

Topology Selection for Self-Organizing Maps

A Utsugi

National Institute of Bioscience and Human-Technology,
1-1 Higashi Tsukuba Ibaraki 305, Japan

March 11, 1997

Abstract

A topology-selection method for self-organizing maps (SOMs) based on empirical Bayesian inference is presented. This method is natural extension of the hyperparameter-selection method presented earlier, in which the SOM algorithm is regarded as an estimation algorithm for a Gaussian mixture model with a Gaussian smoothing prior on the centroid parameters, and optimal hyperparameters are obtained by maximizing their evidence. In the present paper, comparisons between models with different topologies are made possible by further specifying the prior of the centroid parameters with an additional hyperparameter. In addition, a fast hyperparameter-search algorithm using the derivatives of evidence is presented. The validity of the methods presented is confirmed by simulation experiments.

1 Introduction

The virtue of self-organizing maps (SOMs) [1] is their ability to extract intrinsic topological structure hidden in multidimensional data despite the simplicity of their algorithm. In reality, SOMs are too flexible and fit to any data distribution, no matter which topology they postulate. Thus, using SOM, we are faced with the difficulty of determining the best topology from a number of possibilities.

The SOM algorithm is regarded as a vector quantization (VQ) algorithm with a topological constraint, which gives stability and robustness to the original VQ algorithm [2, 3]. By gradually eliminating the constraint, the SOM algorithm ultimately converges to one of the solutions of VQ. In fact, many applications have used SOMs in this manner. Although we can obtain a topological structure as the trace of the algorithmic process, such a structure is too weak to discriminate between different topologies.

It has recently been shown that SOM can also be regarded as an approximate estimation algorithm for a type of Gaussian mixture model. Luttrell [4] showed that a maximum likelihood (ML) estimation algorithm for a Gaussian mixture model is approximated by a SOM algorithm if the centroids of its components can be assumed to lie on a very smooth curve. Moreover, Utsugi [5] derived a SOM algorithm as an approximate maximum *a posteriori* (MAP) estimation algorithm for a Gaussian mixture model with a Gaussian smoothing prior on the centroid parameters along a specified topology. He then presented a statistical method to determine the optimal strength of the topological constraint, which is regarded as a hyperparameter of the stochastic model, rather than eliminate the constraint. This method is based on an empirical Bayesian approach with a Gaussian approximation of the evidence of the hyperparameter [6]. The magnitude of noise on the data is also regarded as a hyperparameter and estimated by the same method.

The aim of the present paper is to advance such an approach to topology selection. The above Bayesian framework is restricted to comparisons between models with a common topology; thus, we cannot use the evidence for selection from models with different topologies. This is because the Gaussian smoothing priors in the models are partially improper: that is, they specify only a subspace of the parameter space. In general, it is impossible to compare models with different improper dimensions, which are the dimensions of subspaces unspecified by the prior, and such a difference is produced by the variation of topologies. In this paper, this comparison is made possible by further specifying the prior of the centroid parameters with an additional hyperparameter.

Moreover, the comparisons between many topologies require a fast hyperparameter-search algorithm. However, we have difficulty in using the derivatives of the evidence for such an algorithm, owing to the complicated dependence of the evidence on the hyperparameters. Thus, we had no alternative but to use the direct search for the maximizer of the evidence. In this paper, by obtaining the derivatives through a further approximation of the evidence we construct a fast hyperparameter-search algorithm.

2 Bayesian framework for SOM

In this section, the Bayesian framework for SOMs is reviewed. The overall theory is divided into the theory of the stochastic model of SOMs and that of its estimation algorithms.

2.1 Stochastic model of SOMs

Initially, we consider a stochastic model underlying the SOM algorithm. Let $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n)'$ be a data set consisting of data points $\mathbf{x}_i = (x_{i1}, \dots, x_{im})' \in \mathbf{R}^m, i = 1, \dots, n$. We assume that each data point is generated by one of the *Gaussian generators* with densities

$$f(\mathbf{x}_i|\mathbf{w}_s, \beta) = \left(\frac{\beta}{2\pi}\right)^{m/2} \exp\left(-\frac{\beta}{2}\|\mathbf{x}_i - \mathbf{w}_s\|^2\right) \quad s = 1, \dots, r \quad (1)$$

where $\mathbf{w}_s = (w_{s1}, \dots, w_{sm})' \in \mathbf{R}^m, s = 1, \dots, r$, are individual centroids and $1/\beta$ is a common variance. By referring to the truth value of the event ‘the s th generator yields \mathbf{x}_i ’ as y_{si} , we obtain a density function for the data set:

$$f(\mathbf{X}|\mathbf{Y}, \mathbf{w}, \beta) = \prod_{i=1}^n \prod_{s=1}^r f(\mathbf{x}_i|\mathbf{w}_s, \beta)^{y_{si}} \quad (2)$$

where $\mathbf{Y} = (y_{si})$ and $\mathbf{w} = (\mathbf{w}'_1, \dots, \mathbf{w}'_r)'$. This is called a *classification likelihood* [7].

Next, we assume a multinomial prior for the binary memberships \mathbf{Y} :

$$f(\mathbf{Y}|\boldsymbol{\mu}) = \prod_{i=1}^n \prod_{s=1}^r \mu_s^{y_{si}} \quad (3)$$

where μ_s is a *prior selection probability* for the s th generator and $\boldsymbol{\mu} = (\mu_1, \dots, \mu_r)'$. In this paper, for simplicity, each μ_s is fixed at $1/r$. The product of the classification likelihood (2) and the multinomial prior (3) gives a *complete likelihood* $f(\mathbf{X}, \mathbf{Y}|\mathbf{w}, \beta)$. Then, by integrating out the missing data \mathbf{Y} from the complete likelihood, we obtain a marginal likelihood of \mathbf{w} :

$$f(\mathbf{X}|\mathbf{w}, \beta) = \prod_{i=1}^n \sum_{s=1}^r \frac{1}{r} f(\mathbf{x}_i|\mathbf{w}_s, \beta). \quad (4)$$

This is called a *Gaussian mixture likelihood* [7, 8]. In a neural network interpretation [9], the Gaussian generators correspond to inner units and their centroids correspond to the synaptic weight vectors of the units. The hard and soft competitive-learning algorithms are ML estimation algorithms for the classification likelihood and the mixture likelihood respectively. Binary memberships and fuzzy memberships (defined in (12)) are regarded as activities of inner units in the respective competitive-learning.

Luttrell [4] found that the activity of each inner unit in the soft competitive learning is a Gaussian function of the distance from the first winner unit in an

inner-unit space when the weight points can be assumed to lie on a very smooth curve. In this case, the soft competitive-learning rule is identical to a SOM learning rule with a Gaussian neighborhood function. In the present paper, however, we treat a topological constraint as a smoothing prior of the weight vectors along a topology of the inner-unit assembly; this makes statistical inference easier in more general cases.

We now regard each of the commutated weight vectors $\mathbf{w}_{(j)} = (w_{1j}, \dots, w_{rj})'$, $j = 1, \dots, m$, as the discretization of a function on a topological space. Next, we let \mathbf{D} be a discretized differential operator on the space. For example, a discretized Laplacian operator on a one-dimensional line segment has entries

$$d_{ij} = \begin{cases} -2 & i + 1 = j \\ 1 & i + 1 = j \pm 1 \\ 0 & \textit{otherwise} \end{cases} \quad i = 1, \dots, r - 2 \quad j = 1, \dots, r. \quad (5)$$

We can change the topology by manipulating the row vectors of \mathbf{D} , each of which represents the smoothness around an inner unit. For example, the elimination of a pair of successive row vectors leads to a disconnection. Using \mathbf{D} , a *Gaussian smoothing prior* is defined as

$$f(\mathbf{w}|\alpha, \mathbf{D}) = \prod_{j=1}^m \left(\frac{\alpha}{2\pi}\right)^{l/2} (\det^+ \mathbf{M})^{1/2} \exp\left(-\frac{\alpha}{2} \|\mathbf{D}\mathbf{w}_{(j)}\|^2\right) \quad (6)$$

where $\mathbf{M} = \mathbf{D}'\mathbf{D}$, $l = \text{rank } \mathbf{M}$ and $\det^+ \mathbf{M}$ denotes the product of the positive eigenvalues of \mathbf{M} . The hyperparameter α represents the strength of the smoothing constraint.

In general, the prior (6) is partially improper if \mathbf{M} is singular; thus, the specification of this stochastic model is incomplete. However, this partial specification is sufficient for hyperparameter selection insofar as a common topology is considered. In section 3, a further specification of the model will be made to permit comparisons between various topologies.

Evidence of hyperparameters.

Using the Gaussian mixture likelihood (4) and the Gaussian smoothing prior (6), we now obtain a marginal likelihood of the hyperparameters α and β :

$$f(\mathbf{X}|\alpha, \beta, \mathbf{D}) = \int f(\mathbf{X}|\mathbf{w}, \beta) f(\mathbf{w}|\alpha, \mathbf{D}) d\mathbf{w} \quad (7)$$

which is also called the *evidence* of hyperparameters. In the empirical Bayesian approach, maximizing this evidence gives the optimal values of the hyperparameters. This is regarded as the MAP estimation of the hyperparameters when the

hyperprior (i.e., the prior of hyperparameters) is assumed to be uniform over a sufficiently wide range. In that cases, the prior of weights, with the hyperparameters replaced by their optimal values, is used for Bayesian inference on the weight parameters. For example, MAP estimates of weights are obtained by maximizing the posterior of weights, which is proportional to the integrand in (7).

The calculation of the evidence is difficult because of the integral in (7); thus, we need an approximate integral method. Here, we use an asymptotic approximation called Gaussian approximation [6] or Laplace’s method [10]. Using the MAP estimates of the weights $\hat{\mathbf{w}}$ and the negative Hessian of the integrand,

$$\mathbf{H}(\mathbf{w}) = -\frac{\partial^2}{\partial \mathbf{w} \partial \mathbf{w}'} \log f(\mathbf{X}, \mathbf{w} | \alpha, \beta, \mathbf{D}) \quad (8)$$

we approximate the log evidence by

$$\begin{aligned} \log f(\mathbf{X}, S_{\hat{\mathbf{w}}} | \alpha, \beta, \mathbf{D}) &= \log \int_{S_{\hat{\mathbf{w}}}} f(\mathbf{X}, \mathbf{w} | \alpha, \beta, \mathbf{D}) d\mathbf{w} \\ &\simeq \log f(\mathbf{X}, \hat{\mathbf{w}} | \alpha, \beta, \mathbf{D}) - \frac{1}{2} \log \det \mathbf{H}(\hat{\mathbf{w}}) + \frac{rm}{2} \log 2\pi \end{aligned} \quad (9)$$

where $S_{\hat{\mathbf{w}}}$ is a region dominated by $\hat{\mathbf{w}}$ in the weight space. The matrix $\mathbf{H}(\mathbf{w})$ is the sum of negative Hesse matrices of the log mixture likelihood and the log smoothing prior. The negative Hessian of the log prior is given by

$$\mathbf{H}_p = \alpha \mathbf{M} \otimes \mathbf{I}_m \quad (10)$$

where \mathbf{I}_m is an identity matrix with size m , and “ \otimes ” denotes the Kronecker product. The negative Hessian of the log likelihood can be also obtained exactly, although it is somewhat complicated [5].

2.2 Algorithms for MAP estimation of weights

The above estimation of the stochastic model requires a pair of algorithms: an algorithm for the MAP estimation of weights and that for the optimal hyperparameters. In this section, some algorithms for MAP estimation of weights are considered. An algorithm for the optimal hyperparameters is considered in section 4.

Gradient ascent algorithm.

The simplest algorithm for MAP estimation is the gradient ascent algorithm for the posterior of the weights, which corresponds to the elastic net algorithm [11].

In this algorithm, the weights are updated towards the steepest direction of ascent of the log posterior:

$$\begin{aligned} \mathbf{d}(\mathbf{w}) &= \frac{\partial}{\partial \mathbf{w}} \log f(\mathbf{X}, \mathbf{w} | \alpha, \beta, \mathbf{D}) \\ &= \beta \sum_{i=1}^n (\mathbf{P}_i \otimes \mathbf{I}_m) (\mathbf{1}_r \otimes \mathbf{x}_i - \mathbf{w}) - \alpha (\mathbf{M} \otimes \mathbf{I}_m) \mathbf{w} \end{aligned} \quad (11)$$

where $\mathbf{P}_i, i = 1, \dots, n$, are diagonal matrices whose entries are *fuzzy memberships*:

$$p_{si} = p(y_{si} = 1 | \mathbf{x}_i, \mathbf{w}, \beta) = \frac{f(\mathbf{x}_i | \mathbf{w}_s, \beta)}{\sum_{s=1}^r f(\mathbf{x}_i | \mathbf{w}_s, \beta)} \quad s = 1, \dots, r \quad (12)$$

and $\mathbf{1}_r$ is an r -dimensional column-vector of 1s. This algorithm has strong localization of computation, which makes it easy to implement on parallel computers. However, steepest ascent directions near a stationary point deviate somewhat from the direction towards the stationary point when the objective function has a narrow shape around the peak. In this case, gradient ascent algorithms are slow.

Expectation-maximization (EM) algorithm.

We can use the EM algorithm [12] for the MAP estimation of our model [5, 13]. This algorithm is also regarded as an approximate Newton-Raphson algorithm using an approximate Hessian instead of \mathbf{H} . This approximate Hessian is

$$\begin{aligned} \mathbf{H}_{\text{EM}}(\mathbf{w}) &= -E_{\mathbf{Y}} \left(\frac{\partial^2}{\partial \mathbf{w} \partial \mathbf{w}'} \log f(\mathbf{X}, \mathbf{Y} | \mathbf{w}, \beta) | \mathbf{X}, \mathbf{w}, \beta \right) + \mathbf{H}_p \\ &= (\beta \mathbf{N} + \alpha \mathbf{M}) \otimes \mathbf{I}_m \end{aligned} \quad (13)$$

where $E_{\mathbf{Y}}(\cdot | \cdot)$ denotes conditional expectation with respect to \mathbf{Y} , and $\mathbf{N} = \sum_{i=1}^n \mathbf{P}_i$. The update rule of the EM algorithm is now expressed in the same form as the Newton-Raphson rule:

$$\mathbf{w}^{(k+1)} = \mathbf{w}^{(k)} + \mathbf{H}_{\text{EM}}^{-1}(\mathbf{w}^{(k)}) \mathbf{d}(\mathbf{w}^{(k)}) \quad (14)$$

where $\mathbf{w}^{(k)}$ is a temporary estimate at the k th step. Unlike \mathbf{H} , The calculation of \mathbf{H}_{EM} and its inversion is not so heavy for its sparsity. However, this requires a special matrix solver for sparse matrices, and thus its implementation on parallel computers may be difficult.

An EM algorithm has better update directions than the corresponded gradient ascent algorithm, since the directions are corrected using the approximate Hessian. However, the normal use of EM algorithms does not lead to fast convergence, since the variation of parameters becomes too small at the last stage. To overcome this difficulty, some acceleration methods for EM algorithms are presented [14].

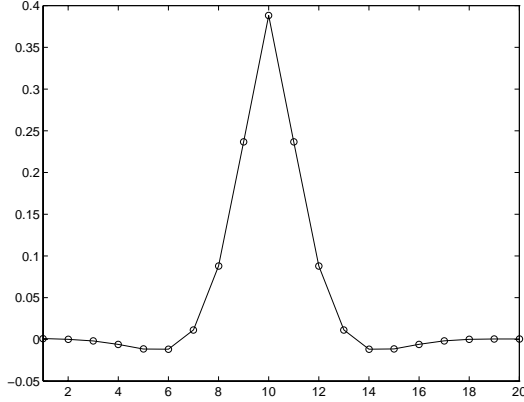


Figure 1: The graph of the tenth row-vector of $\mathbf{K}_{\text{SOM}} = (\beta\bar{\mathbf{N}} + \alpha\mathbf{M})^{-1}$ ($n = 20, r = 20, \alpha = 1, \beta = 1$).

SOM algorithm.

The update rule of the EM algorithm (14) is also expressed as

$$w_{sj}^{(k+1)} = \frac{\sum_{t=1}^r k_{st} n_t \bar{x}_{tj}}{\sum_{t=1}^r k_{st} n_t} \quad s = 1, \dots, r \quad j = 1, \dots, m \quad (15)$$

where k_{st} is an entry in $\mathbf{K}_{\text{EM}} = (\beta\mathbf{N} + \alpha\mathbf{M})^{-1}$, \bar{x}_{tj} is the mean of data weighted by the fuzzy membership:

$$\bar{x}_{tj} = \frac{1}{n_t} \sum_{i=1}^n p_{ti} x_{ij} \quad (16)$$

and $n_t = \sum_i p_{ti}$ [5]. Note that the denominator of (15) is always $1/\beta$. This rule can be regarded as kernel smoothing using the kernel \mathbf{K}_{EM} .

Here, we approximate \mathbf{H}_{EM} by its expectation:

$$\mathbf{H}_{\text{SOM}} = E_{\mathbf{X}}(\mathbf{H}_{\text{EM}}(\mathbf{w}) | \mathbf{w}, \beta) = (\beta\bar{\mathbf{N}} + \alpha\mathbf{M}) \otimes \mathbf{I}_m \quad (17)$$

where $\bar{\mathbf{N}} = \bar{n}\mathbf{I}_r$ and $\bar{n} = n/r$. The associated kernel $\mathbf{K}_{\text{SOM}} = (\beta\bar{\mathbf{N}} + \alpha\mathbf{M})^{-1}$ is now independent of individual data samples; thus, it can be fixed through the learning process. The new update rule (15) using \mathbf{K}_{SOM} is regarded as a batch version of the SOM algorithm [2, 15] using soft, rather than hard, competition (12). By calculating \mathbf{K}_{SOM} concretely for \mathbf{D} in (5), we find that each row vector of this matrix has a Mexican-hat shape (figure 2.2), and thus can be regarded as a lateral interaction matrix (corresponding to a neighborhood function). Unlike the other algorithms, the SOM algorithm does not necessarily converge to the MAP estimates. Nevertheless, it is expected to give a good approximation near a stationary point for a sufficient data size, since $\mathbf{H}_{\text{EM}}(\hat{\mathbf{w}}) \rightarrow \mathbf{H}_{\text{SOM}}$ as $n \rightarrow \infty$.

3 Topology Selection

In our model, a topology of an inner-unit assembly is represented by \mathbf{D} in the smoothing prior. In reality, \mathbf{D} represents a smoothness structure of weights; thus, it implies more than the connectivity among units. In general, according to such a structure the improper dimension of the prior varies. Although improper priors are commonly used in Bayesian inference, comparison between models with different improper dimensions is impossible. In the traditional Bayesian approach (e.g. [16]), all priors are made to be proper using subjective knowledge. In our model, however, it is difficult to introduce subjective priors, since the subspace corresponding to the improper dimension has no easy interpretation. We thus attempt to use an empirical Bayesian approach for topology selection.

Initially, we remark that $\mathbf{M} = \mathbf{D}'\mathbf{D}$ always has $\mathbf{1}_r$ as an eigenvector with zero eigenvalue. We then decompose the vector space for $\mathbf{w}_{(j)}$ into three orthogonal linear subspaces: $\mathbf{R}^r = S_1 \oplus S_2 \oplus S_3$, where S_1 is a subspace spanned by the row vectors of \mathbf{D} , S_2 is a one-dimensional subspace along $\mathbf{1}_r$ and S_3 is the remainder. The smoothing prior (6) is a prior on S_1 only. For S_2 , we can use a subjective prior, since the projection of $\mathbf{w}_{(j)}$ onto S_1 has an easy interpretation in term of the centroid of all weight points. However, we can ignore this prior in many cases, since it is regarded as common among various model structures and flat. For S_3 , we assume a spherical Gaussian prior with variance $1/\xi$:

$$f(\mathbf{w}|\xi, \mathbf{D}) = \prod_{j=1}^m \left(\frac{\xi}{2\pi} \right)^{k/2} \exp \left(-\frac{\xi}{2} \|\mathbf{E}\mathbf{w}_{(j)}\|^2 \right) \quad (18)$$

where \mathbf{E} is the orthogonal projector to S_3 and $k = \dim S_3 = r - l - 1$.¹

A new prior of weights is now given by the product of priors on S_1 and S_3 :

$$f(\mathbf{w}|\alpha, \xi, \mathbf{D}) = f(\mathbf{w}|\alpha, \mathbf{D})f(\mathbf{w}|\xi, \mathbf{D}). \quad (19)$$

From this prior and the mixture likelihood, we can obtain the evidence of α , β and ξ :

$$f(\mathbf{X}|\alpha, \beta, \xi, \mathbf{D}) = \int f(\mathbf{X}|\mathbf{w}, \beta) f(\mathbf{w}|\alpha, \xi, \mathbf{D}) d\mathbf{w}. \quad (20)$$

Using the Gaussian approximation again, we obtain a new log evidence:

$$\varepsilon(\alpha, \beta, \xi) = \log f(\mathbf{X}, S_{\hat{\mathbf{w}}}| \alpha, \beta, \xi, \mathbf{D})$$

¹Alternatively, we can assume a spherical Gaussian prior on $S_2 \oplus S_3$, which is the null space of \mathbf{M} , rather than on S_3 only. In fact, this method leads to similar results to the above method in many cases. However, this method is avoided since it depends upon the choice of the origin.

$$\begin{aligned}
&= -n \log r + \frac{nm}{2} \log \beta + \sum_{i=1}^n \log \sum_{s=1}^r \exp \left(-\frac{\beta}{2} \|\mathbf{x}_i - \hat{\mathbf{w}}_s\|^2 \right) \\
&\quad + \frac{ml}{2} \log \alpha + \frac{m}{2} \log \det^+ \mathbf{M} + \frac{mk}{2} \log \xi \\
&\quad - \frac{1}{2} \sum_{j=1}^m (\alpha \|\mathbf{D}\hat{\mathbf{w}}_{(j)}\|^2 + \xi \|\mathbf{E}\hat{\mathbf{w}}_{(j)}\|^2) - \frac{1}{2} \log \det \tilde{\mathbf{H}} + \text{const.} \quad (21)
\end{aligned}$$

where $\tilde{\mathbf{H}} = \mathbf{H} + \xi \mathbf{W} \otimes \mathbf{I}_m$ and $\mathbf{W} = \mathbf{E}'\mathbf{E}$.

If ξ is regarded as much smaller than α and β , the MAP estimation algorithm for the weights and the evidence for α and β need not be changed. We can thus employ the previous procedure for hyperparameter selection. However, the evaluation of the topology requires an estimate of ξ , given by the maximizer of (21).

Furthermore, we must consider the symmetry of the weight configuration. For example, a line-segment topology has a pair of weight configurations with the identical evidence value, each of which has reverse indexing to the other. On the other hand, a topology obtained by dividing the line segment has at least four equivalent configurations. In particular, a division at the center leads to eight equivalent configurations. The logarithm of number of such equivalent configurations has to be added to the log evidence.

We can now compare models with various connection styles using the new evidence following hyperparameter selection.

4 Fast algorithm for hyperparameter search

In this section, we attempt to obtain a fast hyperparameter-search algorithm using the derivatives of the evidence with respect to the hyperparameters. These derivatives are difficult to obtain owing to the complicated dependence of the evidence on the hyperparameters. Thus, we need a further approximation of the evidence.

First, the variations of the MAP estimates $\hat{\mathbf{w}}$ with the hyperparameters are regarded as small and ignored in the calculation of the derivatives. Such an approximation is adopted in MacKay's hyperparameter-search algorithm [6] for back-propagation learning. Next, the negative Hessian of the log posterior \mathbf{H} is replaced by its approximation. In AutoClass [16], the conditional expectation of a complete likelihood is used as an approximation of the mixture likelihood for model selection. This corresponds to the use of \mathbf{H}_{EM} as an approximation of \mathbf{H} . Here, we use \mathbf{H}_{SOM} .

Using this approximation, we obtain the derivatives of the log evidence with

respect to hyperparameters:

$$\frac{\partial \varepsilon}{\partial \beta} = \frac{nm}{2\beta} - \frac{1}{2} \sum_{i=1}^n \sum_{s=1}^r p_{si} \|\mathbf{x}_i - \hat{\mathbf{w}}_s\|^2 - \frac{1}{2} \text{tr}\{\tilde{\mathbf{H}}_{\text{SOM}}^{-1}(\bar{\mathbf{N}} \otimes \mathbf{I}_m)\} \quad (22)$$

$$\frac{\partial \varepsilon}{\partial \alpha} = \frac{ml}{2\alpha} - \frac{1}{2} \sum_{j=1}^m \|\mathbf{D}\hat{\mathbf{w}}_{(j)}\|^2 - \frac{1}{2} \text{tr}\{\tilde{\mathbf{K}}_{\text{SOM}}^{-1}(\mathbf{M} \otimes \mathbf{I}_m)\} \quad (23)$$

$$\frac{\partial \varepsilon}{\partial \xi} = \frac{mk}{2\xi} - \frac{1}{2} \sum_{j=1}^m \|\mathbf{E}\hat{\mathbf{w}}_{(j)}\|^2 - \frac{1}{2} \text{tr}\{\tilde{\mathbf{K}}_{\text{SOM}}^{-1}(\mathbf{W} \otimes \mathbf{I}_m)\} \quad (24)$$

where

$$\tilde{\mathbf{H}}_{\text{SOM}} = \mathbf{H}_{\text{SOM}} + \xi \mathbf{W} \otimes \mathbf{I}_m = (\beta \bar{\mathbf{N}} + \alpha \mathbf{M} + \xi \mathbf{W}) \otimes \mathbf{I}_m. \quad (25)$$

Then, by setting these derivatives to zeros and regarding ξ to be sufficiently smaller than α and β , we obtain recursive update formulae for the hyperparameters:

$$\hat{\beta}^{-1} = \frac{1}{m(n - \gamma - k - 1)} \sum_{i=1}^n \sum_{s=1}^r p_{si} \|\mathbf{x}_i - \hat{\mathbf{w}}_s\|^2 \quad (26)$$

$$\hat{\alpha}^{-1} = \frac{1}{m\gamma} \sum_{j=1}^m \|\mathbf{D}\hat{\mathbf{w}}_{(j)}\|^2 \quad (27)$$

$$\hat{\xi}^{-1} = \frac{1}{mk} \sum_{j=1}^m \|\mathbf{E}\hat{\mathbf{w}}_{(j)}\|^2 \quad (28)$$

where γ is the *effective number* of weight parameters in S_2 , defined by

$$\gamma = l - \alpha \text{tr}(\mathbf{K}_{\text{SOM}} \mathbf{M}). \quad (29)$$

Using the positive eigenvalues of \mathbf{M} , $\lambda_s, s = 1, \dots, l$, this effective number is calculated by

$$\gamma = \sum_{s=1}^l \frac{\beta \bar{n}}{\beta \bar{n} + \alpha \lambda_s}. \quad (30)$$

In fact, it is sufficient to use the update formula for ξ (28) once only after the search for α and β .

During a search by this algorithm, the calculation of the evidence itself is unnecessary. This also lightens the hyperparameter search significantly, since the calculation of the Hessian and its determinant in the evidence formula (21) was a bottleneck of the direct search algorithm. Moreover, while the calculation of the evidence requires strict convergence of the weight estimation algorithm to avoid negative Hessians, the new algorithm has not such problem. Thus, by virtue of a generous convergence condition for the weight-estimation algorithm, further acceleration is possible. In fact, the weights and the hyperparameters can be estimated simultaneously.

5 Simulation

We study the validity of the above methods for a simple case via simulation experiments. Artificial data are generated from two independent standard Gaussian random series e_{i1} and e_{i2} by

$$\begin{aligned} x_{i1} &= 4(i-1)/(n-1) - 2 + \sigma e_{i1} \\ x_{i2} &= \begin{cases} \sin(2\pi(i-1)/(n-1) - \theta) + \sigma e_{i2} & i \leq n/2 \\ \sin(2\pi(i-1)/(n-1) + \theta) + \sigma e_{i2} & i > n/2 \end{cases} \quad i = 1, \dots, n \end{aligned}$$

where θ represents the gap size at a discontinuous point. Here, we use $\theta = 0, 0.1\pi$ and 0.2π . In addition, we use two levels of noise: low noise ($\sigma = 0.1$) and high noise ($\sigma = 0.3$). We use a data size $n = 100$ in the low-noise condition and data sizes $n = 100$ and 400 in the high-noise condition. The samples of these data are shown in figure 5.

For each data set, three types of models with inner-unit size $r = 20$ are applied. The first model (M1) has a one-dimensional line-segment topology and the Laplacian operator (5). The second and third models (M2 and M3) have topologies obtained by dividing the first model at the fifth and tenth links respectively. Each of their Laplacian operators is obtained by eliminating an associated pair of row vectors from \mathbf{D} in (5).

For the first model, initial weights are given by the principal component analysis of the data set: that is, the weights are positioned at regular intervals on a line segment along the first principal-component vector with the same length as the data range. We use $\alpha = 10^4$ as its initial value for the low-noise condition and $\alpha = 10^3$ for the high-noise condition. Each initial value of β is given by the inverse of mean squared distance between the data points and a straight line fitted to the data set. The initial conditions for the other models are given by the estimates of the weights and hyperparameters obtained for the first model.

From such an initial condition, the update of weights by the EM rule (15) and the update of the hyperparameters by (26) and (27) are iterated alternately. When the relative variations of both α and β are smaller than 10^{-3} or α exceeds its initial value, the hyperparameters are fixed at the current values. Then only the EM algorithm for weights is continued until the sum of squared variations of weights is less than 10^{-25} . Here, we use an acceleration method for the EM algorithm. At the end of this procedure, ξ is obtained using (28), and then the log evidence (21) is calculated.

Optimal weight configurations obtained in this manner are also shown in figure 5. In addition, a graph of averaged log evidence and a peak histogram for 50 different data sets in each condition are illustrated in figure 5. For $\theta = 0$ and $\theta = 0.2\pi$, the peak histograms show the almost perfect performance of topology

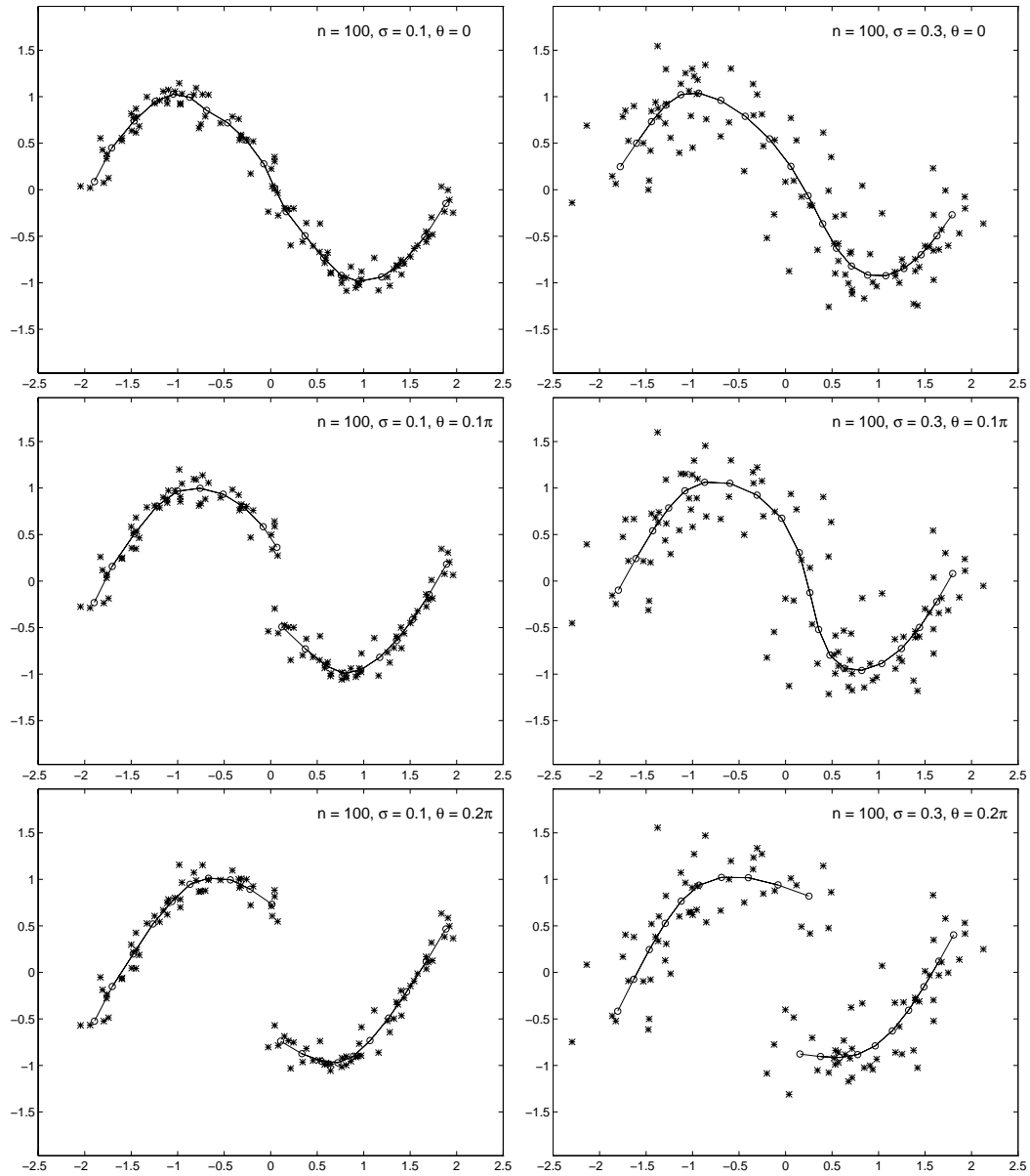


Figure 2: Scatter plots and the best weight-configuration.

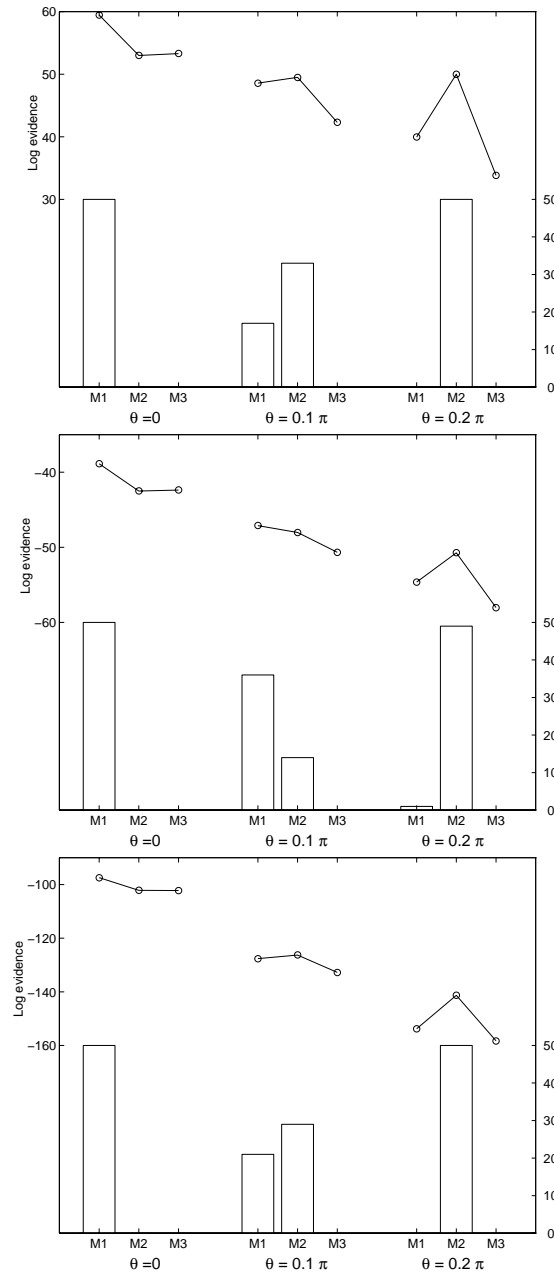


Figure 3: Graphs of the averaged log evidence and peak histograms for 50 data sets. Data sizes and noise levels are (a) $n = 100, \sigma = 0.1$, (b) $n = 100, \sigma = 0.3$ and (c) $n = 400, \sigma = 0.3$, for each of which three sizes of gaps ($\theta = 0, 0.1\pi, 0.2\pi$) are used. The first model (M1) has a line-segment topology with 20 inner units. The second and third models (M2, M3) are obtained by dividing the first model at the fifth and tenth links respectively.

selection. For $\theta = 0.1\pi$, on the other hand, the performance depends upon the conditions. For high noise and small data ($\sigma = 0.3$, $n = 100$), the presented method prefers the simpler model (M1) to the model underlying the data (M2). The corresponding data plot in figure 5 also shows that this discrimination is subtle even in our eyes. In general, the method seems to have a bias toward simple models. This agrees with the result for the hyperparameter selection in [5]. However, the graphs in figure 5 show that the performance of the method is improved by increasing the signal level against noise and size of data.

Moreover, for high noise and small data, M2 sometimes fails to obtain α smaller than its initial value. In such a situation, the direct search can acquire a rational value of α . This seems to show the limitation of the approximation method. However, in such a case, the variation of the evidence over a wide α -range is small, and thus this failure hardly affects the performance of topology selection.

6 Discussion

6.1 Other model selection approaches

So far we have focussed on the empirical Bayesian approach as a model-selection method. As mentioned in section 3, the traditional Bayesian approach using subjective priors is difficult to apply to our model. The empirical Bayesian approach, to be exact, is inconsistent with the Bayesian principle, since it uses data-dependent priors. However, it has been justified based on the minimum description length (MDL) principle [17]. In this view, the negative logarithm of the evidence of a model is called the *stochastic complexity* of the data relative to the model, and is regarded as the minimum description length of the data using the model. The principle that we should search for the model giving the data their shortest code is comprehensible and practical.

Another popular model-selection approach is that of resampling methods, such as cross-validation. Although the application of cross validation to our model is straightforward [5], it is time-consuming. First, it requires many repetitions of learning for each set of hyperparameter values for the stabilization of estimation. Although this is not problem for linear models, for which there exists a convenient method avoiding this repetition, non-linear models such as our model have no such convenient method. Moreover, cross validation scores have no closed function form with respect to hyperparameters, and thus we cannot construct a fast search algorithm using their derivatives.

There are also various simplified criteria for model selection, such as AIC, BIC and MDL criteria. However, these are justified only when there is sufficient data to ignore details of priors.

6.2 Other approximation methods for evidence

We used a Gaussian approximation for the integral in the evidence. We can also use certain Monte Carlo techniques for the integral, such as Gibbs sampler [10]. These methods are time consuming, but may be superior to the Gaussian approximation if data are small.

Another attractive approach for the evidence calculation has recently been presented. This is based on the extension of EM algorithm by Neal and Hinton [18]. In the conventional EM algorithm [12], the expectation of log complete likelihood by the posterior of missing data is calculated at the E-step, and then at the M-step the target parameters maximizing this expectation are obtained as the next temporary estimates. In the extended EM algorithm, any probability distribution increasing a *variational free energy* can be used at the E-step instead of the posterior of missing data. In particular, the posterior of missing data is the maximizer of this free energy. In the approximation method, the variational free energy is maximized within a restricted distribution family, which is selected to make the expectation calculation easy.

In particular, an estimation algorithm for *mixtures of experts* (ME) in [19] using this approximation method bears a strong relation to our method. The ME models are similar to Gaussian mixture models except that their centroids and prior selection probabilities are dependent on their input variables. The Gaussian mixture models are regarded as special cases of the ME models with constant input, and thus the same estimation algorithm is available. Our model can also be integrated into this framework by using a Gaussian smoothing prior for the centroid parameters, rather than the spherical Gaussian prior used in the ME models, and fixing the prior selection probabilities. The new estimation algorithm obtained in this manner is similar to the algorithm presented in the present paper, except for some points: the manner of soft competition is a little reformed from (12) in that it considers the uncertainty of weight estimates; the hyperparameter update rules are different from (26)–(28) but they are also obtained from (22)–(24) using \mathbf{H}_{EM} rather than \mathbf{H}_{SOM} . A preliminary simulation experiment showed that, in many case, these algorithms lead to almost the same solutions, although there are some cases where the new algorithm is slow to converge. The detail comparison of their performance is a future task. Moreover, the new approach is convenient in extending our model to include variable prior selection probabilities and variable hyperparameters among its generators.

7 Conclusion

A topology-selection method for SOMs based on an empirical Bayesian approach was presented. Moreover, a fast hyperparameter-search algorithm using the derivatives of evidence has been presented. The validity of these methods was confirmed by simulation experiments.

In this paper, we focused on the evaluation of topologies. The next step is to develop a search algorithm for optimal topologies. Since we probably cannot expect any effective algorithms to find exact optimal structures, good heuristic methods for structure generation will be required.

References

- [1] Kohonen T 1988 *Self-Organization and Associative Memory (2nd ed.)* (Berlin: Springer-Verlag)
Kohonen T 1990 The self-organizing map *Proc. IEEE* **78** 1464–1480
- [2] Luttrell S P 1990 Derivation of a class of training algorithms *IEEE Trans. Neural Networks* **1** 229–232
- [3] Yair E, Zeger K and Gersho A 1992 Competitive learning and soft competition for vector quantizer design *IEEE Trans. Signal Processing* **40** 294–309
- [4] Luttrell S P 1995 Using self-organizing maps to classify radar range profiles *Proceedings of the 4th International Conference on Artificial Neural Networks* (Cambridge) 335–340
- [5] Utsugi A 1997 Hyperparameter selection for self-organizing maps *Neural Comp.* **9** in press
- [6] MacKay D J C 1992 Bayesian interpolation *Neural Comp.* **4** 415–447
MacKay D J C 1992 A practical Bayesian framework for backprop networks *Neural Comp.* **4** 448–472
- [7] McLachlan G J and Basford K E 1988 *Mixture Models: Inference and Applications to Clustering* (New York: Marcel Dekker)
- [8] Wolfe J H 1970 Pattern clustering by multivariate mixture analysis *Multivariate Behavioral Research* **5** 329–350

- [9] Nowlan S J 1990 Maximum likelihood competitive learning *Advances in Neural Information Processing Systems* 2 ed Touretzky D S (San Mateo, CA.: Morgan Kaufmann) pp 574–582
- [10] Kass R E and Raftery A E 1995 Bayes factors *J. Amer. Statist. Assoc.* **90** 773–795
- [11] Durbin R and Willshaw D 1987 An analogue approach to the traveling salesman problem using an elastic net method *Nature* **326** 689–691
 Durbin R, Szeliski R and Yuille A 1989 An analysis of the elastic net approach to the traveling salesman problem *Neural Comp.* **1** 348–358
- [12] Dempster A P, Laird N M and Rubin D B 1977 Maximum likelihood from incomplete data via the EM algorithm *J. Roy. Statist. Soc. B* **39** 1–38
- [13] Yuille A L, Stolorz P and Utans J 1994 Statistical physics, mixture of distributions and the EM Algorithm *Neural Comp.* **6** 334–340
- [14] Jordan M I and Xu L 1995 Convergence results for the EM approach to mixtures of expert architectures *Neural Networks* **8** 1409–1431
- [15] Mulier F and Cherkassky V 1995 Self-organization as an iterative kernel smoothing process *Neural Comp.* **7** 1141–1153
- [16] Hanson R, Stutz J and Cheeseman P 1991 Bayesian classification theory *NASA Ames Research Center* FIA-90-12-7-01
- [17] Rissanen J 1989 *Stochastic Complexity in Statistical Inquiry* (Singapore: World Scientific)
- [18] Neal R M and Hinton G E 1993 A new view of the EM algorithm that justifies incremental and other variants *preprint* University of Toronto
- [19] Waterhouse S, MacKay D and Robinson T 1996 Bayesian methods for mixtures of experts *Advances in Neural Information Processing Systems* 8 ed Touretzky D S et al (Cambridge: MIT Press) pp 351–357