Sparsity Preserving Canonical Correlation Analysis

Chen Zu and Daoqiang Zhang

Department of Computer Science and Engineering,
Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China
{zuchen,dqzhang}@nuaa.edu.cn

Abstract. Canonical correlation analysis (CCA) acts as a well-known tool to analyze the underlying dependency between the observed samples in multiple views of data. Recently, a locality-preserving CCA, called LPCCA, has been developed to incorporate the neighborhood information into CCA. However, both CCA and LPCCA are unsupervised methods which do not take class label information into account. In this paper, we propose an alternative formulation for integrating both the neighborhood information and the discriminative information into CCA and derive a new method called Sparsity Preserving Canonical Correlation Analysis (SPCCA). In SPCCA, besides considering the correlation between two views from the same sample, the cross correlations between two views respectively from different within-class samples, which are automatically determined by performing sparse representation, are also used to achieve good performance. The experimental results on a series of data sets validate the effectiveness of the proposed method.

Keywords: Canonical correlation analysis (CCA), sparse representation, locality preserving, feature extraction, multi-view dimensionality reduction.

1 Introduction

Canonical correlation analysis (CCA) [1] is a standard method to reveal linear relationships between two views in statistics. It seeks to find two sets of directions, one for each set. Related features are extracted through maximizing the correlation between the two sets of canonical variables.

CCA is often used to find linear relationship of two sets of features (or views) by projecting two-view data into respective canonical subspaces. In the past decades, CCA and its variants have been used in many areas such as pattern recognition [2], image processing [3], image retrieval [4], regression and prediction [5]. However, standard CCA is an unsupervised method, and it can not preserve discriminant information in canonical subspaces. In order to solve this issue, a variant of CCA called discriminant CCA (DCCA) is proposed [6]. DCCA is a supervised feature fusion method, which utilizes class information by maximizing the correlation between feature vectors in the same class and minimizing the correlation between features vectors belonging to different class.

C.-L. Liu, C. Zhang, and L. Wang (Eds.): CCPR 2012, CCIS 321, pp. 56–63, 2012.

Locality-preserving methods which can discover the low dimensional manifold structure embedded in the original high dimensional space have been proposed to deal with nonlinear problem. In recent years, remarkable results have been achieved in nonlinear dimensionality reduction by locality based methods. Typical methods include local principal component analysis [7], locally linear embedding (LLE) [8], and locality preserving projection (LPP) [9], etc. Following similar idea, Sun and Chen proposed a novel locality preserving method for CCA, called LPCCA [10] which can incorporate local structure information into CCA for multi-view data. It has been shown that LPCCA performs better than CCA in some classification tasks. However, CCA and LPCCA only concern the correlation of pair-wise samples in different views, and they are not designed to use the class information of samples, which is very important for discrimination.

In this paper, inspired by the recent success of sparse representation in dimensionality reduction [11][12], we propose a new method called Sparsity Preserving Canonical Correlation Analysis (SPCCA). It's worth noting that our proposed method is different from the one in [12], which only constrains the sparse reconstructive relationship features in the dimension reduced space. In contrast, our method considers not only the correlations between two views from the same sample, but also the cross correlations between two views respectively from different within-class samples, which are determined by performing sparse representation. We compare the proposed SPCCA method with both existing multi-view dimensionality reduction methods in classification tasks. The experimental results on a series of multi-view data sets validate the effectiveness of the proposed method.

The rest of this paper is organized as follows: we derive the proposed method SPCCA in Section 2. In Section 3, we present the experiments to validate the performance of various methods. Finally, we conclude this paper in Section 4.

2 Proposed Method

2.1 Sparse Representation

Given a set of samples $\{\mathbf{x}_i\}_{i=1}^n$, where $\mathbf{x}_i \in \mathbf{R}^m$, let $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n] \in \mathbf{R}^{m \times n}$. For each $\mathbf{x}_i$, we first find a sparse reconstructive weight vector $\mathbf{s}_i$ by the following modified ℓ_1 minimization problem [11]:

$$\begin{aligned} \min_{\mathbf{s}_i} \quad & \|\mathbf{s}_i\|_1 \\ \text{s.t.} \quad & \mathbf{x}_i = \mathbf{X}\mathbf{s}_i \\ & 1 = \mathbf{1}^T \mathbf{s}_i \end{aligned} \tag{1}$$

Here $\mathbf{s}_i = [s_{i1}, \ldots, s_{i,j-1}, 0, s_{i,i+1}, \ldots, s_{in}]^T$ is an vector where the i-th element is zero for removing $\mathbf{x}_i$ from $\mathbf{X}$. $\mathbf{1} \in \mathbf{R}^n$ is a vector of all ones.

The optimization problem of Eq.(1) can be solved by standard linear programming. When we get $\tilde{\mathbf{s}}_i$ which is the optimal solution of Eq.(1), the sparse reconstructive weight matrix $\mathbf{S} = (\tilde{s}_{ij})_{n \times n}$ can be denoted as follows:

$$\mathbf{S} = [\tilde{\mathbf{s}}_\mathbf{1}, \tilde{\mathbf{s}}_\mathbf{2}, \ldots, \tilde{\mathbf{s}}_\mathbf{n}] \tag{2}$$

The reconstructive weight matrix $\mathbf{S}$ is a $n \times n$ matrix, of which the element $\tilde{s}_{ij}$ represents the contribution of each $\mathbf{x}_j$ to reconstruct $\mathbf{x}_i$. Generally speaking, If the element $\tilde{s}_{ij}$ is bigger, the sample $\mathbf{x}_j$ is more important to reconstruct $\mathbf{x}_i$.

2.2 Sparsity Preserving CCA (SPCCA)

The optimization problem of CCA [13] can be proven to be expressed as the following equivalent form:

$$\begin{aligned} \max_{\mathbf{w}_x, \mathbf{w}_y} \quad & \mathbf{w}_x^T \cdot \sum_{i=1}^{n}\sum_{j=1}^{n}(\mathbf{x}_i - \mathbf{x}_j)(\mathbf{y}_i - \mathbf{y}_j)^T \cdot \mathbf{w}_y \\ \text{s.t.} \quad & \mathbf{w}_x^T \cdot \sum_{i=1}^{n}\sum_{j=1}^{n}(\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T \cdot \mathbf{w}_x = 1, \\ & \mathbf{w}_y^T \cdot \sum_{i=1}^{n}\sum_{j=1}^{n}(\mathbf{y}_i - \mathbf{y}_j)(\mathbf{y}_i - \mathbf{y}_j)^T \cdot \mathbf{w}_y = 1 \end{aligned} \tag{3}$$

In order to incorporate the neighborhood information into CCA, the objective function of LPCCA can be written as [10]:

$$\begin{aligned} \max_{\mathbf{w}_x, \mathbf{w}_y} \quad & \mathbf{w}_x^T \cdot \sum_{i=1}^{n}\sum_{j=1}^{n} S_{ij}^x(\mathbf{x}_i - \mathbf{x}_j) S_{ij}^y(\mathbf{y}_i - \mathbf{y}_j)^T \cdot \mathbf{w}_y \\ \text{s.t.} \quad & \mathbf{w}_x^T \cdot \sum_{i=1}^{n}\sum_{j=1}^{n} S_{ij}^x(\mathbf{x}_i - \mathbf{x}_j) S_{ij}^x(\mathbf{x}_i - \mathbf{x}_j)^T \cdot \mathbf{w}_x = 1 \\ & \mathbf{w}_y^T \cdot \sum_{i=1}^{n}\sum_{j=1}^{n} S_{ij}^y(\mathbf{y}_i - \mathbf{y}_j) S_{ij}^y(\mathbf{y}_i - \mathbf{y}_j)^T \cdot \mathbf{w}_y = 1 \end{aligned} \tag{4}$$

Here $\mathbf{S}^x = \{S_{ij}^x\}_{i,j=1}^n$, $\mathbf{S}^y = \{S_{ij}^y\}_{i,j=1}^n$ are similarity matrices. Based on LPCCA, we propose a new method called sparsity preserving canonical correlation analysis (SPCCA) which incorporates both neighborhood information and class information into the objective function. The objective function of SPCCA is given in the following form

$$\begin{aligned} \max_{\mathbf{w}_x, \mathbf{w}_y} \quad & \mathbf{w}_x^T \mathbf{X}(\mathbf{S}^{xy} + \mathbf{S}^x + \mathbf{S}^y)\mathbf{Y}^T \mathbf{w}_y \\ \text{s.t.} \quad & \mathbf{w}_x^T \mathbf{X}\mathbf{S}^{xx}\mathbf{X}^T \mathbf{w}_x = 1 \\ & \mathbf{w}_y^T \mathbf{Y}\mathbf{S}^{yy}\mathbf{Y}^T \mathbf{w}_y = 1 \end{aligned} \tag{5}$$

Where $\mathbf{S}^x \in \mathbf{R}^{n\times n}$, $\mathbf{S}^y \in \mathbf{R}^{n\times n}$ are the sparse reconstructive weight matrices defined in Eq.(2).

In order to use the class information of samples, we compute the $\mathbf{S}^x$ only with the samples having the same label. Denote $\mathbf{X}^k = [\mathbf{x}_1^k, \mathbf{x}_2^k, \ldots, \mathbf{x}_{n_k}^k]$ are samples in the k-th class. For each sample $\mathbf{x}_i^k$, we can get a reconstructive weight vector $\mathbf{p}_i^k$ by using other samples in the same class. And thus n_k vectors are obtained, from which we can get a reconstructive matrix $\mathbf{P}^k = [\mathbf{p}_1^k, \mathbf{p}_2^k, \ldots, \mathbf{p}_{n_k}^k]$. Finally, we combine the c reconstructive matrices to one matrix $\mathbf{S}^x \in \mathbf{R}^{n \times n}$. The elements S_{ij}^x are zero if $\mathbf{x}_i$ and $\mathbf{x}_j$ are in different classes. And if the samples $\mathbf{x}_i$ and $\mathbf{x}_j$ are in the same k-th class, the element S_{ij}^x is equal to the corresponding value of the element in reconstructive matrix $\mathbf{P}^k$, so as $\mathbf{S}^y$. On the other hand, $\mathbf{S}^{xy}$, $\mathbf{S}^{xx}$ and $\mathbf{S}^{yy}$ are denoted from [10], where $\mathbf{S}^{xy} = \mathbf{D}^{xy} - \mathbf{S}^x \circ \mathbf{S}^y$, $\mathbf{S}^{xx} = \mathbf{D}^{xx} - \mathbf{S}^x \circ \mathbf{S}^x$, $\mathbf{S}^{yy} = \mathbf{D}^{yy} - \mathbf{S}^y \circ \mathbf{S}^y$, the symbol $\circ$ denotes an operator such as $(\mathbf{A} \circ \mathbf{B})_{ij} = \mathbf{A}_{ij}\mathbf{B}_{ij}$ for matrices $\mathbf{A}$, $\mathbf{B}$ with the same size and $\mathbf{A}_{ij}$ denotes the ij-th entry of $\mathbf{A}$, $\mathbf{D}_{xx}(\mathbf{D}_{yy}, \mathbf{D}_{xy})$ is a diagonal matrix of size n-by-n, and its i-th diagonal entry equal to the sum of the entries in the i-th row (or the i-th column due to the symmetry) of the matrix $\mathbf{S}^x \circ \mathbf{S}^x$ ($\mathbf{S}^y \circ \mathbf{S}^y$, $\mathbf{S}^x \circ \mathbf{S}^y$). The samples $\mathbf{X}$ and $\mathbf{Y}$ do not need to be zero mean.

Similar as the strategy for solving CCA [1], we obtain the following generalized eigenvalue decomposition problem

$$\begin{bmatrix} & \tilde{\mathbf{C}}_{xy} \\ \tilde{\mathbf{C}}_{xy}^T & \end{bmatrix} \begin{bmatrix} \mathbf{w}_x \\ \mathbf{w}_y \end{bmatrix} = \lambda \begin{bmatrix} \tilde{\mathbf{C}}_{xx} & \\ & \tilde{\mathbf{C}}_{yy} \end{bmatrix} \begin{bmatrix} \mathbf{w}_x \\ \mathbf{w}_y \end{bmatrix} \tag{6}$$

Where $\tilde{\mathbf{C}}_{xy} = \mathbf{X}(\mathbf{S}^{xy} + \mathbf{S}^x + \mathbf{S}^y)\mathbf{Y}^T$, $\tilde{\mathbf{C}}_{xx} = \mathbf{X}\mathbf{S}^{xx}\mathbf{X}^T$, $\tilde{\mathbf{C}}_{yy} = \mathbf{Y}\mathbf{S}^{yy}\mathbf{Y}^T$. Eq. (6) can be computed via the singular value decomposition (SVD) technique. We include the detailed solution (pseudo-code) in Algorithm 1.

After obtaining eigenvectors $\mathbf{w}_{xi}$, $\mathbf{w}_{yi}$ corresponding to d generalized eigenvalue λ_i, $i = 1, \ldots, d$, any pair sample $(\mathbf{x}, \mathbf{y})$ can do feature combination as follows [2]:

$$\mathbf{W}_x^T x + \mathbf{W}_y^T y \tag{7}$$

$$\begin{bmatrix} \mathbf{W}_x^T x \\ \mathbf{W}_y^T y \end{bmatrix} \tag{8}$$

Where $\mathbf{W}_x = [\mathbf{w}_{x1}, \ldots, \mathbf{w}_{xd}]$, $\mathbf{W}_y = [\mathbf{w}_{y1}, \ldots, \mathbf{w}_{yd}]$. The obtained features here are called fused features, which can be used by any subsequent classifier. The two strategies to extract features are denoted as FS1 and FS2, respectively Eq.(7-8). In the rest, we calculate the accuracies of different methods on a series data sets by using FS1 and FS2 respectively. In order to show the effectiveness of the methods, we adopt the average accuracy of both fusion methods including FS1 and FS2.

Algorithm 1. SPCCA

Input : Training sets $\mathbf{X} \in \mathbf{R}^{p\times n}, \mathbf{Y} \in \mathbf{R}^{q\times n}$
Output: Projection matrices $\mathbf{W}_x, \mathbf{W}_y$

1 Construct $\mathbf{S}^x, \mathbf{S}^y$.
2 Define $\mathbf{Q} = \mathbf{S}^{xy} + \mathbf{S}^x + \mathbf{S}^y$;
3 Compute matrices $\tilde{\mathbf{C}}_{xy} = \mathbf{XQY}^T$, $\tilde{\mathbf{C}}_{xx} = \mathbf{XS}^{xx}\mathbf{X}^T$, $\tilde{\mathbf{C}}_{yy} = \mathbf{YS}^{yy}\mathbf{Y}^T$.
4 Compute matrix $\mathbf{H} = \tilde{\mathbf{C}}_{xx}^{-\frac{1}{2}}\tilde{\mathbf{C}}_{xy}\tilde{\mathbf{C}}_{yy}^{-\frac{1}{2}}$;
5 Perform SVD decomposition on $\mathbf{H}$: $\mathbf{H} = \mathbf{UDV}^T$;
6 Choose $\mathbf{U} = [U_1...U_d]$,$\mathbf{V} = [V_1...V_d]$,$d < n$;
7 Obtain $\mathbf{W}_x = \tilde{\mathbf{C}}_{xx}^{-\frac{1}{2}}\mathbf{U}$,$\mathbf{W}_y = \tilde{\mathbf{C}}_{yy}^{-\frac{1}{2}}\mathbf{V}$;

3 Experiments

In this section, we evaluate the performance of the proposed SPCCA algorithm on Multiple Features data set picked out from UCI repository[1] as well as ORL[2] face databases. For performance evaluation, we adopt the nearest neighbor classifier.

3.1 Multiple Feature Data Set

The first data set used is Multiple Features data set, which consists of 2000 examples of 10 handwritten digits ('0-9') with six set of features, 200 examples for each digit. The 6 feature sets include fou (fourier coefficients, 76 features), fac (profile correlations, 216 features), kar (Karhunen-Love coefficients, 64 features), pix(pixel averages in 2x3 windows, 240 features), zer (Zernike moments, 47 features), mor (morphological features, 6 features).

Any two of them can be used as two working views. So there are totally fifteen different combinations. For each combination, we randomly choose 1000 samples in each class for training, the other 1000 samples are left for testing. Thus there are 100 training examples and 100 testing examples for each class. First, we extract features using CCA, LPCCA, DCCA, SPCCA algorithms. Then we perform classification based on the extracted features using nearest neighbor classifier.

Table 1 shows the accuracies of different algorithms on Multiple Features data set, where bold denotes the highest value at that data set. It can be seen from Table 1 that in nearly all cases SPCCA achieves the best performances compared with CCA, LPCCA and DCCA. DCCA is the best on only 2 cases, where both contain mor features. The mor features are of only six dimensions, which is too hard for SPCCA to get enough information for constructing weight matrix correctly. However, even in those two cases SPCCA is still consistently better than CCA and LPCCA.

[1] `http://archive.ics.uci.edu/ml/`
[2] `http://www.cs.uiuc.edu/~dengcai2/Data/FaceData.html`

In order to study the effect of subspace dimensionality on performances of different algorithms, accuracies of different algorithms w.r.t different subspace dimensions are shown in Fig.1. Three combinations are picked, which cover all of the six views in Multiple Feature data set. Fig.1 indicates that SPCCA achieves better accuracies than other methods at almost all dimensions. On the other hand, we notice that with more subspace dimensions the results of LPCCA and SPCCA are usually better, which is different from CCA and DCCA. By simultaneously considering the class and the local information, SPCCA almost gets the best results at all dimensions.

Table 1. Classification accuracy on multiple feature database

X	Y	CCA	LPCCA	DCCA	SPCCA
fac	fou	0.8159	0.9600	0.8458	**0.9842**
fac	kar	0.9392	0.9622	0.9376	**0.9815**
fac	mor	0.7647	0.8080	0.9181	**0.9635**
fac	pix	0.8385	0.9520	0.8260	**0.9782**
fac	zer	0.8556	0.9417	0.9259	**0.9820**
fou	kar	0.8959	0.9635	0.9008	**0.9757**
fou	mor	0.7608	0.7592	**0.8141**	0.8127
fou	pix	0.7401	0.8890	0.7631	**0.9762**
fou	zer	0.8045	0.8467	0.8021	**0.8567**
kar	mor	0.8014	0.8535	0.9051	**0.9282**
kar	pix	0.9180	0.9722	0.8709	**0.9762**
kar	zer	0.9163	0.9695	0.9324	**0.9760**
mor	pix	0.7318	0.7097	0.8945	**0.9440**
mor	zer	0.7307	0.6922	**0.8020**	0.7872
pix	zer	0.8242	0.9482	0.8766	**0.9727**

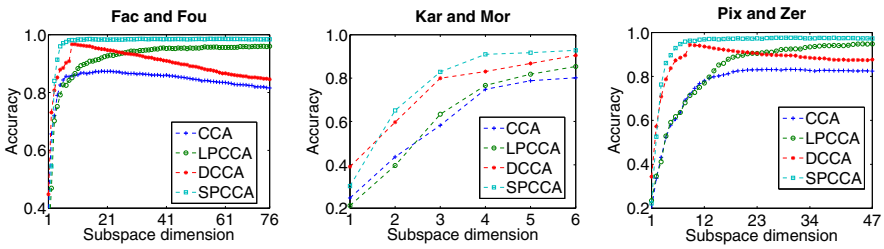

Fig. 1. Recognition accuracies of all methods with different subspace dimensions on parts of Multiple Feature data set

3.2 Face Recognition Data Set

We use ORL face database in face recognition experiments. ORL, also called AT&T database consists of 400 images of 40 person, 10 images for each person. These images are photographed in different times, with changing lightning, facial expressions. Each image is size of 112x92 in 256-gray scale.

The ORL data set has been preprocessed. We first resize each original image to size 64x64 and 32x32, respectively, and then we perform LBP and wavelet methods on 64x64 size of images. We get 4 different views at all. Because original images are easy to get and the other views can be calculated from the original images, the original images are always taken as the first view in the experiments, and the other view are taken in turn as the second view. Thus there are 3 combinations. We randomly partitioned ORL into equal size training and testing sets. There are 10 runs and the averaged result is reported. Because the original features of ORL data sets are very huge, for computational issues such as accelerate computing speed and avoiding out of memory in program running, we first reduce their dimension with PCA before they are directly used in experiments. In order to preserve relevant information, the cumulative eigenvalue ratio(we call it 'energy ratio for short) r is set 0.95.

Table 2 gives the classification results of different algorithms on ORL database. As can be seen from Table 2, SPCCA achieves better performances than CCA and LPCCA on ORL database. and LPCCA is even inferior to CCA in this data set. Also Table 2 shows that SPCCA also outperforms traditional dimensionality reduction methods PCA and LDA. As in Section 3.1 on Multiple Features data set, we also plot the effects of the number of algorithms, and the corresponding results are shown in Fig.2. which again validate the effectiveness of the proposed method SPCCA.

Table 2. Classification accuracy on ORL databases

X	Y	CCA	LPCCA	PCA	LDA	SPCCA
ORL(64x64)	ORL(32x32)	0.8690	0.8250	0.7270	0.6885	**0.9525**
ORL(64x64)	ORL(wav)	0.8845	0.8410	0.7487	0.5992	**0.9500**
ORL(64x64)	ORL(lbp)	0.9197	0.8552	0.8380	0.8985	**0.9662**

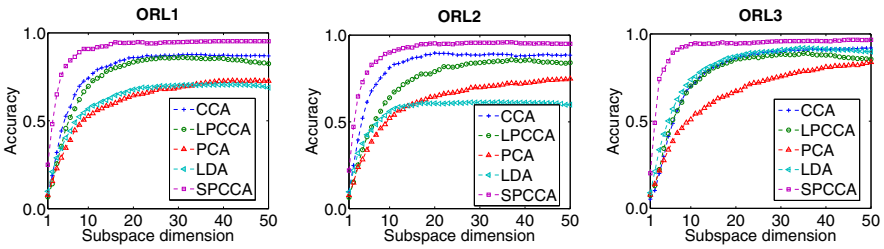

Fig. 2. Recognition accuracies of all methods with different subspace dimensions on ORL data set

4 Conclusion

In this paper, we propose a sparsity preserving CCA method called SPCCA for multi-view dimensionality reduction. Different from traditional CCA, SPCCA considers not only correlations between sample pairs, but also uses cross-correlations between different samples in the same class, by performing sparse representation on two views. The experimental results on a series of data sets show that our method can effectively improve the classification performance compared with conventional CCA based methods. In the future, we will use SPCCA for other tasks such as data visualization, clusting, etc. Also we will introduce the semi-supervised information into our method for further performance improvement.

References

1. Hotelling, H.: Relations Between Two Sets of Variates. Biometrica 28, 322–377 (1936)
2. Sun, Q., Zeng, S., Liu, Y., Heng, P., Xia, D.: A new method of feature fusion and its application in image recognition. Pattern Recognition 38, 2437–2448 (2005)
3. Hel-Or, Y.: The Cannoical Correlations of Color Images and their use for Demosaicing. Technical report, HP Laboratories Israel (2004)
4. Hardoon, D.R., Szedmák, S., Shawe-Taylor, J.: Canonical Correlation Analysis: An Overview with Application to Learning Methods. Neural Computation 16, 2639–2664 (2004)
5. Abraham, B., Merola, G.: Dimensionality reduction approach to multivariate prediction. Computational Statistics & Data Analysis 48, 5–16 (2005)
6. Sun, T., Chen, S., Yang, J., Shi, P.: A Novel Method of Combined Feature Extraction for Recognition. In: Proceedings of the International Conference on Data Mining, pp. 1043–1048. IEEE Press, Piscataway (2008)
7. Kambhatla, N., Leen, T.K.: Dimension Reduction by Local Principal Component Analysis. Neural Computation 9, 1493–1516 (1997)
8. Roweis, S.T., Saul, L.K.: Nonlinear Dimensionality Reduction by Locally Linear Embedding. Science 290, 2323–2326 (2000)
9. He, X., Niyogi, P.: Locality Preserving Projections. In: 17th Annual Conference on Neural Information Processing Systems, pp. 153–160. MIT Press, Cambridge (2004)
10. Sun, T., Chen, S.: Locality preserving CCA with applications to data visualization and pose estimation. Image and Vision Computing 25, 531–543 (2007)
11. Qiao, L., Chen, S., Tan, X.: Sparsity preserving projections with applications to face recognition. Pattern Recognition 43, 331–341 (2010)
12. Hou, S., Sun, Q.: Sparsity Preserving Canonical Correlation Analysis with Application in Feature Fusion. Acta Automatica Sinica 38, 659–665 (2012)
13. Hoegaerts, L., Suykens, J.A.K., Vandewalle, J., Moor, B.D.: Subset based least squares subspace regression in RKHS. Neurocomputing 63, 293–323 (2005)