Journal of Physics: Conference Series **699** (2016) 012018 doi:10.1088/1742-6596/699/1/012018

Asymmetric biclustering with constrained von Mises-Fisher models

Kazuho Watanabe[1], Hsiang-Yun Wu[2], Shigeo Takahashi[3] and Issei Fujishiro[2]

[1] Department of Computer Science and Engineering, Toyohashi University of Technology, 1-1 Hibarigaoka, Tempaku-cho, Toyohashi, Aichi 441-8580, Japan
[2] Department of Information and Computer Science, Keio University, 3-14-1 Hiyoshi, Kohoku-ku, Yokohama, Kanagawa 223-8522, Japan
[3] Department of Computer Science and Engineering, University of Aizu, Tsuruga, Ikki-machi, Aizu-Wakamatsu, Fukushima 965-8580, Japan

E-mail: wkazuho@cs.tut.ac.jp

Abstract. As a probability distribution on the high-dimensional sphere, the von Mises-Fisher (vMF) distribution is widely used for directional statistics and data analysis methods based on correlation. We consider a constrained vMF distribution for block modeling, which provides a probabilistic model of an asymmetric biclustering method that uses correlation as the similarity measure of data features. We derive the variational Bayesian inference algorithm for the mixture of the constrained vMF distributions. It is applied to a multivariate data visualization method implemented with enhanced parallel coordinate plots.

1. Introduction

Finding correlated subspaces, sets of features, is one of promising approaches to understanding high-dimensional data. Visual analysis methods for such a purpose have been developed in the visualization community [1, 2]. Although these methods are effective, the results of the analysis depend heavily on knowledge and observation skill of the users and thus may fail to illuminate important subspaces. To support such data exploration, we developed a multivariate data visualization method, which combines a biclustering technique and parallel coordinate plots [3]. Biclustering, also known as co-clustering or two-mode clustering, performs a simultaneous clustering of the rows (samples or instances) and columns (features or variables) of a data matrix consisting of multivariate data samples [4, 5]. Conventional popular models of biclustering are based on the so-called block models [5]. Although there are extensions of block models using probabilistic inference [6, 7], they treat the clusterings of rows and columns of the data matrix basically symmetrically. This is not necessarily appropriate if a block model is combined with data visualization methods such as parallel coordinate plots, where correlation is considered as a suitable measure of similarity between features (axes), and is in fact used for reordering or contracting axes [8, 9]. Therefore, in our previous work [3], we developed an asymmetric biclustering method, which uses correlation as the similarity measure of features, based on spherical k-means [10, 11]. However, the proposed asymmetric biclustering method assumes that the numbers of clusters are predefined both for rows and columns, and the model selection mechanism has yet to be implemented.

In this paper, we propose a probabilistic model for the asymmetric biclustering method by the von Mises-Fisher (vMF) distribution, which is a probability distribution on a high-dimensional unit sphere. Providing a probabilistic model offers advantages such as dealing with missing data and automatic control of model's complexity by Bayesian methods. However, combining the vMF model with block modeling requires additional constraints on the vMF model that the elements of its mean direction parameter vector are divided into groups and are common within each group. Although the maximum likelihood estimation algorithm was derived for the mixture of (unconstrained) vMF distributions [11] and the variational Bayesian inference algorithms were derived as a deterministic approximation of Bayesian inference [12, 13], their extensions to the constrained vMF distributions are not straightforward due to the constraints. Thus in this paper, we formulate the variational Bayesian inference algorithm for the mixture of the constrained vMF distributions. Since in a small variance limit of the probabilistic model, the algorithm reduces to the asymmetric biclustering method proposed in [3], the derived algorithm is its extension that naturally provides a mechanism for controlling model's complexity, namely, the numbers of data sample clusters and data feature clusters. Variational Bayesian inference offers a criterion for model selection by its objective functional called variational free energy. Furthermore, the variational Bayesian inference algorithm automatically eliminates redundant clusters inheriting the nature of Bayesian inference [14]. Although such a Bayesian framework has been proposed for the usual (symmetric) biclustering model [6], we consider its application to the asymmetric biclustering model based on the mixture of vMF models. The derived inference algorithm is incorporated into the visual analysis framework proposed in [3], which is implemented with enhanced parallel coordinate plots. We demonstrate the model selection ability of the variational Bayesian inference through its applications to synthetic and real datasets.

2. Constrained von Mises-Fisher model for asymmetric biclustering

2.1. Block model

Suppose we are given an $n \times d$ data matrix $\mathbf{X}$ consisting of n samples and d features. Biclustering methods based on the block model [4, 5] divide the data matrix into $K \times L$ submatrices (blocks), each of which has the size $n_k \times d_l$ ($k = 1, \cdots, K$ and $l = 1, \cdots, L$), where n_k is the number of data samples assigned to the kth cluster of samples, and d_l is the number of features assigned to the lth cluster of dimensions. Hence, $\sum_{k=1}^{K} n_k = n$ and $\sum_{l=1}^{L} d_l = d$ hold.

Let $\mathbf{y} = \{y_i\}_{i=1}^{n}$ and $\mathbf{z} = \{z_j\}_{j=1}^{d}$ be latent variables, where $y_i \in \{1, 2, ..., K\}$ indicates the sample cluster assignment of the ith data sample and $z_j \in \{1, 2, ..., L\}$ indicates the feature cluster assignment of the jth feature. Then it holds that $n_k = \sum_{i=1}^{n} \delta_{y_i,k}$ and $d_l = \sum_{j=1}^{d} \delta_{z_j,l}$, where $\delta_{k,l} = 1$ if $k = l$, and $\delta_{k,l} = 0$ otherwise. The 1-of-K and 1-of-L representations of sample and feature cluster assignments are given by $(\delta_{y_i,1}, \cdots, \delta_{y_i,K})$ and $(\delta_{z_j,1}, \cdots, \delta_{z_j,L})$ for the ith sample and jth feature, respectively.

Each block has an associated parameter. Let $\nu_{k,l} \in \mathbb{R}$ be the parameter of the klth block and $\boldsymbol{\nu} = \{\nu_{k,l} : k = 1, ..., K, l = 1, ..., L\}$. Then, the block model estimates $\mathbf{y}$, $\mathbf{z}$ and $\boldsymbol{\nu}$ from the given data matrix $\mathbf{X}$.

2.2. Constrained von Mises-Fisher model

Normalizing each dimension of the data matrix so that each dimension has average zero and unit norm, we denote the ijth element of the normalized data matrix by x_{ij} and the jth column vector by $\mathbf{x}_j = (x_{1j}, \cdots, x_{nj})$, that is, $\frac{1}{n}\sum_{i=1}^{n} x_{ij} = 0$ and $\|\mathbf{x}_j\|^2 = \sum_{i=1}^{n} x_{ij}^2 = 1$ hold for $j = 1, \cdots, d$. This means that each feature lies on the $(n-1)$-dimensional unit sphere, $\mathbf{x}_j \in \mathbb{S}^{n-1}$.

Given the latent variables, we assume the following model on the generation of the data

matrix $\mathbf{X}$,

$$p(\mathbf{X}|\boldsymbol{\nu}, \mathbf{s}, \mathbf{y}, \mathbf{z}) = \prod_{j=1}^{d} p_{\mathrm{vMF}}(\mathbf{x}_j|s_j\boldsymbol{\mu}_{z_j}), \tag{1}$$

where

$$p_{\mathrm{vMF}}(\mathbf{x}|\boldsymbol{\mu}) = \frac{\exp(\kappa\mathbf{x}\cdot\boldsymbol{\mu})}{C_{n-1}(\kappa)} \tag{2}$$

is the von Mises-Fisher (vMF) distribution for $\mathbf{x} = (x_1, ..., x_n) \in \mathbb{S}^{n-1}$ with the mean direction parameter $\boldsymbol{\mu} \in \mathbb{S}^{n-1}$ and the concentration parameter $\kappa > 0$. The normalizaton constant is defined by the function,

$$C_n(\kappa) = \int_{\|\mathbf{x}\|=1} \exp(\kappa\mathbf{x}\cdot\boldsymbol{\mu})d\mathbf{x} = \frac{(2\pi)^{n/2}}{\kappa^{\frac{n}{2}-1}} I_{\frac{n}{2}-1}(\kappa),$$

where I_v is the modified Bessel function of the first kind of order v.[1]

For block modeling, we assume that the mean parameter vectors of features in (1) are defined for $l = 1, ..., L$ by

$$\boldsymbol{\mu}_l = (\nu_{y_1,l}, ..., \nu_{y_n,l}),$$

where $\nu_{k,l} \in \mathbb{R}(k = 1, ..., K)$ is the mean parameter of the klth block. This means that the elements of the mean parameter vector $\boldsymbol{\mu}_l$ are constrained to have common values if the cluster assignments of data samples are the same.

Then the constraint of the vMF distribution in (2) entails the constraints,

$$\|\boldsymbol{\mu}_l\|^2 = \sum_{k=1}^{K} n_k\nu_{k,l}^2 = 1, \tag{3}$$

$$\boldsymbol{\mu}_l \cdot \mathbf{1} = \sum_{k=1}^{K} n_k\nu_{k,l} = 0, \tag{4}$$

for the block mean parameter $\boldsymbol{\nu} = \{\nu_{k,l}\}_{k=1,l=1}^{K,L}$.

Additionally, in the model (1), the variable $\mathbf{s} = \{s_j \in \{-1, +1\}\}_{j=1}^{d}$ is introduced to take into account the sign of correlation between $\mathbf{x}_j$ and $\boldsymbol{\mu}_{z_j}$. That is, the cluster centroid vector of features $\boldsymbol{\mu}_{z_j}$ is positively (negatively, resp.) correlated with the jth feature if $s_j = +1$ ($s_j = -1$, resp.).

Hence, given the data matrix $\mathbf{X}$, we estimate latent variables $\mathbf{y}$, $\mathbf{z}$ and $\mathbf{s}$, and the block mean parameters $\boldsymbol{\nu}$ under the above constraints. Note here that the model (1) assumes the independence over data features instead of data samples. Probabilistic models with such an independence assumption have been proposed, for example, in [15] and [16]. In particular, spectral dimensionality reduction methods such as isomap, locally linear embeddings, and Laplacian eigenmaps can be interpreted through such probabilistic models [15].

Since the constraints (3) and (4) depend on the latent variable $\mathbf{y}$ through $\{n_k\}$, we express the block mean parameter $\nu_{k,l}$ as[2]

$$\nu_{k,l} = \frac{\xi_{k,l}}{n_k R_l}, \quad \text{for } \xi_{k,l} \in \mathbb{R},$$

[1] The usual vMF distribution on the $(n-1)$-dimensional sphere has the normalization constant $C_n(\kappa)$ whereas in (2), we have $C_{n-1}(\kappa)$ because of the additional constraint, $\mathbf{x}\cdot\mathbf{1} = \sum_{i=1}^{n} x_i = 0$, which implies that $\mathbf{x}$ lies on $(n-2)$-dimensional unit sphere.

[2] If $n_k = 0$, let $\nu_{k,l} = \xi_{k,l} = 0$.

where

$$R_l = \sqrt{\sum_{k=1}^{K} \frac{\xi_{k,l}^2}{n_k}}.$$

If $\boldsymbol{\xi} = \{\xi_{k,l}\}_{k=1,l=1}^{K,L}$ satisfies $\sum_{k=1}^{K} \xi_{k,l} = 0$ for $l = 1, \cdots, L$ then the constraints (3) and (4) are satisfied. Hence, we consider $\boldsymbol{\xi}$ as a parameter instead of $\boldsymbol{\nu}$ henceforth, and express the model (1) as $p(\mathbf{X}|\boldsymbol{\xi}, \mathbf{s}, \mathbf{y}, \mathbf{z})$.

2.3. The overall generative model

As in the usual probabilistic models for biclustering [6], we assume the following categorical models for latent variables,

$$p_{\mathrm{cat}}(\mathbf{y}|\boldsymbol{\rho}) = \prod_{i=1}^{n} \rho_{y_i} = \prod_{i=1}^{n} \prod_{k=1}^{K} \rho_k^{\delta_{y_i,k}}, \tag{5}$$

$$p_{\mathrm{cat}}(\mathbf{z}|\boldsymbol{\pi}) = \prod_{j=1}^{d} \pi_{z_j} = \prod_{j=1}^{d} \prod_{l=1}^{L} \pi_l^{\delta_{z_j,l}}, \tag{6}$$

where $\boldsymbol{\rho} = \{\rho_k\}_{k=1}^{K}$ and $\boldsymbol{\pi} = \{\pi_l\}_{l=1}^{L}$ are mixing proportions satisfying $\rho_k \geq 0$ $(k = 1, ..., K)$ and $\sum_{k=1}^{K} \rho_k = 1$, and $\pi_l \geq 0$ $(l = 1, ..., L)$ and $\sum_{l=1}^{L} \pi_l = 1$.

Thus the overall generative model of the data matrix $\mathbf{X}$ is given by marginalizing out the latent variables,

$$\begin{aligned} p(\mathbf{X}|\boldsymbol{\xi}, \mathbf{s}, \boldsymbol{\rho}, \boldsymbol{\pi}) &= \sum_{\mathbf{z}} \sum_{\mathbf{y}} p(\mathbf{X}, \mathbf{y}, \mathbf{z}|\boldsymbol{\xi}, \mathbf{s}, \boldsymbol{\rho}, \boldsymbol{\pi}) \\ &= \sum_{\mathbf{z}} \sum_{\mathbf{y}} p_{\mathrm{vMF}}(\mathbf{X}|\boldsymbol{\xi}, \mathbf{s}, \mathbf{y}, \mathbf{z}) p_{\mathrm{cat}}(\mathbf{y}|\boldsymbol{\rho}) p_{\mathrm{cat}}(\mathbf{z}|\boldsymbol{\pi}), \end{aligned}$$

where $\sum_{\mathbf{y}}$ and $\sum_{\mathbf{z}}$ denote the summations over all possible configurations of latent variables $\mathbf{y}$ and $\mathbf{z}$, respectively.

3. Inference and parameter estimation

In this section, we formulate Bayesian inference for the biclustering model in the previous section. We assume the Dirichlet distributions as prior distributions for the mixing proportions of the data sample and feature clusters, respectively,

$$p_{\mathrm{Dir}}(\boldsymbol{\rho}) = \frac{\Gamma(K\alpha)}{\Gamma(\alpha^K)} \prod_{k=1}^{K} \rho_k^{\alpha-1}, \tag{7}$$

$$p_{\mathrm{Dir}}(\boldsymbol{\pi}) = \frac{\Gamma(L\beta)}{\Gamma(\beta^L)} \prod_{l=1}^{L} \pi_l^{\beta-1}, \tag{8}$$

where $\alpha > 0$ and $\beta > 0$ are hyperparameters, and $\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt$ is the gamma function.

The posterior distribution of the latent variables conditioned on the data matrix, $\boldsymbol{\xi}$, and $\mathbf{s}$ is

$$p(\mathbf{y}, \mathbf{z}|\mathbf{X}, \boldsymbol{\xi}, \mathbf{s}) = \int p(\mathbf{y}, \mathbf{z}, \boldsymbol{\rho}, \boldsymbol{\pi}|\mathbf{X}, \boldsymbol{\xi}, \mathbf{s}) d\boldsymbol{\rho} d\boldsymbol{\pi},$$

where $p(\mathbf{y}, \mathbf{z}, \boldsymbol{\rho}, \boldsymbol{\pi}|\mathbf{X}, \boldsymbol{\xi}, \mathbf{s}) = p(\mathbf{X}, \mathbf{y}, \mathbf{z}|\boldsymbol{\xi}, \mathbf{s}, \boldsymbol{\rho}, \boldsymbol{\pi}) p_{\mathrm{Dir}}(\boldsymbol{\rho}) p_{\mathrm{Dir}}(\boldsymbol{\pi}) / p(\mathbf{X}|\boldsymbol{\xi}, \mathbf{s})$, which is intractable because the likelihood function $p(\mathbf{X}|\boldsymbol{\xi}, \mathbf{s})$ requires the summations $\sum_{\mathbf{y}}$ and $\sum_{\mathbf{z}}$ of K^n and L^d terms, respectively.

3.1. Variational Bayesian inference

To obtain a tractable upper bound for the negative log-likelihood, $-\log p(\mathbf{X}|\boldsymbol{\xi}, \mathbf{s})$, we introduce an approximating posterior $q(\mathbf{y}, \mathbf{z}, \boldsymbol{\rho}, \boldsymbol{\pi})$, that factorizes as

$$q(\mathbf{y}, \mathbf{z}, \boldsymbol{\rho}, \boldsymbol{\pi}) = q(\mathbf{y})q(\mathbf{z})q(\boldsymbol{\rho})q(\boldsymbol{\pi}),$$

where the distribution $q(\mathbf{y})$ further factorizes as $q(\mathbf{y}) = \prod_{i=1}^{n} q(y_i)$. Then we obtain an upper bound for $F(\boldsymbol{\xi}, \mathbf{s}) = -\log p(\mathbf{X}|\boldsymbol{\xi}, \mathbf{s})$ as follows,

$$\begin{aligned} F(\boldsymbol{\xi}, \mathbf{s}) &\leq F(\boldsymbol{\xi}, \mathbf{s}) + \mathrm{KL}[q(\mathbf{y}, \mathbf{z}, \boldsymbol{\rho}, \boldsymbol{\pi})||p(\mathbf{y}, \mathbf{z}|\mathbf{X}, \boldsymbol{\xi}, \mathbf{s})] & (9) \\ &= \left\langle \log \frac{q(\mathbf{y}, \mathbf{z}, \boldsymbol{\rho}, \boldsymbol{\pi})}{p(\mathbf{X}, \mathbf{y}, \mathbf{z}, \boldsymbol{\rho}, \boldsymbol{\pi}|\boldsymbol{\xi}, \mathbf{s})} \right\rangle_{q(\mathbf{y}, \mathbf{z}, \boldsymbol{\rho}, \boldsymbol{\pi})} \equiv \overline{F}[q(\mathbf{y}), q(\mathbf{z}), q(\boldsymbol{\rho}), q(\boldsymbol{\pi}), \boldsymbol{\xi}, \mathbf{s}], & (10) \end{aligned}$$

where $\mathrm{KL}[q||p]$ denotes the Kullback information from q to p, and $\langle\cdot\rangle_q$ the expectation with respect to q. The inequality (9) follows from the non-negativity of the Kullback information or equivalently from Jensen's inequality. The upper bound $\overline{F}$, called the variational free energy, is minimized alternately with respect to one of $q(\mathbf{y})$, $q(\mathbf{z})$, $q(\boldsymbol{\rho})$, $q(\boldsymbol{\pi})$, $\boldsymbol{\xi}$, and $\mathbf{s}$ while the others are fixed. The variational free energy provides an estimate of model's negative log-likelihood. It is used as a model selection criterion by choosing the numbers of clusters, K and L, that minimize the variational free energy among candidate models.

Variational posteriors of mixing proportions: Minimizing $\overline{F}$ with respect to $q(\boldsymbol{\rho})$ and $q(\boldsymbol{\pi})$ while the other variational posteriors and parameters fixed, we have

$$q(\boldsymbol{\rho}) \propto \exp \left\langle \log(p_{\mathrm{cat}}(\mathbf{y}|\boldsymbol{\rho})p_{\mathrm{Dir}}(\boldsymbol{\rho})) \right\rangle_{q(\mathbf{y})},$$

and

$$q(\boldsymbol{\pi}) \propto \exp \left\langle \log(p_{\mathrm{cat}}(\mathbf{z}|\boldsymbol{\pi})p_{\mathrm{Dir}}(\boldsymbol{\pi})) \right\rangle_{q(\mathbf{z})}.$$

It follows from (5), (7), (6), and (8) that

$$\begin{aligned} q(\boldsymbol{\rho}) &\propto \prod_{k=1}^{K} \rho_k^{\overline{n}_k + \alpha - 1}, & (11) \\ q(\boldsymbol{\pi}) &\propto \prod_{l=1}^{L} \pi_l^{\overline{d}_l + \beta - 1}. & (12) \end{aligned}$$

This means that $q(\boldsymbol{\rho})$ and $q(\boldsymbol{\pi})$ are the Dirichlet distributions with hyperparameters $\{\overline{n}_k + \alpha\}$ and $\{\overline{d}_l + \beta\}$, respectively, where

$$\begin{aligned} \overline{n}_k &= \langle n_k \rangle_{q(\mathbf{y})} = \sum_{i=1}^{n} q(y_i = k), & (13) \\ \overline{d}_l &= \langle d_l \rangle_{q(\mathbf{z})} = \sum_{j=1}^{d} q(z_j = l), & (14) \end{aligned}$$

are the expected number of data samples assigned to the kth sample cluster and that of data features assigned to the lth feature cluster, respectively.

Variational posterior of feature cluster assignments and sign of correlation: As a function of only $q(\mathbf{z})$, $\overline{F}$ is minimized when

$$q(\mathbf{z}) \propto \exp\left\langle \log(p_{\mathrm{vMF}}(\mathbf{X}|\mathbf{z},\mathbf{y},\boldsymbol{\xi})p_{\mathrm{cat}}(\mathbf{z}|\boldsymbol{\pi}))\right\rangle_{q(\mathbf{y})q(\boldsymbol{\pi})}.$$

Combined with (2), (6) and (12), this yields that

$$q(\mathbf{z}) = \prod_{j=1}^{d} q(z_j),$$

and

$$q(z_j = l) \propto \exp\left(\kappa s_j \eta_l^{(j)} + \Psi(\overline{d}_l + \beta) - \Psi(d + L\beta)\right), \tag{15}$$

where $\Psi(x) = d\log\Gamma(x)/dx$ is the digamma function, $\eta_l^{(j)} = \mathbf{x}_j \cdot \overline{\boldsymbol{\mu}}_l$, $\overline{\boldsymbol{\mu}}_l$ is the n-dimensional vector whose ith element is $\overline{\nu}_{y_i,l}$, which can be interpreted as the expected cluster centroid vector of the lth feature cluster, and

$$\overline{\nu}_{k,l} = \langle \nu_{k,l} \rangle_{q(\mathbf{y})} \simeq \frac{\xi_{k,l}}{\overline{n}_k \overline{R}_l}, \tag{16}$$

$$\overline{R}_l = \sqrt{\sum_{k=1}^{K} \frac{\xi_{k,l}^2}{\overline{n}_k}}. \tag{17}$$

Here, we have used the approximation that the expectation of $\nu_{k,l}$ with respect to $q(\mathbf{y})$ is computed by the expectation of n_k. Since $\overline{F}$ depends on the sign s_j through $-s_j \sum_{l=1}^{L} q(z_j = l)\eta_l^{(i)}$, we compute it for the two candidates of $q(z_j)$ with $s_j = +1$ and -1, and adopt s_j and $q(z_j)$ yielding smaller $\overline{F}$.

Variational posterior of sample cluster assignments: Similarly, $\overline{F}$ as a function of $q(y_i)$ is minimized by

$$q(y_i) \propto \exp\left\langle \log(p_{\mathrm{vMF}}(\mathbf{X}|\mathbf{z},\mathbf{y},\boldsymbol{\xi})p_{\mathrm{cat}}(\mathbf{y}|\boldsymbol{\rho}))\right\rangle_{q(\mathbf{y}^{\neg i})q(\mathbf{z})q(\boldsymbol{\rho})},$$

where $q(\mathbf{y}^{\neg i}) = \prod_{i' \neq i} q(y_{i'})$. This yields that

$$q(y_i = k) \propto \exp\left(\kappa \gamma_k^{(i)} + \Psi(\overline{n}_k + \alpha) - \Psi(n + K\alpha)\right), \tag{18}$$

where

$$\gamma_k^{(i)} = \sum_{j=1}^{d} s_j x_{ij} \left(\sum_{l=1}^{L} q(z_j = l) \left\langle \nu_{k,l}^{(y_i=k)} \right\rangle_{q(\mathbf{y}^{\neg i})}\right) + \sum_{i' \neq i} \sum_{k'=1}^{K} \sum_{j=1}^{d} s_j x_{i'j} \left(\sum_{l=1}^{L} q(z_j = l) \left\langle \nu_{k',l}^{(y_i=k)} \right\rangle_{q(\mathbf{y}^{\neg i})}\right), \tag{19}$$

$$\begin{aligned}
\left\langle \nu_{k',l}^{(y_i=k)} \right\rangle_{q(\mathbf{y}^{\neg i})} &= \frac{\xi_{k',l}}{(\overline{n}_{k'}^{\neg i} + \delta_{k',k})\overline{R}_l^{(y_i=k)}}, \\
\overline{R}_l^{(y_i=k)} &= \sqrt{\sum_{m=1}^{K} \frac{\xi_{m,l}^2}{\overline{n}_m^{\neg i} + \delta_{m,k}}}, \\
\text{and } \overline{n}_k^{\neg i} &= \sum_{i' \neq i} q(y_{i'} = k) = \overline{n}_k - q(y_i = k).
\end{aligned}$$

Note that the computation of $\gamma_k^{(i)}$ for each i requires the complexity of $O(n)$. If we approximate it by ignoring the dependency of $\nu_{k,l}$ on $\mathbf{y}$, $\gamma_k^{(i)}$ in (19) is replaced by

$$\gamma_k^{(i)} = -\frac{\|\tilde{\mathbf{x}}_i - \overline{\mathbf{m}}_k\|^2}{2}, \tag{20}$$

where $\tilde{\mathbf{x}}_i$ is the ith row vector of the data matrix and $\overline{\mathbf{m}}_k$ is the d-dimensional vector consisting of $s_j \left\langle \nu_{k,z_j} \right\rangle_{q(z_j)}$, which can be interpreted as the expected cluster centroid vector of the kth sample cluster.

Parameter estimation: The variational free energy depends on the block mean parameter $\xi_{k,l}$ through

$$-\kappa \sum_{j=1}^{d} \sum_{i=1}^{n} \sum_{l=1}^{L} \sum_{k=1}^{K} q(z_j = l) q(y_i = k) \left\langle \nu_{k,l} \right\rangle_{q(\mathbf{y})} s_j x_{ij},$$

if we approximate the expectation with respect to $q(\mathbf{y})$ by separately applying it for $\delta_{y_i,k}$ and $\nu_{k,l}$. If $\left\langle \nu_{k,l} \right\rangle_{q(\mathbf{y})}$ in the above sum is further approximated by (16), the Cauchy-Schwarz inequality implies that $\overline{F}$ is minimized by

$$\xi_{k,l} = \sum_{j=1}^{d} \sum_{i=1}^{n} q(z_j = l) q(y_i = k) s_j x_{ij}, \tag{21}$$

up to multiplication of a constant independent of k.

The update rule of the concentration parameter κ is obtained by differentiating $\overline{F}$ with respect to κ and equating it to zero, which, combined with (17) and (21), yields,

$$A_{n-1}(\kappa) \equiv \frac{C'_{n-1}(\kappa)}{C_{n-1}(\kappa)} = \frac{I_{\frac{n-1}{2}}(\kappa)}{I_{\frac{n-1}{2}-1}(\kappa)} = \frac{1}{d} \sum_{l=1}^{L} \overline{R}_l.$$

We use the method proposed in [17] to solve this equation for κ. Moreover, if we introduce a separate concentration parameter κ_l to the lth feature cluster, the above equation is replaced by

$$A_{n-1}(\kappa_l) = \frac{\overline{R}_l}{\overline{d}_l},$$

for $l = 1, ..., L$.

The hyperparameters α and β of the Dirichlet priors can also be estimated by using the Newton-Raphson updates as shown in [6].

In the limit, $\kappa \to \infty$, the probabilistic (soft) cluster assignments given in (15) and (18) become deterministic (hard), and the algorithm reduces to the asymmetric biclustering method proposed in [3], which maximizes the objective function, $\sum_{j=1}^{d} s_j \mathbf{x}_j \cdot \boldsymbol{\mu}_{z_j}$ where $\boldsymbol{\mu}_l = (\nu_{y_1,l}, ..., \nu_{y_n,l})$.

4. Experiments

In this section, we present experimental results of the proposed approach with synthetic and real-world datasets. We used the hyperparameters $\alpha = \beta = 1$ so that the prior distributions (7) and (8) are uniform. We ran five trials of the variational Bayesian algorithm and took the result with the smallest variational free energy. We observed that the variational free energy monotonically decreases in spite of the approximations introduced in Section 3.1 except for the rough approximation given by (20), which we did not use in the experiments below.

4.1. IRIS data

As a low-dimensional real dataset for demonstration, we applied the variational Bayesian inference algorithm to the *Iris* dataset, which has often been used in pattern recognition [18]. The data matrix consists of $150(=n)$ samples and $4(=d)$ features. We observed that the concentration parameter grows unboundedly, if we estimate it for fixed K and L that we considered. Although this means that the deterministic (non-probabilistic) asymmetric biclustering is favored, the model selection mechanism is lost in this limit, while if we set $L = 4$, one of the feature clusters is automatically eliminated during the estimation as is shown in Figure 1 for a different setting of κ.

We fixed $\kappa \in \{100, 150, 200, 250, 300\}$ and examined if the model selection mechanism works. Fixing $L = 4$, we compared the variational free energy for $K \in \{3, 4, \cdots, 8\}$ (Table 1). We can see that as κ grows, the more complex model is favored. This means that κ works as a regularization parameter. If its value is fixed, the numbers of clusters K and L can be determined by the minimization of the vatiational free energy. In addition, because of the prior distributions (7) and (8), redundant components are automatically eliminated if $\overline{n}_k/n$ or $\overline{d}_l/d$ is estimated near zero, or two cluster centers are estimated to be overlapping [14]. In fact, for all the results in Table 1, at least one of the four feature clusters was eliminated and the feature cluster consisting of the two features, *petal length* and *petal width* was obtained as illustrated in Figure 1 for $\kappa = 200$ and $K = 3$ with the clustered parallel coordinate plot proposed in [3], where sample cluster assignments are indicated by colors of polylines and feature clusters are divided by thick separators.

Table 1. Variational free energy (negated) for different number K of sample clusters and the concentration parameter κ. The number of sample clusters with the smallest variational free energy is highlighted for each column.

	$\kappa = 100$	$\kappa = 150$	$\kappa = 200$	$\kappa = 250$	$\kappa = 300$
$K = 3$	**719.028**	**766.249**	**801.694**	827.433	845.829
$K = 4$	717.768	764.923	800.353	831.528	853.828
$K = 5$	717.887	763.457	799.726	**832.243**	856.360
$K = 6$	716.763	757.156	798.302	830.458	**857.001**
$K = 7$	715.442	755.101	796.768	829.424	853.641
$K = 8$	714.556	755.072	795.784	828.766	852.105

In this example, the growth of κ was observed by the minimization of the variational free energy, which is equivalent to the maximum likelihood estimation of κ. A finite κ can be estimated by introducing a prior distribution and incorporating Bayesian inference for it [13]. Also in such Bayesian inference of κ, the model selection results by variational free energy minimization or automatic elimination of redundant components depend on the choice of the prior distribution of κ.

4.2. Synthetic data

We demonstrate our experimental study on a 750 records of 12-dimensional synthetic data employed in [1, 3]. This multivariate data includes four 3-dimensional clusters with 10% noise and two 6-dimensional clusters without noise, while the data samples are uniformly distributed in other dimensions.

The estimation of the concentration parameter chose the deterministic biclustering ($\kappa \to \infty$) also for this dataset. Hence, we fixed $\kappa \in \{200, 225, 250, 275, 300\}$, and compared the variational free energy for $K \in \{3, \cdots, 8\}$ and $L \in \{4, \cdots, 9\}$. Table 2 shows the selected pairs of the numbers of clusters, (K, L), that attained the minimum variational free energy for different κ.

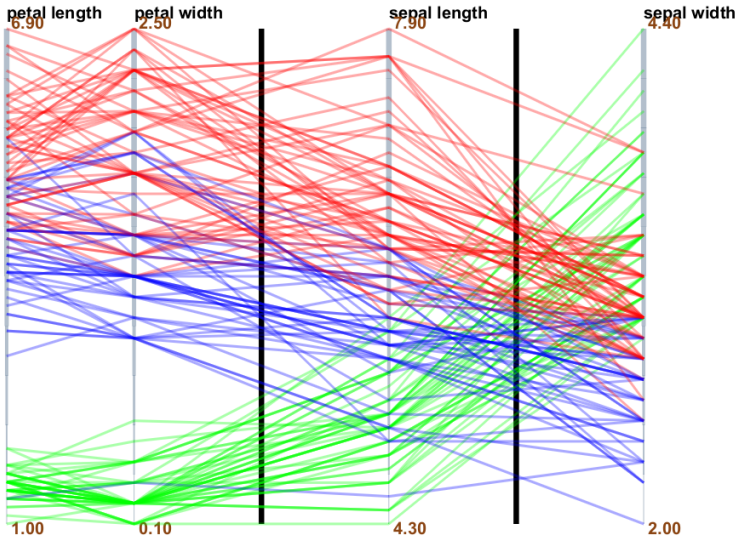

Figure 1. Clustered parallel coordinate plot for $K = 3$ and $L = 4$.

Table 2. Selected numbers of sample and feature clusters

	$\kappa = 200$	$\kappa = 225$	$\kappa = 250$	$\kappa = 275$	$\kappa = 300$
(K, L)	$(4, 3)$	$(4, 5)$	$(6, 8)$	$(5, 8)$	$(6, 7)$

We see the tendency that the greater κ is, the more complex model is selected if the variational free energy is used as the model selection criterion. Figure 2 (left) shows the variational free energy for pairs (K, L), with its sign flipped for presentation purpose, when $\kappa = 250$. Its minimum is attained at $(K = 6, L = 8)$ and $(K = 6, L = 9)$. The result of biclustering for $K = 6$ and $L = 8$ is demonstrated in Figure 2 (right) with the clustered parallel coordinate plot. Although we set $L = 8$, one of the feature clusters is eliminated, and there are seven feature clusters, one cluster with six features and six clusters with each of the other features.

In the clustered parallel coordinate plot (Figure 2 (right)), we can find the three feature clusters with each of the three uniformly distributed dimensions, which can be detected as less correlated clusters by looking at the error of each block defined by

$$\frac{1}{\overline{n}_k \overline{d}_l} \sum_{i=1}^{n} \sum_{j=1}^{d} q(y_i = k) q(z_j = l) \left(x_{ij} - s_j \nu_{k,l}\right)^2 ,$$

for the klth block. Feature or sample clusters with high block errors can be considered mismatched with the block model. The interactive visual data analysis framework developed in [3] enables to keep such clusters aside and continue the analysis for the remaining data matrix.

4.3. Supernova data

We employed the Berkeley supernova dataset [19] as another real-world example, to which the deterministic asymmetric biclustering has been applied in a previous work [20]. The dataset consists of 132 data samples corresponding to supernovae of a subtype, called "Type Ia", and 14 variables including both photometric and spectroscopic features.

We observed the tendency that if the concentration parameter κ is fixed, the more complex model is favored as κ grows also for this dataset. We also found that κ does not grow

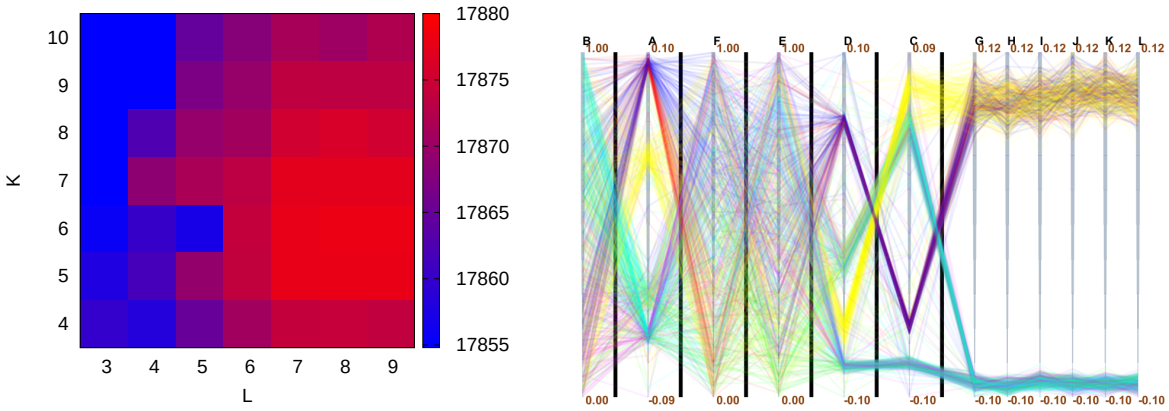

Figure 2. Variational free energy (negated) for different numbers of sample and feature clusters (left). Clustered parallel coordinate plot for $K = 6$ and $L = 8$ (right).

unboundedly if it is estimated for moderate numbers of clusters. Thus, we introduced separate concentration parameters to respective feature clusters and estimated them by the update rule given in Section 3.1. We set the numbers of clusters to $K = L = 6$, from which a detailed interactive analysis of this dataset was conducted in the previous work [20]. Figure 3(left) shows the result with the contracted parallel coordinate plot where axes of each feature cluster are contracted to a single axis by the linear discriminant analysis using sample cluster assignments as class labels [3]. We focus on the feature clusters, {Si6V},{Si6EW, Si6A, Si6FWHM}, {Si5EW, Si5DEW, x1, MBc}, which are consistent with the previous analysis except for the fact that MBc is included in the last cluster. We also observe that the scatter plot between the two feature clusters, {Si6EW, Si6A, Si6FWHM}, {Si5EW, Si5DEW, x1, MBc} appears similar to that discussed in the previous work (Figure 3(right)). Additionally, the feature x1 is negatively correlated with the other three features, Si5EW, Si5DEW, MBc, and included in the cluster with its sign flipped. This suggests that the variables $\mathbf{s}$ in the model (1) are also estimated appropriately.

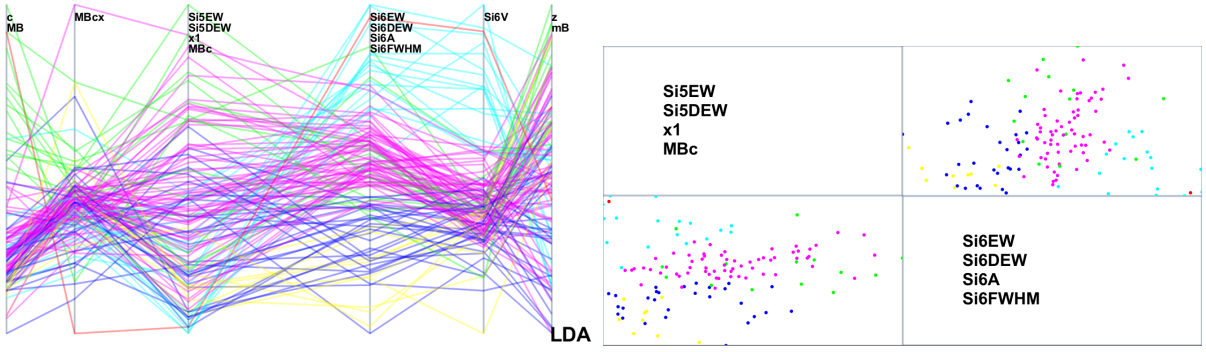

Figure 3. Contracted parallel coordinate plot of the supernova dataset (left). Scatter plot of the third and fourth feature clusters on the left, {Si5EW, Si5DEW, x1, MBc} and {Si6EW, Si6A, Si6FWHM} (right).

5. Discussion and conclusion

In this paper, we derived the variational Bayesian inference algorithm for the mixture of the constrained vMF distributions, which provides a probabilistic model of asymmetric biclustering.

The derived algorithm was applied to the multivariate data visualization tool with enhanced parallel coordinate plots. Through numerical experiments, we demonstrated that the Bayesian formulation offers a selection criterion of the number of feature and sample clusters.

The assumption of the block structure may be too restrictive for some datasets. Our current framework enables to interactively select clustered rows and columns of the data matrix and set aside them from the main target of the analysis. It would be important to model more complex structures directly, such as the nested partitioning proposed for the ordinary biclustering method in [7].

Another type of parallel coordinate plots chooses the ordering, locations and scales of axes so that data samples are aligned as horizontally as possible [21]. It is an important undertaking to explore an extension of our visualization framework in such a direction, in particular, by incorporating probabilistic cluster assignments of data samples and features.

Acknowledgments

This work has been partially supported by MEXT KAKENHI under Grants-in-Aid for Scientific Research on Innovative Areas No. 25120014. We would like to thank Professor Makoto Uemura, Hiroshima Astrophysical Science Center, Hiroshima University for his valuable comments on the analysis of the supernova dataset.

References

[1] Tatu A, Maaß F, Farber I, Bertini E, Schreck T, Seidl T and Keim D 2012 *Proc. IEEE Conference on Visual Analytics Science and Technology* pp 63–72
[2] Yuan X, Ren D, Wang Z and Guo C 2013 *IEEE Transactions on Visualization and Computer Graphics* **19** 2625–2633
[3] Watanabe K, Wu H-Y, Niibe Y, Takahashi S and Fujishiro I 2015 *Proc. 2015 IEEE Pacific Visualization Symp.* pp 287–294
[4] Hartigan J A 1972 *Journal of the American Statistical Association* **67** 123–129
[5] Mechelen I V, Bock H H and Boeck P D 2004 *Statistical methods in medical research* **13** 363–394
[6] Shan H and Banerjee A 2008 *Proc. 2008 IEEE International Conference on Data Mining* pp 530–539
[7] Roy D M and Teh Y W 2009 *Advances in Neural Information Processing Systems*
[8] Peng W, Ward M O and Rundensteiner E A 2004 *Proc. IEEE Conference on Information Visualization* pp 89–96
[9] Nohno K, Wu H-Y, Watanabe K, Takahashi S and Fujishiro I 2014 *Proc. International Conference on Information Visualisation* pp 7–12
[10] Dhillon I S and Modha D S 2001 *Machine Learning* **42** 143–175
[11] Banerjee A, Dhillon I S, Ghosh J and Sra S 2005 *Journal of Machine Learning Research* **6** 1345–1382
[12] Tanabe A, Fukumizu K, Oba S and Ishii S 2004 *Proc. 2004 Workshop on Information-Based Induction Sciences* pp 46–51 (in Japanese)
[13] Gopal S and Yang Y 2014 *Proc. International Conference on Machine Learning* pp 154–162
[14] Attias H 1999 *Proc. 15th Conference on Uncertainty in Artificial Intelligence* pp 21–30
[15] Lawrence N D 2012 *Journal of Machine Learning Research* **13** 1609–1638
[16] Kemp C and Tenenbaum J B 2008 *Proc. National Academy of Sciences* **105** 10687–10692
[17] Sra S 2012 *Computational Statistics* **27** 177–190
[18] Asuncion A and Newman D J 2007 UCI Machine Learning Repository
[19] Silverman J M, Foley R J, Filippenko A V, Ganeshalingam M *et al* 2012 *Monthly Notices of the Royal Astronomical Society* **425** 1789–1818
[20] Uemura M, Kawabata K S, Ikeda S, Maeda K, Wu H-Y, Watanabe K, Takahashi S and Fujishiro I 2016 Data-driven approach to Type Ia supernovae: variable selection on the peak luminosity and clustering in visual analytics *Journal of Physics: Conference Series* in press
[21] Kumasaka N and Shibata R 2008 *Computational Statistics & Data Analysis* **52** 3616–3644