Behavior Analysis of Volume Prototypes in High Dimensionality

Maiko Sato, Mineichi Kudo, and Jun Toyama

Division of Computer Science
Graduate School of Information Science and Technology
Hokkaido University
Kita-14, Nishi-9, Kita-ku, Sapporo 060-0814, Japan
{maiko,mine,jun}@main.ist.hokudai.ac.jp
<http://prml.main.ist.hokudai.ac.jp>

Abstract. In these years, we often deal with an enormous amount of data or a data stream in a large variety of pattern recognition tasks. As a promising approach for economising the amount, we have previously defined a *volume prototype* as a geometric configuration that represents some data points inside and proposed a single-pass algorithm for finding them. In this paper, we analyze the convergence behavior of volume prototypes in high-dimensional cases. In addition, we show the applicability of volume prototypes to high-dimensional classification problems.

1 Introduction

In these years, we often deal with an enormous amount of data or a data stream in a large variety of pattern recognition tasks. Such data require a huge amount of memory space and computation time for processing. So it is preferable to use a more simple expression of data without losing characteristics of original datasets. We have proposed *volume prototypes* [1] as an extension of conventional point prototypes and defined each as a geometric configuration that represents many data points inside.

Volume prototypes can be generated by a single-pass algorithm, and thus they are significantly effective for streaming data that can be accessed only once. As another advantage, volume prototypes typically converge to modes of the underlying distribution, so that we do not need to care the number of prototypes. It suffices to take a sufficiently large number of initial prototypes, unlike mixture models.

In the initial study [1], we conducted some experiments and analyzed a typical behavior of volume prototypes in two-dimensional problems. However, it is still unclear about how they move and how they are effective in high dimensionality. In this paper, we clarify the behavior of volume prototypes in high-dimensional spaces and show the applicability of them to high-dimensional problems.

2 Related Works and Achieved Status

The prototype approaches for massive data are divided into four groups: prototype selection, density estimation by mixture models, clustering and classification. The volume prototype approach is related to all of these groups of studies, especially it is strongly related to clustering and mixture models (Table 1). However, there are clear differences from them. A major difference is that volume prototype approach is at a lower level than clustering or mixture models. That is, volume prototypes are regarded as one of intermediate expressions, more close to data, between data points and components as seen in clustering or mixture models.

Table 1. Comparison of volume prototypes, clustering and mixture models

	Volume prototypes	Clustering	Mixture models
Goal	To condense (large) data appropriately	To analyze relative locations and distributions of clusters	To estimate mechanisms of data generation
Definition	Local prototypes which have volume	Partition of data without overlapping	Sum of component distributions to estimate original distributions
Example

The name *volume prototype* is not new. Kaymak and Setnes [2] already used the terminology, but their volume prototypes are clusters with a boundary. On the contrary, our volume prototypes are closer to data themselves.

Fig. 1. For one two-dimensional data (with five modes), volume prototypes with squared radius $r^2 = \chi^2_2(\theta)$. Here 10^5 artificial data were compressed into 100 or less volume prototypes.

An example of volume prototypes is shown in Fig. 1 for a two-dimensional data. From Fig. 1, we see that 1) the volume prototypes converge to the modes or ridge of distributions, 2) they approximate the distribution from inside if $\theta \leq 0.9$, and 3) only a few volume prototypes can summarize a massive data.

3 Algorithm

In this section, we show the concrete algorithm VP to generate volume prototypes. The single-pass algorithm is shown in Fig. 2.

Inputs:(Unlimited samples) $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N, \dots$

N is the number of samples for **ME Step**.

M is the number of seed prototypes.

N_v is the number of latest samples to be kept.

θ is the value to determine radius r of the acceptance region.

Procedure for obtaining volume prototypes

ME (mode estimation) step:

Repeat M times the following for the first N samples.

1. Choose a random permutation σ to reorder the samples to $\mathbf{x}_{\sigma_1}, \mathbf{x}_{\sigma_2}, \dots, \mathbf{x}_{\sigma_N}$.
2. Initialize a prototype by the way described in the detailed description.
3. Use $\mathbf{x}_{\sigma_2}, \dots, \mathbf{x}_{\sigma_N}$ in order to update the *seed prototype* by the same way as 2. (using only the samples in the acceptance region (1)).

C (Convergence) step:

For each sample $\mathbf{x}_i (i = N + 1, N + 2, \dots)$ do the following.

1. Update the seed prototypes in which $\mathbf{x}_i$ falls by (2)–(4).
 2. Keep the latest N_v samples.
-

Exploit time:

From M prototypes, select greedily $L (\leq M)$ prototypes until no improvement is found for covering N_v samples. Count the number of samples falling in each prototype in N_v samples. Here a sample is divided by k when k prototypes share it.

Fig. 2. Volume Prototype Algorithm (VP)

The outline is as follows. With multiple scans over the first N samples in different orders, we obtain M *seed prototypes* (ME Step). We bring up those seed prototypes by updating them using the samples fell in their acceptance regions (C Step). For *final prototypes*, we carry out prototype selection using a set covering criterion to have L final prototypes.

3.1 Details

Dataset: We consider a data stream $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N, \mathbf{x}_{N+1}, \dots$ of d -dimensional vectors.

Prototype: Let $\mathbf{p} = (\boldsymbol{\mu}, \boldsymbol{\Sigma}, r, n)$ be a prototype. Here, $\boldsymbol{\mu}$ is the prototype center, $\boldsymbol{\Sigma}$ is the covariance matrix, r is the Mahalanobis radius and n is the number of samples inside the prototype. Here, r is a significant parameter that affects the behavior of the prototype growth and is set to $\chi_d^2(\theta)$ ($100 \times \theta$ % point of chi-squared statistics with d degrees of freedom). With $\theta = 0.9$, we expect that 90% of surrounding points are covered.

Included sample set: Let $\mathbf{S}_p$ be the set of data points inside the prototype $\mathbf{p}$, which is specified by $\mathbf{S}_p = \{\mathbf{x} | (\mathbf{x} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \leq r^2\}$. Then $n = |\mathbf{S}_p|$. We approximate n by the number of samples used for updating the prototype.

Acceptance region: We specify the acceptance region $\mathbf{A}_p$ of prototype $\mathbf{p} = (\boldsymbol{\mu}, \boldsymbol{\Sigma}, r, n)$, by the acceptance Mahalanobis radius R as

$$\mathbf{A}_p = \{\mathbf{x} | (\mathbf{x} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \leq R^2\}. \quad (1)$$

Here, $R(\geq r)$ is determined from r and n by $R = r + \sqrt{\frac{\chi_d^2(0.95)}{n}}$.

Here, 0.95 means to take the 95% confidence region of the sample mean in Mahalanobis radius. It is noted that R approaches to r as n goes to infinity.

Updating procedure: When a sample $\mathbf{x}$ falls into the acceptance region $\mathbf{A}_p$ of prototype $\mathbf{p}_t$ at time $t + 1$, $\mathbf{p}_t$ is updated to $\mathbf{p}_{t+1}$ by

$$n_{t+1} = n_t + 1, \quad (2)$$

$$\hat{\boldsymbol{\mu}}_{t+1} = \frac{1}{n_{t+1}} (n_t \hat{\boldsymbol{\mu}}_t + \mathbf{x}), \quad (3)$$

$$\hat{\boldsymbol{\Sigma}}_{t+1} = \frac{1}{n_{t+1}} \left(n_t \hat{\boldsymbol{\Sigma}}_t + \mathbf{x} \mathbf{x}' + n_t \hat{\boldsymbol{\mu}}_t \hat{\boldsymbol{\mu}}_t' - n_{t+1} \hat{\boldsymbol{\mu}}_{t+1} \hat{\boldsymbol{\mu}}_{t+1}' \right). \quad (4)$$

In addition, radius R of the acceptance region is also updated.

Initialization of ME step: We initialize the center $\hat{\boldsymbol{\mu}}^{(0)}$ and the covariance matrix $\hat{\boldsymbol{\Sigma}}^{(0)}$ by $\hat{\boldsymbol{\mu}}^{(0)} = \mathbf{x}_1$ (the first element after random reordering), and $\hat{\boldsymbol{\Sigma}}^{(0)} = \lambda^2 \mathbf{I}$. Here,

$$\lambda^2 = \frac{1}{dN} \sum_{i=1}^N \min_{j \neq i, j \in \{1, 2, \dots, N\}} \|\mathbf{x}_i - \mathbf{x}_j\|^2$$

($\|\cdot\|^2$ is the squared Euclidean norm) and $\mathbf{I}$ is the unit matrix.

In addition, we set the initial value of $n^{(0)}$ to $n^{(0)} = d + 1$.

4 Behavior Analysis

Here, let us make clear the behavior of volume prototypes, especially where they move to, how fast they move and whether they shrink or not, as the number t of updtings increases.

The behavior of a volume prototype is very complicated. The center $\boldsymbol{\mu}_t$ and covariance matrix $\boldsymbol{\Sigma}_t$ are updated to $\boldsymbol{\mu}_{t+1}$ and $\boldsymbol{\Sigma}_{t+1}$ through (2) – (4). The factors $\boldsymbol{\mu}_t$, $\boldsymbol{\Sigma}_t$, r , n_t affect the updating as well as $\mathbf{X}_{t+1}$ generated from the underlying distribution f . Here let us derive the typical behavior under some strong but generally acceptable assumptions.

4.1 Movement of Prototype Centers

First, let us analyze the direction and speed of movement of a volume prototype center. Let a prototype after t th updating be $\mathbf{p}_t = (\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t, r, n_t)$. For simplicity, we assume that $\boldsymbol{\Sigma}_t = \boldsymbol{\Sigma}_0$ and $R^2 = r_\theta^2 = \chi_d^2(\theta)$ for $t = 1, 2, \dots$. Here, $n_t = n_0 + t$. Then, the acceptance region $\mathbf{A}_{p_t}$ of $\mathbf{p}_t$ is given by

$$\mathbf{A}_{p_t} = \{\mathbf{x} \mid (\mathbf{x} - \boldsymbol{\mu}_t)' \boldsymbol{\Sigma}_0^{-1} (\mathbf{x} - \boldsymbol{\mu}_t) \leq r_\theta^2\}.$$

With a sample $\mathbf{X}_{t+1} \in \mathbf{A}_{p_t}$, the center $\boldsymbol{\mu}_t$ is updated to $\boldsymbol{\mu}_{t+1}$ by

$$\begin{aligned} \boldsymbol{\mu}_{t+1} &= \frac{n_0 + t}{n_0 + t + 1} \boldsymbol{\mu}_t + \frac{1}{n_0 + t + 1} \mathbf{X}_{t+1} \\ &= \boldsymbol{\mu}_t + \frac{\mathbf{X}_{t+1} - \boldsymbol{\mu}_t}{n_0 + t + 1} = \boldsymbol{\mu}_t + \frac{\Delta \mathbf{X}_{t+1}}{n_0 + t + 1}. \end{aligned}$$

Then the expected updated center is given by

$$E_{\mathbf{X}_{t+1} \in \mathbf{A}_{p_t}} \boldsymbol{\mu}_{t+1} = \boldsymbol{\mu}_t + \frac{E_{\mathbf{X}_{t+1}} \Delta \mathbf{X}_{t+1}}{n_0 + t + 1}.$$

If $\boldsymbol{\Sigma}_0$ is small enough to allow f to be linearly approximated in $\mathbf{A}_{p_t}$ (Fig. 3) as $f(\mathbf{x}) = f(\boldsymbol{\mu}_t) + \nabla f(\boldsymbol{\mu}_t)'(\mathbf{x} - \boldsymbol{\mu}_t)$, we can have

$$E_{\mathbf{X}_{t+1}} \Delta \mathbf{X}_{t+1} = \frac{r_\theta^2}{d+2} \boldsymbol{\Sigma}_0 \frac{\nabla f(\boldsymbol{\mu}_t)}{f(\boldsymbol{\mu}_t)},$$

(for the proof, see Fukunaga [3](pp.534–535)). Therefore, we have

$$E_{\mathbf{X}_{t+1} \in \mathbf{A}_{p_t}} \boldsymbol{\mu}_{t+1} = \boldsymbol{\mu}_t + \frac{1}{n_0 + t + 1} \frac{r_\theta^2}{d+2} \boldsymbol{\Sigma}_0 \frac{\nabla f(\boldsymbol{\mu}_t)}{f(\boldsymbol{\mu}_t)}. \quad (5)$$

From Eq. (5), we see that 1) the direction of displacement is $\boldsymbol{\Sigma}_0 \nabla f(\boldsymbol{\mu}_t)$, 2) the speed $E\boldsymbol{\mu}_{t+1} - \boldsymbol{\mu}_t$ is proportional to $\frac{r_\theta^2}{d+2}$ and $\frac{\nabla f(\boldsymbol{\mu}_t)}{f(\boldsymbol{\mu}_t)}$, therefore, 3) the speed is low when $\nabla f(\boldsymbol{\mu}_t)$ is small or $f(\boldsymbol{\mu}_t)$ is large depending on the previous position $\boldsymbol{\mu}_t$, and 4) the speed reduces at ratio $1/(n_0 + t + 1)$ for the increase of t , hence, the accumulated amount is approximately $\log t$. Note that $r_\theta^2/(d+2)$ does not

Fig. 3. Movement of a prototype center

Fig. 4. $\chi_d^2(0.9)/(d+2)$ vs. dimensions

decrease linearly to $1/d$ because r_θ depends on both d and θ . Indeed, this value changes in a little strange way as shown in Fig. 4 ($\theta = 0.9$). In $\theta = 0.9$, up to around $d = 8$, the value increases, hence the movement speed of the center increases, but, after $d = 8$, the value decreases toward one, hence the speed decreases, as the dimensionality d increases.

We can obtain more intuition when $\Sigma_0 = \alpha \Sigma_*$ with a constant α (< 1), where Σ_* is the covariance matrix of the underlying normal distribution $f(\mathbf{x}) = N(\mathbf{x}; \mu_*, \Sigma_*)$. Then, for Eq. (5), we have a simpler equation:

$$\begin{aligned} E_{\mathbf{X}_{t+1}} \mu_{t+1} &= \mu_t + \frac{1}{n_0 + t + 1} \frac{r_\theta^2}{d + 2} \Sigma_0 \frac{\nabla f(\mu_t)}{f(\mu_t)} \\ &= \mu_t + \frac{1}{n_0 + t + 1} \frac{\alpha r_\theta^2}{d + 2} \Sigma_* \Sigma_*^{-1} (\mu_* - \mu_t) \quad (\text{for Gaussian } f) \\ &= \mu_t + \frac{1}{n_0 + t + 1} \frac{\alpha r_\theta^2}{d + 2} (\mu_* - \mu_t) \end{aligned} \quad (6)$$

Therefore, the expected μ_{t+1} is the point between the global mean μ_* and μ_t at ratio $\frac{1}{n_0 + t + 1} \frac{\alpha r_\theta^2}{d + 2} : (1 - \frac{1}{n_0 + t + 1} \frac{\alpha r_\theta^2}{d + 2})$ and climbs up to the steepest ascend direction.

Experiment 1

Data: $N(\mathbf{0}, I_d)$, $d = 2, 4, 8, 16, 32, 64$

Sample size: Out of 10^5 samples, the first $N = 10^3$ samples were used in ME step.

The number of prototype: $M = 10$

Radius: $R^2 = r_\theta^2 = \chi_d^2(\theta)$, $\theta = 0.85, 0.90, 0.95$

Procedure: Just after ME step, we have $\mathbf{p}_0 = (\mu_0, \Sigma_0, r, n_0)$, where n_0 depends on the prototype.

Results: The upper row of Fig. 5. Here, the horizontal axis is the logarithm of the number of samples.

From Fig. 5 (the upper row), we can confirm that the speed is linear to $\log t$ within a range in which $\frac{\nabla f(\mu_t)}{f(\mu_t)}$ is almost the same. The speed (the slope) increases up to $d = 8$ and then decreases for larger d . When θ , thus r_θ , increases, they converge more quickly. These observations show the validity of our analysis.

As a result, the centers of volume prototypes move at speed of the reciprocal of the number of samples and converge to one of local peaks of the underlying distribution, and they will stay there (because ∇f is close to zero) or approach to a more higher place (dense part) very slowly.

4.2 Convergence of Covariance Matrix

Next, let us analyze whether volume prototypes vanish or not with a sufficiently large number of samples.

Fig. 5. The behavior of prototypes for $f(\mathbf{x}) = N(\mathbf{x}; \mathbf{0}, \mathbf{I}_d)$. The distance $\|\boldsymbol{\mu}_t\|^2/d$ (the upper row) and $\text{tr} \boldsymbol{\Sigma}_t/d$ (the lower row) are plotted for increasing number of sample in log-scale in $d = 2, 4, 8, 16, 32, 64$. From left to right, $\theta = 0.85, 0.90, 0.95$.

For simplify, let us assume that $f(\mathbf{x}) = N(\mathbf{x}; \mathbf{0}, \boldsymbol{\Sigma}_*)$ and a prototype $\mathbf{p}_t = (\mathbf{0}, \boldsymbol{\Sigma}_t, r_\theta, n_t)$ as the one of which center has already converged to the origin. In addition, we assume that $\boldsymbol{\Sigma}_t = \alpha_t \boldsymbol{\Sigma}_*$, that is, the configuration is the same as $\boldsymbol{\Sigma}_*$ and only the scale is different by α_t .

Then the behavior of $\boldsymbol{\Sigma}_t$ is identical to that of α_t . Let $Y^2 = \mathbf{X}' \boldsymbol{\Sigma}_*^{-1} \mathbf{X}$. Then Y^2 obeys χ_d^2 and $EY^2 = d$. The acceptance region $\mathbf{A}_{p_t} = \{\mathbf{x} | \mathbf{x}' \boldsymbol{\Sigma}_t^{-1} \mathbf{x} \leq r_\theta^2\}$ is rewritten by $\mathbf{A}_{p_t} = \{\mathbf{x} | \mathbf{x}' \boldsymbol{\Sigma}_*^{-1} \mathbf{x} \leq \alpha_t r_\theta^2\}$. Then, with $f(y^2) = \chi_d^2(y^2)$, we have

$$\begin{aligned} E_{\mathbf{X} \in \mathbf{A}_{p_t}} Y^2 &= \int_0^{\alpha_t r_\theta^2} y^2 f(y^2) dy^2 / \int_0^{\alpha_t r_\theta^2} f(y^2) dy^2 \\ &= \frac{2\gamma\left(\frac{d}{2} + 1, \alpha_t r_\theta^2\right)}{\gamma\left(\frac{d}{2}, \alpha_t r_\theta^2\right)}, \end{aligned} \quad (7)$$

where $\gamma(\cdot, \cdot)$ is the incomplete gamma function defined by $\gamma(a, x) = \int_0^x e^{-t} t^{a-1} dt$. Note that $E_{\mathbf{X} \in \mathbf{A}_{p_t}} Y^2 \rightarrow d$ as $r_\theta^2 \rightarrow \infty$. From Eq. (7) and the fact that $E_{\mathbf{X} \in \mathbf{A}_{p_t}} Y^2 = \text{tr}(\boldsymbol{\Sigma}_*)^{-1} E_{\mathbf{X} \in \mathbf{A}_{p_t}} \mathbf{X} \mathbf{X}'$, we have

$$E_{\mathbf{X} \in \mathbf{A}_{p_t}} \mathbf{X} \mathbf{X}' = \frac{2\gamma\left(\frac{d}{2} + 1, \alpha_t r_\theta^2\right)}{d\gamma\left(\frac{d}{2}, \alpha_t r_\theta^2\right)} \boldsymbol{\Sigma}_*. \quad (8)$$

Fig. 6. Acceptance region of prototype **Fig. 7.** Update scheme of covariance matrix $\mathbf{\Sigma}_t$

On the other hand, the covariance $\mathbf{\Sigma}_t$ is updated to $\mathbf{\Sigma}_{t+1}$ by $\mathbf{X}_{t+1}$ as

$$\mathbf{\Sigma}_{t+1} = \frac{n_0 + t}{n_0 + t + 1} \mathbf{\Sigma}_t + \frac{1}{n_0 + t + 1} \mathbf{X}_{t+1} \mathbf{X}_{t+1}'.$$

By taking the expectation over $\mathbf{X}_{t+1} \in \mathbf{A}_{p_t}$, from (8), we have

$$\begin{aligned} \alpha_{t+1} \mathbf{\Sigma}_* &= E_{\mathbf{X}_{t+1} \in \mathbf{A}_{p_t}} \mathbf{\Sigma}_{t+1} = \mathbf{\Sigma}_t + \frac{E \mathbf{X}_{t+1} \mathbf{X}_{t+1}' - \mathbf{\Sigma}_t}{n_0 + t + 1} \\ &= \alpha_t \mathbf{\Sigma}_* + \frac{\frac{2\gamma(\frac{d}{2} + 1, \alpha_t r_\theta^2)}{d\gamma(\frac{d}{2}, \alpha_t r_\theta^2)} - \alpha_t}{n_0 + t + 1} \mathbf{\Sigma}_*. \end{aligned}$$

That is,

$$\alpha_{t+1} = \alpha_t + \frac{\frac{2\gamma(\frac{d}{2} + 1, \alpha_t r_\theta^2)}{d\gamma(\frac{d}{2}, \alpha_t r_\theta^2)} - \alpha_t}{n_0 + t + 1},$$

$$\alpha_{t+1} d = \left(1 - \frac{1}{n_0 + t + 1}\right) \alpha_t d + \frac{1}{n_0 + t + 1} \frac{2\gamma(\frac{d}{2} + 1, \alpha_t r_\theta^2)}{\gamma(\frac{d}{2} + 1, \alpha_t r_\theta^2)}. \quad (9)$$

At time t , $\mathbf{\Sigma}_t$ and $\mathbf{A}_t$ correspond to $\alpha_t d$ and $\alpha_t r_\theta^2$, respectively (Fig. 6). Note that $r_\theta^2 > d$ for large θ , say $\theta = 0.9$. Then the contribution by $\mathbf{X}_{t+1}$ corresponds to $2\gamma(\frac{d}{2} + 1, \alpha_t r_\theta^2) / \gamma(\frac{d}{2}, \alpha_t r_\theta^2)$. Thus a pair $(\alpha_t d, \alpha_t r_\theta^2)$ produces $2\gamma(\frac{d}{2} + 1, \alpha_t r_\theta^2) / \gamma(\frac{d}{2}, \alpha_t r_\theta^2)$.

When $\alpha_t > 1$, since $2\gamma(\frac{d}{2} + 1, \alpha_t r_\theta^2) / \gamma(\frac{d}{2}, \alpha_t r_\theta^2) < d < \alpha_t d < \alpha_t r_\theta^2$, we have $\alpha_{t+1} d < \alpha_t d$ from (9). On the other hand, when α_t is smaller than α^* defined below, we have $\alpha_t d < 2\gamma(\frac{d}{2} + 1, \alpha_t r_\theta^2) / \gamma(\frac{d}{2}, \alpha_t r_\theta^2) < \alpha_t r_\theta^2$, then $\alpha_{t+1} d > \alpha_t d$. The latter situation is illustrated in Fig. 7.

To find the convergence value α^* , it suffices to solve the equation

$$\alpha d = \frac{2\gamma(\frac{d}{2} + 1, \alpha r_\theta^2)}{\gamma(\frac{d}{2}, \alpha r_\theta^2)}.$$

Table 2. Equilibrium value α^* of $\alpha = 2\gamma \left(\frac{d}{2} + 1, \frac{\alpha r_\theta^2}{2} \right) / d\gamma \left(\frac{d}{2}, \frac{\alpha r_\theta^2}{2} \right)$

θ	d					
	2	4	8	16	32	64
0.85	—	0.367	0.628	0.769	0.850	0.901
0.90	0.346	0.634	0.781	0.862	0.911	0.940
0.95	0.715	0.835	0.899	0.936	0.958	0.972

Let us denote $\alpha^* = \alpha^*(d, \theta)$ by making clear the dependency. Several values of $\alpha^*(d, \theta)$ are shown in Table. 2. Here $r_\theta^2 = \chi_d^2(\theta)$. As a result, the covariance matrix Σ_t approaches to $\alpha^*(d, \theta)\Sigma_*$ as $t \rightarrow \infty$.

We conducted the same experiment as Experiment 1. The result is shown in Fig. 5. In Fig. 5 (the lower row), we plotted the value of $\text{tr}\Sigma_t/d$ as an estimate of α_t because $\text{tr}\mathbf{I}_d = d$. From Fig. 5, we can see that α_t approaches to the theoretical equilibrium value $\alpha^*(d, \theta)$.

In total, we can say that volume prototypes converge to local peaks at a speed proportional to the reciprocal of the number of samples and then the covariance matrices converges to $\alpha^*(d, \theta)$ times the local covariance matrices.

5 Experiments

5.1 Classification by Volume Prototypes

Here, we present a simple but natural classifier constructed by volume prototypes. Each volume prototype can be regarded as a component of a mixture model, but is closer to a point as a prototype. Thus according to the following rule we classify a class-unknown sample $\mathbf{x}$ to c^* .

$$c^* = \arg \max_c \max_{k|c} P(c)P(k|c)f(\mathbf{x}|\boldsymbol{\theta}_{c,k}),$$

where $P(c)$ is the prior probability of class c , $P(k|c)$ is the probability of k th prototype in c , and $f(\mathbf{x}, \boldsymbol{\theta}_{c,k})$ is the probabilistic distance of $\mathbf{x}$ to k th prototype of class c . In the estimation, we used $P(k|c) = n_{c,k}/n_c$ ($n_c = \sum_k n_{c,k}$) and $f(\mathbf{x}|\boldsymbol{\theta}_{c,k}) = N(\mathbf{x}|\boldsymbol{\mu}_{c,k}, \Sigma_{c,k})$ for $\mathbf{p} = (\boldsymbol{\mu}_{c,k}, \Sigma_{c,k}, n_{c,k})$ where n_c is the total number of the latest samples of c that were kept in VP algorithm. Here $n_{c,k}$ was modified to satisfy $\sum_k n_{c,k} = n_c$ by taking into account the degree of overlapping among volume prototypes. To estimate the classification rate of the classifier, we used 10-fold cross validation.

We used a multi-featured dataset of 10 handwritten numerals represented in terms of four different feature sets [4]. The number of prototypes was set to $M = 10$, the first $N = 20$ samples were used for ME step and the remaining 180 samples were used for C step. Algorithm VP was invoked class by class. The radius was $r^2 = \chi_d^2(\theta)$ with $\theta = 0.9$. The latest sample number was $N_v = 180$ because no more sample was available.

Table 3. Recognition rate in each feature set. The figure of parentheses is the dimensionality.

Classifier	Recognition rate (%)			
	mfeat-zer (47)	mfeat-kar (64)	mfeat-fou (76)	mfeat-pix (240)
VP Clf ($\theta = 0.90$)	83.5	82.5	97.3	97.9
SVM	80.9	82.0	97.5	97.9

(a) mfeat-zer (10-class) (b) mfeat-zer (2-class) (c) mfeat-pix (10-class) (d) mfeat-pix (2-class)

Fig. 8. Volume prototypes in each dataset (10 classes). The number of volume prototypes is 10 in each class. The radius is $r^2 = \chi_d^2(0.9)$. The groups of various marks are the centers of prototypes and the group of ellipsoids with a same color is the volume prototypes of the class. The covariance of datapoints is drawn as a black bold ellipsoid.

The result is shown in Table 3 and Fig. 8. For comparison, the results by SVM with a RBF kernel is also shown.

From Fig. 8, we see that 1) 10 prototypes are closely located but they represent well the spread of the class distribution (Fig. 8 (a), (c)), 2) the sizes are a little smaller than that of the covariance matrix of all the samples of the class (two of them are chosen for Fig. 8, (b), (d)). Note that the prototype sizes look larger than the real size when they are projected on a 2-dimensional space, because the volume of an ellipsoid increases exponentially in dimensionality. As a whole, volume prototypes well represent the configuration of dataset. The recognition rates are comparable to those of SVM. This means that the volume prototypes keep the characteristics of the original data.

6 Conclusion

We have clarified the behavior of volume prototypes toward large and high-dimensional datasets. A volume prototype moves toward a dense part of samples, e.g., a mode, and the covariance matrix of a prototype converges to a certain size depending on the local structure of the underlying distribution. The size of converged prototype is affected by the dimensionality and the user-specified radius for prototypes.

The efficiency of the classifier with volume prototypes is also shown. There are various possible applications of volume prototypes in pattern recognition tasks, in virtue of the high compression rate and the single-pass algorithm for

generating them. In the future, we will particularly investigate the availability of volume prototypes to clustering and mixture model estimation.

References

1. Tabata, K., Kudo, M.: Information compression by volume prototypes. The IEICE Technical Report, PRMU 106(429), 25–30 (2006) (in Japanese)
2. Kaymak, U., Setnes, M.: Fuzzy clustering with volume prototypes and adaptive cluster merging. *IEEE Transactions on Fuzzy Systems* 10(6), 705–712 (2002)
3. Fukunaga, K.: *Introduction to Statistical Pattern Recognition*. Academic Press, London (1990)
4. Asuncion, A., Newman, D.: *UCI machine learning repository* (2007)