus, their estimation becomes unstable for small data. Although we can introduce the regularization of the parameters, this smoothness representation is not natural. On the other hand, our model represents the smoothness directly using a spline regularizer. Thus, under empirical Bayesian inference, our model is expected to acquire a finer estimate than the nonlinear PCA net.

4.2.3 Generative topographic mapping

Bishop et al. (1997) recently proposed a new method to extract topological structure from data, which is called generative topographic mapping (GTM). This method is also based on a Gaussian mixture model with constrained centroids and an EM algorithm. However, GTM represents a smooth mapping between the inner space and the data space using a generalized linear network. Thus, the performance of this method depends on the choice of a basis function system in the generalized linear network.

It is known that the spline regularization method is also expressed by a generalized linear network (Poggio and Girosi, 1990). Thus, GTM and our model may be two different model representations of an identical principle. However, our model representation is more suitable for empirical Bayesian inference, because the smoothness is expressed directly by a prior probability on the parameters.

5 Conclusion

A stochastic model for SOM with a multi-dimensional inner space was constructed as a mixture model with thin-plate spline regularization on the centroids. Then, estimation on the model was attempted using artificial data. In the experiment, we observed that the model has a difficulty caused by constant component selection probabilities. To overcome this difficulty, we constructed a mixture model with variable component selection probabilities. For the stabilization of the estimation, a smoothing prior on the component selection probabilities was introduced.

Although introducing variable component selection probabilities was effective for reducing the deformation of the coordinate systems on low-curvature manifolds, the deformation on high-curvature manifolds was not removed. This is a limitation in fitting a manifold with a single coordinate system into data. Instead, we can consider a model with multiple local coordinate systems. However, such a complicated model will require an efficient topology search algorithm.
References


Appendix A

We obtain the Hesse matrix of the negative log posterior with respect to the parameters. First, the derivatives of the log posterior with respect to the parameters are obtained. These have the same forms as the derivatives of \( Q \), given by (3.15) and (3.18), except that the posterior component selection probabilities \( p_{si} \) in the formulae are calculated using \( \theta \) rather than \( \theta^* \). Thus, in calculating the Hesse matrix by differentiating the derivatives, we must consider the differentiation of \( p_{si} \). We now obtain the Hesse matrix consisting of the following submatrices:

\[
\begin{align*}
-\frac{\partial^2 G}{\partial w_s \partial w_t'} &= -\beta^2 \sum_{i=1}^{n} (\delta_{st} p_{si} - p_{si} p_{ti}) (x_i - w_s) (x_i - w_t)' + (\beta n_s \delta_{st} + h_{st}) I_m \\
-\frac{\partial^2 G}{\partial \mu \partial \mu'} &= -\sum_{i=1}^{n} (\text{diag } p_i - p_i p_t' - V) + H_{PM} \\
-\frac{\partial^2 G}{\partial \mu_s \partial w_t'} &= -\beta \sum_{i=1}^{n} (\delta_{st} p_{si} - p_{si} p_{ti}) (x_i - w_t)
\end{align*}
\]

where \( G \) is the log posterior and \( h_{st} \) are the entries in \( H_{PW} \). If all \( p_{si} \) have binary values and satisfy a probability condition \( \sum_{s=1}^{r} p_{si} = 1 \), the above Hesse matrix becomes identical to the Hesse matrix of \(-Q\), given by (3.17) and (3.19). Since \( p_{si} \) are the estimates of the binary memberships \( y_{si} \), they are near to binary values if each component has sufficient identity. In this case, we can substitute the Hesse matrix of \(-Q\) for that of \(-G\) approximately.

Appendix B

Here, we obtain a hyperparameter search algorithm by an approximate Newton-Raphson algorithm. First, we obtain the gradients of the log evidence with respect to the hyperparameters:

\[
\frac{\partial e}{\partial \beta} = \frac{nm}{2\beta} - \frac{1}{2} \sum_{i=1}^{n} \sum_{s=1}^{r} p_{si} \| x_i - \hat{w}_s \|^2 - \frac{m}{2} \text{tr}(H_{QW}^{-1} N)
\]


\[
\frac{\partial e}{\partial \alpha_o} = \frac{m}{2} \text{tr}(H_{PW}^+ M_o) - \frac{1}{2} \sum_{j=1}^{m} \| D_o w_{(j)} \|^2 - \frac{m}{2} \text{tr}(H_{QW}^{-1} M_o) \quad \text{(A.5)}
\]

\[
\frac{\partial e}{\partial \gamma_o} = \frac{1}{2} \text{tr}(H_{PM}^+ M_o) - \frac{1}{2} \| D_o \hat{\mu} \|^2 - \frac{1}{2} \text{tr}(H_{QM}^+ M_o) \quad \text{(A.6)}
\]

where \( o = 1, 2 \) and \( H^+ \) denotes the Moore-Penrose generalized inverse of \( H \). In this calculation, the variation of \( \theta \) with the hyperparameters is neglected. Next, we obtain the Hesse matrix of the log evidence, whose diagonal is given by

\[
\frac{\partial^2 e}{\partial \beta^2} = -\frac{nm}{2\beta^2} + \frac{m}{2} \text{tr}(H_{QW}^{-1} N)^2 \quad \text{(A.7)}
\]

\[
\frac{\partial^2 e}{\partial \alpha_o^2} = -\frac{m}{2} \text{tr}\{(H_{PW}^+ M_o)^2 - (H_{QW}^{-1} M_o)^2\} \quad \text{(A.8)}
\]

\[
\frac{\partial^2 e}{\partial \gamma_o^2} = -\frac{1}{2} \text{tr}\{(H_{PM}^+ M_o)^2 - (H_{QM}^+ M_o)^2\}. \quad \text{(A.9)}
\]

For simplicity, we employ only the diagonal in the Hesse matrix. Although we can now make a Newton-Raphson algorithm for the hyperparameters, we should use that for the logarithms of the hyperparameters, because the hyperparameters are restricted to positive values. The gradients and the Hesse matrix with respect to the log hyperparameters can be calculated using formulae, such as

\[
\frac{\partial e}{\partial \log \beta} = \beta \frac{\partial e}{\partial \beta} \quad \text{(A.10)}
\]

\[
\frac{\partial^2 e}{\partial (\log \beta)^2} = \beta \frac{\partial e}{\partial \beta} + \beta^2 \frac{\partial^2 e}{\partial \beta^2}. \quad \text{(A.11)}
\]

We now obtain the both updating rules for the parameters and the hyperparameters. Strictly speaking, the hyperparameter-updating rule should be used after obtaining the MAP estimate of the parameters under fixed hyperparameters. We actually use these rules alternately and this accelerates the whole of the estimation procedure.