A Bayesian Approach for Feature and Model Selection in Mixture-Based Clustering

C. Constantinopoulos, M.K. Titsias, A. Likas

11 – 2004

Preprint, no 11 – 7 / 2004

Department of Computer Science
University of Ioannina
45110 Ioannina, Greece

A Bayesian Approach for Feature and Model Selection in Mixture-Based Clustering

C. Constantinopoulos
Department of Computer Science
University of Ioannina
GR 45110 Ioannina, Greece
ccostas@cs.uoi.gr

M. K. Titsias
School of Informatics
University of Edinburgh
Edinburgh EH1 2QL, UK
M.Titsias@sms.ed.ac.uk

A. Likas
Department of Computer Science
University of Ioannina
GR 45110 Ioannina, Greece
arly@cs.uoi.gr

Abstract

We present a Bayesian method for mixture model training that simultaneously treats the feature selection and the model selection problem. The method is based on the integration of a mixture model formulation that takes into account the saliency of the features and a Bayesian approach to mixture learning that can be used to estimate the number of mixture components. The motivation of the proposed method was the empirical observation that the existence of irrelevant features deteriorates the performance of the Bayesian approach for estimating the number of components, therefore feature relevance should be taken into account. The proposed learning algorithm follows the variational framework and it can simultaneously optimize over the number of components, the saliency of the features and the parameters of the mixture model. We provide experimental results on several data sets indicating the robustness of the proposed method in the presence of irrelevant features.

1 Introduction

Mixture models constitute a widely used approach for clustering and density estimation. The estimation of the parameters of mixture models with a predefined number of components is usually achieved by maximizing the likelihood using EM algorithm or several variants [1]. A very important problem in mixture learning deals with the selection of the number of components. It has been addressed using several approaches like cross-validation, statistical criteria and Bayesian methods [2, 3] which are of particular interest for us.

Apart from the selection of the number of components, another problem that naturally arises, especially in high dimensional data, deals with the detection of the most salient features for the task. A solution to this problem can be obtained by incorporating a feature selection process into the training algorithm so that some features that are not useful for partitioning the data into clusters (e.g they are just noise) can be discarded. Notice that when we estimate a mixture model, choosing the features and finding the number of components are strongly dependent problems. Clearly for different feature subsets we might get different number of clusters, e.g. see [4] for a discussion. This suggests that choosing the features and selecting the number of clusters should be addressed simultaneously.

The problem of feature selection in mixture models is difficult due to the absence of class information. Law et al. [5] have described a mixture model that incorporates a feature saliency determination process where each feature is useful up to a probability, so when this probability obtains a close to zero value the feature is effectively removed from consideration. This approach is attractive since it does not require an explicit search over the possible subsets of the features which generally is infeasible. To estimate the number of components, Law et al. [5, 6] employ the MML criterion in the training procedure.

To address both feature and model selection, in this paper we present a Bayesian variational framework for training the above mixture model that maximizes a lower bound of a marginal likelihood using the EM algorithm. This algorithm follows the variational framework of Corduneanu and Bishop [3] for training a Gaussian mixture model, and suitably integrates the model proposed in [5], so that it can simultaneously optimize over the number of components, feature saliency and the parameters of the mixture model. The motivation for the proposed method was the empirical observation that the performance of the variational method suggested in [3], deteriorates considerably by the existence of irrelevant features. Therefore feature relevance should be taken into account in the Bayesian approach for model selection.

In our experiments we compare the proposed variational Bayesian mixture model with the method of [3] (assuming diagonal covariances) and show that our method is more robust in estimating the number of mixture components in the presence of irrelevant features.

In section 2 we describe the Bayesian mixture model with feature saliency, and give a training algorithm based on variational learning. Comparative experiments are described in section 3. Finally, section 4 provides related work and discusses the strengths and drawbacks of the proposed method.

2 A Bayesian Mixture Model with Feature Relevance

In this section we discuss a Bayesian method for learning mixture models where the number of components and feature saliency are automatically specified. Particularly, in section 2.1 we define the Bayesian mixture model with feature saliency, and in section 2.2 we present a variational training method.

2.1 Bayesian Framework

Assume a set of data $X = \{x^n | n = 1, \ldots, N\}$, where each x^n is a feature vector in the d-dimensional space. We wish to cluster these data based on training a mixture model. We further assume that each component density of the mixture is factorized over the features, so that the features are considered to be independent given a component. Some of the features might be irrelevant for clustering while others may be more useful. Instead of assuming that there is a deterministic separation between useful and noisy features, we assume that a feature is useful up to a probability. Thus given some component, we assume that a feature of x is drawn from a mixture of two univariate sub–components, as in [5]. The first sub–component, that is different for each mixture component, generates 'useful' data, while the other sub–component, that is common to all mixture components, generates

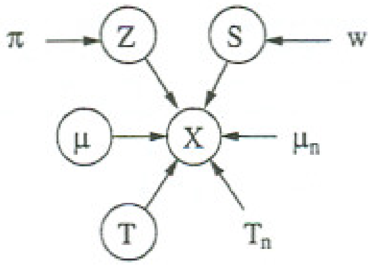

Figure 1: Graphical model for the generation of the observed data assuming a mixture density model and allowing noisy features.

'noisy' data.

In our work the above model is integrated in the Bayesian framework suggested in [3] for estimating the number of components. The proposed approach assumes that X is generated from the graphical model illustrated in Figure 1. This model implies a dependence of the observed variable x^n on the j-th mixture component through the hidden variables z_j^n, where $z_j^n \in \{0, 1\}$ and $\sum_j z_j^n = 1$. If x^n is generated from the j-th component, then the value of z_j^n is one, otherwise is zero. The saliency of features is expressed through the hidden variables s_i^n, where $s_i^n \in \{0, 1\}$. If the value of s_i^n is one, then i-th feature of x^n has been generated from the 'useful' sub–component, otherwise it is generated from the 'noisy' sub–component.

Given the sets of hidden variables $Z = \{z_j^n\}$ and $S = \{s_i^n\}$, the data is assumed to be independently drawn from a Gaussian distribution

$$p(X|Z, \mu, T, S, \mu_n, T_n) = \prod_{n=1}^{N} \prod_{j=1}^{J} \left[\prod_{i=1}^{d} \mathcal{N}(x_i^n; \mu_{ji}, \tau_{ji})^{s_i^n} \mathcal{N}(x_i^n; \mu_{ni}, \tau_{ni})^{1-s_i^n} \right]^{z_j^n} . \quad (1)$$

The set $\mu = \{\mu_{ji}\}$ accumulates the means of the 'useful' sub–components, and $T = \{\tau_{ji}\}$ the inverse variances. Correspondingly, $\mu_n = \{\mu_{ni}\}$ and $T_n = \{\tau_{ni}\}$ are the parameters of the 'noisy' sub–component. The distribution of the hidden variables Z given the mixing probabilities $\pi = \{\pi_j\}$, and of the hidden variables S given the probabilities $w = \{w_i\}$ (feature saliencies) are given by

$$p(Z|\pi) = \prod_{n=1}^{N} \prod_{j=1}^{J} \pi_j^{z_j^n}, \quad (2)$$

$$p(S|w) = \prod_{n=1}^{N} \prod_{i=1}^{d} w_i^{s_i^n} (1 - w_i)^{1-s_i^n} . \quad (3)$$

The distribution of the observed data given the parameters can be obtained by marginalizing out the hidden variables Z and S from $p(X, Z, S|\pi, \mu, T, w, \mu_n, T_n)$

$$\begin{aligned} p(X|\pi, \mu, T, w, \mu_n, T_n) &= \prod_{n=1}^{N} \sum_{j=1}^{J} \pi_j \prod_{i=1}^{d} [w_i \mathcal{N}(x_i^n; \mu_{ji}, \tau_{ji}) \\ &+ (1 - w_i)\mathcal{N}(x_i^n; \mu_{ni}, \tau_{ni})] . \quad (4) \end{aligned}$$

This is the usual quantity that the maximum likelihood framework maximizes over the parameters. However this objective function cannot be used for selecting the number of components. Thus it is not useful in our case since we do not know the number of components.

In [5] they address this problem by applying the MML criterion through a component–wise version of the EM algorithm that enforces a pruning behaviour, regarding the components of the model. On the contrary, we adopt the Bayesian approach of [3]. In particular, we introduce Gaussian and Gamma priors for μ and T respectively

$$
\begin{aligned}
p(\mu) &= \prod_{j=1}^{J}\prod_{i=1}^{d}\mathcal{N}(\mu_{ji}; m_i, c), &(5)\\
p(T) &= \prod_{j=1}^{J}\prod_{i=1}^{d}\mathcal{G}(\tau_{ji}; \alpha, \beta), &(6)
\end{aligned}
$$

and marginilize them out. The hyperparameters m and c control the distribution of the mean vectors, and take fixed values. The mean m is set to the mean of all data, and the inverse variance c takes a small value to ensure that all possible means have non-zero probability. The hyperparameters α and β control the density of the inverse variance components, and take near zero values, so that the prior is broad (quite non-informative).

2.2 Variational Learning

The learning method we propose estimates the parameters of the model through maximization of the marginal likelihood $p(X|\pi, w, \mu_n, T_n)$, defined as

$$
p(X|\pi, w, \mu_n, T_n) = \sum_{Z,S}\int p(X, Z, \mu, T, S|\pi, w, \mu_n, T_n)\mathrm{d}\mu\mathrm{d}T, \qquad (7)
$$

with respect to the mixing probabilities π, feature saliencies w and the parameters of the noise components. Note that by assuming suitable prior distributions on the component parameters and marginalizing them out, we expect to smooth the likelihood surface (eq. (4)) and obtain a marginal likelihood that is more robust to overfitting. This methodology was proposed in [3] to optimize over the mixing probabilities π and infer the number of components in a typical mixture model with remarkable results.

Since the integration in equation (7) is intractable, we resort to the maximization of a lower bound $\mathcal{L}$ of the marginal likelihood

$$
\begin{aligned}
\mathcal{L}(Q) &= \sum_{Z,S}\int Q(Z, \mu, T, S)\log\frac{p(X, Z, \mu, T, S|\pi, w, \mu_n, T_n)}{Q(Z, \mu, T, S)}\mathrm{d}\mu\mathrm{d}T &(8)\\
&\leq \log p(X|\pi, w, \mu_n, T_n). &(9)
\end{aligned}
$$

The bound $\mathcal{L}(Q)$ contains a distribution $Q(Z, \mu, T, S)$, that approximates the posterior distribution $p(Z, \mu, T, S|X, \pi, w, \mu_n, T_n)$, and is constrained to be a product of the form $Q(Z, \mu, T, S) = Q_Z(Z)Q_\mu(\mu)Q_T(T)Q_S(S)$. In order to maximize $\mathcal{L}(Q)$ an iterative procedure is adopted that consists of two steps: first maximization of the bound with respect to Q, and subsequently with respect to π, w, μ_n and T_n.

The method does not assume any specific form for the factors of Q, instead it maximizes $\mathcal{L}(Q)$ with respect to the functional form of Q_Z, Q_μ, Q_T and Q_S. Using standard variational analysis techniques we find

$$
\begin{aligned}
Q_Z(Z) &= \prod_{n=1}^{N}\prod_{j=1}^{J} r_{jn}^{z_j^n}, &(10)\\
Q_\mu(\mu) &= \prod_{j=1}^{J}\prod_{i=1}^{d}\mathcal{N}(\mu_{ji}; m_{ji}^{v}, c_{ij}^{v}), &(11)
\end{aligned}
$$

$$Q_T(T) = \prod_{j=1}^{J} \prod_{i=1}^{d} \mathcal{G}(\tau_{ji}; \alpha_{ji}^{v}, \beta_{ji}^{v}), \tag{12}$$

$$Q_S(S) = \prod_{n=1}^{N} \prod_{i=1}^{d} \rho_{in}^{s_i^n} (1 - \rho_{in})^{1 - s_i^n}. \tag{13}$$

The variational parameters $r_{jn}, m_{ji}^{v}, c_{ji}^{v}, \alpha_{ji}^{v}, \beta_{ji}^{v}$ and ρ_{in} emerge from the maximization, and determine the densities involved in Q. The variational parameters themselves are defined using the expected values of z_j^n, μ_{ji}, τ_{ji}, s_i^n and functions of them. Using the functional forms of Q_Z, Q_μ, Q_T and Q_S, we can derive the expectations and use them in the definitions of the variational parameters, obtaining the following equations

$$r_{jn} = \frac{\tilde{r}_{jn}}{\sum_{j=1}^{J} \tilde{r}_{jn}}, \tag{14}$$

$$\begin{aligned} \tilde{r}_{jn} = {} & \pi_j \exp \left\{ \sum_{i=1}^{d} (1 - \rho_{in}) \log \mathcal{N}(x_i^n; \mu_{ni}, \tau_{ni}) + \frac{1}{2} \sum_{i=1}^{d} \rho_{in} \left(\psi(\alpha_{ji}^{v}) - \log \beta_{ji}^{v} \right) \right. \\ & \left. - \frac{1}{2} \sum_{i=1}^{d} \rho_{in} \frac{\alpha_{ji}^{v}}{\beta_{ji}^{v}} \left[(x_i^n)^2 + \left(m_{ji}^{v} \right)^2 - 2 x_i^n m_{ji}^{v} + \frac{1}{c_{ji}^{v}} \right] \right\}, \end{aligned} \tag{15}$$

$$m_{ji}^{v} = \frac{c\, m_i + \frac{\alpha_{ji}^{v}}{\beta_{ji}^{v}} \sum_{n=1}^{N} r_{jn} \rho_{in} x_i^n}{c + \frac{\alpha_{ji}^{v}}{\beta_{ji}^{v}} \sum_{n=1}^{N} r_{jn} \rho_{in}}, \tag{16}$$

$$c_{ji}^{v} = c + \frac{\alpha_{ji}^{v}}{\beta_{ji}^{v}} \sum_{n=1}^{N} r_{jn} \rho_{in}, \tag{17}$$

$$\alpha_{ji}^{v} = \alpha + \frac{1}{2} \sum_{n=1}^{N} r_{jn} \rho_{in}, \tag{18}$$

$$\beta_{ji}^{v} = \beta + \frac{1}{2} \sum_{n=1}^{N} r_{jn} \rho_{in} \left[(x_i^n)^2 + \left(m_{ji}^{v} \right)^2 - 2 x_i^n m_{ji}^{v} + \frac{1}{c_{ji}^{v}} \right], \tag{19}$$

$$\rho_{in} = \frac{\tilde{\rho}_{in}}{\tilde{\rho}_{in} + (1 - w_i) \exp \left\{ -\frac{1}{2} \tau_{ni} \left(x_i^n - \mu_{ni} \right)^2 + \frac{1}{2} \log \tau_{ni} \right\}}, \tag{20}$$

$$\begin{aligned} \tilde{\rho}_{in} = {} & w_i \exp \left\{ \frac{1}{2} \sum_{j=1}^{J} r_{jn} \left(\psi(\alpha_{ji}^{v}) - \log \beta_{ji}^{v} \right) \right. \\ & \left. - \frac{1}{2} \sum_{j=1}^{J} r_{jn} \frac{\alpha_{ji}^{v}}{\beta_{ji}^{v}} \left[(x_i^n)^2 + \left(m_{ji}^{v} \right)^2 - 2 x_i^n m_{ji}^{v} + \frac{1}{c_{ji}^{v}} \right] \right\}, \end{aligned} \tag{21}$$

where $\psi(x) = \mathrm{d} \log \Gamma(x) / \mathrm{d}x$. The maximization of $\mathcal{L}$ with respect to Q aims to find a tight bound of the log marginal likelihood. Although an exact maximization of $\mathcal{L}$ with respect to the variational parameters is impossible, as they are coupled together in a non-linear way, we can still improve the bound by iteratively updating the parameters using equations (14)–(21).

After the maximization of $\mathcal{L}$ with respect to Q, the second step of the method requires maximization of $\mathcal{L}$ with respect to π_j, w_i, μ_{ni} and τ_{ni}, providing the following update rules

$$\pi_j = \frac{1}{N}\sum_{n=1}^{N} r_{jn}, \tag{22}$$

$$w_i = \frac{1}{N}\sum_{n=1}^{N} \rho_{in}, \tag{23}$$

$$\mu_{ni} = \frac{\sum_{n=1}^{N} \rho_{in} x_i^n}{\sum_{n=1}^{N} \rho_{in}}, \tag{24}$$

$$\frac{1}{\tau_{ni}} = \frac{\sum_{n=1}^{N} \rho_{in}(x_i^n - \mu_{ni})^2}{\sum_{n=1}^{N} \rho_{in}}. \tag{25}$$

The above two-step procedure is repeated until convergence. Convergence can be monitored through inspection of the variational bound. The above algorithm has the property that it does not allow for Gaussians with similar parameters to fit the same cluster. Consequently, one of them dominates and the other gets removed.

3 Experimental Results

In order to evaluate how feature saliency affects mixture learning and clustering performance, we compared our method with the method of Corduneanu and Bishop [3], assuming diagonal covariance matrices. To derive the update equations for such a model, we fix feature saliencies $\rho_{in} = 1, \forall\, i, n$, and omit update equations (20), (21), and (23)–(25). In this way the noise component is not taken into account, so we can deduce the importance of irrelevant features regarding clustering. We considered synthetic datasets so that the true number of components was known in advance. In each experiment both methods started from the same initial parameter values provided by running the k–means algorithm. Also the k–means algorithm was applied with random initialization in each experiment.

The first dataset consisted of eight hundred data from a mixture of four equiprobable Gaussians: $\mathcal{N}([0,3]^T, I)$, $\mathcal{N}([1,9]^T, I)$, $\mathcal{N}([6,4]^T, I)$ and $\mathcal{N}([7,10]^T, I)$. Also eight 'noisy' features have been added, sampled from $\mathcal{N}(0,1)$. The same dataset was used in [5]. We started each method with forty components. We repeated the experiment ten times, and both methods always detected the four clusters. This was expected as the number of features is not quite large, and there is a 'hard' separation between features that contain only useful data and features with only noisy data.

In the second series of experiments, we compared the two methods on the dataset described in [7]: two thousand samples were drawn equiprobably from two 20-dimensional Gaussians: $\mathcal{N}(\mu_1, I)$ and $\mathcal{N}(\mu_2, I)$, where $\mu_1 = \mu$, $\mu_2 = -\mu$, and μ is a vector whose i-th component is $(1/i)^{1/2}$. Notice that as dimension increases the cluster means come closer, and the two clusters merge to form only one. Consequently as the number of features increases, they come more irrelevant to the separation of the clusters. We repeated the experiment ten times, starting each method with forty components. We observed that our method always detected correctly the two clusters as well as the descending salience of the features, as the dimension increases. On the contrary, the method of [3] three times detected two clusters, five times detected three clusters, and two times detected four clusters. The average saliency of features is presented in Table 1, the corresponding standard deviation was less than $2 \cdot 10^{-3}$ for all features.

We also conducted comparative experiments for clustering shapes, using a dataset inspired by the experiments in [8]. We created three hundred 9×9 grayscale images. Each image

Table 1: Average feature saliency for Trunk data

feature	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
saliency	0.56	0.39	0.32	0.28	0.24	0.21	0.21	0.16	0.17	0.16	0.17	0.16	0.13	0.13	0.14	0.12	0.12	0.13	0.10	0.10

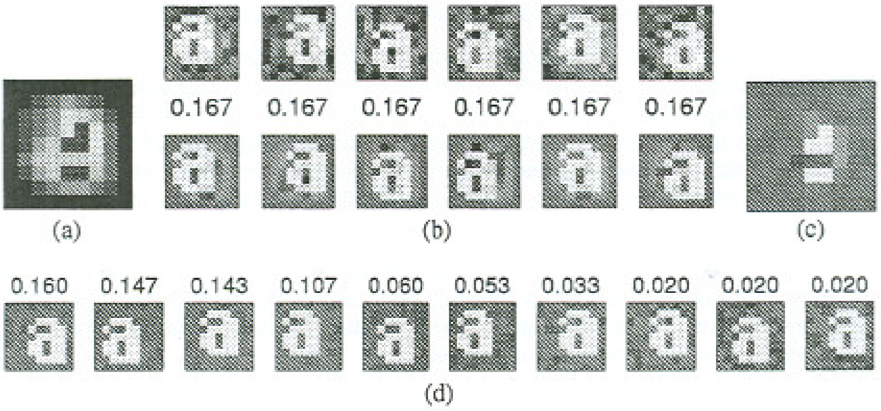

Figure 2: (a) The saliency of each feature displayed in grayscale, with black corresponding to zero and white to one. (b) Top row: a representative image of each cluster. Bottom row: the mean vectors of the six components detected by the proposed method, along with their respective mixing probabilities. (c) The mean vector of the 'noisy' component. (d) The mean vectors of the ten components with the greater mixing probabilities using method [3]

contained the shape of character 'a' placed in one of six different positions, so that the pixels across the image border were always background. The intensities of the background pixels were drawn from a Gaussian $\mathcal{N}(0.4, 12 \cdot 10^{-3})$, the foreground pixels from a Gaussian $\mathcal{N}(0.85, 0.4 \cdot 10^{-3})$, and all intensities were normalized in $[0, 1]$. Figure 2 displays the data and results from a typical experiment, with the top row in Figure 2(b) providing one representative image of each cluster. In each experiment, both methods were started with fifty components. The experiment was repeated ten times, and our method always detected six clusters, corresponding to the six different placements of the character. At the same time, only forty six features found to have saliency greater than 10^{-5}. Figure 2(a) visually displays the saliency of the features, while the bottom row of Figure 2(b) displays the mean vector of each of the six components with the respective mixing probabilities above each image. The mean vector of the 'noisy' component is displayed in Figure 2(c). The other method resulted in a wide range of components, varying from eighteen to forty three. This indicates the extra difficulty that irrelevant features introduce to the clustering problem. The mean vectors of the ten components with the greater mixing probabilities are displayed in Figure 2(d) with the respective mixing probability above each image.

4 Conclusions

We have presented a Bayesian variational algorithm for mixture learning that can automatically determine the number of components and the saliency of features. Our experiments show that this algorithm outperforms the variational Bayesian method of [3] in the presence of irrelevant features and this illustrates the importance of incorporating a feature selection process in learning mixture models.

Other Bayesian methods for feature selection for clustering have also been proposed in the literature. In [9] demonstrated that the marginal likelihood in multinomial mixtures can be used as a criterion for choosing the feature subset and finding the optimal number of clusters. However this method is still limited since it searches over feature subsets. Another Bayesian method described in [10] uses a shrinkage prior and performs a MAP estimation procedure which outputs an importance weighting of the features. However this method does attempt to integrate out parameters and also does not specify the number of clusters.

Our approach can also be compared with the method described in [6] where the MML criterion is used for estimating the number of components. Our approach is theoretically more appealing, since it has been developed in a Bayesian framework and the underlying assumptions are clearly described in the graphical model of Figure 1. On the other hand, the MML approach is based on a statistical criterion and is obtained after several assumptions and simplifications. In addition, as stated in [6], the MML approach can be viewed as a MAP approach with improper priors on π and w.

The main restriction of the proposed method is that the features are assumed to be conditionally independent given the component. We plan to elaborate further on this by generalizing our method so that covariance of the useful features can be modelled and simultaneously the feature saliency can be estimated.

References

[1] B. G. McLachlan and D. Peel. *Finite Mixture Models.* Wiley, New York, 2000.

[2] H Attias. A variational Bayesian framework for graphical models. In *Advances in Neural Information Processing Systems 12*. MIT Press, 2000.

[3] A. Corduneanu and C. M. Bishop. Variational bayesian model selection for mixture distributions. In T. Richardson and T. Jaakkola, editors, *Eighth International Conference on Artificial Intelligence and Statistics*, pages 27–34. Morgan Kaufmann, 2001.

[4] J. Dy and C. Brodley. Feature subset selection and order identification for unsupervised learning. In P. Langley, editor, *Proceedings of the Seventeenth International Conference on Machine Learning*, pages 247–254. Morgan Kaufmann, 2000.

[5] M. H. Law, M. A. T. Figueiredo, and A. K. Jain. Feature selection in mixture-based clustering. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural Information Processing Systems 15*. MIT Press, 2003.

[6] M. H. Law, M. A. T. Figueiredo, and A. K. Jain. Simultaneous feature selection and clustering using a mixture model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (accepted), 2004.

[7] G. V. Trunk. A problem of dimensionality: A simple example. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1(3), 1979.

[8] B. J. Frey and N. Jojic. Estimating mixture models of images and inferring spatial transformations using the EM algorithm. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1999.

[9] S. Vaithyanathan and B. Dom. Generalized model selection for unsupervised learning in high dimensions. In S. A. Solla, T. K. Leen, and K.-R. Müller, editors, *Advances in Neural Information Processing Systems 12*, Cambridge, MA, 2000. MIT Press.

[10] P. Gustafson, P. Carbonetto, N. Thompson, and N. de Freitas. Bayesian feature weighting for unsupervised learning, with application to object recognition. In *9th International Conference on Artificial Intelligence and Statistics*, 2003.