On Some Convergence Properties of the Subspace Constrained Mean Shift

Y. Aliyari Ghassabeh*, T. Linder, G. Takahara

Department of Mathematics and Statistics, Queen's University, Kingston, ON, K7L 3N6

Abstract

Subspace constrained mean shift (SCMS) is a non-parametric, iterative algorithm that has recently been proposed to find principal curves and surfaces based on a new definition involving the gradient and Hessian of a kernel probability density estimate. Although simulation results using synthetic and real data have demonstrated the usefulness of the SCMS algorithm, a rigorous study of its convergence is still missing. This paper aims to take initial steps in this direction by showing that the SCMS algorithm inherits some important convergence properties of the mean shift (MS) algorithm. In particular, the monotonicity and convergence of the density estimate values along the sequence of output values of the algorithm is shown. Also, it is shown that the distance between consecutive points of the output sequence converges to zero, as does the projection of the gradient vector onto the subspace spanned by the $D - d$ largest eigenvectors of the local inverse covariance matrix. These last two properties provide theoretical guarantees for stopping criteria. By modifying the projection step, three variations of the SCMS algorithm are proposed and the running times and performance of the resulting algorithms are compared.

Keywords: Unsupervised learning, subspace constrained mean shift, dimensionality reduction, principal curves, principal surfaces, convergence.

1. Introduction

Dimensionality reduction and manifold-learning techniques provide compact and meaningful representations which facilitate compression, classification, and visualization of high dimensional data. In many applications it is a realistic assumption that the observed high dimensional data have an intrinsically low dimensional structure, so that the data points lie on or near a low dimensional manifold, embedded in the high dimensional space. A multitude of different algorithms have been introduced to find or approximate such a low-dimensional manifold; see, e.g., [1] for an overview.

*Corresponding author. Phone +16135332390; Fax: +16135332964.
Email addresses: aliyari@mast.queensu.ca (Y. Aliyari Ghassabeh), linder@mast.queensu.ca (T. Linder), takahara@mast.queensu.ca (G. Takahara)

In some situations, the observed data can be modeled as low-dimensional ”clean” data corrupted by high-dimensional noise. In this case, applying common linear or nonlinear dimensionality reduction techniques on the noisy observations may not lead to a meaningful low dimensional representation. Partly to overcome this problem, nonlinear generalizations of principal components, called principal curves (and surfaces) have been proposed. The first formal definition of a principal curve was given by Hastie and Stuetzle [2]. According to their definition, a principal curve is a smooth (one-dimensional) curve that passes through the ”middle of a data set.” More formally, a smooth, parameterized curve that does not intersect itself and has finite length inside any bounded ball is a principal curve of a probability distribution if each of its points is the (conditional) mean of the distribution given the set of points that project to it.

Several definitions of principal curves and algorithms to construct them have been proposed based on, or inspired by, Hastie and Stuetzle’s original definition (see [3], [4], [5], [6], [7], [8] among others). The aim of these new definitions and algorithms was to address some of the shortcomings of the original (and subsequent) definition(s) and to extend the range of potential applications. Recently, an interesting new definition of principal curves and surfaces has been proposed by Ozertem and Erdogmus [9]. According to this definition, given a smooth (at least twice continuously differentiable) probability density function (pdf) f on $\mathbb{R}^D$, a d-dimensional principal surface ($d < D$) is the collection of all points where the gradient of f is orthogonal to exactly $D - d$ eigenvectors of the Hessian of f, and the eigenvalues corresponding to these eigenvectors are negative. Thus each point on the principal surface is a local maximum of the pdf in a $(D - d)$-dimensional affine subspace and the principal surface is a d-dimensional ridge of the pdf. An attractive property of this new definition is that the smoothness of principal curves and surfaces is not stipulated by their definition, but rather it is inherited from the smoothness of the underlying pdf or its estimate.

To estimate principal curves/surfaces based on the new definition, [9] proposed the so-called subspace constrained mean shift (SCMS) algorithm. It is a generalization of the well-known mean shift (MS) algorithm [10], [11], [12] that iteratively tries to find modes of a pdf (estimated from data samples) in a local subspace. On synthetic data sets the performance of the SCMS algorithm is comparable to (and in some situations better than) the principal curve algorithms of Hastie and Stuetzle [2] and Kégl *et al.* [7], and it is computationally less demanding. Moreover, in contrast to most previous principal curve algorithms, the SCMS algorithm can naturally handle loops and self-intersections, and it easily generalizes from principal curves to surfaces. Applications to time-series denoising and independent component analysis (among others) were also presented in [9]. Recently, the present authors have successfully applied a version of the SCMS algorithm to vector quantization of noisy sources [13].

Based on an assertion in [12] that the MS algorithm converges, Ozertem and Erdogmus claimed that their SCMS algorithm converges to a principal curve/surface. However, Li *et al*. [14] pointed out a seemingly fundamental mistake in the proof of the convergence of the MS algorithm in [12]. Thus it seems that, similar to most previous principal curve algorithms (with the exception of [7] and [8]), no optimality properties for the SCMS algorithm have been proved.

The purpose of this paper is to investigate some convergence properties of the SCMS algorithm. While we cannot prove that the sequence produced by the algorithm converges (let alone to a principal curve/surface), we show a convergence result concerning the estimated pdf values along the output sequence, which is indicative of the ridge property of the newly defined principal curves. We also show that the two stopping criteria proposed in [9] indeed ensure that the algorithm stops after a finite number of steps. Since these criteria are based on the fact that any point on the principal curve/surface is a fixed point of the SCMS algorithm, these results can be considered as steps toward proving the optimality of the SCMS algorithm, or an improved version of it. In addition, we introduce three variations of the SCMS algorithm for which our convergence results also apply. The performance of these algorithms is compared through simulations.

2. Locally Defined Principal Curves and Surfaces

Let f be a pdf on $\mathbb{R}^D$ that is at least twice continuously differentiable with gradient ∇f and Hessian $\boldsymbol{H}$. For $d \in \{0, 1, \ldots, D-1\}$, Ozertem and Ergodmus defined the d-dimensional principal surfaces associated with the pdf f as follows:

Definition 1 ([9])**.** *The d-dimensional principal surface $\mathcal{P}^d$ associated with pdf f is the collection of all points $\boldsymbol{x} \in \mathbb{R}^D$ such that the gradient $\nabla f(\boldsymbol{x})$ is orthogonal to exactly $D-d$ eigenvectors of the Hessian $\boldsymbol{H}(\boldsymbol{x})$, and the eigenvalues of $\boldsymbol{H}(\boldsymbol{x})$ corresponding to these $D-d$ orthogonal eigenvectors are negative.*

For the one-dimensional ($d = 1$) case, this definition simplifies to the following: the one-dimensional principal surface (principal curve) $\mathcal{P}^1$ is the collection of all points $\boldsymbol{x} \in \mathbb{R}^D$ at which the gradient of the pdf is an eigenvector of the Hessian of the pdf, and the rest of the eigenvectors of the Hessian have negative eigenvalues.

Clearly, all points on a d-dimensional principal surface in Definition 1 are local maxima of the pdf in a local affine orthogonal $D-d$-dimensional subspace. For example, a principal curve is a ridge of the pdf, and every point on the principal curve is a local maximum of the pdf in the affine subspace orthogonal to the curve. Thus Ozertem and Ergodmus' definition replaces Hastie and Stuetzle's requirement that every point on the principal curve be the conditional expectation of the pdf in a local orthogonal subspace with the requirement that the pdf have a local maximum in a local orthogonal subspace.

For Gaussian distribution the principal surfaces of Definition 1 coincide with subspaces spanned by the eigenvectors of the covariance matrix, making connections with principal component analysis [9]. Further existence issues and properties of the new definition for principal surfaces were not treated in detail in [9], but an effective iterative algorithm was given. This algorithm is based on the well known mean shift (MS) procedure, which we quickly review before turning to the subspace constrained mean shift algorithm (SCMS) of [9].

3. The Mean Shift Algorithm

The MS algorithm is a non-parametric, iterative technique for locating modes of a pdf obtained via a kernel density estimate (see, e.g., [15]) from a given data set. These modes play an important role in many machine learning applications, such as classification [12], image segmentation [16], and object tracking [17].

The MS algorithm iteratively updates its mode estimate to a weighted average of the neighboring data points to find a stationary point of the estimated pdf [12]. Specifically, let $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\} \subset \mathbb{R}^D$ denote the observed data. The kernel density estimate with kernel K and bandwidth $h > 0$ is given by

$$\hat{f}(\boldsymbol{x}) = \frac{1}{nh^D} \sum_{i=1}^{n} K\left(\frac{\boldsymbol{x} - \boldsymbol{x}_i}{h}\right), \tag{1}$$

where $K : \mathbb{R}^D \to \mathbb{R}$ is a non-negative function satisfying $\int_{\mathbb{R}^D} K(\boldsymbol{x})\, d\boldsymbol{x} = 1$. Let $\|\boldsymbol{x}\|$ denote the Euclidean norm of the vector $\boldsymbol{x}$. Radially symmetric kernels are defined by $K(\boldsymbol{x}) = ck(\|\boldsymbol{x}\|^2)$, where c is a normalization factor and $k : [0, \infty) \to [0, \infty)$ is called the profile of the kernel, which is assumed to be nonnegative and nonincreasing such that $\int_{\mathbb{R}^D} k(\|\boldsymbol{x}\|^2)\, d\boldsymbol{x} < \infty$. The estimated pdf using the profile k and the bandwidth h has the form

$$\hat{f}(\boldsymbol{x}) = \frac{c}{nh^D} \sum_{i=1}^{n} k\left(\left\|\frac{\boldsymbol{x} - \boldsymbol{x}_i}{h}\right\|^2\right). \tag{2}$$

Assuming that k is differentiable with derivative k', taking the gradient of (2) and equating it to zero yields that modes of the estimated pdf are roots of the function

$$\boldsymbol{m}(\boldsymbol{x}) = \frac{\sum_{i=1}^{n} \boldsymbol{x}_i g\left(\left\|\frac{\boldsymbol{x}-\boldsymbol{x}_i}{h}\right\|^2\right)}{\sum_{i=1}^{n} g\left(\left\|\frac{\boldsymbol{x}-\boldsymbol{x}_i}{h}\right\|^2\right)} - \boldsymbol{x} \tag{3}$$

where $g(t) = -k'(t)$. The vector $\boldsymbol{m}(\boldsymbol{x})$ is called the mean shift vector [12]. The MS algorithm starts from an arbitrary point in $\mathbb{R}^D$ (typically one of the data points) and its mode estimate $\boldsymbol{y}_j$ in the jth iteration is updated as

$$\boldsymbol{y}_{j+1} = \boldsymbol{y}_j + \boldsymbol{m}(\boldsymbol{y}_j). \tag{4}$$

The algorithm iterates this step until $\|\boldsymbol{y}_{j+1} - \boldsymbol{y}_j\| < \epsilon$, where ϵ is some predefined positive threshold.

If the profile k is convex and bounded, the proof of Theorem 1 in [12] implies that $\|\boldsymbol{y}_{j+1} - \boldsymbol{y}_j\| \to 0$ as $j \to \infty$ (the algorithm always stops) and that $\{\hat{f}(\boldsymbol{y}_j); j = 1, 2, \ldots\}$ is an increasing and convergent sequence. However, an error was pointed out in [14] in the proof of the main statement of Theorem 1 in [12] claiming the convergence of the sequence $\{\boldsymbol{y}_i; i = 1, 2, \ldots\}$. Carreira-Perpiñán [18] showed that the MS algorithm with the Gaussian kernel $K(\boldsymbol{x}) = c\, e^{-\|\boldsymbol{x}\|^2}$ is an instance of the EM algorithm and claimed that this fact implies the convergence of $\{\boldsymbol{y}_j\}$. However, without additional conditions the EM algorithm may not converge (see [19] or [20]), and so it appears that the convergence of the MS algorithm has not yet been proved. Incidentally, the error in the original proof for the convergence of the EM algorithm in [21] and the error in the proof of the convergence of the MS algorithm in [12] are both due to the same incorrect use of the triangle inequality.

On the positive side, if $\hat{f}$ has a finite number of stationary points, the convergence of the MS algorithm is not hard to prove (see Theorem 2 in [14]). Unfortunately, a general and useful condition for the finiteness of the set of stationary points of $\hat{f}$ for commonly used kernels such as the Gaussian still seems to be missing (although [18] makes the plausible claim, without proof, that the set of stationary points is always finite for the Gaussian kernel). The following result considers the special case $D = 1$, which may admittedly be of limited interest in applications.

Proposition 1. *For $D = 1$ the mode estimate sequence $\{y_j\}$ generated by the MS algorithm using the profile $k(x) = e^{-x}$ associated with the Gaussian kernel converges to a stationary point of $\hat{f}(x)$.*

Proof. Since $K(x) = c\,e^{-x^2}$, the derivative of the kernel pdf estimate, $\hat{f}'(x)$, is proportional to

$$\sum_{i=1}^{n} (x_i - x) \exp\Big(-\frac{(x - x_i)^2}{h^2}\Big),$$

which is easily seen to be a real analytic function that is not constant on $\mathbb{R}$. Hence the set of stationary points $S = \{x \in \mathbb{R} : \hat{f}'(x) = 0\}$ has no limit points. However, S is a bounded set since one clearly has $S \subset [m, M]$, where $m = \min_{1 \le i \le n} x_i$ and $M = \max_{1 \le i \le n} x_i$, implying that S is finite. Thus $\{y_j\}$ converges to a point in S by Theorem 2 in [14]. □

4. Subspace Constrained Mean Shift Algorithms

Under some regularity conditions, the set of local maxima of a pdf is exactly the zero-dimensional principal manifold $\mathcal{P}^0$ resulting from Definition 1 for $d = 0$. The SCMS algorithm [9] generalizes the MS algorithm to estimate higher order principal curves and surfaces ($d \ge 1$). Similar to the MS algorithm, the SCMS algorithm starts from a finite data set sampled from the probability distribution, forms a kernel density estimate $\hat{f}$ based on the data, and in each iteration it evaluates the MS vector. However, the SCMS algorithm projects the mean shift vector to the local (affine) subspace spanned by the $D - d$ eigenvectors corresponding to the $D - d$ largest eigenvalues of the so-called local inverse covariance matrix of the pdf estimate at that point, given by

$$\hat{\mathbf{\Sigma}}^{-1}(\boldsymbol{x}) = -\hat{\boldsymbol{H}}(\boldsymbol{x})\hat{f}(\boldsymbol{x})^{-1} + \nabla \hat{f}(\boldsymbol{x}) \nabla \hat{f}(\boldsymbol{x})^T \hat{f}(\boldsymbol{x})^{-2}, \tag{5}$$

where $\hat{\boldsymbol{H}}(\boldsymbol{x})$ and $\nabla \hat{f}(\boldsymbol{x})$ are the Hessian and gradient of the pdf estimate at $\boldsymbol{x}$, respectively. Here and throughout the paper bold lowercase letters denote column vectors of appropriate dimensions, and $\boldsymbol{x}^T$ is the transpose of $\boldsymbol{x}$. Note that $\hat{\mathbf{\Sigma}}^{-1}$ is the negative Hessian of the logarithm of $\hat{f}$. The motivation for this definition and its connections to Definition 1 and principal component analysis are discussed in detail in [9]. Note that $\hat{\mathbf{\Sigma}}^{-1}$ is well defined and symmetric if $\hat{f}$ is positive and twice continuously differentiable everywhere. The SCMS algorithm can be summarized as follows.

1, Set $\epsilon > 0$, $j = 1$, and initialize the SCMS algorithm to an arbitrary point $\boldsymbol{y}_1$.

2. Evaluate the mean shift vector $\boldsymbol{m}(\boldsymbol{y}_j)$ using (3).

3. Evaluate the gradient, the Hessian matrix, and the local inverse covariance matrix $\hat{\mathbf{\Sigma}}^{-1}$ given in (5) at $\boldsymbol{y}_j$. Perform the eigendecomposition of $\hat{\mathbf{\Sigma}}_j^{-1} = \hat{\mathbf{\Sigma}}^{-1}(\boldsymbol{y}_j)$ and find its eigenvalues and eigenvectors.

4. Let $\boldsymbol{V}_j = [\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{D-d}]$ be the $D \times (D-d)$ matrix whose columns are the $D-d$ orthonormal eigenvectors corresponding to the $D-d$ largest eigenvalues of $\hat{\mathbf{\Sigma}}_j^{-1}$.

5. Compute $\boldsymbol{y}_{j+1} = \boldsymbol{V}_j \boldsymbol{V}_j^T \boldsymbol{m}(\boldsymbol{y}_j) + \boldsymbol{y}_j$.

6. Stop if $\|\boldsymbol{y}_{j+1} - \boldsymbol{y}_j\| < \epsilon$; otherwise increment j by 1 and go to step 2.

Remark. In [9], the stopping rule $\|\boldsymbol{y}_{j+1} - \boldsymbol{y}_j\| < \epsilon$ was suggested as an alternative to the recommended rule

$$\frac{\|\boldsymbol{V}_{j+1}^T \nabla \hat{f}(\boldsymbol{y}_{j+1})\|}{\|\nabla \hat{f}(\boldsymbol{y}_{j+1})\|} < \epsilon$$

which is meant to check if the gradient is (nearly) orthogonal to the subspace spanned by the columns of $\boldsymbol{V}_j$. However, this criterion seems to be problematic (e.g., the denominator is zero if the algorithm starts at a stationary point). We will later consider the following simpler stopping rule of a similar flavor:

6'. Stop if $\|\boldsymbol{V}_{j+1}^T \nabla \hat{f}(\boldsymbol{y}_{j+1})\| < \epsilon$; otherwise increment j by 1 and go to step 2.

Typically, n instances of the SCMS algorithm are run, each time initialized to one of the n data points. The resulting n output points are considered as a discrete approximation to the underlying principal curve or surface; see the illustrative example in Figure 1. In both the MS and the SCMS algorithms the stopping threshold ϵ is set manually so that a good tradeoff between running time and approximation accuracy is achieved. The problem of selecting the bandwidth h for the MS algorithm is discussed in detail in [12], and variable-bandwidth, locally-adaptive MS algorithms are introduced and investigated in [22]. The bandwidth selection problem for the SCMS algorithm is discussed in detail in [9], but it is not clear that the automatic rules suggested from the literature of kernel density estimation are in any way optimal when applied in conjunction with the SCMS algorithm.

In [9] extensive simulation results on artificial data demonstrated the ability of the algorithm to well approximate principal curves and surfaces. As well, promising applications of the SCMS algorithm to time-varying MIMO channel equalization and time series signal denoising were discussed. We note here that an algorithm for manifold

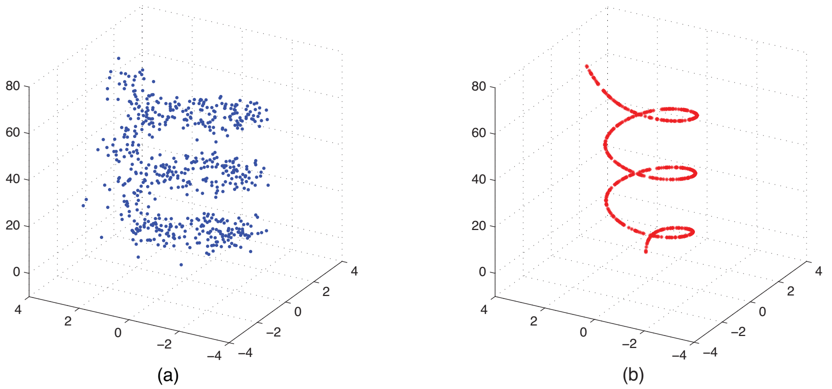


Figure 1: (a) $n = 600$ data points were generated by adding 3-dimensional standard Gaussian noise samples to 600 points uniformly sampled on a spiral in $\mathbb{R}^3$. (b) The output of the SCMS algorithm using $D = 3$, $d = 1$, the Gaussian kernel with bandwidth $h = 3$, and stopping threshold $\epsilon = 0.005$.

denoising that is somewhat similar in spirit to SCMS, but which is based on the blurring version of the MS procedure, was given by Wang and Carreira-Perpiñán [25].

On the theoretical side, [9] claimed that the SCMS algorithm will converge to a point on the principal surface with appropriate dimensionality. This claim was based on the assumption that the MS algorithm always converges which, as we discussed, has so far been unproven. In addition, it does not seem clear at all that the convergence of MS actually implies the convergence of SCMS, let alone its convergence to the principal surface.

The next proposition states three convergence results relating to the density estimate values produced by the SCMS algorithm and the two stopping criteria presented earlier. The proof is given in the Appendix.

Proposition 2. *Assume the kernel pdf estimator $\hat{f}$ is defined as in* (2) *with a radially symmetric kernel K having profile k which is positive, nonincreasing, convex, and such that the function $t \mapsto k(t^2)$ is twice continuously differentiable at all $t \in \mathbb{R}$. Let $\{\boldsymbol{y}_j\}$ denote the sequence of points generated by the SCMS algorithm with arbitrary initialization. Then the following hold:*

(i) *The sequence $\{\hat{f}(\boldsymbol{y}_j)\}$ is nondecreasing and convergent.*

(ii) $\lim_{j\to\infty} \|\boldsymbol{y}_{j+1} - \boldsymbol{y}_j\| = 0.$

(iii) $\lim_{j\to\infty} \|\boldsymbol{V}_j^T \nabla \hat{f}(\boldsymbol{y}_j)\| = 0.$

Remarks.

(a) Parts (i) and (ii) of the proposition are analogous to what is proved in Theorem 1 of [12] for the MS algorithm, with some proof ideas being also similar. All three statements indicate (but by no means prove) the ability of the SCMS algorithm to converge to the principal surface of dimension d. In particular, (i) is related to the "ridge" property of locally defined principal curves and surfaces, (ii) and (iii) provide useful stopping criteria, while (iii) is related to the fact that at any point $\boldsymbol{y}$ of $\mathcal{P}^d$ one must have $\boldsymbol{V}(\boldsymbol{y})^T \nabla \hat{f}(\boldsymbol{y}_j) = \mathbf{0}$, where $\boldsymbol{V}(\boldsymbol{y})$ is the $D \times (D-d)$ matrix whose columns are the $D-d$ orthonormal eigenvectors corresponding to the $D-d$ largest eigenvalues of $\hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{y})$.

(b) The differentiability condition on the profile k ensures that $\hat{f}$ is twice continuously differentiable so that all quantities used in the SCMS updates are well defined no matter how the algorithm is initialized. The condition that the kernel K is integrable and the conditions on k imposed in the proposition imply that k is bounded, its derivative k' is nondecreasing and negative on $[0,\infty)$, and both $k(x)$ and $k'(x)$ converge to zero as $x \to \infty$. The profile $k(x) = e^{-x}$ of the widely used Gaussian kernel satisfies these conditions.

(c) At the price of complicating the notation, the proof can straightforwardly be extended to more general kernel density estimates of the form

$$\hat{f}(\boldsymbol{x}) = \frac{c}{nh^D} \sum_{i=1}^{n} k\left(\left\| \frac{\boldsymbol{x} - \boldsymbol{x}_i}{h} \right\|_{\boldsymbol{K}_i}^2 \right)$$

where $\|\boldsymbol{y}\|_{\boldsymbol{K}_i}^2 = \boldsymbol{y}^T \boldsymbol{K}_i \boldsymbol{y}$, with $\boldsymbol{K}_i$, $i = 1, \ldots, n$ being symmetric and positive definite $D \times D$ matrices. The potential usefulness of considering such more general estimates, which may account better for anisotropy and local scale information in the data sample, has been argued in [14].

5. Simulation Results

An inspection of the proof of Proposition 2 shows that all three statements remain valid if $\boldsymbol{V}_j$, $j = 1, 2, \ldots$, is an arbitrary sequence of $D \times (D-d)$ matrices having orthonormal columns. Thus for the convergence results to hold, $\boldsymbol{V}_j$ does not have to be the matrix whose columns are the $D-d$ orthonormal eigenvectors corresponding to the largest eigenvalues of $\hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{y}_j)$.

Of course, for the outputs of the algorithm to be meaningful the columns of $\boldsymbol{V}_j$ should be (nearly) orthogonal to the gradient of $\hat{f}$ at points on the d-dimensional principal surface of $\hat{f}$. The choice of $\hat{\boldsymbol{\Sigma}}^{-1}$ was motivated in [9] by Definition 1 and the connection to principal components when the underlying pdf is Gaussian. In this case the local inverse covariance matrix (of the Gaussian pdf, not estimated from data) is just the inverse covariance matrix of the Gaussian pdf up to a constant at any point with eigendirections the principal component directions. In practice, the density estimate $\hat{f}$ is never Gaussian so the use of $\hat{\boldsymbol{\Sigma}}^{-1}$ seems less well motivated for the SCMS algorithm than simply using the estimated Hessian $\hat{\boldsymbol{H}}$, which is a more natural choice in the context of Definition 1, as well as requiring slightly fewer operations to compute. At points $\boldsymbol{x}$ on the d-dimensional principal surface of $\hat{f}$, the gradient $\nabla\hat{f}(\boldsymbol{x})$ is orthogonal to exactly $D-d$ eigenvectors of $\hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{x})$ and to exactly $D-d$ eigenvectors of $\hat{\boldsymbol{H}}(\boldsymbol{x})$, and these two sets of eigenvectors are the same (see [9]). The eigenvalues of $\hat{\boldsymbol{H}}(\boldsymbol{x})$ associated with these eigenvectors are $-\hat{f}(\boldsymbol{x})$ times the corresponding eigenvalues of $\hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{x})$ and so we form $\boldsymbol{V}_j$ from the $D-d$ eigenvectors of $\hat{\boldsymbol{H}}(\boldsymbol{y}_j)$ corresponding to the $D-d$ *smallest* eigenvalues of $\hat{\boldsymbol{H}}(\boldsymbol{y}_j)$.

In this section, we compare the use in the SCMS algorithm of $\hat{\boldsymbol{\Sigma}}^{-1}$, $\hat{\boldsymbol{H}}$, and two local estimates (local to $\boldsymbol{y}_j$) of the covariance matrix of $\hat{f}$ due to Wang and Carreira-Perpiñán [25]. In the resulting three variations of the original SCMS algorithm, the mean shift vectors and output updates are computed using (3) and Step 5 of the SCMS algorithm, respectively, but instead of the local inverse covariance matrix in (5), three different matrices are used. Let $\{\boldsymbol{y}_j^1,\ldots,\boldsymbol{y}_j^n\}$ denote the set of outputs after the jth iteration, where $\boldsymbol{y}_j^{(i)}$ is the output of the algorithm when it is initialized to the ith data point $\boldsymbol{x}_i$, $i=1,\ldots,n$. In the jth iteration, the proposed matrices at a point $\boldsymbol{x}$ (set to one of the points $\boldsymbol{y}_j^{(i)}$) are

(i) The Hessian of $\hat{f}$,

$$\hat{\boldsymbol{H}}(\boldsymbol{x})=\frac{c}{nh^{2+D}}\sum_{i=1}^{n}\left(-\boldsymbol{I}+\frac{2(\boldsymbol{x}-\boldsymbol{x}_i)(\boldsymbol{x}-\boldsymbol{x}_i)^T}{h^2}\right)\exp\left(-\frac{\|\boldsymbol{x}-\boldsymbol{x}_i\|^2}{2h^2}\right),$$

where c is the kernel profile normalization factor and $\boldsymbol{I}$ is the $D\times D$ identity matrix;

(ii) The estimated local covariance matrix using the κ nearest *data points*,

$$\hat{\boldsymbol{\Sigma}}_\kappa(\boldsymbol{x})=\frac{1}{\kappa-1}\sum_{\boldsymbol{x}_i\in N_\kappa(\boldsymbol{x})}(\boldsymbol{x}_i-\boldsymbol{m}_\kappa(\boldsymbol{x}))(\boldsymbol{x}_i-\boldsymbol{m}_\kappa(\boldsymbol{x}))^T,$$

where $N_\kappa(\boldsymbol{x})$ is the set of the κ nearest neighbors of $\boldsymbol{x}$ in the observed data set $\{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n\}$, and $\boldsymbol{m}_\kappa(\boldsymbol{x})$ is the average over members of $N_\kappa(\boldsymbol{x})$;

(iii) The estimated local covariance matrix using the κ nearest *outputs*,

$$\hat{\boldsymbol{\Sigma}}_{\kappa,j}(\boldsymbol{x})=\frac{1}{\kappa-1}\sum_{\boldsymbol{y}_j^{(i)}\in N_{\kappa,j}(\boldsymbol{x})}(\boldsymbol{y}_j^{(i)}-\boldsymbol{m}_{\kappa,j}(\boldsymbol{x}))(\boldsymbol{y}_j^{(i)}-\boldsymbol{m}_{\kappa,j}(\boldsymbol{x}))^T,$$

where $N_{\kappa,j}(\boldsymbol{x})$ is the set of the κ nearest neighbors of $\boldsymbol{x}$ among the outputs $\{\boldsymbol{y}_j^1,\ldots,\boldsymbol{y}_j^n\}$ at the jth iteration and $\boldsymbol{m}_{\kappa,j}(\boldsymbol{x})$ is the average over members of $N_{\kappa,j}(\boldsymbol{x})$. In this case we update all the outputs in each iteration.

For each matrix above the matrix $\boldsymbol{V}_j$ at Step 4 of the SCMS algorithm is given by

$$\boldsymbol{V}_j=[\boldsymbol{v}_{d+1},\ldots,\boldsymbol{v}_D],$$

where $\boldsymbol{v}_i, i=d+1,\ldots,D$ are the $D-d$ eigenvectors corresponding to the $D-d$ *smallest* eigenvalues. The projection step and termination criterion are the same as in Steps 5 and 6, respectively, in the SCMS algorithm. Proposition 2 guarantees that each of the resulting three SCMS algorithm variations stops after a finite number of iterations.

The projection of the MS vectors onto the subspace spanned by the eigenvectors of the Hessian matrix corresponding to the $D - d$ smallest eigenvalues complies with Definition 1, since a point $\boldsymbol{x}$ is located on the d-dimensional principal surface if the gradient at $\boldsymbol{x}$ is orthogonal to the $D - d$ smallest eigenvectors of the Hessian at $\boldsymbol{x}$ and the corresponding eigenvalues are negative [26]. The matrices in (ii) and (iii) follow Wang and Carreira-Perpiñán [25]. There the authors computed the blurred MS vectors using the blurring version of the MS algorithm [27] and then a corrector projective step is computed to constrain the motion to be orthogonal to the underlying manifold.

Although using only the κ nearest neighbors instead of the whole data set to estimate the projection matrix does not change the theoretical complexity in each iteration, in practice with a finite data set the running time significantly reduces. A good value of κ will in general depend on the structure of the underlying manifold. In our simulations we chose κ to be between 4 and 6 percent of the number of observations, but setting κ in general is beyond the scope of this paper. We note that the authors in [25] suggest that κ typically should grow sublinearly with the sample size n.

In the rest of this section, we present a simulation example using the original SCMS algorithm and our three variations on the two and three dimensional spiral. The input data are generated as

$$x_i = u_i + e_i, \qquad i = 1, \ldots, n,$$

where the u_i's are independently and uniformly selected on the two or three dimensional spiral, called the generative curve, and the e_i's are independent, zero mean spherical Gaussian random vectors of appropriate dimension, independent of the u_i's and with component variance σ^2. We used $\epsilon = 0.01$ in the stopping criterion in Step 6 of the SCMS algorithm in all runs. For the two dimensional spiral we used $n = 1000$ data samples, $\sigma^2 = 1$ for the noise variance, $h = 2$ for the bandwidth of the kernel density estimator, and $\kappa = 50$ nearest neighbors for computing the two variations of the local covariance matrix. For the three dimensional spiral we used $n = 600$, $\sigma^2 = 0.6$, $h = 3$, and $\kappa = 40$. For performance evaluation we computed the average squared Euclidean distance between the output points and the closest points on the generative curve, and the average running time, in seconds. All simulations were run using Matlab on a desktop computer with an Intel Core i7-870 processor.

2-d Spiral	SCMS	Hessian	Cov. 1	Cov. 2
Running time (sec.)	11.34	11.34	3.91	3.85
Av. Squared Euclidean Distance	0.074	0.075	0.077	0.077
3-d Spiral	SCMS	Hessian	Cov. 1	Cov. 2
Running time (sec.)	109.89	111.56	19.53	17.89
Av. Squared Euclidean Distance	0.273	0.299	0.152	0.152

Table 1: Performance results for the two and three dimensional spirals.

Table 1 shows the results for the two and three dimensional spirals using the original SCMS algorithm and the three variations using the Hessian, the local covariance matrix using the original data points (Cov. 1), and the local covariance matrix using the output points in each iteration (Cov. 2) in place of the inverse covariance matrix. Performance in terms of closeness to the generative curve is similar for all 4 variations though, interestingly, use of the local covariance matrices gives no worse performance. In terms of runtime, the local covariance matrices perform significantly better, as expected. Adaptive optimization of the local neighborhood size κ should yield improved performance. We note that we also tested all four algorithms on a two dimensional circle with very similar results.

Figures 2 and 3 show the generative curve, the simulated data points, and the output points from the four versions of the algorithm, for the two dimensional and the three dimensional spiral, respectively. All four versions of the algorithm show similar performance visually.

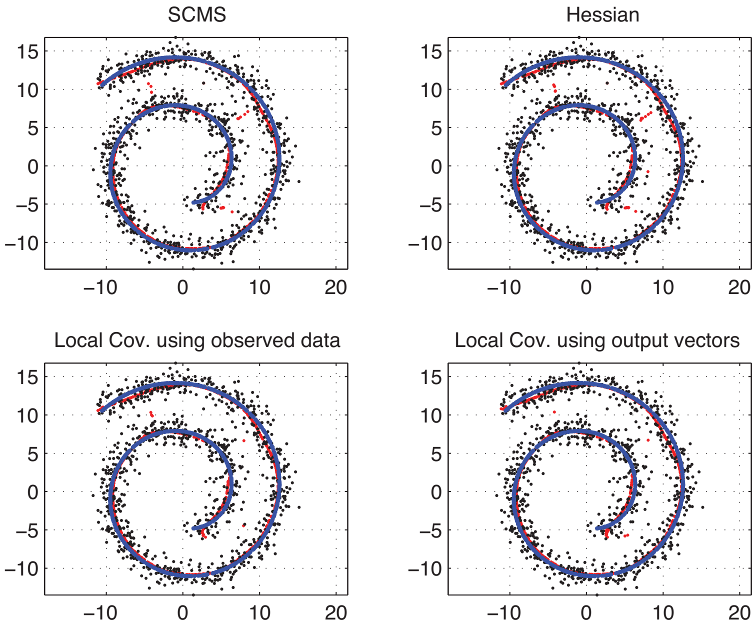
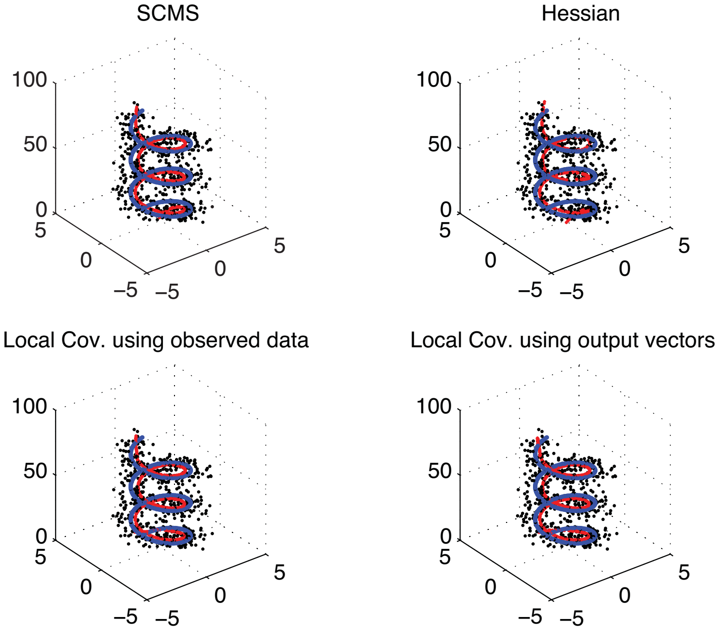


Figure 2: The blue points are $n = 1000$ samples uniformly selected on the two dimensional spiral generative curve, the red points are the outputs of each algorithm, and the black points are the observed data points generated by adding independent, zero mean Gaussian noise to the points on the generative curve.



SCMS

Hessian

Local Cov. using observed data

Local Cov. using output vectors

Figure 3: The blue points are $n = 600$ samples uniformly selected on the three dimensional spiral generative curve, the red points are the outputs of each algorithm, and the black points are the observed data points generated by adding independent, zero mean Gaussian noise to the points on the generative curve.

6. Discussion

We studied the SCMS algorithm for finding principal curves and proved convergence result indicating that it inherits some important convergence properties of the MS algorithm. The more challenging problem of proving the convergence of the sequence generated by the SCMS algorithm is the subject of future research. Further along this line, the study of the optimality of the SCMS algorithm (i.e., its convergence to a principal curve/surface), seems to necessitate a more careful study of the definition of locally defined principal curves and surfaces. In particular, it is likely that existence issues should be resolved and differential geometric properties studied, before optimality issues can be addressed.

Appendix

Proof of Proposition 2. The subspace constrained mean shift sequence $\{\boldsymbol{y}_j\}$ is defined recursively by

$$\boldsymbol{y}_{j+1} = \boldsymbol{V}_j \boldsymbol{V}_j^T \boldsymbol{m}(\boldsymbol{y}_j) + \boldsymbol{y}_j, \tag{6}$$

where

$$\boldsymbol{m}(\boldsymbol{y}_j) = \frac{\sum_{i=1}^{n} \boldsymbol{x}_i g(\|\frac{\boldsymbol{y}_j - \boldsymbol{x}_i}{h}\|^2)}{\sum_{i=1}^{n} g(\|\frac{\boldsymbol{y}_j - \boldsymbol{x}_i}{h}\|^2)} - \boldsymbol{y}_j, \tag{7}$$

with $\boldsymbol{y}_1$ being an arbitrary starting point. Here $g(x) = -k'(x)$, where k is the profile of kernel K and $\boldsymbol{V}_j$ is the $D \times (D-d)$ matrix having orthonormal columns which are eigenvectors corresponding to the largest eigenvalues of the local inverse covariance matrix $\hat{\boldsymbol{\Sigma}}^{-1}$ evaluated at $\boldsymbol{y}_j$.

Since the profile k is bounded, the sequence $\{\hat{f}(\boldsymbol{y}_j)\}$ is bounded, so it suffices to show that the sequence is nondecreasing to prove convergence. The convexity of k implies that $k(t_2) - k(t_1) \geq g(t_1)(t_1 - t_2)$ for all $t_1, t_2 \geq 0$, where $g = -k'$. This and the definition of $\hat{f}$ yield

$$\begin{aligned} \hat{f}(\boldsymbol{y}_{j+1}) - \hat{f}(\boldsymbol{y}_j) &= \frac{c}{nh^D} \sum_{i=1}^{n} \Big(k\Big(\Big\| \frac{\boldsymbol{y}_{j+1} - \boldsymbol{x}_i}{h} \Big\|^2 \Big) - k\Big(\Big\| \frac{\boldsymbol{y}_j - \boldsymbol{x}_i}{h} \Big\|^2 \Big) \Big) \\ &\geq \frac{c}{nh^{D+2}} \sum_{i=1}^{n} g\Big(\Big\| \frac{\boldsymbol{y}_j - \boldsymbol{x}_i}{h} \Big\|^2 \Big) \big(\|\boldsymbol{y}_j - \boldsymbol{x}_i\|^2 - \|\boldsymbol{y}_{j+1} - \boldsymbol{x}_i\|^2 \big) \\ &= C_j \sum_{i=1}^{n} p_j(i) \big(\|\boldsymbol{y}_j - \boldsymbol{x}_i\|^2 - \|\boldsymbol{y}_{j+1} - \boldsymbol{x}_i\|^2 \big), \end{aligned} \tag{8}$$

where

$$p_j(i) = \frac{g\big(\|\frac{\boldsymbol{y}_j - \boldsymbol{x}_i}{h}\|^2\big)}{\sum_{k=1}^{n} g\big(\|\frac{\boldsymbol{y}_j - \boldsymbol{x}_k}{h}\|^2\big)}, \quad i = 1, \ldots, n$$

and

$$C_j = \frac{c}{nh^{D+2}} \sum_{i=1}^{n} g\Big(\Big\| \frac{\boldsymbol{y}_j - \boldsymbol{x}_i}{h} \Big\|^2 \Big).$$

Since $g(t) > 0$ for all $t \geq 0$, $p_j(1), \ldots, p_j(n)$ are well defined, positive, and sum to 1. In fact, the name "mean shift" derives from the fact that the mean shift of $\boldsymbol{y}_j$, given in (3), can be written in terms of an expectation; namely

$$\boldsymbol{m}(\boldsymbol{y}_j) = \sum_{i=1}^{n} p_j(i)(\boldsymbol{x}_i - \boldsymbol{y}_j) = E[\boldsymbol{Z}_j],$$

where $\boldsymbol{Z}_j$ is an $\mathbb{R}^D$-valued random vector with discrete distribution given by $\Pr(\boldsymbol{Z}_j = \boldsymbol{x}_i - \boldsymbol{y}_j) = p_j(i),\ i = 1, \ldots, n$. Thus, letting $\boldsymbol{T}_j = \boldsymbol{V}_j \boldsymbol{V}_j^T$, the SCMS update step can be rewritten as

$$\boldsymbol{y}_{j+1} - \boldsymbol{y}_j = \boldsymbol{T}_j \boldsymbol{m}(\boldsymbol{y}_j) = \boldsymbol{T}_j E[\boldsymbol{Z}_j]. \tag{9}$$

Let $\boldsymbol{W}_j$ be a $D \times D$ matrix representing any orthogonal projection onto the null space of $\boldsymbol{T}_j$. Then $\boldsymbol{x} = \boldsymbol{T}_j \boldsymbol{x} + \boldsymbol{W}_j \boldsymbol{x}$ for all $\boldsymbol{x} \in \mathbb{R}^D$, and $\boldsymbol{T}_j \boldsymbol{x}$ and $\boldsymbol{W}_j \boldsymbol{y}$ are orthogonal for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^D$. We can rewrite the last sum in (8) as follows

$$\begin{aligned}
&\sum_{i=1}^{n} p_j(i)\big(\|\boldsymbol{x}_i - \boldsymbol{y}_j\|^2 - \|\boldsymbol{x}_i - \boldsymbol{y}_{j+1}\|^2\big) \\
&\qquad = \quad E\big[\|\boldsymbol{Z}_j\|^2\big] - E\big[\big\|\boldsymbol{Z}_j - \boldsymbol{T}_j E[\boldsymbol{Z}_j]\big\|^2\big] \\
&\qquad = \quad E\big[\|\boldsymbol{W}_j \boldsymbol{Z}_j\|^2 + \|\boldsymbol{T}_j \boldsymbol{Z}_j\|^2\big] - E\big[\|\boldsymbol{W}_j \boldsymbol{Z}_j\|^2 + \big\|\boldsymbol{T}_j \boldsymbol{Z}_j - \boldsymbol{T}_j E[\boldsymbol{Z}_j]\big\|^2\big] \\
&\qquad = \quad E\big[\|\boldsymbol{T}_j \boldsymbol{Z}_j\|^2\big] - E\big[\big\|\boldsymbol{T}_j \boldsymbol{Z}_j - E[\boldsymbol{T}_j \boldsymbol{Z}_j]\big\|^2\big] \\
&\qquad = \quad \big\|E[\boldsymbol{T}_j \boldsymbol{Z}_j]\big\|^2 = \|\boldsymbol{y}_{j+1} - \boldsymbol{y}_j\|^2,
\end{aligned}$$

where in the penultimate equality we applied the identity $E[Z^2] = \mathrm{Var}[Z] + (E[Z])^2$, valid for real random variables with finite variance, to the components of $\boldsymbol{T}_j \boldsymbol{Z}_j$. Combining this with (8), we obtain

$$\hat{f}(\boldsymbol{y}_{j+1}) - \hat{f}(\boldsymbol{y}_j) \geq C_j \|\boldsymbol{y}_{j+1} - \boldsymbol{y}_j\|^2, \tag{10}$$

where $C_j > 0$, which implies that $\{\hat{f}(\boldsymbol{y}_j)\}$ is nondecreasing and thus convergent, proving part (i) of the proposition.

To prove part (ii), we note that $k(x) > 0$ for all $x \geq 0$ implies that $\hat{f}(\boldsymbol{y}_1) > 0$, so part (i) yields $\min\{\hat{f}(\boldsymbol{y}_j) : j \geq 1\} = \hat{f}(\boldsymbol{y}_1) > 0$. But this in turn implies that $\{\boldsymbol{y}_j\}$ is a bounded sequence, since otherwise it would have a subsequence $\{\boldsymbol{y}_{j_k}\}$ such that $\lim_{k\to\infty} \|\boldsymbol{y}_{j_k}\| = \infty$ which, in view of $\lim_{x\to\infty} k(x) = 0$, would give $\lim_{k\to\infty} \hat{f}(\boldsymbol{y}_{j_k}) = 0$, contradicting our uniform positive lower bound on the $\hat{f}(\boldsymbol{y}_j)$.

In view of the above, there exists $R > 0$ such that $\|\boldsymbol{y}_j - \boldsymbol{x}_i\| \leq R$ for all $j \geq 1$ and $i = 1, \ldots, n$. Since $g = -k'$ is nonincreasing on $[0, \infty)$, we obtain

$$C_j = \frac{c}{nh^{D+2}} \sum_{k=1}^{n} g\Big(\Big\|\frac{\boldsymbol{y}_j - \boldsymbol{x}_k}{h}\Big\|^2\Big) \geq \frac{c}{h^{D+2}} g\Big(\frac{R^2}{h^2}\Big) = C,$$

where $C > 0$ since $g(x) > 0$ for all $x \geq 0$. Thus (10) implies

$$\|\boldsymbol{y}_{j+1} - \boldsymbol{y}_j\|^2 \leq C^{-1}\big(\hat{f}(\boldsymbol{y}_{j+1}) - \hat{f}(\boldsymbol{y}_j)\big),$$

and since $\lim\limits_{j\to\infty} \big(\hat{f}(\boldsymbol{y}_{j+1}) - \hat{f}(\boldsymbol{y}_{j+1})\big) = 0$ by part (i), we obtain $\lim\limits_{j\to\infty} \|\boldsymbol{y}_{j+1} - \boldsymbol{y}_j\| = 0$.

Finally, to show (iii) we note that by definition (2) of $\hat{f}$,

$$\begin{aligned}
\nabla \hat{f}(\boldsymbol{y}_j) &= \frac{2c}{nh^{D+2}} \sum_{i=1}^{n} (\boldsymbol{x}_i - \boldsymbol{y}_j) g\Big(\Big\|\frac{\boldsymbol{y}_j - \boldsymbol{x}_i}{h}\Big\|^2\Big) \\
&= \frac{2c}{nh^{D+2}} \left[\sum_{i=1}^{n} g\Big(\Big\|\frac{\boldsymbol{y}_j - \boldsymbol{x}_i}{h}\Big\|^2\Big)\right] \left[\frac{\sum_{i=1}^{n} \boldsymbol{x}_i g(\|\frac{\boldsymbol{x}_i - \boldsymbol{y}_j}{h}\|^2)}{\sum_{i=1}^{n} g(\|\frac{\boldsymbol{x}_i - \boldsymbol{y}_j}{h}\|^2)} - \boldsymbol{y}_j\right] \\
&= \frac{2c}{nh^{D+2}} \left[\sum_{i=1}^{n} g\Big(\Big\|\frac{\boldsymbol{y}_j - \boldsymbol{x}_i}{h}\Big\|^2\Big)\right] \boldsymbol{m}(\boldsymbol{y}_j).
\end{aligned}$$

Therefore,

$$\|\boldsymbol{V}_j^T \nabla \hat{f}(\boldsymbol{y}_j)\| = \frac{2c}{nh^{D+2}} \left[\sum_{i=1}^{n} g\Big(\Big\| \frac{\boldsymbol{y}_j - \boldsymbol{x}_i}{h} \Big\|^2 \Big) \right] \|\boldsymbol{V}_j^T \boldsymbol{m}(\boldsymbol{y}_j)\|.$$

Since $\boldsymbol{V}_j$ has orthonormal columns and $\boldsymbol{T}_j = \boldsymbol{V}_j \boldsymbol{V}_j^T$, we have $\|\boldsymbol{T}_j \boldsymbol{m}(\boldsymbol{y}_j)\| = \|\boldsymbol{V}_j^T \boldsymbol{m}(\boldsymbol{y}_j)\|$. This and (9) yield

$$\|\boldsymbol{V}_j^T \nabla \hat{f}(\boldsymbol{y}_j)\| = \frac{2c}{nh^{D+2}} \Big[\sum_{i=1}^{n} g\Big(\Big\| \frac{\boldsymbol{y}_j - \boldsymbol{x}_i}{h} \Big\|^2 \Big) \Big] \|\boldsymbol{y}_{j+1} - \boldsymbol{y}_j\|$$

so part (iii) follows from part (ii) and the fact that the conditions on k ensure that $g = -k'$ is bounded. □

Acknowledgment

This research was supported in part by the Natural Sciences and Engineering Research Council (NSERC) of Canada.

References

References

[1] J. A. Lee, M. Verleysen, Nonlinear Dimensionality Reduction, Springer-Verlag, 2007.

[2] T. Hastie, W. Stuetzle, Principal curves, Journal of the American Statistical Association 85 (1989) 502–516.

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

[4] R. Tibshirani, Principal curves revisited, Statistics and Computation 2 (1992) 183–190.

[5] K. Y. Chang, J. Ghosh, A unified model for probabilistic principal surfaces, IEEE Trans. on Pattern Analysis and Machine Intelligence 23 (2001) 22–41.

[6] P. Delicado, Another look at principal curves and surfaces, Journal of Multivariate Analysis 77 (2001) 84–116.

[7] B. Kegl, A. Krzyzak, T. Linder, K. Zeger, Learning and design of principal curves, IEEE Trans. on Pattern Analysis and Machine Intelligence 22 (2000) 281–297.

[8] S. Sandilya, S. Kulkarni, Principal curves with bounded turn, IEEE Trans. on Information Theory 48 (2002) 2789–2793.

[9] U. Ozertem, D. Erdogmus, Locally defined principal curves and surfaces, Journal of Machine Learning Research 12 (2011) 1249–1286.

[10] K. Fukunaga, L. D. Hostetler, Estimation of the gradient of a density function, with applications in pattern recognition, IEEE Trans. on Inform. Theory 21 (1975) 32–40.

[11] Y. Cheng, Mean shift, mode seeking and clustering, IEEE Trans. on Pattern Analysis and Machine Intelligence 17 (1995) 790–799.

[12] D. Comaniciu, P. Meer, Mean shift: A robust approach toward feature space analysis, IEEE Trans. on Pattern Analysis and Machine Intelligence 24 (2002) 603–619.

[13] Y. A. Ghassabeh, T. Linder, G. Takahara, On noisy source vector quantization via a subspace constrained mean shift algorithm, in: Proc. 26th Biennial Symp. on Communications, Kingston, Canada, pp. 107–110, 2012.

[14] X. Li, Z. Hu, F. Wu, A note on the convergence of the mean shift, Pattern Recognition 40 (2007) 1756–1762.

[15] B. W. Silverman, Density Estimation for Statistics and Data Analysis, Chapman and Hall, 1986.

[16] J. Wang, B. Thiesson, Y. Xu, M. Cohen, Image and video segmentation by anisotropic kernel mean shift, in: Proc. European Conf. on Computer Vision, Prague, Czech Republic, pp. 238–250, 2004.

[17] A. Yilmaz, Object tracking by asymmetric kernel mean shift with automated scale and orientation selection, in: Proc. IEEE Conf. on Computer Vision and Pattern Recognition, Minnesota, USA, pp. 18–23, 2007.

[18] M. A. Carreira-Perpiñán, Gaussian mean shift is an EM algorithm, IEEE Trans. on Pattern Analysis and Machine Intelligence 29 (2007) 767–776.

[19] R. A. Boyles, On the convergence of the EM algorithm, Journal of the Royal Statistical Society: Series B 45 (1983) 47–50.

[20] C. F. J. Wu, On the convergence properties of the EM algorithm, The Annals of Statistics 11 (1983) 95–103.

[21] A. P. Dempster, N. M. Laird, D. B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, Journal of the Royal Statistical Society: Series B 39 (1977) 1–38.

[22] D. Comaniciu, V. Ramesh, P. Meer, The variable bandwidth mean shift and data-driven scale selection, in: Proc. 8th Intl. Conf. Computer Vision, Princeton, USA, pp. 438–445.

[23] S. Ray, B.G. Lindsay, The topography of multivariate normal mixtures, The Annals of Statistics 33 (2005) 2042–2065.

[24] M. Golubitsky, V. Guillemin, Stable Mappings and their Singularities, Springer-Verlag, 1973.

[25] W. Wang, M. A. Carreira-Perpiñán, Manifold blurring mean shift algorithms for manifold denoising, in: Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR 2010), San Francisco, USA, pp. 1759–1766, 2010.

[26] D. Eberly, Ridges in Image and Data Analysis, Kluwer, 1996.

[27] M. A. Carreira-Perpiñán, Fast nonparametric clustering with Gaussian blurring mean-shift, in: 23rd Int. Conf. Machine Learning (ICML 2006), Pittsburgh, USA, pp. 153–160, 2006.