Clustering on a Subspace of Exponential Family Using Variational Bayes Method

Kazuho Watanabe
University of Tokyo
Kashiwa, Japan

Shotaro Akaho
The National Institute of Advanced Industrial Science and Technology
Tsukuba, Japan

Masato Okada
University of Tokyo, RIKEN
Wako, Japan

Abstract The e-PCA has been proposed to reduce the dimension of the parameters of probability distributions using Kullback information as a distance between two distributions. It also provides a framework for dealing with various data types such as binary and integer for which the Gaussian assumption on the data distribution is inappropriate. In this paper, we introduce a latent variable model for the e-PCA. Assuming the discrete distribution on the latent variable leads to mixture models whose parameters are constrained to the lower-dimensional subspace of exponential family distributions. We derive a learning algorithm for those mixture models based on the variational Bayes method applying Laplace’s method to carry out clustering on an arbitrary subspace. Combined with the estimation of the subspace, the resulting algorithm performs simultaneous dimensionality reduction and clustering.

Keywords: e-PCA, Mixture Model, Variational Bayes

1 Introduction

Exponential principal component analysis (e-PCA) has been proposed as a dimensionality reduction method that extracts a low dimensional subspace in the space of probability distributions [1]. This method provides a framework for introducing appropriate distance measures for special data types such as binary and integer. The original PCA uses the squared Euclidian distance based on the Gaussian assumption of the data. Extending this assumption to the exponential family distribution introduces more appropriate distance measures for various data types [6]. Such extension has been applied to clustering methods as well [4]. In fact, these techniques are required in neuroscience applications which commonly deal with the binary or integer data representing the presence or the counts of “spikes” in neuronal cells[8].

Interpreting dimensionality reduction and clustering methods as estimation in statistical models has provided various extensions of them. The probabilistic PCA, proposed as a latent variable model for the PCA, became the basis for several methods such as the Bayesian estimation for it, called the Bayesian PCA [5][9]. The probabilistic PCA assumes the Gaussian distributions on the data and the latent variables and makes the calculation of marginalization and expectation tractable. However, this is not the case for other distributions.

In this paper, we introduce a latent variable model for the e-PCA. Assuming the discrete distribution on the latent variable derives mixture models whose parameters are constrained to a low dimensional subspace. This constrained mixture model enables to perform clustering on the low dimensional subspace of exponential family distributions.

We derive a learning algorithm for the mixture models based on the variational Bayes method. For a fixed subspace, however, the algorithm requires intractable integration. We apply a Laplace approximation method to it to achieve the clustering on an arbitrary subspace. Furthermore, combined with the estimation of the subspace, this technique enables to perform simultaneous clustering and dimensionality reduction. The results of experiments conducted for integer and binary data show that the derived algorithm extracts the structure of data effectively and has high generalization ability.

2 Exponential Family and e-PCA

A statistical model \( p(x|\theta) \) of a random vector \( x \in \mathbb{X} \) is called the exponential family if its probability density function (or probability function) has the follow-
ing form\textsuperscript{1},
\begin{equation}
p(x|\theta) = \exp\{\theta \cdot F(x) + F_0(x) - G(\theta)\},
\end{equation}
where \(\theta = (\theta_1, \cdots, \theta_M)^T \in \Theta\) is called the natural parameter and \(\theta \cdot F(x)\) is the inner product of \(\theta\) and \(F(x) = (F_1(x), \cdots, F_M(x))^T\), that is, \(\theta \cdot F(x) = \sum_{j=1}^M \theta_j F_j(x)\). The function \(G(\theta)\) ensures that \(p(x|\theta)\) is a probability density function. The expectation parameter is defined by \(\eta = (\eta_1, \eta_2, \cdots, \eta_M)\) where \(\eta_j = \int F_j(x)p(x|\theta)dx\) is the expectation of \(F_j(x)\) with respect to the distribution (1).

The exponential family has an important property that there is a bijection between \(\eta\) and \(F\). Let \(L\) be the so-called “e-flat” subspace \([1][2]\). It is known \(L\) in \(\Theta\). Here \(\eta\) is the parameter and \(\theta\) is the corresponding natural parameter, that is, \(\begin{equation}
F(x^{(i)}) = \eta(\theta^{(i)}),
\end{equation}\) then the probability density of \(x^{(i)}\) is transformed to that of \(\theta^{(i)}\) \(\in \Theta^n\) \(\in \mathbb{R}^L\) by \(\begin{equation}
\bar{p}(\theta^{(i)}|\theta(\theta^{(i)})) = p(x^{(i)}|\bar{\eta}(\theta^{(i)}))[\eta(\theta^{(i)})]I(\theta^{(i)}),
\end{equation}\)
where \(w^{(i)}\) is the latent variable corresponding to \(x^{(i)}\) and \(I(\theta^{(i)})\) is the Jacobian of the above transformation (not necessary when \(x\) is discrete).

Assuming prior distribution \(p(w^{(i)}\) from eqs.\textsuperscript{(4)} and \textsuperscript{(6)}, we obtain the posterior distribution over \(W^n\) as follows,
\begin{equation}
p(W^n|\Theta^n) \propto \exp \left( -\sum_{i=1}^n K(\bar{\theta}(\theta^{(i)})) \right) \prod_{i=1}^n p(w^{(i)}).
\end{equation}

The e-PCA solution corresponds to the maximizer of the posterior distribution (the maximum a posteriori estimator of \(W^n\)) when the prior is uniform. In the following section, we assume the discrete distribution of the latent variables \(W^n\) and marginalize them to derive a mixture model. Section 5 describes the entire model and gives a method to estimate \(W^n\).

3.1 Probabilistic interpretation of e-PCA

The probabilistic PCA assumes the joint probability distribution of the datum \(x\) and the latent variable \(w = (w_1, w_2, \cdots, w_L) \in R^L\) as follows,
\begin{equation}
p(x, w) = p(x|\bar{\theta}(w))p(w),
\end{equation}
where \(\bar{\theta}(w)\) is defined by eq.(2). Throughout this paper, we suppose that \(p(x|\theta)\) in eq.(5) is an exponential family distribution given in eq.(1). The subsequent discussion before Section 5 assumes that \(U\) and \(u_0\) are fixed.

Suppose \(n\) training samples \(X^n = \{x^{(i)}\}_{i=1}^n\) are given. If we consider \(F(x^{(i)})\) as a sample of the expectation parameter and \(\theta^{(i)}\) as the corresponding natural parameter, then \(p(x^{(i)}|\bar{\theta}(w^{(i)}))\) is the expectation of \(F_j(x)\) with respect to the distribution \(\begin{equation}
p(x^{(i)}|\bar{\theta}(w)) = \sum_{i=1}^M \theta_j F_j(x)\),
\end{equation}\)

and marginalizing \(w^{(i)}\) in eq.(5) is an exponential family distribution given in eq.(1).

In this section, generalizing the probabilistic PCA, we interpret the e-PCA in terms of probabilistic inference.

\textsuperscript{1}When \(x\) is discrete, integration with respect to \(x\) is replaced by summation.
\textsuperscript{2}The Kullback information between the two distributions, \(p(x|\theta)\) and \(p(x|\tilde{\theta})\) is defined by
\begin{equation}
K(\theta|\tilde{\theta}) = \int p(x|\theta) \log \frac{p(x|\theta)}{p(x|\tilde{\theta})} dx.
\end{equation}

In the case of the exponential family, it is given by
\begin{equation}
K(\theta|\tilde{\theta}) = (\theta - \tilde{\theta})\eta(\tilde{\theta}) - G(\theta) + G(\tilde{\theta}).
\end{equation}
mixture model. Each component \( p(x|\theta(w_k)) \) is the exponential family distribution eq.(1) whose parameter \( \theta(w_k) \) defined by eq.(2) is constrained to the \( L \)-dimensional subspace.

This mixture model was derived based on non-parametric maximum likelihood estimation of the latent distribution \( p(w) \) and an expectation maximization (EM) algorithm was given for parameter estimation [10]. We derive the variational Bayesian estimation algorithm for this model. 

4 Variational Bayes Method

The next section derives the variational Bayesian algorithm for this model.

4.1 Optimal \( q(\omega) \) for fixed \( q(Z^n) \)

We define

\[
q(\omega) = \frac{\Gamma(\sum_{k=1}^{K} \phi_k)}{\prod_{k=1}^{K} \Gamma(\phi_k)} \prod_{k=1}^{K} a_k^{\phi_k - 1},
\]

(8)

\[
p(w) = \exp\{\xi_0(w_k \cdot \alpha_0 - G(\tilde{\theta}(w_k))) - \Phi(\alpha_0, \xi_0)\},
\]

(9)

where \( p(a) \) of \( a = \{a_k\}_{k=1}^{K} \) is the Dirichlet distribution with hyperparameter \( \phi_0 > 0 \) and \( p(w_k) \) has \( \mathbf{U}, u_k \) and \( \alpha_0 \in \mathbb{R}^L, \xi_0 \in \mathbb{R} \) as hyperparameters. The function \( \Phi(\alpha, \xi) \) of \( \xi \in \mathbb{R} \) and \( \alpha \in \mathbb{R}^L \) is defined by

\[
\Phi(\alpha, \xi) = \log \int \exp\{\xi \cdot w - G(\tilde{\theta}(w))\} dw.
\]

(10)

As in the EM and variational Bayesian algorithm for the usual mixture model, we introduce another latent (hidden) variable \( z \) that is 1 if the datum \( x \) is generated from the \( k \)th component and 0 otherwise. Then we have the following joint probability distribution of the observed data \( x \) and hidden variable \( z = (z_1, z_2, \cdots, z_K) \):

\[
p(x, z|\omega) = \prod_{k=1}^{K} (a_k p(x|\theta_k))^z_k.
\]

(11)

4.2 Optimal \( q(Z^n) \) for fixed \( q(\omega) \)

From eq.(10), as for \( (w_k)_q(w_k) \) and \( (G(\tilde{\theta}(w_k)))_q(w_k) \), we have

\[
(w_k)_q(w_k) = \frac{\partial \Phi(\alpha_k, \xi_k)}{\xi_k} \frac{\partial \Phi(\alpha_k, \xi_k)}{\partial \alpha_k}.
\]

(18)

\[
\{G(\tilde{\theta}(w_k))\}_q(w_k) = \alpha_k \cdot (w_k)_{q(w_k)} - \frac{\partial \Phi(\alpha_k, \xi_k)}{\partial \xi_k}.
\]

(19)

We also have \( \log a_k_\mathcal{Q}(a) = \Psi(\phi_k) - \Psi(\sum_{k=1}^{K} \phi_k) \) where

\[
\Psi(x) = (\log \Gamma(x))' = \frac{\Gamma'(x)}{\Gamma(x)}.
\]

(14)

is the di-gamma function. From (11), (13) and the above equations, the optimal \( q(Z^n) \) for the fixed \( q(\omega) \) is given by

\[
q(Z^n) \propto \prod_{i=1}^{n} \prod_{k=1}^{K} \exp(z^{(i)}_k s^{(i)}_k)
\]

(13)
where we have defined
\[ s_k^{(i)} = \Psi(\phi_k) - \Psi(\sum_{k=1}^{K} \phi_k) + \frac{\partial \Phi(\alpha_k, \xi_k)}{\partial \xi_k} \]
\[ + \left( \frac{1}{\xi_k} \frac{\partial \Phi(\alpha_k, \xi_k)}{\partial \alpha_k} \right) \cdot (U^T F(x^{(i)}) - \alpha_k). \tag{20} \]

Its mean is given by
\[ (z_k^{(i)})_{q(z^n)} = q(z_k^{(i)}) = 1 = \frac{e^{s_k^{(i)}}}{\sum_{k=1}^{K} e^{s_k^{(i)}}}. \tag{21} \]

The resulting variational Bayesian algorithm, initializing \{\phi_k\}, \{\alpha_k\} and \{\xi_k\}, iterates the following procedures alternately,
- compute \( (z_k^{(i)})_{q(z^n)} \) by (20) and (21),
- update \{\phi_k\}, \{\alpha_k\} and \{\xi_k\} by (14) and (17).

Note that the algorithm requires computing \( \frac{\partial \Phi}{\partial \theta} \) and \( \frac{\partial \Phi}{\partial \xi} \) in eq.(20) (and (18)).

\section{4.3 Full dimensional case}

When \( U \) is the unit matrix, \( u_0 = 0 \) and \( \hat{\theta}(w_k) = w_k \in \mathbb{R}^M \), we have the usual unconstrained mixture model. In this case, we can obtain the explicit forms of the function \( \Phi(\alpha, \xi) \) in eq.(10) for some examples of the exponential family [12]. For example, in the case of binomial component (with independent elements), defined by \( \theta_j = \log \frac{p_j}{1-p_j} \), \( F(x) = x \in \{0, 1, \cdots, N\}^M \), \( G(\theta) = \sum_{j=1}^{M} N \log(1 + e^{\theta_j}) \), \( F_0(x) = \sum_{j=1}^{M} \log \frac{N}{N-x_j} \), \( \Phi(\alpha, \xi) \) is given by
\[ \Phi(\alpha, \xi) = \sum_{j=1}^{M} \log \frac{\Gamma(\xi_\alpha_j) \Gamma(\xi(N - \alpha_j))}{\Gamma(\xi N)}. \]

We use the binomial mixture model in the experiments below also for the reduced dimensional case.

\section{4.4 Laplace approximation}

When the rank of the matrix \( U \) is \( L < M \) in general, we no longer have explicit forms of the function \( \Phi, \frac{\partial \Phi}{\partial \alpha} \) and \( \frac{\partial \Phi}{\partial \xi} \). In this section, we provide an approximation scheme based on Laplace’s method.

We assume that the exponent in the integral of eq.(10) takes the maximum at \( w \). Expanding the exponent in Taylor series around \( w \), we approximate it by taking the series up to the second-order term,
\[ \xi\{(\alpha \cdot w - G(U \hat{w} + u_0)) - \frac{1}{2} (w - \hat{w})^T J(w)(w - \hat{w})\}. \]

Here \( J(w) = U^T \frac{\partial G(U w + u_0)}{\partial \theta} U \) and \( w \) is given by solving the equation
\[ U^T \frac{\partial G(U w + u_0)}{\partial \theta} = \alpha \tag{22} \]
with respect to \( w \). Then, we have the approximation for the function \( \Phi \),
\[ \Phi(\alpha, \xi) \simeq \xi(\alpha \cdot \hat{w} - G(U \hat{w} + u_0)) - \frac{L}{2} \log \xi - \frac{1}{2} \log |J|. \]

Its partial derivatives are approximated by
\[ \frac{\partial \Phi}{\partial \alpha} = \xi \hat{w} - \frac{1}{2} \frac{\partial \log |J|}{\partial \alpha}, \tag{23} \]
\[ \frac{\partial \Phi}{\partial \xi} = \alpha \cdot \hat{w} - G(U \hat{w} + u_0) - \frac{L}{2} \frac{1}{\xi}. \tag{24} \]

Here we have applied eq.(22) to derive eq.(23).

The equation (22) to obtain \( \hat{w} \) is solved by an iterative method that initializes \( w \) with \( \hat{w} \) and repeats updating \( \hat{w} \) by adding
\[ d \hat{w} = J^{-1} \left(U^T U \alpha - U^T \eta (U \hat{w} + u_0)\right). \]

The derivation is omitted here [1].

We evaluate the term \( \frac{\partial \log |J|}{\partial \alpha_k} = (J^{-1})_{k1} \) as follows,
\[ \frac{\partial \log |J|}{\partial \alpha_k} = \sum_{k=1}^{L} (J^{-1})_{k1} \text{Tr} \left( \frac{\partial J}{\partial u_k} J^{-1} \right), \]
for \( l = 1, 2, \cdots, L. \)

\section{5 Dimensionality Reduction}

In the previous sections, we derived the variational Bayesian learning algorithm for the mixture model whose component parameter vectors are constrained to the \( L \)-dimensional subspace. Estimating the basis vectors \( U \) and the displacement vector \( u_0 \) as well enables dimensionality reduction simultaneously performing clustering. This section provides a method for obtaining low-dimensional representations of the data and that for estimating the subspace.

\subsection{5.1 Low dimensional representation}

We have discussed the finite mixture model obtained by assuming the discrete distributions on the latent variables, that are the low dimensional representations of the data, \( W^n = \{w^{(i)}\} \) and marginalizing them. Although we have obtained the posterior distribution of the cluster centers \( w^K = \{w_k\} \), we need to estimate the low dimensional representations \( W^n \) in some way. This section provides a method to estimate them.

Let \( a = \{a_k\}_{k=1}^{K}, \ w^K = \{w_k\}_{k=1}^{K} \) and \( p(a), \ p(w^K) \) be the prior distributions for them. The
joint probability distribution of the all variables \( \Theta^n, W^n, Z^n, w^K \), a \( \Theta^n \) is the set of the natural parameters whose corresponding expectation parameters are the training samples \( X^n \) is given by

\[
\prod_{i=1}^{n} \left[ p(\theta(i)|w(i)) p(w(i)|z(i), w^K) p(z(i)|\Theta) \right] p(a)p(w^K),
\]

where \( p(\theta(i)|w(i)) \) is the distribution of \( \theta \) defined by eq.(6), \( p(w(i)|z(i), w^K) = \prod_{k=1}^{K} \delta(w(i) - w_k)^2 \) and \( p(z(i)|\Theta) = \prod_{k=1}^{K} a_k^{z(i)} \). Hence, marginalizing \( a, w^K, Z^n \), we obtain the posterior distribution of \( W^n \),

\[ p(W^n|\Theta^n) \propto \prod_{i=1}^{n} \left[ p(\theta(i)|w(i)) p(w(i)|w(i)) \right]. \]

From eqs.(6) and (9), \( p(W^n|\Theta^n) \) is given by

\[
\prod_{i=1}^{n} \exp \{ \xi_0(w(i) \cdot \alpha(i) - G(\theta(w(i))) - \Phi(\alpha(i), \xi_0) \}, \]

where \( \xi_0 = 1 + \xi_0 \) and \( \alpha(i) = \frac{U^F w(i) + \xi_0}{1 + \xi_0} \). The posterior mean of \( w(i) \) is

\[
\langle w(i) \rangle_{p(w(i)|\phi(i))} = \frac{1}{1 + \xi_0} \frac{\partial \Phi(\alpha(i), 1 + \xi_0)}{\partial \alpha},
\]

which can be used as the low dimensional representation of each datum.

5.2 Estimation of basis vectors

We have developed the clustering method on the fixed subspace defined by the basis vectors \( U \) and \( u_0 \). One way to estimate the basis vectors is to apply the e-PCA to all the given data. Another way is to apply the e-PCA to cluster centers and is expected to take into account the cluster structure of the data set. In this case, the steepest descent method for minimizing

\[
\sum_{k=1}^{K} K(\theta(v_k)|\theta(\langle w_k \rangle_{q(w_k)})) \text{ updates the basis vector } u_k \rightarrow \hat{u}_k \text{ by } \hat{u}_k = u_k + \epsilon_u \sum_{k=1}^{K} \eta_k (v_k - \eta (\langle w_k \rangle_{q(w_k)})) \langle w_k \rangle_{q(w_k)},
\]

where \( \{ n_k \} \) and \( \{ v_k \} \) are defined in eq.(14) and \( \epsilon_u \) is a small constant 3. We have put \( w_{00} = 1 \).

We can perform dimensionality reduction and clustering simultaneously by updating the basis vectors with the above rule after once updating \( q(Z^n) \) and \( q(\omega) \) with the variational Bayes method described in Section 4.

6 Experiments

In this section, we present the results of experiments where the derived algorithm was applied to count and binary data.

6.1 Count data

In the first experiment, we used the binomial mixture model for 10-dimensional count data with the basis vectors \( U \) and \( u_0 \) fixed. The component 10-dimensional binomial distribution is constituted of independent elements as the one presented in Section 4.3. For the maximum likelihood (ML) estimation of the constrained mixture model, a learning method using an EM algorithm was derived [10]. In order to compare the variational Bayesian (VB) estimation method derived in the previous sections with the ML and the maximum a posteriori (MAP) method, we applied the respective methods to the model and evaluated the training and generalization errors.

We set, as the target true distribution, the 2 component binomial mixture model whose parameters are given by \( a = (a_1, a_2)^T = (0.8, 0.2)^T, \eta(\theta_1) = 2 \cdot 1, \eta(\theta_2) = 6 \cdot 1 \) where 1 is the 10-dimensional vector whose elements are all one. We denote this set of parameters as a vector \( \omega^* \). We considered reducing the dimensionality to \( L = 1 \) and fixed \( u_1 = \frac{1}{\sqrt{10}}, u_0 = 0 \). The ML, MAP and VB estimation methods were applied to the mixture models consisting of \( K = 2, 4, 6, 8 \) components.

The update rule for the MAP estimation is obtained by letting

\[
\alpha_k \propto \left\{ \begin{array}{ll}
\phi_k - 1 & (\phi_k > 1), \\
0 & (\phi_k \leq 1),
\end{array} \right.
\]

replacing the terms \( \Psi(\phi_k) - \Psi(\sum_{k=1}^{K} \phi_k) \) by \( \log \alpha_k \) in eq.(20) and rewriting eqs.(23),(24) as

\[
\frac{\partial \Phi}{\partial \alpha} = \xi \omega, \quad \frac{\partial \Phi}{\partial \xi} = \alpha \cdot \omega - G(U \omega + u_0),
\]

(In the MAP estimation, these are not approximation but exact expression). Furthermore, setting the hyper-parameters as \( \phi_0 = 1, \xi_0 = 0 \) reduces to the update rule for the ML method. We set as \( \phi_0 = 0.1, \xi_0 = 2/(\sqrt{MN}), \alpha_0 = \sqrt{MN}/2 \) in the VB and MAP methods. In each method, \( n_k, v_k \) were initialized by adding random numbers to \( n/K, \frac{1}{n} \sum_{i=1}^{n} x_i \).

The training error was evaluated by the log-likelihood ratio,

\[
\frac{\sum_{i=1}^{n} \log p(x_i|\omega^*)}{p(x_i|\hat{\omega})},
\]
where \( \hat{\omega} \) is the estimator of the parameter in each method. In the VB method, we used \( \hat{\omega} = \langle \omega \rangle_{q(\omega)} \). The generalization error was evaluated by,

\[
\frac{1}{n} \sum_{i=1}^{n} \log \frac{p(\tilde{x}_i|\omega^*)}{p(\tilde{x}_i|X^n)} ,
\]

where \( \{\tilde{x}_i\}_{i=1}^{n} \) is the set of test data and \( p(x|X^n) \) is the predictive distribution. In the ML and MAP methods, we defined \( p(x|X^n) = p(x|\hat{\omega}) \) using the estimator \( \hat{\omega} \) of each estimation. In the VB method, we used \( \langle p(x|\omega) \rangle_{q(\omega)} \)

\[
= \sum_{k=1}^{K} \langle a_k \rangle_{q(\omega)} e^{\Phi(a_k, \xi_k) - \Phi(a_k, \xi_k) + \alpha_k \Phi(a_k) + f_0(a_k)},
\]

where \( \xi_k = 1 + \xi_k \) and \( \alpha_k = \frac{\text{U}^{T} \Phi(a_k) + \xi_k \alpha_k}{\xi_k + 1} \).

Dividing 10000 data generated from the true distribution into \( n = 50 \) training and \( \tilde{n} = 9950 \) test samples, we used 200-fold cross validation to evaluate the training and generalization errors and averaged over 200 sets of the samples. Figure 1 and 2 show the averages of the training and generalization errors respectively for various \( K \).

Figure 1: Averages of training errors of ML (solid line), MAP (dotted line) and VB (dashed line) methods for different numbers of components \( K \) with standard errors.

Figure 2: Averages of generalization errors of ML (solid line), MAP (dotted line) and VB (dashed line) methods for different numbers of components \( K \) with standard errors.

Table 1: estimated number of components

<table>
<thead>
<tr>
<th></th>
<th>Full Dimensional</th>
<th>Reduced Dimensional</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gaussian</td>
<td>12.9 ± 1.8</td>
<td>4.0 ± 0.6</td>
</tr>
<tr>
<td>Bernoulli</td>
<td>5.6 ± 1.6</td>
<td>4.0 ± 0.8</td>
</tr>
</tbody>
</table>

In the MAP and VB estimation, when the model has redundant components \( K > 2 \), their mixing proportions \( a_k \) are estimated to be almost 0 and they are pruned automatically, due to the effect of the prior distribution. We observed this phenomenon in the experiment. As a result, these two methods tend to overfit less than the ML estimation as can be seen in Figure 1 and 2. Moreover, the VB estimation shows the smaller generalization errors with smaller variance than the MAP estimation. This may be because the regularization term influences the prior distribution too much and pruned even necessary components in the MAP estimation.

6.2 Binary data

In the second experiment, we investigated the property of the derived algorithm for the 16-dimensional binary data used in [12][10]. Three ‘prototype’ vectors were generated as the three true cluster centers with each bit randomly drawn from Bernoulli(1/2). Then 600 data were generated by taking 200 copies of each prototype and inverting each bit with probability 0.15.

We first applied the full dimensional Gaussian and Bernoulli (binomial with \( N = 1 \)) mixture models to investigate how the quality of the clustering depends on the appropriateness of the model. We used the isotropic Gaussian mixture model and gave its variance parameter \( \sigma^2 = 0.1275 \) which is the true variance of each cluster in order to compare the Bernoulli mixture with the Gaussian mixture that is expected to give the best performance. The two models were compared in terms of the estimated number of components. We applied the variational Bayesian algorithm to both models with \( K = 30 \) components and estimated the number of components by counting the components with \( n_k > 1 \) after the convergence. The hyperparameters were set to \( \phi_0 = 1 \), \( \alpha_0 = 1/2 \) and \( \xi_0 = 2 \). In Table 1 (left two columns), we present the numbers of components estimated by the two models averaged over ten data sets.

The Bernoulli mixture model shows much better result than the Gaussian mixture model. This indicates that the appropriate model for the special data type is advantageous to capture the true structure of the data.

We next applied the derived variational Bayesian algorithm reducing the dimensionality to \( L = 2 \) and simultaneously estimating the basis vectors, \( \text{U} \) and \( \text{u}_0 \). Figure 3 demonstrates an example of the projected data by the method described in Section 5.1. Also shown are the cluster centers estimated by the Bernoulli mixture model with the respective mixing proportions indicated beside them.
Figure 3: Projection of binary data consisting of three types (+, △, *) and the cluster centers (open square) estimated by the VB method with the mixing proportion of each cluster indicated beside them.

Table 2: generalization errors

<table>
<thead>
<tr>
<th>Method</th>
<th>$L = 2$, VB</th>
<th>$L = 2$, ML</th>
<th>$L = 16$, VB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Error</td>
<td>22.9 ± 5.3</td>
<td>27.2 ± 7.2</td>
<td>36.7 ± 10.0</td>
</tr>
</tbody>
</table>

We used the Bernoulli and Gaussian mixtures with $K = 30$ components and estimated the number of components in the same way as above. We set the hyperparameters to $\phi_0 = 1$, $\alpha_0 = (0.5, 0.5)$ and $\xi_0 = 0.5$. Table 1 (right two columns) shows the estimated number of components averaged over the same 10 data sets. The Bernoulli mixture still showed more accurate estimation than the Gaussian mixture model whose accuracy was much improved by reducing the dimensionality.

To compare the ML and the VB estimation methods in the simultaneous dimension reduction and clustering, we also applied the EM algorithm to the Bernoulli mixture model and evaluated the generalization errors, eq.(29). The ML method does not have any mechanism to prune the redundant components and it actually used all the 30 components in every trial. In Table 2, we present the generalization errors of respective methods averaged over 10 data sets. Also presented for comparison is that for the full dimensional Bernoulli mixture model with the VB algorithm applied. The reduced dimensional models have lower errors than the full dimensional one. This indicates that reducing the dimensionality enhances the generalization ability of the model. Furthermore, the VB algorithm shows the lowest error indicating that the pruning of redundant components improves generalization as well.

7 Discussion & Conclusion

We developed a method for simultaneous clustering and dimensionality reduction based on the probabilistic model of the e-PCA. We derived the variational Bayesian estimation algorithm by providing a Laplace approxima-

### References


