ROBUST PRINCIPAL CURVES BASED ON MAXIMUM CORRENTROPY CRITERION

CHUN-GUO LI[1,2], BAO-GANG HU[1]

[1]NLPR/LIAMA, Institute of Automation,Chinese Academy of Sciences, Beijing, P.R.China, 100190
[2]Machine Learning Center, Faculty of Mathematics and Computer Science, Hebei University, Baoding 071002, China
E-MAIL: cgli@nlpr.ia.ac.cn, hubaogang@gmail.com

Abstract:

Principal curves are curves which pass throught the 'middle' of a data cloud. They are sensitive to variances of data clouds. In this paper, we propose a robust principal curve model - Correntropy based Principal Curve (CPC) model, based on maximum correntropy criterion (MCC). CPC model approximate the principal curve with k-segments polygonal line. Employing the half-quadratic technique, CPC model are optimized in a iteratively way. CPC model are insensitive to variances and outliers of data clouds. Extensive experiments on sythetic and real-life datasets illustrate the robustness of CPC model in learning principal curves.

Keywords:

Pringcipal curves; MCC; ITL; Polygonal line

1. Introduction

Principal curves (PCs) are one of nonlinear extension of principal component analysis (PCA). PCA is known as a linear technology to find the mainstreams of given data clouds. It plays the role of linear mapping in original data space [1, 2]. Kernel principal components analysis (KPCA) employ kernel trick [3] to deal with nonlinear problem, but KPCA share the same motivation with linear PCA. [4] studied the principle of the extension of linear regression to nonlinear regression and defined PCs as 'curves' passing through the centers of data clouds, instead of 'lines' in PCA. This definition extended the very intrinsic motivation of PCA technology. Based on this definition, many algorithms came forth to learn the best PCs for given data clouds. [4] employed EM algorithm to find the key points and interpolate them to smooth the connecting polygonal line. But the nonparameteric methods and assumptiongs of smoothness and differentiability lead to a series of problems, such as bias, inefficiency, non-uniqueness, and no convergency [5]. Afterwards, researchers devoted themselves to seeking for resolutions to those problems [6–12]. After all, they share one drawbacks that the produced PCs are not located in the 'middle' of those data clouds with big variances. An outlier might make the PCs deviate from the middle of a data cloud.

Information theoretic learning (ITL) was initiated in the late 1990s. It has a theoretical foundation in machine learning. ITL uses information-theoretic descriptors to substitute the conventional statistical descriptors of variance and covariance [13]. In ITL, Maximum correntropy criterion (MCC) has the robust learning property [14,15], and hence MCC is applied broadly in learning problems with outliers and with big variances [16,17]. In [17], adopting MCC as the objective function for robust PCA, authors proposed HQ-PCA to improve face recognition accuracy.

Reminding that PCs are nonlinear extensions of PCA, we propose a correntropy-based principal curve (CPC) model which adopts MCC to measure the fitness of a curve passing through the center of a data cloud. A k-segments polygonal line [9,10] is taken to approximate the latent PC. Half-quadratic technique [18] is used to transform the optimization problem into an iterative optimizing problem. K-segments CPC learning algorithm is designed and carried out in synthetic data sets. To state the robust learning property of the proposed algorithm, we compare it with soft k-segments PC learning algorithm. The main contributions of this paper are as follows: (1) by introducing correntropy into PC definitions, CPC model is more robust in learning PCs for data clouds with large variances; (2) half-quadratic technology helps to accelerate the convergence rate in an iterative way. K-segments CPC algorithm also has several interesting perspectives. On the one hand, it utilizes half-quadratic technique to make the optimization problem more facile. On the other hand, outliers impact little on recovering

the generating curve which makes the proposed model much more robust.

The rest of this paper is organized as follows. First, PC definitions and MCC are introduced as related works in Section 2. In Section 3, we propose CPC model and k-segments approximation, and derive optimizing procedure via half-quadratic technique. K-segments CPC learning algorithm is also stated in Section 3. Experimental results illustrate the robust performance of the proposed algorithm in Section 4, prior to summary of this paper in Section 5.

2. Related Works

Recently, most researchers turned to the methodologies of PC applications, e.g. [12, 22, 23]. But applications would become facile with the developments of PCs learning algorithms. This section reviewed the developments of PCs first. Then MCC in ITL is introduced. Applications of MCC in PCA are described briefly in the end of this section.

2.1. Principal Curve Definitions and Learning Algorithms

PCA is originally defined as the orthogonal projection into a lower dimensional linear space, in which the projected data explain the original variance the most [1]. Equivalently, PCA is also defined as the linear projection which minimizes the average projection cost defined as reconstruction errors [2].

Given a data set $X = (x_1, x_1, \cdots, x_n)$, where $x_i \in R^d$, PCA can be mathematically described as solving the following optimization problem:

$$\min_{\mu,U} \sum_{i=1}^{n} \parallel x_i - \mu - U\lambda_i \parallel^2 \tag{1}$$

where $U \in R^{d\times m}(d < m)$ is a projection matrix which constitutes the bases of a m-dimensional subspace, μ is the center of the data in the m-dimensional space, $\lambda_i = U^T(x_i - \mu)$ is the principal component of the i-th point under the projection matrix, $\parallel \cdot \parallel$ is the L2 norm. Specifically, d equals 1 for the first PCA, which means that a line $\mu + U\lambda$ would minimize the objective function. The probabilistic formulation of PCA was proposed independently by [26] and [25].

Principal curves (PCs) are defined as smooth curves passing through the middle of a multidimensional data set [4]. They are nonlinear generalizations of the first principal component analysis. [4] were the pioneers in PCs learning. They analogically studied the motivations of PCA and linear regression. They found that nonlinear regression generalized linear regression and another parallel definition should be given to generalize the linear problem PCA. Subsequently, they proposed the earliest PC definition – HSPC [5]. HSPC substitutes a curve expression $x = f(\lambda) + e$ for a line $x = \mu + U\lambda + e$ in PCA definitions. MSE is taken as the cost function to be minimized. Mathematically, to find a curve f means to find the solution of the following problem:

$$\begin{aligned} \min_{f} \quad & \textstyle\sum_{i=1}^{n} \parallel x_i - f(\lambda_i) \parallel^2 \\ s.t. \quad & \parallel f' \parallel \equiv 1 \end{aligned} \tag{2}$$

where $\lambda_i = \lambda_f(x_i)$ and $\parallel f' \parallel \equiv 1$ require a smooth curve as the optimal solution. HSPC employs the EM algorithm to learn PCs and initializes it with PCA. In E-step, conditional expectations are calculated to find the centers for those points that project into a small area of λ_i. Scatterplot smoother is used to find a curve along those centers. In M-step, all the points are projected onto the new curve and λ_i are updated. E-step and M-step take place by turn until stopping criterions are met. HSPC open a new window for nonlinear extension of linear PCA, although the original defintion bring forth lots of problems which attract researcher to seek for resolutions, such as [8]. [10] proposed soft k-segments algorithm which divided PC into k-segments polygonal line. Each segment is learned by PCA in a defined Voronoi region. Polygonal line tends to be nonsmooth due to segments connection, which goes against the smoothness of PC definitions.

2.2. HQ-PCA

Recently, the concept of correntropy [15] was proposed for ITL. Correntropy is defined as a generalized similarity measure between two random variables A and B [14]:

$$V_\sigma(A, B) = E(k_\sigma(A - B)) \tag{3}$$

where $k_\sigma(\cdot)$ is the kernel function which is Gaussian function in this paper, and $E(\cdot)$ denotes the mathematical expectation. When observations of (A, B) are obtained, $\{(a_1, b_1), (a_2, b_2), \cdots, (a_n, b_n)\}$, correntropy can be estimated by:

$$\tilde{V}_\sigma(A, B) = \frac{1}{n}\sum_{i=1}^{n} k_\sigma(a_i - b_i). \tag{4}$$

The maximum of error correntropy in Eq. (4) is called maximum correntropy criterion (MCC). MCC is equivalent to robust least square criterion [13], but MCC has a clear theoretical foundation. Different from MSE, MCC employs kernel techniques to perform a nonlinear projection from the input space

to a higher dimensional space. MCC emphasizes the difference between A and B along the line $A = B$ and exponentially attenuates contributions away from the line. This characteristic overcomes the shortcomings of MSE that values away from the $A = B$ line produces the quadratic increase of error due to the second moment. It makes the correntropy more robust than MSE, especially for distributions with outliers and nonzero-mean.

MCC became very popular in machine learning since its proposal [16, 17]. Based on MCC, [17] proposed robust PCA which he named HQ-PCA with the use of half-quadratic technique in the optimization problem. HQ-PCA did not assume the Gaussian noise distribution and could learn PCA for non-Gaussian signal with outliers. The optimization problem of HQ-PCA is described mathematically as

$$\max_{\mu,U} \frac{1}{n} \sum_{i=1}^{n} g(x_i - \mu - U\lambda_i) \tag{5}$$

where $g(\cdot)$ was the Gaussian kernel function. Half-quadratic technique made the algorithm as a form of weighted PCA. Hence the algorithm can be solved iteratively and convergence rate is accelerated. What is more, the local property of MCC makes HQ-PCA more robust for outliers in the training data sets than MSE objective function.

3. K-Segments CPC Algorithm

Considering correntropy can efficiently cope with outliers and noise, we introduce correntropy criterion into PC learning. Employing the half-quadratic technique, the optimization problem is translated into iterative optimizing procedure – k-segments CPC algorithm. In this section, we follow the notations in Section 2.

3.1. K-Segments Correntropy-based Principal Curve

Reminding the PC motivation, it would be ideal that a curve can thread through all the points when any noise is absent. However, a variety of noise is always present in observing data. Therefore the second best thing is to get a curve passing through the center of the data cloud. To generalizing the HQ-PCA definition, we take MCC instead of MSE criterion in PCs model. Then, PC optimization problem is reformulated as CPC model:

$$\max_{f} \frac{1}{n} \sum_{i=1}^{n} g(x_i - f(\lambda_i)) \tag{6}$$

where $g(\cdot) = \exp(-\frac{\|\cdot\|^2}{2\sigma^2})$. Eq. (6) is defined as a generalized problem of HQ-PCA. When f is a linear function $f(\lambda_i) = \mu + U\lambda_i$, the problem degenerates into HQ-PCA. Otherwise, the problem is a PC learning problem.

Remembering the soft k-segments PC learning algorithm [9, 10], we approximate a PC with a polygonal line consisting of k segments. We divide the data space into k Voronoi Regions (VRs) and find the shortest segment of the optimal line v_i in each VR V_i. For the line v_i in the i-th VR, the orthogonal distance of point x to the line can be denoted by $x - \lambda v_i$, where λ is the projection coordinate under the linear projection $v_i^T x$. Then CPC model (6) can be simplified into

$$\max_{v} J(v) = \max_{v} \sum_{i=1}^{k} \sum_{x \in V_i} g(x_i - \lambda v_i) \tag{7}$$

$$= \max_{v} \sum_{i=1}^{k} J_i \tag{8}$$

where $v = (v_1, v_2, \cdots, v_k)$ is segment matrices in VRs and the scaling factor $\frac{1}{n}$ is neglected.

We claim that the optimal curve of Eq. (15) would be a local solution. Firstly, CPC model works independently with each pair of a sample and its estimated point. Secondly, the model emphasizes the errors produced near the line $A = B$ and exponentially shrink the errors away the line. Lastly, k-segments polygonal line is used to approximate the curve. Each segment would have the optimal orientation in the current VR.

To address Eq. (15), we employ the half-quadratic technique which is often used to solve nonlinear optimization problems in ITL [19–21]. Based on the convex conjugate function [18], one can easily derive the following proposition.

Proposition 1 *There exists a convex conjugate function φ of $g(y)$ such that*

$$g(y) = \max_{p}(p\frac{\| y \|^2}{\sigma^2} - \varphi(p)) \tag{9}$$

And for a fixed y, the maximum is reached at $p = -g(y)$.

Substituting Eq. (9) into Eq. (15), we get an augmented objective function in an increased dimensional parameter space:

$$\max_{v} J(v) = \max_{p,v} \tilde{J}(v, p)$$

$$= \max_{p,v} \sum_{i=1}^{k} \sum_{x \in V_i} (p\frac{\| x - \lambda v_i \|^2}{\sigma_i^2} - \varphi(p)) \tag{10}$$

where

$$\begin{aligned} p &= -g(\| x - \lambda v_i \|) \\ \lambda &= v_i^T x \\ \sigma_i^2 &= \frac{1}{s_i} \sum_{x \in V_i} \| x - \lambda v_i \|^2 \end{aligned}$$

p and λ are accompanying with x, and s_i is the number of points in the i-th VR. According to Proposition 1, for a fixed v, the following equation holds:

$$J(v) = \max_p \tilde{J}(v, p) \tag{11}$$

It follows that

$$\max_v J(v) = \max_{v,p} \tilde{J}(v, p). \tag{12}$$

i.e. maximizing $J(v)$ is identical to maximizing its augmented function $\tilde{J}(v, p)$. For the optimal solution $p = -g(y)$, $J(v)$ and $\tilde{J}(v, p)$ achieve the same maximum point at the same optimal solution v.

Recalling the optimal solution of $\tilde{J}(v, p)$ given v, we can easily calculate a maximizer (v, p) in an alternative maximizing way:

$$p^{(t+1)} = -g(x - \lambda v_i^{(t)}) \tag{13}$$

$$v_i^{(t+1)} = \arg\max_{v_i} \sum_{i=1}^{k} \sum_{x \in V_i} \frac{\| x - \lambda^{(t)} v_i \|^2 p^{(t+1)}}{(\sigma_i^2)^{(t)}} \tag{14}$$

Here we neglect the item of $\varphi(p)$ due to its consistence for different v for given p. The maximizing items in Eq. (14) are quadratic programming problems (QPs) with no constraints. The optimization can be achieved via setting differential formulae with v_i to zeros, and we get an explicit solution to the above QP:

$$v_i = \sum_{\hat{x} \in V_i} \frac{\hat{\lambda}}{\sum_{x \in V_i} \hat{\lambda}^2} \hat{x}, i = 1, 2, \cdots, k \tag{15}$$

3.2. k-Segments CPC Algorithm

K-segments CPC algorithm summarizes the optimizing procedure of Eq. (7). In Step 6, the kernel size (bandwidth) is computed as in Eq. (10). In Step 8, the length of the line should suit the range of the current VR. Furthermore, additional algorithm is needed to connect the outcome segments to form a polygonal line. For comparison reason, we adopt travel salesman optimization as in paper [9, 10]. All the segments are connected with the shortest length to be a polygonal line.

Algorithm 1 k-Segments CPC Algorithm

Input:
X: data matrix
k: number of segments
ε: a small positive value of threshold for increased entropy

Output: $v = (v_1, v_2, \cdots, v_k)$: segments inserted finally
λ: projected coordinates in the principal curve

1: Find the first PC of all the data
2: **while** change of correntropy is bigger than ε_1, or segments number is smaller than k **do**
3: Find the points that can increase the correntropy most to insert a segment;
4: Redivide the VRs according to the orthogonal distance of the points to k segments;
5: Find the first PC in each VR;
6: Compute $p = -g(x - \lambda v_i)$ in each VR;
7: Replace x with $\hat{x} = x\sqrt{-\frac{p}{\sigma^2}}$ and λ with $\hat{\lambda} = \lambda\sqrt{-\frac{p}{\sigma^2}}$;
8: Get the optimal line direction via Eq. (15);
9: **end while**

Comparing to soft k-segments algorithm in paper [9, 10], k-segments CPC algorithm removes the estimation bias although it still has a model bias. The original soft k-segments algorithm estimated the PDF of the latent variable x. Although k-segments CPC algorithm has a model bias just the same as original soft k-segments algorithm, it does not require the distribution of the data set. Furthermore, k-segments CPC algorithm has the advantage of robustness to outliers and noise with big variances. The original k-segments algorithm cannot produce a proper polygonal line when outliers are present while k-segments CPC algorithm can. Experimental results in the following section will illustrate the advantage.

4. Experiment Results with Synthetic Data Sets

To verify the performance of k-segments CPC algorithm, we carry out groups of numerical experiments on synthetic data and real world data. Synthetic data with big variance and with outliers are drawn similar to [10]. Real world data are downloaded from GAPMINDER (http://www.gapminder.org/) which are the world statistical website.

4.1. Data Sets with Normal Noise and with Outliers

Figure 1 shows the better performance of k-segments CPC algorithm than soft k-segments algorithm. We present the results with the biggest variances that can mainly keep the shapes

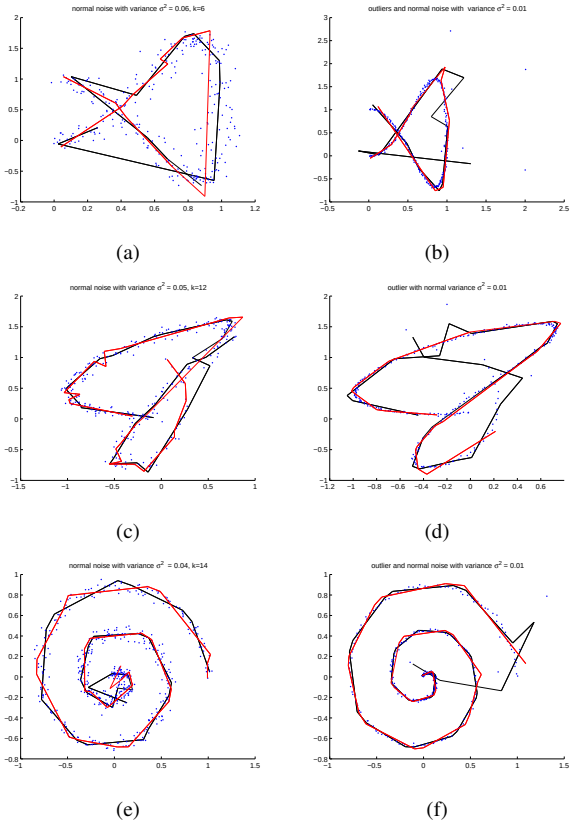

Figure 1. Principal curves constructed on data sets with normal noise of big variances (left side) and with outliers (right side) by soft k-segments (in black) and k-segments CPC (in red).

of generating curves. When variances are bigger, the basic shapes of generating curves are lost. The biggest noise variances which can mainly keep the shapes of generating curves are called the preserving noise variances (PNVs for short). For example, in Figure 1, the curve with one cross point (the first row) can still preserve its basic shape with noise variance 0.06, while the spiral curve (the third row) preserves its shape with noise variance no more than 0.04. From Figure 1, one can see that k-segments CPC algorithm can recover the structure of datasets, especially for subfigure (a) which the structure is distorted by soft k-segments algorithm.

We also take experiments on datasets with outliers and experimental results are shown in Figure 1. We plug in outliers with number of 2 percent of number of samples. We add normal noise with small variance ($\sigma^2 = 0.01$) to generating curve, and mix in outliers to produce the final data sets. The other parameters are set to the same as that no outliers are present. Soft k-segment algorithm tends to fit those outliers which deviate away greatly from the generating curve. Therefore, it fails to approximate the generating curves properly. However, k-segments CPC algorithm refrains from misleading by virtue of kernel function. Outliers affect little for the proposed algorithm to retrieve the generating curve from the given data sets. The approximating performance changes less for k-segments CPC algorithm than soft k-segments. Experimental results with outliers enforce the statements that k-segments CPC algorithms promises a robust PC learning.

4.2. Real world Datasets

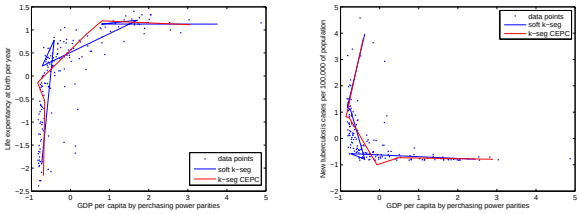

Figure 2. Principal curves constructed on quality of life index by soft k-segments (in black) and k-segments CPC (in red).

Real-life datasets are with big variances in common. Gapminder website collects the most important data that shows the world's most important trends. We study the relationship of GDP with life expentancy and new tuberculosis. From the distribution of data points in 2, we could see that GDP is monotone with each indicator. k-segments CPC algorithm can profile the main trends while soft k-segments algorithm produce a circuit, which is unreasonable.

5. Conclusions

Informatiion theoretic learning becomes popular in machine learning and pattern recognition. It considers the higher statistical moments of the datasets. Thus ITL provides more information of datasets for data processing. Employing correntropy, this paper propose a robust principal curve learning algorithm - k-segments CPC algorithm. The algorithm can retrieve the latent structure of datasets, even with big variances or outliers.

Experiments on synthetic datasets and real-life datasets illustrate the robustness of k-segments CPC algorithm. What is more, the polygonal lines should be smoothed further which is our future work.

Acknowledgements

The authors would like to thank the advice coming from machine learning crew in NLPR. This work is supported in part by NSFC (No. 61075051 and No. 61273196), the Natural Science Foundation of Hebei Province (No. F2011201063), the Science and Technology Research and Development Orientation Project of Baoding City (No. 12ZG005).

References

[1] H. Hotelling, "Analysis of a complex of statistical variables into principal components", J. of Educ. Psychol, vol. 24, pp. 417-441, 1933.

[2] K. Pearson, "On lines and planes of closes fit to systems of points in space", The Lond. Edinb. and Dublin Philos. Mag. and J. of Sci. Sixth Ser., vol. 2, pp. 559-572, 1901.

[3] C. Bishop, Pattern Recognition and Machine Learning, New York: Springer, 2006.

[4] T. Hastie, W. Stuetzle, "Principal Curves", J. of the Am. Stat. Assoc., vol. 84, no. 406, pp. 502-516, 1989.

[5] J. Zhang, J. Wang, "An Overview of Principal Curves", Chin. J. of Comput., vol. 26, no. 3, pp. 1-18, 2003.

[6] T. Duchamp, W. Stuetzle, "Geometric properties of principal curves in the plane", Lect. Notes in Statistics, vol. 109, pp. 135-152, 1996.

[7] B. Kégl, A. Krzyżak, T. Linder, K. Zeger, "Learning and Design of Principal Curves", IEEE Trans. on Pattern Anal. and Mach. Intell., vol. 22, no. 3, pp. 281-297, 2000.

[8] R. Tibshirani, "Principal Curves Revisited", Stat. and Comput., vol. 2, pp. 183-190, 1992.

[9] J.J. Verbeek, N. Vlassis, B. Kröse, "A k-segments algorithm for finding principal curves", Pattern Recognit. Lett., vol. 23, no. 8, pp. 1009-1017, 2002.

[10] J.J. Verbeek, N. Vlassis, B. Kröse, "A Soft k-segments Algorithm for Principal Curves", Lect. Notes in Comput. Sci., vol. 2130, pp. 450-456, 2001.

[11] H. Wang, T.C.M. Lee, "Automatic Parameter Selection for a k-segments Algorithm for Computing Principal Curves", Pattern Recognit. Lett., vol. 27, pp. 1142-1150, 2006.

[12] H. Wang, T.C.M. Lee, "Extraction of Curvilinear Features from Noisy Point Patterns Using Principal Curves", Pattern Recognit. Lett., vol. 29, pp. 2078-2084, 2008.

[13] J.C. Principe, Information Theoretic Learning: Renyis Entropy and Kernel Perspectives, New York: Springer, 2010.

[14] W. Liu, P.P. Pokharel, J.C. Principe, "Correntropy: Properties and applications in non-Gaussian signal processing", IEEE Trans. on Signal Process., vol. 55, no. 11, pp. 5286-5298, 2007.

[15] I. Santamaria, P. Pokharel, J.C. Principe, "Generalized correlation function: Definition, Properties and Application to Blind Equalization", IEEE Trans. on Signal Process., vol. 54, no. 6, pp. 2187-2197, 2006.

[16] R. He, W. Zheng, B. Hu," Maximum correntropy criterion for robust face recognition", IEEE Trans. on Pattern Anal. and Mach. Intell., vol. 33, no. 8, pp. 1561-1576, 2011.

[17] R. He, B. Hu, W. Zheng, X. Kong, "Robust Principal Component Analysis Based on Maximum Correntropy Criterion", IEEE Trans. on Image Process, vol. 20, no. 6, pp. 1485-1494, 2011.

[18] R. Rockfellar, Convex Analysis, New Jersery: Princetion University Press, 1970.

[19] R. Dahyot, P. Charbonnier, F. Heitz, "Robust visual recognition of colour images". Proceedings of CVPR, pp. 1685-1690, 2000.

[20] X. Yuan, B. Hu, "Robust feature extraction via information theoretic learning", Int. Conf. on Mach. Learn., Montreal, Canada, pp. 1193-1200, 2009.

[21] R. He, B. Hu, W. Zheng, Y. Guo, "Two-stage sparse representation for robust recognition on large-scale database", Proc. of AAAI, pp. 1-6, 2010.

[22] D. Miao, Q. Tang, W. Fu, "Fingerprint minutiae Extraction Based on Principal Curves", Pattern Recognit. Lett., vol. 28, pp. 2184-2189, 2007.

[23] J. Gibert, E. Valveny, H. Bunke, "Feature Selection on Node Statistics Based Embedding of Graphs", Pattern Recognit. Lett., vol. 33, pp. 1980-1990, 2012.

[24] R. He, B. Hu, X. Yuan, W. Zheng, "Principal component analysis based on non-parametric maximum entropy", Neurocomputing, vol. 73, pp. 840-1852, 2010.

[25] S. Roweis, "EM algorithms for PCA and SPCA", Adv. in Neural Inf. Process. Syst., vol. 10, pp. 626-632, 1998.

[26] M.E. Tipping, C.M. Bishop, "Probabilistic principal component analysis", J. of the Royal Stat. Soc. Series B., vol. 21, no. 3, pp. 611-622, 1999.