Von Mises-Fisher Mean Shift for Clustering on a Hypersphere

Takumi Kobayashi
Information Technology Research Institute
AIST
1-1-1 Umezono, Tsukuba, Japan
Email: takumi.kobayashi@aist.go.jp

Nobuyuki Otsu
Fellow
AIST
1-1-1 Umezono, Tsukuba, Japan
Email: otsu.n@aist.go.jp

***Abstract*—We propose a method of clustering sample vectors on a hypersphere. Sample vectors are normalized in many cases, especially when applying kernel functions, and thus lie on a (unit) hypersphere. Considering the constraint of the hypersphere, the proposed method utilizes the von Mises-Fisher distribution in the framework of mean shift. It is also extended to the kernel-based clustering method via kernel tricks to cope with complex distributions. The algorithms of the proposed methods are based on simple matrix calculations. In the experiments, including a practical motion clustering task, the proposed methods produce favorable clustering results.**

***Keywords*-clustering, hypersphere, mean shift, von Mises-Fisher distribution**

I. Introduction

Clustering sample vectors is a fundamental procedure for pattern cognition and is employed in various applications, such as data mining, image segmentation and pattern classification. For the clustering, we empirically estimate the distribution of samples without any supervision, which belongs to the framework of unsupervised learning. One of the most popular methods is the k-means clustering method [1]. The k-means method, however, requires the number of clusters *a priori* based on the users' experienced knowledge. On the other hand, mean shift [2] is also widely employed, and its derivatives, such as medoid shift [3] and quick shift [4], have been proposed in recent years. The method of mean shift operates without prior knowledge about the number of clusters, and performs clustering of the sample vectors by seeking local maxima of the probability distribution defined in Euclidean space. The points of the local maxima are considered to be the centers of the clusters (modes).

In actual pattern recognition problems, the sample vector is composed of features extracted from input data, e.g., still images and motion images, and they are often normalized in L2 norm in order to increase the robustness to various changes, such as illumination changes. For example, histogram-based features, such as SIFT [5], are ordinarily normalized; in particular, most of the kernel functions used in kernel-based methods implicitly assume that sample vectors have unit norm in the kernel feature space. In these cases, the sample vectors span not the whole feature space, rather a restricted region, i.e., a unit hypersphere. Banerjee et al. [6] applied the EM method to cluster the samples on the hypersphere under the assumption of a mixture distribution, although the method based on EM requires the number of the clusters as in the k-means method.

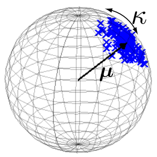


Figure 1. von Mises-Fisher distribution on a three-dimensional sphere.

In this paper, we propose a method to cluster sample vectors on the hypersphere without requiring the number of clusters. We explicitly exploit the structure of the hypersphere by employing the von Mises-Fisher distribution [7] defined on the sphere, and then follow the scheme of mean shift. The method is also extended to the kernel-based clustering method via kernel tricks.

II. Distribution on the hypersphere

On the hypersphere, sample vectors are characterized by their directions (angles). In a discipline of directional statistics, the von Mises-Fisher (vMF) distribution [7] (Fig. 1) is commonly used, and the probability density function is defined as follows:

$$\mathcal{M}(\boldsymbol{y};\boldsymbol{\mu},\kappa)=C_{\mathcal{M}}(\kappa)\exp(\kappa\boldsymbol{y}'\boldsymbol{\mu}), \tag{1}$$

where $\boldsymbol{y}$ is a d-dimensional unit vector ($\boldsymbol{y}\in\boldsymbol{R}^d$, $||\boldsymbol{y}||=1$), $\boldsymbol{\mu}$ is a unit vector orienting the center of the distribution, κ is a parameter to control the concentration of the distribution to the vector $\boldsymbol{\mu}$, and $C_{\mathcal{M}}(\kappa)$ is a normalization constant.

The vMF distribution is based on the monotonically increasing convex function exp and the inner product between $\boldsymbol{y}$ and $\boldsymbol{\mu}$. According to this concept, we can generalize the vMF distribution defined on the hypersphere as follows:

$$\mathcal{F}(\boldsymbol{y};\boldsymbol{\mu},\kappa)=C_{\mathcal{F}}(\kappa)f(\boldsymbol{y}'\boldsymbol{\mu};\kappa), \tag{2}$$

where f is a monotonically increasing convex function called profile function with a parameter κ. In this formu-

1051-4651/10 $26.00 © 2010 IEEE
DOI 10.1109/ICPR.2010.522

Algorithm 1 : vMF Mean Shift

```
[Normalization] $||\boldsymbol{x}_i|| = 1, \forall i.$
    for $i = 1$ to $N$ do
        [Init] Start from the $i$-th sample: $\boldsymbol{y}_i^{(0)} = \boldsymbol{x}_i$.
        [Update] Update the target points (vectors) until convergence by the following formula:
            $$\boldsymbol{y}_i^{(t+1)} \leftarrow \frac{\sum_j^N \boldsymbol{x}_j g(\boldsymbol{x}_j' \boldsymbol{y}_i^{(t)}; \kappa)}{||\sum_j^N \boldsymbol{x}_j g(\boldsymbol{x}_j' \boldsymbol{y}_i^{(t)}; \kappa)||}. \quad (6)$$
    end for
[Postprocessing] Merge close convergent points ($\boldsymbol{y}_i^{(\infty)'} \boldsymbol{y}_j^{(\infty)} > 1 - \epsilon$) into the same cluster.
```

lation, the vMF is defined as $f(x; \kappa) \equiv \exp(\kappa x)$, and we call the distribution in Eq.(2) a vMF distribution.

III. vMF MEAN SHIFT

Using the above vMF distribution, we propose a method of mean shift on the sphere (*vMF mean shift*) by imposing the constraint that all samples are laid on the unit sphere.

A. Linear method

The empirical probability distribution can be estimated by applying the Parzen window method [1] to the given samples. We employ the vMF distribution in Eq.(2) for the window function at each sample vector $\boldsymbol{x}_i$ ($||\boldsymbol{x}_i|| = 1$), and thus the probability distribution is obtained by

$$p(\boldsymbol{y}) = \frac{1}{N} \sum_i^N \mathcal{F}(\boldsymbol{y}; \boldsymbol{x}_i, \kappa) = \frac{C_\mathcal{F}}{N} \sum_i^N f(\boldsymbol{y}' \boldsymbol{x}_i; \kappa), \quad (3)$$

where $||\boldsymbol{y}|| = 1$ and N is the number of samples. Local maxima $\boldsymbol{y}^*$ of this probability distribution p is calculated via derivatives of the objective function with the Lagrange multiplier η for imposing the constraint $||\boldsymbol{y}|| = 1$:

$$\text{Objective function: } L = p(\boldsymbol{y}) - \frac{\eta}{2}(||\boldsymbol{y}||^2 - 1), \quad (4)$$

$$\boldsymbol{y}^* = \frac{1}{\eta N} \sum_i^N \boldsymbol{x}_i \mathcal{G}(\boldsymbol{y}; \boldsymbol{x}_i, \kappa) = \frac{\sum_i^N \boldsymbol{x}_i g(\boldsymbol{y}' \boldsymbol{x}_i; \kappa)}{||\sum_i^N \boldsymbol{x}_i g(\boldsymbol{y}' \boldsymbol{x}_i; \kappa)||}, \quad (5)$$

where $\mathcal{G}$, g are derivatives of $\mathcal{F}$,f, respectively. Thus, the algorithm of vMF mean shift is given in Algorithm 1. Note that κ is a single parameter of the method.

B. Kernel-based method

In the previous section, the hypersphere that the sample vectors lie on is embedded in Euclidean space. On the other hand, because the proposed method is mainly based on inner products between sample vectors, we can extend the vMF mean shift to the kernel-based clustering method via kernel tricks.

Algorithm 2 : Kernel vMF Mean Shift

```
[Normalization] $\boldsymbol{K} = diag(\boldsymbol{K})^{-1/2} \boldsymbol{K} diag(\boldsymbol{K})^{-1/2}$
[Init] Start from each sample: $\boldsymbol{A}^{(0)} = \boldsymbol{I}$, where $\boldsymbol{I}$ is an identity matrix.
[Update] Update the target points (coefficients) until convergence by the following formula:
    $$\boldsymbol{W}^{(t)} \leftarrow g_M(\boldsymbol{K}\boldsymbol{A}^{(t)}; \kappa) \quad (7)$$
    $$\boldsymbol{A}^{(t+1)} \leftarrow \boldsymbol{W}^{(t)} diag(\boldsymbol{W}^{t'} \boldsymbol{K} \boldsymbol{W}^{(t)})^{-1/2} \quad (8)$$
[Postprocessing] Merge close convergent points ($\boldsymbol{a}_i^{(\infty)'} \boldsymbol{K} \boldsymbol{a}_j^{(\infty)} > 1 - \epsilon$) into the same cluster.
```

The vMF distribution in Eq.(2) can also be defined in the kernel feature space by simply replacing inner products with kernel functions:

$$\mathcal{F}(\boldsymbol{\phi}_y; \boldsymbol{\phi}_\mu, \kappa) = C_\mathcal{F} f(\boldsymbol{\phi}_y' \boldsymbol{\phi}_\mu; \kappa) = C_\mathcal{F} f(k(\boldsymbol{y}, \boldsymbol{\mu}); \kappa),$$

where $\boldsymbol{\phi}_y, \boldsymbol{\phi}_\mu$ are normalized vectors in the kernel feature space, and thus k is a normalized kernel function: $k(\boldsymbol{x}, \boldsymbol{y}) \leftarrow \frac{k(\boldsymbol{x}, \boldsymbol{y})}{\sqrt{k(\boldsymbol{x}, \boldsymbol{x}) k(\boldsymbol{y}, \boldsymbol{y})}}$. The frequently used Gaussian kernel is inherently normalized, and the normalized kernel function has also produced better performance in recognition problems [8].

In addition, the updated vector is described as a linear combination of sample vectors in Eq.(6) as in the representer theorem [9]. Thus, the updated vector for the i-th sample in the kernel feature space is represented by

$$\boldsymbol{\phi}_{y_i} \propto \sum_j^N a_{ij} \boldsymbol{\phi}_{x_j} = \boldsymbol{\phi}_X \boldsymbol{a}_i, \quad (9)$$

where $\boldsymbol{\phi}_X = [\boldsymbol{\phi}_{x_1}, \cdots, \boldsymbol{\phi}_{x_N}]$ and $\boldsymbol{a}_i$ are linear coefficients. Thus, the update formula in Eq.(6) leads to

$$\begin{aligned} \boldsymbol{\phi}_X \boldsymbol{a}_i^{(t+1)} &\leftarrow \frac{\sum_j^N \boldsymbol{\phi}_{x_j} g(\boldsymbol{\phi}_{x_j}' \boldsymbol{\phi}_X \boldsymbol{a}_i^{(t)}; \kappa)}{||\sum_j^N \boldsymbol{\phi}_{x_j} g(\boldsymbol{\phi}_{x_j}' \boldsymbol{\phi}_X \boldsymbol{a}_i^{(t)}; \kappa)||} \\ &= \frac{\sum_j^N \boldsymbol{\phi}_{x_j} g(\boldsymbol{k}_j' \boldsymbol{a}_i^{(t)}; \kappa)}{||\sum_j^N \boldsymbol{\phi}_{x_j} g(\boldsymbol{k}_j' \boldsymbol{a}_i^{(t)}; \kappa)||} \\ &= \frac{\boldsymbol{\phi}_X g_M(\boldsymbol{K} \boldsymbol{a}_i^{(t)}; \kappa)}{\sqrt{g_M(\boldsymbol{K} \boldsymbol{a}_i^{(t)}; \kappa)' \boldsymbol{K} g_M(\boldsymbol{K} \boldsymbol{a}_i^{(t)}; \kappa)}} \quad (10) \end{aligned}$$

$$\therefore \boldsymbol{a}_i^{(t+1)} \leftarrow \frac{\boldsymbol{w}_i^{(t)}}{\sqrt{\boldsymbol{w}_i^{(t)'} \boldsymbol{K} \boldsymbol{w}_i^{(t)}}}, \; \boldsymbol{w}_i^{(t)} = g_M(\boldsymbol{K} \boldsymbol{a}_i^{(t)}; \kappa), \quad (11)$$

where $\boldsymbol{k}_j$ is the j-th column vector of the (normalized) kernel Gram matrix $\boldsymbol{K}$, and g_M denotes the matrix function applying g to all elements of the matrix, $\{g_M(\boldsymbol{A})\}_{ij} = g(A_{ij})$. The algorithm of the kernel vMF mean shift can be simply described by using matrix calculations, as shown in Algorithm 2.

C. Discussion

We prove the convergence of the kernel vMF mean shift, which can also be applied to that of the (linear) vMF mean shift. Since the probability distribution function p is upper-bounded, what we prove is only that p is monotonically increasing by the updates:

$$\begin{aligned}
&p(\boldsymbol{\phi}_X\boldsymbol{a}^{(t+1)}) - p(\boldsymbol{\phi}_X\boldsymbol{a}^{(t)})\\
&= \frac{C_{\mathcal{F}}}{N}\sum_i^N f(\boldsymbol{k}_i'\boldsymbol{a}^{(t+1)};\kappa) - f(\boldsymbol{k}_i'\boldsymbol{a}^{(t)};\kappa)\\
&\geq \frac{C_{\mathcal{F}}}{N}\sum_i^N \{\boldsymbol{k}_i'(\boldsymbol{a}^{(t+1)} - \boldsymbol{a}^{(t)})\} g(\boldsymbol{k}_i'\boldsymbol{a}^{(t)};\kappa) \qquad (12)\\
&= \frac{C_{\mathcal{F}}}{N}(\boldsymbol{a}^{(t+1)} - \boldsymbol{a}^{(t)})'\boldsymbol{K} g_M(\boldsymbol{K}\boldsymbol{a}^{(t)};\kappa)\\
&= \frac{C_{\mathcal{F}}\sqrt{\boldsymbol{w}^{(t)'}\boldsymbol{K}\boldsymbol{w}^{(t)}}}{N}(\boldsymbol{a}^{(t+1)} - \boldsymbol{a}^{(t)})'\boldsymbol{K}\boldsymbol{a}^{(t+1)}\\
&= \frac{C_{\mathcal{F}}\sqrt{\boldsymbol{w}^{(t)'}\boldsymbol{K}\boldsymbol{w}^{(t)}}}{2N}(\boldsymbol{a}^{(t+1)} - \boldsymbol{a}^{(t)})'\boldsymbol{K}(\boldsymbol{a}^{(t+1)} - \boldsymbol{a}^{(t)})\\
&\geq 0, \qquad (13)
\end{aligned}$$

where we use convexity of f in Eq.(12), and positive definiteness of the kernel Gram matrix $\boldsymbol{K}$ in Eq.(13).

In the kernel vMF mean shift, when the update is converged, the following equation holds:

$$g_M(\boldsymbol{K}\boldsymbol{a}^{(\infty)};\kappa) = \lambda\boldsymbol{a}^{(\infty)}, \qquad (14)$$

where $\lambda = \sqrt{g_M(\boldsymbol{K}\boldsymbol{a}^{(\infty)};\kappa)'\boldsymbol{K} g_M(\boldsymbol{K}\boldsymbol{a}^{(\infty)};\kappa)}$ and $\boldsymbol{a}^{(\infty)}$ is a fixed point. This is similar to an eigenvalue problem of the kernel Gram matrix, $\boldsymbol{K}\boldsymbol{a} = \lambda\boldsymbol{a}$, in which the eigenvectors are considered as fixed points of the matrix $\boldsymbol{K}$. Therefore, the method of the kernel vMF mean shift is a sort of nonlinear extension of spectral clustering [10] which also finds the eigenvectors (fixed points) of the kernel Gram matrix (similarity matrix).

IV. Experimental results

We apply the proposed methods to several clustering problems. In the experiments, the profile function f and its derivative g are set as

$$f(x;\kappa) = \begin{cases} 0 & 0 \leq x \leq \kappa \\ \frac{1}{2}(x-\kappa)^2 & \kappa \leq x \leq 1 \end{cases}, \qquad (15)$$

$$g(x;\kappa) = \begin{cases} 0 & 0 \leq x \leq \kappa \\ x-\kappa & \kappa \leq x \leq 1 \end{cases}. \qquad (16)$$

The profile f is smooth and thus the derivative g is continuous.

The proposed methods have a single parameter κ that controls the concentration of vMF distribution. In this study, the parameter value is determined based on directional statistics of sample vectors, similarly to the standard deviation of a Gaussian distribution:

$$\kappa = \cos\left\{\frac{1}{2}\boldsymbol{E}_{ij}\cos^{-1}(k(\boldsymbol{x}_i,\boldsymbol{x}_j))\right\}. \qquad (17)$$

We focus on the angle between sample vectors, which corresponds to the metric on the hypersphere.

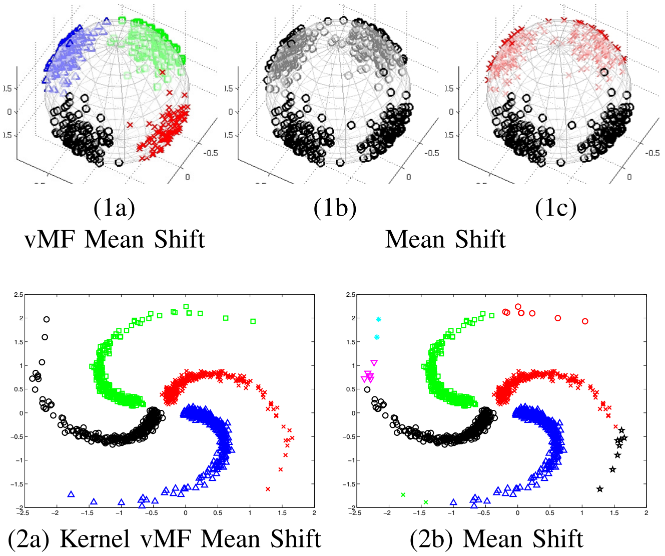
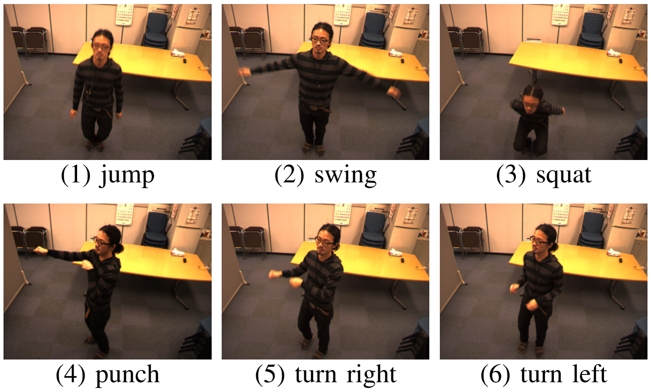

(1a) (1b) (1c)
vMF Mean Shift Mean Shift
(2a) Kernel vMF Mean Shift (2b) Mean Shift

Figure 2. Two types of toy examples.

A. Toy examples

The first task is to cluster sample vectors that simply form four modes on a 3D unit sphere (Fig. 2(1)). The proposed method extracts the correct mode structure as shown in Fig. 2(1a). In the standard mean shift, the parameter value greatly affects the clustering results; the parameter value derived from the standard deviation of the samples causes a single cluster (Fig. 2(1b)), and that derived from Eq.(17) even leads to the incorrect result (Fig. 2(1c)). To obtain the same result as in Fig. 2(1a), we have to carefully determine the parameter in the mean shift. The vMF mean shift, which exploits the constraint for the hypersphere, produces favorable results by using the statistical parameter value in Eq.(17).

Next, we apply the kernel vMF mean shift to cluster sample vectors on a complex distribution consisting of four spirals (Fig. 2(2)). We employ the Gaussian kernel for geodesic distance [11] and, by using such kernel functions, the kernel vMF mean shift can deal with complex distributions. As shown in Fig. 2(2a), the proposed method retrieves the four correct modes. The mean shift, however, loses some of the mode structures in this complex distribution (Fig. 2(2b)).



(1) jump (2) swing (3) squat

(4) punch (5) turn right (6) turn left

Figure 3. Examples of motions.

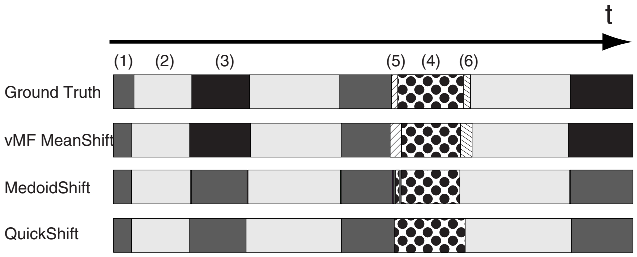


Figure 4. Results of motion clustering.

B. Motion clustering

We can apply linear vMF mean shift to a more practical problem such as the clustering of human motions in motion images. The motion images consist of 1100 frames containing four basic cyclic motions: *jump, swing, squat* and *punch*, and two transient motions: *turn left* and *turn right*, as shown in Fig. 3. We extract motion features at every time point (frame) by applying the CHLAC method [12] with the time window of 60 frames. The feature vectors are then normalized (to unit in L2-norm) to increase robustness to environmental variations, such as illumination changes. Based on the human motion features, frames in motion images are clustered into segments along the time axis.

For comparison, quick shift [4] and medoid shift [3] are applied with the parameter values that are statistically determined in a manner similar to Eq.(17). The clustering results are shown in Fig. 4. The proposed method successfully detects all motion clusters, not only four basic motions but also two short transient motions, whereas both the quick shift and medoid shift confuse the *jump* and *squat* motions and miss the two transient motions. Note that, in this experiment, the profile function and definition of κ (Eq.(17)) in the vMF mean shift are the same as those in Sec.IV-A.

V. Conclusion

We have proposed a method for clustering sample vectors on a hypersphere. The method incorporates the vMF distribution on the hypersphere into the framework of mean shift. We also extended the method to the kernel-based clustering method by utilizing kernel tricks. The unit-norm constraint that the proposed methods assume is frequently imposed in feature vectors and kernel functions. We applied the methods to several clustering problems including motion clustering, and the proposed methods exhibited the favorable results.

References

[1] R. Duda, P. Hart, and D. Stork, *Pattern Classification*. Wiley-Interscience, 2000.

[2] D. Comaniciu and P. Meer, “Mean shift: A robust approach toward feature space analysis,” *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 24, no. 5, pp. 603–619, 2002.

[3] Y. Sheikh, E. Khan, and T. Kanade, “Mode-seeking by medoidshifts,” in *ICCV*, 2007.

[4] A. Vedaldi and S. Soatto, “Quick shift and kernel methods for mode seeking,” in *ECCV*, 2008, pp. 705–718.

[5] D. Lowe, “Distinctive image features from scale invariant features,” *International Journal of Computer Vision*, vol. 60, pp. 91–110, 2004.

[6] A. Banerjee, I. S. Dhillon, J. Ghosh, and S. Sra, “Clustering on the unit hypersphere using von mises-fisher distributions,” *Journal of Machine Learning Research*, vol. 6, no. 12, pp. 1345–1382, 2005.

[7] K.V.Mardia and P. Jupp, *Directional Statistics (2nd edition)*. John Wiley and Sons Ltd., 2000.

[8] K. Hotta, “Non-linear feature extraction by linear pca using local kernel,” in *ICPR*, 2008.

[9] B. Scholkopf and A. Smola, *Learning with Kernels*. MIT Press, 2001.

[10] A. Ng, M. Jordan, and Y. Weiss, “On spectral clustering: Analysis and an algorithms,” *Advances in Neural Information Processing Systems*, vol. 14, 2001.

[11] J. B. Tenenbaum, V. Silva, and J. C. Langford, “A global geometric framework for nonlinear dimensionality reduction,” *Science*, vol. 290, no. 5500, pp. 2319–2323, 2000.

[12] T. Kobayashi and N. Otsu, “A three-way auto-correlation based approach to motion recognition,” *Pattern Recognition Letters*, vol. 30, no. 3, pp. 185–192, 2009.