Journal of Classification
https://doi.org/10.1007/s00357-019-09353-1

Modified Subspace Constrained Mean Shift Algorithm

Youness Aliyari Ghassabeh[1] · **Frank Rudzicz**[1,2,3,4]

Published online: 11 February 2020

Abstract
A subspace constrained mean shift (SCMS) algorithm is a non-parametric iterative technique to estimate principal curves. Principal curves, as a nonlinear generalization of principal components analysis (PCA), are smooth curves (or surfaces) that pass through the middle of a data set and provide a compact low-dimensional representation of data. The SCMS algorithm combines the mean shift (MS) algorithm with a projection step to estimate principal curves and surfaces. The MS algorithm is a simple iterative method for locating modes of an unknown probability density function (pdf) obtained via a kernel density estimate. Modes of a pdf can be interpreted as zero-dimensional principal curves. These modes also can be used for clustering the input data. The SCMS algorithm generalizes the MS algorithm to estimate higher order principal curves and surfaces. Although both algorithms have been widely used in many real-world applications, their convergence for widely used kernels (e.g., Gaussian kernel) has not been sown yet. In this paper, we first introduce a modified version of the MS algorithm and then combine it with different variations of the SCMS algorithm to estimate the underlying low-dimensional principal curve, embedded in a high-dimensional space. The different variations of the SCMS algorithm are obtained via modification of the projection step in the original SCMS algorithm. We show that the modification of the MS algorithm guarantees its convergence and also implies the convergence of different variations of the SCMS algorithm. The performance and effectiveness of the proposed modified versions to successfully estimate an underlying principal curve was shown through simulations using the synthetic data.

Keywords Mean shift algorithm · Subspace constrained mean shift algorithm · Convergent sequence · Principal curves · Principal surfaces · Clustering

✉ Youness Aliyari Ghassabeh
aliyari@cs.toronto.edu

Frank Rudzicz
frank@cs.toronto.edu

[1] Department of Computer Science, University of Toronto, Toronto, Canada

[2] Li Ka Shing Knowledge Institute, St Michael's Hospital, Toronto, Canada

[3] Surgical Safety Technologies Inc, Toronto, Canada

[4] Vector Institute for Artificial Intelligence, Toronto, Canada

1 Introduction

Principal curves and surfaces are nonlinear generalization of principal component analysis (PCA) (Jolliffe 2002). They provide a new low-dimensional representation of the input data by mapping the high-dimensional observations onto a low-dimensional manifold, embedded in the high-dimensional space. The new low-dimensional representation facilitates tasks such as dimensionality reduction and data visualization. The first formal definition of principal curves was given by Hastie and Stuetzle (1989). According to their definition, a principal curve is a smooth, self-consistent, parameterized curve that passes through the middle of data set to provide a nonlinear summary of the data. Since Hastie and Stuetzle's groundbreaking work, several other definitions of principal curves and algorithms to construct them have been proposed based on, or inspired by, the original definition (see Banfield and Raftery (1992), Chang and Gosh (2001), Delicado (2001), Biau and Fischer (2012), and Tibshirani (1992) and Kegl et al. (2000), among others).

The modes of a probability density function (pdf) play an important role in many machine learning applications, such as image segmentation (Tao et al. 2007), object tracking in video sequences (Comaniciu et al. 2000), and clustering (Yuan et al. 2012). The collection of these modes can be viewed as a zero-dimensional principal curve. The mean shift (MS) algorithm is a simple non-parametric technique that iteratively tries to find modes of a pdf (estimated from data samples). The mean shift (MS) algorithm is an iterative, non-parametric technique that was introduced by Fukunaga and Hostetler (1975) to estimate modes of a pdf and it was generalized by Cheng (1995). The MS algorithm became popular in the machine learning community when its applications for clustering and image segmentation were revealed by Comaniciu and Meer (2002). Although the MS algorithm has been successfully used in many machine learning applications ranging from clustering to object tracking, a rigorous proof for its convergence is still missing in the literature (Ghassabeh 2016). The authors in Comaniciu and Meer (2002) incorrectly claimed that the sequence generated by the MS algorithm is a convergent sequence. Later it was shown that a crucial step in the their convergence proof was not correct (Li et al. 2007; Ghassabeh 2015). Later, the convergence of the generated MS sequence was claimed by showing that the MS algorithm with the Gaussian kernel is an instance of the expectation maximization (EM) algorithm (Carreira-Perpiñán 2007). However, without additional conditions, the generated sequence of parameter estimates by the EM algorithm may not necessarily converge (Wu 1982). On the positive side, the authors in Li et al. (2007) and Ghassabeh (2015) showed that the sequence of the estimated modes (zero-dimensional principal curves) generated by the MS algorithm is a convergent sequence if an estimated pdf has a finite number of modes (or equivalently has isolated modes (Arias-Castro et al. 2016)). However, this convergence proof has not been generalized for widely used kernels (e.g., Gaussian) (Ghassabeh 2016).

Ozertem and Erdogmus (2011) introduced subspace constrained mean shift (SCMS) algorithm to estimate principal curves and surfaces. The SCMS algorithm is a generalization of the MS algorithm, which iteratively tries to find modes of a pdf (estimated from data samples) in a local subspace. In other words, in contrast to the MS algorithm that finds zero-dimensional principal curves as the modes of a pdf, the SCMS algorithm estimates a principal curve or surface (with dimensions higher that zero) by looking for modes of a pdf projected in a subspace. The SCMS algorithm has been successfully used in applications such as time-series denoising (Ozertem and Erdogmus 2011), vector quantization of noisy sources (Ghassabeh et al. 2012a), and dimensionality reduction of noisy data (Ghassabeh et al. 2012b). Although extensive simulation results demonstrated the power of the SCMS algorithm to estimate the underlying principal curve/surface, the convergence of

the sequence generated by the SCMS algorithm has not been proved yet. The authors in Ozertem and Erdogmus (2011) claimed the convergence of the SCMS algorithm based on the assumption that the MS algorithm always converges that, as mentioned above, has not been proven yet.

In this paper, we first present the modified version of the MS algorithm that guarantees the convergence of the generated sequence by the MS algorithm. The converged generated sequence by the modified MS algorithm can be considered as zero-dimensional principal curve. Then, we present the modified SCMS algorithm by combining the modified MS algorithm with the original SCMS algorithm and its three variations. The convergence of the modified MS algorithm implies the convergence of the modified SCMS algorithm for any initial starting point. We show that the estimated pdf values along the output sequence generated by the proposed modified SCMS algorithm are a convergent sequence. We also show that the difference between two consecutive members of the generated sequence by the modified SCMS algorithm converges to zero. The effectiveness of the proposed modified SCMS algorithm to estimate a principal curve is shown through simulations. In the next section, we briefly review the MS and SCMS algorithms. The modified MS and proposed modified SCMS algorithms are presented in Section 3. Simulation results to support the theoretical results and to show the effectiveness of the proposed algorithms for finding a principal curve are given in Section 4. Section 5 is devoted to the concluding remarks.

2 Mean Shift and Subspace Constrained Mean Shift Algorithms

In this section, we briefly review the MS and SCMS algorithms and show how they can be used to estimate principal curves and surfaces (zero-dimensional principal curves for the MS algorithm).

2.1 MS Algorithm

The MS algorithm starts from one of the data points, as the initial mode, and iteratively shifts it to a weighted average of neighboring points to find stationary point of the estimated pdf. The MS algorithm does not require any prior knowledge of the number of clusters and there is no assumption for the shape of the clusters. The MS algorithm behaves like an instance of the gradient ascent algorithm (Fashing and Tomasi 2005) with an adaptive step size. The iterations for each starting point continue until the norm of the difference between two consecutive mode estimates becomes less than some predefined threshold. The resulting estimated modes in this procedure can be taken as the cluster center (or zero-dimensional principal curves) (Comaniciu and Meer 2002). Furthermore, all data points associated with the same mode are considered members of the same cluster.

A D-variate kernel $K : \mathbb{R}^D \to \mathbb{R}$ is a non-negative real-valued function that satisfies the following conditions (Wand and Jones 1995)

$$\int_{\mathbb{R}^D} K(\boldsymbol{x})d\boldsymbol{x} = 1, \quad \lim_{\|\boldsymbol{x}\|\to\infty} \|\boldsymbol{x}\|^D K(\boldsymbol{x}) = 0,$$
$$\int_{\mathbb{R}^D} \boldsymbol{x} K(\boldsymbol{x})d\boldsymbol{x} = 0, \quad \int_{\mathbb{R}^D} \boldsymbol{x}\boldsymbol{x}^T K(\boldsymbol{x})d\boldsymbol{x} = c_K \boldsymbol{I},$$

where c_K is a constant and $\boldsymbol{I}$ is the identity matrix. Let $\boldsymbol{x}_i \in \mathbb{R}^D, i = 1, \ldots, n$ be a sequence of n independent and identically distributed (iid) random variables.

The kernel density estimate $\hat{f}$ at an arbitrary point $\boldsymbol{x}$ using a kernel $K(\boldsymbol{x})$ is given by

$$\hat{f}(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n} K_{\mathbf{H}}(\boldsymbol{x} - \boldsymbol{x}_i), \tag{1}$$

where $K_{\mathbf{H}}(\boldsymbol{x}) = |\mathbf{H}|^{-1/2} K(\mathbf{H}^{-1/2}\boldsymbol{x})$, $\mathbf{H}$ is a symmetric positive definite $D \times D$ matrix called the bandwidth matrix, and $|\mathbf{H}|$ denotes the determinant of $\mathbf{H}$. A special class of kernels, called radially symmetric kernels, has been widely used for pdf estimation. Radially symmetric kernels are defined by $K(\boldsymbol{x}) = c_{k,D} k(\|\boldsymbol{x}\|^2)$, where $c_{k,D}$ is a normalization factor that causes $K(\boldsymbol{x})$ to integrate to one and $k : [0, \infty) \to [0, \infty)$ is called the profile of the kernel. The profile of a kernel is assumed to be a non-negative, non-increasing, and piecewise continuous function that satisfies $\int_0^\infty k(x)dx < \infty$. Symmetric kernels are defined by $K(\boldsymbol{x}) = c_{k,D} k(\|\boldsymbol{x}\|^2)$, where $c_{k,D}$ is a normalization factor that causes $K(\boldsymbol{x})$ to integrate to one and $k : [0, \infty) \to [0, \infty)$ is called the *profile* of the kernel. The profile of a kernel is assumed to be a non-increasing, non-negative, and piecewise continuous function that satisfies $\int_0^\infty k(x)dx < \infty$. Using the profile k, and the bandwidth h, the kernel density $\hat{f}(\boldsymbol{x})$ in Eq. 1 changes to the following well-known form (Silverman 1986)

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

Assuming that the profile k is differentiable with derivative k', by taking the gradient of Eq. 2, we obtain

$$\begin{aligned}\nabla \hat{f}_{h,k}(\boldsymbol{x}) = \frac{2c_{k,D}}{nh^{D+2}}&\left[\sum_{i=1}^{n} g\left(\left\|\frac{\boldsymbol{x} - \boldsymbol{x}_i}{h}\right\|^2\right)\right] \\ &\times\left[\frac{\sum_{i=1}^{n} \boldsymbol{x}_i g\left(\left\|\frac{\boldsymbol{x}-\boldsymbol{x}_i}{h}\right\|^2\right)}{\sum_{i=1}^{n} g\left(\left\|\frac{\boldsymbol{x}-\boldsymbol{x}_i}{h}\right\|^2\right)} - \boldsymbol{x}\right],\end{aligned} \tag{3}$$

where $g(x) = -k'(x)$. The second term in the above representation is called the mean shift (MS) vector $\boldsymbol{m}_{h,g}(\boldsymbol{x})$ (Comaniciu and Meer 2002).

From Eq. 3 and by equating the gradient function to zero, it can be observed that the modes of the estimated pdf are the fixed points of the following function

$$\boldsymbol{m}_{h,g}(\boldsymbol{x}) + \boldsymbol{x} = \frac{\sum_{i=1}^{n} \boldsymbol{x}_i g\left(\left\|\frac{\boldsymbol{x}-\boldsymbol{x}_i}{h}\right\|^2\right)}{\sum_{i=1}^{n} g\left(\left\|\frac{\boldsymbol{x}-\boldsymbol{x}_i}{h}\right\|^2\right)}, \tag{4}$$

where $\boldsymbol{m}_{h,g}(\boldsymbol{x})$ is the MS vector defined in Eq. 3. To solve Eq. 4, the MS algorithm starts from one of the data points and the mode estimate $\boldsymbol{y}_j$ in the jth iteration is updated by

$$\boldsymbol{y}_{j+1} = \boldsymbol{y}_j + \boldsymbol{m}(\boldsymbol{y}_j) = \frac{\sum_{i=1}^{n} \boldsymbol{x}_i g\left(\left\|\frac{\boldsymbol{y}_j-\boldsymbol{x}_i}{h}\right\|^2\right)}{\sum_{i=1}^{n} g\left(\left\|\frac{\boldsymbol{y}_j-\boldsymbol{x}_i}{h}\right\|^2\right)}. \tag{5}$$

The MS algorithm repeatedly updates the mode estimate, $\boldsymbol{y}_j$, using Eq. 5 until the norm of the difference between two consecutive estimates becomes less than some predefined

threshold, i.e., $\|\boldsymbol{y}_{j+1} - \boldsymbol{y}_j\| < \epsilon$ for some $j \in \mathbb{N}$. The final points when the algorithm stops are modes of the pdf, which also can be considered as zero-dimensional principal curves. The MS algorithm is summarized in **Algorithm 1**. Typically all n instances of the MS algorithm are run in parallel, with the ith instance initialized to the ith input data point.

Algorithm 1 MS algorithm for mode estimation and clustering

Input : Bandwidth h, profile function $g(x)$, threshold ϵ, and data set $\mathcal{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}, n \geq 2$.

Output: Estimated modes of the pdf, $\boldsymbol{y}_1^*, \ldots, \boldsymbol{y}_k^*$, where k is the number of modes. The estimated modes are zero-dimensional principal curves that can be interpreted as centers of clusters.

begin

  for *(i=1 to n)* **do**

    Initialization: $j \leftarrow 1$; $\boldsymbol{y}_j \leftarrow \boldsymbol{x}_i$;

    `/* Initialize the mode estimate sequence` $\boldsymbol{y}_1$ `to be one of the observed data points. */`

    repeat

      `/* Evaluate the mean shift vector using (??) */`

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

      `/* Update the mode estimate using (5) */`

      $\boldsymbol{y}_{j+1} = \boldsymbol{m}(\boldsymbol{y}_j) + \boldsymbol{y}_j$;

    until *Until* $\|\boldsymbol{y}_{j+1} - \boldsymbol{y}_j\| < \epsilon$ *for some predefined threshold* ϵ;

    $\boldsymbol{y}_i^* \leftarrow \boldsymbol{y}_j$;

  Merge the estimated cluster centers that are closer than h ;

  Eliminate clusters that attract small number of data points;

2.2 SCMS Algorithm

Ozertem and Erogmus defined a d-dimensional principal surface in $\mathbb{R}^D$[1] as the set of points that are local maximum of a pdf in a local orthogonal $D-d$-dimensional subspace (Ozertem and Erdogmus 2011). They proposed the SCMS algorithm to find points that satisfy that definition. The SCMS algorithm generalizes the MS algorithm to estimate higher order principal curves and surfaces ($d \geq 1$). Similar to the MS algorithm, the SCMS algorithm starts from one of the input points, as the initial estimate, it evaluates the MS vector in each iteration. In order to estimate the modes on the d-dimensional subspace, it projects the calculated MS vector in the previous step to the subspace spanned by the $D-d$ eigenvectors corresponding to the $D-d$ largest eigenvalues of the local inverse covariance matrix of the estimated pdf at that point. The local inverse covariance matrix at an arbitrary point $\boldsymbol{x}$ is estimated by Ozertem and Erdogmus (2011)

$$\hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{x}) = -\hat{\boldsymbol{H}}(\boldsymbol{x})\hat{f}(\boldsymbol{x})^{-1} + \nabla\hat{f}(\boldsymbol{x})\nabla\hat{f}(\boldsymbol{x})^T\hat{f}(\boldsymbol{x})^{-2}, \tag{6}$$

[1]It is called a principal curve for $D = 1$, and becomes a mode of the pdf for $D = 0$

where $\hat{\boldsymbol{H}}(\boldsymbol{x})$ and $\nabla \hat{f}(\boldsymbol{x})$ are the Hessian and gradient of the pdf estimate at $\boldsymbol{x}$.[2] The steps of the SCMS algorithm to estimate a d-dimensional principal curve/surface are summarized in **Algorithm 2** (Ghassabeh et al. 2013). Typically, n instances of the SCMS algorithm can run in parallel, each time initialized to one of the input data points. The resulting n output points are considered as a discrete approximation of the underlying principal curve (for $d = 1$) or surface (for $d \geq 2$). Ghassabeh et al. (2013) presented three new variations of the SCMS algorithm by replacing local inverse covariance matrix, $\hat{\boldsymbol{\Sigma}}^{-1}$, by the Hessian matrix $\hat{\boldsymbol{H}}$, and two local estimates (local to $\boldsymbol{y}_j$) of the covariance matrix of $\hat{f}$ as follows:

(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*,

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

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*,

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

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.

The resulting three variations of the original SCMS algorithm calculate the mean shift vectors using Eq. 5. But for the projection step instead of the local inverse covariance matrix in Eq. 6, three different matrices are used and the mean shift vector is projected into subspace spanned by the eigenvectors of the above matrices. In other words, for each matrix above the projection matrix $\boldsymbol{V}_j$ in 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 (Ghassabeh et al. 2013). The projection step and termination criterion remain the same as in the original SCMS algorithm.

[2]Note that for the special case of Gaussian distribution $\hat{f} \sim N(\mu, \boldsymbol{\Sigma})$, the local inverse covariance matrix in Eq. 6 becomes equal to the inverse covariance matrix, i.e., $\hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{x}) = \boldsymbol{\Sigma}^{-1}$.

Algorithm 2 The SCMS algorithm to find a principal curve/surface

Input : Bandwidth h, profile function $g(x)$, threshold ϵ, $d \geq 1$ dimension of the principal curve/surface, and data set $\mathcal{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$, $n \geq 2$.

Output: Estimated points $\boldsymbol{y}_1^*, \ldots, \boldsymbol{y}_n^*$, on the d-dimensional principal curve or surface.

begin

  for *(i=1 to n)* **do**

    Initialization: $j \leftarrow 0$; $\boldsymbol{y}_j \leftarrow \boldsymbol{x}_i$;

    /* The SCMS algorithm is initialized by one the observed data points. */

    repeat

      /* Evaluate the mean shift vector using (??) */

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

      /* Evaluate the gradient, the Hessian matrix, and the local inverse covariance matrix $\hat{\boldsymbol{\Sigma}}^{-1}$ using (6) at $\boldsymbol{y}_j$ */

      $\hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{y}_j) = -\hat{\boldsymbol{H}}(\boldsymbol{y}_j)\hat{f}(\boldsymbol{y}_j)^{-1} + \nabla\hat{f}(\boldsymbol{y}_j)\nabla\hat{f}(\boldsymbol{y}_j)^T\hat{f}(\boldsymbol{y}_j)^{-2}$;

      /* 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 local inverse covariance matrix $\hat{\boldsymbol{\Sigma}}_j^{-1}$. */

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

      $j = j + 1$;

    until $\|\boldsymbol{y}_{j+1} - \boldsymbol{y}_j\| < \epsilon$, *for some* $j \in \mathbb{N}$;

    $\boldsymbol{y}_i^* \leftarrow \boldsymbol{y}_j$;

3 Modified MS and SCMS Algorithms

From Eq. 5, it can be observed that the mode estimate sequence, $\{\boldsymbol{y}_j\}$, generated by the MS algorithm is always inside the convex hull of the data set. Therefore, $\{\boldsymbol{y}_j\}$ is a bounded sequence and it can be shown that it satisfies (Ghassabeh 2016)

$$\lim_{k\to\infty} \|\boldsymbol{y}_{j+1} - \boldsymbol{y}_j\| = 0. \tag{7}$$

Note that the above two properties are not enough to imply the convergence of the sequence $\{\boldsymbol{y}_j\}_{j=1,2,\ldots}$ generated by the MS algorithm (Ghassabeh 2016; Li et al. 2007). We slightly modified the MS algorithm to guarantee the convergence of the generated sequence. Similar to the original version, the proposed modified MS algorithm starts from one of the input data points and computes the mean shift vector using Eq. 4. Then, it updates the mode estimate by assigning the computed MS vector to the closest input data point. In other words, the

proposed algorithm updates the mode estimate in each iteration by assigning it to one of the data points as follows (Ghassabeh and Rudzicz 2018)

$$\tilde{\boldsymbol{y}}_{j+1} = \boldsymbol{y}_j + \boldsymbol{m}(\boldsymbol{y}_j), \tag{8}$$

$$\boldsymbol{y}_{j+1} = \underset{\boldsymbol{x}\in\{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n\}}{\operatorname{argmin}} \|\tilde{\boldsymbol{y}}_{j+1} - \boldsymbol{x}\|, \tag{9}$$

where $\boldsymbol{m}(\boldsymbol{y}_j)$ is the MS vector in Eq. 4, and $\{\boldsymbol{x}_1 \ldots, \boldsymbol{x}_n\}$ is the set of input data. The proposed modified MS algorithm is summarized in **Algorithm 3**.

Algorithm 3 Modified MS algorithm for clustering

Input : Bandwidth h, profile function $g(x)$, and data set $\mathcal{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$, $n \geq 2$.
Output: Estimated modes of the pdf, $\boldsymbol{y}_1^*, \ldots, \boldsymbol{y}_k^*$, where k is the number of modes .

begin
  for *(i=1 to n)* **do**
    Initialization: $j \leftarrow 1$; $\boldsymbol{y}_j \leftarrow \boldsymbol{x}_i$;
    /* Initialize the mode estimate sequence $\boldsymbol{y}_1$ to be one of the observed data points. */
    repeat
      /* Evaluate the mean shift vector using (??) */
      $\boldsymbol{m}(\boldsymbol{y}_j) = \frac{\sum_{i=1}^n \boldsymbol{x}_i g\left(\|\frac{\boldsymbol{y}_j - \boldsymbol{x}_i}{h}\|^2\right)}{\sum_{i=1}^n g\left(\|\frac{\boldsymbol{y}_j - \boldsymbol{x}_i}{h}\|^2\right)} - \boldsymbol{y}_j$;
      /* Update the mode estimate using (5) */
      $\tilde{\boldsymbol{y}}_{j+1} = \boldsymbol{m}(\boldsymbol{y}_j) + \boldsymbol{y}_j$;
      /* Find the closest data point to $\tilde{y}_{j+1}$ */
      $\boldsymbol{y}_{j+1} = \arg\min_{\boldsymbol{x}\in\{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n\}} \|\tilde{\boldsymbol{y}}_{j+1} - \boldsymbol{x}\|$;
    until *Until convergence occurs, i.e.,* $\boldsymbol{y}_{j+1} = \boldsymbol{y}_j$ *for some* $j \in \mathbb{N}$;
    $\boldsymbol{y}_i^* \leftarrow \boldsymbol{y}_j$;
  Merge the estimated modes that are closer than h as zero-dimensional principal curve or center of clusters for a clustring task;
  /* For the clustring task, eliminate clusters that attract small number of data points. */

The sequence generated by the modified MS algorithm, $\{\boldsymbol{y}_j\}_{j=1,2,\ldots}$, is a convergent sequence. Also, similar to the original MS algorithm, the estimated pdf values along the generated sequence, $\{\hat{f}(\boldsymbol{y})_j\}_{j=1,2,\ldots}$, are an increasing and convergent sequence (Ghassabeh and Rudzicz 2018). The next theorem states two convergence results relating to the density estimate values produced by the modified MS algorithm. The proof is given in the Appendix.

Theorem 1 *Assume a kernel pdf estimate $\hat{f}$ with bandwidth h, and a radially symmetric kernel K having profile k which is positive, strictly decreasing, convex, and continuously differentiable on $\mathbb{R}$ (as is defined in (2)). Let $\{\boldsymbol{y}_j\}$ denote the sequence of points generated by the modified MS algorithm with arbitrary initialization. Then, the following holds:*

(i) *The density estimate values along the sequence of output values of the modified MS algorithm is a monotonically increasing and convergent sequence, i.e., $\{\hat{f}_{h,k}(\boldsymbol{y}_j)\}_{j=1,2,\ldots}$ is monotonically increasing and convergent.*

(ii) *The mode estimate sequence,* $\{\boldsymbol{y}_j\}_{j=1,2,\ldots,}$ *generated by the modified MS algorithm is a convergent sequence.*

In the original MS algorithm, the stopping threshold ϵ is set manually so that a good tradeoff between running time and approximation accuracy is achieved. However, the convergence is guaranteed in the proposed version, and there is no need to set a predefined threshold as the stopping criterion for the algorithm. It is clear that setting a threshold in general cannot be used as a reliable measure of closeness to the convergence point (if there is a convergence point). It is possible that the difference between two consecutive mode estimates becomes less than the predefined threshold and the algorithm terminates the iterations, but both points are far from the possible convergence point.

The authors in Ozertem and Erdogmus (2011) claimed the convergence of the SCMS algorithm based on the assumption that the MS algorithm always converges, which, as we discussed, has so far been unproven. We propose the modified SCMS algorithm to estimate principal curves and surface by replacing the mean shift vector calculation in the original version by the modified MS vector calculated through the modified MS algorithm. To put it succinctly, the proposed modified SCMS algorithm differs from the original step only at the way it updates the mean shift vector in each iteration, i.e., it uses the modified MS algorithm. In this way, the convergence of the proposed algorithm is guaranteed for any initial point, since from Theorem 1 it can be observed that the modified MS step generates is a convergent sequence. The next theorem summarizes three convergence results relating to the density estimate values produced by the proposed modified SCMS algorithm. The proof is given in the Appendix.

Theorem 2 *Assume a kernel pdf estimate* $\hat{f}$ *with bandwidth h, and a radially symmetric kernel K having profile k, which is positive, strictly decreasing, convex, and continuously differentiable on* $\mathbb{R}$ *(as is defined in (2)). Let* $\{\boldsymbol{y}_j\}$ *denote the sequence of points generated by the SCMS algorithm with arbitrary initialization. Then, the following holds:*

(i) *The sequence* $\{\hat{f}(\boldsymbol{y}_j)\}$ *is non-decreasing and convergent.*
(ii) $\lim_{j\to\infty} \|\boldsymbol{y}_{j+1} - \boldsymbol{y}_j\| = 0.$
(iii) $\lim_{j\to\infty} \|\boldsymbol{V}_j^T \nabla \hat{f}(\boldsymbol{y}_j)\| = 0.$

Note that the above theorem is analogous to what is proved in proposition 2 in Ghassabeh et al. (2013) for the original SCMS algorithm, just the MS vectors are replaced by the modified MS vectors to guarantee the convergence of the whole process. Although the local inverse covariance matrix $\hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{y}_j)$ is used during the proof of theorem 2, all three statements remain valid if the projection matrix $\boldsymbol{V}_j$, $j = 1, 2, \ldots$, 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 necessarily the matrix whose columns are the $D - d$ orthonormal eigenvectors corresponding to the largest eigenvalues of the local inverse covariance matrix $\hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{y}_j)$. As a result, the three matrices introduced at the end of Section 2 can be used to create the projection matrix $\boldsymbol{V}_j$, $j = 1, 2, \ldots$ as well. Therefore, the proposed modified SCMS algorithm can be presented in four variations that differ only at the projection matrix. The four projection matrices can be derived using any of the following four matrices:

(i) The local inverse covariance matrix,
(ii) The Hessian of $\hat{f}$,

(iii) The estimated local covariance matrix using the κ nearest *data points*,
(iv) The estimated local covariance matrix using the κ nearest *outputs*.

The proposed modified SCMS algorithm is summarized in **Algorithm 4**. Use of two local estimates of the covariance matrix, introduced at Section 2, may reduce the computational cost. 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 generally depend on the structure of the underlying manifold. In our simulations, we chose κ to be between 4 and 6% of the number of observations, but setting κ in general is beyond the scope of this paper.

Algorithm 4 The proposed modified SCMS algorithm to find a principal curve/surface

Input : Bandwidth h, profile function $g(x)$, threshold ϵ, $d \geq 1$ dimension of the principal curve/surface, and data set $\mathcal{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$, $n \geq 2$.
Output: Estimated points $\boldsymbol{y}_1^*, \ldots, \boldsymbol{y}_n^*$, on the d-dimensional principal curve or surface.

begin
  for *(i=1 to n)* **do**
    Initialization: $j \leftarrow 0$; $\boldsymbol{y}_j \leftarrow \boldsymbol{x}_i$;
    /* The SCMS algorithm is initialized by one the observed data points. */
    repeat
      /* Evaluate the mean shift vector using (4) */
      $\boldsymbol{m}(\boldsymbol{y}_j) = \frac{\sum_{i=1}^n \boldsymbol{x}_i g\left(\| \frac{\boldsymbol{y}_j - \boldsymbol{x}_i}{h} \|^2\right)}{\sum_{i=1}^n g\left(\| \frac{\boldsymbol{y}_j - \boldsymbol{x}_i}{h} \|^2\right)} - \boldsymbol{y}_j$;
      /* Evaluate the gradient, the Hessian matrix, and the local inverse covariance matrix $\hat{\boldsymbol{\Sigma}}^{-1}$ using (6) at $\boldsymbol{y}_j$ */
      $\hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{y}_j) = -\hat{\boldsymbol{H}}(\boldsymbol{y}_j)\hat{f}(\boldsymbol{y}_j)^{-1} + \nabla\hat{f}(\boldsymbol{y}_j)\nabla\hat{f}(\boldsymbol{y}_j)^T\hat{f}(\boldsymbol{y}_j)^{-2}$;
      /* note that $\hat{\boldsymbol{\Sigma}}^{-1}$ can be replaced by any of the alternative three matrices introduced at the end of Section 2. */
      /* 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 local inverse covariance matrix $\hat{\boldsymbol{\Sigma}}_j^{-1}$. */
      $\tilde{\boldsymbol{y}}_{j+1} = \boldsymbol{V}_j\boldsymbol{V}_j^T\boldsymbol{m}(\boldsymbol{y}_j) + \boldsymbol{y}_j$;
      /* Find the closest data point in $\mathcal{X}$ to $\tilde{y}_{j+1}$ */
      $\boldsymbol{y}_{j+1} = \arg\min_{\boldsymbol{x}\in\{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n\}} \|\tilde{\boldsymbol{y}}_{j+1} - \boldsymbol{x}\|$;
    until *Until convergence occurs, i.e.,* $\boldsymbol{y}_{j+1} = \boldsymbol{y}_j$ *for some* $j \in \mathbb{N}$;
  $\boldsymbol{y}_i^* \leftarrow \boldsymbol{y}_j$;

Table 1 Performance results of four variations of the proposed modified SCMS algorithm for estimating the principal curve on a noisy circle

2-D circle	SCMS	Hessian	Cov. 1	Cov. 2
Running time (s)	32.704	34.570	13.593	12.517
Av. squared Euclidean distance	0.814	0.812	0.786	0.779

4 Simulation Results

In this section, we demonstrate the effectiveness of the proposed modified SCMS algorithm for finding a principal curve. Simulations using synthetic data are provided to support the theoretical results in Section 3. The input data set $\mathcal{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{500}\}$ has the following form $\boldsymbol{x}_i = \boldsymbol{u}_i + \boldsymbol{e}_i, i = 1, \ldots, 500$, where $\boldsymbol{u}_i \in \mathbb{R}^2, i = 1, \ldots, 500$ are uniformly selected data points on a circle with unit radius (they can be interpreted as clean unobserved data), $\boldsymbol{e}_i, i = 1, \ldots, 500$ is bivariate zero-mean Gaussian noise with identity covariance matrix times 0.45. The bandwidth h is set to 0.4, dimension of the principal curve $d = 1$, $\epsilon = 0.01$ is used as the stopping criteria, and the number of nearest neighbors for last two variations the proposed SCMA algorithm is set to $\kappa = 40$. The average squared Euclidean distance between the output points and the closest point on the generative circle is used to measure the performance of each variations of the proposed modified SCMS algorithm, and the average running time is measured in seconds. Data points in $\mathcal{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{500}\}$ are fed into each algorithms and each variation is initialized with input data. Table 1 summarizes the performance of four variations of the proposed modified SCMA algorithm to estimate the principal curve on a two-dimensional noisy circle. The different variations of the proposed algorithm use different projection matrices as follows, SCMS (local inverse

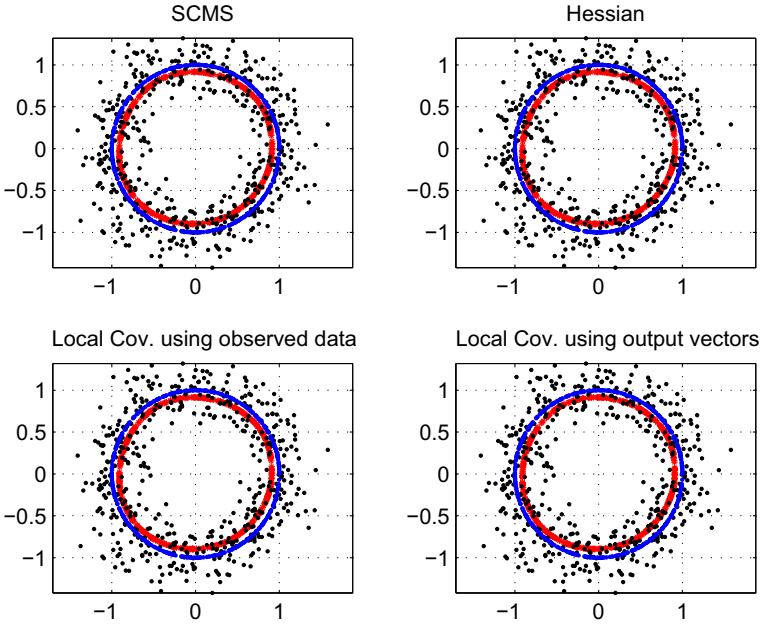

Fig. 1 First $n = 500$ samples, denoted by blue, are uniformly selected on the circle, the black points are the observed input data points (inputs to the proposed algorithm) generated by adding independent, zero-mean Gaussian noise to the blue points on the generative curve, and the red points are the outputs of different variations of the proposed SCMA algorithm

covariance matrix), the Hessian matrix, 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). It can be observed from Table 1 that performance of all four proposed variation is similar in terms of closeness to the generative original circle. In terms of runtime, as it is expected, the local covariance matrices (using original data points or using the output points) perform significantly better than two other cases.

The generative curve, the simulated data points, and the generated output points from the four versions of the proposed modified algorithm are shown in Fig. 1. From Fig. 1, it can be observed that all four versions of the proposed modified version show similar performance visually and all four are able to successfully estimate the underlying one-dimensional principal curve.

5 Conclusion

The MS and SCMS algorithms have been widely used in many machine learning applications, but the convergence of the generated sequences has not been proved yet. The MS algorithm has been used to estimate modes of a pdf, which play an important role in applications such as clustering. The SCMS algorithm generalized the MS algorithm and was used to estimate a low-dimensional manifold embedded in a high-dimensional space. In this paper, we first present the modified MS algorithm and then combine it with the original SCMS algorithm and introduce the modified SCMS algorithm. The convergence properties of the generated sequences for both modified MS and modified SCMS algorithms are investigated and it was shown that the modified versions generate convergent sequences. Convergence property of the generated sequences implies that there is no need to set a stopping threshold for the proposed modified versions and the generated sequence converges after a finite number of iterations (when the number of samples n is finite). We also show that the estimated pdf along the generated sequence by the modified SCMS algorithm is a monotonically increasing and convergent sequence. By changing the projection matrix in the projection step, four different variations of the modified SCMS algorithm are presented. Finally, simulations using synthetic data are provided to support the theoretical results. Through simulations, we show that four variations of the proposed modified SCMS algorithm can be successfully used to estimate the underlying principal curve.

Appendix

Proof of Theorem 1 Let $\mathcal{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ denote the input data set. Let $\boldsymbol{y}_{j+1} \neq \boldsymbol{y}_j$. To prove part (i), we show $\hat{f}_{h,k}(\boldsymbol{y}_{j+1}) > \hat{f}(\boldsymbol{y}_j)$. From Eq. 2, we have

$$\begin{aligned}
&\hat{f}_{h,k}(\boldsymbol{y}_{j+1}) - \hat{f}_{h,k}(\boldsymbol{y}_j) \\
&= \frac{c_{k,D}}{nh^D}\left[\sum_{i=1}^{n} k\left(\left\|\frac{\boldsymbol{y}_{j+1}-\boldsymbol{x}_i}{h}\right\|^2\right) - k\left(\left\|\frac{\boldsymbol{y}_j-\boldsymbol{x}_i}{h}\right\|^2\right)\right] \\
&\geq \frac{c_{k,D}}{nh^{D+2}}\sum_{i=1}^{n} k'\left(\left\|\frac{\boldsymbol{y}_j-\boldsymbol{x}_i\|^2}{h}\right\|\right) \\
&\quad \left(\|\boldsymbol{y}_{j+1}-\boldsymbol{x}_i\|^2 - \|\boldsymbol{y}_j-\boldsymbol{x}_i\|^2\right), \qquad (10)
\end{aligned}$$

where the last inequality comes from the convexity of the profile function k, i.e., $k(x_2) - k(x1) \geq k'(x_1)(x_2 - x_1)$. By the triangle inequality, we have

$$\|\boldsymbol{y}_{j+1} - \tilde{\boldsymbol{y}}_{j+1}\| \leq \|\boldsymbol{y}_{j+1} - \boldsymbol{x}_i\| + \|\tilde{\boldsymbol{y}}_{j+1} - \boldsymbol{x}_i\|, i = 1, 2, \ldots, n, \tag{11}$$

where $\tilde{\boldsymbol{y}}_{j+1}$ is given in Eq. 8. Using Eqs. 10 and 11, we obtain

$$\begin{aligned}\hat{f}_{h,k}(\boldsymbol{y}_{j+1}) - \hat{f}_{h,k}(\boldsymbol{y}_j) \geq \frac{c_{k,D}}{nh^{D+2}} \sum_{i=1}^{n} k'\left(\|\frac{\boldsymbol{y}_j - \boldsymbol{x}_i}{h}\|^2\right) \\ \left(\|\boldsymbol{y}_{j+1} - \tilde{\boldsymbol{y}}_{j+1}\|^2 - \|\tilde{\boldsymbol{y}}_{j+1} - \boldsymbol{x}_i\|^2 \right. \\ \left. -2\|\boldsymbol{y}_{j+1} - \tilde{\boldsymbol{y}}_{j+1}\|\|\tilde{\boldsymbol{y}}_{j+1} - \boldsymbol{x}_i\| - \|\boldsymbol{y}_j - \boldsymbol{x}_i\|^2\right).\end{aligned} \tag{12}$$

From Eq. 13, we have $\|\boldsymbol{y}_{j+1} - \tilde{\boldsymbol{y}}_{j+1}\|^2 - \|\tilde{\boldsymbol{y}}_{j+1} - \boldsymbol{x}_i\|^2 \leq 0$ for $\boldsymbol{x}_i \in \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$, and as a result we have

$$\begin{aligned}\sum_{i=1}^{n} k'\left(\|\frac{\boldsymbol{y}_j - \boldsymbol{x}_i}{h}\|^2\right) \\ \left(\|\boldsymbol{y}_{j+1} - \tilde{\boldsymbol{y}}_{j+1}\|^2 - \|\tilde{\boldsymbol{y}}_{j+1} - \boldsymbol{x}_i\|^2\right) > 0,\end{aligned} \tag{13}$$

where the above inequality is true since the profile k is a strictly decreasing function and $k'(x) < 0$. Furthermore, we have

$$\begin{aligned}\sum_{i=1}^{n} k'\left(\|\frac{\boldsymbol{y}_j - \boldsymbol{x}_i}{h}\|^2\right) \\ \left(-2\|\boldsymbol{y}_{j+1} - \tilde{\boldsymbol{y}}_{j+1}\|\|\tilde{\boldsymbol{y}}_{j+1} - \boldsymbol{x}_i\| - \|\boldsymbol{y}_j - \boldsymbol{x}_i\|^2\right) > 0.\end{aligned} \tag{14}$$

Combining Eqs. 12, 13, and 14, we obtain

$$\hat{f}_{h,k}(\boldsymbol{y}_{j+1}) - \hat{f}_{h,k}(\boldsymbol{y}_j) > 0, \tag{15}$$

which implies the sequence $\{\hat{f}_{h,k}(\boldsymbol{y}_j)\}_{j=1,2,\ldots}$ is an increasing sequence. From Eq. 13, it is obvious that $\boldsymbol{y}_j \in \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$, $j = 1, 2, \ldots$, and since n is finite then $\hat{f}_{h,k}(\boldsymbol{y}_j)$, given in Eq. 2, is bounded. Thus, as long as $\boldsymbol{y}_{j+1} \neq \boldsymbol{y}_j$, the sequence $\{\hat{f}_{h,k}(\boldsymbol{y}_j)\}_{j=1,2,\ldots}$ is a bounded and strictly increasing sequence, which two previous conditions imply the convergence of $\{\hat{f}_{h,k}(\boldsymbol{y}_j)\}$.

To prove part (ii), first note that the modified MS algorithm starts from one of the data points, and in each iteration the cluster center estimate is assigned to be one of the data points. The algorithm stops when two consecutive estimates become equal, i.e., $\boldsymbol{y}_{j+1} = \boldsymbol{y}_j$ for some $j \geq 1$. From part (a), in each iteration, each data point can be assigned to the cluster center estimate at most one time; otherwise, $\hat{f}_{h,k,}(\boldsymbol{y}_{j+k}) = \hat{f}_{h,k}(\boldsymbol{y}_j)$ for some $k \geq 1$ which contradicts part (a). Since the number of data samples, n, is finite, after a finite number of iterations, the convergence for the sequence $\{\boldsymbol{y}_j\}$ occurs. □

Proof of Theorem 2 Let $\mathcal{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$, $n \geq 2$ denote the set of observed data points. The subspace constrained mean shift sequence $\{\boldsymbol{y}_j\}$ is defined recursively by

$$\boldsymbol{y}_{j+1} = \boldsymbol{V}_j \boldsymbol{V}_j^T \boldsymbol{m}(\boldsymbol{y}_j) + \boldsymbol{y}_j, \tag{16}$$

where

$$\boldsymbol{m}(\boldsymbol{y}_j) = \frac{\sum_{i=1}^{n} \boldsymbol{x}_i g\left(\left\|\frac{\boldsymbol{y}_j - \boldsymbol{x}_i}{h}\right\|^2\right)}{\sum_{i=1}^{n} g\left(\left\|\frac{\boldsymbol{y}_j - \boldsymbol{x}_i}{h}\right\|^2\right)} - \boldsymbol{y}_j, \tag{17}$$

with $\boldsymbol{y}_j \in \mathcal{X}$ being one of the input data points, as required by the modified MS algorithm. 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 that are eigenvectors corresponding to the largest eigenvalues of the local inverse covariance matrix $\hat{\boldsymbol{\Sigma}}^{-1}$, defined in Eq. 6, evaluated at $\boldsymbol{y}_j$ that is one of the input data point according to the modified MS algorithm.

Since the profile k is bounded, the sequence $\{\hat{f}(\boldsymbol{y}_j)\}$ is bounded, so it suffices to show that the sequence is non-decreasing to prove the convergence. Since it is assumed that k is a convex function, we have $k(t_2) - k(t_1) \geq g(t_1)(t_1 - t_2)$ for all $t_1, t_2 \geq 0$, where $g = -k'$. This combined by the definition of $\hat{f}$ in Eq. 2 yields

$$\begin{aligned}
\hat{f}(\boldsymbol{y}_{j+1}) - \hat{f}(\boldsymbol{y}_j) &= \frac{c}{nh^D} \sum_{i=1}^{n} \left(k\left(\left\|\frac{\boldsymbol{y}_{j+1} - \boldsymbol{x}_i}{h}\right\|^2\right) - k\left(\left\|\frac{\boldsymbol{y}_j - \boldsymbol{x}_i}{h}\right\|^2\right) \right) \\
&\geq \frac{c}{nh^{D+2}} \sum_{i=1}^{n} g\left(\left\|\frac{\boldsymbol{y}_j - \boldsymbol{x}_i}{h}\right\|^2\right) \left(\|\boldsymbol{y}_j - \boldsymbol{x}_i\|^2 - \|\boldsymbol{y}_{j+1} - \boldsymbol{x}_i\|^2 \right) \\
&= C_j \sum_{i=1}^{n} p_j(i) \left(\|\boldsymbol{y}_j - \boldsymbol{x}_i\|^2 - \|\boldsymbol{y}_{j+1} - \boldsymbol{x}_i\|^2 \right),
\end{aligned} \tag{18}$$

where c is the normalization factor,

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

and

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

Since by assumption k is strictly decreasing, then $g(t) = -k'(t) > 0$ for all $t \geq 0$, $p_j(1), \ldots, p_j(n)$ are well defined, positive, and sum to 1. Therefore, the mean shift vector at (26) can be rewritten as

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

where $\boldsymbol{Z}_j$ is a random vector in $\mathbb{R}^D$ with discrete probability distribution function given by $\Pr(\boldsymbol{Z}_j = \boldsymbol{x}_i - \boldsymbol{y}_j) = p_j(i)$, $i = 1, \ldots, n$, and $\boldsymbol{y}_j \in \mathcal{X}$, $j = 1, \ldots,$. Thus, letting $\boldsymbol{T}_j = \boldsymbol{V}_j \boldsymbol{V}_j^T$, the update step in the proposed modified SCMS algorithm can be rewritten as

$$\boldsymbol{y}_{j+1} - \boldsymbol{y}_j = \boldsymbol{T}_j \boldsymbol{m}(\boldsymbol{y}_j) = \boldsymbol{T}_j E[\boldsymbol{Z}_j]. \tag{19}$$

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 Eq. 18 as follows

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

where in the last equality, we applied the identity $E[Z^2] = \mathrm{Var}[Z] + (E[Z])^2$, which is valid for real random variables with finite variance, to the components of $\boldsymbol{T}_j\boldsymbol{Z}_j$. Combining this with Eq. 18, we obtain

$$\hat{f}(\boldsymbol{y}_{j+1}) - \hat{f}(\boldsymbol{y}_j) \geq C_j\|\boldsymbol{y}_{j+1} - \boldsymbol{y}_j\|^2, \tag{20}$$

where $C_j > 0$ and $\|\boldsymbol{y}_{j+1} - \boldsymbol{y}_j\|^2 \geq 0$[3] which imply that $\{\hat{f}(\boldsymbol{y}_j)\}$ is non-decreasing and thus convergent, proving part (i) of the theorem.

To prove part (ii), we note that $k(x) > 0$ for all $x \geq 0$. Therefore, (2) implies that $\hat{f}(\boldsymbol{y}_1) > 0$, $\boldsymbol{y}_1 \in \mathcal{X}$, 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 non-increasing on $[0, \infty)$, we obtain

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

where $C > 0$ since $g(x) > 0$ for all $x \geq 0$. Thus, Eq. 20 implies

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

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

[3]The equality $\|\boldsymbol{y}_{j+1} - \boldsymbol{y}_j\|^2 = 0$ happens only when the convergence occurs that Theorem 1 guarantees it for the modified MS algorithm.

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

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

Therefore,

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

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 Eq. 19 yield

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

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

References

Arias-Castro, E., Mason, D., Pelletier, B. (2016). Errata: on the estimation of the gradient lines of a density and the consistency of the mean-shift algorithm. *Journal of Machine Learning Research*, *17*, 14.

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

Biau, G., & Fischer, A. (2012). Parameter selection for principal curves. *IEEE Trans. on Information Theory*, *58*, 1924–1939.

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

Chang, K.Y., & Gosh, A. (2001). A unified model for probabilistic principal surfaces. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, *23*, 22–41.

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

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

Comaniciu, D., Ramesh, V., Meer, P. (2000). Real-time tracking of non-rigid objects using mean shift. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition (CVPR2000)* (pp. 142–149). USA: Hilton Head Island.

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

Fashing, M., & Tomasi, C. (2005). Mean shift is a bound optimization. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, *27*, 471–474.

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

Ghassabeh, Y.A. (2015). Asymptotic stability of equilibrium points of mean shift algorithm. *Machine Learning*, *98*, 359–368.

Ghassabeh, Y.A. (2016). A sufficient condition for the convergence of the mean shift algorithm with Gaussian kernel. *Journal of Multivariate Analysis*, *135*, 1–10.

Ghassabeh, Y.A., Linder, T., Takahara, G. (2012a). On noisy source vector quantization via a subspace constrained mean shift algorithm. In *Proceedings of the 26th Biennial Symp. on Communications* (pp. 107–110). Canada: Kingston.

Ghassabeh, Y.A., Linder, T., Takahara, G. (2012b). On the convergence and applications of mean shift type algorithms. In: Proceedings of 25th IEEE Canadian Conference on Electrical & Computer Engineering (CCECE). Montreal, Canada, pp. 1-5.

Ghassabeh, Y.A., Linder, T., Takahara, G. (2013). On some convergence properties of the subspace constrained mean shift. *Pattern Recognition*, *46*, 3140–3147.

Ghassabeh, Y.A., & Rudzicz, F. (2018). Modified mean shift algorithm. *IET Image Processing*, *12*, 2172–2177.

Hastie, T., & Stuetzle, W. (1989). Principal curves. *Journal of the American Statistical Association*, *84*, 502–516.

Jolliffe, I.T. (2002). *Principal component analysis*, 1st edn. New York: Springer-Verlag.

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

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

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

Silverman, B.W. (1986). *Density estimation for statistics and data analysis*, 1st edn. New York: Chapman and Hall.

Tao, W., Jin, H., Zhang, Y. (2007). Color image segmentation based on mean shift and normalized cuts. *IEEE Trans. on Systems, Man, and Cybernetics Part B: Cybernetics*, *37*, 1382–1389.

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

Wand, M.P., & Jones, M. (1995). *Kernel smoothing*. London: Chapman and Hall.

Wu, C.F.J. (1982). On the convergence properties of the eM algorithm. *The Annals of Statistics*, *11*, 95103.

Yuan, X.T., Hu, B.G., He, R. (2012). Agglomerative mean-shift clustering. *IEEE Trans. on Knowledge and Data Engineering*, *24*, 209–219.