SELF-CONSISTENT LOCALLY DEFINED PRINCIPAL SURFACES

Deniz Erdogmus, Umut Ozertem

CSEE Department, Oregon Health and Science University, Portland, Oregon, USA

ABSTRACT

Principal curves and surfaces play an important role in dimensionality reduction applications of machine learning and signal processing. Vaguely defined, principal curves are *smooth curves* that pass through the *middle* of the data distribution. This intuitive definition is ill posed and to this day researchers have struggled with its practical implications. Two main causes of these difficulties are: (i) the desire to build a self-consistent definition using global statistics (for instance conditional expectations), and (ii) not decoupling the definition of the principal curve from the data samples. In this paper, we introduce the concept of principal sets, which are the union of all principal surfaces with a particular dimensionality. The proposed definition of principal surfaces provides rigorous conditions for a point to satisfy that can be evaluated using only the gradient and Hessian of the probability density at the point of interest. Since the definition is decoupled from the data samples, any density estimator could be employed to obtain a probability distribution expression and identify the principal surfaces of the data under this particular model.

Index Terms— principal curves and surfaces, dimensionality reduction, unsupervised learning, optical character recognition

1. INTRODUCTION

Intuitively, principal curves are smooth curves that pass through the middle of the data. This definition has been formulated mathematically by Hastie and Stuetzle in which self-consistent principal curves are defined to coincide with the *local* conditional expectation of the data/distribution at every orthogonal hyperplane to the curve itself [1]. It is probably safe to say that this definition has so far drowe the development of almost all principal curve identification algorithms. A natural consequence of the definition above is that the principal curve becomes a stationary point of the variational optimization problem that minimizes the projection means squared error when high dimensional data is approximated by their projections on the principal curve [1].

The literature is relatively rich with various propositions of algorithms for determining principal curves based on definitions that are variants of the Hastie-Stuetzle definition. Tibshirani provides a definition based on mixture models and estimation is carried out using expectation-maximization [2]. Kegl and colleagues define a regularized version of Hastie's definition by constraining the total length of the parametric curve to be *fit* to the data [3]. Sandilya and Kulkarni similarly regularize the principal curve by imposing bounds on the turns of the curve [4].

The principal surface concept is closely related to nonlinear principal component analysis (in the context of neural networks) and manifold learning (in the context of machine learning). There exist various approaches and algorithms for determining parametric or nonparametric solutions to these two problems ranging from autoassociative neural network models to spectral manifold unfolding techniques, which we will not review here for brevity.

In summary, Hastie's definition is quite elegant and forms a strong basis for many, possibly all, of the approaches one can find in the literature. This definition and the resulting variants attempt to define principal curves and surfaces using data statistics that are conditioned locally to the hyperplane orthogonal to the curve (conditional expectation or squared projection error). This definition that utilizes global statistics (through hyperplanes that extend to infinity) create the difficulty that when nonlinear principal curves make turns such that the orthogonal hyperplanes of two different points on the curve intersect, it becomes ambiguous to which point on the curve the intersection would be projected to. The source of the difficulty can actually be traced back to the definition of Hastie, which was most likely inspired by the desire to utilize a definition of principal curves that is relevant to least squares regression and the minimum squared error projection property of linear principal component analysis. Furthermore, the ease of estimating conditional expectations over hyperplanes using sample averages is attractive for practical algorithm design purposes.

In this paper, we propose an alternative self-consistent definition of principal curves and surfaces. The definition utilizes the concept of *likelihood maximization* in its more general sense rather than the least squares approach exploited in existing approaches, which emerges from the maximization of symmetric unimodal distributions (such as Gaussian – many illustrations in fact consider synthetic datasets that are radially perturbed by a Gaussian distribution from some constructed principal curve). The intuition behind this is that the principal curves/surfaces pass through *high-density* regions of the data (in other words, they must be some form of local maxima), which might not necessarily be the *middle* literally. Furthermore, in the primary definition of principal curves, we do not believe that smoothness considerations should be included, because smoothness is an issue that arises due to regularization concerns in the presence of only a finite number of data samples. In our view, principal curves/surfaces should be uniquely defined by the data distribution according to the likelihood maximization principle mentioned above and the smoothness of the distribution itself would naturally impose any necessary smoothness constraints on the corresponding principal surfaces.

Authors thank Jose Principe, Sudhir Rao, and Allan de Medeiros for valuable comments and providing the mixture of Gaussians example. This work is supported by NSF grants ECS-0524835 and ECS-0622239.

The definition we propose also possesses the following desirable properties that are mutually agreed on by various authors in the literature: (*i*) existence of the principal surfaces is theoretically guaranteed; (*ii*) definition provides a unique solution for any dimensional principal surface; (*iii*) one need not study global statistics to determine a local portion of the principal curve/surface, local density information is sufficient (and of course necessary). In this paper, we do not make any attempts to develop a computationally efficient exact or approximate algorithm that identifies principal curves or surfaces. Based on the definition, numerical integration of a partial differential equation emerges as a natural (but not necessarily efficient) approach to determine these curves and surfaces. Therefore, we demonstrate results utilizing such an algorithm.

The proposed definition when applied to feature (skeleton) extraction for optical character recognition (OCR) yields the desired skeletons. OCR is a particularly relevant application because while the majority of alphanumeric characters contain some form of intersection, existing definitions of algorithms for principal curves typically fail to handle such situations (save for Kegl's heuristic yet effective approach [3]).

2. DEFINITION OF PRINCIPAL SURFACES

A mathematically rigorous definition of principal surfaces must be defined in terms of probability density functions (pdf) and not in terms of data points or sample statistics. Consequently, in the rest of the paper, we propose our definition and develop properties with the understanding that a pdf expression that describes the data is available, whether it is known or estimated parametrically or nonparametrically from the data.

Given a random vector $\mathbf{x}\in\Re^n$, let $p(\mathbf{x})$ be its pdf, $\mathbf{g}(\mathbf{x})$ be the transpose of the local gradient of this pdf, and $\mathbf{H}(\mathbf{x})$ be the local Hessian of this pdf. To avoid mathematical complications, in the current treatment we assume that the data distribution $p(\mathbf{x})$ is at least twice differentiable, so that both $\mathbf{g}(\mathbf{x})$ and $\mathbf{H}(\mathbf{x})$ are continuous. The definition we will propose requires differentiating strict local maxima where all relevant eigenvalues of the Hessian matrix are strictly negative and local maxima where the Hessian is allowed to have a mixture of negative and zero eigenvalues.

Definition 2.1. A point $\mathbf{x}$ is an element of the *d*-dimensional principal set, denoted by $\wp^d$, iff $\mathbf{g}(\mathbf{x})$ is orthogonal (null inner product) to at least (*n*-*d*) eigenvectors of $\mathbf{H}(\mathbf{x})$ *and* $p(\mathbf{x})$ is a strict local maximum in the subspace spanned by these (*n*-*d*) eigenvectors.

Lemma 2.1. Strict local maxima of $p(\mathbf{x})$ constitute $\wp^0$.

Proof. If $\mathbf{x}^*$ is a strict local maximum of of $p(\mathbf{x})$, then $\mathbf{g}(\mathbf{x}^*)=\mathbf{0}$ and all eigenvalues of $\mathbf{H}(\mathbf{x}^*)$ are strictly negative. Let $\{\mathbf{q}_1(\mathbf{x}^*),\ldots,\mathbf{q}_n(\mathbf{x}^*)\}$ be the orthonormal eigenvectors of $\mathbf{H}(\mathbf{x}^*)$. Clearly $\mathbf{g}^T(\mathbf{x}^*)\mathbf{q}_i(\mathbf{x}^*)=0$ for $i=1,\ldots,n$; hence $\mathbf{g}(\mathbf{x}^*)$ is orthogonal to all *n* eigenvectors of $\mathbf{H}(\mathbf{x}^*)$. Furthermore, by hypothesis, $p(\mathbf{x}^*)$ is a local maximum in the space spanned by these *n* eigenvectors, which is $\Re^n$. Consequently $\mathbf{x}^*\in\wp^0$. □

This lemma states that the modes of a pdf, called principal points, form the 0-dimensional principal set. Note that the modes of a pdf provide a natural clustering solution for data with a certain pdf. In fact, the widely used mean-shift algorithm [5,6] utilizes this property of the modes of a kernel density estimate to arrive at a clustering solution nonparametrically for a given data set.

Lemma 2.2. $\wp^d\subset\wp^{d+1}$.

Proof. Let $\mathbf{x}^*\in\wp^d$. Let $\{\mathbf{q}_1(\mathbf{x}^*),\ldots,\mathbf{q}_n(\mathbf{x}^*)\}$ be the orthonormal eigenvectors of $\mathbf{H}(\mathbf{x}^*)$. Without loss of generality assume that the subspace of interest is defined by eigenvectors indexed from *d*+1 to *n*. Therefore, by definition $\mathbf{g}^T(\mathbf{x}^*)\mathbf{q}_i(\mathbf{x}^*)=0$ for $i=d+1,\ldots,n$ and $p(\mathbf{x}^*+\boldsymbol{\delta})<p(\mathbf{x}^*)$ for all $\boldsymbol{\delta}$ such that $\boldsymbol{\delta}=\sum_{i=d+1}^{n}\alpha_{i-d}\mathbf{q}_i(\mathbf{x}^*)$, where $\boldsymbol{\alpha}\in B_\varepsilon^{n-d}(\mathbf{0})$ for any $\varepsilon>0$. Consequently, $\mathbf{g}^T(\mathbf{x}^*)\mathbf{q}_i(\mathbf{x}^*)=0$ for $i=d+2,\ldots,n$ and $p(\mathbf{x}^*+\boldsymbol{\delta}')<p(\mathbf{x}^*)$ for $\boldsymbol{\delta}'=\sum_{i=d+2}^{n}\beta_{i-d-1}\mathbf{q}_i(\mathbf{x}^*)$, where $\boldsymbol{\beta}\in B_\varepsilon^{n-d-1}(\mathbf{0})$ for any $\varepsilon>0$. Hence $\mathbf{x}^*\in\wp^{d+1}$; therefore, $\wp^d\subset\wp^{d+1}$. □

In plain terms, this lemma states that low dimensional principal sets are subsets of higher dimensional principal sets. In the context of continuous and twice differentiable pdfs this implies that principal curves must pass through the local maxima, two-dimensional principal surfaces must contain principal curves, etc. Consequently, a deflation or inflation procedure could be employed to discover these surfaces sequentially (as done in some PCA algorithms).

Lemma 2.3. Let $\mathbf{x}^*\in\wp^d$. Let $\{\mathbf{q}_1(\mathbf{x}^*),\ldots,\mathbf{q}_n(\mathbf{x}^*)\}$ be the eigenvectors of $\mathbf{H}(\mathbf{x}^*)$. Without loss of generality let $S_{//}(\mathbf{x}^*)=span\{\mathbf{q}_1(\mathbf{x}^*),\ldots,\mathbf{q}_d(\mathbf{x}^*)\}$, $S_\perp(\mathbf{x}^*)=\Re^n$-$S_{//}(\mathbf{x}^*)$, and assume that $\mathbf{g}^T(\mathbf{x}^*)\mathbf{q}_i(\mathbf{x}^*)=0$ for $i=d+1,\ldots,n$. Then $S_{//}(\mathbf{x}^*)$ is tangent to $\wp^d$ at $\mathbf{x}^*$.

Proof. Consider the following truncated Taylor approximation: $p(\mathbf{x}^*+\boldsymbol{\delta})\approx p(\mathbf{x}^*)+\mathbf{g}^T(\mathbf{x})\boldsymbol{\delta}+\boldsymbol{\delta}^T\mathbf{H}(\mathbf{x})\boldsymbol{\delta}/2+O(\|\boldsymbol{\delta}\|^3)$. Let $\mathbf{Q}(\mathbf{x}^*)=[\mathbf{q}_1(\mathbf{x}^*),\ldots,\mathbf{q}_n(\mathbf{x}^*)]$, where $\mathbf{Q}^T(\mathbf{x}^*)\mathbf{Q}(\mathbf{x}^*)=\mathbf{I}$. We have $\mathbf{H}(\mathbf{x}^*)=\mathbf{Q}(\mathbf{x}^*)\boldsymbol{\Lambda}(\mathbf{x}^*)\mathbf{Q}^T(\mathbf{x}^*)$. For brevity, from now on, we will drop the argument $(\mathbf{x}^*)$ from relevant functions when the point of evaluation is clear from the context. Define *parallel* and *orthogonal* components of the local Hessian as follows:

$$\mathbf{H}_{//}=\sum_{i=1}^{d}\lambda_i\mathbf{q}_i\mathbf{q}_i^T \qquad \mathbf{H}_{\perp}=\sum_{i=d+1}^{n}\lambda_i\mathbf{q}_i\mathbf{q}_i^T \tag{1}$$

Since by hypothesis $\mathbf{g}^T\mathbf{q}_i=0$ for $i=d+1,\ldots,n$, we can express the local gradient as a linear combination of the remaining eigenvectors: $\mathbf{g}=\sum_{i=1}^{d}\alpha_i\mathbf{q}_i$. For an arbitrary vector $\boldsymbol{\delta}=\boldsymbol{\delta}_{//}+\boldsymbol{\delta}_\perp=\sum_{i=1}^{d}\beta_i\mathbf{q}_i+\sum_{i=d+1}^{n}\beta_i\mathbf{q}_i$, clearly, $\mathbf{g}^T\boldsymbol{\delta}=\mathbf{g}^T\boldsymbol{\delta}_{//}$. Also similarly $\boldsymbol{\delta}^T\mathbf{H}\boldsymbol{\delta}=\boldsymbol{\delta}_{//}{}^T\mathbf{H}_{//}\boldsymbol{\delta}_{//}+\boldsymbol{\delta}_\perp{}^T\mathbf{H}_\perp\boldsymbol{\delta}_\perp$. Substituting these last two expressions in the Taylor series approximation:

$$p(\mathbf{x}^*+\boldsymbol{\delta})\approx p(\mathbf{x}^*)+\mathbf{g}^T\boldsymbol{\delta}_{//}+(\boldsymbol{\delta}_{//}^T\mathbf{H}_{//}\boldsymbol{\delta}_{//}+\boldsymbol{\delta}_\perp^T\mathbf{H}_\perp\boldsymbol{\delta}_\perp)/2 \tag{2}$$

Consequently, at $\mathbf{x}^*$, a perturbation $\boldsymbol{\delta}_{//}\in S_{//}(\mathbf{x}^*)$ yields $p(\mathbf{x}^*+\boldsymbol{\delta}_{//})\approx p(\mathbf{x}^*)+\mathbf{g}^T\boldsymbol{\delta}_{//}+\boldsymbol{\delta}_{//}^T\mathbf{H}_{//}\boldsymbol{\delta}_{//}/2$. If infinitesimally small, this perturbation yields a point $\mathbf{x}^*+\boldsymbol{\delta}_{//}$ where $\mathbf{g}(\mathbf{x}^*+\boldsymbol{\delta}_{//})\approx\mathbf{g}(\mathbf{x}^*)+\mathbf{H}_{//}(\mathbf{x}^*)\boldsymbol{\delta}_{//}$, $\mathbf{H}(\mathbf{x}^*+\boldsymbol{\delta}_{//})\approx\mathbf{H}_{//}(\mathbf{x}^*)$. Thus, $\mathbf{g}^T(\mathbf{x}^*+\boldsymbol{\delta}_{//})\mathbf{q}_i(\mathbf{x}^*)=0$ for $i=d+1,\ldots,n$ and $\mathbf{x}^*+\boldsymbol{\delta}_{//}$ is a local maximum in $S_\perp(\mathbf{x}^*+\boldsymbol{\delta}_{//})\approx S_\perp(\mathbf{x}^*)$ from (2). Hence, $\mathbf{x}^*+\boldsymbol{\delta}_{//}\in\wp^d$; therefore, $S_{//}(\mathbf{x}^*)$ is locally tangent to $\wp^d$ at $\mathbf{x}^*$. □

Note that Lemma 2.3 defines the *d*-dimensional principal subset utilizing only local gradient and Hessian information that can be calculated at any point in the vector space from the given probability distribution. Lemma 2.2 creates a unifying perspective between clustering and manifold learning as dimensionality reduction tools that can be utilized for data compression and denoising. Finally, note that the definition so far is not concerned about dividing a *d*-dimensional principal set into disjoint surface

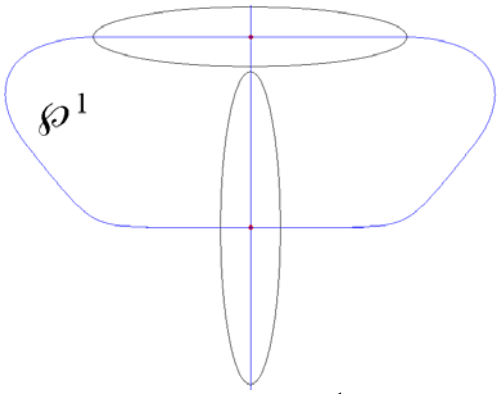

Figure 1. An attempt to illustrate the $\wp^1$ of the mixture of two Gaussians shown by ellipses arranged as a T. The local first principal curve of the horizontal ellipse merges with the local second principal curve of the vertical ellipse. The first 1-dimensional principal component would intuitively include the T-shaped local first principal curves of the two ellipses. However, such a *curve* would then lose intuitive appeal by self-intersecting and not being parameterized by one coordinate.

components that are ranked in terms of some data statistics (as opposed to the natural ranking of linear principal components by the eigenvalues of the data covariance matrix). For instance, we have not yet addressed the question of how to define a global *first principal curve* for a given distribution. Locally however, the eigenvalues of the Hessian could be utilized to rank the orthogonal directions in the tangent space $S_{//}(\mathbf{x}^*)$ at any point $\mathbf{x}^*$. We will discuss this briefly later.

Principal Curves. Note that according to the definition proposed above, $\wp^1$ (union of all principal curves) consists of all points $\mathbf{x}$ where $\mathbf{g}(\mathbf{x})$ is an eigenvector of $\mathbf{H}(\mathbf{x})$. This provides some guidance towards building algorithms to discover principal curves. In this paper, due to lack of space, we will not investigate potential algorithm designs.

Example. To illustrate the concept, let us consider the trivial example of a Gaussian distribution with zero mean and covariance $\mathbf{\Sigma}$. We have the following:

$$\begin{aligned} p(\mathbf{x}) &= C_{\mathbf{\Sigma}} e^{-\mathbf{x}^T \mathbf{\Sigma}^{-1} \mathbf{x}/2} \\ \mathbf{g}(\mathbf{x}) &= -p(\mathbf{x}) \mathbf{\Sigma}^{-1} \mathbf{x} \\ \mathbf{H}(\mathbf{x}) &= p(\mathbf{x})[\mathbf{\Sigma}^{-1} \mathbf{x}\mathbf{x}^T \mathbf{\Sigma}^{-1} - \mathbf{\Sigma}^{-1}] \end{aligned} \quad (3)$$

Let $\mathbf{\Sigma}^{-1} = \sum_{i=1}^{n} \gamma_i^{-1} \mathbf{v}_i \mathbf{v}_i^T$. Since the general case of *d*-dimensional principal surfaces (hyperplanes) is computationally cumbersome, we will not go through this most general case. Instead consider the easier principal curves (lines) by letting $\mathbf{x} = \alpha \mathbf{v}_k$. For these points that are aligned with the eigenvectors of the covariance matrix, we can calculate that the gradient is $\mathbf{g}(\mathbf{x}) = -\alpha p(\mathbf{x}) \gamma_k^{-1} \mathbf{v}_k$ and the Hessian is $\mathbf{H}(\mathbf{x}) = p(\mathbf{x})[\alpha^2 \gamma_k^{-2} \mathbf{v}_k \mathbf{v}_k^T - \mathbf{\Sigma}^{-1}]$. Note that any $\mathbf{v}_j$ is an eigenvector of this particular Hessian matrix; specifically we have:

$$\mathbf{H}(\alpha \mathbf{v}_k) \mathbf{v}_j = p(\mathbf{x})(\alpha^2 \gamma_j^{-2} \delta_{kj} - \gamma_j^{-1}) \mathbf{v}_j \quad (4)$$

From (4) we see that if $j \neq k$ (any directional orthogonal to the local gradient), then the eigenvalue of the Hessian becomes negative (specifically, $-\gamma_j^{-1} < 0$). Therefore, $\mathbf{x} = \alpha \mathbf{v}_k$ is a local maximum in the subspace spanned by the eigenvectors orthogonal to the gradient, hence $\mathbf{x} \in \wp^1$. In other words, all points that lie on one of the eigenvectors of the data covariance matrix $\mathbf{\Sigma}$ is on some principle curve (seen to be a line in this case as we would expect). Hence, linear principal components of a Gaussian distributed data arise naturally from the proposed definition.

Ranking portions of the principal surfaces. In general, it is difficult to designate a portion/subset of $\wp^1$ as the first, second, third,… principal curve. The reason is that counterintuitive scenarios might arise in nonlinear principal surfaces and local information might not always indicate global rank. To illustrate this fact, we qualitatively study a mixture of two Gaussians shaped like T (see Fig. 1). The principal curves of a pdf will form a graph where the modes are the nodes. The connecting edges could pass through other stationary points of the pdf which are not local maxima (for instance saddle points – note that a saddle point with one positive and *n*-1 negative eigenvalues could lie in $\wp^1$ provided that the gradient at this point is aligned with the eigenvector of the positive eigenvalue. This saddle point would not be in $\wp^0$, however. In fact, such saddle points are essential for the graph to be connected smoothly by facilitating sharp turns in the principal curves. These saddle points also create the problem observed in Figure 1, by potentially merging two principal curve segments emanating from two modes, but are locally ranked differently. We also know that mixtures of Gaussians in high-dimensional spaces could have more modes than the number of components [7]. These additional modes, not at the center of a component, would also behave similarly. Other degenerate cases are possible. Since in general ranking components of $\wp^1$ into ordered principal curves is not possible, we will not attempt to resolve this problem in this short paper and will address the issue in another publication.

Tracing $\wp^1$. Using an inflation approach and numerical integration, we can determine the one-dimensional principal set. Determining the modes of a pdf is easy. In practice, one could start from some number of reasonable initial points and use gradient ascent or a fixed point algorithm to find the corresponding modes (hopefully all modes of the pdf). In Gaussian mixture distributions, including kernel density estimates, such an algorithm could start from the component means as done in mean shift. Once the modes are determined, starting from each mode and *shooting* a trajectory in the direction of each eigenvector at the mode of interest, one can start tracing the edges of the principal curve graph. A numerical integration algorithm such as Runge-Kutta order 4 (RK4) could be utilized to numerically determine the next point on the curve starting from the current point and moving in the direction of the corresponding local eigenvector of the Hessian. With a small step size and patience, we have obtained reasonable approximations to the principal curves of various Gaussian mixture densities.

Note however, that our goal in this paper is to propose a self-consistent and mathematically rigorous definition of principal curves that utilizes local information. Therefore, an attempt to develop an efficient algorithm that identifies the principal curves is not made. It should be noted that, in fact an algorithm that identifies the curves would be relatively easy to determine when compared with the task of determining a reliable and computationally feasible method of projecting an arbitrary test data point on to the principal curves or surfaces of interest. The latter is the relevant challenge since the purpose of determining

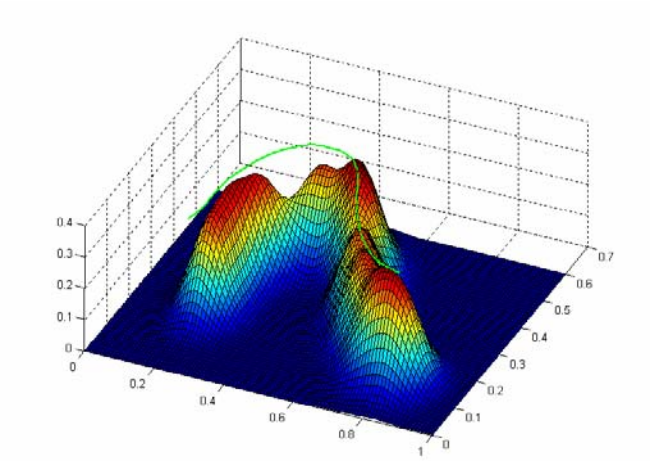

Figure 2. The *first* principal curve of a mixture of 10 Gaussians. Note that this is an easy case where the first principal curve direction is easily identified by local Hessian eigenvalues. (See in color.)

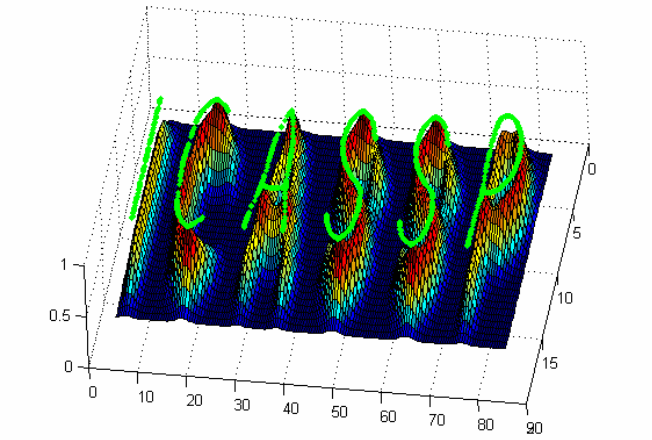

Figure 3. The *first* principal set of the ICASSP dataset. Note that this data set is a mixture of many characters (some self-intersecting), therefore the intuitive single smooth curve passing through the data cannot be identified, however, the 1-dimensional principal set still exists. Only the dominant portions of $\wp^1$ are shown by tracing the direction corresponding to the largest local Hessian eigenvalue. (See in color.)

these curves is to seek projections for data compression and denoising.

3. EXPERIMENTAL RESULTS

In this section we present experimental results obtained using the proposed principal curve definition. Here, we use the RK4 numerical integration method; we trace the principal curves by starting a trajectory at each of the modes of the pdf and tracing the eigenvector of the local Hessian that has the largest eigenvalue. In our examples, Gaussian mixture models (GMM) and Gaussian-kernel density estimation method (for OCR) is employed, therefore relevant modes can be identified by mean-shift [5,6].

The GMM consists of 10 Gaussian components, and the numerical integration algorithm is initialized to the peak of one of these Gaussian components. The probability density and the corresponding principal curve are depicted in Figure 2.

Optical character recognition is one of the most promising applications of principal curves, and the principal curves literature includes many OCR applications that use the principal curve as a skeleton-feature extraction step. For this reason, we provide OCR skeleton extraction results for our principal curve definition using the ICASSP dataset. Note that, this dataset also includes some letters that force the principal curve to intersect itself. The kernel density estimate and the principal curve are shown in Figure 3.

4. DISCUSSION

This paper contributes a self-consistent principal surface definition, which is uniquely-defined through local gradient Hessian information about the data distribution. The definition avoids smoothness concerns by decoupling the principal curve definition from algorithmic estimation aspects of the problem. Given a probability density function, the proposed principal surfaces become local maxima in their orthogonal subspaces, therefore the intuition behind principal surfaces is changed from *passing from the middle of the data* to *passing from the high-density ridge of the data*. This corresponds to selecting principal curves that have a maximum likelihood property rather than a least-squares representation error property. Various complications arising from the conditional-expectation-based definition of Hastie are avoided by this local information oriented definition.

5. REFERENCES

[1] T. Hastie, W. Stuetzle, "Principal Curves," Journal of the American Statistical Association, vol. 84, no. 406, pp. 502-516, 1989.

[2] R. Tibshirani, "Principal Curves Revisited," Statistics and Computation, vol. 2, pp. 183-190, 1992.

[3] B. Kegl, A. Kryzak, T. Linder, K. Zeger, "Learning and Design of Principal Curves," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, no. 3, pp. 281-297, 2000.

[4] S. Sandilya, S.R. Kulkarni, "Principal Curves with Bounded Turn," IEEE Transactions on Information Theory, vol. 48, no. 10, pp. 2789-2793, 2002.

[5] K. Fukunaga, L.D. Hostetler, "The Estimation of the Gradient of a Density Function, with Applications in Pattern Recognition," IEEE Transactions on Information Theory, vol. 21, pp. 3240, 1975.

[6] U. Ozertem, D. Erdogmus, T. Lan, "Mean Shift Spectral Clustering for Perceptual Image Segmentation," Proceedings of ICASSP 2006, vol. 2, pp. 2.117-2.120, 2006.

[7] M.A. Carreira-Perpinan, C.K.I. Williams, "On the number of modes of a Gaussian Mixture," Proceedings of Scale-Space Methods in Computer Vision, pp. 625-640, 2003.