Kernel Relative Principal Component Analysis for Pattern Recognition

Yoshikazu Washizawa[1], Kenji Hikida[2], Toshihisa Tanaka[3,4], and Yuhikiko Yamashita[5]

[1] Toshiba Solutions Corporation, Tokyo, 183-8511, Japan
washizawa.yoshikazu@toshiba-sol.co.jp
[2] Wireless Terminals, Texas Instruments Japan Limited, Tokyo 160-8366, Japan
hikida@ti.com
[3] Department of Electrical and Electronic Engineering
Tokyo University of Agriculture and Technology, Tokyo 184-8588, Japan
tanakat@cc.tuat.ac.jp
[4] Laboratory for Advanced Brain Signal Processing, RIKEN Brain Science Institute
Saitama 351-0198, Japan
[5] Graduate School of Science and Engineering, Tokyo Institute of Technology
Tokyo 152-8552, Japan
yamasita@ide.titech.ac.jp

Abstract. Principal component analysis (PCA) is widely used in signal processing, pattern recognition, etc. PCA was extended to the relative PCA (RPCA). RPCA provides principal components of a signal while suppressing effects of other signals. PCA was also extended to the kernel PCA (KPCA). By using a mapping from the original space to a higher dimensional space and its kernel, we can perform PCA in the higher dimensional space. In this paper, we propose the kernel RPCA (KRPCA) and give its solution. Similarly to KPCA, the order of matrices that we should calculate for the solution is the number of samples, that is 'kernel trick'. We provide experimental results of an application to pattern recognition in order to show the advantages of KRPCA over KPCA.

1 Introduction

Principal component analysis (PCA) or Karhunen-Loève transform is widely used in signal processing, pattern recognition, etc [1], [2], [3]. We can extract important components of a signal that minimize the mean square error between the extracted and the original signals. Consider that input vectors are in a D-dimensional real vector space R^D with the inner product $\langle \boldsymbol{f}, \boldsymbol{g} \rangle$ and the norm $\|\boldsymbol{f}\| = \sqrt{\langle \boldsymbol{f}, \boldsymbol{f} \rangle}$ for vectors $\boldsymbol{f}$ and $\boldsymbol{g}$. Let P be a matrix. PCA is defined by minimizing the following criterion

$$E_{\boldsymbol{f}} \|P\boldsymbol{f} - \boldsymbol{f}\|^2 \tag{1}$$

under the condition that rank$(P) \leq d$ for any $d \leq D$, where $E_{\boldsymbol{f}}$ is the ensemble average with respect to a signal $\boldsymbol{f}$. Let $R_{\boldsymbol{f}} = E_{\boldsymbol{f}} \boldsymbol{f}\boldsymbol{f}^T$ be the correlation matrix of $\boldsymbol{f}$, where $\boldsymbol{f}^T$ is the transpose of $\boldsymbol{f}$. The vector $\boldsymbol{\phi}_n$ is given as the eigen vector corresponding to the n-th largest eigenvalue of $R_{\boldsymbol{f}}$. The solution of the PCA P is given as

A. Fred et al. (Eds.): SSPR&SPR 2004, LNCS 3138, pp. 1105–1113, 2004.

$$P = \sum_{n=1}^{d} \boldsymbol{\phi}_n \boldsymbol{\phi}_n^T. \tag{2}$$

We consider the case that there exist two signals $\boldsymbol{f}$ and $\boldsymbol{g}$ and the principal component of $\boldsymbol{f}$ is obtained while the effect of $\boldsymbol{g}$ is suppressed. When $\boldsymbol{g}$ is noise, we can extract the principal components of a signal $\boldsymbol{f}$ in the presence of noise. When $\boldsymbol{g}$ is a pattern in other categories, we can extract the features of $\boldsymbol{f}$ that are not closed to the features of $\boldsymbol{g}$.

By extending eq.(1) the relative PCA (RPCA) or the relative Karhunen-Loève transform was proposed [4], [5]. Consider that a matrix X minimizes the following criterion:

$$E_{\boldsymbol{f}} \|X\boldsymbol{f} - \boldsymbol{f}\|^2 + \alpha E_{\boldsymbol{g}} \|X\boldsymbol{g}\|^2 \tag{3}$$

under the condition that rank$(X) \leq d$ for any d. The parameter $\alpha (> 0)$ controls the weight for suppressing the effect of $\boldsymbol{g}$. With a large α, the effect of $\boldsymbol{g}$ is suppressed well. With a small α, the approximation error between $X\boldsymbol{f}$ and $\boldsymbol{f}$ is decreased.

The solution of RPCA is given as the form:

$$X = \sum_{n=1}^{d} \boldsymbol{\phi}_n (\boldsymbol{\varphi}_n)^T. \tag{4}$$

Here, we describe its solution when $R_{\boldsymbol{f}} + \alpha R_{\boldsymbol{g}}$ is not singular. Let μ_n ($\mu_1 \geq \mu_2 \geq \cdots \geq \mu_D$) be eigenvalues and let $\boldsymbol{\psi}_i$ be corresponding eigen vectors of $R_{\boldsymbol{f}}(R_{\boldsymbol{f}} + \alpha R_{\boldsymbol{g}})^{-1} R_{\boldsymbol{f}}$. Then, the solution is given as $\boldsymbol{\phi}_n = \boldsymbol{\psi}_n$ and $\boldsymbol{\varphi}_n = (R_{\boldsymbol{f}} + \alpha R_{\boldsymbol{g}})^{-1} R_{\boldsymbol{f}} \boldsymbol{\psi}_n$. From eq.(4) the n-th relative principal component of $\boldsymbol{f}$ is given as $\langle \boldsymbol{f}, \boldsymbol{\varphi}_n \rangle \boldsymbol{\phi}_n$. When $d = D$, X is reduced to Wiener filter, which provides the best approximation of a signal in sense of mean square error. A similar criterion with eq.(3) was provided for the rank reduced Wiener filter. However, the reason why they restrict the rank is not for obtaining principal components but for robustness [6].

The advantages of RPCA over PCA were shown by experiments of data compression in the presence of noise [4] and handwritten character recognition [5].

As for PCA with two kinds of signals, Fisher discriminant [7] and Oriented PCA (OPCA) [3] were proposed. OPCA is defined by vectors $\boldsymbol{\phi}_n$ that minimize

$$\frac{E_f \langle f, \boldsymbol{\phi}_n \rangle^2}{E_g \langle g, \boldsymbol{\phi}_n \rangle^2} \tag{5}$$

under the condition that $\langle \boldsymbol{\phi}_m, R_f \boldsymbol{\phi}_n \rangle = 0$ and $\langle \boldsymbol{\phi}_m, R_g \boldsymbol{\phi}_n \rangle = 0$ for $m \neq n$. The theoretical advantage of RPCA over them is in that the criterion of RPCA compares the approximation error between principal components and the original signal directly. Then, it provides the principal components of which accuracy is guaranteed.

PCA was extended to another direction based on a kernel. The kernel PCA (KPCA) is defined as follows [8], [9], [10], [11], [12], [13]. Let Φ be a mapping from a input vector space R^D to a real Hilbert space $\mathcal{H}$. The inner product $\langle \boldsymbol{x}, \boldsymbol{y} \rangle$ is also defined for vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ in $\mathcal{H}$. Furthermore, we assume $\langle \Phi(\boldsymbol{f}), \Phi(\boldsymbol{g}) \rangle = k(\boldsymbol{f}, \boldsymbol{g})$, where $k(\boldsymbol{f}, \boldsymbol{g})$ is a kernel function. Let $\{\boldsymbol{f}_l\}_{l=1}^{L}$ be a set of samples of a signal. When $\mathcal{H}$ is a vector space, a sample correlation matrix $S_{\boldsymbol{f}}$ in $\mathcal{H}$ is given as

$$S_{\boldsymbol{f}} = \frac{1}{L}\sum_{l=1}^{L} \Phi(\boldsymbol{f}_l)\Phi(\boldsymbol{f}_l)^T. \tag{6}$$

Let V^n be the eigen vector corresponding to the n-th largest eigenvalue of $S_{\boldsymbol{f}}$. For an input vector $\boldsymbol{f}$, the magnitude of the n-th principal component is given as $\langle V^n, \Phi(\boldsymbol{f})\rangle$. However, for obtaining the magnitude we don't need to calculate the eigenvalue problem in $\mathcal{H}$, which is a very high or infinite dimensional space. The advantage of the theory of KPCA is that we can reduce the dimension for the calculation to the number of samples L. Fisher discriminant was also extended to the kernel Fisher discriminant [14], [15], [16].

Since a signal contains noise in almost all cases or suppression of the effect of other categories is useful in many cases, the kernelization of RPCA is very important as well as the kernelization of PCA. In this paper, we propose the kernel relative PCA (KRPCA) which is a kernel based extension of RPCA. We provide the definition and its solution in this paper. The dimension of a space for the calculation is the number of samples. Different from other many kernelized problems such as support vector machine (SVM), the solution that minimizes the criterion of KRPCA is given by a closed form, that is, it can be provided by eigen vectors and inversion of matrices similarly to KPCA. Furthermore, since it provides the relative principal components, it can be applied not only to discrimination but also to many problems such as feature extraction and dimensional reduction. In order to show the advantage of KRPCA, we show experimental results of an application to pattern recognition.

2 Kernel Relative PCA

In this section, we provide the definition and a solution of KRPCA.

Schatten product $\boldsymbol{x} \otimes \overline{\boldsymbol{y}}$ for vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ in $\mathcal{H}$ is defined as a linear operator from $\mathcal{H}$ to $\mathcal{H}$ such that $(\boldsymbol{x} \otimes \overline{\boldsymbol{y}})\boldsymbol{z} = \langle \boldsymbol{z}, \boldsymbol{y}\rangle \boldsymbol{x}$ for any $\boldsymbol{z} \in \mathcal{H}$ [17]. It is an abstract notation of $\boldsymbol{x}\boldsymbol{y}^T$ for vectors in a Hilbert space.

Let $\{\boldsymbol{f}_l\}_{l=1}^{L}$ be a set of samples, of which principal components are extracted. Let $\{\boldsymbol{g}_m\}_{m=1}^{M}$ be a set of samples of which effect is suppressed. We define the criterion of KRPCA X as minimizing

$$J = \frac{1}{L}\sum_{l=1}^{L} \|X\Phi(\boldsymbol{f}_l) - \Phi(\boldsymbol{f}_l)\|^2 + \frac{\alpha}{M}\sum_{m=1}^{M} \|X\Phi(\boldsymbol{g}_m)\|^2 \tag{7}$$

under the conditions that $\mathrm{rank}(X) \leq d$ and its null space includes the orthogonal subspace of the subspace spanned by $\{\boldsymbol{f}_l\}_{l=1}^{L}$ and $\{\boldsymbol{g}_m\}_{m=1}^{M}$.

For brief, let $\boldsymbol{f}_{i+L} = \boldsymbol{g}_i$ $(i = 1, 2, \cdots, M)$ and $N = L + M$. Since all vectors in the criterion(7) are in a subspace spanned by $\Phi(\boldsymbol{f}_i)$ $(i = 1, 2, \cdots, L, L+1, \cdots, N)$, we can assume that X is expressed with an (N, N)-matrix $A = (a_{ij})$ as

$$X = \sum_{i=1}^{N}\sum_{j=1}^{N} a_{ij}\Phi(\boldsymbol{f}_i) \otimes \overline{\Phi(\boldsymbol{f}_j)} \tag{8}$$

Let $\mathbf{0}_{mn}$ be (m, n)-matrix of which all elements are zero. We define (L, L)-matrix K_0, (N, L)-matrix K_1, (N, M)-matrix K_2, (N, N)-matrix $\tilde{K}_1$, (N, N)-matrix $\tilde{K}_2$, (N, N)-matrix K, and (N, N)-matrix $\tilde{K}$ as

$$K_0 = \begin{pmatrix} k(\boldsymbol{f}_1, \boldsymbol{f}_1) & \cdots & k(\boldsymbol{f}_1, \boldsymbol{f}_L) \\ \vdots & \ddots & \vdots \\ k(\boldsymbol{f}_L, \boldsymbol{f}_1) & \cdots & k(\boldsymbol{f}_L, \boldsymbol{f}_L) \end{pmatrix},$$

$$K_1 = \begin{pmatrix} k(\boldsymbol{f}_1, \boldsymbol{f}_1) & \cdots & k(\boldsymbol{f}_1, \boldsymbol{f}_L) \\ \vdots & \ddots & \vdots \\ k(\boldsymbol{f}_N, \boldsymbol{f}_1) & \cdots & k(\boldsymbol{f}_N, \boldsymbol{f}_L) \end{pmatrix}, \quad K_2 = \begin{pmatrix} k(\boldsymbol{f}_1, \boldsymbol{f}_{L+1}) & \cdots & k(\boldsymbol{f}_1, \boldsymbol{f}_N) \\ \vdots & \ddots & \vdots \\ k(\boldsymbol{f}_N, \boldsymbol{f}_{L+1}) & \cdots & k(\boldsymbol{f}_N, \boldsymbol{f}_N) \end{pmatrix},$$

$$\tilde{K}_1 = \left[\frac{1}{\sqrt{L}} K_1 \quad \mathbf{0}_{NM} \right], \quad \tilde{K}_2 = \left[\mathbf{0}_{NL} \quad \sqrt{\frac{\alpha}{M}} K_2 \right], \quad K = [K_1 \quad K_2], \quad \tilde{K} = \tilde{K}_1 + \tilde{K}_2.$$

It is clear that $K_1 K_1^T = L \tilde{K}_1 \tilde{K}_1^T$, $\alpha K_2 K_2^T = M \tilde{K}_2 \tilde{K}_2^T$, and $\tilde{K}_1 \tilde{K}_1^T + \tilde{K}_2 \tilde{K}_2^T = \tilde{K} \tilde{K}^T$. For a matrix B there exists an unique matrix C such that $BCB = B$, $CBC = C$, $(BC)^T = BC$, and $(CB)^T = CB$. The matrix C is called the Moore-Penrose generalized inverse matrix and denoted by $B^\dagger$ [18]. A symmetric matrix B is called the *positive semi-definite* if and only if $\langle B\boldsymbol{x}, \boldsymbol{x} \rangle \geq 0$ for any $\boldsymbol{x}$. For a positive semi-definite matrix B, there exists a positive semi-definite matrix C such that $CC = B$. We denote the matrix C by $B^{1/2}$. We define a matrix A_0 as

$$A_0 = K^\dagger \tilde{K}_1 \tilde{K}_1^T (\tilde{K} \tilde{K}^T)^\dagger. \tag{9}$$

Let λ_i ($\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_N$) be eigenvalues of $K^{1/2} A_0 \tilde{K} (K^{1/2} A_0 \tilde{K})^T$, which is a symmetric matrix, and let $\boldsymbol{u}_i$ be the corresponding eigen vectors such that $\{\boldsymbol{u}_n\}_{n=1}^N$ is an orthonormal basis. When $\lambda_n \neq 0$, let $\boldsymbol{v}_n = ((K^{1/2} A_0 \tilde{K})^T \boldsymbol{u}_n) / \sqrt{\lambda_n}$. When $\lambda_n = 0$, we can chose any $\boldsymbol{v}_n$ such that $\{\boldsymbol{v}_n\}_{n=1}^N$ is an orthonormal basis. The i-th element of an N-dimensional vector $\boldsymbol{w}$ is denoted by $(\boldsymbol{w})_i$.

Theorem 1. *A KRPCA X is given as*

$$X = \sum_{i=1}^N \sum_{j=1}^N \left(\sum_{n=1}^d \left((K^{1/2})^\dagger \boldsymbol{u}_n \right) \tilde{\boldsymbol{v}}_n^T \right)_{ij} \Phi(\boldsymbol{f}_i) \otimes \overline{\Phi(\boldsymbol{f}_j)}, \tag{10}$$

where

$$\tilde{\boldsymbol{v}}_n = \sqrt{\lambda_n} (\tilde{K}^\dagger)^T \boldsymbol{v}_n. \tag{11}$$

For an input vector $\boldsymbol{f}$*, let*

$$\boldsymbol{h} = (k(\boldsymbol{f}, \boldsymbol{f}_1), k(\boldsymbol{f}, \boldsymbol{f}_2), \cdots, k(\boldsymbol{f}, \boldsymbol{f}_N))^T. \tag{12}$$

Then, $X\Phi(\boldsymbol{f})$ *is given as*

$$X\Phi(\boldsymbol{f}) = \sum_{i=1}^N \sum_{n=1}^d \langle \boldsymbol{h}, \tilde{\boldsymbol{v}}_n \rangle \left((K^{1/2})^\dagger \boldsymbol{u}_n \right)_i \Phi(\boldsymbol{f}_i). \tag{13}$$

The n-th kernel relative principal component of $\boldsymbol{f}$ *with respect to* $\mathbf{g}$ *is given as*

$$\sum_{i=1}^{N} \langle \boldsymbol{h}, \tilde{\boldsymbol{v}}_n \rangle \left((K^{1/2})^{\dagger} \boldsymbol{u}_n \right)_i \Phi(\boldsymbol{f}_i). \tag{14}$$

Furthermore, we have

$$\|X\Phi(\boldsymbol{f})\|^2 = \sum_{n=1}^{d} |\langle \tilde{\boldsymbol{v}}_n, \boldsymbol{h} \rangle|^2 \tag{15}$$

and

$$\|X\Phi(\boldsymbol{f}) - \Phi(\boldsymbol{f})\|^2 = k(\boldsymbol{f}, \boldsymbol{f}) + \sum_{n=1}^{d} \{ |\langle \tilde{\boldsymbol{v}}_n, \boldsymbol{h} \rangle|^2 - \langle \tilde{\boldsymbol{v}}_n, \boldsymbol{h} \rangle \langle (K^{1/2})^{\dagger} \boldsymbol{u}_n, \boldsymbol{h} \rangle \}. \tag{16}$$

Outline of the Proof
We can expand the sum in eq.(7) and simplify to

$$J = \|K^{1/2} A \tilde{K} - K^{1/2} A_0 \tilde{K}\|_2^2 - \mathrm{tr}[\tilde{K}^T A_0^T K A_0 \tilde{K}] + \frac{1}{L} \mathrm{tr}[K_0]. \tag{17}$$

where $\|\cdot\|_2$ is the Frobenius norm of a matrix. From eq.(17), J is minimum subject to $\mathrm{rank}(A) \leq d$ if and only if $\|K^{1/2} A \tilde{K} - K^{1/2} A_0 \tilde{K}\|_2^2$ is minimum with the condition. From the definitions of $\boldsymbol{u}_n$ and $\boldsymbol{v}_n$, by considering SVD of $K^{1/2} A_0 \tilde{K}$, J is minimum subject to $\mathrm{rank}(A) \leq d$ if and only if

$$K^{1/2} A \tilde{K} = \sum_{n=1}^{d} \sqrt{\lambda_n} \boldsymbol{u}_n \boldsymbol{v}_n^T. \tag{18}$$

Note that the sum of (18) is truncated by d. Then, we have

$$A = (K^{1/2})^{\dagger} \sum_{n=1}^{d} \boldsymbol{u}_n \tilde{\boldsymbol{v}}_n^T. \tag{19}$$

Then, eq.(10) is proved. Furthermore, $X\Phi(\boldsymbol{f})$ is given as

$$X\Phi(\boldsymbol{f}) = \sum_{i=0}^{N} (A\boldsymbol{h})_i \Phi(\boldsymbol{f}_i) = \sum_{i=1}^{N} \sum_{n=1}^{d} \langle \boldsymbol{h}, \tilde{\boldsymbol{v}}_n \rangle \left((K^{1/2})^{\dagger} \boldsymbol{u}_n \right)_i \Phi(\boldsymbol{f}_i). \tag{20}$$

The rest of the proof is clear. □

3 Application to Pattern Recognition

In order to show the advantage of KRPCA, we provide an experimental result of handwritten character recognition.

Let Ω_c be the learning sample set for a category c ($c = 1, 2, \cdots, N_C$). The matrix P_c and the operator P'_c of PCA and KPCA for the category Ω_c are decided as minimizing

$$\frac{1}{|\Omega_c|} \sum_{\boldsymbol{f} \in \Omega_c} \|P_c \boldsymbol{f} - \boldsymbol{f}\|^2, \qquad \frac{1}{|\Omega_c|} \sum_{\boldsymbol{f} \in \Omega_c} \|P'_c \Phi(\boldsymbol{f}) - \Phi(\boldsymbol{f})\|^2$$

subject to rank(P_c) and rank(P'_c) are fixed, respectively, where $|\Omega_c|$ is the number of samples in Ω_c.

The matrix X_c and the operator X'_c of the RPCA and the KRPCA for a category c is decided as minimizing with a parameter α

$$\frac{1}{|\Omega_c|}\sum_{\boldsymbol{f}\in\Omega_c}\|X_c\boldsymbol{f}-\boldsymbol{f}\|^2+\alpha\frac{1}{|\Omega_x|}\sum_{\boldsymbol{g}\in\Omega_x}\|X_c\boldsymbol{g}\|^2, \tag{21}$$

$$\frac{1}{|\Omega_c|}\sum_{\boldsymbol{f}\in\Omega_c}\|X'_c\Phi(\boldsymbol{f})-\Phi(\boldsymbol{f})\|^2+\alpha\frac{1}{|\Omega_x|}\sum_{\boldsymbol{g}\in\Omega_x}\|X'_c\Phi(\boldsymbol{g})\|^2 \tag{22}$$

subject to rank(X_c) and rank(X'_c) are fixed, respectively, where Ω_x is the set of samples which are suppressed. We call Ω_x the suppression set. In the above criterion, the notations $\boldsymbol{f}$ and $\boldsymbol{g}$ express patterns in the own and the others categories, respectively.

An unknown pattern $\boldsymbol{h}$ is discriminated as the category c for each method, when for all $b \neq c$ we have

$$\|P_c\boldsymbol{h}-\boldsymbol{h}\|^2 < \|P_b\boldsymbol{h}-\boldsymbol{h}\|^2, \qquad \|P'_c\Phi(\boldsymbol{h})-\Phi(\boldsymbol{h})\|^2 < \|P'_b\Phi(\boldsymbol{h})-\Phi(\boldsymbol{h})\|^2,$$
$$\|X_c\boldsymbol{h}-\boldsymbol{h}\|^2 < \|X_b\boldsymbol{h}-\boldsymbol{h}\|^2, \qquad \|X'_c\Phi(\boldsymbol{h})-\Phi(\boldsymbol{h})\|^2 < \|X'_b\Phi(\boldsymbol{h})-\Phi(\boldsymbol{h})\|^2.$$

In cases of PCA and KPCA since P_c and P'_c are orthogonal projection matrix and operator, $\|P_c\boldsymbol{h}-\boldsymbol{h}\|^2$ and $\|P'_c\Phi(\boldsymbol{h})-\Phi(\boldsymbol{h})\|^2$ are minimum if and only if $\|P_c\boldsymbol{h}\|^2$ and $\|P'_c\Phi(\boldsymbol{h})\|^2$ are maximum, respectively. Usually the latter rules are used as the discriminant laws.

In point of view of the learning set, KPCA and KRPCA use the same learning set. In point of view of calculation complexity, the dimension of the space where we have to calculate is the number of learning samples used for evaluations. For KPCA and KRPCA the numbers are given as $|\Omega_c|$ and $|\Omega_c| + |\Omega_x|$, respectively. Therefore, it is difficult to use all samples, which do not belong to Ω_c, for the suppression set Ω_x. Then, we fix $|\Omega_x|$ as N_x. Consider a value $t(\boldsymbol{g}) = \|P'_c\Phi(\boldsymbol{g})\|/\|P'_b\Phi(\boldsymbol{g})\|$ for a pattern $\boldsymbol{g}$ in Ω_b ($b \neq c$). Let t_{N_x} be the N_x-th largest value among $t(\boldsymbol{h})$ for all patterns $\boldsymbol{h}$ in Ω_b ($b \neq c$). We add the pattern $\boldsymbol{g}$ to the suppression set Ω_x with respect to Ω_c when $t(\boldsymbol{g})$ is not less than t_{N_x}.

3.1 Data

We use US Postal Service database (USPS) which contains 7291 training patterns and 2007 test patterns collected from real-life zip codes. It has ten categories from '0' to '9' ($N_c = 10$). For a preprocessing we use the weighted direction index histogram method and the variable transformation [19].

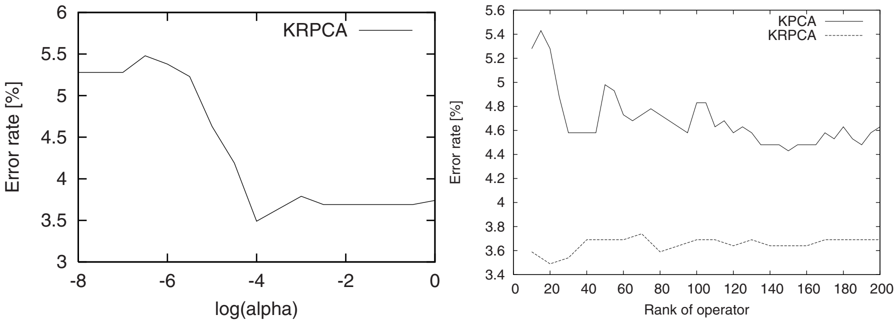
3.2 Result

The following kernel is used in this experiment for KPCA and KRPCA.

$$k(\boldsymbol{f},\boldsymbol{g}) = (\langle\boldsymbol{f},\boldsymbol{g}\rangle + 1)^{20}. \tag{23}$$

Table 1. Result of handwritten character recognition.

METHOD	ERROR RATE (%)	RANK	α	$\lvert\Omega_x\rvert$
PCA	5.38	9	–	
RPCA	4.91	12	$10^{-2.5}$	25
KPCA	4.43	150	–	
KRPCA	3.49	20	$10^{-4.0}$	150



(a)

(b)

Fig. 1. (a) Parameter α and error rate by KRPCA(rank=20) ($|\Omega_x| = 150$). (b) Rank and error rate in KRPCA($\alpha = 10^{-4.0}$, $|\Omega_x| = 150$) and KPCA.

We show the best error rate of the test set for each method among various ranks and values of the parameter α, $|\Omega_x|$ in Table 1. We show the relation between the parameter α and the error rate of KRPCA in Figure 1 (a). We also show the relation between the ranks and the error rates of KPCA and KRPCA in Figure 1 (b).

We can see from Table 1 that KRPCA performs the best recognition rate. We can also see from Figures 1 (a) and (b) that KRPCA outperforms KPCA for any $\alpha > 10^{-4}$ and for any rank of operators, respectively.

4 Discussion

4.1 Computational Complexity

In the construction stage, the computational complexity to calculate a matrix of KRPCA is several times higher than that of KPCA. The dominant parts of computations are as follows.

- KPCA

the kernel function:	L^2 times
the eigen value problem of an (L, L)-matrix:	1 times

- KRPCA

the kernel function:	N^2 times
the inversion of an (N, N)-matrix:	2 times
the SVD of an (N, N)-matrix:	1 times
the square root of a symmetric (N, N)-matrix:	1 times

In the recognition stage, since many terms in eq.(16) can be calculated in advance, the dominant parts of computations are as follows.

- KPCA
  the kernel function: L times
  multiplication of elements: DL times
- KRPCA
  the kernel function: N times
  multiplication of elements: $2DN$ times

where D is the rank of each operator. In case of the experiment in Section 3, $L = 729$ and $N = 879$ in average and $D = 150$ for KPCA and $D = 20$ for KRPCA to achieve minimum error rates. In this case, the computational complexity of KRPCA is less than that of the KPCA in the recognition stage.

4.2 Comparison with Other Kernel Machines

SVM has the problem that training needs enormous computational cost because the optimization problem becomes very large. It depends on the complexity of the problem and the number of samples belonging to both its own and all rival classes. If the total number of samples is very large, kernel fisher discriminant (KFD) also has the same problem. On the other hand, KPCA and KRPCA are trained by the samples of its own class only and by those and some samples of rival classes, respectively, and can be solved only by matrix computations. Therefore, even if the number of classes is very large, KPCA and KRPCA can be obtained easily compared to SVM and KFD. Since SVD has a convex criterion and SVM does not have, sample selection for SVM also needs much computational cost. In practice, the KRPCA with a few thousands of samples for a class can be obtained by a present personal computer. Furthermore, since KRPCA extracts features of which mean square error is minimized, it can be used not only for discrimination but also for analysis.

5 Conclusion

We proposed the theory of the kernel relative principal component analysis (KRPCA). KRPCA can extract the principal components of a signal while suppressing the effects of other signals. We provided its definition and a solution. In order to show the advantage of KRPCA, we provided an experimental example of handwritten character recognition.

Acknowledgments

The authors wish to thank members in USPS for providing the date of the handwritten numeral characters. This research was partially supported by the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for Scientific Research (C), 1550101, 2003.

References

1. Watanabe, S., Pakvasa, N.: Subspace method in pattern recognition. Proc. 1st Int. J. Conf on Pattern Recognition, Washington DC (1973) 25–32
2. Oja, E.: Subspace Methods of Pattern Recognition. Research Studies Press, Hertfordshire (1983)
3. Diamantaras, K.I., Kung, S.Y.: Principal Component Neural Networks. John Wiley & Sons, Inc., Yew York (1996)
4. Yamashita, Y., Ogawa, H.: Relative Karhunen-Loève transform. IEEE Trans. on Signal Processing **44** (1996) 371–378
5. Ikeno, Y., Yamashita, Y., Ogawa, H.: Relative Karhunen-Loève transform method for pattern recognition. Proc. of the 14th International Conference on Pattern Recognition, Brisben, Austraria **2** (1998) 1031–1033
6. Scharf, L.: The SVD and reduced rank signal processing. Signal Processing **25** (1991) 113–133
7. Fisher, R.A.: The use of multiple measurements in taxonomic problems. Annal Eugenics **7** (1936) 179–188
8. Schölkopf, B., Smola, A., Müller, K.: Nonlinear component analysis as a kernel eigenvalue problem. Neural Computation **10** (1998) 1299–1319
9. Mika, S., Schölkopf, B., Smola, A.J., Müller, K.R., Scholz, M., Rätsch, G.: Kernel PCA and de-noising in feature spaces. In Kearns, M.S., Solla, S.A., Cohn, D.A., eds.: Advances in Neural Information Processing Systems 11, MIT Press (1999) 536–542
10. Schölkopf, B., Mika, S., Burges, C., Knirsch, P., Müller, K.R., Rätsch, G., Smola, A.: Input space vs. feature space in kernel-based methods. IEEE Transactions on Neural Networks **10** (1999) 1000–1017
11. Mika, S., Rätsch, G., Weston, J., Schölkopf, B., Smola, A.J., Müller, K.R.: Invariant feature extraction and classification in kernel spaces. In Kearns, M.S., Solla, S.A., Cohn, D.A., eds.: Advances in Neural Information Processing Systems 12, MIT Press (2000) 526–532
12. Maeda, E., Murase, H.: Kernel based nonlinear subspace method for multi-category classification. Tech. Rep., Information Science Laboratory **ISRL-98-1** (1998)
13. Tsuda, K.: Subspace classifier in the Hilbert space. Pattern Recognition Letters **20** (1999) 513–519
14. Mika, S., Rätsch, G., Weston, J., Schölkopf, B., Müller, K.R.: Fisher discriminant analysis with kernels. In Y.-H. Hu, J. Larsen, E.W., Douglas, S., eds.: Neural Networks for Signal Processing IX, IEEE (1999) 41–48
15. Mika, S., Rätsch, G., Müller, K.R.: A mathematical programming approach to the kernel Fisher algorithm. In T.K. Leen, T.D., Tresp, V., eds.: Advances in Neural Information Processing Systems 13, MIT Press (2001) 591–597
16. Mika, S., Smola, A.J., Schölkopf, B.: An improved training algorithm for kernel Fisher discriminants. In Jaakkola, T., Richardson, T., eds.: Proceedings of Eighth International Workshop on Artificial Intelligence and Statistics, Morgan Kaufmann (2001) 98–104
17. Schatten, R.: Norm Ideals of Completely Continuous Operators. Springer-Verlag, Berlin (1970)
18. Ben-Israel, A., Greville, T.N.E.: Generalized Inverses: Theory and Applications. John Wiley & Sons, New York (1974)
19. T. Wakabayashi, S. Tsuruoka, F.K., Miyake, Y.: Increasing the feature size in handwritten numeral recognition to improve accuracy. Systems and Computers in Japan (Scripta Technica) **26** (1995) 35–44