Journal of Machine Learning Research 12 (2011) 241-274 Submitted 3/09; Revised 10/10; Published 1/11

Locally Defined Principal Curves and Surfaces

Umut Ozertem UMUT@YAHOO-INC.COM
Yahoo! Labs
701 First Ave.
Sunnyvale, CA 94086, USA

Deniz Erdogmus ERDOGMUS@ECE.NEU.EDU
Department of Electrical and Computer Engineering
409 Dana Research Center, 360 Huntington Avenue
Northeastern University
Boston, MA 02115, USA

Editor: Saharon Rosset

Abstract

Principal curves are defined as self-consistent *smooth* curves passing through the *middle* of the data, and they have been used in many applications of machine learning as a generalization, dimensionality reduction and a feature extraction tool. We redefine principal curves and surfaces in terms of the gradient and the Hessian of the probability density estimate. This provides a geometric understanding of the principal curves and surfaces, as well as a unifying view for clustering, principal curve fitting and manifold learning by regarding those as principal manifolds of different intrinsic dimensionalities. The theory does not impose any particular density estimation method can be used with any density estimator that gives continuous first and second derivatives. Therefore, we first present our principal curve/surface definition without assuming any particular density estimation method. Afterwards, we develop practical algorithms for the commonly used kernel density estimation (KDE) and Gaussian mixture models (GMM). Results of these algorithms are presented in notional data sets as well as real applications with comparisons to other approaches in the principal curve literature. All in all, we present a novel theoretical understanding of principal curves and surfaces, practical algorithms as general purpose machine learning tools, and applications of these algorithms to several practical problems.

Keywords: unsupervised learning, dimensionality reduction, principal curves, principal surfaces, subspace constrained mean-shift

1. Introduction

Principal components analysis (PCA) -also known as Karhunen-Loeve Transform- is perhaps the most commonly used dimensionality reduction method (Jolliffe, 1986; Jackson, 1991), which is defined using the linear projection that maximizes the variance in the projected space (Hotelling, 1933). For a data set, principal axes are the set of orthogonal vectors onto which the variance of the projected data points remains maximal. Another closely related property of PCA is that, for Gaussian distributions, the principal line is also self-consistent. That is, any point on the principal line is the conditional expectation of the data on the

orthogonal hyperplane. In fact, this forms the basic idea behind the original principal curve definition by Hastie (1984); Hastie and Stuetzle (1989).

Due to the insufficiency of linear methods for dimensionality reduction, many nonlinear projection approaches have been studied. A common approach is to use a mixture of linear models (Bishop, 1997). Mixture models are attractive, since they arevsimple and analyzable as linear methods; however, assuming a *suitable* model order, they are able to provide much more powerful tools as compared to linear methods. Although model order selection is a tough discrete optimization problem, and mixture methods suffer from the problems introduced by improper selection of model order, there are principled ways to approach this problem such as Dirichlet process mixtures (Ferguson, 1973). Techniques based on local PCA include most well-known examples for mixture models (Fukunaga and Olsen, 1971; Meinicke and Ritter, 1999; Kambhatla and Leen, 1994, 1997).

Another common way of developing nonlinear projections is to use generalized linear models (McCullagh and Nelder, 1989; Fahrmeir and Tutz, 1994). This is based on the idea of constructing the nonlinear projection as a linear combination of nonlinear basis functions. All reproducing kernel Hilbert space techniques such as the well-known kernel PCA (Schölkopf et al., 1998) and kernel LDA (Baudat and Anouar, 2000) belong to this family. The main idea here is to map the data into a high dimensional space and perform the original linear method in this space, where the dot products are computed via a kernel function using the so-called *kernel trick*. More recent methods in this category replace the widely used Gaussian kernel with similarity metrics stemming from a weighted neighborhood graph. These methods are referred to as graph-based kernel methods (Shawe-Taylor and Singer, 2004; Ham et al., 2004).

If the data dimensionality is very high, the most successful methods are manifold learning algorithms, which are based on generating the locality information of data samples using a data proximity graph. Most well known methods that fall into this category include Isomap, local linear embedding, Laplacian eigenmaps, and maximum variance unfolding (Tenenbaum et al., 2000; Roweis and Saul, 2000; Belkin and Niyogi, 2003; Weinberger and Saul, 2006). The idea of defining geodesic distances using the data neighborhood graphs assumes that the graph does not have any *gaps* in the manifold, as well as the graph also does not go *outside* the data manifold. This requires a careful tuning of the parameters of graph construction (K or ϵ, as in the case of most commonly used K-nearest neighbor or ϵ-ball graphs), since the efficiency of the dimensionality reduction methods depend on the quality of the neighborhood graph.

At the time, Hastie and Stuetzle's proposition of self consistent principal curves (Hastie, 1984; Hastie and Stuetzle, 1989) pointed out a different track for nonlinear dimensionality reduction. They defined self-consistency over the *local* conditional data expectations, and generalized the self-consistency property of the principal line into nonlinear structures to introduce the concept of principal curves. Hastie and Stuetzle define the principal curve as *an infinitely differentiable finite length curve that passes through the middle of the data*. Self-consistency means that every point on the curve is the expected value of the data points projecting onto this point.

Hastie and Stuetzle's major theoretical contributions are the following: (i) they show that if a straight line is self-consistent, it is a principal component (ii) based on the MSE criterion, self-consistent principal curves are saddle points of the distance function. They

use this second property to develop an algorithm that starts from the principal line and iteratively finds the principal curve by minimizing the average squared distance of the data points and the curve (Hastie, 1984; Hastie and Stuetzle, 1989). Although they cannot prove the convergence of their algorithm, Hastie and Stuetzle claim that principal curves are by definition a fixed point of their algorithm, and if the projection step of their algorithm is replaced with least squares line fitting, the algorithm converges to the principal line. Since there is no proof of convergence for Hastie-Stuetzle algorithm, existence of principal curves could only be proven for special cases such as elliptical distributions or distributions concentrated around a smooth curve, until Duchamp and Stuetzle's studies on principal curves on the plane (Duchamp and Stuetzle, 1996a,b).

Banfield and Raftery extend the Hastie-Stuetzle principal curve algorithm to closed curves and and propose an algorithm that reduces the estimation bias (Banfield and Raftery, 1992). Tibshirani approaches the problem from a mixture model point-of-view, and provides an algorithm that uses expectation maximization (Tibshirani, 1992). Delicado's proposition uses another property of the principal line rather than self-consistency (Delicado, 1998). Delicado's method is based on the total variance and conditional means and finds the principal curve of *oriented points* of the data set. Stanford and Raftery propose another approach that improves on the outlier robustness capabilities of principal curves (Stanford and Raftery, 2000). Probabilistic principal curves approach, which uses a cubic spline over a mixture of Gaussians to estimate the principal curves/surfaces (Chang and Grosh, 2002), is known to be among the most successful methods to overcome the common problem of bias introduced in the regions of high curvature. Verbeek and coworkers used local principal lines to construct principal curves (Verbeek et al., 2002), and a soft version of the algorithm is also available (Verbeek et al., 2001), known as K-segments and soft K-segments methods.

Algorithmically, Manifold Parzen Windows method (Vincent and Bengio, 2003; Bengio et al., 2006) the most similar method in the literature to our approach. They use a kernel density estimation (and in their later paper, a Gaussian mixture model with a regularized covariance) based density estimate that takes the leading eigenvectors of the local covariance matrices into account. Many principal curve approaches in the literature, including the original Hastie-Stuetzle algorithm, are based on the idea of minimizing mean square projection error. An obvious problem with such approaches is overfitting, and there are different methods in the literature to provide regularization. Kegl and colleagues provide a regularized version of Hastie's definition by bounding the total length of the principal curve to avoid overfitting (Kegl et al., 2000), and they also show that principal curves of bounded length always exist, if the data distribution has finite second moments. Sandilya and Kulkarni define the regularization in another way by constraining bounds on the turns of the principal curve (Sandilya and Kulkarni, 2002). Similar to Kegl's principal curve definition of bounded length, they also show that principal curves with bounded turn always exist if the data distribution has finite second moments. Later, Kegl later applies this algorithm to skeletonization of handwritten digits by extending it into the Principal Graph algorithm (Kegl and Kryzak, 2002). At this point, note that the original Hastie-Stuetzle definition requires the principal curve not to intersect itself, which is quite restrictive, and perhaps, Kegl's Principal Graph algorithm is the only approach in the principal curves literature that can handle self-intersecting data.

Overall, the original principal curve definition by Hastie and Stuetzle forms a strong basis for many, possibly all, principal curve algorithms. The idea of using least squares regression or minimum squared projection error properties of linear principal component analysis to build a nonlinear counterpart brings the problem of overfitting. Hence, algorithms based on these definitions have to introduce a regularization term. Here we take a bold step by defining the principal curves with no explicit smoothness constraint at all; we assume that smoothness of principal curves/surfaces is inherent in the smoothness of the underlying probability density (estimate). Providing the definition in terms of data probability density allows us to link open ended problems of principal curve fitting literature -like optimal regularization constraints and outlier robustness- to well established principles in density estimation literature.

In this paper we emphasize the following messages: (i) principal curves and surfaces are geometrically interesting structures of the theoretical probability distribution that underlies the data as opposed to the particular data set realization, (ii) optimal density estimation (in some sense) does not necessarily result in optimal principal surface estimation. The first point illuminates the fact that one should not seek to solve a problem such as manifold learning without precisely characterizing the sought solution; defining the sought manifold as the solution to one's optimality criterion of choice is incorrect, the solution should be defined geometrically first, and then it should be approximated and its optimality properties should be discovered, leading to optimal approximation algorithms. The second point highlights the fact that a maximum likelihood density estimate, for instance, might not lead to a maximum likelihood estimate of the principal surfaces. Statistically optimal and consistent estimation procedures for the latter must be sought by the community.

The following sections try to address the first issue mentioned above but the second issue will be left as future work; we are confident that the community will eventually propose much better algorithms for identifying principal surfaces than the ones we provide, given the framework presented here. Consequently, the subspace constrained mean shift algorithm presented later is not implied to be optimal in any statistical sense - its choice in this paper is merely due to (i) the familiarity of our audience with the mean shift clustering algorithm (which suffers from all the drawbacks we suffer, such as curse of dimensionality for kernel density estimation), (ii) the fact that it includes parametric mixture distributions as a special case of the formulation (i.e., the same formulas apply to both kernel density and mixture model estimates with minor modifications), (iii) the convergence of the algorithm to a point on the principal surface with appropriate dimensionality is guaranteed for any initial point, since mean-shift is a convergent procedure.

2. Principal Curves/Surfaces

We start with an illustration to give some intuition to our approach, and then we provide a formal definition of the principal curves and surfaces, study special cases and connections to PCA, existence conditions, limitations and ambiguities. All this will be conducted in terms of the gradient and the Hessian of the data pdf, and throughout this section, the data pdf is assumed to be known or can be estimated either parametrically or non-parametrically from the data samples. In the next section we will go back to the data samples themselves while we develop a practical algorithm.

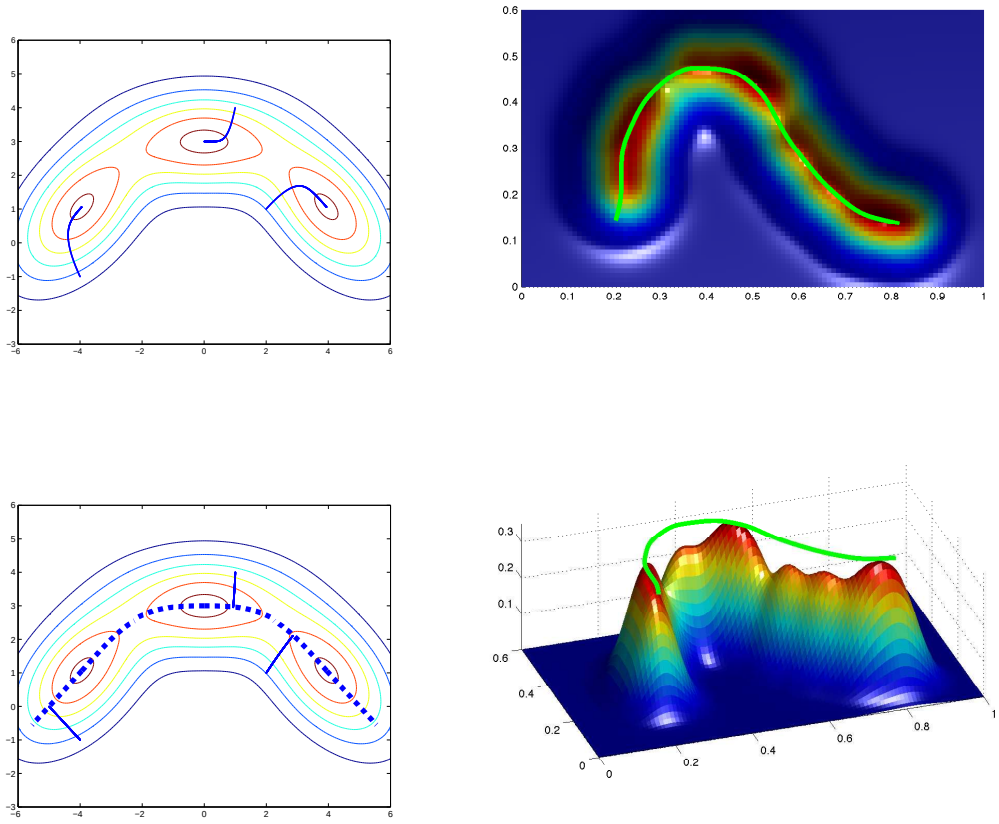


Figure 1: An illustration of the principal curve on a two Gaussian mixtures.

.

2.1 An Illustration

Before we go into the details of the formal definition, we will present a simple illustration. Our principal curve definition essentially corresponds to the *ridge* of the probability density function. Principal curve definitions in the literature are based on local expectations and self-consistency. Hastie's self-consistency principle states that every point on the principal curve is the expected value of the points in the orthogonal subspace of the principal curve at that point - and this orthogonal space rotates along the curve. *In our view, every point on the principal surface is the local maximum, not the expected value, of the probability density in the local orthogonal subspace.*

Consider the modes (local maxima) of the pdf. On the modes, the gradient of the pdf is equal to zero and the eigenvectors of the Hessian matrix are all negative, so that the pdf is decreasing in all directions. The definition of the ridge of the pdf can be given very similarly in terms of the gradient and the Hessian of the pdf. On the ridge of the pdf, one of the eigenvectors of the Hessian is parallel with the gradient. Furthermore, the eigenvalues of the all remaining eigenvectors (which in fact span the orthogonal space of the principal

curve) are all negative, so that the pdf is decreasing in all these directions; hence the point is on a ridge, not in a valley.

Figure 1 presents two illustrations on two Gaussian mixtures. On the left, a comparison of the proposed principal curve projection, and the trajectories of the gradient of the pdf is presented. Consider a Gaussian mixture with 3 components with the pdf contour plot shown. Following the local gradient (top left) essentially coincides with well-known mean shift algorithm (Cheng, 1995; Comaniciu and Meer, 2002), and maps the points to the modes of the pdf, whereas following the eigenvectors of the local covariance (bottom left) gives an orthogonal projection onto the principal curve. The principal curve -the ridge- of this 3-component Gaussian mixture is also shown with the dashed line. On the right, we present the principal curve of a 7-component Gaussian mixture from two different points of view.

2.2 Formal Definition of Principal Curves and Surfaces

We assert that principal surfaces are geometrically well defined structures that underly the theoretical, albeit usually unknown, probability distribution function of the data; consequently, one should define principal surfaces with the assumption that the density is known - finite sample estimators of these surfaces is a question to be answered based on this characterization. Inspired by differential geometry where principal lines of curvature are well-defined and understood, we define the principal curves and surfaces in terms of the first and second order derivatives of the *assumed probability density function*. Next, we define critical, principal, and minor surfaces of all dimensions and point out facts relating to these structures - proofs are generally trivial and are omitted for most statements.

Given a random vector $\mathbf{x} \in \mathbb{R}^n$, let $p(\mathbf{x})$ be its pdf, $\mathbf{g}(\mathbf{x})$ be the transpose of the local gradient, and $\mathbf{H}(\mathbf{x})$ be the local Hessian of the probability density function. To avoid mathematical complications, we assume that the data distribution $p(\mathbf{x}) > 0$ for all $\mathbf{x}$, and is at least twice differentiable. Also let $\{(\lambda_1(\mathbf{x}), q_1(\mathbf{x})), \ldots, (\lambda_n(\mathbf{x}), q_n(\mathbf{x}))\}$ be the eigenvalue-eigenvector pairs of $\mathbf{H}(\mathbf{x})$, where the eigenvalues are sorted such that $\lambda_1(\mathbf{x}) > \lambda_2(\mathbf{x}) > \ldots > \lambda_n(\mathbf{x})$ and $\lambda_i \neq 0$.[1]

Definition 1. A point $\mathbf{x}$ is an element of the d-dimensional critical set, denoted by $\mathcal{C}^d$ iff the inner product of $\mathbf{g}(\mathbf{x})$ with at least (n-d) eigenvectors of $\mathbf{H}(\mathbf{x})$ is zero.

The definition above is an intentional extension of the familiar notion of critical points in calculus; thus local maxima, minima, and saddle points of the pdf become the simplest special case.

Fact 1. $\mathcal{C}^0$ consists of and only of the critical points (where gradient is zero) of $p(\mathbf{x})$. Furthermore, $\mathcal{C}^d \subset \mathcal{C}^{d+1}$.

In practice, this fact points to the possibility of designing dimension reduction algorithms where each data is projected to a critical manifold of one lower dimension sequentially (deflation). Alternatively, one could trace out critical curves starting off from critical points (inflation). This property of linear PCA has been extensively used in the design of on-line algorithms in the 90's (Kung et al., May 1994; Wong et al., 2000; Hegde et al., 2006).

1. Strict inequalities are assumed here for the theoretical analysis, because in the case of repeated eigenvalues local uncertainties similar to those in PCA will occur. We also assume non-zero eigenvalues for the Hessian of the pdf. These assumptions are not critical to the general theme of the paper and generalized conclusions can be relatively easily obtained. These ambiguities will later be discussed in Section 2.6.

Definition 2. A point $\mathbf{x} \in \mathcal{C}^d - \mathcal{C}^{d-1}$ is called a regular point of $\mathcal{C}^d$. Otherwise, it is an irregular point.

Fact 2. If $\mathbf{x}$ is a regular point of $\mathcal{C}^d$, then there exists an index set $I_\perp \subset \{1, \ldots, n\}$ with cardinality $|I_\perp| = (n-d)$ such that $\langle \mathbf{g}(\mathbf{x}), \mathbf{q}_i(\mathbf{x}) \rangle = 0$ iff $i \in I_\perp$. If $\mathbf{x}$ is an irregular point of $\mathcal{C}^d$, then $|I_\perp| > (n-d)$.

Regular points of a critical set are the set of points that are not also in the lower dimensional critical sets. At regular points, the gradient is orthogonal to exactly $(n-d)$ eigenvectors of the Hessian, thus these points locally lie on a surface with an intrinsic dimensionality of d. Naturally, these surfaces have their tangent and orthogonal spaces locally.

Definition 3. Let $\mathbf{x}$ be a regular point of $\mathcal{C}^d$ with $I_\perp$. Let $I_\parallel = \{1, \ldots, n\} - I_\perp$. The tangent subspace is $\mathcal{C}^d_\parallel(\mathbf{x}) = span\{\mathbf{q}_i(\mathbf{x}) | i \in I_\parallel\}$ and the normal/orthogonal subspace is $\mathcal{C}^d_\perp(\mathbf{x}) = span\{\mathbf{q}_i(\mathbf{x}) | i \in I_\perp\}$.

Definition 4. A regular point $\mathbf{x}$ of $\mathcal{C}^d$ with $I_\perp$ is (assuming no zero-eigenvalues exist for simplicity):

1. a regular point in the principal set $\mathcal{P}^d$ iff $\lambda_i(\mathbf{x}) < 0 \ \forall i \in I_\perp$; that is, $\mathbf{x}$ is a local maximum in $\mathcal{C}^d_\perp(\mathbf{x})$.

2. a regular point in the minor set $\mathcal{M}^d$ iff $\lambda_i(\mathbf{x}) > 0 \ \forall i \in I_\perp$; that is, $\mathbf{x}$ is a local minimum in $\mathcal{C}^d_\perp(\mathbf{x})$.

3. a regular point in the saddle set $\mathcal{S}^d$ otherwise; that is, $\mathbf{x}$ is a saddle in $\mathcal{C}^d_\perp(\mathbf{x})$.

Regular and irregular points in these special cases are defined similarly. Also, tangent and orthogonal subspaces are defined identically.

Clearly, $(\mathcal{P}^d, \mathcal{M}^d, \mathcal{S}^d)$ is a partition of $\mathcal{C}^d$. In practice, while principal surfaces might be useful in dimension reduction as in manifold learning, minor surfaces, valleys in the probability density function, can be useful in semi-supervised learning. A common theme in semi-supervised learning employs the so-called cluster hypothesis, where the valleys in the data probability density function have to be identified (Chapelle et al., 2006), like in the well-known Low Density Separation algorithm (Chapelle and Zien, 2005). Note that allowing zero-eigenvalues would result in local plateaus in pdf, and allowing repeated eigenvalues would result in ill-defined regular points. While conceptually the consequences are clear, we avoid discussing all possible such circumstance for now for the sake of simplicity. We give a detailed discussion on these limitations in Section 2.6.

By construction, we have $\mathbf{x} \in \mathcal{P}^0$ iff $\mathbf{x}$ is a local maximum of $p(\mathbf{x})$; $\mathbf{x} \in \mathcal{M}^0$ iff $\mathbf{x}$ is a local minimum of $p(\mathbf{x})$; $\mathbf{x} \in \mathcal{S}^0$ iff $\mathbf{x}$ is a saddle point of $p(\mathbf{x})$. Furthermore, $\mathcal{P}^d \subset \mathcal{P}^{d+1}$ and $\mathcal{M}^d \subset \mathcal{M}^{d+1}$.[2] In mean shift clustering, projections of data points to $\mathcal{P}^0$ are used to find the solution (Cheng, 1995; Comaniciu and Meer, 2002). In fact, the *attraction basin*[3] of each mode of the pdf can be taken as a local chart that has a curvilinear orthogonal

2. Observe this inclusion property by revisiting Figure 1, as the major principal curve (show in the figures on the right) passes through all local maxima of the Gaussian mixture density.
3. The attraction basin is defined as the set of points in the feature space such that initial conditions chosen in this set evolve to a particular attractor -modes of the pdf for this particular case. In the context of mean-shift the underlying criterion is the KDE of the data. In this case, attraction basins are regions bounded by minor curves, and the attractors are the modes of the pdf.

coordinate system determined by the eigenvectors of the Hessian of the pdf (or a nonlinear function of it - consequences of the choice of the nonlinear function will be discussed soon).

Note that the definitions and properties above allow for piecewise smooth principal surfaces and opportunities are much broader than techniques that seek a *globally smooth optimal manifold*, which does not generally exist according to our interpretation of the geometry. Figure 2 illustrates a simple density where a globally smooth curve (for instance a principle line) can not provide a satisfactory underlying manifold; in fact such a case would likely be handled using local PCA - a solution which essentially approximates the principal curve definition we advocate above.

At this point we note that due to the assumption of a second-order continuously differentiable pdf model, the Hessian matrix and its eigenvectors and eigenvalues are continuous everywhere. Consequently, at any point on the d-dimensional principal set (or critical or minor sets) in a small open ball around this point, the points in the principal set form a continuous surface. Considering the union of open balls around points in the $d-1$-dimensional principal surface, we can note that the continuouty of the d-dimensional surface implies continuity of the $d-1$-dimensional subsurface as well as the 1-dimensional projection trajectories in the vicinity. Furthermore, if we assume that the pdf models are three-times continuously differentiable, the projection trajectories (following local Hessian eigenvectors) are not only locally continuous, but also locally continuously differentiable). In general, the order of continuous differentiability of the underlying pdf model is reflected to the emerging principal surfaces and projection trajectories accordingly.

2.3 Principal Surfaces of a Nonlinear Function of the PDF

In this section we show that for a pdf the *set of points* that constitute $\mathcal{P}^d$ is identical to the *set of points* that constitute $\mathcal{P}_f^d$ of the function $f(p(\mathbf{x}))$ where $f(\xi)$ is monotonically increasing. The same conclusion can be drawn and shown similarly for the minor and critical surfaces; details of this will not be provided here.

Consider $\mathbf{x}$, a regular point of $\mathcal{P}^d$ with pdf $p(\mathbf{x})$ and its gradient-transpose $\mathbf{g}(\mathbf{x})$ and Hessian $\mathbf{H}(\mathbf{x})$. Then, the eigenvectors and eigenvalues of the Hessian at this point can be partitioned into the parallel and orthogonal subspace contributions: $\mathbf{H}(\mathbf{x}) = \mathbf{Q}_\parallel \Lambda_\parallel \mathbf{Q}_\parallel^T + \mathbf{Q}_\perp \Lambda_\perp \mathbf{Q}_\perp^T$, where the parallel subspace is spanned by d eigenvectors in the columns of $\mathbf{Q}_\parallel$ and the orthogonal subspace is spanned by $(n-d)$ eigenvectors in $\mathbf{Q}_\perp$. At a regular point the gradient is in the tangent space, therefore, $\mathbf{g}(\mathbf{x}) = \mathbf{Q}_\parallel \beta$ for some suitable vector β of linear combination coefficients. The gradient-transpose and Hessian of the function $f(p(\mathbf{x}))$ are:

$$\begin{aligned}
\mathbf{g}_f(\mathbf{x}) &= f\prime(p(\mathbf{x}))\mathbf{g}(\mathbf{x}) \\
&= f\prime(p(\mathbf{x}))\mathbf{Q}_\parallel \beta \ , \\
\mathbf{H}_f(\mathbf{x}) &= f\prime(p(\mathbf{x}))\mathbf{H}f(\mathbf{x}) + f\prime\prime(p(\mathbf{x}))\mathbf{g}(\mathbf{x})\mathbf{g}^T(\mathbf{x}) \\
&= \left(f\prime(p(\mathbf{x}))\mathbf{Q}_\parallel \Lambda_\parallel \mathbf{Q}_\parallel^T + f\prime\prime(p(\mathbf{x}))\mathbf{Q}_\parallel \beta\beta^T \mathbf{Q}_\parallel^T\right) + f\prime(p(\mathbf{x}))\mathbf{Q}_\perp \Lambda_\perp \mathbf{Q}_\perp^T \ .
\end{aligned}$$

We observe that at $\mathbf{x}$ the gradient $\mathbf{g}_f(\mathbf{x})$ is also in the original d-dimensional tangent space. Further, the orthogonal subspace and the sign of its eigenvalues remain unchanged (since $f\prime(\xi) > 0$). This shows that if $\mathbf{x}$ is a regular point of $\mathcal{P}^d$, then it is also a regular point of

$\mathcal{P}_f^d$. The converse statement can also be shown by switching the roles of the two functions and considering the inverse of f as the nonlinear mapping.

Note that we simply proved that the principal surface (as a set of points) of a given dimension remains unchanged under monotonic transformations of the pdf. If one projects points in higher dimensional surfaces to lower dimensional principal surfaces following trajectories traced by the Hessian of $f(p((x)))$, these projection trajectories will depend on f. This brings us to the connection with PCA.

2.4 Special Case of Gaussian Distributions, Connections to PCA

For a jointly Gaussian pdf, choosing $f(\xi) = \log(\xi)$ yields a quadratic function of $\mathbf{x}$, thus the local Hessian $\mathbf{H}_{log}(\mathbf{x}) = -(1/2)\Sigma^{-1}$ becomes independent of position. Consequently, the local Hessian eigendirections form linear trajectories and principal surfaces become hyperplanes spanned by the eigenvectors of the Gaussian's covariance matrix. If this connection to PCA is desired, that is, if the density becomes Gaussian, principal surface projections of points coincide with those one would obtain via linear PCA, then the choice log $p(\mathbf{x})$ becomes attractive. Otherwise, one can seek choices of f that brings other benefits or desirable properties. For this reason, using log as the nonlinearity, we introduce the concept of local covariance matrix.

Definition 5. The local covariance-inverse of a pdf at any point $\mathbf{x}$ is given by -2 times the Hessian of the logarithm of the pdf. Specifically, in terms of the gradient-transpose and the Hessian of the pdf, this corresponds to $\mathbf{\Sigma}^{-1}(\mathbf{x}) = -p^{-1}(\mathbf{x})\mathbf{H}(\mathbf{x}) + p^{-2}\mathbf{g}(\mathbf{x})\mathbf{g}^T(\mathbf{x})$. If we assume that its eigenvalue-vector pairs are $\{\gamma_i(\mathbf{x}), \mathbf{v}_i(\mathbf{x})\}$ for $i \in \{1, \ldots, n\}$ and if the eigenvalues (some of which might be negative) are sorted as follows: $\gamma_1 < \ldots < \gamma_n$, the local ordering of critical directions from most principal to least follows the same indexing scheme (i.e., γ_n is the first to go when projecting to lower dimensions).

2.5 Existence of Principal Curves and Surfaces

Considering Hastie's principal curve definition, the existence proof of principal curves is limited to some special cases, such as elliptical or spherical distributions concentrated around a smooth curve. It should also be noted that this definition of the principal curve requires the principal curve not to intersect itself. The principal curve definition of Kegl et al. (2000) and Sandilya and Kulkarni (2002) are theoretically more appealing in this context, since by their definition, the principal curve always exists if the distribution has finite second moments.

According to our definition, the principal curve exists as long as the data probability density is twice differentiable, such that the Hessian is nonzero. There is no restriction of finite moments, which is an improvement on existing methods. However, also note that by our definition the principal curve does not exist for uniform distributions.[4] In practice, however, since we will build our algorithms based on KDE with Gaussian kernels or GMM, even if the true underlying distribution is uniform, KDE or GMM guarantee that the gradient and Hessian are continuous.

4. Note that one can always convolve a distribution with a spherical Gaussian or other circularly symmetric unimodal kernel to introduce continuous first and second derivatives without distorting the geometry of the principal surfaces.

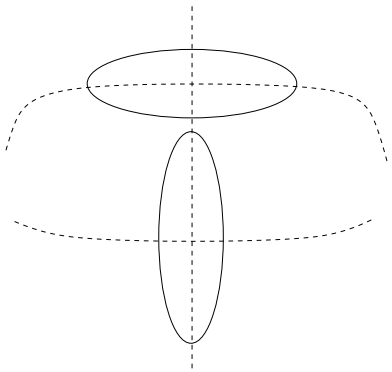


Figure 2: A T-shaped Gaussian mixture

.

2.6 Local Ranking of the Principal Curves and Ambiguities

In PCA, the ordering of the principal component directions are naturally given by sorting the corresponding eigenvalues of the covariance matrix in a descending order. Note that, since it coincides with PCA for Gaussian distributions, our principal curve definition also has the ambiguity that occurs in PCA; the principal surface of a spherically symmetric distribution is not well-defined.

Conditional expectation or mean squared projection error based definitions have driven the principal curves research, but in general, the definition is limited to the nonlinear counterpart of the first principal component. In fact, there is no definition of *second, third, etc. principal curve* in the literature that we are aware of. Considering the connection to PCA, one can see that our principal curve definition is not limited to the nonlinear counterpart of the first principal component, under the assumption that the Hessian matrix has distinct eigenvalues, one can obtain the *local* ordering for any d-dimensional principal manifold.

In general, data densities may take complex forms and counterintuitive scenarios may arise. Hence, generally, local information may not always indicate the global rank, and a global ordering in a principal set of given dimensionality may not be possible. To illustrate this fact, consider again the T-shaped Gaussian mixture consisting of two components. Note that both branches of this principal graph correspond to the leading eigenvector of the local covariance at different portions of the feature space and a global ranking is not possible.

3. Subspace Constrained Mean Shift (SCMS)

Consider the fact that $\mathcal{P}^0$, principal surface of dimensionality zero, is by construction the local maxima points of the $p(\mathbf{x})$. This presents a strong connection to clustering, since mapping to the local maxima points of the data pdf is a widely accepted clustering solution, achieved by the well-known mean shift algorithm (Cheng, 1995; Comaniciu and Meer, 2002). In this section we present a subspace constrained likelihood maximization idea that stems from Definition 4; a point on $\mathcal{P}^d$ is a local maximum in the orthogonal space. We provide an algorithm which is very similar to mean-shift in spirit. This lays an algorith-

mic connection between clustering and principal curve/surface fitting that accompanies the theoretical connection.

Mean-shift assumes an underlying KDE probability density of the data and implements a fixed-point iteration that maps the data points to the closest mode (local maximum) of the pdf, and the mean-shift update at any point on the feature space is parallel with the gradient of the KDE (Cheng, 1995; Comaniciu and Meer, 2002). A point is on the one dimensional principal surface iff the local gradient is an eigenvector of the local Hessian -since the gradient has to be orthogonal to the other $(n-1)$ eigenvectors- and the corresponding $(n-1)$ eigenvalues are negative. Again via the same underlying KDE assumption, a simple modification of the mean-shift algorithm by constraining the fixed-point iterations in the orthogonal space of corresponding $(n-1)$ eigenvector directions at the current point in the trajectory leads to an update that converges to the principal curves and not to the local maxima. For this case, the orthogonal space of corresponding $(n-1)$ eigenvector directions of the local covariance is the parallel space of the leading eigenvector of the local covariance. The algorithm could be modified to converge to the d-dimensional principal manifold P^d trivially, by selecting the constrained subspace as the subspace spanned by corresponding $(n-d)$ eigenvectors of the local covariance to constrain the mean-shift iterations into the subspace spanned by d leading eigenvectors of the local covariance. To provide both parametric and nonparametric variations, we will present an algorithm that can be used for well-known KDE and GMM density estimators.

Consider the data samples $\{\mathbf{x}_i\}_{i=1}^N$, where $\mathbf{x}_i \in \Re^n$. The KDE of this data set (using Gaussian kernels) is given as

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

where Σ_i is the kernel covariance for $\mathbf{x}_i$; $G_{\Sigma_i}(\mathbf{y}) = C_{\Sigma_i} e^{-\mathbf{y}^T \Sigma_i^{-1} \mathbf{y}/2}$. Note that for in (1) we use the general case of anisotropic variable (data-dependent) kernel functions. For isotropic kernels one can use a scalar value instead of a full covariance, or for fixed kernel functions one can constrain the data dependency and drop the sample index i. Again for the general case, the gradient and the Hessian of the KDE are

$$\begin{aligned} \mathbf{g}(\mathbf{x}) &= -N^{-1} \sum_{i=1}^{N} c_i \mathbf{u}_i \ , \\ \mathbf{H}(\mathbf{x}) &= N^{-1} \sum_{i=1}^{N} c_i (\mathbf{u}_i \mathbf{u}_i^T - \Sigma_i^{-1}) \ , \\ \mathbf{\Sigma}^{-1}(\mathbf{x}) &= -p^{-1}(\mathbf{x}) \mathbf{H}(\mathbf{x}) + p^{-2} \mathbf{g}(\mathbf{x}) \mathbf{g}^T(\mathbf{x}) \\ \textit{where} \quad \mathbf{u}_i &= \Sigma_i^{-1}(\mathbf{x} - \mathbf{x}_i) \quad \textit{and} \quad c_i = G_{\Sigma_i}(\mathbf{x} - \mathbf{x}_i) \ . \end{aligned} \tag{2}$$

Let $\{(\gamma_1(\mathbf{x}), \mathbf{v}_1(\mathbf{x})), \ldots, (\gamma_n(\mathbf{x}), \mathbf{v}_n(\mathbf{x}))\}$ be the eigenvalue-eigenvector pairs of $\mathbf{\Sigma}^{-1}(\mathbf{x})$ as defined in (2) ordered from smallest to largest and the mean-shift update emerging from (2) is

$$\mathbf{x} \leftarrow \mathbf{m}(\mathbf{x}) = (\textstyle\sum_{i=1}^{N} c_i \Sigma_i^{-1})^{-1} \sum_{i=1}^{N} c_i \Sigma_i^{-1} \mathbf{x}_i \tag{3}$$

1. Initialize the trajectories to a mesh or data points and set $t = 0$. Input the Gaussian kernel bandwidth σ (or kernel covariance matrix for anisotropic Gaussian kernels) to the algorithm.
2. For every trajectory evaluate $\mathbf{m}(\mathbf{x}(t))$ using (2) and (3).
3. Evaluate the gradient, the Hessian, and perform the eigendecomposition of $\mathbf{\Sigma}^{-1}(\mathbf{x}(t)) = \mathbf{V}\mathbf{\Gamma}\mathbf{V}$ (in specific cases, the full eigendecomposition could be avoided).
4. Let $\mathbf{V}_\perp = [\mathbf{v}_1 \dots \mathbf{v}_{n-d}]$ be the $(n-d)$ largest eigenvectors of Σ^{-1}
5. $\tilde{\mathbf{x}}(k) = \mathbf{V}_\perp \mathbf{V}_\perp^T \mathbf{m}(\mathbf{x})$
6. If $\mid \mathbf{g}^T(\mathbf{x})\mathbf{V}_\perp^T\mathbf{g}(\mathbf{x}) \mid /(\|\mathbf{g}(\mathbf{x})\| \cdot \|\mathbf{V}_\perp^T\mathbf{g}(\mathbf{x})\|) < \epsilon$ then *stop*, else $\mathbf{x}(t+1) \leftarrow \tilde{\mathbf{x}}$, increment t and go to step 2.

Table 1: KDE-based SCMS Algorithm

At $\mathbf{x}$, the subspace mean-shift update is performed by projecting $\mathbf{x}$ into the constrained space $\tilde{\mathbf{x}}_k = (\mathbf{V}_\perp \mathbf{V}_\perp^T \mathbf{m}(\mathbf{x}))$. The stopping criterion can be constructed from definition directly to check if the gradient is orthogonal to the subspace spanned by the selected $n-d$ eigenvectors when projecting the data from n to d dimensions: $\mid \mathbf{g}^T(\mathbf{x})\mathbf{V}_\perp^T\mathbf{g}(\mathbf{x}) \mid /(\|\mathbf{g}(\mathbf{x})\| \cdot \|\mathbf{V}_\perp^T\mathbf{g}(\mathbf{x})\|) < \epsilon$. For the special case of $d = 1$, an equivalent stopping criterion is that the gradient becomes an eigenvector of the Hessian, so one can employ: $\mid \mathbf{g}^T(\mathbf{x})\mathbf{H}\mathbf{g}(\mathbf{x}) \mid /(\|\mathbf{g}(\mathbf{x})\| \cdot \|\mathbf{H}\mathbf{g}(\mathbf{x})\|) > 1 - \epsilon$. Alternatively, the more traditional (but rather more risky) stopping criterion of $\|\tilde{\mathbf{x}}_k - \mathbf{x}_k\| < \epsilon$ can be used.

The iterations can be used to find the principal curve projection of any arbitrary point of interest in the feature space.[5] To find the principal curve projections of the data samples, a suitable way is to initialize the projection trajectories to the data samples themselves, as in mean-shift clustering. The general version of SCMS algorithm that converges to the d-dimensional principal manifold is presented in Table 1, and SCMS principal curve algorithm can simply be obtained by setting $d = 1$.

Following the derivation of the KDE with Gaussian kernel functions, using SCMS for GMM density estimates is trivial, by replacing the data samples with Gaussian mixture centers and the kernel bandwidth/covariance with the Gaussian mixture bandwidth/covariances. From now on, we will refer these as KDE-SCMS and GMM-SCMS, and we will present results based on both KDE and GMM density estimates in the next section.

3.1 Properties of KDE-SCMS

Before proceeding to experiments we would like to briefly discuss some properties of SCMS. We believe these properties are important since they connect many open-ended questions in principal curves literature to well-studied results in density estimation. Outlier robustness

5. Note that these fixed-point-update-based projections are relatively coarse approximations and more accurate projections can be obtained via numerical integration of the corresponding differential equations, for instance using Runge-Kutta order-4 method.

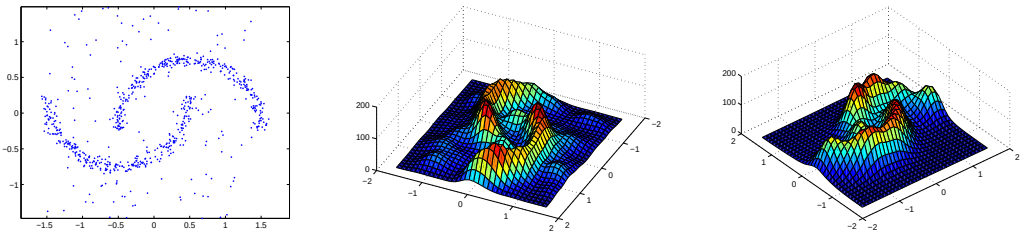


Figure 3: Curves buried in noise (left) and finite bandwidth (middle) and variable bandwidth (right) KDE

and regularization properties are just some examples of the properties that are adopted from the particular density estimation method -KDE in our case. Similar algorithms that stem from the definitions in Section 2 can be designed for other density estimation methods as well. The properties presented here are a few of many possibilities to illustrate the connections.

3.1.1 Computational Load

The computational complexity of KDE-SCMS is $O(N^2 \times n^3)$, where N is the number of samples, and n is the data dimensionality. The n^3 dependency comes from the eigendecomposition of the Hessian matrix. For GMM-SCMS, the complexity becomes $O(N \times m \times n^3)$, where m is the number of Gaussians in the mixture density estimate.[6] Note that the computational load required by SCMS is only slightly higher than the mean-shift algorithm that has been practically implemented in many application domains. The literature is rich in approaches to accelerate mean shift, all of which are directly applicable for our algorithm as well. These methods vary from simple heuristics to more principled methods like Fast Gaussian Transform (Yang et al., 2003b), quasi-Newton methods (Yang et al., 2003a) or Gaussian Blurring Mean Shift (Carreira-Perpinan, 2006). The cubic computational dependency may become the bottleneck for very high dimensional data. One solution to this problem might be to look for the d leading eigenvalues of the Hessian matrix sequentially, instead of the full eigendecomposition (as in Hegde et al., 2006), which will drop the complexity down to $O(N^2 \times d^3)$ where d is the target dimensionality ($d = 1$ for principal curves). However note that if this is the case, the computational bottleneck is not the only problem. If we have $d^3 \gg N^2$, density estimation will also suffer from *the curse of dimensionality* and our approach -that is based on the density estimation- will fail. In the experimental result section we will show results with such high dimensional data.

3.1.2 Statistical Consistency

Our algorithmic approach has used the powerful kernel density estimation (KDE) technique to estimate principal surface structures that underly data densities. The convergence prop-

6. Note that this excludes the computational load required for the expectation-maximization training to fit the GMM.

erties of KDE have been well understood and bandwidth selection, especially in the case of fixed-bandwidth models, have been rigorously investigated leading to a variety of criteria for KDE construction with optimal density estimate convergence properties. In principal surfaces, however, we rely on the accurate estimation of the first and second derivatives of the multivariate data density along with the density itself. Consequently, an important question that one needs to ask (the authors thank the reviewer who posed this question) is whether the first and second order derivatives of the KDE will converge to the true corresponding derivatives, thus leading to the convergence of the principal surface structures of the KDE to those of the actual data density. Literature on the convergence properties of KDE in estimating derivatives of densities is relatively less developed - however, some work exists on the convergence of KDE derivatives in probability using isotropic kernels with dependent data and general bandwidth sequences (Hansen, 2008). In particular, a timely result on general kernel bandwidth matrices for fixed-bandwidth KDE derivatives, albeit slightly more restrictive since it uses convergence in the mean squared error sense, partly answers this question for us under relatively reasonable assumptions considering typical machine learning applications involving manifold learning (Chacon et al., 2011).

Specifically and without going into too much detail, Chacon et al. (2011) demonstrate that under the assumptions that the (unstructured but fixed) kernel bandwidth matrix converges to zero fast enough, and the underlying density and the kernel have continuous square integrable derivatives up to the necessary order or more (density must have square integrable derivatives 2 orders more than the kernel), and that the kernel has a finite covariance, the integrated mean squared error between the vector of order-r derivatives of the KDE converge to those of the true density of the data (from Theorems 1-3). The order of convergence for the integrated mean squared error has been given, from Theorems 2 & 3, as: $o(n^{-4/(d+2r+4)}) + o(n^{-1}|\mathbf{H}|^{-1/2}tr^r(\mathbf{H}^{-1}) + tr2\mathbf{H})$

This demonstrates that as the number of samples N goes to infinity, given a *sufficiently smooth* density and kernel, the derivatives will also converge. Consequently, principal surfaces characterized by first and second derivatives as in our definition will also converge.

3.1.3 Outlier Robustness

Outlier robustness is another key issue in principal curve literature. Principal curve definitions that involve conditional sample expectations and mean squared projection error do not incorporate any data likelihood prior; hence, they treat each data sample equally. Such approaches are known to be sensitive to noise, and presence of outlier data samples, of course, will bias the principal curve towards outliers. Stanford and Raftery present an algorithm that improves upon the outlier robustness of the earlier approaches (Stanford and Raftery, 2000).

Outlier robustness is a well-known property of variable bandwidth KDE. In this approach, a data dependent kernel function is evaluated for each sample such that the width of the kernel is directly proportional with the likelihood that sample is an outlier. This can be implemented in various 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. Hence, the kernel bandwidth increases for the samples that are in a sparse neigh-

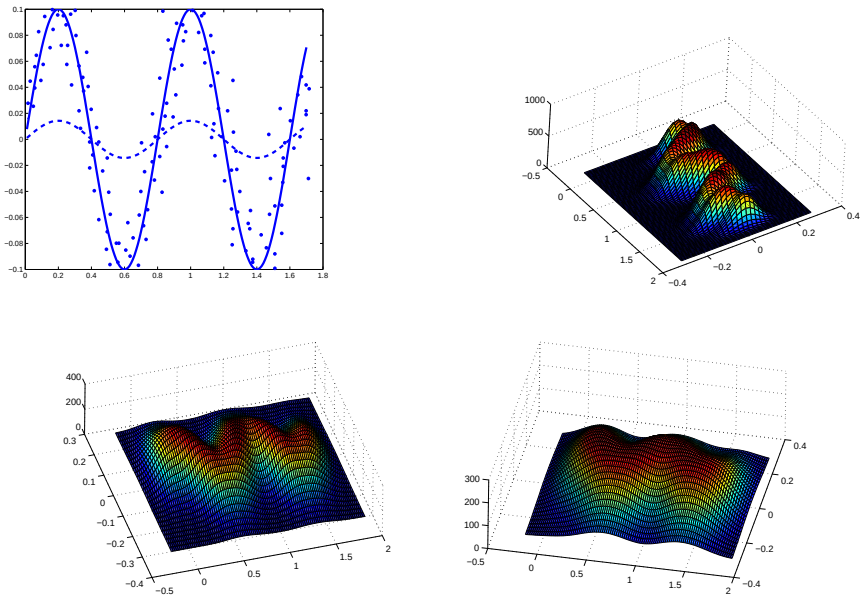


Figure 4: Mean projection error vs. overfitting tradeoff as a kernel bandwidth selection problem. Three density estimates are presented - a narrow bandwidth (left) Maximum Likelihood kernel bandwidth (middle) and a wide kernel bandwidth (right)

borhood of data samples. Figure 3 (left) presents a data set consisting of two crescent-like clusters buried in noise. In fact, this data set is similar to the illustration that Stanford and Raftery use as they propose their noise robust principal curve approach (Stanford and Raftery, 2000). We present the fixed and variable bandwidth -using K-nearest neighbor method (i) mentioned above and selecting $K = N^{1/4}$- KDE of the data set in Figure 3 in middle and right, respectively. Note that in the resulting density estimate the variable size KDE eliminates the effects of the outliers without oversmoothing or distorting the pdf significantly in the support of the data. Selecting the kernel functions in a data dependent manner, can make KDE-based SCMS robust to outliers in the data. However, additional computational load of variable kernel bandwidth evaluations may increase the overall computational complexity.

3.1.4 Regularization and Overfitting

If a problem is formulated over sample expectations or minimization of the average projection error, the issue of overfitting arises. In the context of principal curves and surfaces, most explicitly, Kegl brings up this question in his PhD dissertation (Kegl, 1999). Considering the data set and principal curves in Figure 4 (left), Kegl asks, which of the curves is the right one. "Is the solid curve following the data too closely, or is the dashed curve

generalizing too much?" In general, of course, this is an open ended question and the answer depends on the particular application.

Still, density estimation methods can provide many approaches to define the regularization, varying from heuristics to theoretically well-founded approaches like maximum likelihood. In other words, instead of trying for different length (Kegl et al., 2000) or curvature (Sandilya and Kulkarni, 2002) parameters, density estimation can provide purely data-driven approaches, where the regularization parameters are learned from the data directly using cross-validation.

Figure 4 shows density estimates obtained using KDE for different kernel bandwidth selections for the data set presented. In SCMS, the trade-off between projection error and overfitting can be adjusted by setting the kernel width. One can select the kernel bandwidth manually by observing the data or exploiting domain specific knowledge. This is, of course, not much different than observing the data and selecting a suitable length or curvature constraint. However, the real advantage here is the rich literature on how to select the kernel function from the data samples directly. There are many theoretically well-founded ways of optimizing the kernel width according to maximum likelihood or similar criteria (Silverman, 1986; Parzen, 1962; Comaniciu, 2003; Sheather and Jones, 1991; Jones et al., 1996; Raykar and Duraiswami, 2006).

Furthermore, anisotropic and/or variable size kernel functions naturally implement many types of constraints that cannot be defined by any length or bound of turn. By selecting anisotropic kernel functions, one can define the regularization constraint at different scales along different directions. This can also be achieved by lenghth/curvature constraints by scaling the data differently among different dimensions. However, data-dependent variable bandwidth kernels can define varying constraints throughout the space. This is not possible to achieve by a constant curvature or length penalty of any sort.

In summary, our KDE based principal curve projection algorithm 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, which cannot be given by any length or curvature constraint whatsoever. Although this still cannot ultimately answer the open-ended question on the trade-off between the regularization and projection error, it provides a principled way to approach the problem and proves to be effective in many real applications as we will show next.

4. Experimental Results

This section consists of three parts. In the first part, we provide comparisons with some earlier principal curve algorithms in the literature. We perform simulations on notional data sets and give performance and computation times. In the second part, we focus on real applications, where we briefly mention some applications with pointers to our recent publications and also provide results in some areas that principal curves has (feature extraction for OCR) or has not been (time-frequency distribution sharpening, MIMO channel equalization) used before. In these applications we use SCMS *directly*. Surely, pre- and post-processing steps can be added to improve performance of these applications, however our aim is to show the versatility of the approach not to optimize every implementation detail. In the third and final part, we focus on the limitations of the method.

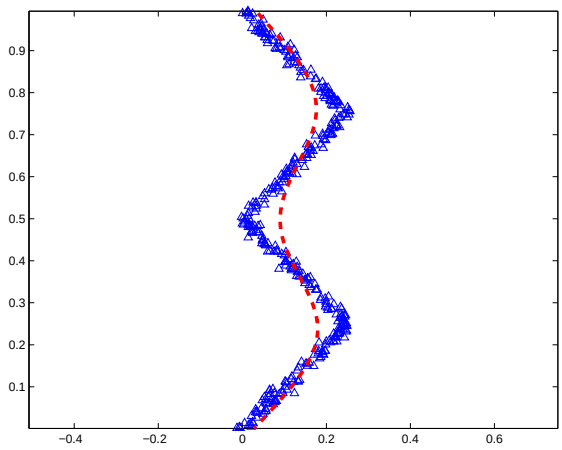


Figure 5: Zig-zag data set, and Hastie-Stuetzle principal curve

Same as the principal line, principal curves -in our definition- extend to infinity. In general though, what one is really interested in is not the whole structure, but the projections of samples onto the underlying structure. Therefore, throughout this section, rather than populating samples on the curve that extend to infinity, we prefer representing the principal curve with *the data samples projected onto the principal curve*, so that the curves in the plots remain in the support of the data. For the same reason, although the underlying structure is continuous (and can be populated into any desired density), the presented principal curves sometimes do not *look* continuous where the data is sparse.

4.1 Comparisons with Other Principal Curve Methods

In this section we present comparisons with original Hastie-Stuezle principal curve method (Hastie, 1984; Hastie and Stuetzle, 1989) and the Polygonal Line Algorithm by Kegl et al. (Kegl et al., 2000), and we provide both computation time and performance comparisons.

4.1.1 Zig-Zag Data Set

Zig-Zag data set has been used in an earlier principal curve paper by Kegl et al. (2000) (This data set is provided by Kegl). Figure 5 shows the data samples and result of Hastie's algorithm. Figure 6 presents the results of Kegl's polygonal line algorithm for different penalty coefficients. The length penalty coefficient is equal to 0.1, 0.3, 0.5, and 0.7, respectively. Polygonal Line algorithm with the right length penalty seems to be working the best for this dataset with high curvature on the corners.

In Figure 7 we compare results of the SCMS algorithm based on three different density estimates: (i) KDE with constant bandwidth, (ii) KDE with variable (data-dependent) covariance (iii) Gaussian mixture with 4 components. For (i) and (ii), the bandwidth and covariance of the Gaussian kernel are selected according to the leave-one-out maximum likelihood criterion (Duda et al., 2000). For the Gaussian mixture model, the *correct* model

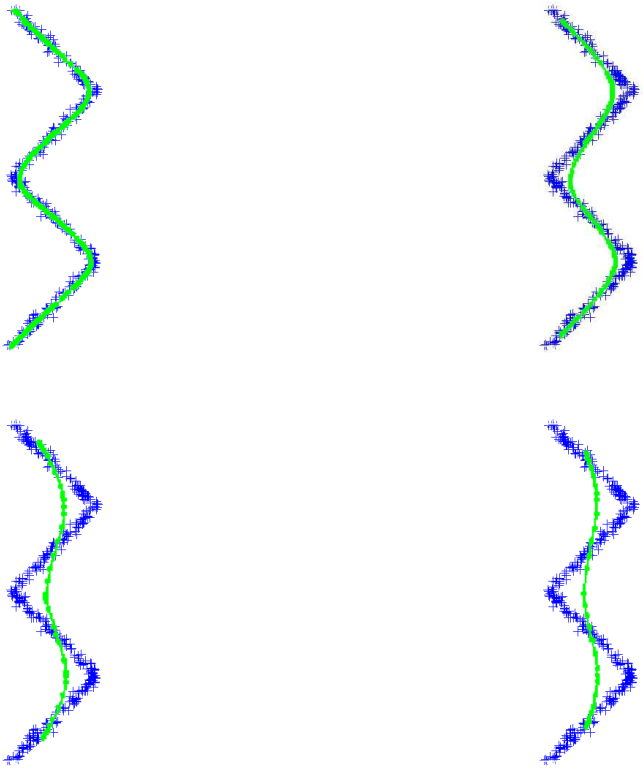


Figure 6: Zig-zag data set, and result of the Polygonal Line Algorithm

order is assumed to be known and a standard expectation-maximization algorithm is used to estimate the parameters (Duda et al., 2000).

Here all density estimates lead to very similar results. Since it allows one to learn the kernel covariances elongated with the data, (ii) gives a sharper KDE estimate as compared to (i). However, since there is no significant difference between the principal curve projections of these two, (ii) might be regarded as somewhat overfitting, since too many parameters (d^2 additional parameters per sample, as the constant kernel bandwidth is replaced by a full data-dependent covariance) are learned, leading to no significant changes. The result shown in (iii) is a good example which shows that good results can be obtained if the parametric family fits the distribution very well. Of course, as you can imagine, the GMM based results might have been much worse for an unsuitable selection of the number of components, or if EM converges to a suboptimal result due to poor initialization, whereas KDE is much more robust in this sense. Also note an analogy to Kegl's approach, using GMM-SCMS leads to a piecewise linear structure if the Gaussian components are sufficiently far (in Mahalanobis distance sence) from each other. In the vicinity of the Gaussian component centers, except when components significantly overlap or get close, the principal curves can be approximated well linearly by piecewise local components.

Note that theoretically the principal curves in (i) and (ii) extend to infinity on both ends; and for the GMM based example in (iii), each component crosses and extends to infinity.

Here -and also for the rest of the paper- we present the data projected onto principal curve only, that depicts the portion of the principal curve in the support of the input data. The nature of the curves outside this region is obvious from the definition and the density estimate plots.

4.1.2 Spiral Data Set

Since many principal curve algorithms are based on the idea of starting with the principal line and adding complexity to the structure (for example adding a vertex to piecewise linear curve) to minimize mean projected error, a data set that folds onto itself may lead to counterintuitive results, and spiral data set is a benchmark data set that has been used in manifold learning and principal curve algorithm literature (Kegl et al., 2000; Vincent and Bengio, 2003) (again, this data set is provided by Kegl).

Similar to the previous example, we start with the results of Hastie-Stuetzle algorithm and Kegl's polygonal line algorithm. Figure 8 shows the data samples and the result of Hastie's algorithm. Figure 9 presents the results of Kegl's polygonal line algorithm for different penalty coefficients. The length penalty coefficient is equal to 0.1, 0.2, 0.4, and 0.5, respectively.

As in the previous example, in SCMS uses the leave-one-out ML kernel bandwidth for this data set. Figure 10 shows the same spiral data set along with the results of KDE-SCMS. Comparing Figure 9 and Figure 10, one can see that both Polygonal Line algorithm -with suitable parameters- and our locally defined principal curve can achieve satisfactory results. Therefore, we create a more challenging scenario, where the spiral this time has some substantial noise around the underlying generating curve and has fewer samples. Figure 11 shows the result of KDE-SCMS, and Figure 12 shows results of Polygonal Line algorithm for different penalty coefficients; 0.05, 0.1, 0.2, and 0.3.

On the noisy spiral data set, we also provide quantitative results for different noise levels and compare the computation times. At each noise level, we find the principal curve using both methods using the same noisy data set, and afterwards we take another 200 samples from the same generating curve and add same amount of radial noise to use as the test set. We present the MSE between the projection of the test samples and their original points on the noiseless generating curve. Results for KDE-SCMS, and Polygonal Line algorithm are presented in Table 2 along with corresponding running times for 50 Monte Carlo runs of this experiment. Since results are presented for the leave-one-out ML kernel bandwidth, the running times for SCMS include this ML training as well. For the Polygonal Line algorithm we performed a manual parameter tuning for each noise level and best results are presented.

Overall, as the noise level increases, the computation time of SCMS increases, presumably due to more iterations being required for convergence; still, the computation time is much less than that of the Polygonal Line algorithm. In terms of MSE between the estimated and the true curve, SCMS provides similar or better performance as compared to the Polygonal Line algorithm. For some noise levels the difference in performance is very small; however, note that the real advantage of SCMS is that it provides the similar/better results nonparametrically -as compared to the best result of several runs of the Polygonal Line algorithm with different parameters.

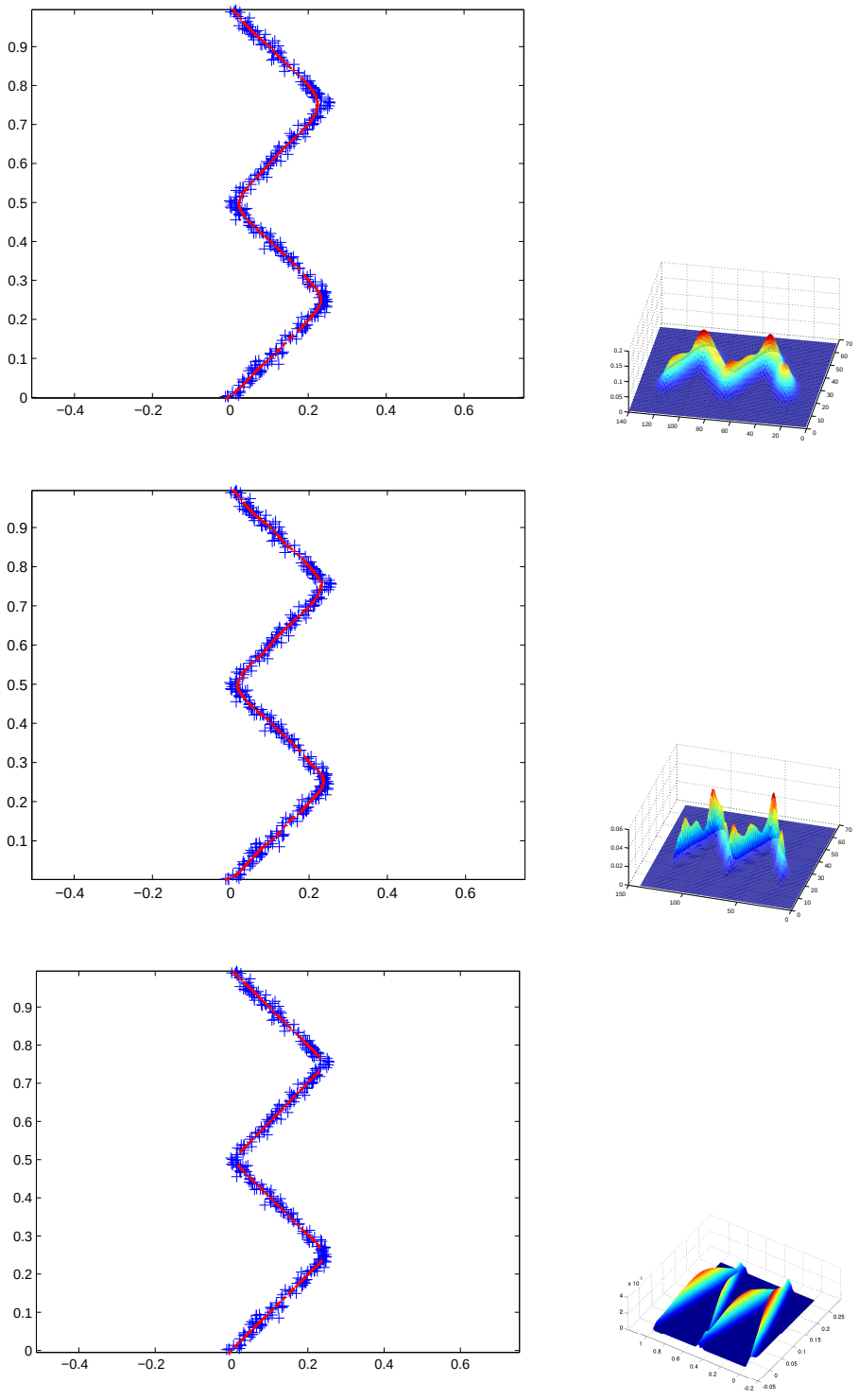


Figure 7: Zig-zag data set, and its principal curve projections obtained for KDE with isotropic constant bandwidth (top), KDE with anisotropic and data-dependent covariance (middle), and Gaussian mixture with 4 components (bottom). The underlying density estimates are shown on the right column.

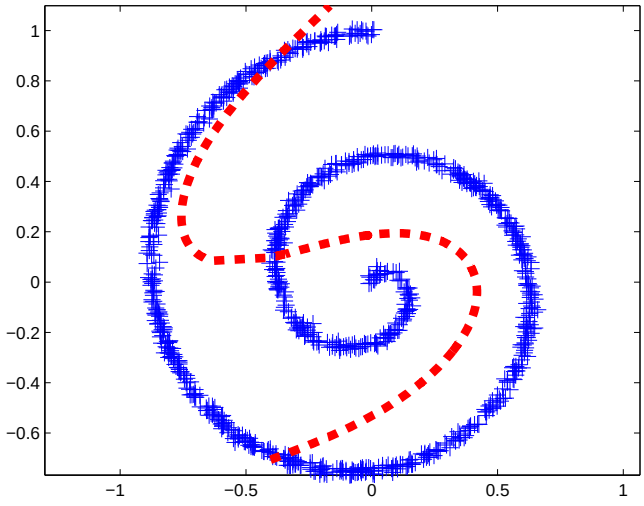


Figure 8: Spiral data set, and Hastie-Stuetzle principal curve

	Computation time	Mean squared projection error	σ_{noise}
SCMS	3.237 sec.	0.003184	0.005
PL	18.422 sec.	0.017677	0.005
SCMS	3.508 sec.	0.011551	0.01
PL	20.547 sec.	0.024497	0.01
SCMS	3.986 sec.	0.062832	0.02
PL	22.671 sec.	0.066665	0.02
SCMS	6.257 sec.	0.194560	0.04
PL	27.672 sec.	0.269184	0.04
SCMS	7.198 sec.	0.433269	0.06
PL	19.093 sec.	0.618819	0.06
SCMS	8.813 sec	0.912748	0.08
PL	19.719 sec	1.883287	0.08

Table 2: Computation Time and MSE Performance Comparisons

4.1.3 Loops, Self-intersections, and Bifurcations

Since they are specifically designed to fit *smooth* curves to the data, traditional principal curve fitting approaches in the literature have difficulties if there are loops, bifurcations and self intersections in the data. Perhaps the most efficient algorithm in this context is Kegl's principal graph algorithm (Kegl and Kryzak, 2002), where Kegl modifies his polygonal line algorithm (Kegl et al., 2000) with a table of predefined rules to handle these irregularities. On the other hand, in the presence of such irregularities, our definition yields a principal graph -a collection of smooth curves. Since the ridges of the pdf can intersect each other, KDE-SCMS can handle such data sets with no additional effort/parameter. Results of

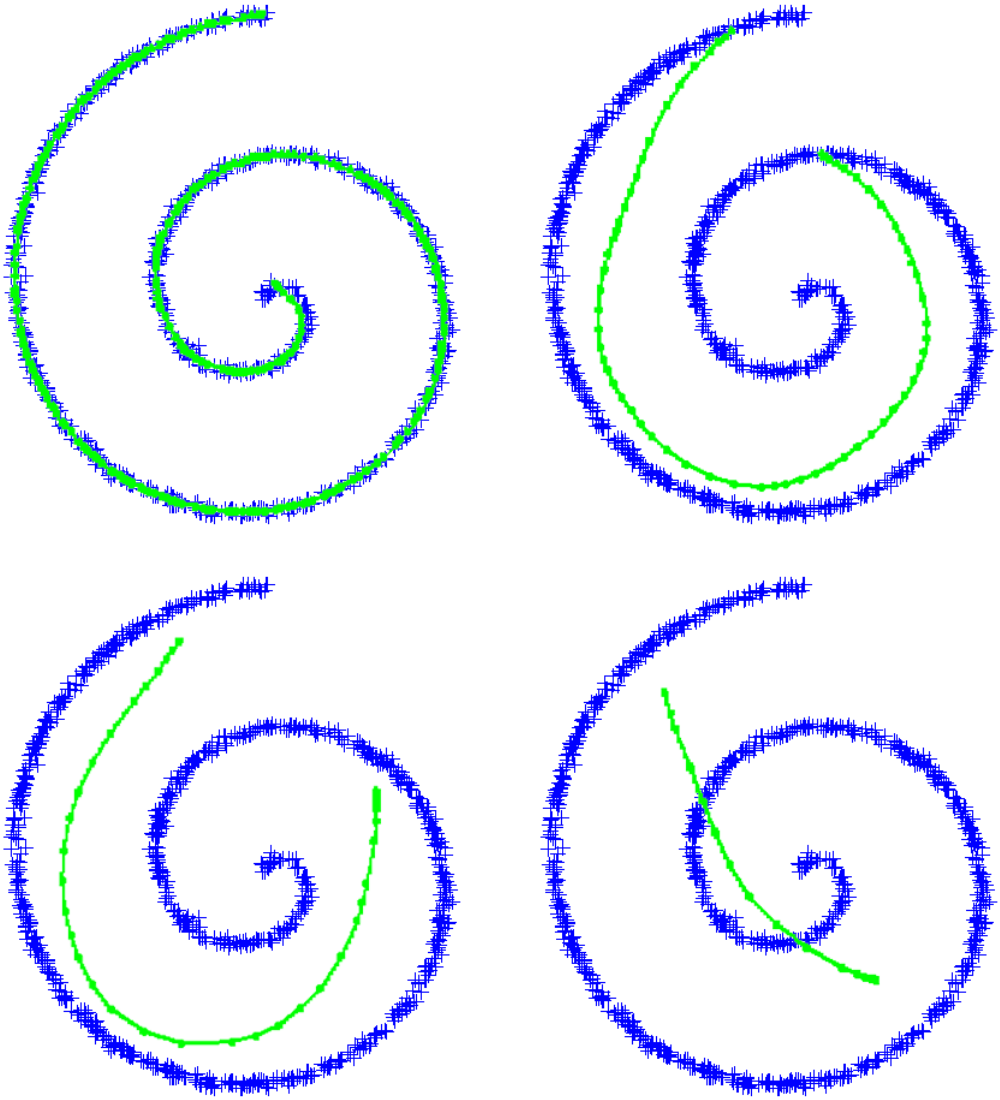


Figure 9: Spiral data set, and result of the Polygonal Line Algorithm

KDE-SCMS on a synthetically-created snow crystal data set that has a number of loops, self intersections, and bifurcation points is presented in Figure 13.

4.1.4 Extending the Definition to Higher Dimensional Manifolds

The generalization of principal curves to principal surfaces and higher order manifolds is naturally achieved with our definition. Here we present the results of KDE-SCMS for $d = 1$ and $d = 2$ for a three-dimensional helix data set in Figure 14. (For $d = 2$, we present the surface built by the Delaunay triangulations Delaunay, 1934 of the principal surface projections for better visualization.) Here, note that the covariance of the helix data around the principal curve is not symmetric, and the horizontal dimension has a higher variance

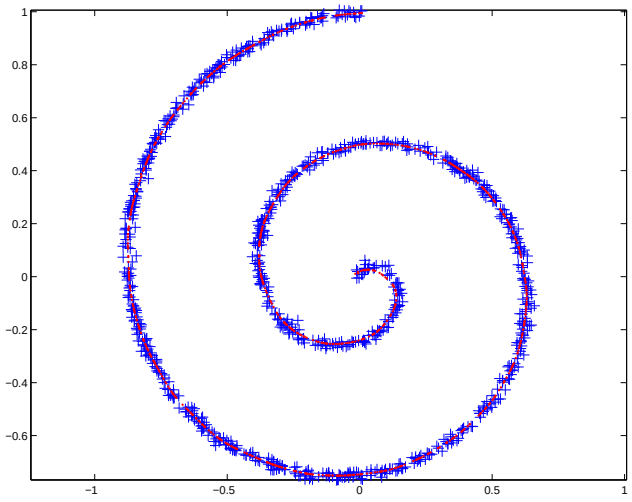


Figure 10: Spiral data set, and KDE-SCMS principal curve

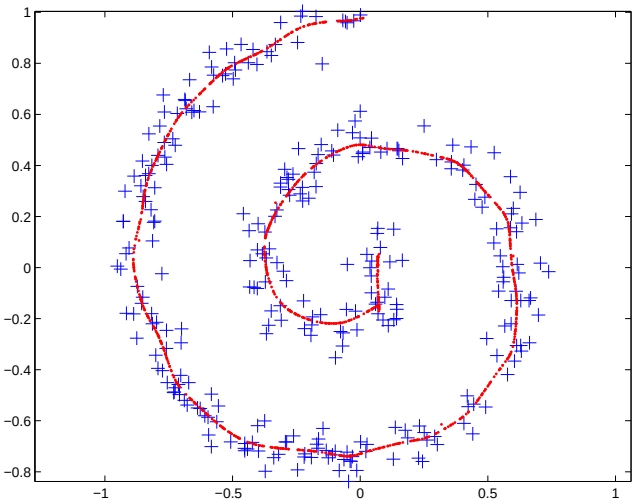


Figure 11: Noisy spiral data set, and KDE-SCMS principal curve

(and this is why the principal surface is spanned along this dimension). If the helix had been symmetric around the principal curve, the principal surface would have been ill-defined.

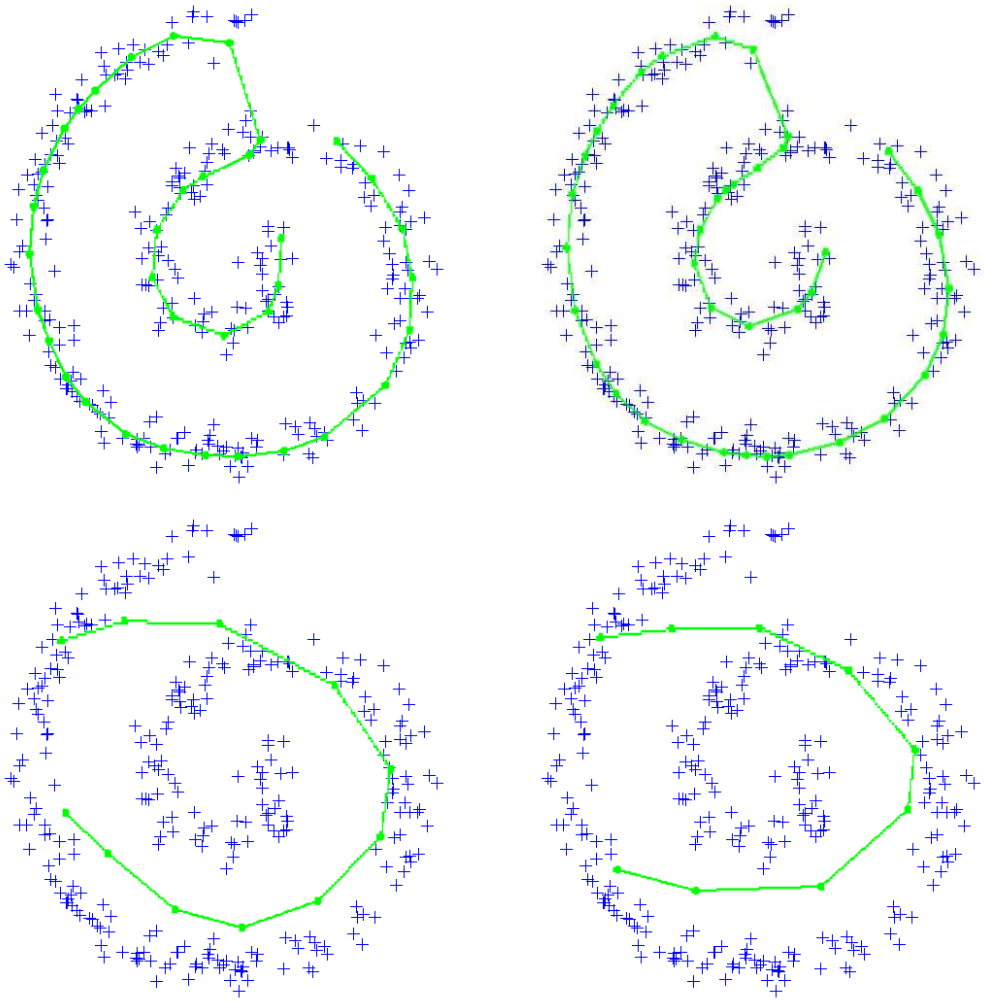


Figure 12: Noisy spiral data set, and result of the Polygonal Line Algorithm

4.2 Applications of Principal Curves

In the following, we will present a number of applications of our approach; on time series denoising, independent components analysis, time-frequency reassignment, channel equalization, and optical character skeletonization.

4.2.1 Time Series Signal Denoising

KDE-SCMS finds use in many applications of time series denoising. In general, the feature space for such problems can be constructed using the time index as one of the features, yielding an embedded structure of the -possibly multidimensional- time signal. In such spaces, we show that KDE-SCMS can successfully be used for denoising (Ozertem and Erdogmus, 2009; Ozertem et al., 2008). In the following, first we will briefly mention our previous work on applying KDE-SCMS to signal denoising applications, and proceed with preliminary results in two other application domains.

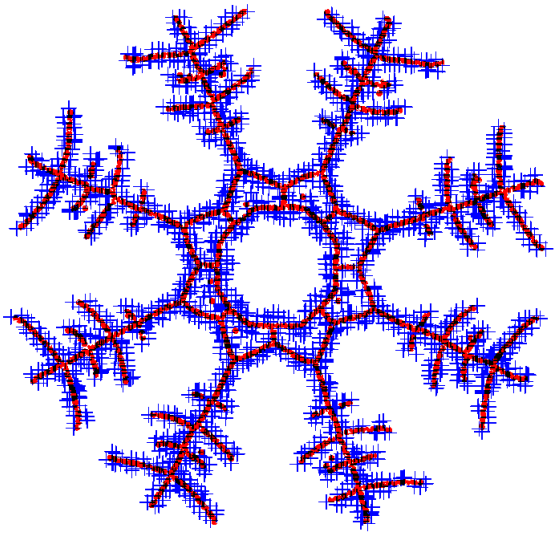


Figure 13: Snow crystal data set, and KDE-based SCMS result

We proposed to use principal curve projections as a nonparametric denoising filter at the preprocessing stage of time warping algorithms, which in general are prone to noise (Ozertem and Erdogmus, 2009). In this setting, time embedding is used in the scatter plot of the pair of signals that we want to find the time warping function in between. We use a slightly different variant of KDE-based SCMS for this purpose that exploits the application specific case that the time embedding dimension is not a random variable, and shown improvement in time series classification and clustering.

A common problem in signal denoising is that if the signal has a blocky, in other words, a piecewise-smooth structure, traditional frequency domain filtering techniques may lead to oversmoothings in discontinuities. One idea to overcome this is to take discrete wavelet transform (DWT), do the filtering (or thresholding) in this domain and recover the smoothed signal by taking inverse DWT. The shortcoming of this is high frequency artifacts (similar to Gibbs-effect) at both ends of the discontinuities. We show that KDE-SCMS can be used for this purpose (Ozertem et al., 2008). Since at the discontinuities, KDE will not be much affected by the signal samples of the other end of the discontinuity, the algorithms leads to a piecewise-smooth denoising result without introducing oversmoothings or any artifacts at the discontinuities.

4.2.2 Nonlinear Independent Component Analysis

The proposed principal surface definition can be viewed in a differential geometric framework as follows: at each point $\mathbf{x}$, the solutions to the differential equations that characterize curve

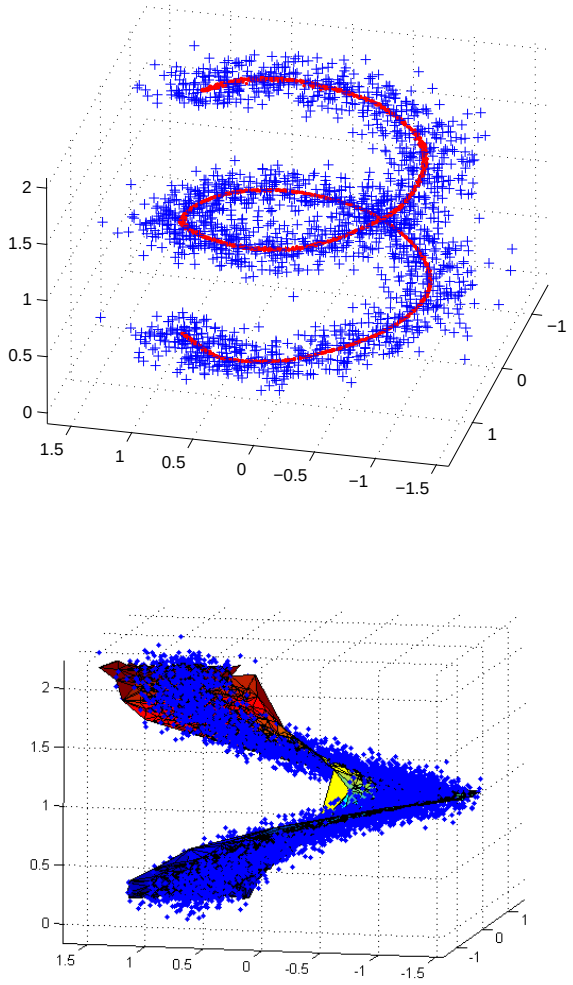
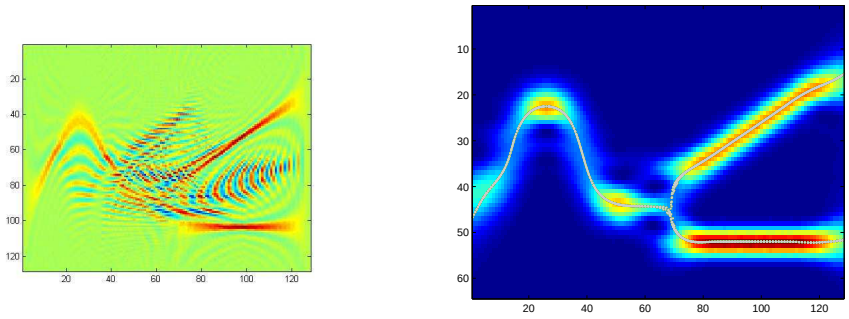


Figure 14: Helix data set, and KDE-based SCMS result for $d = 1$ (top) and $d = 2$ (bottom)

whose tangents are the eigenvectors of the local covariance of the pdf form a local curvilinear coordinate system that is isomorphic to an Euclidean space in some open ball around $\mathbf{x}$. The trajectories that take a point $\mathbf{x}$ to its projection on the d-dimensional principal surface can be used to obtain these curvilinear coordinates that specify the point with respect to some reference critical point that can be assumed to be the origin. Consequently, for instance, for $\mathbf{x}$, the lengths of curves during its projection from n-dimensional space to the $(n-1)$-dimensional principal surface, and then subsequently to $(n-2), \ldots, 1$, and eventually to a local maximum (the one that has been recognized as the origin) could, in some cases when a global manifold unfolding is possible, lead to a nonlinear coordinate vector. This manifold unfolding strategy can be used in many applications including visualization and



(a) The Wigner-Ville distribution

(b) Smoothed Wigner-Ville distribution and its principal curve

Figure 15: Wigner-Ville distribution in time-frequency domain, its smoothed version and principal curve of the smoothed distribution

nonlinear blind source separation. As we do not aim to focus on the manifold unwrapping aspects of the proposed framework in this paper (because that, in principle, requires solving differential equations accurately and the proposed algorithm is not at a desired level of accuracy for that purpose), we simply point out that the definition presented here allows for a principled coordinate unfolding strategy as demonstrated in nonlinear blind source separation (Erdogmus and Ozertem, 2007). Developing fast feedforward approximations (via parametric or nonparametric mappings) to this manifold unwrapping strategy remains as a critical future work.

4.2.3 Time-Frequency Distribution Reassignment

Time-frequency reassignment is a known problem in signal processing literature and yields another example, where KDE-SCMS can be applied directly. As any other bilinear energy distribution, the spectrogram is faced with an unavoidable trade-off between the reduction of misleading interference terms and a sharp localization of the signal components. To reduce the smoothing effects introduced by the window function in short-term Fourier transform, reassignment methods are used to sharpen the time-frequency representation by using the rate of change of phase of the signal, which finds numerous applications in speech signal processing and signal roughness analysis (Fulop and Fitz, 2007; K. Fitz and L. Haken and P. Christensen, 2000). Parameter envelopes of spectral components are obtained by *following ridges on the smooth time-frequency surface*, using the reassignment method (Auger and Flandrin, May 1995) to improve the time and frequency estimates for the envelope breakpoints. Figure 15 shows our preliminary results for a synthetic time frequency surface with multiple components in some time intervals that yield cross interference terms. Wigner-Ville distribution of the signal, and the smoothed Wigner-Ville distribution, where the cross-terms in the original spectogram are eliminated are shown in Figure 15(a). Figure 15(b) shows the principal curve of this time-frequency surface obtained by KDE-SCMS.

Furthermore, in the presence of the *auto-cross terms*, a much more challenging scenario appears (Ozdemir and Arikan, 2000; Ozdemir et al., 2001). In these cases a rotation invariant reassignment method is required and traditional methods that are based on the rate of change of the phase cannot answer this need. KDE-SCMS, on the other hand, is still directly applicable to this problem because it is invariant to rotations in the input data.

4.2.4 Time-varying MIMO Channel Equalization

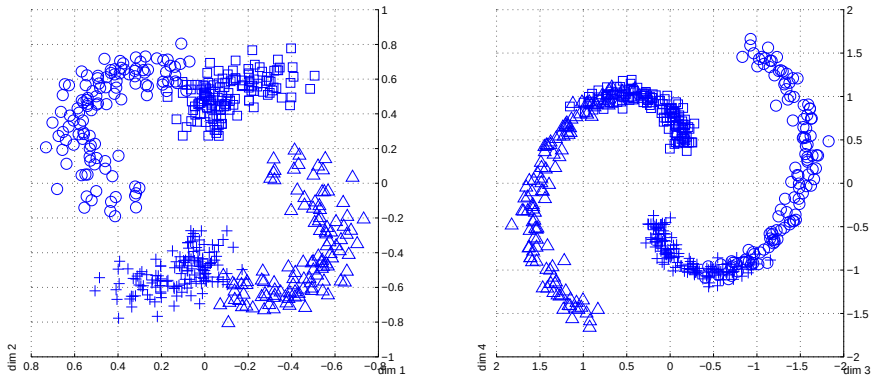
Recently, multiple-input multiple-output wireless communication systems have drawn considerable attention, and there are reliable and computationally inexpensive symbol detection algorithms in the literature (Foschini et al., Nov 1999). On the other hand, applications in time-varying environments pose a harder problem to the changing channel state, and some supervised algorithms have been proposed to tackle this issue, where an initialization phase is used in the beginning for training purpose (Rontogiannis et al., May 2006; Karami and Shiva, 2006; Choi et al., Nov. 2005).

Blind channel equalization approaches in the literature are based on clustering (Chen et al., Jul 1993). However, these approaches mostly focus on time-invariant single-input single-output channels. Recently, a spectral clustering technique is proposed that extends the applications into time-varying multiple-input multiple-output channels as well (Van Vaerenbergh et al., 2007; Vaerenbergh and Santamaria, 2008). Van Vaerenbergh and Santamaria introduce the time embedding into the feature space before employing the clustering algorithm to *untangle* the clusters. The same idea proves to be effective in Post-Nonlinear Blind Source Separation as well (Vaerenbergh and Santamaría, 2006).

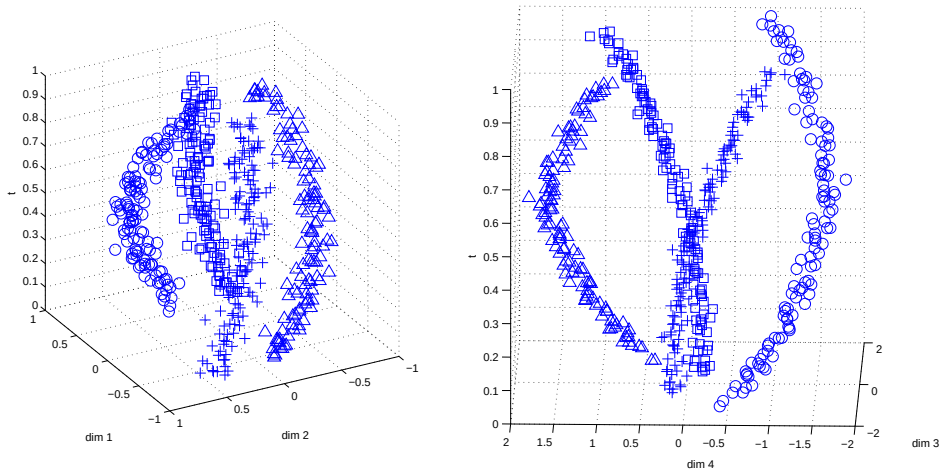
The original clustering problem in four dimensions presented in Figure 16(a). The fast time-varying nature of the channel poses a very difficult clustering problem with overlapping clusters. With the time embedding, the overlapping clusters become intertwined threads as shown in Figure 16(b) with two three-dimensional subspace projections of the data. Van Vaerenbergh and Santamaria employ a spectral clustering algorithm to solve the channel equalization problem with no supervision. At this point, one can improve noise robustness of the clustering by using the fact that the clusters are curves in the feature space by using the spectral clustering of the principal curve projections instead of the data samples. Figure 17 shows a result of KDE-SCMS for the same data set at signal to noise ratio of 5dB, along with the average normalized MSE (and $\pm$ one standard deviation) between the actual noisefree signal and the principal curve projection over 20 Monte Carlo runs. The principal curve projection result can give a good estimate of the noisefree signal even in signal to noise ratio levels even lower than 0dB -where the noise power is greater than the signal power itself.

4.2.5 Skeletonization of Optical Characters

Optical character skeletonization can be used for two purposes: feature extraction for optical character recognition and compression. Principal curves have been used for this application (Kegl and Kryzak, 2002). One significant problem with applying principal curve algorithms to skeletonization of optical characters is that, by definition, algorithms are seeking for a *smooth curve*. In general, data may have loops, self intersections, and bifurcation points, which is obviously the case for optical characters. Kegl's principal graph algorithm is



(a) Four dimensional data, coming from four symbols



(b) Four dimensional data with time embedding

Figure 16: Symbol clustering problem for a MIMO channel

perhaps the only method in the literature that can successfully handle such irregularities (Kegl and Kryzak, 2002). In this approach, Kegl reshapes his polygonal line algorithm (Kegl et al., 2000) to handle loops, and self intersections by modifying it with a table of rules and adding preprocessing and postprocessing steps. Using the handwritten digits data set provided by Kegl, we show the results of KDE-SCMS. Figure 18 shows the binary

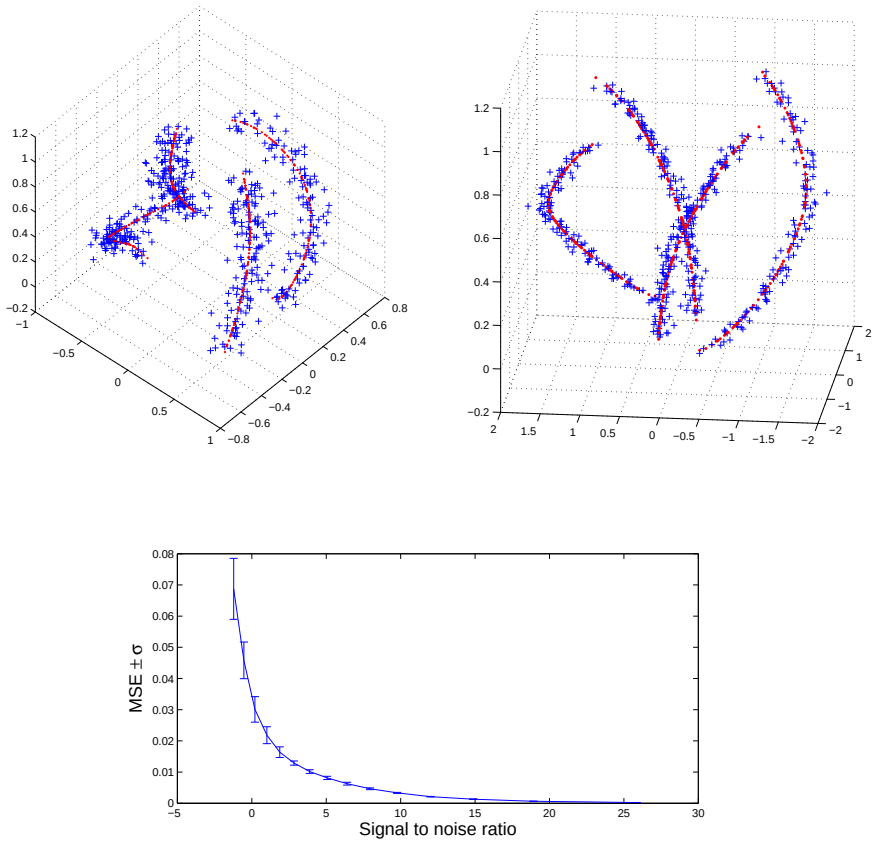


Figure 17: Signal samples and their principal curve projections, normalized MSE vs signal to noise ratio (in dB).

images along with the principal curve projection of the pixels. SCMS gives satisfactory results without any rule or model based special treatment for the self intersections.

4.3 Limitations, Finite Sample Effects, and the Curse of Dimensionality

Since our principal curve definition assumes the pdf to be given, it depends on the reliability of the preceding density estimation step, which in general may not be an easy task. Stated by Bellman as *the curse of dimensionality* (Bellman, 1961), it is a very well-known fact that density estimation becomes a much harder problem as the dimensionality of the data increases. Therefore, before we move on to applications on real data, in this section we will present the performance of our principal curve fitting results for various density estimates with different number of samples and dimensions.

The first comparison is with principal line estimation based on eigendecomposition of the data covariance, where the true underlying probability distribution is Gaussian. The second comparison examines the model order estimation using a Gaussian mixture model,

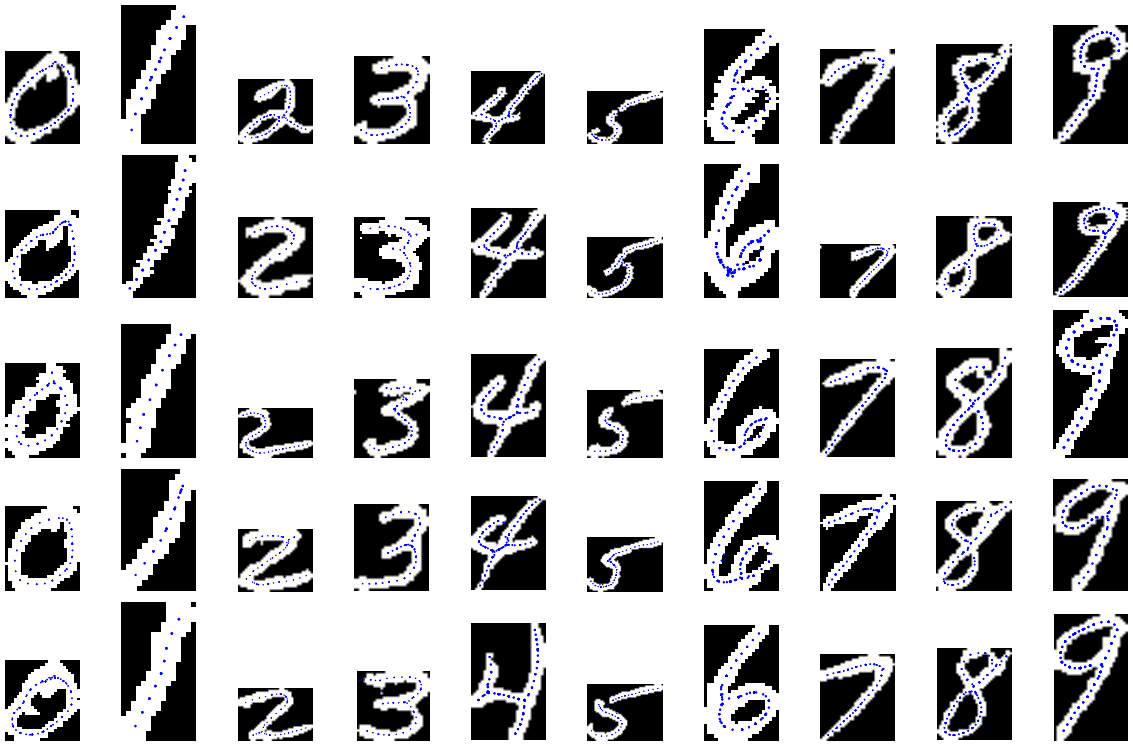


Figure 18: SCMS results in optical characters

which in the limiting case, where the number of Gaussian mixtures is equal to the number of samples, converges to KDE. In all comparisons presented below principal curve projections are obtained by the KDE-SCMS algorithm using the leave-one-out ML kernel bandwidth.

4.4 Comparison with Eigenvector Estimation

As mentioned before, the reason why we prefer to use KDE is its ability to adapt to different complex shapes that data may take. Indeed, results previously presented in this section show that KDE based principal curve estimation proves to be efficient in adapting to many real-life data distributions of a diverse set of applications. However, one well-known disadvantage of KDE is the required number of samples as the dimensionality of the data increases. Here we discuss the case where the true underlying probability density is Gaussian; hence, the claim of *the requirement to adapt to complex shapes in the data* is an obvious overstatement. In this scenario, we will compare the principal line estimator based on PCA to the principal curve based on KDE, for different number of dimensions.

Consider the data set $\{\mathbf{x}_{i=1}^N\}$ Gaussian distributed in d-dimensional space, where $\mathbf{v}$ denotes the true principal line of this distribution, and $\mathbf{v}_*$ denotes the principal line obtained by sample PCA. What we are going to compare here is the following:

1. mean squared distance between the projection of the data samples onto the true first eigenvector and the estimated first principal component, $E\{\|\mathbf{v}^T\mathbf{x} - \mathbf{v}_*^T\mathbf{x}\|\}$

2. mean squared distance between the projection of the data samples onto the true eigenvector and the principal curve projection $\widetilde{\mathbf{x}}$, $E\{\|\mathbf{v}^T\mathbf{x} - \widetilde{\mathbf{x}}\|\}$.

Figure 19 presents the MSE of the principal line (dashed curve) and principal curve (solid curve) projections for 2, 3, 4, 5, 10, 20, 30,and 40 dimensions, and average log MSE for 100 Monte Carlo simulations is shown. For all cases the MSE decreases for both methods as the number of samples increase. Principal line projection always results in better accuracy and the performance of principal curve projections drop exponentially for increasing dimensions.

4.5 Effects of the Model Order Estimation

An important problem in parametric density estimation is model order selection. In the real applications presented above, we work with KDE-SCMS to provide a general purpose nonparametric algorithm, and to avoid model order selection problems. However, using a parametric model has two main advantages:

1. As opposed to $O(N^2)$ complexity of the KDE-SCMS, the computational complexity of GMM-SCMS is $O(MN)$, where M is the number of mixtures in the Gaussian mixture and N is the number of samples, since typically $M \ll N$.

2. As also implied in the previous section, with the comparison against PCA on a Gaussian data set, a parametric approach with a *suitable* model order, the algorithm would need less samples to achieve good principal curve estimates.

Here we will evaluate the stability of principal curve estimation with GMM-SCMS for improper model order selections in the GMM density estimation step, and compare the principal curve projection results for a Gaussian mixture with 3 components. Since the true underlying density is known to have 3 components, we measure the performance as of principal curve projection results for different number of components in the density estimate as the distance to the principal curve projections obtained with three components

$$\begin{aligned} J_d &= E\{(\widetilde{\mathbf{x}}_3(\mathbf{x}) - \widetilde{\mathbf{x}}_d(\mathbf{x}))^2\} \ , \\ where \ \ d &= 1, 2, 3, 4, 5, 6, 10, 15, 25, 50, 100, 200, 400 \ . \end{aligned}$$

The data set $\mathbf{x}$ has 400 samples in 2-dimensional space. Figure 20 shows a realization of the Gaussian mixture, and Figure 21 presents the performance of the principal curve projections for different number of components in the Gaussian mixture estimation, and results of 50 Monte Carlo simulations is shown. Note that for increasing model orders, if the GMM has more number of components than the true underlying distribution, the generalization performance of the principal curve does not change significantly.

5. Discussions

We proposed a novel definition that characterizes the principal curves and surfaces in terms of the gradient and the Hessian of the density estimate. Unlike traditional machine learning papers on manifold learning, which tend to focus on criteria such as reconstruction error of available samples, we focus on the definition of the underlying manifold from a more (differential—though not emphasized here) geometric point of view. There are strong

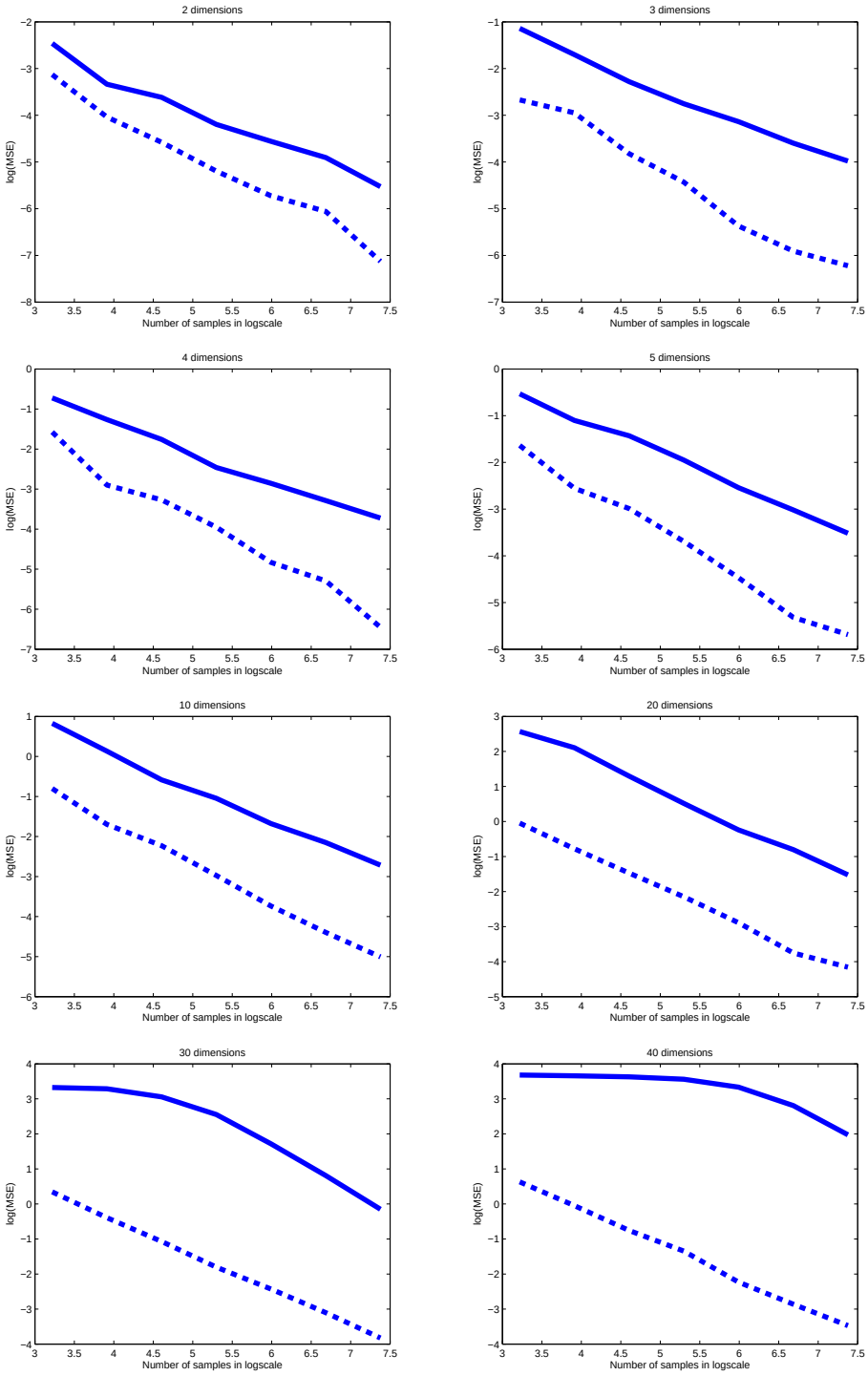


Figure 19: Mean projection error in log_e scale for principal line (dashed) and principal curve (solid). Average of 100 Monte Carlo simulations is shown.

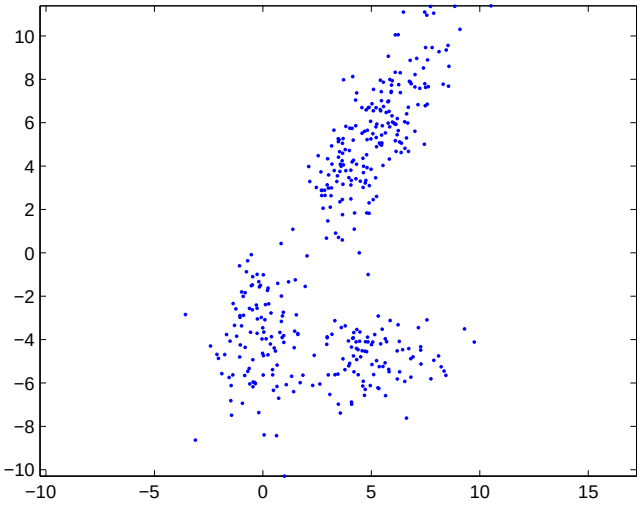


Figure 20: One realization of the 3-component Gaussian mixture data used in performance evaluations

connections between our definition and the literature. If the ridge cross-section is a unimodal and symmetric density, our definition coincides with the original Hastie & Stuetzle definition. There is a strong connection to Kegl's piecewise linear curve proposition, when the underlying density is selected to be a Gaussian mixture. However, the connections are less obvious when considering a principal curve or manifold definition that is not explicit (e.g., the principal curve is the solution to some optimization problem without an analytical expression or property).

Providing the definition in terms of the probability density estimate of the data allows us to exclude any smoothness or regularization constraints from the definition, and adopt them from the density estimation literature directly. Although this cannot ultimately answer the question of the trade-off between generalization and overfitting, using the connection to density estimation yields data-driven nonparametric solutions for handling regularization and outlier robustness. In the definition, we also do not assume any parametric model and since the ridges of the pdf can intersect each other, handling self-intersecting data structures requires no additional effort.

An important property of the definition is that it yields a unified framework for clustering, principal curve fitting and manifold learning. Similar to PCA, for an n-dimensional data set, our definition contains all the d-dimensional principal manifolds, where $d < n$. Theoretically, the principal set of $d = 0$ yields the modes of the probability density, which coincides with a widely accepted clustering solution. We accompany this with an algorithmic connection by showing that principal curves can be achieved using the SCMS idea, very similar to the well-known mean-shift clustering algorithm. KDE-based SCMS implementation is significantly faster than the most commonly used method in principal curves literature.

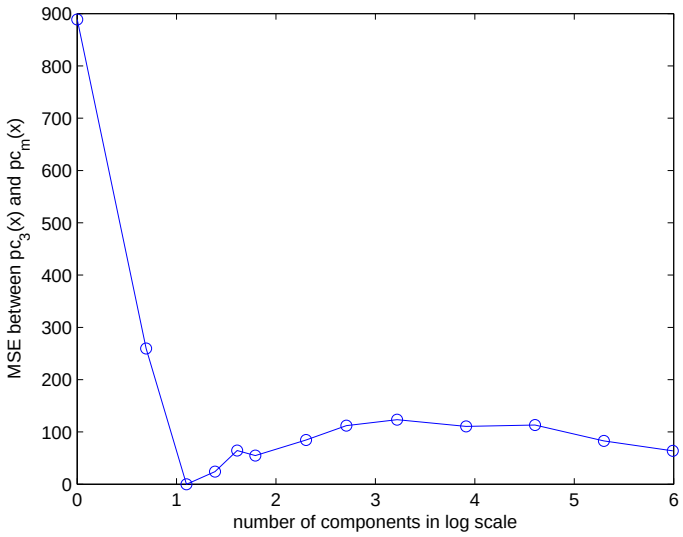


Figure 21: Principal curve projections for different number of components in the density estimate in log_e scale. Specifically, $d = 1, 2, 3, 4, 5, 6, 10, 15, 25, 50, 100, 200, 400$.

Besides, it does not require significantly more time or memory storage as compared to mean shift, which already has been used in many practical application domains.

In high dimensional spaces, density estimation becomes impractical due to the curse of dimensionality. Therefore, similar to existing methods in principal curves literature, the proposed method is not an alternative for proximity graph based manifold learning methods like Isomap, Laplacian eigenmaps etc. Still, we show that there are many real applications in lower dimensional spaces suitable for KDE-based SCMS. We show results on a family of applications in time series signal processing, as well as an earlier proposed application of principal curves (OCR).

Acknowledgments

The authors would like to thank Jose C. Principe, Miguel A. Carreira-Perpinan, Sudhir Rao, Engin Tola, and M. Emre Sargin for valuable discussions, and B. Kegl for providing some of the data sets and implementations of the Polygonal Line algorithm used for the comparisons in the experiments. This work is partially supported by NSF grants ECS-0524835, and ECS-0622239.

References

F. Auger and P. Flandrin. Improving the readability of time-frequency and time-scale representations by the reassignment method. *IEEE Transactions on Signal Processing*, 43(5):1068–1089, May 1995.

J. D. Banfield and A. E. Raftery. Ice floe identification in satellite images using mathematical morphology and clustering about principal curves. *Journal of the American Statistical Association*, 87(417):7–16, 1992.

G. Baudat and F. Anouar. Generalized discriminant analysis using a kernel approach. *Neural Computation*, 12(10):2385–2404, 2000.

M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6):1373–1396, June 2003.

R. E. Bellman. *Adaptive Control Processes.* Princeton University Press, Princeton, NJ, 1961.

Y. Bengio, H. Larochelle, and P. Vincent. Non-local manifold parzen windows. In *Advances in Neural Information Processing Systems 18*, pages 115–122. MIT Press, 2006.

C. M. Bishop. *Neural Networks for Pattern Recognition, 1st Ed.* Clarendon Press, Oxford, 1997.

M. A. Carreira-Perpinan. Fast nonparametric clustering with gaussian blurring mean-shift. In *ICML '06: Proceedings of the 23rd international conference on Machine learning*, pages 153–160, New York, NY, USA, 2006. ACM. ISBN 1595933832.

J. E. Chacon, T. Duong, and M. P. Wand. Asymptotics for general multivariate kernel density derivative estimators. *Statistica Sinica, in press*, 2011.

K. Chang and J. Grosh. A unified model for probabilistic principal surfaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(1):59–74, 2002.

O. Chapelle and A. Zien. Semi-supervised classification by low density separation. In R. G. Cowell and Z. Ghahramani, editors, *Proc. of the Tenth Int. Workshop on Artificial Intelligence and Statistics (AISTATS 2005)*, pages 57–64, Barbados, January 6–8 2005.

O. Chapelle, B. Schölkopf, and A. Zien, editors. *Semi-Supervised Learning.* MIT Press, Cambridge, MA, 2006. URL `http://www.kyb.tuebingen.mpg.de/ssl-book`.

S. Chen, B. Mulgrew, and P. M. Grant. A clustering technique for digital communications channel equalization using radial basis function networks. *IEEE Transactions on Neural Networks*, 4(4):570–590, Jul 1993.

Y. Cheng. Mean shift, mode seeking, and clustering. *IEEE Trans. Pattern Analysis and Machine Intelligence,*, 17(8):790–799, 1995.

J. Choi, H. Yu, and Y. H. Lee. Adaptive mimo decision feedback equalization for receivers with time-varying channels. *Signal Processing, IEEE Transactions on [see also Acoustics, Speech, and Signal Processing, IEEE Transactions on]*, 53(11):4295–4303, Nov. 2005.

D. Comaniciu. An algorithm for data-driven bandwidth selection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(2):281–288, 2003.

D. Comaniciu and P. Meer. Mean shift: a robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(5):603–619, 2002.

B. N. Delaunay. Sur la sphère vide. *Bulletin of Academy of Sciences of the USSR*, (6): 793–800, 1934.

P. Delicado. *Principal curves and principal oriented points*. 1998. URL `http://www.econ.upf.es/deehome/what/wpapers/postscripts/309.pdf`.

T. Duchamp and W. Stuetzle. Geometric properties of principal curves in the plane. *Robust Statistics, Data Analysis, and Computer Intensive Methods: In Honor of Peter Huber's 60th Birthday*, 109:135–152, 1996a.

T. Duchamp and W. Stuetzle. Extremal properties of principal curves in the plane. *The Annals of Statistics*, 24(4):1511–1520, 1996b.

R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. Wiley-Interscience Publication, 2000.

D. Erdogmus and U. Ozertem. Nonlinear coordinate unfolding via principal curve projections with application to bss. In *14th International Conference on Neural Information Processing*, 2007.

L. Fahrmeir and G. Tutz. *Multivariate Statistical Modelling Based on Generalized Linear Models*. Springer-Verlag, New York, 1994.

T. S. Ferguson. A bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1(2):209–230, 1973.

G. J. Foschini, G. D. Golden, R. A. Valenzuela, and P. W. Wolniansky. Simplified processing for high spectral efficiency wireless communication employing multi-element arrays. *IEEE Journal on Selected Areas in Communications*, 17(11):1841–1852, Nov 1999.

K. Fukunaga and D.R. Olsen. An algorithm for finding intrinsic dimensionality of data. *IEEE Transactions on Computers*, 20(2):176–183, 1971.

S. A. Fulop and K. Fitz. Separation of components from impulses in reassigned spectrograms. *Acoustical Society of America Journal*, 121:1510–1517, 2007.

J. Ham, D. Lee, S. Mika, and B. Scholkopf. A kernel view of the dimensionality reduction of manifolds. *In Proceedings of the Twenty First International Conference on Machine Learning (ICML-04)*, pages 369–376, 2004.

B. E. Hansen. Uniform convergence rates for kernel estimation with dependent data. *Econometric Theory*, 24(03):726–748, June 2008. URL `http://ideas.repec.org/a/cup/etheor/v24y2008i03p726-748_08.html`.

T. Hastie. *Principal Curves and Surfaces*. PhD thesis, Stanford University, 1984.

T. Hastie and W. Stuetzle. Principal curves. *Journal of American Statistical Association*, 84:502–516, 1989.

A. Hegde, J. C. Principe, D. Erdogmus, U. Ozertem, Y. N. Rao, and H. Peddaneni. Perturbation-based eigenvector updates for on-line principal components analysis and canonical correlation analysis. *Journal of VLSI Signal Processing Systems*, 45(1-2):85–95, 2006.

H. Hotelling. Analysis of a complex of statistical variables into principal components. *J. Educ. Psychol.*, 24, 1933.

J. E. Jackson. *A User's Guide to Principal Components*. John Wiley and Sons, New York, 1991.

I. T. Jolliffe. *Principal Components Analysis*. Springer-Verlag, Berlin, 1986.

M. C. Jones, J. S. Marron, and S. J. Sheather. A brief survey of bandwidth selection for density estimation. *Journal of the American Statistical Association*, 91(433):401–407, 1996.

K. Fitz and L. Haken and P. Christensen. Transient preservation under transformation in an additive sound model. *Proceedings of International Computer Music Conference*, pages 392–395, 2000.

N. Kambhatla and T. K. Leen. Fast non-linear dimension reduction. In *Neural Information Processing Systems*, pages 152–159, 1994.

N. Kambhatla and T. K. Leen. Dimension reduction by local principal component analysis. *Neural Computation*, 9(7):1493–1516, 1997.

E. Karami and M. Shiva. Decision-directed recursive least squares mimo channels tracking. *EURASIP Journal of Wireless Communication Networks*, 2006(2):7–7, 2006.

B. Kegl. *Principal Curves: Learning, Design, And Applications*. PhD thesis, Concordia University, Montreal, Canada, 1999.

B. Kegl and A. Kryzak. Piecewise linear skeletonization using principal curves. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(1):59–74, 2002.

B. Kegl, A. Kryzak, T. Linder, and K. Zeger. Learning and design of principal curves. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(3):281–297, 2000.

S. Y. Kung, K. I. Diamantaras, and J. S. Taur. Adaptive principal component extraction (apex) and applications. *IEEE Transactions on Signal Processing*, 42(5):1202–1217, May 1994.

P. McCullagh and J. A. Nelder. *Generalized Linear Models*. Chapman & Hall/CRC, London, 1989.

P. Meinicke and H. Ritter. Local pca learning with resolution-dependent mixtures of gaussians. *Proceedings of 9th International Conference on Artificial Neural Networks*, pages 497–502, 1999.

A. K. Ozdemir and O. Arikan. A high resolution time frequency representation with significantly reduced cross-terms. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2:693–696, 2000.

A. K. Ozdemir, L. Durak, and O. Arikan. High resolution time-frequency analysis by fractional domain warping. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 6:3553–3556, 2001.

U. Ozertem and D. Erdogmus. Principal curve time warping. *IEEE Transactions on Signal Processing*, 57(6):2041–2049, 2009.

U. Ozertem, D. Erdogmus, and O. Arikan. Piecewise smooth signal denoising via principal curve projections. In *IEEE Int. Conf. on Machine Learning for Signal Processing*, pages 426 – 431, 2008.

E. Parzen. On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33(3):1065–1076, 1962.

V. C. Raykar and R. Duraiswami. Fast optimal bandwidth selection for kernel density estimation. In J. Ghosh, D. Lambert, D. Skillicorn, and J. Srivastava, editors, *Proceedings of the sixth SIAM International Conference on Data Mining*, pages 524–528, 2006.

A.A. Rontogiannis, V. Kekatos, and K. Berberidis. A square-root adaptive v-blast algorithm for fast time-varying mimo channels. *IEEE Signal Processing Letters*, 13(5):265–268, May 2006.

S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, December 2000.

S. Sandilya and S. R. Kulkarni. Principal curves with bounded turn. *IEEE Transactions on Information Theory*, 48(10):2789–2793, 2002.

B. Schölkopf, A. Smola, and K. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10(5):1299–1319, 1998.

John Shawe-Taylor and Yoram Singer, editors. *Regularization and Semi-supervised Learning on Large Graphs*, volume 3120 of *Lecture Notes in Computer Science*, 2004. Springer.

S. J. Sheather and M. C. Jones. A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society. Series B (Methodological)*, 53(3):683–690, 1991.

B. W. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman & Hall/CRC, April 1986. ISBN 0412246201.

D. C. Stanford and A. E. Raftery. Finding curvilinear features in spatial point patterns: Principal curve clustering with noise. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(6):601–609, 2000.

J. B. Tenenbaum, V. de Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, December 2000.

R. Tibshirani. Principal curves revisited. *Statistics and Computation*, 2:183–190, 1992.

S. Van Vaerenbergh and I. Santamaría. A spectral clustering approach to underdetermined post-nonlinear blind source separation of sparse sources. *IEEE Transactions on Neural Networks*, 17(3):811–814, May 2006.

S. Van Vaerenbergh and I. Santamaria. A spectral clustering approach for blind decoding of mimo transmissions over time-correlated fading channels. In E. Hines and M. Martinez, editors, *Intelligent Systems: Techniques and Applications*. Shaker Publishing, 2008.

S. Van Vaerenbergh, E. Estébanez, and I. Santamaría. A spectral clustering algorithm for decoding fast time-varying BPSK MIMO channels. In *15th European Signal Processing Conference*, Poznan, Poland, September 2007.

J. J. Verbeek, N. A. Vlassis, and B. Kröse. A soft k-segments algorithm for principal curves. In *ICANN '01: Proceedings of the International Conference on Artificial Neural Networks*, pages 450–456, London, UK, 2001. Springer-Verlag. ISBN 3-540-42486-5.

J. J. Verbeek, N. Vlassis, and B. Kröse. A k-segments algorithm for finding principal curves. *Pattern Recognition Letters*, 23(8):1009–1017, 2002.

P. Vincent and Y. Bengio. Manifold parzen windows. In *Advances in Neural Information Processing Systems 15*, pages 825–832, 2003.

K. Q. Weinberger and L. K. Saul. Unsupervised learning of image manifolds by semidefinite programming. *International Journal of Computer Vision*, 70(1):77–90, 2006.

A. S. Wong, K. Wong, and C. Wong. A practical sequential method for principal component analysis. *Neural Processing Letters*, 11(2):107–112, 2000.

C. Yang, R. Duraiswami, D. Dementhon, and L. Davis. Mean-shift analysis using quasi-newton methods. In *Proceedings of the International Conference on Image Processing*, pages 447–450, 2003a.

C. Yang, R. Duraiswami, N. A. Gumerov, and L. Davis. Improved fast gauss transform and efficient kernel density estimation. In *Ninth IEEE International Conference on Computer Vision*, pages 664–671 vol.1, 2003b.