Nonlinear kernel density principal component analysis with application to climate data

Seppo Pulkkinen

Received: 12 December 2013 / Accepted: 27 October 2014 / Published online: 9 December 2014

Abstract Principal component analysis (PCA) is a well-established tool for identifying the main sources of variation in multivariate data. However, as a linear method it cannot describe complex nonlinear structures. To overcome this limitation, a novel nonlinear generalization of PCA is developed in this paper. The method obtains the nonlinear principal components from ridges of the underlying density of the data. The density is estimated by using Gaussian kernels. Projection onto a ridge of such a density estimate is formulated as a solution to a differential equation, and a predictor-corrector method is developed for this purpose. The method is further extended to time series data by applying it to the phase space representation of the time series. This extension can be viewed as a nonlinear generalization of singular spectrum analysis (SSA). Ability of the nonlinear PCA to capture complex nonlinear shapes and its SSA-based extension to identify periodic patterns from time series are demonstrated on climate data.

Keywords Principal component analysis · Singular spectrum analysis · Exploratory data analysis · Kernel density · Ridge · Predictor-corrector method

S. Pulkkinen (✉)
Turku Centre for Computer Science (TUCS), University of Turku, 20014 Turku, Finland
e-mail: seppo.pulkkinen@utu.fi

S. Pulkkinen
Finnish Meteorological Institute,
P.O. Box 503, 00101 Helsinki, Finland

1 Introduction

In practical applications, one is often dealing with high-dimensional data that is confined to some low-dimensional subspace. Since its introduction by Pearson (1901), *principal component analysis* (PCA, e.g. Jolliffe 1986) has become a ubiquitous tool for identifying such subspaces. The method uses an orthogonal transformation to separate the directions of maximal variance. PCA and its variants have appeared in various contexts such as *empirical orthogonal functions* (EOF) in climate analysis (e.g. Weare et al. 1976), *proper orthogonal decomposition* (POD) in fluid mechanics (e.g. Berkooz et al. 1993) and the *Karhunen-Loève transform* (KLT) in the theory of stochastic processes (e.g. Loève 1955).

However, as a linear method, PCA is insufficient for describing complex nonlinear data. Several nonlinear extensions have been developed to overcome this limitation. The most prominent of these are the neural network-based nonlinear PCA (NLPCA, e.g. Hsieh 2004; Kramer 1991; Monahan 2001; Scholz et al. 2005, 2008) and kernel PCA (KPCA, e.g. Schölkopf et al. 1997). These methods, however, have shortcomings. NLPCA requires a large number of user-supplied parameters that need to be carefully tuned for the application at hand. Furthermore, the transformation of the input data into the high-dimensional kernel space in KPCA incurs a significant computational cost. A careful choice of kernel function is also needed when using KPCA.

Some variants of PCA, where the principal components are obtained by restricting the analysis to local neighbourhoods of the data points, have been developed (e.g. Kambhatla and Leen 1997; Einbeck et al. 2005, 2008). However, this approach leads to the problem of determining a global coordinate system. A well-known approach to this problem is local tangent space alignment (LTSA, Zhang and Zha 2004) that determines a coordinate system by solving an eigenvalue problem constructed from the local principal component coordinates. However, this method and other neighbourhood-based methods are in general sensitive to noise and the choice of the neighbourhoods.

The contribution of this paper is the development of *kernel density principal component analysis* (KDPCA). The proposed method builds on the idea of using *ridges* of the underlying density of the data to estimate nonlinear structures (Ozertem and Erdogmus 2011). This idea has later been refined by Pulkkinen et al. (2014) and Pulkkinen (2015). In the proposed approach, the ridges are interpreted as nonlinear counterparts of principal component hyperplanes. The density is estimated by using Gaussian *kernels*.

In the linear PCA, principal component *scores* (i.e. coordinates) of a given sample point are obtained as projections along principal component axes. Generalizing the concept of a principal component axis, the projections in KDPCA are done along curvilinear trajectories onto ridges of a Gaussian kernel density estimate. Based on the theory of ridges, it is shown that such projections can be done in a well-defined coordinate system. A projection trajectory is formulated as a solution to a differential equation, and a predictor-corrector algorithm is developed for tracing its solution curve.

A strategy for choosing the kernel *bandwidth* is critical for the practical applicability of KDPCA. Use of an automatic bandwidth selector for this purpose is demonstrated. It is also shown that the nonlinear principal components are well-defined for any sufficiently large bandwidth. In addition, they converge to the linear ones when the kernel bandwidth approaches infinity, thus making the linear PCA as a special case of KDPCA.

Finally, KDPCA is extended to time series analysis. In analogy with the well-known *singular spectrum analysis* (SSA, e.g. Golyandina et al. 2001; Vautard et al. 1992), it is applied to the *phase space* representation of the time series. This approach addresses the main shortcoming of the linear SSA. That is, being based on the linear PCA, it cannot separate different components of a time series when its trajectory in the phase space forms a closed loop. This is the case for quasiperiodic (i.e. approximately periodic) time series that form an important special class appearing in many applications. Examples include climate analysis (e.g. Hsieh and Hamilton 2003; Hsieh 2004) and medical applications such as electrocardiography and electroencephalography (e.g. Rangayyan 2002).

The remaining of this paper is organized as follows. In Sect. 2 we recall the linear PCA. Section 3 is devoted to development of KDPCA, and in Sect. 4 it is extended to time series data. Test results on a synthetic dataset, simulated climate dataset and an atmospheric time series are given in Sect. 5. The computational complexity of KDPCA is also analyzed and a comparison with related methods is given. Finally, Sect. 6 concludes this paper. The more involved proofs are deferred to Appendix.

2 The linear PCA

As the proposed method is a generalization of the linear PCA (e.g. Jolliffe 1986), we briefly recall the theoretical background of this method in this section.

The linear PCA attempts to capture the variability of a given data

$$\boldsymbol{Y} = [\boldsymbol{y}_1 \quad \boldsymbol{y}_2 \quad \cdots \quad \boldsymbol{y}_n]^T \in \mathbb{R}^{n \times d}$$

by transforming the data into a lower-dimensional coordinate system via an orthogonal transformation. In this coordinate system, the axes point along directions of maximal variance.

For the formulation of PCA, we denote the mean-centered samples by

$$\tilde{\boldsymbol{y}}_i = \boldsymbol{y}_i - \hat{\boldsymbol{\mu}}, \quad \text{where } \hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{y}_i. \tag{1}$$

Assume that the mean-centered samples $\tilde{\boldsymbol{y}}_i$ are transformed into an m-dimensional space via the mapping

$$\boldsymbol{\theta}_i(\boldsymbol{A}) = \boldsymbol{A}^T \tilde{\boldsymbol{y}}_i,$$

where $\boldsymbol{A}$ is a $d \times m$ matrix with $0 < m < d$ and with orthonormal columns. Conversely, for the given coordinates $\boldsymbol{\theta}_i$ in the m-dimensional space, the corresponding *reconstruction* (i.e. projection onto the hyperplane spanned by the m first principal components) of $\boldsymbol{y}_i$ in the input space is obtained as

$$\hat{\boldsymbol{y}}_i(\boldsymbol{A}) = \hat{\boldsymbol{\mu}} + \boldsymbol{A}\boldsymbol{\theta}_i. \tag{2}$$

With the above definitions, it can be shown that finding the matrix $\boldsymbol{A}$ that minimizes the reconstruction error is equivalent to maximizing the variance in the transformed coordinate system (Jolliffe 1986). That is,

$$\min_{\boldsymbol{A} \in O(d,m)} \sum_{i=1}^{n} \|\hat{\boldsymbol{y}}_i(\boldsymbol{A}) - \hat{\boldsymbol{\mu}} - \tilde{\boldsymbol{y}}_i\|^2 = \max_{\boldsymbol{A} \in O(d,m)} \sum_{i=1}^{n} \|\boldsymbol{\theta}_i(\boldsymbol{A})\|^2,$$

where $O(d, m)$ denotes the set of $d \times m$ matrices having orthonormal columns. Furthermore, any i-th principal component corresponds to the direction of the i-th largest variance, and these directions form an orthogonal set.

The solution to the above optimization problems is the matrix $\boldsymbol{V}_m = [\boldsymbol{v}_1 \quad \boldsymbol{v}_2 \quad \cdots \quad \boldsymbol{v}_m]$, where the column vectors $\boldsymbol{v}_i$ are the (normalized) eigenvectors of the $d \times d$ sample *covariance* matrix

$$\hat{\boldsymbol{\Sigma}}_Y = \frac{1}{n-1}\sum_{i=1}^{n} (\boldsymbol{y}_i - \hat{\boldsymbol{\mu}})(\boldsymbol{y}_i - \hat{\boldsymbol{\mu}})^T \tag{3}$$

corresponding to the m largest eigenvalues. Thus, projection of the mean-centered sample set $\tilde{\boldsymbol{Y}}$ onto the m-dimensional subspace corresponding to the directions of largest variance is given by

$$\boldsymbol{\Theta} = \tilde{\boldsymbol{Y}} \boldsymbol{V}_m. \tag{4}$$

In statistical literature, the coordinates $\boldsymbol{\Theta} \in \mathbb{R}^{n \times m}$ obtained in this way are called principal component *scores* (e.g. Jolliffe 1986).

3 Nonlinear kernel density PCA

In this section we develop the kernel density principal component analysis (KDPCA). The method is based on estimation of the underlying density of the data with Gaussian kernels. It is shown that the nonlinear principal component scores of given sample points can be obtained one by one by successively projecting them onto ridges of the underlying density. The projection curves are defined as a solution to a differential equation, and a predictor-corrector method is developed for this purpose.

3.1 Statistical model and ridge definition

In order to formally define the underlying density of the data points, we assume that they are sampled from an m-dimensional *manifold* M embedded in $\mathbb{R}^d$. The sampling is done with normally distributed additive noise having variance σ^2. Denoting the points on M by the random variable $\boldsymbol{Z}$, the model is written as

$$\boldsymbol{X} = \boldsymbol{Z} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}_d(\boldsymbol{0}, \sigma^2 \boldsymbol{I}).$$

The marginal density for the observed variable $\boldsymbol{X}$ is

$$p(\boldsymbol{x}) = \int_M p(\boldsymbol{x}, z) dz = \int_M p(\boldsymbol{x}|z) w(z) dz, \tag{5}$$

where

$$p(\boldsymbol{x}|z) = \frac{1}{(\sqrt{2\pi}\sigma)^d} \exp\left(-\frac{\|\boldsymbol{x} - z\|^2}{2\sigma^2}\right)$$

and w denotes the density of $\boldsymbol{Z}$ supported on M. The idea is to estimate the manifold M from m-dimensional ridges of the marginal density p. A detailed theoretical analysis of this approach is given in Genovese et al. (2014).

We adapt the definition of a ridge set from Pulkkinen et al. (2014). An r-dimensional ridge point of a probability density is a local maximum in a subspace spanned by a subset of the eigenvectors of its Hessian matrix. These eigenvectors correspond to the $d - r$ algebraically smallest eigenvalues. The one-dimensional ridge set (i.e. ridge curve) of the density of a point set sampled from the above model is illustrated in Fig. 1.

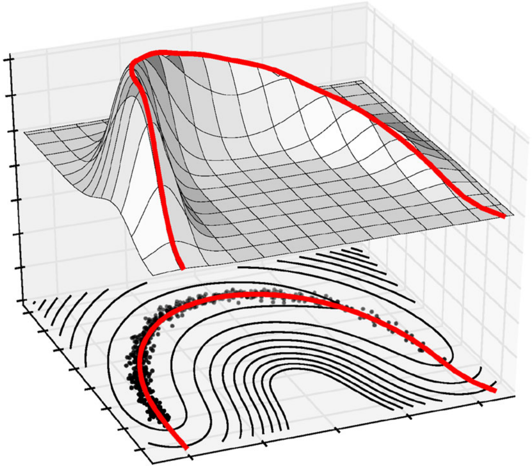

Fig. 1 Ridge curve of the density of a point set that is distributed around a curve

Definition 3.1 *A point* $\boldsymbol{x} \in \mathbb{R}^d$ *belongs to the* r*-dimensional ridge set* $\mathcal{R}_p^r$, where $0 \leq r < d$, of a twice differentiable probability density $p : \mathbb{R}^d \to \mathbb{R}$ if

$$\nabla p(\boldsymbol{x})^T \boldsymbol{v}_i(\boldsymbol{x}) = 0, \quad i > r, \tag{6a}$$
$$\lambda_{r+1}(\boldsymbol{x}) < 0, \tag{6b}$$
$$\lambda_1(\boldsymbol{x}) > \lambda_2(\boldsymbol{x}) > \cdots > \lambda_{r+1}(\boldsymbol{x}), \quad \text{if } r > 0, \tag{6c}$$

where $\lambda_1(\boldsymbol{x}) \geq \lambda_2(\boldsymbol{x}) \geq \cdots \geq \lambda_d(\boldsymbol{x})$ and $\{\boldsymbol{v}_i(\boldsymbol{x})\}_{i=1}^d$ denote the eigenvalues and the corresponding eigenvectors of $\nabla^2 p(\boldsymbol{x})$, respectively.

The following result shows a connection between ridge sets and linear principal components when the underlying density of the data is normal. This result follows trivially from the following lemma (Ozertem and Erdogmus 2011) and the fact that the logarithm of a normal density with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$ is a quadratic function whose gradient and Hessian are

$$\nabla \log p(\boldsymbol{x}) = -\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}) \quad \text{and} \quad \nabla^2 \log p = -\boldsymbol{\Sigma}^{-1},$$

respectively.

Lemma 3.1 *If* $p : \mathbb{R}^d \to \mathbb{R}$ *is twice differentiable, then* $\mathcal{R}_{\log p}^r = \mathcal{R}_p^r$ *for all* $r = 0, 1, 2, \ldots, d - 1$.

Proposition 3.1 *Let* $p : \mathbb{R}^d \to \mathbb{R}$ *be a* d*-variate normal density with mean* $\boldsymbol{\mu}$ *and positive definite covariance matrix*

$\boldsymbol{\Sigma}$. Denote the eigenvalues of $\boldsymbol{\Sigma}$ by $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$ and the corresponding eigenvectors by $\{\boldsymbol{v}_i\}_{i=1}^d$. Then for any $0 \leq r < d$ such that $\lambda_1 > \lambda_2 > \cdots > \lambda_{r+1}$ we have

$$\mathcal{R}_p^r = \begin{cases} \{\boldsymbol{\mu}\}, & r = 0, \\ \{\boldsymbol{\mu}\} + \operatorname{span}(\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_r), & r = 1, 2, \ldots, d-1. \end{cases}$$

For a linear model with $\boldsymbol{Z} = \boldsymbol{\mu} + \boldsymbol{W\Phi}$,

$$\boldsymbol{X} = \boldsymbol{Z} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\Phi} \sim \mathcal{N}_m(\boldsymbol{0}, \boldsymbol{I}), \quad \boldsymbol{\varepsilon} \sim \mathcal{N}_d(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$$

and a matrix $\boldsymbol{W} \in \mathbb{R}^{d \times m}$ having orthogonal columns, the marginal density (5) is normal. Namely, the observed variable $\boldsymbol{X}$ is distributed according to

$$\boldsymbol{X} \sim \mathcal{N}_d(\boldsymbol{\mu}, \boldsymbol{W}\boldsymbol{W}^T + \sigma^2 \boldsymbol{I})$$

(Tipping and Bishop 1999). From the above we observe that the m-dimensional ridge set of the density of $\boldsymbol{X} - \boldsymbol{\mu}$ coincides with the subspace spanned by the columns of $\boldsymbol{W}$.

Proposition 3.1 and the above observation suggest an approach for computing the principal component scores $\boldsymbol{\theta}$, as defined in Sect. 2, of a given point having an underlying density p. The idea is to project the point onto $\mathcal{R}_{\log p}^m$ in the subspace spanned by the eigenvectors $\{\boldsymbol{v}_i\}_{i=m+1}^d$ of $\nabla^2 \log p$ and then obtain projection coordinates along the first m eigenvectors. The remaining $d - m$ components, that are interpreted as noise, are discarded. The point in $\mathcal{R}_{\log p}^0$, that is the maximum of $\log p$, is chosen as the origin of the coordinate system. This idea will be generalized to the nonlinear case in the next section.

However, when the manifold M is nonlinear, the estimate for M obtained from ridges of the marginal density (5) is biased. In Pulkkinen (2015), this is analyzed with a representative example In this example, the manifold M is the unit circle parametrized as $\boldsymbol{f}(\Phi) = (\cos(\Phi), \sin(\Phi))$ and Φ is uniformly distributed on the interval $[0, 2\pi]$. With normally distributed noise, the bias is small, as illustrated in Fig. 2 with $\sigma = 0.2$. It occurs towards the center of curvature and is proportional to the ratio between σ and the curvature radius. In addition, for any compact and closed manifold M, the ridge estimate converges to M when σ tends to zero (Genovese et al. 2014).

3.2 Obtaining principal component scores from ridge sets

Based on Proposition 3.1, we now develop the theoretical basis for computing the first m nonlinear principal component scores of a given point $\boldsymbol{y}$. The idea is to obtain the scores one by one by successively projecting the point onto lower-dimensional ridge sets of its underlying density (5). The projections are done along eigenvector curves that are defined by a differential equation. The arc lengths of the curves are interpreted as the principal component scores. As a special case of this approach, we obtain an orthogonal projection onto a linear PCA hyperplane.

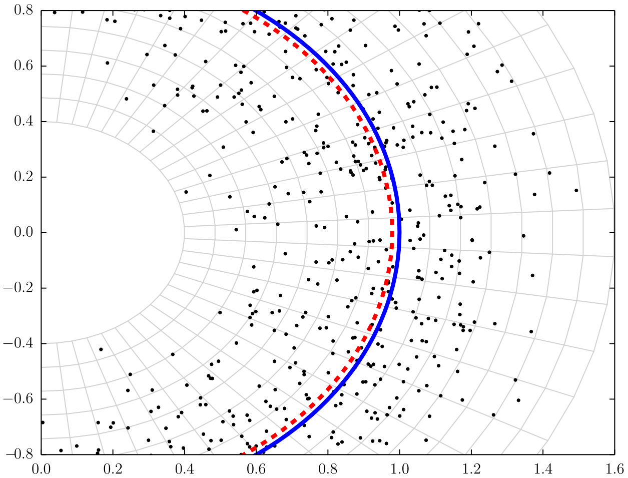

Fig. 2 Ridge curve of a circular distribution (*dashed line*) and the true generating curve (*solid line*)

For now, we assume that a given point $\boldsymbol{y} \in \mathbb{R}^d$ has already been projected onto an m-dimensional ridge set of its underlying density p with $m \leq d$. For $r = 1, 2, \ldots, m$, we define a projection curve $\boldsymbol{\gamma}_r : \mathbb{R} \to \mathbb{R}^d$ onto the $r-1$-dimensional ridge set as a solution to the initial value problem

$$\frac{d}{dt}\left[\boldsymbol{P}_r(\boldsymbol{\gamma}_r(t))\nabla \log p(\boldsymbol{\gamma}_r(t))\right] = \boldsymbol{0}, \qquad t \geq 0,$$
$$\boldsymbol{\gamma}_r(0) = \boldsymbol{x}_0, \quad \boldsymbol{x}_0 \in \mathcal{R}_{\log p}^r \setminus \mathcal{R}_{\log p}^{r-1}, \tag{7}$$

where $\boldsymbol{P}_r(\cdot) = \boldsymbol{I} - \boldsymbol{v}_r(\cdot)\boldsymbol{v}_r(\cdot)^T$ and $\{\boldsymbol{v}_i(\cdot)\}_{i=1}^d$ denote the eigenvectors corresponding to the eigenvalues $\lambda_1(\cdot) \geq \lambda_2(\cdot) \geq \cdots \geq \lambda_d(\cdot)$ of $\nabla^2 \log p$.

We begin with a special case that motivates the above definition and shows its connection to the linear PCA projection. Namely, for any d-dimensional normal density p, a ridge point $\boldsymbol{x}_0 \in \mathcal{R}_{\log p}^r$, where $1 \leq r \leq m$, can be projected onto the lower-dimensional ridge set $\mathcal{R}_{\log p}^{r-1}$ by following the solution curve of (7) that is a straight line parallel to the eigenvector $\boldsymbol{v}_r$. This property follows trivially from the definitions of the normal density and the ridge set because $\log p$ is in this case a quadratic function.

Proposition 3.2 *Let p be a d-variate normal density with symmetric and positive definite covariance matrix $\boldsymbol{\Sigma}$ and let $1 \leq r \leq d$. If the eigenvalues of $\boldsymbol{\Sigma}$ satisfy the condition $\lambda_1 > \lambda_2 > \cdots > \lambda_{r+1}$, then for any solution curve $\boldsymbol{\gamma}_r$ of the initial value problem (7) we have*

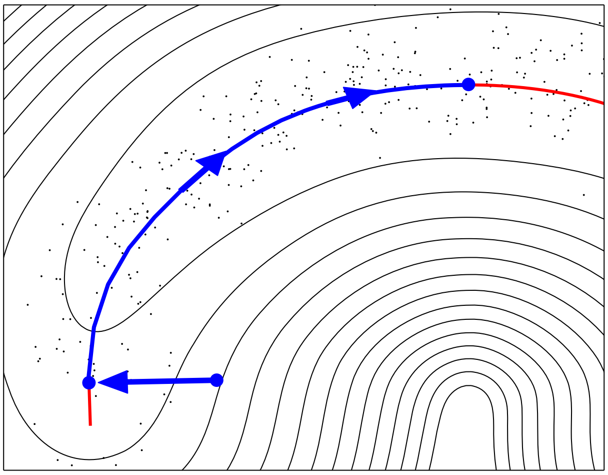

Fig. 3 Obtaining principal component scores of a point $\boldsymbol{y}$ by successive projections along ridge sets of $\log p$

$$\boldsymbol{\gamma}_r'(t)/\|\boldsymbol{\gamma}_r'(t)\| = \pm \boldsymbol{v}_r$$

for all $t \geq 0$. *Furthermore, if the sign of* $\boldsymbol{\gamma}_r'$ *is chosen such that*

$$\boldsymbol{\gamma}_r'(t)^T \nabla \log p(\boldsymbol{\gamma}_r(t)) > 0 \quad \textit{for all } t \geq 0, \tag{8}$$

then $\log p$ *has a unique maximum point* $\boldsymbol{x}^* \in \mathcal{R}_{\log p}^{r-1}$ *along the curve* $\boldsymbol{\gamma}_r$.

However, when the density p defined by Eq. (5) is not normal, we need the following assumption. It is needed to guarantee that the ridge sets of p induce a well-defined coordinate system. This assumption is reasonable when the data follows some unimodal distribution with clearly distinguishable major and minor axes as in Figs. 1 and 3. A higher-dimensional example will be given in Sect. 5.

Assumption 3.1 Define the set

$$U_{\boldsymbol{y}} = \{\boldsymbol{x} \in \mathbb{R}^d \mid \log p(\boldsymbol{x}) \geq \log p(\boldsymbol{y})\}.$$

Let $\lambda_1(\cdot) \geq \lambda_2(\cdot) \geq \cdots \geq \lambda_d(\cdot)$ denote the eigenvalues of $\nabla^2 \log p$. Assume that for all $\boldsymbol{x} \in U_{\boldsymbol{y}}$ we have

$$0 > \lambda_1(\boldsymbol{x}) > \lambda_2(\boldsymbol{x}) > \cdots > \lambda_{m+1}(\boldsymbol{x}) \tag{9}$$

and that $U_{\boldsymbol{y}}$ is compact and connected.

The density p defined by Eq. (5) is a C^∞-function. Thus, compactness and connectedness of the set $U_{\boldsymbol{y}}$ together with condition (9) guarantees unimodality of $\log p$ in the set $U_{\boldsymbol{y}}$. Damon (1998) and Miller (1998) give a rigorous treatment of ridge sets of C^∞-functions in a differential geometric framework. When Assumption 3.1 is satisfied, their results guarantee that the r-dimensional ridge set of $\log p$ induces a connected manifold in $U_{\boldsymbol{y}}$ for any $1 \leq r \leq m$. In addition, condition (9) implies differentiability of the Hessian eigenvectors (e.g. Magnus 1985), which is essential for the definition of the initial value problem (7).

Remark 3.1 When p is multimodal, a separate coordinate system can be obtained for each disjoint component of a superlevel set $\mathcal{L}_c = \{\boldsymbol{x} \in \mathbb{R}^d \mid \log p(\boldsymbol{x}) \geq c\}$ in which condition (9) is satisfied.

When the density p is not normal, obtaining an expression for the tangent vector $\boldsymbol{\gamma}_r'(t)$ is nontrivial. However, by utilizing the formula for the derivatives of eigenvectors (e.g. Magnus 1985), Eq. (7) for a general density p can after some calculation be rewritten as

$$\boldsymbol{A}_r(\boldsymbol{\gamma}_r(t))\boldsymbol{\gamma}_r'(t) = \boldsymbol{0}, \tag{10}$$

where

$$\begin{aligned}
\boldsymbol{A}_r(\boldsymbol{x}) =& \boldsymbol{P}_r(\boldsymbol{x})\nabla^2 \log p(\boldsymbol{x}) - \boldsymbol{F}_r(\boldsymbol{x}), \\
\boldsymbol{F}_r(\boldsymbol{x}) =& \boldsymbol{v}_r(\boldsymbol{x})^T \nabla \log p(\boldsymbol{x}) \nabla \boldsymbol{v}_r(\boldsymbol{x})^T \qquad (11)\\
& + \boldsymbol{v}_r(\boldsymbol{x}) \nabla \log p(\boldsymbol{x})^T \nabla \boldsymbol{v}_r(\boldsymbol{x}) \qquad (12)
\end{aligned}$$

and

$$\nabla \boldsymbol{v}_r(\boldsymbol{x}) = \left[\lambda_r(\boldsymbol{x})\boldsymbol{I} - \nabla^2 \log p(\boldsymbol{x})\right]^+ \nabla^3 \log p(\boldsymbol{x}) \boldsymbol{v}_r(\boldsymbol{x}),$$

and the operator "+" denotes the Moore-Penrose pseudoinverse (e.g. Golub and Loan 1996).

For a general density p, projection onto the ridge set $\mathcal{R}_{\log p}^{r-1}$ by maximizing $\log p$ along the curve $\boldsymbol{\gamma}_r$ requires additional justification. To this end, we first note that under Assumption 3.1 choosing the sign of $\boldsymbol{\gamma}_r'(t)$ according to (8) is sufficient to guarantee that $\boldsymbol{\gamma}_r$ with $r = m, m-1, \ldots, 1$ yield projection curves in a well-defined coordinate system. This is illustrated in Fig. 3. Second, we show that when $\boldsymbol{\gamma}_r$ approaches a ridge point $\boldsymbol{x}^* \in \mathcal{R}_{\log p}^{r-1}$, the tangent vector $\boldsymbol{\gamma}_r'$ becomes parallel to the eigenvector $\boldsymbol{v}_r$.

Proposition 3.3 *Let* $1 \leq r \leq d$ *and let* $\boldsymbol{\gamma}_r'$ *denote the normalized tangent vector of a solution curve of* (7). *If Assumption* 3.1 *and condition* (8) *are satisfied and*

$$\lim_{t \to t^*} \boldsymbol{v}_r(\boldsymbol{\gamma}_r(t))^T \nabla \log p(\boldsymbol{\gamma}_r(t)) = 0 \tag{13}$$

for some $t^* > 0$, *then*

$$\lim_{t \to t^*} |\boldsymbol{\gamma}_r'(t)^T \boldsymbol{v}_r(\boldsymbol{\gamma}_r(t))| = 1. \tag{14}$$

Proof Define the set

$$U = \{\boldsymbol{x} \in \mathbb{R}^d \mid \lambda_1(\boldsymbol{x}) < 0 \text{ and} \tag{15}$$

$$\lambda_1(\boldsymbol{x}) > \lambda_2(\boldsymbol{x}) > \cdots > \lambda_{r+1}(\boldsymbol{x})\}. \tag{16}$$

The range of the matrix in the second term of $\boldsymbol{F}_r(\boldsymbol{x})$ defined by Eq. (12), that is

$$\boldsymbol{G}(\boldsymbol{x}) = \boldsymbol{v}_r(\boldsymbol{x})\nabla \log p(\boldsymbol{x})^T \nabla \boldsymbol{v}_r(\boldsymbol{x}),$$

is clearly spanned by the vector $\boldsymbol{v}_r(\boldsymbol{x})$ for all $\boldsymbol{x} \in U$. Furthermore, $\boldsymbol{v}_r(\boldsymbol{x})$ is uniquely determined by condition (16). We also note that the range of the first term of the matrix $\boldsymbol{A}_r(\boldsymbol{x})$ defined by Eq. (11), that is

$$\boldsymbol{B}(\boldsymbol{x}) = \boldsymbol{P}_r(\boldsymbol{x})\nabla^2 \log p(\boldsymbol{x}),$$

is the set $\{\boldsymbol{w} \in \mathbb{R}^d \mid \boldsymbol{w}^T \boldsymbol{v}_r(\boldsymbol{x}) = 0\}$ for all $\boldsymbol{x} \in U$. This follows from the definition of the matrix $\boldsymbol{P}_r(\boldsymbol{x})$, the eigendecomposition of $\nabla^2 \log p(\boldsymbol{x})$ and condition (15) that guarantees nonsingularity of $\nabla^2 \log p(\boldsymbol{x})$.

On the other hand, by the limit (13) the first term of the matrix $\boldsymbol{F}_r(\boldsymbol{\gamma}_r(t))$ defined by Eq. (12), that is

$$\boldsymbol{v}_r(\boldsymbol{\gamma}(t))^T \nabla \log p(\boldsymbol{\gamma}_r(t))\nabla \boldsymbol{v}_r(\boldsymbol{\gamma}_r(t))^T$$

converges to zero as t approaches t^*. In view of the above observation that the ranges of the matrices $\boldsymbol{B}(\boldsymbol{x})$ and $\boldsymbol{G}(\boldsymbol{x})$ are orthogonal for all $\boldsymbol{x} \in U$, Eqs. (10)–(12) and Assumption 3.1 together with condition (8) thus imply that

$$\lim_{t \to t^*} \boldsymbol{B}(\boldsymbol{\gamma}_r(t))\boldsymbol{\gamma}_r'(t) = \boldsymbol{0} \quad \text{and} \quad \lim_{t \to t^*} \boldsymbol{G}(\boldsymbol{\gamma}_r(t))\boldsymbol{\gamma}_r'(t) = \boldsymbol{0}.$$

The claim follows from the first of the above limits because the range of the symmetric matrix $\boldsymbol{B}(\boldsymbol{x})$ is orthogonal to its null space. □

Proposition 3.3 implies the following properties that motivate seeking for a lower-dimensional ridge point by maximizing $\log p$ along the curve $\boldsymbol{\gamma}_r$.

Proposition 3.4 *If $\boldsymbol{\gamma}_r$ is a solution to (7) for some $1 \leq r \leq d$ and Assumption 3.1 and condition (8) are satisfied, then either $\boldsymbol{\gamma}_r(t) \in \mathcal{R}^r_{\log p} \setminus \mathcal{R}^{r-1}_{\log p}$ for all $t \geq 0$ or $\lim_{t \to t^*} \boldsymbol{\gamma}_r(t) \in \mathcal{R}^{r-1}_{\log p}$ for some $t^* > 0$. In the latter case, $\log p$ attains its local maximum along $\boldsymbol{\gamma}_r$ at the limit point $\boldsymbol{\gamma}_r(t^*)$.*

Proof By Eq. (7), the choice of $\boldsymbol{x}_0$ and the definition of the matrix $\boldsymbol{P}_r(\cdot)$, for all $i \neq r$ and $t \geq 0$ we have

$$\boldsymbol{v}_i(\boldsymbol{\gamma}_r(t))^T \nabla \log p(\boldsymbol{\gamma}_r(t)) = c_i$$

for some constants $c_i \neq 0$. By Assumption 3.1, condition (8) and Definition 3.1 this implies that either $\boldsymbol{\gamma}_r(t) \in \mathcal{R}^r_{\log p} \setminus \mathcal{R}^{r-1}_{\log p}$ for all $t \geq 0$ or $\lim_{t \to t^*} \boldsymbol{\gamma}_r(t) \in \mathcal{R}^{r-1}_{\log p}$ for some $t^* > 0$. In the latter case we have

$$\boldsymbol{v}_r(\boldsymbol{\gamma}_r(t^*))^T \nabla \log p(\boldsymbol{\gamma}_r(t^*)) = 0.$$

Thus, the limit (14) implies that

$$\lim_{t \to t^*} \frac{d}{dt} \log p(\boldsymbol{\gamma}_r(t)) = \lim_{t \to t^*} \nabla \log p(\boldsymbol{\gamma}_r(t))^T \boldsymbol{\gamma}_r'(t) = 0.$$

Furthermore, by condition (6b) the point $\boldsymbol{\gamma}_r(t^*)$ is a local maximum of $\log p$ along $\boldsymbol{\gamma}_r$. □

Finally, the arc length of a curve $\boldsymbol{\gamma}_r$ gives the (curvilinear) distance of its starting point to the ridge set $\mathcal{R}^{r-1}_{\log p}$. Assume that we have projected the given point $\boldsymbol{y}$ onto the ridge set $\mathcal{R}^m_{\log p}$. Starting from such a point, computing the arc lengths successively for $r = m, m-1, \ldots, 1$ then yields the first m principal component scores of $\boldsymbol{y}$. When Assumption 3.1 is satisfied, imposing the conditions (cf. Proposition 3.2)

$$\boldsymbol{\gamma}_r'(t)^T \nabla \log p(\boldsymbol{\gamma}_r(t)) > 0 \quad \text{and} \quad \|\boldsymbol{\gamma}_r'(t)\| = 1 \tag{17}$$

for all $r = m, m-1, \ldots, 1$ and $t \geq 0$ guarantees that the curves $\boldsymbol{\gamma}_r$ lie in the set $U_{\boldsymbol{y}}$.

Denote the projection of the point $\boldsymbol{y}$ onto the set $\mathcal{R}^m_{\log p}$ as $\tilde{\boldsymbol{y}}$ and the starting points of the curves $\boldsymbol{\gamma}_r$ as $\boldsymbol{x}_0^r$. The m first principal component scores of $\boldsymbol{y}$ are then obtained recursively as

$$\theta_r = s_r^* \int_0^{t_r^*} \|\boldsymbol{\gamma}_r'(t)\| dt, \quad r = m, m-1, \ldots, 1, \tag{18}$$

where

$$\boldsymbol{x}_0^r = \begin{cases} \tilde{\boldsymbol{y}}, & r = m, \\ \boldsymbol{\gamma}_{r+1}(t_{r+1}^*), & 1 \leq r < m. \end{cases}$$

Here we assume that for each r there exists $t_r^* \geq 0$ such that $\boldsymbol{\gamma}_r(t_r^*) \in \mathcal{R}^{r-1}_{\log p}$. The multiplier $s_r^* = \lim_{t \to t_r^*-} s_r(t)$, where

$$s_r(t) = \begin{cases} 1, & \text{if } \boldsymbol{\gamma}_r'(t)^T \boldsymbol{v}_r(\boldsymbol{\gamma}_r(t)) > 0, \\ -1, & \text{otherwise}, \end{cases} \tag{19}$$

is introduced to ensure that the principal component score θ_r has the correct sign.

3.3 Density estimation with Gaussian kernels

In practice, the target density p is unknown and it needs to be estimated from the observations that we shall denote as $\boldsymbol{Y} = \{\boldsymbol{y}_i\}_{i=1}^n$. To this end, we use Gaussian kernels, and in this section we recall the necessary theoretical results. We also show that the bandwidth can be interpreted as a free scale parameter and that choosing a sufficiently large bandwidth yields a unimodal kernel density with connected ridge sets. This is a key result since the target density might not in all cases satisfy the these assumptions. Finally, we show that the linear PCA can be obtained as a special case of the above approach.

Definition 3.2 *The Gaussian kernel density estimate* $\hat{p}_{\boldsymbol{H}}$ obtained by drawing a set of samples $\boldsymbol{Y} = \{\boldsymbol{y}_i\}_{i=1}^n \subset \mathbb{R}^d$ from a probability density $p : \mathbb{R}^d \to \mathbb{R}$ is

$$\hat{p}_{\boldsymbol{H}}(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n} K_{\boldsymbol{H}}(\boldsymbol{x} - \boldsymbol{y}_i), \tag{20}$$

where the *kernel* $K_{\boldsymbol{H}} : \mathbb{R}^d \to]0, \infty[$ is the *Gaussian* function

$$K_{\boldsymbol{H}}(\boldsymbol{x}) = \frac{1}{\sqrt{(2\pi)^d |\boldsymbol{H}|}} \exp\left(-\frac{1}{2}\boldsymbol{x}^T \boldsymbol{H}^{-1}\boldsymbol{x}\right) \tag{21}$$

with symmetric and positive definite *bandwidth* matrix $\boldsymbol{H}$.

The problem of finding an optimal bandwith matrix $\boldsymbol{H}$ is well-studied in the literature. Under mild assumptions on the target density p, asymptotic convergence of several bandwidth selectors for the target density and its derivatives can be proven (e.g. Chacón et al. 2011). That is, for a given order k, the k-th derivative of the estimator $\hat{p}_{\boldsymbol{H}}$ converges to the k-th derivative of p in probability as the number of samples n tends to infinity. Extending these results, Genovese et al. (2014) show that any ridge set of a kernel density estimate $\hat{p}_{h^2\boldsymbol{I}}$ converges to the ridge set of p as $n \to \infty$ for an appropriately chosen sequence of bandwidths h. In addition, several bandwidth selectors have been implemented in the ks package for R (Duong 2007).

Using the function $\hat{p}_{\boldsymbol{H}}$ as a density estimate might not be appropriate in all cases. The target density p might not be unimodal or have connected ridge sets, as assumed in Sect. 3.2. To address this issue, the bandwidth $\boldsymbol{H}$ can be given an alternative interpretation when parametrized as $\boldsymbol{H} = h^2\boldsymbol{I}$. Namely, the following result shows that a ridge point lies on a locally defined principal component hyperplane. This hyperplane is determined by a weighted sample mean and the eigenvectors of a weighted sample covariance matrix, where the weights are Gaussian functions. For the following, we introduce the notation $\hat{p}_h = \hat{p}_{h^2\boldsymbol{I}}$.

Theorem 3.1 *Let* $\hat{p}_h : \mathbb{R}^d \to \mathbb{R}$ *be a Gaussian kernel density, let* $0 < r < d$ *and denote the eigenvectors of* $\nabla^2 \log \hat{p}_h(\cdot)$ *corresponding to the* r *greatest eigenvalues by* $\{\boldsymbol{v}_i(\cdot)\}_{i=1}^r$. *Define*

$$\tilde{\boldsymbol{\mu}}(\boldsymbol{x}) = \sum_{i=1}^{n} c_i(\boldsymbol{x})\boldsymbol{y}_i, \tag{22}$$

$$\tilde{\boldsymbol{\Sigma}}(\boldsymbol{x}) = \sum_{i=1}^{n} c_i(\boldsymbol{x})[\boldsymbol{y}_i - \tilde{\boldsymbol{\mu}}(\boldsymbol{x})][\boldsymbol{y}_i - \tilde{\boldsymbol{\mu}}(\boldsymbol{x})]^T, \tag{23}$$

where

$$c_i(\boldsymbol{x}) = \frac{\exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{y}_i\|^2}{2h^2}\right)}{\sum_{j=1}^{n}\exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{y}_j\|^2}{2h^2}\right)}, \quad i = 1, 2, \ldots, n.$$

Assume that the eigenvalues of $\nabla^2 \log \hat{p}_h(\boldsymbol{x})$ *satisfy the condition* $\lambda_1(\boldsymbol{x}) > \lambda_2(\boldsymbol{x}) > \cdots > \lambda_{r+1}(\boldsymbol{x})$. *Then*

$$\nabla \log \hat{p}_h(\boldsymbol{x})^T \boldsymbol{v}_i(\boldsymbol{x}) = 0 \quad \textit{for all } i > r$$

if and only if

$$\boldsymbol{x} - \tilde{\boldsymbol{\mu}}(\boldsymbol{x}) \in \operatorname{span}(\tilde{\boldsymbol{v}}_1(\boldsymbol{x}), \tilde{\boldsymbol{v}}_2(\boldsymbol{x}), \ldots, \tilde{\boldsymbol{v}}_r(\boldsymbol{x})),$$

where $\{\tilde{\boldsymbol{v}}_i(\boldsymbol{x})\}_{i=1}^r$ *denote the eigenvectors of* $\tilde{\boldsymbol{\Sigma}}(\boldsymbol{x})$ *corresponding to the* r *greatest eigenvalues.*

Proof First, we note the formulae

$$\nabla \log \hat{p}_h(\boldsymbol{x}) = \frac{\nabla \hat{p}_h(\boldsymbol{x})}{\hat{p}_h(\boldsymbol{x})}$$

and

$$\nabla^2 \log \hat{p}_h(\boldsymbol{x}) = \frac{\nabla^2 \hat{p}_h(\boldsymbol{x})}{\hat{p}_h(\boldsymbol{x})} - \frac{\nabla \hat{p}_h(\boldsymbol{x})\nabla \hat{p}_h(\boldsymbol{x})^T}{\hat{p}_h(\boldsymbol{x})^2}.$$

By a straightforward calculation we then obtain that

$$h^2 \nabla \log p_h(\boldsymbol{x}) = -[\boldsymbol{x} - \tilde{\boldsymbol{\mu}}(\boldsymbol{x})] \tag{24}$$

and

$$h^4 \nabla^2 \log p_h(\boldsymbol{x}) + h^2 \boldsymbol{I} = \tilde{\boldsymbol{\Sigma}}(\boldsymbol{x}). \tag{25}$$

By Eq. (25), the matrices $\nabla^2 \log \hat{p}_h$ and $\tilde{\boldsymbol{\Sigma}}(\boldsymbol{x})$ have the same eigenvectors. Hence, by Eq. (24) the condition that

$$[\boldsymbol{x} - \tilde{\boldsymbol{\mu}}(\boldsymbol{x})]^T \tilde{\boldsymbol{v}}_i(\boldsymbol{x}) = 0 \quad \text{for all } i > r$$

is equivalent to

$$\nabla \log \hat{p}_h(\boldsymbol{x})^T \boldsymbol{v}_i(\boldsymbol{x}) = 0 \quad \text{for all } i > r,$$

from which the claim follows by the orthogonality of the eigenvectors $\tilde{\boldsymbol{v}}_i(\boldsymbol{x})$. □

Ridges of the Gaussian kernel density $\hat{p}_h$ can be used in an exploratory fashion by adjusting the bandwidth h. As suggested by Theorem 3.1, this parameter determines the scale of the structures sought from the data. However, differently to the normal density in Proposition 3.2, the density

$\hat{p}_h$ is not generally unimodal or have connected ridge sets as requred by Assumption 3.1. For instance, it becomes multimodal when h is too small.

Fortunately, Assumption 3.1 for $\hat{p}_h$ and the superlevel set

$$U_{\boldsymbol{Y}} = \bigcup_{i=1}^{n} U_{\boldsymbol{y}_i} = \bigcup_{i=1}^{n} \{\boldsymbol{x} \in \mathbb{R}^d \mid \log \hat{p}_h(\boldsymbol{x}) \geq \log \hat{p}_h(\boldsymbol{y}_i)\}$$

can be satisfied by choosing a sufficiently large h. This is guaranteed by the following theorem that is proven in Appendix. Although $\hat{p}_h$ does not necessarily reflect the underlying density with such bandwidth choices, its ridges can still capture nonlinear structure that cannot be described by the linear PCA.

Assumption 3.2 The $r+1$ greatest eigenvalues of the sample covariance matrix $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{Y}}$ defined by equation (3) satisfy the conditions $\lambda_1 > \lambda_2 > \cdots > \lambda_{r+1} > 0$.

Theorem 3.2 *Under Assumption 3.2 for $r = m$ and given $\boldsymbol{Y} = \{\boldsymbol{y}_i\}_{i=1}^N$, for a Gaussian kernel density $\hat{p}_h$ there exists $h_0 > 0$ such that Assumption 3.1 is satisfied for $U_{\boldsymbol{Y}}$ and $\hat{p}_h$ for all $h \geq h_0$.*

An important special case arises when h approaches infinity. At this limit, any r-dimensional ridge set of the Gaussian kernel density approaches the r-dimensional PCA hyperplane, which can be readily observed from Eqs. (22)–(25). A rigorous proof of this property is deferred to Appendix. As a result, the linear principal components are a special case of those obtained from kernel density ridges.

Theorem 3.3 *For $\boldsymbol{Y} = \{\boldsymbol{y}_i\}_{i=1}^N \subset \mathbb{R}^d$, let $\hat{p}_h : \mathbb{R}^d \to \mathbb{R}$ be a Gaussian kernel density, let $0 \leq r < d$ and let Assumption 3.2 be satisfied. Define the set*

$$S_\infty^r = \left\{ \hat{\boldsymbol{\mu}} + \sum_{i=1}^{r} \alpha_i \boldsymbol{v}_i \mid \boldsymbol{\alpha} \in \mathbb{R}^r \right\},$$

where $\hat{\boldsymbol{\mu}}$ denotes the sample mean (1) and $\{\boldsymbol{v}_i\}_{i=1}^r$ denote the eigenvectors of the sample covariance matrix $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{Y}}$ corresponding to the eigenvalues $\{\lambda_i\}_{i=1}^r$. Then for any compact set $U \subset \mathbb{R}^d$ such that $U \cap S_\infty^r \neq \emptyset$ and $\varepsilon > 0$ there exists $h_0 > 0$ such that

$$\left.\begin{array}{l} \mathrm{dist}(\mathcal{R}_{\hat{p}_h}^r \cap U, S_\infty^r) < \varepsilon, \\ \mathrm{dist}(S_\infty^r \cap U, \mathcal{R}_{\hat{p}_h}^r) < \varepsilon \end{array}\right\} \quad \textit{for all } h \geq h_0,$$

where

$$\mathrm{dist}(S_1, S_2) = \sup_{\boldsymbol{x} \in S_1} \inf_{\boldsymbol{y} \in S_2} \|\boldsymbol{x} - \boldsymbol{y}\|.$$

3.4 Algorithm for estimating principal component scores

Based on the theory developed in Sects. 3.2 and 3.3, we now develop the algorithm for computing the nonlinear principal component score estimates

$$\hat{\boldsymbol{\Theta}} = [\hat{\boldsymbol{\theta}}_1 \quad \hat{\boldsymbol{\theta}}_2 \quad \cdots \quad \hat{\boldsymbol{\theta}}_n]^T \in \mathbb{R}^{n \times m}$$

of a given sample set

$$\boldsymbol{Y} = [\boldsymbol{y}_1 \quad \boldsymbol{y}_2 \quad \cdots \quad \boldsymbol{y}_n]^T \in \mathbb{R}^{n \times d}$$

for a given $0 < m \leq d$. This amounts to first projecting the samples $\boldsymbol{y}_i$ onto the ridge set $\mathcal{R}_{\log \hat{p}_{\boldsymbol{H}}}^m$ of the kernel density $\log \hat{p}_{\boldsymbol{H}}$ and then successively projecting them onto the lower-dimensional ridge sets $\mathcal{R}_{\log \hat{p}_{\boldsymbol{H}}}^r$ until $r = 0$. The latter projections are done by tracing the curves denoted by $\hat{\boldsymbol{\gamma}}_r$ by using a predictor-corrector method As a by-product, the principal component scores are obtained from a numerical approximation of the integral (18).

A pseudocode of the algorithm is listed as Algorithm 1. It involves the initial projection onto the ridge set $\mathcal{R}_{\log \hat{p}_{\boldsymbol{H}}}^m$ (lines 2 and 3), and after that $m \times n$ loops. Each iteration for $r = m, m-1, \ldots, 1$ projects each of the n sample points onto the ridge set $\mathcal{R}_{\log \hat{p}_{\boldsymbol{H}}}^{r-1}$. The intermediate projections are stored in the variables $\{\boldsymbol{x}_i^*\}_{i=1}^n$. For the initial ridge projection and the corrector steps, the algorithm utilizes the trust region Newton method developed by Pulkkinen et al. (2014) (the GTRN algorithm). This method is briefly described at the end of this subsection.

In the following, we describe the steps for carrying out one ridge projection (i.e. one iteration of the loop over the index i) for a given r. The starting point $\boldsymbol{x}_0$ for $\hat{\boldsymbol{\gamma}}_r$ is chosen as $\boldsymbol{x}_i^*$ representing the projection of the sample point $\boldsymbol{y}_i$ onto the set $\mathcal{R}_{\log \hat{p}_{\boldsymbol{H}}}^r$. Assuming that there exists a monotoneously increasing sequence $\{t_k\}$ such that $\hat{\boldsymbol{\gamma}}_r(t_{k^*}) \in \mathcal{R}_{\log \hat{p}_{\boldsymbol{H}}}^{r-1}$ for some k^*, we introduce the notation $\boldsymbol{x}_k = \hat{\boldsymbol{\gamma}}_r(t_k)$ for the iterates along the curve $\hat{\boldsymbol{\gamma}}_r$. With this notation, an approximation to the integral (18) is given by

$$\hat{\theta}_r = \int_0^{t_r^*} \|\hat{\boldsymbol{\gamma}}_r'(t)\| dt \approx \sum_{k=1}^{k^*} \|\hat{\boldsymbol{\gamma}}_r(t_k) - \hat{\boldsymbol{\gamma}}_r(t_{k-1})\|$$
$$= \sum_{k=1}^{k^*} \|\boldsymbol{x}_k - \boldsymbol{x}_{k-1}\|.$$

The algorithm uses a predictor–corrector method to generate the iterates $\boldsymbol{x}_k$. At the predictor step (line 23), the algorithm proceeds along a tangent vector $\boldsymbol{u}_k = \hat{\boldsymbol{\gamma}}_r'(t_k)$ solved from Eq. (10). That is,

$$\tilde{\boldsymbol{x}}_k = \boldsymbol{x}_k + \tau s_k \boldsymbol{u}_k,$$

Algorithm 1: Nonlinear principal component scores

```
input : sample points $\boldsymbol{Y} = [\boldsymbol{y}_1 \ \ \boldsymbol{y}_2 \ \ \cdots \ \ \boldsymbol{y}_n]^T \in \mathbb{R}^{n \times d}$
        Gaussian kernel density $\hat{p}_{\boldsymbol{H}} : \mathbb{R}^d \rightarrow \mathbb{R}$
        ridge dimension $0 < m \leq d$
        step size $\tau > 0$
output: principal component scores
        $\hat{\boldsymbol{\Theta}} = [\hat{\boldsymbol{\theta}}_1 \ \ \hat{\boldsymbol{\theta}}_2 \ \ \cdots \ \ \hat{\boldsymbol{\theta}}_n]^T \in \mathbb{R}^{n \times m}$
1  $\hat{\boldsymbol{\Theta}} \leftarrow \boldsymbol{0}$
2  for $i = 1, 2, \ldots, n$ do
3      $\boldsymbol{x}_i^* \leftarrow$ GTRN($\log \hat{p}_{\boldsymbol{H}}, m, \boldsymbol{y}_i, \tau, 10^{-5}$)
4  for $r = m, m-1, \ldots, 1$ do
5      for $i = 1, 2, \ldots, n$ do
6          $\boldsymbol{x}_0 \leftarrow \boldsymbol{x}_i^*$
7          for $k = 0, 1, \ldots$ do
8              Obtain the tangent $\boldsymbol{u}_k$ from (10), $\|\boldsymbol{u}_k\| = 1$.
9              if $\boldsymbol{u}_k^T \nabla \log \hat{p}_{\boldsymbol{H}}(\boldsymbol{x}_k) > 0$ then
10                 $s_k \leftarrow 1$
11             else
12                 $s_k \leftarrow -1$
13             if $k > 0$ then
14                 if $s_{k-1}\boldsymbol{u}_{k-1}^T \boldsymbol{u}_k s_k < 0$ then
15                     $\bar{\boldsymbol{x}} \leftarrow (\boldsymbol{x}_{k-1} + \boldsymbol{x}_k)/2$
16                     $\boldsymbol{x}_i^* \leftarrow$ GTRN($\log \hat{p}_{\boldsymbol{H}}, r-1, \bar{\boldsymbol{x}}, 0.5\tau, 10^{-5}$)
17                     $\hat{\theta}_{i,r} \leftarrow \hat{\theta}_{i,r} + \|\boldsymbol{x}_i^* - \boldsymbol{x}_{k-1}\|$
18                     if $(\boldsymbol{x}_i^* - \boldsymbol{x}_{k-1})^T \boldsymbol{v}_r(\boldsymbol{x}_i^*) < 0$ then
19                         $\hat{\theta}_{i,r} \leftarrow -\hat{\theta}_{i,r}$
20                     Return to line 5.
21                 else
22                     $\hat{\theta}_{i,r} \leftarrow \hat{\theta}_{i,r} + \|\boldsymbol{x}_k - \boldsymbol{x}_{k-1}\|$
23             $\tilde{\boldsymbol{x}}_k \leftarrow \boldsymbol{x}_k + \tau s_k \boldsymbol{u}_k$
24             $\boldsymbol{x}_{k+1} \leftarrow$ GTRN($\log \hat{p}_{\boldsymbol{H}}, r, \tilde{\boldsymbol{x}}_k, 0.5\tau, 10^{-5}$)
```

where $\tau > 0$ is some user-supplied step size, $\|\boldsymbol{u}_k\| = 1$ and the multiplier

$$s_k = \begin{cases} 1, & \text{if } \boldsymbol{u}_k^T \nabla \log \hat{p}_{\boldsymbol{H}}(\boldsymbol{x}_k) > 0 \\ -1, & \text{otherwise} \end{cases}$$

(lines 9–12) is introduced to impose conditions (17). To project the predictor estimate $\tilde{\boldsymbol{x}}_k$ back to the ridge set $\mathcal{R}^r_{\log \hat{p}_{\boldsymbol{H}}}$, the algorithm takes a corrector step (line 24).

A stopping criterion is imposed to terminate the tracing of the curve $\hat{\boldsymbol{\gamma}}_r$ when a maximum of $\log \hat{p}_{\boldsymbol{H}}$ along $\hat{\boldsymbol{\gamma}}_r$ is encountered (line 14). For $k > 0$, the condition

$$s_{k-1}\boldsymbol{u}_{k-1}^T \boldsymbol{u}_k s_k < 0$$

tests whether the gradient changes sign along the curve. When this condition is met, the algorithm projects the midpoint of the current and previous iterate onto a nearby ridge point $\boldsymbol{x}_i^* \in \mathcal{R}^{r-1}_{\log \hat{p}_{\boldsymbol{H}}}$ (line 16). At lines 18–19, the algorithm computes the sign s_r^* for the integral (18) by approximately testing condition (19). The inner iteration (i.e. iteration of the loop over the index k) is then terminated, and the point $\boldsymbol{x}_i^*$ is retained as a starting point for projection onto a lower-dimensional ridge set.

Tests for unimodality or connectedness of ridge sets are not included in Algorithm 1 for simplicity. Unimodality can be tested by finding all modes of the density $\hat{p}_{\boldsymbol{H}}$ by using the GTRN algorithm, as described in Pulkkinen (2015). Disconnectedness can be tested if a curve $\hat{\boldsymbol{\gamma}}_r$ crosses a point $\boldsymbol{x}$ where $\lambda_{r+1}(\boldsymbol{x}) = 0$ or $\lambda_i(\boldsymbol{x}) = \lambda_j(\boldsymbol{x})$ for some $i, j = 1, 2, \ldots, r+1$ such that $i \neq j$, where $\lambda_i(\cdot)$ denote the eigenvalues of $\nabla^2 \log \hat{p}_{\boldsymbol{H}}$ (Miller 1998). When multimodality or a disconnected ridge set is detected, the algorithm can be restarted with larger h or smaller initial ridge dimension m.

An alternative approach to amend the above situation is to relax the requirement of having a single coordinate system. This can in particular occur when $\hat{p}_{\boldsymbol{H}}$ is used as a density estimate and the target density p does not satisfy the above assumptions. In such case, the points can be given as many coordinate systems as there are clusters in the data identified by the modes of $\hat{p}_{\boldsymbol{H}}$ (see e.g. Pulkkinen 2015 for the case $m = 1$). However, this does not address the more difficult case when the density is unimodal but not all of its first m ridge sets are connected.

The GTRN algorithm (Pulkkinen et al. 2014) utilized in Algorithm 1 implements a Newton-type method for projecting a d-dimensional point onto an r-dimensional ridge set of a probability density. The method successively maximizes a quadratic model of the objective function. The maximization is constrained within a *trust region* to guarantee convergence. To obtain a ridge projection, it is done in the subspace spanned by the Hessian eigenvectors corresponding to the $d - r$ smallest eigenvalues. That is, at each iteration l the next iterate $z_{l+1} = z_l + \boldsymbol{s}_l$ is obtained by solving the subproblem

$$\max_{\boldsymbol{s}} Q_l(\boldsymbol{s}) \quad \text{s.t.} \begin{cases} \|\boldsymbol{s}\| \leq \Delta_l, \\ \boldsymbol{s} \in \operatorname{span}(\boldsymbol{v}_{r+1}(z_l), \boldsymbol{v}_{r+2}(z_l), \ldots, \boldsymbol{v}_d(z_l)), \end{cases}$$

where Q_l denotes the quadratic model at the current iterate z_l, $\{\boldsymbol{v}_i(z_l)\}_{i=r+1}^d$ denote the eigenvectors and $\Delta_l \leq \Delta_{\max}$ denotes the current trust region radius that is updated after each iteration. For each call of GTRN, Algorithm 1 uses the experimentally chosen $\Delta_{\max} = 0.5\tau$ (τ for the initial projection) and stopping criterion threshold $\varepsilon_{pr} = 10^{-5}$.

Remark 3.2 The GTRN algorithm can be viewed as an approximate solution method to an initial value problem of the form (7), where $\boldsymbol{P}_r(\cdot) = \sum_{i=1}^r \boldsymbol{v}_i(\cdot)\boldsymbol{v}_i(\cdot)^T$. As Algorithm 1, GTRN yields an orthogonal projection when applied to the logarithm of a normal density. Differently to Algorithm 1, projection of a d-dimensional point onto an r-dimensional ridge set with this algorithm only requires continuity of the first r Hessian eigenvectors. That is, when the $r + 1$ greatest eigenvalues are distinct in the set $U_{\boldsymbol{Y}}$.

Remark 3.3 Consistent orientation of the eigenvectors $\boldsymbol{v}_r(\boldsymbol{x}_i^*)$ at the projected points is necessary for the principal component scores $\hat{\theta}_{i,r}$ to have correct signs. However, in practice the signs of the eigenvectors depend on the numerical algorithm for computing them. Therefore, the implementation of Algorithm 1 uses an Euclidean minimum spanning tree (e.g. Jaromczyk and Toussaint 1992) to align the eigenvectors after each iteration of the outer loop.

4 Nonlinear extension of SSA to time series data

In this section, the KDPCA method developed in Sect. 3 is extended to time series data. The method, that we call KDSSA, is based on the singular spectrum analysis (SSA) that is an extension of the linear PCA. In SSA, a time series is embedded in a multidimensional *phase space*. This is done by constructing a *trajectory matrix* from time-lagged copies of the time series. Formally, the trajectory matrix of a time series $\boldsymbol{x} = (x_1, x_2, \dots, x_n)$ is defined as

$$\boldsymbol{Y}_{\boldsymbol{x},L} = \begin{bmatrix} x_1 & x_2 & x_3 & \cdots & x_L \\ x_2 & x_3 & x_4 & \cdots & x_{L+1} \\ x_3 & x_4 & x_5 & \cdots & x_{L+2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{n-L+1} & x_{n-L+2} & x_{n-L+3} & \cdots & x_n \end{bmatrix}, \qquad (26)$$

where L is some user-supplied time window length.

Applying the linear PCA to the above matrix, one can obtain the principal components and the reconstructed time series by using the formulae given by Vautard et al. (1992). Generalizing their approach, we minimize the reconstruction error

$$E(\boldsymbol{x}) = \sum_{i=1}^{n-L+1} \sum_{j=1}^{L} \left(\tilde{y}_{i,j} - x_{i+j-1}\right)^2 \qquad (27)$$

using the first m nonlinear principal components, where $m \leq L$. Here the vectors $\tilde{\boldsymbol{y}}_i$ denote the projections of the row vectors $\boldsymbol{y}_i$ of $\boldsymbol{Y}_{\boldsymbol{x},L}$ onto the m-dimensional ridge set of their Gaussian kernel density.

A straightforward calculation shows that by equating the gradient $\nabla E(\boldsymbol{x})$ to zero, we obtain the formulae

$$x_i^* = \begin{cases} \dfrac{1}{L} \displaystyle\sum_{j=1}^{L} \tilde{y}_{i-j+1,j}, & L \leq i \leq n-L+1 \\ \dfrac{1}{i} \displaystyle\sum_{j=1}^{i} \tilde{y}_{i-j+1,j}, & 1 \leq i \leq L-1 \\ \dfrac{1}{n-i+1} \displaystyle\sum_{j=i-n+L}^{L} \tilde{y}_{i-j+1,j}, & n-L+2 \leq i \leq n \end{cases} \qquad (28)$$

for the elements of the reconstructed time series such that $E(\boldsymbol{x}^*)$ minimizes the reconstruction error (27).

In this paper the nonlinear SSA is applied to quasiperiodic time series (i.e. noisy time series having some underlying periodic pattern). The motivation is as follows. Assuming that a time series follows the model

$$X(t) = f(t) + \varepsilon(t)$$

for some periodic function f and ε representing the noise, it is reasonable to model the trajectory samples (i.e. the rows of the matrix $\boldsymbol{Y}_{\boldsymbol{x},L}$) as a point set that is randomly distributed around a closed curve (cf. Fig. 10 in Sect. 5).

When the aim is to obtain a noise-free time series from a reconstructed phase space trajectory, only an approximate projection onto the ridge curve (i.e. one-dimensional ridge set) of the trajectory density suffices. The GTRN algorithm developed by Pulkkinen et al. (2014) is appropriate for this purpose. On the other hand, a parametrization of the reconstructed trajectory can be obtained by the algorithm developed by Pulkkinen (2015). Differently to Algorithm 1, this algorithm yields a continuous parametrization even when the trajectory density is multimodal, provided that the set of its ridge curves forms a single closed loop. Both of the aforementioned approaches are demonstrated in the next section.

5 Practical applications

This section is devoted to practical applications of KPDCA. The method is applied to a synthetic dataset and climate model output that exhibit highly nonlinear behaviour. In addition, its SSA-based extension is applied to an atmospheric time series. Finally, computational complexity analysis and comparison with related methods is given.

5.1 Test setup

Algorithm 1 as well as the algorithms developed by Pulkkinen et al. (2014) and Pulkkinen (2015) used in the tests were implemented in Fortran 95. Algorithm 1 was run with $m = d$ and $\tau = 0.05h$. For the nonlinear SSA, the above algorithms were run with their default parameters, except for GTRN the parameters $\Delta_{\max}$ and ε_{pr} were chosen as $0.25h$ and 10^{-4}, respectively.

5.2 Synthetic dataset

In the first experiments we consider a synthetic dataset generated from the model described in Sect. 3.1. Here the three-dimensional data points represented by the random variable $\boldsymbol{X}$ are sampled from a two-dimensional surface with additive

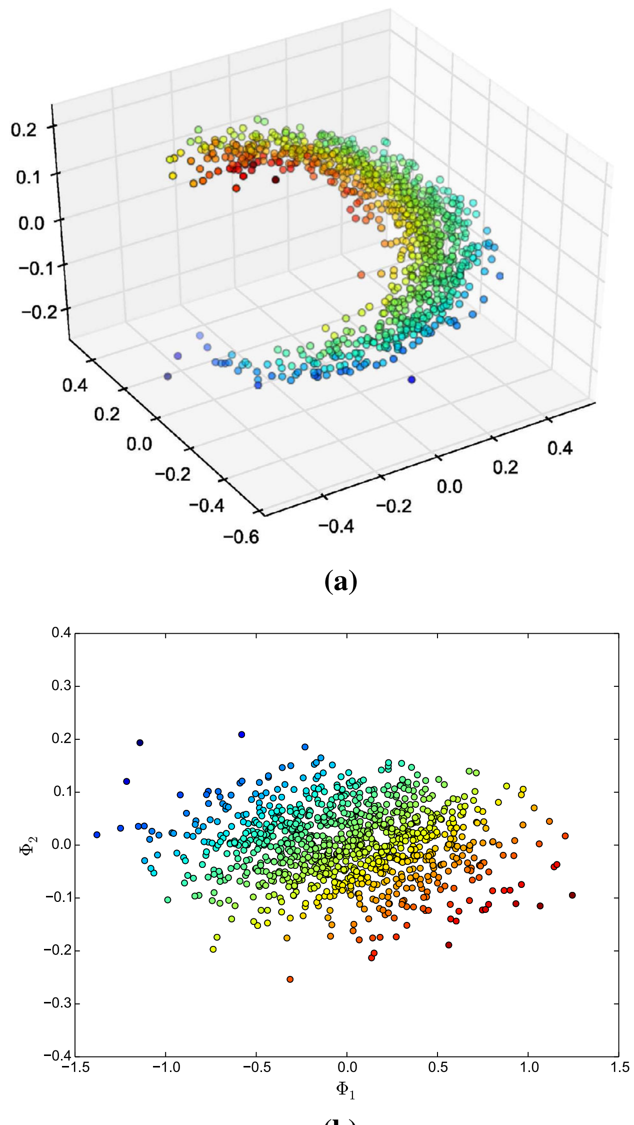

Fig. 4 The synthetic dataset. **a** Input coordinates. **b** Surface coordinates

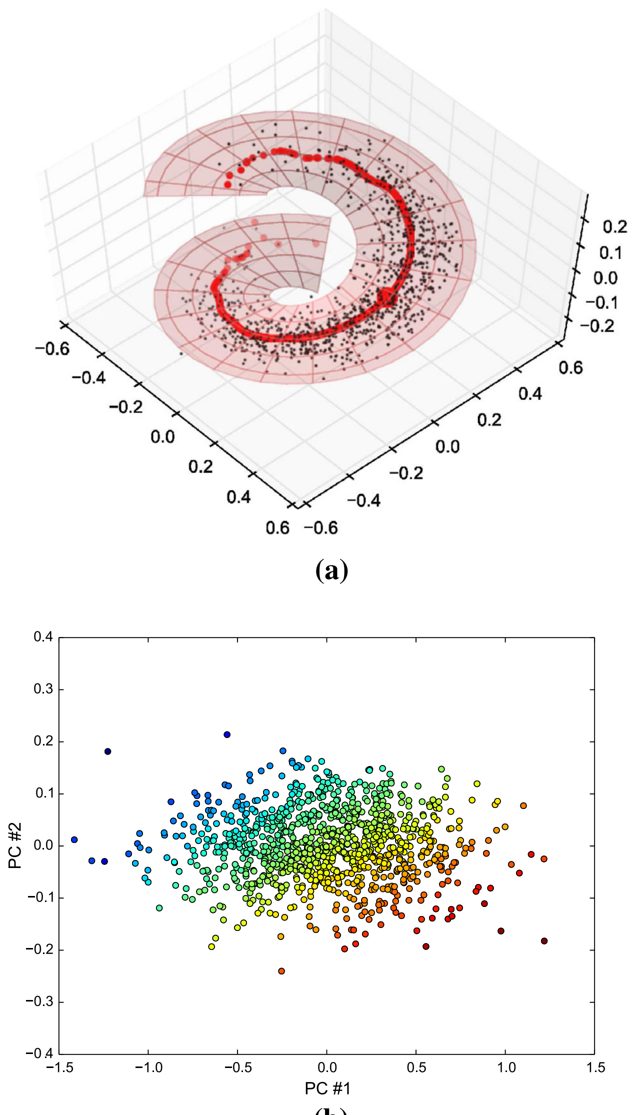

Fig. 5 KDPCA applied to the simulated dataset. **a** 1-d reconstruction, the mode of the kernel density and the surface parametrized by $\boldsymbol{f}$. **b** Scores

normally distributed noise. The model is given by

$$\boldsymbol{X} = \boldsymbol{f}(\boldsymbol{\Phi}) + \boldsymbol{\varepsilon}, \quad \boldsymbol{\Phi} \sim \mathcal{N}_2(\boldsymbol{0}, \boldsymbol{\Sigma}), \quad \boldsymbol{\Sigma} = \begin{pmatrix} 0.2 & 0 \\ 0 & 0.07 \end{pmatrix}^2$$

with

$$\boldsymbol{f}(\boldsymbol{\phi}) = \left((0.35 + \phi_2)\cos(\phi_1), (0.35 + \phi_2)\sin(\phi_1), \frac{0.8}{4\pi}\phi_1 \right)$$

and $\boldsymbol{\varepsilon} \sim \mathcal{N}_2(\boldsymbol{0}, 0.01^2 \boldsymbol{I})$. The point set containing 5000 samples and its coordinates along the surface parametrized by $\boldsymbol{f}$ (scaled by arc length) are plotted in Fig. 4.

The one-dimensional KDPCA reconstruction of the above dataset and the principal component scores are plotted in Fig. 5. The bandwidth matrix $\boldsymbol{H}$ was chosen by using the `Hns` plug-in selector described in Chacón et al. (2011) and implemented in Duong (2007). In this case, the automatic bandwidth selector gives the desired result. Furthermore, the principal component scores give an accurate representation of the manifold coordinates shown in Fig. 4 (the orientation of the coordinate system is adjusted to match that of Fig. 4).

For comparison, the principal component scores obtained with the hierarchical inverse NLPCA described in Scholz et al. (2005, 2008) are shown in Fig. 6. This method is a variant of a neural network-based PCA. Nowadays, such methods have become popular particularly in climate analysis (e.g. Kramer 1991; Monahan 2001; Hsieh 2004). These methods

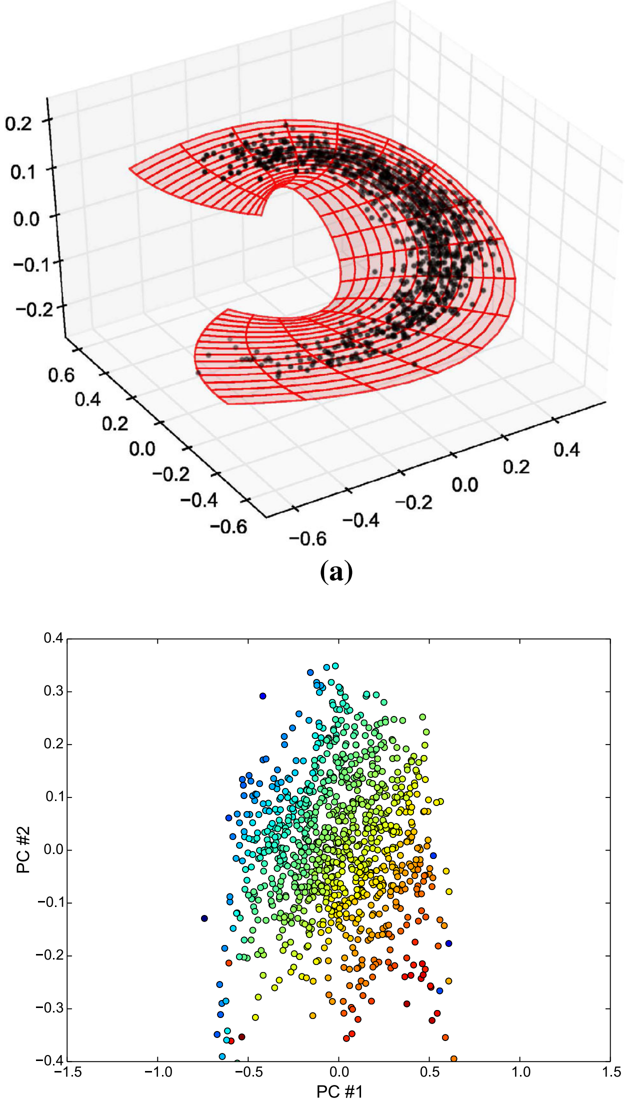

Fig. 6 NLPCA applied to the simulated dataset. **a** Fitted surface. **b** Scores

fit a surface to the given data by using a series of nonlinear mappings. The *weights* determining the shape of the surface are obtained by optimization of a least-squares goodness of fit criterion. Typically, the cost function has a number of local solutions, and thus multiple trial runs are needed to obtain the optimal weights.

Figure 6 shows the best result from five experiments with randomly chosen initial weights. Weight decay coefficient (i.e. regularization parameter) 0.1 was used in order to avoid overfitting. From Fig. 6 we observe that the scores obtained with NLPCA roughly reflect the structure of the dataset. However, there is significant distortion in the scores and the mapping $\boldsymbol{\Psi} : \mathbb{R}^2 \to \mathbb{R}^3$ from the principal component space to the input space. The latter can be seen from Fig. 6 showing a uniform grid in $\mathbb{R}^2$ mapped to $\mathbb{R}^3$ via $\boldsymbol{\Psi}$.

5.3 Application to simulated climate model data

In the second experiment, KDPCA was applied to a simulated sea surface temperature dataset. This dataset is provided by the National Oceanic and Atmospheric Administration (NOAA). The data was obtained from the Coupled Model Intercomparison Project phase 3 (CMIP3) simulations of the GFDL-CM2.1 climate model (Delworth et al. 2006). All preprocessing steps were done as in Ross (2008) and Ross et al. (2008), where this dataset has been analyzed in detail. The preprocessed dataset consisting of 6, 000 samples represents temperature *anomalies* (i.e. data from which seasonality has been removed by subtracting the monthly mean values).

To make estimation of the nonlinear principal components computationally feasible, the high-dimensional data ($d = 10073$, one dimension for each ocean grid point) was first projected onto the first ten principal components obtained by PCA. As these principal components explain 87.3 % of variance, a significant amount of information was not lost by carrying out this preprocessing step.

For the GFDL-CM2.1 dataset, the bandwidth choosers implemented in Duong (2007) failed to yield a unimodal density with connected ridge sets. Therefore the kernel bandwidth was chosen as $\boldsymbol{H} = h^2\boldsymbol{I}$ with $h = 40$. This choice is approximately $\sqrt{\lambda_1}$, where λ_1 denotes the greatest eigenvalue of the matrix $\boldsymbol{H}$ obtained by the `Hns` bandwidth selector.

The dataset and its first principal component obtained from kernel density ridge are plotted in Fig. 7. This figure shows cross-sections of the data and the principal component curve along the first linear principal component axes. Projection of the GFDL-CM2.1 data onto the surface spanned by the first two principal components obtained by Algorithm 1 is shown in Fig. 8. The corresponding principal component scores are plotted in Fig. 9.

Compared to the linear principal component projection shown in Fig. 7, it is clear that the nonlinear principal components represent the "unfolded" dataset and they are better able to capture the variance in the data. Comparison of explained variances of the first eight linear and nonlinear principal components listed in Table 1 also supports this claim. The variance explained by KDPCA is more concentrated towards the first principal component than the variance explained by PCA.[1]

A typical application of principal component analysis (and its nonlinear extensions) is to explain the variance in the given data by some small set of variables. This has been done in Ross (2008) and Ross et al. (2008) for the GFDL-CM2.1 data, and the two main sources of variation were identified. The first principal component correlates with the so-called

[1] The explained variances for the nonlinear principal components were obtained from the covariance matrix of the corresponding scores.

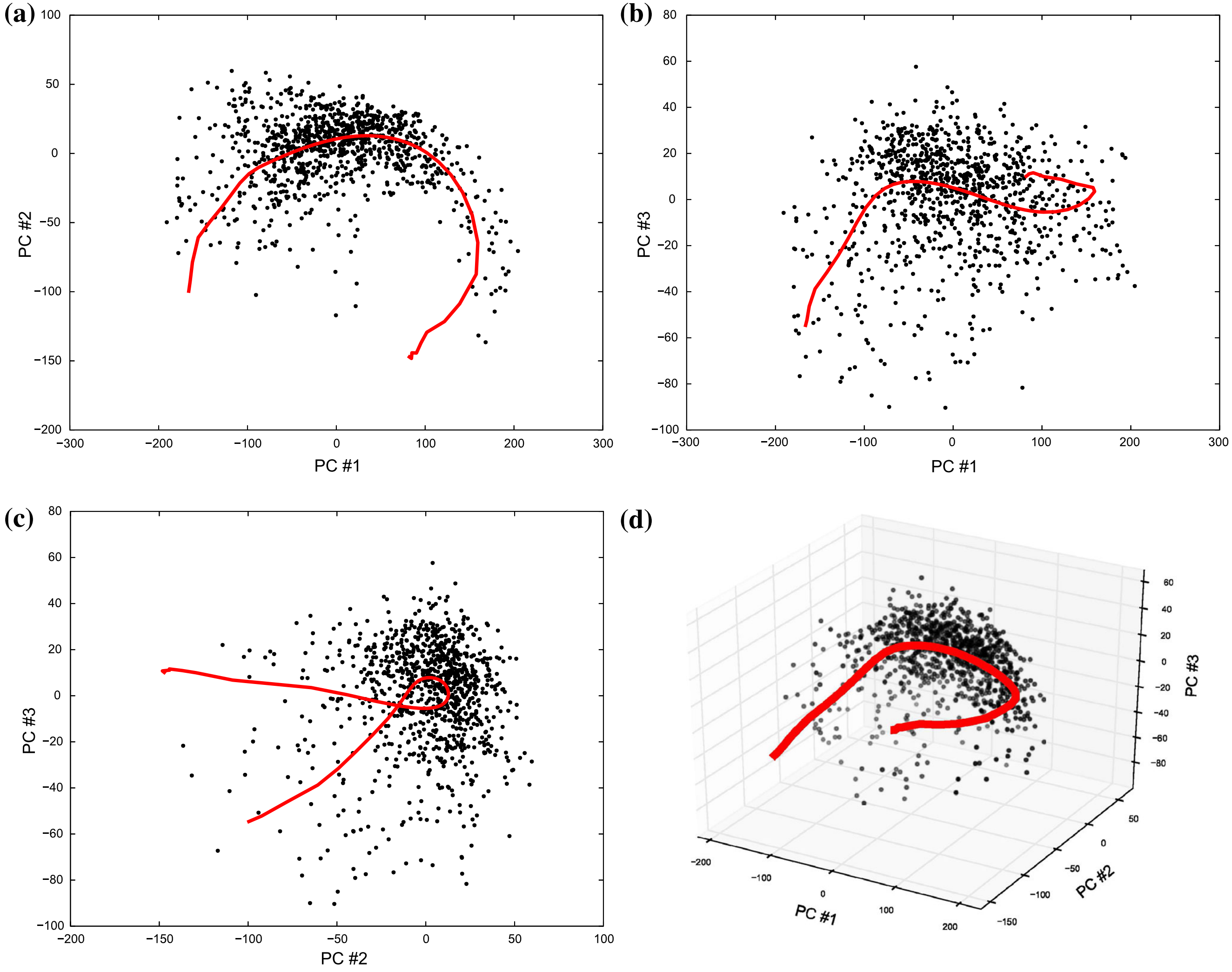

Fig. 7 The first nonlinear principal component obtained from the GFDL-CM2.1 dataset (only a subset of the *curve* is drawn)

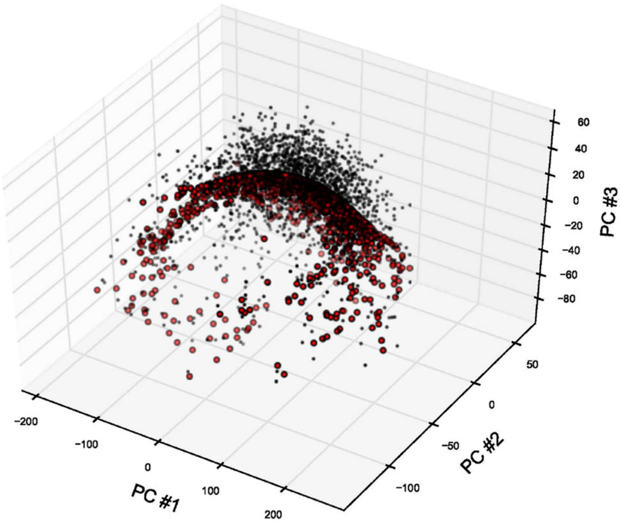

Fig. 8 Projection of the GFDL-CM2.1 dataset onto the surface spanned by its first two nonlinear principal components

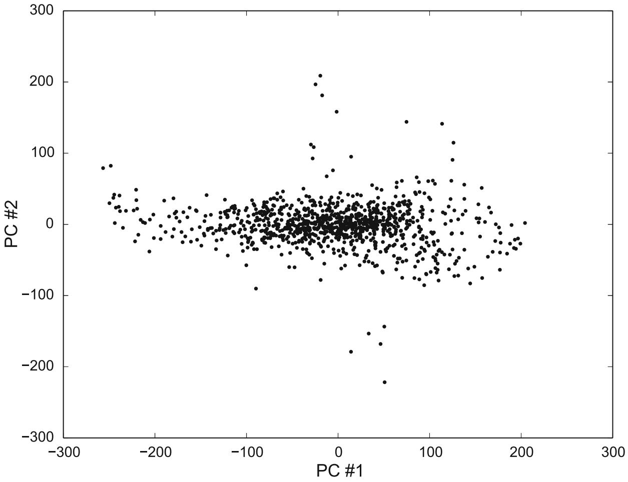

Fig. 9 Two first nonlinear principal component scores obtained from the GFDL-CM2.1 dataset

Table 1 Explained variances of the eight first linear and nonlinear principal components, GFDL-CM2.1 dataset

	1 (%)	2 (%)	3 (%)	4 (%)	5 (%)	6 (%)	7 (%)	8 (%)
PCA	60.0	10.6	6.1	2.5	2.1	1.7	1.3	1.1
KDPCA	66.2	10.4	3.9	2.1	1.4	1.1	0.9	0.8

NINO3 index that is related to the El Niño Southern Oscillation (ENSO) phenomenon. The second one correlates with the Pacific warm water volume. The analysis done here could be carried out further, but we do not attempt repeat the earlier experiments by Ross (2008) and Ross et al. (2008), as using KDPCA would yield similar results than the earlier nonlinear PCA extensions. Of more interest are the differences between KDPCA and the previously proposed methods. A discussion of potential advantages of using KDPCA is given in Sect. 5.6.

5.4 Application to atmospheric time series

The quasi-biennial oscillation (QBO) is one of the most well-studied atmospheric phenomena. The QBO is a quasiperiodic oscillation of the equatorial zonal wind between easterlies and westerlies in the tropical stratosphere with a mean period of 28–29 months. Motivated by an earlier neural network-based nonlinear SSA approach of Hsieh and Hamilton (2003), the nonlinear SSA (KDSSA) described in Sect. 4 was applied to a QBO time series. The time series is provided by the institute of meteorology at the University of Berlin. It consists of monthly mean zonal winds between 1953–2013 constructed from balloon observations at seven different pressure levels corresponding to the altitude range 20–30 km. Here we use a simplified test setup and analyze only the observations from the 30 Hpa level, resulting to a univariate time series.

The trajectory matrix (26) was obtained from the QBO time series with $L = 18$ months. The linear PCA was first applied in the 18-dimensional phase space so that the first four principal components were retained. These principal components explain 95.2 % of the variance, and thus a significant amount of information was not lost by doing this step. The resulting samples were then projected onto the kernel density ridges by using the GTRN algorithm (Pulkkinen et al. 2014). The bandwidth was chosen as $\boldsymbol{H} = h^2\boldsymbol{I}$ with $h = 260$ by the heuristic rule used in Sect. 5.3. The reconstructed time series was obtained by transforming the projected samples from the four-dimensional space back to the 18-dimensional phase space and using the formulae (28).

The trajectory samples and their kernel density ridge projections in the reduced four-dimensional phase space are plotted in Fig. 10. This figure shows a cross-section along the first three linear principal components. Due to the underlying periodic structure present in the time series, its reconstructed phase space trajectory forms a closed loop that passes through the middle of the point cloud. The QBO time series and the reconstructed time series obtained by using the reconstructed phase space trajectory are plotted in Fig. 11. For comparison, the reconstructed time series obtained by using the first linear SSA component and the first and second linear components combined are also plotted in this figure.

The conclusion from Figs. 10 and 11 is that the nonlinear SSA is able to capture the underlying periodic structure in the QBO time series. It is clear that the closed loop found by the nonlinear approach, as shown in Fig. 10, cannot be described by any combination of linear principal components. Consequently, it can be seen from Fig. 11 that the linear SSA reconstruction by using only the first principal component is inadequate to describe the structure of the time

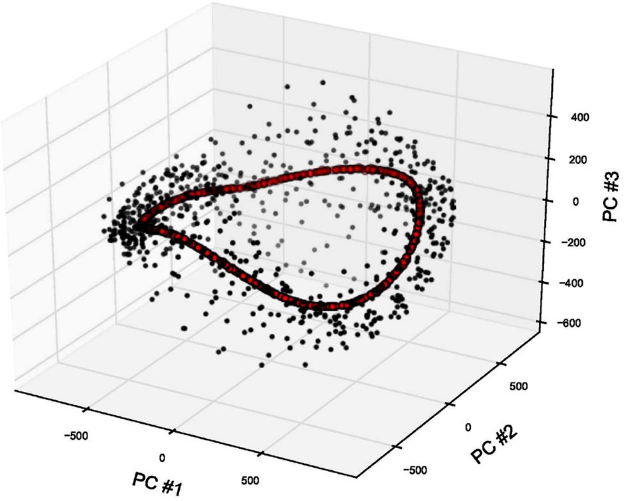

Fig. 10 Phase space trajectory of the QBO time series and the reconstructed trajectory *curve* obtained by kernel density ridge projection

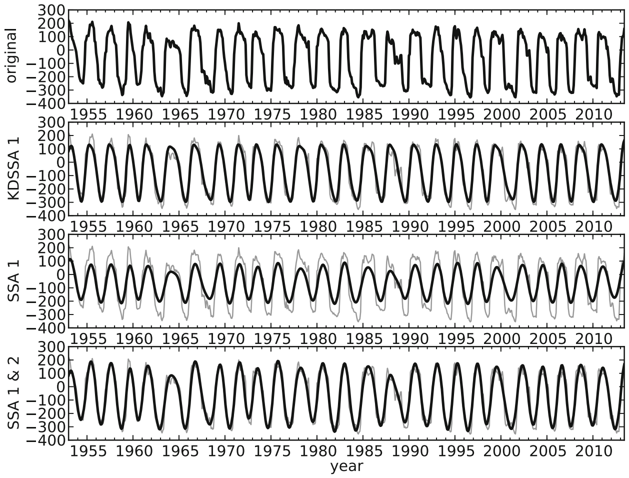

Fig. 11 The QBO time series at the 30 Hpa level and the reconstructed time series obtained by using the first KDSSA component, the first linear SSA component and the two first linear SSA components combined. The original time series is plotted in *gray* in the lower figures

series. On the other hand, by adding more principal components in the analysis, the linear SSA only captures noise and not the underlying periodic pattern.

In Sect. 5.4, the principal component scores (i.e. the coordinates along the nonlinear principal components) were of main interest. Also, in the nonlinear SSA tracking the coordinates of a time series along its reconstructed trajectory curve in the phase space may provide useful information. Namely, when the time series is close to periodic, anomalously short or long cycles can be identified by carrying out such analysis. For the QBO time series, this has been done in Hsieh and Hamilton (2003) by using the neural network-based NLPCA.

Obtaining the coordinates of a time series along its reconstructed phase space trajectory is also possible by using the ridge-based approach. In order to demonstrate this, an approximate parametrization of the trajectory was obtained by using the algorithm developed by Pulkkinen (2015). The projection coordinates were obtained for each sample by finding the nearest point along line segments connecting the trajectory points and computing its distance along the approximate curve to a point fixed as the origin.

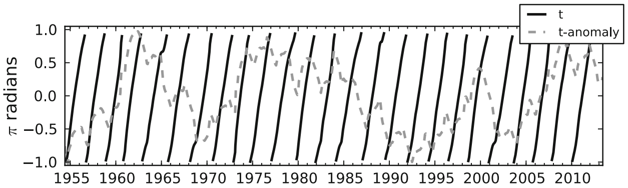

Fig. 12 The first nonlinear principal component coordinate of the QBO time series (t) and the deviation from the fitted regression line (t-anomaly)

Due to its very regular period, the QBO time series progresses along its reconstructed phase space trajectory at a nearly constant rate. This can be seen from Fig. 12 showing the trajectory coordinate t scaled to the interval $[-\pi, \pi]$ as a function of time. In addition, following Hsieh and Hamilton (2003), anomalies (i.e. deviations from the constant rate) were calculated. This was done by fitting a regression line to the t-time series obtained by concatenating the individual cycles and then subtracting the regression line from the concatenated time series. The normalized t-anomaly time series obtained in this way is also plotted in Fig. 12.

Comparison of the t-anomaly time series to the t-time series and Fig. 11 shows its relation to fluctuations from the mean period length. Namely, up- and downward trends in the t-anomalies correspond to abnormally short and long cycles, respectively. This can be seen, for instance, by comparing the long periods during 1962–1969, 1984–1993 and 2000–2002 and the short periods during 1955–1962, 1969–1975 and 2004–2009 to the t-anomaly time series.

5.5 Complexity Analysis

This subsection is devoted to discussion of computational complexity of Algorithm 1 and comparison of KDPCA with existing nonlinear dimensionality reduction methods. After the initial projection step by using the GTRN algorithm (Pulkkinen et al. 2014) having computational cost $\mathcal{O}(n^2d^2 + nd^3)$, the computational cost of Algorithm 1 is $\mathcal{O}(n^2d^3m + d^3nm + n^2dm)$, which is explained in the following paragraphs.

The main source of computational cost is the evaluation of the Gaussian kernel density and its derivatives. For each of the m projection steps, this needs to be done for all n sample points a number of times that depends on the chosen step size τ. For the third derivative that dominates the computational cost, the cost of a single evaluation is $\mathcal{O}(nd^3)$. This makes the total complexity of evaluations $\mathcal{O}(n^2d^3m)$. When d is small, this cost can be reduced by order of n by using the fast Gauss transform or related techniques (Greengard and Strain 1991; Yang et al. 2003).

Computation of the tangent vector in Algorithm 1 and obtaining the trust region step in the corrector involve eigendecomposition of a $d \times d$ matrix (Pulkkinen et al. 2014). The cost of this operation is $\mathcal{O}(d^3)$, and this is done $\mathcal{O}(nm)$ times in the algorithm, making the total cost of eigendecompositions $\mathcal{O}(d^3nm)$. Finally, the cost of traversing the Euclidean minimum spanning tree by using a basic implementation is $\mathcal{O}(n^2d)$. This is done m times in the algorithm, after each projection of all the sample points, and thus the total cost of traversal of such trees is $\mathcal{O}(n^2dm)$.

Computational efficiency of Algorithm 1 can be improved by replacing the projection curve tangent by a Hessian eigenvector (cf. Propositions 3.2 and 3.3). In practice, this leads to slightly worse approximations for the higher-dimensional principal component scores (the first principal component is not affected). This approximation reduces the evaluation cost by order of d since third derivatives are not needed.

When only the first nonlinear principal component (i.e. principal curve) is sought, a significant speedup can be achieved by using a specialized algorithm developed by Pulkkinen (2015). Using this algorithm requires choosing the kernel bandwidth so that the ridge curve set of the density consists of one connected curve. Under this assumption, it suffices to use one starting point, and the total computational cost of tracing the ridge curve is $\mathcal{O}(nd^3)$. The principal component scores can be obtained from projections onto the line segments forming the approximate curve as in Sect. 5.4 at a cost of $\mathcal{O}(ndk)$, where k is the number of line segments.

5.6 Comparison to other methods

The neural network-based methods are among the most popular nonlinear PCA methods applied in climate analysis. However, they have several shortcomings. Some of them are discussed below with comparison to KDPCA.

- NLPCA involves minimization of a cost function that generally has several local minima. This problem is typically addressed by using multiple starting points, which may incur a high computational cost. KDPCA is not affected by this issue because it does not attempt to minimize a single cost function. Instead, it performs local maximizations from each sample point. The projection curves are uniquely defined when the ridge sets are connected.
- The principal components obtained by KDPCA have a statistical interpretation. This is not the case for NLPCA that is based on an artificially constructed neural network. In fact, the NLPCA principal curves and surfaces are not guaranteed to follow regions of high concentration of the data points. Examples of this are given by Christiansen (2005). Due to this issue, drawing statistical inferences from the NLPCA output should be done with extreme caution.
- NLPCA uses artificial penalty terms to avoid overfitting. Despite this, the density of the data along the first nonlinear principal component can exhibit spurious multimodality (Christiansen 2005). This can occur even when the underlying density of the data is close to normal. On the other hand, KDPCA performs no worse than the linear PCA when the kernel bandwidth is chosen sufficiently large.
- When using NLPCA, the type of a curve (open or closed) to be fitted to the data needs to be specified a priori in the neural network structure. KDPCA can determine this automatically when the principal curve is traced by using the algorithm developed by Pulkkinen (2015).
- The curves fitted by NLPCA are not parametrized by arc length. This may introduce a significant bias to reconstructions and principal component scores. When drawing statistical inferences from a curve fitted by NLPCA, arc length reparametrization should be done to remove the bias (Newbigging et al. 2003). However, this approach has not been generalized to higher dimensions. On the other hand, KDPCA produces an arc length parametrization for principal component curves and surfaces of any dimension due to its construction.

KDPCA has also certain advantages compared to other commonly used nonlinear dimensionality reduction methods. This is because it seems to perform well in the presence of noise and it operates directly in the input space.

- Graph-based methods such as Isomap (Tenenbaum et al. 2000), Hessian eigenmaps (Donoho and Grimes 2003) and maximum variance unfolding (Weinberger and Saul 2006) are based on the assumption that the data lies directly on a low-dimensional manifold. Thus, they are sensitive to noise, and blindly applying such methods to noisy data may lead to undesired results.
- The aforementioned methods and kernel-based methods such as KPCA do not produce a reconstruction of the data in the original input space. This would be a very desired feature, for instance, in climate analysis where plots of reconstructed grid data or time series provide information about the main sources of variation.

The main difficulty in KDPCA is the choice of kernel bandwith $\boldsymbol{H}$. When the data follows the model described in Sect. 3.1, the number of samples is sufficiently large and the data dimension is small, an automatic bandwidth selector can be used. This was successfully done in Sect. 5.2. Unfortunately, the above requirements might be unrealistic for practical applications, as observed in Sects. 5.3 and 5.4. As a result, the kernel density may become multimodal or have disconnected ridge sets. In such cases, the bandwidth parametrization $\boldsymbol{H} = h^2\boldsymbol{I}$ can be used and a sufficiently large h can be chosen by visual inspection of principal components in two or three dimensions (cf. Figs. 7, 8 and 10). As shown in Sects. 5.3 and 5.4, a plug-in bandwidth estimate can be utilized for obtaining a first guess.

6 Conclusions and discussion

Principal component analysis (PCA) is a well-established tool for exploratory data analysis. However, as a linear method it cannot describe complex nonlinear structure in the given data. To address this deficiency, a novel nonlinear generalization of the linear PCA was developed in this paper.

The proposed KDPCA method is based on the idea of using ridges of the underlying density of the data to estimate nonlinear structures. It was shown that the principal component coordinates of a given point set can be obtained one by one by successively projecting the points onto lower-dimensional ridge sets of the density. Such a projection was defined as a solution to a differential equation. A predictor-corrector method using a Newton-based corrector was developed for this purpose.

Gaussian kernels were used for estimation of the density from the data. This choice has several advantages. First, this choice allows automatic estimation of an appropriate bandwidth from the data. This was demonstrated by numerical experiments, although the currently available bandwidth estimation methods have only limited applicability. Second, a fundamental result was derived showing that by choosing the bandwidth as $\boldsymbol{H} = h^2\boldsymbol{I}$ and letting h approach infinity, KDPCA reduces to the linear PCA when desired.Third, the theory of ridge sets ensures that any ridge set of a Gaussian kernel density has a well-defined coordinate system when h is sufficiently large.

Based on the linear singular spectrum analysis (SSA), KDPCA was extended to time series data. It was shown that when a time series is (quasi)periodic, the first nonlinear principal component of its phase space representation can be used to reconstruct the underlying periodic pattern from noise. Though the periodicity assumption is restrictive, such time series are relevant for many practical applications. Examples include climate analysis and medical applications such as electrocardiography and electroencephalography.

The proposed KDPCA method and its SSA-based variant were applied to a highly nonlinear dataset obtained from a climate model and to an atmospheric time series. The method is superior to the linear PCA in capturing the complex nonlinear structure of the data. It also has several advantages compared to the existing nonlinear dimensionality reduction methods. In particular, KDPCA requires only one parameter, that is, the kernel bandwidth $\boldsymbol{H}$. When parametrized as $\boldsymbol{H} = h^2\boldsymbol{I}$, the bandwidth has an intuitive interpretation as a scale parameter.

While KDPCA showed convincing results on the test datasets, its applicability to real-world data remains to be fully confirmed. When the data is noisy and sparse, which is typical for observational data, the additional information obtained by KDPCA might not justify its high computational cost. However, using the techniques discussed in Sect. 5.5 could significantly improve the scalability of KDPCA to large datasets. Computational difficulties due to high dimensionality of the data can also be circumvented. In many situations, the variance is confined to some low-dimensional subspace that can be identified by using a simpler method as a preprocessing step (as in Sects. 5.3 and 5.4).

Acknowledgments The author was financially supported by the TUCS Graduate Programme, the Academy of Finland grant No. 255718 and the Finnish Funding Agency for Technology and Innovation (Tekes) grant No. 3155/31/2009. He would also like to thank Prof. Marko Mäkelä, Doc. Napsu Karmitsa and Prof. Hannu Oja (University of Turku) and the reviewers for their valuable comments.

Appendix

In this appendix we give proofs of Theorems 3.2 and 3.3. In the following, we assume that the given set of sample points $\boldsymbol{Y} = \{\boldsymbol{y}_i\}_{i=1}^N \subset \mathbb{R}^d$ is fixed. The proofs are carried out by making the following simplifying assumption that can be made without loss of generality.

Assumption 7.1 The points $\boldsymbol{y}_i$ satisfy the condition $\sum_{i=1}^n \boldsymbol{y}_i = \boldsymbol{0}$.

First, we recall the density estimate defined by equations (20) and (21) with the bandwidth parametrization $\boldsymbol{H} = h^2\boldsymbol{I}$. That is,

$$\hat{p}_h(\boldsymbol{x}) = \frac{1}{(2\pi)^{\frac{d}{2}} h^d n} \sum_{i=1}^n \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{y}_i\|^2}{2h^2}\right). \tag{29}$$

The following limits hold for the logarithm of the Gaussian kernel density estimate and its derivatives as h approaches infinity. Uniform convergence in a given compact set U can be verified by showing that the functions are uniformly bounded (i.e. bounded with respect to all $\boldsymbol{x} \in U$ and all $h \geq h_0$ for some $h_0 > 0$) and that they have a uniform Lipschitz constant in U for all $h \geq h_0$. Under these conditions, this follows from the Arzelà-Ascoli theorem (e.g. Renardy and Rogers 2004). The proof of the following lemma is omitted due to space constraints.

Lemma 7.1 *Let $\hat{p}_h : \mathbb{R}^d \to \mathbb{R}$ be a Gaussian kernel density estimate and assume* 7.1. *Then*

$$\lim_{h\to\infty} h^2 \log\left[(2\pi)^{\frac{d}{2}} h^d \hat{p}_h(\boldsymbol{x})\right] = -\frac{1}{2n}\sum_{i=1}^{n} \|\boldsymbol{x} - \boldsymbol{y}_i\|^2, \quad (30)$$

$$\lim_{h\to\infty} h^2 \nabla \log \hat{p}_h(\boldsymbol{x}) = -\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x} - \boldsymbol{y}_i) = -\boldsymbol{x}, \quad (31)$$

$$\lim_{h\to\infty} \left[h^4 \nabla^2 \log \hat{p}_h(\boldsymbol{x}) + h^2 \boldsymbol{I}\right] = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{y}_i \boldsymbol{y}_i^T \quad (32)$$

for all $\boldsymbol{x} \in \mathbb{R}^d$. Furthermore, convergence to these limits is uniform in any compact set.

The following two lemmata facilitate the proof of Theorem 3.3.

Lemma 7.2 *Let $\hat{p}_h : \mathbb{R}^d \to \mathbb{R}$ be a Gaussian kernel density estimate and let Assumptions* 3.2 *and* 7.1 *be satisfied. Denote the eigenvalues of $\nabla^2 \log \hat{p}_h$ by $\lambda_1(\cdot; h) \geq \lambda_2(\cdot; h) \geq \cdots \geq \lambda_d(\cdot; h)$ and the corresponding eigenvectors by $\{\boldsymbol{w}_i(\cdot; h)\}_{i=1}^d$. Then for any compact set $U \subset \mathbb{R}^d$ there exists $h_0 > 0$ such that*

$$\lambda_1(\boldsymbol{x}; h) < 0, \quad (33)$$

$$\lambda_i(\boldsymbol{x}; h) \neq \lambda_j(\boldsymbol{x}; h) \quad (34)$$

for all $\boldsymbol{x} \in U, h \geq h_0$ and $i, j = 1, 2, \ldots, r+1$ such that $i \neq j$. Furthermore, if we define

$$\boldsymbol{W}(\boldsymbol{x}; h) = [\boldsymbol{w}_1(\boldsymbol{x}; h) \quad \boldsymbol{w}_2(\boldsymbol{x}; h) \quad \cdots \quad \boldsymbol{w}_r(\boldsymbol{x}; h)]$$

and

$$\boldsymbol{V} = [\boldsymbol{v}_1 \quad \boldsymbol{v}_2 \quad \cdots \quad \boldsymbol{v}_r],$$

where $\{\boldsymbol{v}_i\}_{i=1}^r$ denote the eigenvectors of the matrix $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{Y}}$ defined by Eq. (3) *corresponding to its r greatest eigenvalues, then for all $\varepsilon > 0$ there exists $h_0 > 0$ such that*

$$\|\boldsymbol{W}(\boldsymbol{x}; h) - \boldsymbol{V}\| < \varepsilon \quad \textit{for all } \boldsymbol{x} \in U \textit{ and } h \geq h_0. \quad (35)$$

Proof Let $\tilde{\lambda}_1 \geq \tilde{\lambda}_2 \geq \cdots \geq \tilde{\lambda}_d$ denote the eigenvalues of the matrix $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{Y}}$ and let $\{h_k\}$ be some sequence such that $\lim_{k\to\infty} h_k = \infty$. By uniform convergence to the limit (32) under Assumption 7.1 and continuity of eigenvalues of a matrix as a function of its elements (e.g. Ortega 1990 Theorem 3.1.2), for all $\varepsilon > 0$ there exists k_0 such that

$$\left|h_k^4 \lambda_i(\boldsymbol{x}; h_k) + h_k^2 - \frac{n-1}{n}\tilde{\lambda}_i\right| < \varepsilon \quad (36)$$

for all $i = 1, 2, \ldots, r+1$, $\boldsymbol{x} \in U$ and $k \geq k_0$. Consequently, condition (33) holds for all $\boldsymbol{x} \in U$ for any sufficiently large h by Assumption 3.2. It also follows from Assumption 3.2, condition (36) and the reverse triangle inequality that for all $\varepsilon > 0$ and $i, j = 1, 2, \ldots, r+1$ such that $i \neq j$ and $|\tilde{\lambda}_i - \tilde{\lambda}_j| > \varepsilon$ there exists k_1 such that

$$\begin{aligned}
&h_k^4 |\lambda_i(\boldsymbol{x}; h_k) - \lambda_j(\boldsymbol{x}; h_k)| \\
&\geq \left| |h_k^4 \lambda_i(\boldsymbol{x}; h_k) - h_k^2| - |h_k^4 \lambda_j(\boldsymbol{x}; h_k) - h_k^2| \right| > \frac{n-1}{n}\varepsilon
\end{aligned}$$

for all $\boldsymbol{x} \in U$ and $k \geq k_1$. This implies condition (34). Similarly, condition (35) follows from uniform convergence to the limit (32) under Assumption 7.1, condition (34) and continuity of eigenvectors as a function of matrix elements when the eigenvalues are distinct (e.g. Ortega 1990 Theorem 3.1.3). □

Lemma 7.3 *Assume* 3.2 *and* 7.1 *and define the function*

$$\tilde{\boldsymbol{W}}(\boldsymbol{x}; h) = \boldsymbol{I} - \boldsymbol{W}(\boldsymbol{x}; h)\boldsymbol{W}(\boldsymbol{x}; h)^T,$$

where the function $\boldsymbol{W}$ is defined as in Lemma 7.2, and the set S_∞^r as in Theorem 3.3. Then the limit

$$\lim_{h\to\infty} h^2 \|\tilde{\boldsymbol{W}}(\boldsymbol{x}; h)\nabla \log \hat{p}_h(\boldsymbol{x})\| \quad (37)$$

exists for all $\boldsymbol{x} \in \mathbb{R}^d$. Furthermore, $\boldsymbol{x} \in S_\infty^r$ if and only if the limit (37) *is zero.*

Proof By the limits (31) and (35) the limit (37) exists for all $\boldsymbol{x} \in \mathbb{R}^d$. Furthermore, for any $\boldsymbol{x} \in \mathbb{R}^d$, the condition that the limit (37) is zero is equivalent to the condition that

$$\boldsymbol{v}_i^T \boldsymbol{x} = 0 \quad \text{for all } i = r+1, r+2, \ldots, d,$$

where the vectors $\boldsymbol{v}_i$ are defined as in Lemma 7.2. By the orthogonality of the vectors $\boldsymbol{v}_i$, the definition of the set S_∞^r and Assumption 7.1, this condition is equivalent to the condition that $\boldsymbol{x} \in S_\infty^r$. □

For the proof of Theorem 3.3, we define the set

$$S_h^r = \{\boldsymbol{x} \in \mathbb{R}^d \mid \|\tilde{\boldsymbol{W}}(\boldsymbol{x}; h)\nabla \log \hat{p}_h(\boldsymbol{x})\| = 0\}, \quad (38)$$

where the function $\tilde{\boldsymbol{W}}$ is defined as in Lemma 7.3. Under Assumption 7.1, we prove both claims of Theorem 3.3 by the following two lemmata.

Lemma 7.1 *Let $U \subset \mathbb{R}^d$ be a compact set such that $U \cap S_\infty^r \neq \emptyset$ for some $0 \leq r < d$. If Assumptions* 3.2 *and* 7.1 *are satisfied, then for all $\varepsilon > 0$ there exists $h_0 > 0$ such that*

$$\sup_{\boldsymbol{x} \in S_h^r \cap U} \inf_{\boldsymbol{y} \in S_\infty^r} \|\boldsymbol{x} - \boldsymbol{y}\| < \varepsilon \quad \textit{for all } h \geq h_0. \quad (39)$$

Proof The proof is by contradiction. Let $0 \leq r < d$ and let $U \subset \mathbb{R}^d$ be a compact set such that $U \cap S^r_\infty \neq \emptyset$. Assume that there exists $\varepsilon_1 > 0$ such that for all $h_0 > 0$ there exists $h \geq h_0$ such that condition (39) is not satisfied. This implies that for all $h_0 > 0$ there exists $h \geq h_0$ such that

$$\inf_{\boldsymbol{y} \in S^r_\infty} \|\boldsymbol{x} - \boldsymbol{y}\| \geq \varepsilon_1 \quad \text{for some } \boldsymbol{x} \in S^r_h \cap U. \tag{40}$$

Let $\{\boldsymbol{x}_k\}$ denote a sequence of such points $\boldsymbol{x}$ with the corresponding sequence h_k. Since the set $S^r_h \cap U$ is compact by the compactness of U and the continuity of $\widetilde{\boldsymbol{W}}(\cdot, h)$ in U for any sufficiently large h, the sequence $\{\boldsymbol{x}_k\}$ has a convergent subsequence $\{\boldsymbol{z}_k\}$ whose limit point we shall denote as $\boldsymbol{z}^*$. Clearly $\boldsymbol{z}^* \notin S^r_\infty$ by condition (40). Thus, by Lemma 7.3 we deduce that for some $c > 0$,

$$\lim_{k \to \infty} \|\boldsymbol{F}(\boldsymbol{z}^*; h_k)\| = c, \quad \boldsymbol{F}(\boldsymbol{x}; h) = h^2 \tilde{\boldsymbol{W}}(\boldsymbol{x}; h) \nabla \log \hat{p}_h(\boldsymbol{x}).$$

In view of the definition (38), the above limit implies that there exists $\varepsilon_2 > 0$ and k_0 such that for all $k \geq k_0$,

$$\|\boldsymbol{F}(\boldsymbol{z}^*; h_k) - \boldsymbol{F}(\boldsymbol{y}; h_k)\| \geq \varepsilon_2 \quad \text{for all } \boldsymbol{y} \in S^r_{h_k} \cap U. \tag{41}$$

On the other hand, if we define the function $\boldsymbol{F}^*(\boldsymbol{x}) = -(\boldsymbol{I} - \boldsymbol{V}\boldsymbol{V}^T)\boldsymbol{x}$, the triangle inequality yields

$$\begin{aligned}
&\|\boldsymbol{F}(\boldsymbol{z}^*; h_k) - \boldsymbol{F}(\boldsymbol{y}; h_k)\| \\
&\quad \leq \|\boldsymbol{F}(\boldsymbol{z}^*; h_k) - \boldsymbol{F}^*(\boldsymbol{z}^*)\| + \|\boldsymbol{F}(\boldsymbol{y}; h_k) - \boldsymbol{F}^*(\boldsymbol{z}^*)\| \\
&\quad \leq \|\boldsymbol{F}(\boldsymbol{z}^*; h_k) - \boldsymbol{F}^*(\boldsymbol{z}^*)\| + \|\boldsymbol{F}(\boldsymbol{y}; h_k) - \boldsymbol{F}^*(\boldsymbol{y})\| \\
&\qquad + \|\boldsymbol{F}^*(\boldsymbol{y}) - \boldsymbol{F}^*(\boldsymbol{z}^*)\|.
\end{aligned}$$

Combining this with the inequality

$$\|\boldsymbol{F}^*(\boldsymbol{y}) - \boldsymbol{F}^*(\boldsymbol{z}^*)\| \leq \|\boldsymbol{I} - \boldsymbol{V}\boldsymbol{V}^T\| \|\boldsymbol{y} - \boldsymbol{z}^*\| = \|\boldsymbol{y} - \boldsymbol{z}^*\|$$

and noting the convergence of $\boldsymbol{F}(\cdot; h_k)$ to the function $\boldsymbol{F}^*$ (that is uniform in U) as $k \to \infty$ (by Lemmata 7.1 and 7.2), we deduce from (41) that for all $\varepsilon_2 > \varepsilon_3 > 0$ there exists k_1 such that

$$\|\boldsymbol{z}^* - \boldsymbol{y}\| + \varepsilon_3 \geq \|\boldsymbol{F}(\boldsymbol{z}^*; h_k) - \boldsymbol{F}(\boldsymbol{y}; h_k)\| \geq \varepsilon_2 \tag{42}$$

for all $\boldsymbol{y} \in S^r_{h_k} \cap U$ and $k \geq k_1$.

Condition (42) implies that for all $0 < \varepsilon_3 < \varepsilon_2$ there exists k_1 such that

$$\inf_{\boldsymbol{y} \in S^r_{h_k} \cap U} \|\boldsymbol{z}^* - \boldsymbol{y}\| \geq \varepsilon_2 - \varepsilon_3 \quad \text{for all } k \geq k_1. \tag{43}$$

On the other hand, for all $\varepsilon > 0$ we have $\boldsymbol{z}_k \in B(\boldsymbol{z}^*; \varepsilon)$ for any sufficiently large k due to the assumption that $\boldsymbol{z}_k$ converges to $\boldsymbol{z}^*$. If we choose $0 < \varepsilon < \varepsilon_2$, then the sequence $\{\boldsymbol{x}_k\}$, whose subsequence is $\{\boldsymbol{z}_k\}$, has an element $\boldsymbol{x}_k \notin S^r_{h_k} \cap U$ for some k by condition (43). This leads to a contradiction with the construction of the sequence $\{\boldsymbol{x}_k\}$, which states that $\boldsymbol{x}_k \in S^r_{h_k} \cap U$ for all k. □

Lemma 7.5 *Let $\hat{p}_h$ be a Gaussian kernel density estimate, let $0 \leq r < d$, let Assumptions 3.2 and 7.1 be satisfied and define the set S^r_∞ as in Theorem 3.3. Then for any compact set $U \subset \mathbb{R}^d$ such that $U \cap S^r_\infty \neq \emptyset$ and $\varepsilon > 0$ there exists $h_0 > 0$ such that*

$$\sup_{\boldsymbol{x} \in S^r_\infty \cap U} \inf_{\boldsymbol{y} \in \mathcal{R}^r_{\log \hat{p}_h}} \|\boldsymbol{x} - \boldsymbol{y}\| < \varepsilon \quad \textit{for all } h \geq h_0. \tag{44}$$

Proof Let $0 \leq r < d$ and let $\{\boldsymbol{v}_i\}_{i=r+1}^d$ denote a set of orthonormal eigenvectors of the matrix $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{Y}}$ corresponding to the $d - r$ smallest eigenvalues. The vectors $\{\boldsymbol{v}_i\}_{i=r+1}^d$ are uniquely determined up to the choice of basis because the eigenvectors $\{\boldsymbol{v}_i\}_{i=1}^r$ spanning their orthogonal complement are uniquely determined by Assumption 3.2. Define the sets

$$D_{\boldsymbol{x},\varepsilon} = \left\{ \boldsymbol{x} + \sum_{i=r+1}^{d} \alpha_{i-r} \boldsymbol{v}_i \mid \sum_{i=1}^{r} \alpha_i^2 \leq \varepsilon^2 \right\}$$

and

$$D_\varepsilon = \bigcup_{\boldsymbol{x} \in S^r_\infty \cap U} D_{\boldsymbol{x},\varepsilon}$$

for some orthonormal eigenvectors $\{\boldsymbol{v}_i\}_{i=r+1}^d$ spanning the orthogonal complement of $\operatorname{span}(\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_r)$.

Let $\{\boldsymbol{u}_i(\cdot; h)\}_{i=1}^d$ denote a set of orthonormal vectors that are orthogonal to the eigenvectors $\{\boldsymbol{w}_i(\cdot; h)\}_{i=1}^r$ of $\nabla^2 \log \hat{p}_h$ corresponding to the r greatest eigenvalues. Define the functions

$$\boldsymbol{F}(\boldsymbol{x}; h) = h^2 \boldsymbol{U}(\boldsymbol{x}; h)^T \nabla \log \hat{p}_h(\boldsymbol{x})$$

and

$$\tilde{\boldsymbol{F}}_{\boldsymbol{x}_0}(\boldsymbol{y}; h) = h^2 \boldsymbol{U}(\bar{\boldsymbol{V}}\boldsymbol{y} + \boldsymbol{x}_0; h)^T \nabla \log \hat{p}_h(\bar{\boldsymbol{V}}\boldsymbol{y} + \boldsymbol{x}_0),$$

where

$$\boldsymbol{U}(\boldsymbol{x}; h) = [\boldsymbol{u}_1(\boldsymbol{x}; h) \quad \boldsymbol{u}_2(\boldsymbol{x}; h) \quad \cdots \quad \boldsymbol{u}_{d-r}(\boldsymbol{x}; h)]$$

and $\bar{\boldsymbol{V}} = [\boldsymbol{v}_{r+1} \quad \boldsymbol{v}_{r+2} \quad \cdots \quad \boldsymbol{v}_d]$ assuming that the orientation is chosen so that $\det(\bar{\boldsymbol{V}}) = 1$. To fix the orientation of the vectors $\boldsymbol{u}_i(\boldsymbol{x}; h)$, we impose the constraint

$$\boldsymbol{U}(\boldsymbol{x}; h) = \arg \min_{\boldsymbol{U}' \in Q_{\boldsymbol{x},h}} \|\boldsymbol{U}' - \bar{\boldsymbol{V}}\|_F. \tag{45}$$

Here $\|\cdot\|_F$ denotes the Frobenius norm,

$$Q_{\boldsymbol{x},h} = \{\boldsymbol{U}' \in O(d, d-r) \mid \boldsymbol{U}'^T \boldsymbol{W}(\boldsymbol{x}; h) = \boldsymbol{0},\\ \det(\boldsymbol{U}') = 1)\},$$

$O(d, d-r)$ denotes a $d \times (d-r)$ matrix having orthonormal columns and the matrix $\boldsymbol{W}(\boldsymbol{x}; h)$ is defined as in Lemma 7.2. It can be shown that the function $\boldsymbol{U}(\cdot; h)$ is well-defined for any $h > 0$.[2] Spanning the orthogonal complement of the columns of $\boldsymbol{W}(\cdot; h)$, the columns of $\boldsymbol{U}(\cdot; h)$ are also continuous in a given compact set when $\boldsymbol{W}(\cdot; h)$ is continuous. That is, when condition (34) is satisfied in such a set by Lemma 7.2.

The above definitions and condition (35) in the compact set D_ε imply that for all $\varepsilon_1, \varepsilon_2 > 0$ there exists $h_0 > 0$ such that

$$\|\boldsymbol{U}(\boldsymbol{x}; h) - \bar{\boldsymbol{V}}\| < \varepsilon_2 \quad \text{for all } \boldsymbol{x} \in D_{\varepsilon_1} \text{ and } h \geq h_0.$$

Consequently, uniform convergence to the limit (31) as $h \to \infty$ by Lemma 7.3 together with the property that

$$\bar{\boldsymbol{V}}^T(\bar{\boldsymbol{V}}\boldsymbol{y} + \boldsymbol{x}_0) = \boldsymbol{y} \quad \text{for all } \boldsymbol{y} \in \mathbb{R}^{d-r} \text{ and } \boldsymbol{x}_0 \in S_\infty^r$$

following from Assumption 7.1 implies that for all $\varepsilon_1, \varepsilon_2 > 0$ there exists $h_0 > 0$ such that

$$\|\tilde{\boldsymbol{F}}_{\boldsymbol{x}_0}(\boldsymbol{y}; h) - (-\boldsymbol{y})\| < \varepsilon_2$$
$$\text{for all } \boldsymbol{x}_0 \in S_\infty^r \cap U, \boldsymbol{y} \in \tilde{D}_{\varepsilon_1} \text{ and } h \geq h_0,$$

where $\tilde{D}_\varepsilon = \{\boldsymbol{y} \in \mathbb{R}^{d-r} \mid \|\boldsymbol{y}\| \leq \varepsilon\}$.

By the above condition, for any $0 < \varepsilon_2 < \varepsilon_1$ there exists $h_0 > 0$ such that for all $h \geq h_0$ and $\boldsymbol{x}_0 \in S_\infty^r \cap U$ we have $-\tilde{\boldsymbol{F}}_{\boldsymbol{x}_0}(\boldsymbol{y}; h)^T \boldsymbol{y} > 0$ for all $\boldsymbol{y} \in \partial \tilde{D}_{\varepsilon_1}$, where ∂ denotes the boundary of a set. On the other hand, $-\boldsymbol{y}$ is the inward-pointing normal vector of the disk $\tilde{D}_{\varepsilon_1}$ at any $\boldsymbol{y} \in \partial \tilde{D}_{\varepsilon_1}$. Together with the continuity of $\tilde{\boldsymbol{F}}_{\boldsymbol{x}_0}(\cdot; h)$ in $\tilde{D}_{\varepsilon_1}$ when h is sufficiently large, the well-known results from topology (e.g. Whittlesey 1963) then imply that $\tilde{\boldsymbol{F}}_{\boldsymbol{x}_0}(\cdot; h)$ has at least one zero point $\boldsymbol{y}^*$ in the interior of $\tilde{D}_{\varepsilon_1}$ for all $\boldsymbol{x}_0 \in S_\infty^r \cap U$ and $h \geq h_0$. Clearly, for any such $\boldsymbol{y}^*$ and $\boldsymbol{x}_0$ the point $\boldsymbol{x}^* = \bar{\boldsymbol{V}}\boldsymbol{y}^* + \boldsymbol{x}_0$ lies in the set $D_{\boldsymbol{x}_0,\varepsilon}$ and $\boldsymbol{F}(\boldsymbol{x}^*; h) = \tilde{\boldsymbol{F}}_{\boldsymbol{x}_0}(\boldsymbol{y}^*; h) = \boldsymbol{0}$.

From the above we conclude that for all $\varepsilon > 0$ there exists $h_0 > 0$ such that for all $\boldsymbol{x}_0 \in S_\infty^r \cap U$ condition (6a) holds for $\log \hat{p}_h$ at least at one point in $D_{\boldsymbol{x}_0,\varepsilon}$ for all $h \geq h_0$. On the other hand, for all $\varepsilon > 0$ conditions (6b) and (6c) are satisfied in the compact set D_ε for all sufficiently large h by conditions (33) and (34). Hence, we have proven that for all $\varepsilon > 0$ condition (44) holds for all sufficiently large h. □

Proof *(Theorem 3.3)* Follows directly from Lemmata 7.4 and 7.5 by the property that $\mathcal{R}_{\hat{p}_h}^r = \mathcal{R}_{\log \hat{p}_h}^r \subseteq S_h^r$ for all $0 \leq r < d$ and $h > 0$ by Lemma 3.1 and Definition 3.1. □

Next, we prove Theorem 3.2 under Assumption 7.1 by using the following lemma.

Lemma 7.6 *Let $\hat{p}_h$ be a Gaussian kernel density estimate, assume 7.1 and define the set*

$$U_h = \bigcup_{i=1}^n \{\boldsymbol{x} \in \mathbb{R}^d \mid \log \hat{p}_h(\boldsymbol{x}) \geq \log \hat{p}_h(\boldsymbol{y}_i)\}.$$

Then for some $r > \max_{i=1,2,\ldots,n} \|\boldsymbol{y}_i\|$ there exists $h_0 > 0$ such that $U_h \subseteq B(\boldsymbol{0}; r)$ for all $h \geq h_0$.

Proof The proof is by contradiction. Assume that for all $r > r_0 = \max_{i=1,2,\ldots,n} \|\boldsymbol{y}_i\|$ and $h_0 > 0$ there exists $h \geq h_0$ such that $\boldsymbol{x} \in U_h \setminus B(\boldsymbol{0}; r)$. Let $\{\boldsymbol{x}_k\}$, $\{r_k\}$ and $\{h_k\}$ denote sequences satisfying these properties such that $\{r_k\}$ and $\{h_k\}$ are monotoneusly increasing. This implies that

$$\|\boldsymbol{x}_k\| > r_k > r_0 = \max_{i=1,2,\ldots,n} \|\boldsymbol{y}_i\| \quad \text{for all } k \geq k_0 \tag{46}$$

and also that for all $k \geq k_0$,

$$\log \hat{p}_{h_k}(\boldsymbol{x}_k) \geq \log \hat{p}_{h_k}(\boldsymbol{y}_j) \quad \text{for some } j \in \{1, 2, \ldots, n\}. \tag{47}$$

By Assumption 7.1 and condition (46) we have that $\|\boldsymbol{x}_k - \boldsymbol{y}_i\| \geq \|\boldsymbol{x}_k\| - r_0$ for all $k \geq k_0$ and $i = 1, 2, \ldots, n$. Consequently,

$$h_k^2 \log\left[\frac{1}{n}\sum_{i=1}^n \exp\left(-\frac{\|\boldsymbol{x}_k - \boldsymbol{y}_i\|^2}{2h_k^2}\right)\right]$$
$$\leq h_k^2 \log\left[\exp\left(-\frac{(\|\boldsymbol{x}_k\| - r_0)^2}{2h_k^2}\right)\right] = -\frac{(\|\boldsymbol{x}_k\| - r_0)^2}{2}$$

for all $k \geq k_0$. By equation (29), this implies that

$$h_k^2[\log \hat{p}_{h_k}(\boldsymbol{x}_k) + \log[(2\pi)^{\frac{d}{2}} h_k^d]] \leq -\frac{(\|\boldsymbol{x}_k\| - r_0)^2}{2} \tag{48}$$

for all $k \geq k_0$. On the other hand, by the limit (30), Assumption 7.1 and the choice of r_0 we have

$$\lim_{k\to\infty} h_k^2[\log \hat{p}_{h_k}(\boldsymbol{y}_j) + \log[(2\pi)^{\frac{d}{2}} h_k^d]]$$
$$= -\frac{1}{2n}\sum_{i=1}^n \|\boldsymbol{y}_j - \boldsymbol{y}_i\|^2$$
$$= -\frac{1}{2n}\left(\sum_{i=1}^n \|\boldsymbol{y}_j\|^2 - 2\sum_{i=1}^n \boldsymbol{y}_j^T \boldsymbol{y}_i + \sum_{i=1}^n \|\boldsymbol{y}_i\|^2\right) \geq -r_0^2 \tag{49}$$

[2] Problem (45) can be equivalently formulated as an *orthogonal Procrustes problem*. With the matrices defined above, this problem has a unique solution (e.g. Higham 2008).

for all $j = 1, 2, \ldots, n$. Plugging the limits (48) and (49) into inequality (47) then leads to a contradiction for any sufficiently large k since $\lim_{k\to\infty} \|\boldsymbol{x}_k\| = \infty$ by condition (46) and the assumption that the sequence $\{r_k\}$ is monotoneusly increasing. □

Proof (Theorem 3.2) By Lemma 7.6 there exists

$$r > \max_{i=1,2,\ldots,n} \|\boldsymbol{y}_i\|$$

such that $U_h \subseteq B(\mathbf{0}; r)$ for all sufficiently large h. Thus, condition (9) for all $\boldsymbol{x} \in U_h$ and such h follows from Assumption 3.2, compactness of the set $B(\mathbf{0}; r)$ and Lemma 7.2. Finally, compactness and connectedness of the set U_h for all sufficiently large h follows from the strict concavity of $\log \hat{p}_h$ in $B(\mathbf{0}; r) \supseteq U_h$ by condition (33). □

References

Berkooz, G., Holmes, P., Lumley, J.L.: The proper orthogonal decomposition in the analysis of turbulent flows. Annu. Rev. Fluid Mech. **25**, 539–575 (1993)

Chacón, J.E., Duong, T., Wand, M.P.: Asymptotics for general multivariate kernel density derivative estimators. Stat. Sin. **21**, 807–840 (2011)

Christiansen, B.: The shortcomings of nonlinear principal component analysis in identifying circulation regimes. J. Clim. **18**(22), 4814–4823 (2005)

Damon, J.: Generic structure of two-dimensional images under Gaussian blurring. SIAM J. Appl. Math. **59**(1), 97–138 (1998)

Delworth, T.L., Broccoli, A.J., Rosati, A., Stouffer, R.J., Balaji, V.: GFDL's CM2 global coupled climate models. Part I: formulation and simulation characteristics. J. Clim. **19**(5), 643–674 (2006)

Donoho, D.L., Grimes, C.: Hessian eigenmaps: locally linear embedding techniques for high-dimensional data. Proc. Natl. Acad. Sci. **100**(10), 5591–5596 (2003)

Duong, T.: ks: Kernel density estimation and kernel discriminant analysis for multivariate data in R. J. Stat. Softw. **21**(7), 1–16 (2007)

Einbeck, J., Tutz, G., Evers, L.: Local principal curves. Stat. Comput. **15**(4), 301–313 (2005)

Einbeck, J., Evers, L., Bailer-Jones, C.: Representing complex data using localized principal components with application to astronomical data. In: Gorban, A.N., Kégl, B., Wunsch, D.C., Zinovyev, A.Y. (eds.) Principal Manifolds for Data Visualization and Dimension Reduction, volume 58 of Lecture Notes in Computational Science and Engineering, pp. 178–201. Springer, Berlin, Heidelberg (2008)

Genovese, C.R., Perone-Pacifico, M., Verdinelli, I., Wasserman, L.: Nonparametric ridge estimation. Ann. Stat. **42**(4), 1511–1545 (2014)

Golub, G.H., Van Loan, C.F.: Matrix Computations, 3rd edn. The Johns Hopkins University Press, Baltimore (1996)

Golyandina, N., Nekrutkin, V., Zhigljavsky, A.A.: Analysis of Time Series Structure: SSA and Related Techniques. Chapman and Hall/CRC Press, London (2001)

Greengard, L., Strain, J.: The fast Gauss transform. SIAM J. Sci. Stat. Comput. **12**(1), 79–94 (1991)

Higham, N.J.: Functions of Matrices: Theory and Computation. SIAM, Philadelphia (2008)

Hsieh, W.W.: Nonlinear multivariate and time series analysis by neural network methods. Rev. Geophys. **42**(1), 1–25 (2004)

Hsieh, W.W., Hamilton, K.: Nonlinear singular spectrum analysis of the tropical stratospheric wind. Quart. J. R. Meteorol. Soc. **129**(592), 2367–2382 (2003)

Jaromczyk, J.W., Toussaint, G.T.: Relative neighborhood graphs and their relatives. Proc. IEEE **80**(9), 1502–1517 (1992)

Jolliffe, I.T.: Principal Component Analysis. Springer-Verlag, Berlin (1986)

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

Kramer, M.A.: Nonlinear principal component analysis using autoassociative neural networks. AIChE J. **37**(2), 233–243 (1991)

Loève, M.: Probability Theory: Foundations, Random Sequences. van Nostrand, Princeton (1955)

Magnus, J.R.: On differentiating eigenvalues and eigenvectors. Econ. Theory **1**(2), 179–191 (1985)

Miller, J.: Relative critical sets in R^n and applications to image analysis. PhD thesis, University of North Carolina (1998)

Monahan, A.H.: Nonlinear principal component analysis: tropical Indo-Pacific sea surface temperature and sea level pressure. J. Clim. **14**(2), 219–233 (2001)

Newbigging, S.C., Mysak, L.A., Hsieh, W.W.: Improvements to the nonlinear principal component analysis method, with applications to ENSO and QBO. Atmos.-Ocean **41**(4), 291–299 (2003)

Ortega, J.M.: Numerical Analysis: A Second Course. SIAM, Philadelphia (1990)

Ozertem, U., Erdogmus, D.: Locally defined principal curves and surfaces. J. Mach. Learn. Res. **12**, 1249–1286 (2011)

Pearson, K.: On lines and planes of closest fit to systems of points in space. Philos. Mag. Ser. 6 **2**(11), 559–572 (1901)

Pulkkinen, S.: Ridge-based method for finding curvilinear structures from noisy data. Comput. Stat. Data Anal. **82**, 89–109 (2015)

Pulkkinen, S., Mäkelä, M.M., Karmitsa, N.: A generative model and a generalized trust region Newton method for noise reduction. Comput. Optim. Appl. **57**(1), 129–165 (2014)

Rangayyan, R.M.: Biomedical Signal Analysis: A Case-Study Approach. IEEE Press, New York (2002)

Renardy, M., Rogers, R.C.: An introduction to partial differential equations. In: Marsden, J.E., Sirovich, L., Antman, S.S. (eds.) Texts in Applied Mathematics, vol. 13, 2nd edn. Springer-Verlag, New York (2004)

Ross, I.: Nonlinear dimensionality reduction methods in climate data analysis. PhD thesis, University of Bristol, United Kingdom (2008)

Ross, I., Valdes, P.J., Wiggins, S.: ENSO dynamics in current climate models: an investigation using nonlinear dimensionality reduction. Nonlinear Process. Geophys. **15**, 339–363 (2008)

Schölkopf, B., Smola, A., Müller, K.-R.: Kernel principal component analysis. In: Gerstner, W., Germond, A., Hasler, M., Nicoud, J.-D. (eds.) Artificial Neural Networks-ICANN'97, volume 1327 of Lecture Notes in Computer Science, pp. 583–588. Springer, Berlin (1997)

Scholz, M., Kaplan, F., Guy, C.L., Kopka, J., Selbig, J.: Non-linear PCA: a missing data approach. Bioinformatics **21**(20), 3887–3895 (2005)

Scholz, M., Fraunholz, M., Selbig, J.: Nonlinear principal component analysis: Neural network models and applications. In: Gorban, A.N., Kégl, B., Wunsch, D.C., Zinovyev, A.Y. (eds.) Principal Manifolds for Data Visualization and Dimension Reduction, volume 58 of Lecture Notes in Computational Science and Engineering, pp. 44–67. Springer, Berlin (2008)

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

Tipping, M.E., Bishop, C.M.: Probabilistic principal component analysis. J. R. Stat. Soc. **61**(3), 611–622 (1999)

Vautard, R., Yiou, P., Ghil, M.: Singular-spectrum analysis: a toolkit for short, noisy chaotic signals. Physica D **58**(1–4), 95–126 (1992)

Weare, B.C., Navato, A.R., Newell, E.R.: Empirical orthogonal analysis of Pacific sea surface temperatures. J. Phys. Oceanogr. **6**(5), 671–678 (1976)

Weinberger, K., Saul, L.: Unsupervised learning of image manifolds by semidefinite programming. Int. J. Comput. Vis. **70**(1), 77–90 (2006)

Whittlesey, E.F.: Fixed points and antipodal points. Am. Math. Mon. **70**(8), 807–821 (1963)

Yang, C., Duraiswami, R., Gumerov, N.A., Davis, L.: Improved fast gauss transform and efficient kernel density estimation. In Ninth IEEE International Conference on Computer Vision. volume 1, pp. 664–671. Nice, France (2003)

Zhang, Z., Zha, H.: Principal manifolds and nonlinear dimensionality reduction via tangent space alignment. SIAM J. Sci. Comput. **26**(1), 313–338 (2004)