PRINCIPAL GRAPHS AND PIECEWISE LINEAR SUBSPACE CONSTRAINED MEAN-SHIFT

Umut Ozertem, Deniz Erdogmus

Computer Science and Electrical Engineering Department,
Oregon Health and Science University,
Portland, OR, 97329 USA

ABSTRACT

Principal curves have been defined as self-consistent smooth curves that pass through the middle of data. One of the important problems with most existing principal curve algorithms is that they are seeking for a *smooth curve*. In reality, data may take complicated shapes, which may include loops, self-intersections, and and bifurcation points; hence, a smooth curve passing through the data may not be a good representor of the data. Generally, there is, in fact, a principal graph, a collection of smooth curves that represents the dataset. We propose a nonparametric principal graph algorithm, and apply it to optical character recognition, where handling the above mentioned irregularities like loops and self-intersections is a serious problem that appear in many characters.

1. INTRODUCTION

By definition, the principal line, namely the first principal component, is the best linear representor of the data in projected mean square error sense. Principal curves stem from the reinterpretation of the principal line and the definition of self consistency. The term self-consistency was introduced by Hastie and Stuetzle [1] to describe the property that each point on a smooth curve or surface is the mean of all points that project orthogonally onto it. Using the definition of self consistency, they provide a reinterpretation of the principal line. Since every point on the principal line is the expected value of the points that orthogonally project onto this point, the principal line is self consistent, and they generalize this property by defining principal curves as *self-consistent smooth curves passing through the middle of data* [1]. Based on the self-consistency criterion that they define, starting from the first principal component, they develop an iterative algorithm to find the principal curve. However, there is no proof of convergence for Hastie's algorithm, which makes the theoretical analysis impossible. It should also be noted that this definition of the principal curve requires the the principal curve not to intersect itself, which is quite restrictive.

It is probably safe to say that Hastie and Stuetzle's local conditional expectation based definition has been the basis behind almost all principal curve algorithms so far. There are various algorithms in the literature, based on principal curve definitions that are derivatives of the original Hastie-Stuetzle definition. Tibshirani approaches the problem from a mixture models point of view, and provides an algorithm that uses expectation maximization [2]. Sandilya and Kulkarni provide a regularized version of Hastie's definition by constraining bounds on the turns of the principal curve to avoid overfitting [3]. Kegl and colleagues define the regularization in another way by bounding the total length of the principal curve [4]. Later, Kegl also applies this algorithm to skeletonization of handwritten digits by extending it into the principal graph algorithm [5]. Stanford and Raftery propose another approach that improves on the outlier robustness capabilities of principal curves [6]. Probabilistic principal curves approach, which uses a cubic spline of mixture of Gaussians to estimate the principal curves/surfaces [7], is known to be among the most successful methods to overcome the a common problem of the principal curve algorithms; the bias introduced in the regions of high curvature.

One significant problem among the different approaches mentioned above is that, by definition, they are seeking for a *smooth curve*. In general, data may have loops, self intersections, and bifurcation points, in which case there does not exist a *smooth curve passing through the data* that can represent the data sufficiently. In the presence of such irregularities, there still exists a principal graph, a collection of smooth curves, that can represent the data statistics sufficiently.

In fact, Kegl's principal graph algorithm is perhaps the only method in the literature that can successfully handle such irregularities [5]. In this approach, Kegl reshapes his polygonal line algorithm [4] to handle loops, and self in-

The authors would like to thank Balazs Kegl for providing the optical character dataset. This work is partially funded by NSF grants ECS-0524835, and ECS-0622239.

tersections by modifying it with a table of rules and adding preprocessing and postprocessing steps. The polygonal line algorithm, which is the basis of the principal graph algorithm, is based on the idea of bounding the length of the principal curve to prevent overfitting, but there is an important problem: the optimal length to bound the principal curve is unknown. Therefore to prevent overfitting -and too much generalization on the other side- one needs to rerun the algorithm several times with different length penalty and convergence parameters until the satisfactory results of obtained. At this point, modifying this algorithm with a table of predefined rules that include parameters and thresholds is nothing but further parameterizing the problem.

We recently proposed a mathematically more rigorous definition of principal curves in terms of the data probability density, which leads to a nonparametric algorithm that naturally handles the loops and self-intersections with no additional effort [8]. We provide a subspace constrained mean shift algorithm to find the projection onto the principal curve [9]. In some applications like denoising, compression or dimensionality reduction, one needs to know the projection of all data samples onto the principal curve/graph. However, there are also applications that the principal curve/graph is enough, where a computationally cheaper approximation of the principal curve/graph suffice. For instance, optical character recognition is one such application.

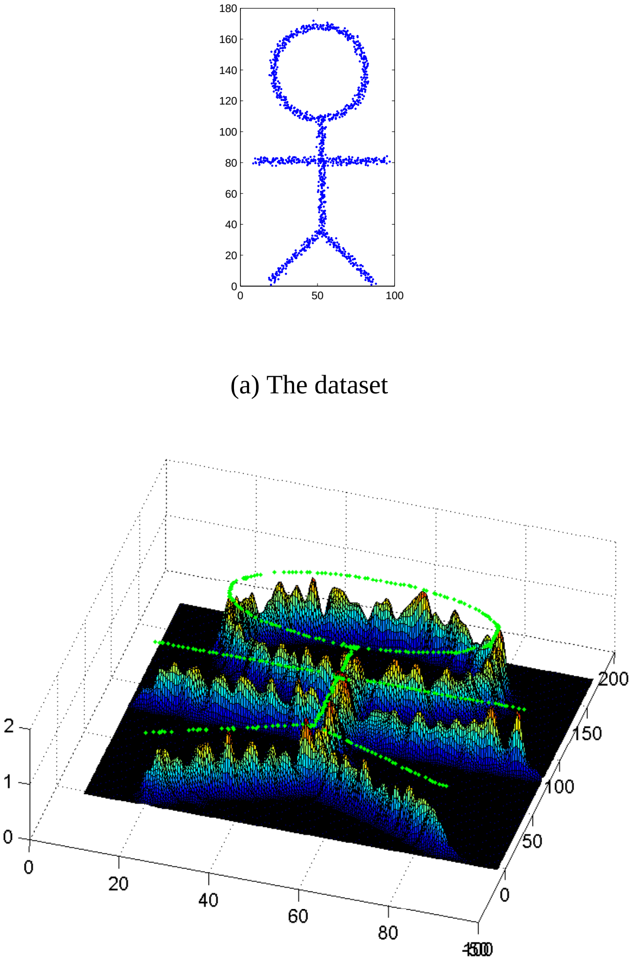
We will develop a computationally inexpensive piecewise linear approximate of our principal graph definition. We will first review our definition briefly, and then provide the piecewise linear subspace constrained mean shift algorithm. We present results of the algorithm on notional datasets, and apply it to optical character recognition as well.

2. LOCALLY DEFINED PRINCIPAL GRAPHS

Before we proceed to the proposed algorithm, we will briefly review our definition of principal graphs. We will start with the general definition of principal manifolds of intrinsic dimensionality d, and continue with the one-dimensional principal manifolds, namely the principal graphs.

A point is in the d dimensional principal manifold iff the gradient of the pdf is orthogonal to at least (n-d) eigenvectors of the Hessian of the pdf, and the eigenvalues corresponding to these (n-d) orthogonal eigenvectors are negative. For one-dimensional principal manifolds, simply substitute $d = 1$, and the general definition further simplifies to the following statement: *A point is on the principal graph iff the gradient of the pdf is an eigenvector of the Hessian of the pdf and the remaining eigenvectors of the Hessian have negative eigenvalues.*

We present an illustration in Figure 1. Kernel density estimate of the data probability distribution, and the data points projected onto the principal curve are shown on a dataset that has a loop, a self-intersection, and a bifurcation point.



(a) The dataset

(b) KDE of the data pdf, along with the principal graph

Fig. 1. A simple illustration of the principal graph.

In the next section, we will develop a fast nonparametric algorithm particulary based on kernel density estimation. At this point, note that the definition of the principal graph is independent of the density estimator used, and one can also use other density estimation techniques to develop various algorithms based on this definition. Wherever suitable, in some applications one can use Gaussian mixture model based density estimates [10] that would lead to computationally cheaper algorithms very similar to the one that we will develop here.

3. PIECEWISE LINEAR SUBSPACE CONSTRAINED MEAN SHIFT

In this section, we will develop a computationally efficient piecewise linear principal graph approximate, but before we move on to the algorithm we will briefly mention the reasons why we use KDE, and the selection of kernel bandwidth, an important implementation step for KDE.

3.1. Natural Connections to KDE

One can see many natural connections between the open ended problems in principal curve and surface fitting literature, and studies in KDE literature. These not only yield direct answers to known problems in principal curve fitting, but also will help to approach these problems in a more principled way.

One obvious connection is the challenge of focusing to a specific region while finding the principal curve; in other words, solving for the principal curve in a region of interest. This is an important challenge since the original principal surface formulation itself is stated to be inefficient for large sample sizes [1]. In KDE, using kernel functions of finite support would handle this problem automatically, and the support of the kernel function clearly defines which samples are necessary for the density estimate, hence for the principal curve/graph, in a given region in the feature space.

Outlier robustness is another important issue in principal curve literature. Principal curve approaches that are based on least squares type methods are very sensitive to noise. Therefore, outliers in the data require special attention, and Stanford and Raftery present results on outlier robustness along these lines by applying their principal curve clustering algorithm to reconnaissance images [6]. Considering our approach, since everything imposed on the probability density estimate is directly imposed on the principal curve as well, one can use variable bandwidth KDE to increase noise robustness. In this approach, data dependent kernel functions are evaluated for each sample such that the width of the kernel is directly proportional with the likelihood of that sample's being an outlier. This can be implemented in several ways, and the most commonly used methods are the K-nearest neighbor based approaches, namely: (*i*) the mean/median distance to the *K*-nearest neighbor data points, (*ii*) sum of the weights of *K*-nearest neighbor data points in a weighted KDE. At this point, note that to obtain an asymptotically unbiased and consistent density estimate the neighborhood parameter *K* should satisfy

$$\lim_{N\to\infty} K = \infty \; , \; \lim_{N\to\infty} \frac{K}{N} = 0 \qquad (1)$$

Selecting the kernel functions in a data dependent manner, by definition, makes our principal curve algorithm based on KDE robust to outliers in the data.

3.2. Selection of the Kernel Bandwidth

Principal curve fitting methods exhibit the problem of overfitting. Overfitting is an issue that arises if the problem is defined in terms of a finite number of samples. By defining the problem in terms of the data density, we assume that the required regularization constraints should be enforced by the density estimation step directly.

Considering our KDE based algorithms, this trade off can be adjusted by setting the kernel width. There is a rich literature about how to select the kernel function, and one can use the literature and select a kernel that optimizes certain criteria about the data, which is much more principled as compared to bounding the length or curvature of the curve to provide the required regularization - *the optimal length or curvature is never known*.

In KDE literature, one of the most common kernel optimization approaches is to use the leave-one-out cross validation maximum likelihood procedure. Consider the random vector $\mathbf{x}$ with samples $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$. The kernel density estimate that leaves the i^{th} sample out of density estimation is given by

$$p_i(\mathbf{x}) = \frac{1}{N} \sum_{j=1,\; j\neq i}^{N} K_\sigma(\mathbf{x} - \mathbf{x}_j) \qquad (2)$$

and the objective of the kernel bandwidth optimization problem is defined by maximizing the log-likelihood function over all samples.

$$\max_\sigma \quad \sum_{i=1}^{N} \, log p_i(\mathbf{x}_i) \qquad (3)$$

This optimization problem can be solved using a line search. Combining the ML kernel with the variable width KDE approach mentioned previously is also straightforward. Let C_i^{KNN} denote the mean distance to *K*-nearest neighbor samples. One can select the variable bandwidth of the i^{th} sample as $\sigma_i = \alpha C_i^{KNN}$, where α is a global scale constant optimized using the ML procedure given above. An anisotropic counterpart of the above isotropic selection method can easily be obtained by defining C_i^{KNN} as the covariance of *K*-nearest neighbor data sampled instead of their mean distance, yielding $\Sigma_i = \alpha C_i^{KNN}$.

KDE based implementation has also other advantages. Anisotropic and/or variable size kernel functions naturally implement many types of constraints that cannot be defined by any bound on the length or the curvature of the curve. Anisotropic kernels yield regularization constraints at different scales among different directions, and variable bandwidth kernel functions define varying constraints throughout the space, wherever necessary as the scale of the data is changing throughout the space. In summary, KDE not only connects the trade off between the projection error and generalization into well studied results of density estimation field, it also allows one to derive data-dependent constraints that vary throughout the space from the data directly.

4. IMPLEMENTATION OF THE PIECEWISE LINEAR SCMS ALGORITHM

Reviewing the definition of the principal graph, a point is on a critical graph iff the local gradient is an eigenvector of the local Hessian, since the gradient has to be orthogonal to the other $(n-1)$ eigenvectors. Furthermore, for this point to be in the principal curve, the corresponding $(n-1)$ eigenvalues must be negative. Under the assumption of a KDE, a modification of the mean-shift algorithm by constraining the fixed-point iterations to the directions of local curvature at the current point in the trajectory leads to an update that converges to the principal curves and not to the local maxima. This is the SCMS algorithm that we proposed earlier [8]. The algorithm is conceptually simple, but computationally demanding. The bottleneck of the computational requirement of evaluating the eigendecomposition of the local covariance at each step for all samples. In the fast version we present here, we get rid of the eigendecomposition of the covariance, and perform the projection for a representative small subset of points -not necessarily data points- in the feature space to obtain the principal curve approximate.

First simplification is to get rid of the covariance evaluation and its eigendecomposition at each iteration. Instead, one can use the eigendecomposition of the covariance at the initial point and assume that the projection direction is not changing significantly along the trajectory, or even cheaper, one can also use the gradient at the initial point. One important observation here is that, if the sought principal curve is a flat ridge, then the gradient of the initial update of each sample is very close to the sought eigenvector of the Hessian, which yields a computationally much cheaper approximate of the required subspace. Consider the example in Figure 2, where the trajectories of the gradient (red) and the trajectories of the ideal principal curve projection, the greatest eigenvector of local covariance (blue) are shown, along with the principal curve (green) of this three component Gaussian mixture probability density.

The second simplification is to build the piecewise linear principal graph over a representative set of samples, instead of evaluating the projections of all data samples. To achieve this, we propose to employ a clustering of the data and combine the cluster means through the principal graph. By definition, the modes -the local maxima- of the pdf are in the principal graph [8, 9]. At this point, since it maps all points to the mode of the associated attraction basin, mean shift [11] becomes a suitable way to achieve the required clustering.

We will start with the KDE of this data set (using Gaussian kernels for illustration)

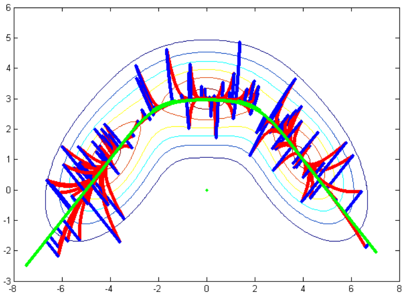


Fig. 2. Gradient (red) and principal curve projection trajectories (blue) for a Gaussian mixture pdf.

Table 1. PL-SCMS Algorithm

1. Select the kernel size (using (3) or any other method)
2. Run mean-shift iterations in (5), to find the modes
3. For all mode pairs project the midpoint of the line that connects these modes onto the principal curve by doing the following:
   - Evaluate the local covariance $\Sigma^{-1}(\mathbf{x})$ given in (6)
   - Find the greatest eigenvector of the local covariance $\mathbf{v}$, select the constrained subspace direction as either $\mathbf{v}$ or -$\mathbf{v}$, depending on which of these has a positive inner product with the gradient given in (6): $\mathbf{d} = \mathbf{v}\, sign(\mathbf{g}^T\mathbf{v})$
   - Until convergence: iterate (5), and project the update onto the constrained subspace direction $\mathbf{d}$.
4. Go back to step 3 if the required depth is not obtained.

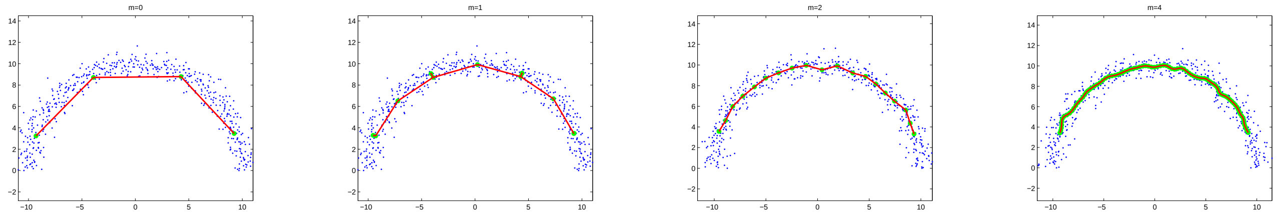
$$p(\mathbf{x}) = (1/N)\sum_{i=1}^{N} G_{\Sigma_i}(\mathbf{x}-\mathbf{x}_i) \quad (4)$$

where Σ_i is the covariance for the Gaussian kernel, and the mean shift update is given by

$$\mathbf{x} \leftarrow (\sum_{i=1}^{N}\Sigma_i^{-1}G_{\Sigma_i}(\mathbf{x}-\mathbf{x}_i))^{-1}\sum_{i=1}^{N}\Sigma_i^{-1}G_{\Sigma_i}(\mathbf{x}-\mathbf{x}_i)\mathbf{x}_i \quad (5)$$

Iterating 5 until convergence, the modes, namely the zero dimensional principal set, is obtained.

Once the modes of the pdf is obtained, the next task is to combine these modes. Consider the gradient and the



(a) (b) (c) (d)

Fig. 3. Principal curve approximations on the semicircle dataset with depths 0, 1, 2, and 4.

Hessian of the KDE,

$$\begin{array}{c}\mathbf{g}(\mathbf{x}) = -N^{-1}\sum_{i=1}^{N} c_i \mathbf{u}_i \\ \mathbf{H}(\mathbf{x}) = N^{-1}\sum_{i=1}^{N} c_i(\mathbf{u}_i\mathbf{u}_i^T - \Sigma_i^{-1}) \\ \text{where} \quad \mathbf{u}_i = \Sigma_i^{-1}(\mathbf{x}-\mathbf{x}_i) \quad , \quad c_i = G_{\Sigma_i}(\mathbf{x}-\mathbf{x}_i), \\ \text{and} \quad \mathbf{\Sigma}^{-1}(\mathbf{x}) = -p^{-1}(\mathbf{x})\mathbf{H}(\mathbf{x}) + p^{-2}\mathbf{g}(\mathbf{x})\mathbf{g}^T(\mathbf{x})\end{array} \qquad (6)$$

where the subspace constrained mean shift is nothing but the same mean shift iteration given in (5), projected onto the subspace selected by the gradient or greatest eigenvector of covariance matrix, evaluated at the initial point, and the convergence condition is checking if the gradient is parallel to one of the eigenvectors by evaluating

$$|\mathbf{g}(\mathbf{x})^T\mathbf{H}(\mathbf{x})\mathbf{g}(\mathbf{x})|/\|\mathbf{H}(\mathbf{x})\mathbf{g}(\mathbf{x})\|\|\mathbf{g}(\mathbf{x})\| > 1-\epsilon \qquad (7)$$

where ϵ is close to zero, depending on the required accuracy of the principal curve - typically 0.01. Also, a computationally cheaper alternative for the stopping criterion that does not use the gradient and the Hessian might be $\|\mathbf{V}\mathbf{V}^T\mathbf{m}(\mathbf{x}(k+1)) - x(k)\| < \epsilon$.

We select the subset of points to be projected in a sequential way, very similar to Kegl's polygonal line algorithm: for each pair of mode, we will project the midpoint of the line segment that connects two projected points in the piecewise linear approximate. This can be performed up to a *depth* of *m*, leading to $2^m + 1$ points in between every pair of mode.

If the data is not self intersecting, performing the piecewise linear approximation may not be necessary for all pairs of modes. A cheaper alternative would be to check the connectivity of the mode graph, and compute the piecewise approximate of the closest pair of mode in the list of unconnected modes. We will present examples for each case. Table 1 summarizes the PL-SCMS algorithm[1].

5. EXPERIMENTAL RESULTS

This section presents our results on notional and real data. For the toy data example, we will present the intermediate steps as well; for the OCR examples only the final results are given. In all experiments, we will use isotropic fixed-bandwidth kernel functions, where the bandwidth of the Gaussian kernel is determined with the ML procedure as in (3).

5.1. Semi-circle dataset

This dataset has 500 samples, drawn from a uniform distribution along a circle with a Gaussian perturbation in the radial direction. Figure 4a shows the data samples (blue) along with the modes (green), that is the result of step 2 in Table 1, and leads to the principal curve approximation of depth 0. Figure 4b, 4c, and 4d presents the principal curve approximates of depth 1, 2, and 4, respectively. The piecewise linear principal curve approximations are shown with green.

5.2. Feature Extraction for OCR

Skeletonization of optical characters is a natural application for principal graphs [5], and here we present how our algorithm performs in this application. The dataset used in this experiment consists of handwritten digits and this dataset is provided by Kegl. For this experiment, we use the ML kernel bandwidth given in Section 3.2, and the depth parameter $m = 2$. Results are presented in Figure 4.

6. CONCLUSIONS

The piecewise linear approximate provided by our approach is similar to the outcome of Kegl's principal graph algorithm; however, the difference is that once you estimate the probability density of the data, the principal curve estimate will not overfit the data no matter how dense you would like to populate the points on the principal curve. The irregularities like loops, self intersections, and bifurcation points do not require special attention and are handled naturally by the definition. The resulting algorithm has only a single parameter, the kernel bandwidth, the optimal value of

[1] Upon acceptance, the MATLAB implementation of the algorithm will be made available on the author's web page.

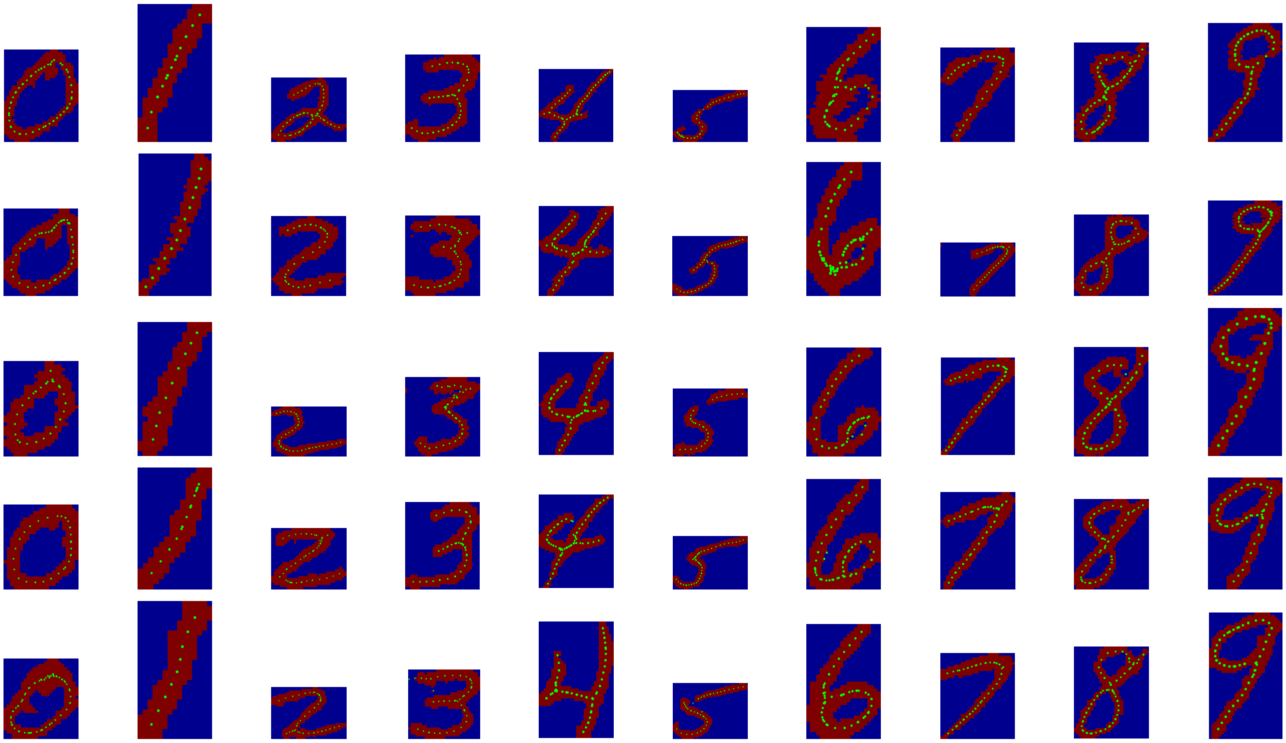


Fig. 4. Principal graph results in optical characters

which can be obtained by an inexpensive ML training prior to principal curve extraction.

In summary, the piecewise linear subspace constrained mean shift provides a computationally much cheaper accurate estimate of the principal graph that we define. In the cases, where the projection of all data samples onto the principal curve is required like in denoising, compression, and dimensionality reduction, this approach might be insufficient, because it is specifically designed for applications where a back projection of the data is not required and the principal graph itself is sufficient, like in the case of optical character recognition. The experimental results provided on notional datasets and the optical character recognition application show effectiveness and robustness of the proposed approach.

7. REFERENCES

[1] T. Hastie and W. Stuetzle, "Principal curves," *Jour. Am. Statistical Assoc.*, vol. 84, pp. 502–516, 1989.

[2] R. Tibshirani, "Principal curves revisited," *Statistics and Computation*, vol. 2, pp. 183–190, 1992.

[3] S. Sandilya and S. R. Kulkarni, "Principal curves with bounded turn," *IEEE Trans. on Information Theory*, vol. 48, no. 10, pp. 2789–2793, 2002.

[4] B. Kegl, A. Kryzak, T. Linder, and K. Zeger, "Learning and design of principal curves," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 22, no. 3, pp. 281–297, 2000.

[5] B. Kegl and A. Kryzak, "Piecewise linear skeletonization using principal curves," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 24, no. 1, pp. 59–74, 2002.

[6] D. C. Stanford and A. E. Raftery, "Finding curvilinear features in spatial point patterns: Principal curve clustering with noise," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 22, no. 6, pp. 601–609, 2000.

[7] K. Chang and J. Grosh, "A unified model for probabilistic principal surfaces," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 24, no. 1, pp. 59–74, 2002.

[8] Deniz Erdogmus and Umut Ozertem, "Self-consistent locally defined principal surfaces," in *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, 2007, pp. II549–II552.

[9] Umut Ozertem and Deniz Erdogmus, "Local conditions for critical and principal manifolds," in *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, 2008.

[10] Deniz Erdogmus and Umut Ozertem, "Nonlinear coordinate unfolding via principal curve projections with application to nonlinear bss," in *Proceedings of International Conference on Neural Information Processing*, 2007.

[11] Y. Cheng, "Mean shift, mode seeking, and clustering," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 17, no. 8, pp. 790–799, 1995.