Seppo Pulkkinen

A Probabilistic Approach to Nonlinear Principal Component Analysis With Applications

Turku Centre *for* Computer Science

TUCS Technical Report
No 1091, November 2013

A Probabilistic Approach to Nonlinear Principal Component Analysis With Applications

Seppo Pulkkinen
University of Turku, Department of Mathematics and Statistics
FI-20014 Turku, Finland
seppo.pulkkinen@utu.fi

TUCS Technical Report
No 1091, November 2013

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; nonlinear; kernel density; nonparametric methods; ridge; predictor-corrector method

TUCS Laboratory
Turku Optimization Group (TOpGroup)

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 [28], *principal component analysis* (PCA, e.g. [18]) 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. [39]), *proper orthogonal decomposition* (POD) in fluid mechanics (e.g. [3]) and the *Karhunen-Loève transform* (KLT) in the theory of stochastic processes (e.g. [21]).

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. [15, 20, 24, 35]) and kernel PCA (KPCA, e.g. [36]). 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. [8,9,19]). 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, [42]) 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 [27]. This idea has later been refined in [29] and [30]. 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 ridge sets, 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 ap-

plicability of KDPCA. To this end, it will be shown that the nonlinear principal components converge to the linear ones when the kernel bandwidth approaches infinity. Consequently, the robustness of the linear PCA is always attained by choosing a sufficiently large bandwidth, but there is a tradeoff between robustness and ability to describe nonlinear structure.

Finally, KDPCA is extended to time series analysis. In analogy with the well-known *singular spectrum analysis* (SSA, e.g. [11, 38]), 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. [15, 16]) and medical applications such as electrocardiography and electroencephalography (e.g. [31]).

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

2 The linear PCA

As the proposed method is a generalization of the linear PCA (e.g. [18]), 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 new coordinate system via an orthogonal transformation. In the new 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 [18]. 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

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

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

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

In statistical literature, the coordinates $\boldsymbol{\Theta} \in \mathbb{R}^{d\times m}$ obtained in this way are called principal component *scores* (e.g. [18]).

3 Nonlinear kernel density PCA

In this section we develop the kernel density principal component analysis (KD-PCA). The method is based on estimation of the underlying density of the data by 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 density estimate. The projection curves are defined as a solution to a differential equation, and predictor-corrector method is developed for this purpose.

3.1 Ridge definition

We adapt the definition of a ridge set from [30]. 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 is illustrated in Figure 1.

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*

$$\nabla p(\boldsymbol{x})^T \boldsymbol{v}_i(\boldsymbol{x}) = 0, \quad i > r, \tag{5a}$$

$$\lambda_{r+1}(\boldsymbol{x}) < 0, \tag{5b}$$

$$\lambda_1(\boldsymbol{x}) > \lambda_2(\boldsymbol{x}) > \cdots > \lambda_{r+1}(\boldsymbol{x}), \quad \text{if } r > 0, \tag{5c}$$

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.

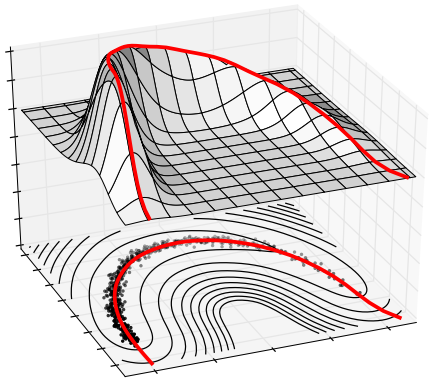

Figure 1: Ridge curve of the density of a point set that is distributed around a curve.

The following result shows a connection between the ridge set and the linear principal components when the underlying density of the data is normal. This result follows trivially from the following lemma (see [27]) 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*

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

Proposition 3.1 suggests an approach for estimating the principal component scores $\boldsymbol{\theta}$ 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 $\nabla^2 \log p$ for some $0 < m \leq d$ 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 Section 3.3.

3.2 Density estimation and choice of bandwidth

Here we use Gaussian kernels to estimate the density from the given data. In what follows, we establish a connection between ridge sets of this density and the linear principal components. Furthermore, we show that a linear PCA hyperplane is obtained as a special case of such a ridge set.

Definition 3.2. *The* Gaussian kernel density estimate $\hat{p}_h$ *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*

$$\hat{p}_h(\boldsymbol{x}) = \frac{1}{n} \sum_{i=1}^{n} K_h(\boldsymbol{x} - \boldsymbol{y}_i), \tag{6}$$

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

$$K_h(\boldsymbol{x}) = \frac{1}{(2\pi)^{\frac{d}{2}} h^d} \exp\left(-\frac{\|\boldsymbol{x}\|^2}{2h^2}\right) \tag{7}$$

with bandwidth $h > 0$.

The following result establishes a connection between linear principal components and ridges of a Gaussian kernel density. It essentially 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.

Theorem 3.1. *Let $\hat{p}_h : \mathbb{R}^d \to \mathbb{R}$ be a Gaussian kernel density estimate, 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*

$$\tilde{\boldsymbol{\mu}}(\boldsymbol{x}) = \sum_{i=1}^n c_i(\boldsymbol{x}) \boldsymbol{y}_i, \tag{8}$$

$$\tilde{\boldsymbol{\Sigma}}(\boldsymbol{x}) = \sum_{i=1}^n c_i(\boldsymbol{x}) [\boldsymbol{y}_i - \tilde{\boldsymbol{\mu}}(\boldsymbol{x})][\boldsymbol{y}_i - \tilde{\boldsymbol{\mu}}(\boldsymbol{x})]^T, \tag{9}$$

where

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

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 (cf. proof of Lemma A.1 in Appendix A)

$$h^2 \nabla \log p_h(\boldsymbol{x}) = -[\boldsymbol{x} - \tilde{\boldsymbol{\mu}}(\boldsymbol{x})] \tag{10}$$

and

$$h^4 \nabla^2 \log p_h(\boldsymbol{x}) + h^2 \boldsymbol{I} = \tilde{\boldsymbol{\Sigma}}(\boldsymbol{x}). \tag{11}$$

By equation (11), the matrices $\nabla^2 \log \hat{p}_h$ and $\tilde{\boldsymbol{\Sigma}}(\boldsymbol{x})$ have the same eigenvectors. Hence, by equation (10) 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 a Gaussian kernel density 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.

An important special case arises when h approaches infinity. At this limit, the r-dimensional ridge set of the density approaches the r-dimensional PCA hyperplane, which can be readily observed from equations (8)–(11). A rigorous proof of this property is deferred to Appendix A. Thus, by choosing a large h we achieve the robustness of PCA but, on the other hand, compromise the ability to describe nonlinear structure in the data.

Assumption 3.1. *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. *Let $\hat{p}_h : \mathbb{R}^d \to \mathbb{R}$ be a Gaussian kernel density estimate, let $0 \leq r < d$ and let Assumption 3.1 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

$$\left.\begin{array}{l} \operatorname{dist}(\mathcal{R}^r_{\hat{p}_{h^2 \boldsymbol{I}}} \cap U, S_\infty^r) < \varepsilon, \\ \operatorname{dist}(S_\infty^r \cap U, \mathcal{R}^r_{\hat{p}_{h^2 \boldsymbol{I}}}) < \varepsilon \end{array}\right\} \quad \text{for all } h \geq h_0,$$

where

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

3.3 Obtaining principal component scores from ridge sets

Based on Proposition 3.1, we now develop the theoretical basis for estimating the first m nonlinear principal component scores of a given point set. The idea is to obtain the scores one by one by successively projecting the points onto lower-dimensional ridge sets of the underlying density that is estimated by Gaussian kernels. 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.

For now, we assume that a given point has already been projected onto an m-dimensional ridge set of its underlying density p with some $m \leq d$. 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

$$\frac{d}{dt}\left[\boldsymbol{P}_r(\boldsymbol{\gamma}_r(t))\nabla \log p(\boldsymbol{\gamma}_r(t))\right] = \mathbf{0}, \quad t \geq 0, \tag{12}$$
$$\boldsymbol{\gamma}_r(0) = \boldsymbol{x}_0, \quad \boldsymbol{x}_0 \in \mathcal{R}^r_{\log p} \setminus \mathcal{R}^{r-1}_{\log p},$$

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 $\boldsymbol{x}_0 \in \mathcal{R}^r_{\log p}$, where $1 \leq r \leq m$, can be projected onto the lower-dimensional ridge set $\mathcal{R}^{r-1}_{\log p}$ by following the solution curve of (12) 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 (12) we have*

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

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

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

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

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. [22]), equation (12) can after some calculation be rewritten as

$$\boldsymbol{A}_r(\boldsymbol{\gamma}_r(t))\boldsymbol{\gamma}_r'(t) = \mathbf{0}, \tag{13}$$

where

$$\boldsymbol{A}_r(\boldsymbol{x}) = \boldsymbol{P}_r(\boldsymbol{x})\nabla^2 \log p(\boldsymbol{x}) - \boldsymbol{F}_r(\boldsymbol{x}), \tag{14}$$
$$\boldsymbol{F}_r(\boldsymbol{x}) = \boldsymbol{v}_r(\boldsymbol{x})^T \nabla \log p(\boldsymbol{x}) \nabla \boldsymbol{v}_r(\boldsymbol{x})^T + \boldsymbol{v}_r(\boldsymbol{x}) \nabla \log p(\boldsymbol{x})^T \nabla \boldsymbol{v}_r(\boldsymbol{x}) \tag{15}$$

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. [10]).

For a general density p, projection onto the ridge set $\mathcal{R}^{r-1}_{\log p}$ can still be done by maximizing $\log p$ along the curve $\boldsymbol{\gamma}_r$, but this requires additional justification. To this end, we first show that when $\boldsymbol{\gamma}_r$ approaches a ridge point $\boldsymbol{x}^* \in \mathcal{R}^{r-1}_{\log p}$, the tangent vector $\boldsymbol{\gamma}_r'$ becomes parallel to the eigenvector $\boldsymbol{v}_r$. Here we need a technical assumption that will be justified later in this subsection.

Assumption 3.2. *The eigenvalues of* $\nabla^2 \log p$ *satisfy the conditions*

(i) $\lambda_1(\boldsymbol{\gamma}_r(t)) > \lambda_2(\boldsymbol{\gamma}_r(t)) > \cdots > \lambda_{r+1}(\boldsymbol{\gamma}_r(t))$,
(ii) $\lambda_1(\boldsymbol{\gamma}_r(t)) < 0$

for all $t \geq 0$.

Proposition 3.3. *Let* $1 \leq r \leq d$ *and let* $\boldsymbol{\gamma}_r'$ *denote the normalized tangent vector of a solution curve of* (12)*. If Assumption 3.2 is satisfied and*

$$\lim_{t \to t^*} \boldsymbol{v}_r(\boldsymbol{\gamma}_r(t))^T \nabla \log p(\boldsymbol{\gamma}_r(t)) = 0 \tag{16}$$

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

$$\lim_{t \to t^*} |\boldsymbol{\gamma}_r'(t)^T \boldsymbol{v}_r(\boldsymbol{\gamma}_r(t))| = 1. \tag{17}$$

Proof. Define the set

$$U = \{\boldsymbol{x} \in \mathbb{R}^d \mid \lambda_1(\boldsymbol{x}) < 0 \text{ and} \tag{18}$$
$$\lambda_1(\boldsymbol{x}) > \lambda_2(\boldsymbol{x}) > \cdots > \lambda_{r+1}(\boldsymbol{x})\}. \tag{19}$$

The range of the matrix in the second term of $\boldsymbol{F}_r(\boldsymbol{x})$ defined by equation (15), 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 (19). We also note that the range of the first term of the matrix $\boldsymbol{A}_r(\boldsymbol{x})$ defined by equation (14), 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 (18) that guarantees nonsingularity of $\nabla^2 \log p(\boldsymbol{x})$.

On the other hand, by the limit (16) the first term of the matrix $\boldsymbol{F}_r(\boldsymbol{\gamma}_r(t))$ defined by equation (15), 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 $\boldsymbol{x} \in U$, equations (13)–(15) and Assumption 3.2 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 (12) for some $1 \leq r \leq d$ and Assumption 3.2 is satisfied, then either $\boldsymbol{\gamma}_r(t) \in \mathcal{R}^r_{\log p} \backslash \mathcal{R}^{r-1}_{\log p}$ 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 equation (12), 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 3.2 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 (17) 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 (5b) the point $\boldsymbol{\gamma}_r(t^*)$ is a local maximum of $\log p$ along $\boldsymbol{\gamma}_r$. □

Recall that our aim is to use projection curves $\boldsymbol{\gamma}_r$ defined by equation (12) to obtain the first m nonlinear principal component scores of the given sample points $\boldsymbol{y}_i$. This will be done by using the kernel density $\log \hat{p}_h$ defined by (6) and (7) as the objective function. Differently to the normal density in Proposition 3.2, this density is not guaranteed to be unimodal or have connected ridge sets. For instance, when h is too small, the density becomes multimodal.

Unimodality of the density and connectedness of its ridge sets are essential here. This is because as in the linear PCA, our aim is to describe the data in a single well-defined coordinate system. Hence, we assume the following.

Assumption 3.3. *Define the set $U_h = \bigcup_{i=1}^n \mathcal{L}_i^h$, where*

$$\mathcal{L}_i^h = \{\boldsymbol{x} \in \mathbb{R}^d \mid \log \hat{p}_h(\boldsymbol{x}) \geq \log \hat{p}_h(\boldsymbol{y}_i)\}.$$

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

$$0 > \lambda_1(\boldsymbol{x}) > \lambda_2(\boldsymbol{x}) > \cdots > \lambda_{m+1}(\boldsymbol{x}) \tag{20}$$

and that set U_h is connected.

Damon [5] and Miller [23] give a rigorous treatment of ridge sets of C^∞-functions in a differential geometric framework. Under the above assumption, their results guarantee that the r-dimensional ridge set of the density $\log \hat{p}_h$ forms a connected manifold in the set U_h for any $1 \leq r \leq m$. Furthermore, $\log \hat{p}_h$ is unimodal in U_h. In addition, Assumption 3.3 implies continuity of the Hessian eigenvectors (e.g. [26], Theorem 3.1.3), which is essential for the definition of the initial value problem (12). Assumption 3.3 also entails Assumption 3.2 when the curves $\boldsymbol{\gamma}_r$ lie in U_h.

Assumption 3.3 can be satisfied by choosing a sufficiently large h. The following result is proven in Appendix A.

Theorem 3.3. *Under Assumption 3.1 for $r = m$, for any Gaussian kernel density estimate $\hat{p}_h$ there exists $h_0 > 0$ such that Assumption 3.3 is satisfied for all $h \geq h_0$.*

Finally, the arc length of a curve $\boldsymbol{\gamma}_r$ gives the (curvilinear) distance of its starting point to the ridge set $\mathcal{R}^{r-1}_{\log \hat{p}_h}$. Assume that we have projected a given sample point $\boldsymbol{y}_i$ onto the ridge set $\mathcal{R}^{m}_{\log \hat{p}_h}$. 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 $\boldsymbol{y}_i$. When Assumption 3.3 is satisfied, imposing the conditions (cf. Proposition 3.2)

$$\left.\begin{array}{l} \boldsymbol{\gamma}_r'(t)^T \nabla \log \hat{p}_{\boldsymbol{H}}(\boldsymbol{\gamma}_r(t)) > 0, \\ \|\boldsymbol{\gamma}_r'(t)\| = 1 \end{array}\right\} \text{ for all } \left\{\begin{array}{l} r = m, m-1, \ldots, 1, \\ t \geq 0 \end{array}\right. \tag{21}$$

guarantees that the curves $\boldsymbol{\gamma}_r$ lie in the set U_h. This ensures that the ridge projections are well-defined.

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

$$\theta_r = s_r^* \int_0^{t_r^*} \|\boldsymbol{\gamma}_r'(t)\| dt, \quad \boldsymbol{x}_0^r = \left\{\begin{array}{ll} \tilde{\boldsymbol{y}}_i, & r = m, \\ \boldsymbol{\gamma}_{r+1}(t_{r+1}^*), & 1 \leq r < m \end{array}\right. \tag{22}$$

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

$$s_r(t) = \left\{\begin{array}{ll} 1, & \text{if } \boldsymbol{\gamma}_r'(t)^T \boldsymbol{v}_r(\boldsymbol{\gamma}_r(t)) > 0, \\ -1, & \text{otherwise,} \end{array}\right. \tag{23}$$

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

3.4 Algorithm for estimating principal component scores

Based on the theory developed in Subsection 3.3, we are now ready to develop the algorithm for estimating the nonlinear principal component scores

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

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 $\mathcal{R}^m_{\log \hat{p}_h}$ and then successively projecting them onto the lower-dimensional ridge sets $\mathcal{R}^r_{\log \hat{p}_h}$ until $r = 0$. The latter projections are done by tracing the curves $\boldsymbol{\gamma}_r$ by using a predictor-corrector method. As a by-product, the principal component scores are obtained from a numerical approximation of the integral (22).

A pseudocode of the algorithm is listed as Algorithm 1. It involves the initial projection onto the ridge set $\mathcal{R}^m_{\log \hat{p}_h}$ (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 $\mathcal{R}^{r-1}_{\log \hat{p}_h}$. 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 in [30] (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 $\boldsymbol{\gamma}_r$ is chosen as $\boldsymbol{x}^*_i$ representing the projection of the sample point $\boldsymbol{y}_i$ onto the set $\mathcal{R}^r_{\log \hat{p}_h}$. Assuming that there exists a monotoneously increasing sequence $\{t_k\}$ such that $\boldsymbol{\gamma}_r(t_{k^*}) \in \mathcal{R}^{r-1}_{\log \hat{p}_h}$ for some k^*, we introduce the notation $\boldsymbol{x}_k = \boldsymbol{\gamma}_r(t_k)$ for the iterates along the curve $\boldsymbol{\gamma}_r$. With this notation, an approximation to the integral (22) is given by

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

The algorithm uses a predictor-corrector method to generate the iterates $\boldsymbol{x}_k$. At the predictor step (line 18), the algorithm proceeds along a tangent vector $\boldsymbol{u}_k = \boldsymbol{\gamma}'_r(t_k)$ solved from equation (13). That is,

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

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

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

(line 9) is introduced to impose conditions (21). To project the predictor estimate $\tilde{\boldsymbol{x}}_k$ back to the ridge set $\mathcal{R}^r_{\log \hat{p}_h}$, the algorithm takes a corrector step (line 19).

A stopping criterion is imposed to terminate the tracing of the curve $\boldsymbol{\gamma}_r$ when a maximum of $\log \hat{p}_h$ along $\boldsymbol{\gamma}_r$ is encountered (line 11). For $k > 0$, the condition

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

Algorithm 1: Nonlinear principal component scores

input : sample points $\boldsymbol{Y} = [\boldsymbol{y}_1 \quad \boldsymbol{y}_2 \quad \cdots \quad \boldsymbol{y}_n]^T \in \mathbb{R}^{n\times d}$
Gaussian kernel density estimate $\hat{p}_h : \mathbb{R}^d \to \mathbb{R}$
ridge dimension $0 < m \leq d$
step size $\tau > 0$

output: principal component scores $\boldsymbol{\Theta} = [\boldsymbol{\theta}_1 \quad \boldsymbol{\theta}_2 \quad \cdots \quad \boldsymbol{\theta}_n]^T \in \mathbb{R}^{n\times m}$

1 $\boldsymbol{\Theta} \leftarrow \boldsymbol{0}$
2 **for** $i = 1, 2, \ldots, n$ **do**
3 $\boldsymbol{x}_i^* \leftarrow \text{GTRN}(\hat{p}_h, m, \boldsymbol{y}_i, \tau, 10^{-5})$
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 vector $\boldsymbol{u}_k$ from (13) such that $\|\boldsymbol{u}_k\| = 1$.
9 **if** $\boldsymbol{u}_k^T \nabla \log \hat{p}_h(\boldsymbol{x}_k) > 0$ **then** $s_k \leftarrow 1$ **else** $s_k \leftarrow -1$
10 **if** $k > 0$ **then**
11 **if** $s_{k-1}\boldsymbol{u}_{k-1}^T \boldsymbol{u}_k s_k < 0$ **then**
12 $\boldsymbol{x}_i^* \leftarrow \text{GTRN}(\hat{p}_h, r-1, (\boldsymbol{x}_{k-1} + \boldsymbol{x}_k)/2, 0.5\tau, 10^{-5})$
13 $\theta_{i,r} \leftarrow \theta_{i,r} + \|\boldsymbol{x}_i^* - \boldsymbol{x}_{k-1}\|$
14 **if** $(\boldsymbol{x}_i^* - \boldsymbol{x}_{k-1})^T \boldsymbol{v}_r(\boldsymbol{x}_i^*) < 0$ **then** $\theta_{i,r} \leftarrow -\theta_{i,r}$
15 Terminate the inner loop.
16 **else**
17 $\theta_{i,r} \leftarrow \theta_{i,r} + \|\boldsymbol{x}_k - \boldsymbol{x}_{k-1}\|$
18 $\tilde{\boldsymbol{x}}_k \leftarrow \boldsymbol{x}_k + \tau s_k \boldsymbol{u}_k$
19 $\boldsymbol{x}_{k+1} \leftarrow \text{GTRN}(\hat{p}_h, r, \tilde{\boldsymbol{x}}_k, 0.5\tau, 10^{-5})$

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 $\boldsymbol{x}_i^* \in \mathcal{R}^{r-1}_{\log \hat{p}_h}$ (line 12). At line 14, the algorithm computes the sign s_r^* for the integral (22) by approximately testing condition (23). 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.

The `GTRN` algorithm developed in [30] 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 $\boldsymbol{z}_{l+1} = \boldsymbol{z}_l + \boldsymbol{s}_l$ is obtained by solving the

subproblem

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

where Q_l denotes the model function at the current iterate $\boldsymbol{z}_l$, $\{\boldsymbol{v}_i(\boldsymbol{z}_l)\}_{i=r+1}^d$ 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 3.1. *The `GTRN` algorithm can be viewed as an approximate solution method to an initial value problem of the form (12), 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 by this algorithm only requires continuity of the first r Hessian eigenvectors. That is, when the $r+1$ greatest eigenvalues are distinct in U_h.*

Remark 3.2. *Tests for connectedness of the ridge sets are not included in Algorithm 1 for the sake of simplicity. This can be done by testing if the traced curve crosses a point $\boldsymbol{x}$ where $\lambda_1(\boldsymbol{x}) = 0$ 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 $\nabla^2 \log \hat{p}_h$ (see [23, 29]). When either of these conditions is met, the algorithm can be restarted with a larger h or smaller initial ridge dimension m.*

Remark 3.3. *A consistent orientation of the eigenvectors $\boldsymbol{v}_r(\boldsymbol{x}_i^*)$ at the projected points is necessary for the principal component scores $\theta_{i,r}$ 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. [17]) 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 Section 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 $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ is defined as

$$\boldsymbol{Y}_{\boldsymbol{x},L} = \begin{bmatrix} x_1 & x_2 & x_3 & \cdots & x_L \\ x_2 & x_3 & x_4 & \cdots & x_{L+1} \\ x_3 & x_4 & x_5 & \cdots & x_{L+2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{n-L+1} & x_{n-L+2} & x_{n-L+3} & \cdots & x_n \end{bmatrix}, \tag{24}$$

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. [38]. Generalizing their approach, we minimize the reconstruction error

$$E(\boldsymbol{x}) = \sum_{i=1}^{n-L+1} \sum_{j=1}^{L} (\tilde{y}_{i,j} - x_{i+j-1})^2 \tag{25}$$

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 estimate.

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

$$x_i^* = \begin{cases} \dfrac{1}{L} \displaystyle\sum_{j=1}^{L} \tilde{y}_{i-j+1,j}, & L \leq i \leq n-L+1 \\ \dfrac{1}{i} \displaystyle\sum_{j=1}^{i} \tilde{y}_{i-j+1,j}, & 1 \leq i \leq L-1 \\ \dfrac{1}{n-i+1} \displaystyle\sum_{j=i-n+L}^{L} \tilde{y}_{i-j+1,j}, & n-L+2 \leq i \leq n \end{cases} \tag{26}$$

for the elements of the reconstructed time series such that $E(\boldsymbol{x}^*)$ minimizes the reconstruction error (25).

Remark 4.1. *SSA can be extended to multivariate time series [38] and to time-series in a two-dimensional grid [12]. As these modifications differ from the univariate SSA only by the construction of the trajectory matrix and the formulae for obtaining the reconstructed time series, we do not consider the multivariate case here.*

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. Figure 5 in Section 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 in [30] is appropriate for this purpose. On the other hand, a parametrization of the reconstructed trajectory can be obtained by the algorithm developed in [29]. 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 the proposed KDPCA method. The method is applied to a climate model output that exhibits a highly nonlinear behaviour. In addition, its SSA-based extension is applied to an atmospheric time series. Finally, computational complexity analysis and comparison to related methods is given.

5.1 Test setup

All numerical tests were done on a machine with a 3.0GHz Core 2 Duo processor and 6GB system memory. Algorithm 1 as well as the algorithms developed in [29] and [30] 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 algorithms of [29] and [30] 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 Application to simulated climate model data

In the first test, 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 [1,6]. The chosen simulation output is from the pre-industrial control (`picntrl`) simulation experiment. All preprocessing steps were done as in [33] and [34], where this dataset has been analyzed in detail.

The simulated sea surface temperatures are computed on a latitude-longitude grid. The subregion 20° S-20° N, 125° E-65° W representing the pacific ocean was chosen from the original data. The temperatures are monthly values, and

the length of the simulation is 500 years, starting from the first century AD. This yields a dataset of 6000 samples. As a preprocessing step, seasonality was removed from the data by subtracting the monthly mean values of the raw data for each month. In climatological literature, such mean values are called *anomalies*.

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. The kernel bandwidth was chosen as $h = 40$.

The GFDL-CM2.1 dataset and its first principal component obtained by KDPCA are plotted in Figure 2. 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 KDPCA is shown in Figure 3. The corresponding principal component scores are plotted in Figure 4.

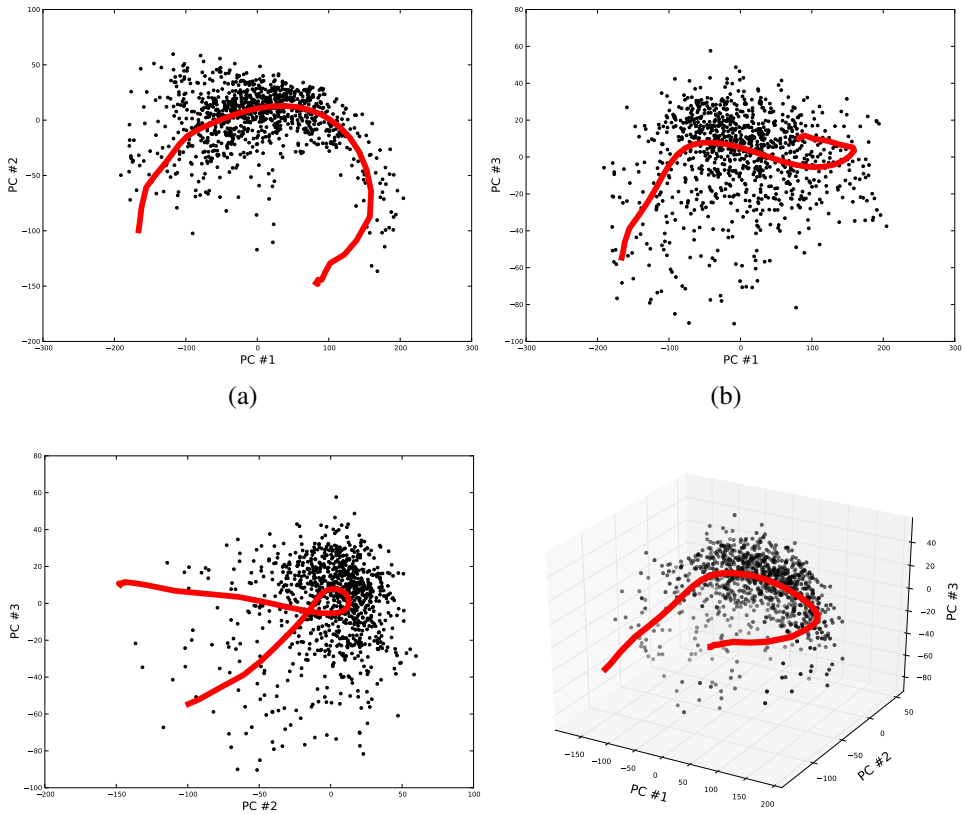

Figure 2: The first nonlinear principal component obtained from the GFDL-CM2.1 dataset (only a subset of the curve is drawn).

Compared to the linear principal component projection shown in Figure 2a, 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. Here the explained variances for the nonlinear principal components were obtained from the covariance matrix of the corresponding scores.

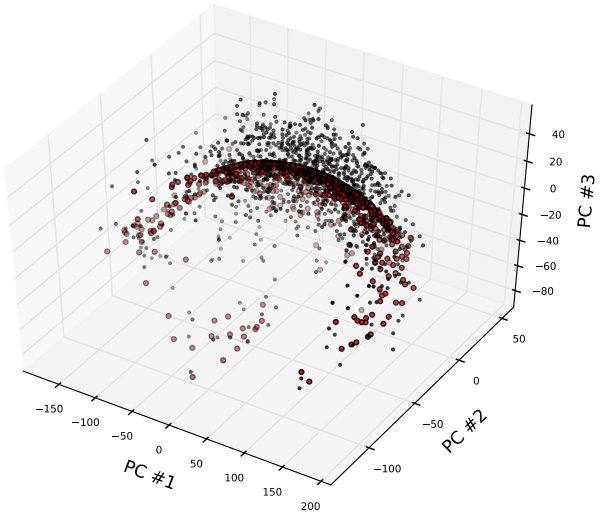

Figure 3: Projection of the GFDL-CM2.1 dataset onto the surface spanned by its first two nonlinear principal components.

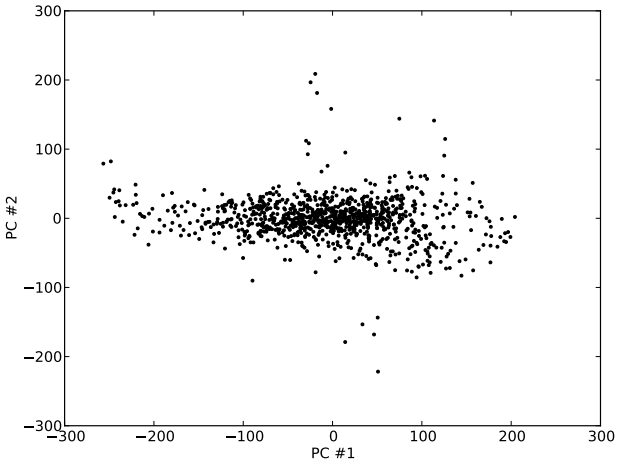

Figure 4: Two first nonlinear principal component scores obtained from the GFDL-CM2.1 dataset.

	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 %

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

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 [33] and [34] for the GFDL-CM2.1 data, and the two main sources of variation were identified. The first principal component correlates with the so-called 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 [33] and [34], 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 Subsection 5.5.

5.3 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 to 29 months. Motivated by an earlier neural network-based nonlinear SSA approach in [16], the nonlinear SSA (KDSSA) described in Section 4 was applied to a QBO time series. The time series is provided by the institute of meteorology at the University of Berlin [2]. 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 (24) 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 described in [30]. The bandwidth was chosen as $h = 200$. 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 (26).

The trajectory samples and their kernel density ridge projections in the reduced four-dimensional phase space are plotted in Figure 5. 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 Figure 6. 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.

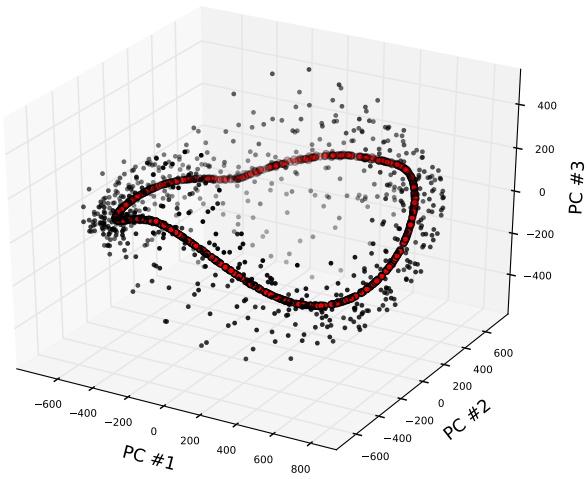

Figure 5: Phase space trajectory of the QBO time series and the reconstructed trajectory curve obtained by kernel density ridge projection.

The conclusion from Figures 5 and 6 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 Figure 5, cannot be described by any combination of linear principal components. Consequently, it can be seen from Figure 6 that the linear SSA reconstruction by using only the first principal component is inadequate to describe the structure of the time 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 Subsection 5.3, 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 [16] 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 us-

ing the algorithm developed in [29]. 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.

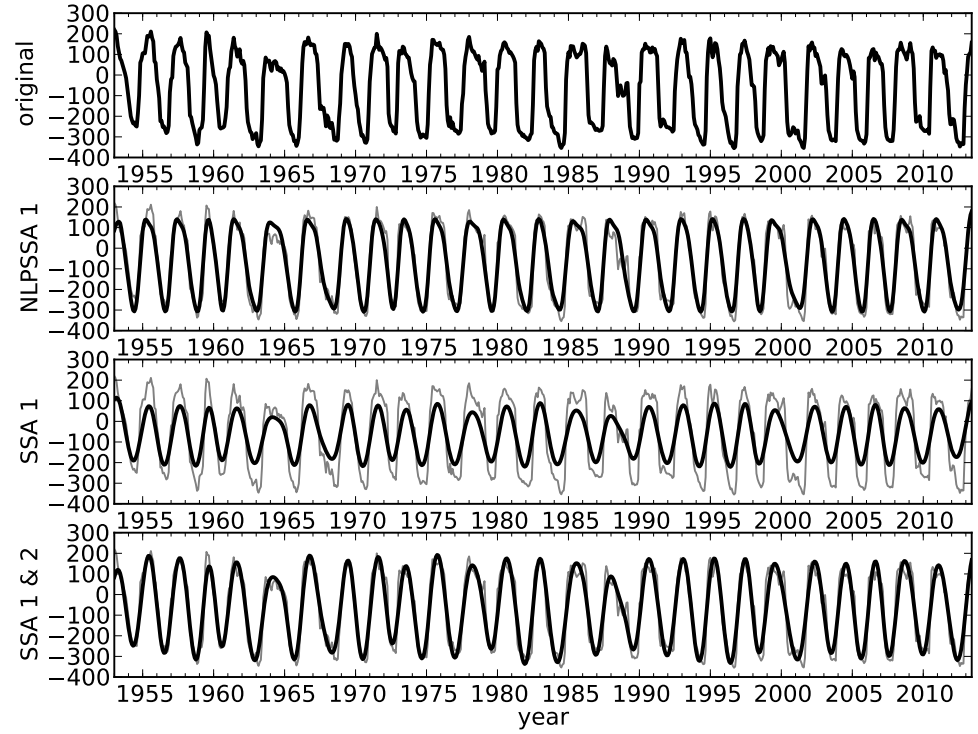

Figure 6: The QBO time series at the 30 Hpa level and the reconstructed time series obtained by using the first KDSSA component, the first linear SSA component and the two first linear SSA components combined. The original time series is plotted in gray in the lower figures.

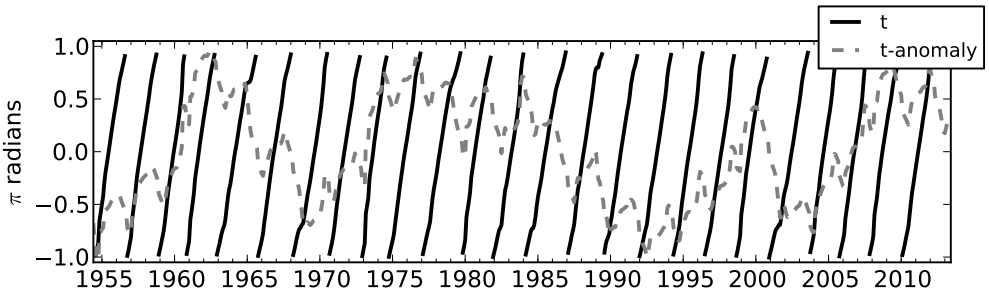

Figure 7: The first nonlinear principal component coordinate of the QBO time series (t) and the deviation from the fitted regression line (t-anomaly).

Due to its very regular period, the QBO time series progresses along its re-

constructed phase space trajectory at a nearly constant rate. This can be seen from Figure 7 showing the trajectory coordinate t scaled to the interval $[-\pi, \pi]$ as a function of time. In addition, following the approach of [16], 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 Figure 7.

Comparison of the t-anomaly time series to the t-time series and Figure 6 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.

5.4 Complexity Analysis

This subsection is devoted to discussion of computational complexity of KDPCA and comparison with existing nonlinear dimensionality reduction methods. After the initial projection step by using the `GTRN` algorithm having computational cost $\mathcal{O}(n^2d^2 + nd^3)$ (see [30]), the computational cost of KDPCA 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 [13] or related techniques.

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 [30]. 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 in [29]. 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 Subsection 5.3 at a cost of $\mathcal{O}(ndk)$, where k is the number of line segments.

5.5 Comparison to Other Methods

Though the neural network-based NLPCA (e.g. [15, 20, 24, 35]) has become popular particularly in climate analysis, it has several shortcomings. Some of them are discussed below.

- NLPCA involves minimization of a complicated cost function that generally has a large number of local minima. This problem is typically addressed by using a large number of 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 probabilistic 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 in [4]. Due to this issue, drawing statistical inferences from the NLPCA output should be done with extreme caution.
- NLPCA uses a number of artificial penalty terms to avoid overfitting. Despite this, the density of the data along the first nonlinear principal component can exhibit spurious multimodality [4]. This can occur even when the underlying density of the data is close to normal. On the other hand, KDPCA always attains the robustness of the linear PCA when the kernel bandwidth h is chosen sufficiently large.
- When using NLPCA, the type of the 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 in [29].
- The curves fitted by NLPCA are not parametrized by arc length. This may introduce a significant bias to the principal component scores. When drawing statistical inferences from a curve fitted by NLPCA, arc length reparametrization should be done to remove the bias [25]. 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 [37], Hessian eigenmaps [7] and maximum variance unfolding [40] 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.

KDPCA has also some shortcomings that it shares with NLPCA. Namely, the bandwidth parameter h plays a similar role to the penalty parameters in NLPCA. KDPCA breaks down when h is too small. In this case, the ridges of the density estimate no longer form a connected set, and the density becomes multimodal. The issue of choosing a sufficiently large h is difficult. In the absence of a more sophisticated method, it can be done by visual inspection of the obtained principal components in two or three dimensions (cf. Figures 2 and 5). A possible approach could be to use some computationally cheap plugin estimate as an initial guess for the bandwidth.

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 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, a fundamental result was derived showing that by letting h approach infinity, KDPCA reduces to the linear PCA and achieves

its robustness when desired. Second, the theory of ridge sets ensures that any ridge set of a Gaussian kernel density estimate has a well-defined coordinate system when h is sufficiently large. Third, this choice allows automatic estimation of an appropriate bandwidth from the data, though this approach was not pursued in this paper.

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 h. The bandwidth has an intuitive interpretation as a scale parameter.

While KDPCA showed impressive results on the test datasets, its applicability to real-world data remains to be fully confirmed. When the dataset 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 fast Gauss transform mentioned in Subsection 5.4 could significantly improve the scalability of KDPCA to large datasets. Computational difficulties due to high dimensionality of the input data can also be circumvented. In many situations, the variance is confined to some low-dimensional linear subspace that can be identified by using a simpler method as a preprocessing step. Examples of this were given in Subsections 5.2 and 5.3, where KDPCA was applied to a low-dimensional projection obtained by the linear PCA.

Acknowledgements. The author was financially supported by the TUCS Graduate Programme. He would also like to thank Prof. Marko Mäkelä, Doc. Napsu Karmitsa and Prof. Hannu Oja for their valuable comments.

References

[1] CM2.X Coupled Climate Models. http://nomads.gfdl.noaa.gov/CM2.X. visited on 18/11/2013.

[2] The quasi-biennial-oscillation (QBO) data serie. http://www.geo.fu-berlin.de/en/met/ag/strat/produkte/qbo. visited on 18/11/2013.

[3] G. Berkooz, P. Holmes, and J. L. Lumley. The proper orthogonal decomposition in the analysis of turbulent flows. *Annual Review of Fluid Mechanics*, 25:539–575, 1993.

[4] B. Christiansen. The shortcomings of nonlinear principal component analysis in identifying circulation regimes. *Journal of Climate*, 18(22):4814–4823, 2005.

[5] J. Damon. Generic structure of two-dimensional images under Gaussian blurring. *SIAM Journal on Applied Mathematics*, 59(1):97–138, 1998.

[6] T. L. Delworth and coauthors. GFDL's CM2 global coupled climate models. part I: Formulation and simulation characteristics. *Journal of Climate*, 19(5):643–674, 2006.

[7] D. L. Donoho and C. Grimes. Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data. *Proceedings of the National Academy of Sciences*, 100(10):5591–5596, 2003.

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

[9] J. Einbeck, G. Tutz, and L. Evers. Local principal curves. *Statistics and Computing*, 15(4):301–313, 2005.

[10] G. H. Golub and C. F. Van Loan. *Matrix Computations*. The Johns Hopkins University Press, Baltimore, third edition, 1996.

[11] N. Golyandina, V. Nekrutkin, and A. A. Zhigljavsky. *Analysis of Time Series Structure: SSA and related techniques*. Chapman and Hall/CRC Press, Boca Raton, Florida, USA, 2001.

[12] N. E. Golyandina and K. D. Usevich. 2D-extension of singular spectrum analysis: algorithm and elements of theory. In V. Olshevsky and E. Tyrtyshnikov, editors, *Matrix Methods: Theory, Algorithms, Applications*. World Scientific Publishing, Singapore, 2010.

[13] L. Greengard and J. Strain. The fast Gauss transform. *SIAM Journal on Scientific and Statistical Computing*, 12(1):79–94, 1991.

[14] N. J. Higham. *Functions of Matrices: Theory and Computation*. SIAM, Philadelphia, 2008.

[15] W. W. Hsieh. Nonlinear multivariate and time series analysis by neural network methods. *Reviews of Geophysics*, 42(1):1–25, 2004.

[16] W. W. Hsieh and K. Hamilton. Nonlinear singular spectrum analysis of the tropical stratospheric wind. *Quarterly Journal of the Royal Meteorological Society*, 129(592):2367–2382, 2003.

[17] J. W. Jaromczyk and G. T. Toussaint. Relative neighborhood graphs and their relatives. *Proceedings of the IEEE*, 80(9):1502–1517, 1992.

[18] I. T. Jolliffe. *Principal Component Analysis*. Springer-Verlag, Berlin, 1986.

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

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

[21] M. Loève. *Probability theory: foundations, random sequences*. van Nostrand, Princeton, New Jersey, USA, 1955.

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

[23] J. Miller. *Relative Critical Sets in R^n and Applications to Image Analysis*. PhD thesis, University of North Carolina, 1998.

[24] A. H. Monahan. Nonlinear principal component analysis: Tropical Indo–Pacific sea surface temperature and sea level pressure. *Journal of Climate*, 14(2):219–233, 2001.

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

[26] J. M. Ortega. *Numerical Analysis: A Second Course*. SIAM, Philadelphia, 1990.

[27] U. Ozertem and D. Erdogmus. Locally defined principal curves and surfaces. *Journal of Machine Learning Research*, 12:1249–1286, April 2011.

[28] K. Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine Series 6*, 2(11):559–572, 1901.

[29] S. Pulkkinen. Ridge curve approach to extraction of curvilinear structures from noisy data. TUCS Technical Report TR1082, Turku Centre for Computer Science (`http://tucs.fi`), Turku, Finland, 2013.

[30] S. Pulkkinen, M. M. Mäkelä, and N. Karmitsa. A generative model and a generalized trust region Newton method for noise reduction. *Computational Optimization and Applications*, 2013. doi: 10.1007/s10589-013-9581-4.

[31] R. M. Rangayyan. *Biomedical Signal Analysis: A Case-Study Approach*. IEEE Press/Wiley, New York, 2002.

[32] M. Renardy and R. C. Rogers. An introduction to partial differential equations. In J. E. Marsden, L. Sirovich, and S. S. Antman, editors, *Texts in Applied Mathematics*, 13. Springer-Verlag, New York, second edition, 2004.

[33] I. Ross. *Nonlinear Dimensionality Reduction Methods in Climate Data Analysis*. PhD thesis, University of Bristol, United Kingdom, 2008.

[34] I. Ross, P. J. Valdes, and S. Wiggins. ENSO dynamics in current climate models: an investigation using nonlinear dimensionality reduction. *Nonlinear Processes in Geophysics*, 15:339–363, 2008.

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

[36] B. Schölkopf, A. Smola, and K.-R. Müller. Kernel principal component analysis. In W. Gerstner, A. Germond, M. Hasler, and J.-D. Nicoud, editors, *Artificial Neural Networks — ICANN'97*, volume 1327 of *Lecture Notes in Computer Science*, pages 583–588. Springer, Berlin Heidelberg, 1997.

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

[38] R. Vautard, P. Yiou, and M. Ghil. Singular-spectrum analysis: A toolkit for short, noisy chaotic signals. *Physica D: Nonlinear Phenomena*, 58(1-4):95–126, 1992.

[39] B. C. Weare, A. R. Navato, and E. R. Newell. Empirical orthogonal analysis of Pacific sea surface temperatures. *Journal of Physical Oceanography*, 6(5):671–678, 1976.

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

[41] E. F. Whittlesey. Fixed points and antipodal points. *The American Mathematical Monthly*, 70(8):807–821, 1963.

[42] Z. Zhang and H. Zha. Principal manifolds and nonlinear dimensionality reduction via tangent space alignment. *SIAM Journal on Scientific Computing*, 26(1):313–338, 2004.

A Proofs of Technical Results

In this appendix we give proofs of Theorems 3.2 and 3.3. The proofs are carried out by making the following simplifying assumption that can be made without loss of generality.

Assumption A.1. *The sample 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 (6) and (7), that is,

$$\hat{p}_h(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^n K_h(\boldsymbol{x}-\boldsymbol{y}_i) = \frac{1}{(2\pi)^{\frac{d}{2}}h^d n}\sum_{i=1}^n \exp\left(-\frac{\|\boldsymbol{x}-\boldsymbol{y}_i\|^2}{2h^2}\right). \tag{27}$$

The gradient and Hessian of this function are

$$\nabla\hat{p}_h(\boldsymbol{x}) = -\frac{1}{h^2 n}\sum_{i=1}^n (\boldsymbol{x}-\boldsymbol{y}_i)K_{h^2\boldsymbol{I}}(\boldsymbol{x}-\boldsymbol{y}_i) \tag{28}$$

and

$$\nabla^2\hat{p}_h(\boldsymbol{x}) = \frac{1}{h^2 n}\sum_{i=1}^n \left[\frac{(\boldsymbol{x}-\boldsymbol{y}_i)(\boldsymbol{x}-\boldsymbol{y}_i)^T}{h^2} - \boldsymbol{I}\right] K_{h^2\boldsymbol{I}}(\boldsymbol{x}-\boldsymbol{y}_i), \tag{29}$$

respectively. For the logarithm of $\hat{p}_h$ we have the formulae

$$\nabla\log\hat{p}_h(\boldsymbol{x}) = \frac{\nabla\hat{p}_h(\boldsymbol{x})}{\hat{p}_h(\boldsymbol{x})} \quad \text{and} \quad \nabla^2\log\hat{p}_h(\boldsymbol{x}) = \frac{\nabla^2\hat{p}_h(\boldsymbol{x})}{\hat{p}_h(\boldsymbol{x})} - \frac{\nabla\hat{p}_h(\boldsymbol{x})\nabla\hat{p}_h(\boldsymbol{x})^T}{\hat{p}_h(\boldsymbol{x})^2}. \tag{30}$$

In what follows, we establish limiting values for the logarithm of the Gaussian kernel density estimate and its derivatives as the bandwidth h approaches infinity. Furthermore, we show that convergence to these limits is uniform. To this end, we need uniform boundedness of the following functions and existence of a uniform Lipschitz constant.

Proposition A.1. *Let $\hat{p}_h : \mathbb{R}^d \to \mathbb{R}$ be a Gaussian kernel density estimate, let $U \subset \mathbb{R}^d$ be compact and let Assumption A.1 be satisfied. Then there exist constants $M_i > 0$ such that for any $h_0 > 0$ the functions*

$$\begin{aligned} F_0(\boldsymbol{x};h) &= h^2\log\left[(2\pi)^{\frac{d}{2}}h^d\hat{p}_h(\boldsymbol{x})\right],\\ \boldsymbol{F}_1(\boldsymbol{x};h) &= h^2\nabla\log\hat{p}_h(\boldsymbol{x}),\\ \boldsymbol{F}_2(\boldsymbol{x};h) &= h^4\nabla^2\log\hat{p}_h(\boldsymbol{x}) + h^2\boldsymbol{I} \end{aligned}$$

satisfy the conditions

$$\|\boldsymbol{F}_i(\boldsymbol{x};h)\| \le M_i$$

for all $\boldsymbol{x} \in U$, $i = 0, 1, 2$ *and* $h \geq h_0$. *Furthermore, there exist constants* $L_i > 0$ *such that*

$$\|\boldsymbol{F}_i(\boldsymbol{x}; h) - \boldsymbol{F}_i(\boldsymbol{y}; h)\| \leq L_i \|\boldsymbol{x} - \boldsymbol{y}\|$$

for all $\boldsymbol{x}, \boldsymbol{y} \in U$, $i = 0, 1, 2$ *and* $h \geq h_0$.

The proof of the above proposition is omitted due to space constraints. The upper bounds M_i can be obtained by straightforward but rather tedious calculation. Likewise, the Lipschitz constants L_i can be obtained by showing that the derivatives of the functions $\boldsymbol{F}_i(\cdot; h)$ with respect to h are uniformly bounded for all $h \geq h_0 > 0$ in any compact set.

The following result that guarantees uniform convergence of a function sequence is a direct extension of the Arzelà-Ascoli theorem (e.g. [32]). The original formulation of the Arzelà-Ascoli theorem is given for equicontinuous functions, but the result of the theorem also holds when the function sequence is Lipschitz continuous with uniformly bounded Lipschitz constant.

Theorem A.1. *Let* $\{\boldsymbol{f}_k\}_{k \in \mathbb{N}}$ *be a sequence of functions from some set* $\Omega \subset \mathbb{R}^m$ *to* $\mathbb{R}^d$ *and assume that* $\{\boldsymbol{f}_k\}$ *converges (pointwise) to a limit function* $\boldsymbol{f}^*$. *If there exists* $M > 0$ *such that*

$$\|\boldsymbol{f}_k(\boldsymbol{x})\| \leq M$$

and $L > 0$ *such that*

$$\|\boldsymbol{f}_k(\boldsymbol{x}) - \boldsymbol{f}_k(\boldsymbol{y})\| \leq L \|\boldsymbol{x} - \boldsymbol{y}\|$$

for all $k \in \mathbb{N}$, $\boldsymbol{x} \in \Omega$ *and* $\boldsymbol{y} \in \Omega$, *then the sequence* $\{\boldsymbol{f}_k\}$ *converges uniformly to* $\boldsymbol{f}^*$ *in* Ω.

Now we are ready to prove the following result that will be utilized in the subsequent proofs.

Lemma A.1. *Let* $\hat{p}_h : \mathbb{R}^d \to \mathbb{R}$ *be a Gaussian kernel density estimate and assume A.1. Then*

$$\lim_{h \to \infty} h^2 \log \left[(2\pi)^{\frac{d}{2}} h^d \hat{p}_h(\boldsymbol{x}) \right] = -\frac{1}{2n} \sum_{i=1}^{n} \|\boldsymbol{x} - \boldsymbol{y}_i\|^2, \tag{31}$$

$$\lim_{h \to \infty} h^2 \nabla \log \hat{p}_h(\boldsymbol{x}) = -\frac{1}{n} \sum_{i=1}^{n} (\boldsymbol{x} - \boldsymbol{y}_i) = -\boldsymbol{x} \tag{32}$$

and

$$\lim_{h \to \infty} \left[h^4 \nabla^2 \log \hat{p}_h(\boldsymbol{x}) + h^2 \boldsymbol{I} \right] = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{y}_i \boldsymbol{y}_i^T \tag{33}$$

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

Proof. Let $\boldsymbol{x} \in \mathbb{R}^d$. For the logarithm of a sum of Gaussians we have the limit (the proof is omitted)

$$\lim_{h \to \infty} h^2 \log \left[\frac{1}{n} \sum_{i=1}^{n} \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{y}_i\|^2}{2h^2} \right) \right] = -\frac{1}{2n} \sum_{i=1}^{n} \|\boldsymbol{x} - \boldsymbol{y}_i\|^2.$$

Consequently, taking the logarithm of (27) and the limit $h \to \infty$ and using the above limit yields (31).

Substitution of equations (27) and (28) into (30) yields

$$\nabla \log \hat{p}_h(\boldsymbol{x}) = \frac{\nabla \hat{p}_h(\boldsymbol{x})}{\hat{p}_h(\boldsymbol{x})} = -\frac{1}{h^2} \frac{\sum_{i=1}^n (\boldsymbol{x} - \boldsymbol{y}_i) K_{h^2 \boldsymbol{I}}(\boldsymbol{x} - \boldsymbol{y}_i)}{\sum_{i=1}^n K_{h^2 \boldsymbol{I}}(\boldsymbol{x} - \boldsymbol{y}_i)}.$$

The limit (32) follows from multiplying the above expression by h^2, taking the limit $h \to \infty$ and using the limit

$$\lim_{h \to \infty} \exp\left(-\frac{r^2}{2h^2}\right) = 1 \quad \text{for all } r \in \mathbb{R}. \tag{34}$$

On the other hand, from equations (27)–(29) we obtain that

$$\begin{aligned} \frac{\nabla^2 \hat{p}_h(\boldsymbol{x})}{\hat{p}_h(\boldsymbol{x})} &= \frac{1}{h^2} \frac{\sum_{i=1}^n \left[\frac{(\boldsymbol{x}-\boldsymbol{y}_i)(\boldsymbol{x}-\boldsymbol{y}_i)^T}{h^2} - \boldsymbol{I}\right] K_{h^2 \boldsymbol{I}}(\boldsymbol{x} - \boldsymbol{y}_i)}{\sum_{i=1}^n K_{h^2 \boldsymbol{I}}(\boldsymbol{x} - \boldsymbol{y}_i)} \\ &= \frac{1}{h^4} \frac{\sum_{i=1}^n (\boldsymbol{x} - \boldsymbol{y}_i)(\boldsymbol{x} - \boldsymbol{y}_i)^T K_{h^2 \boldsymbol{I}}(\boldsymbol{x} - \boldsymbol{y}_i)}{\sum_{i=1}^n K_{h^2 \boldsymbol{I}}(\boldsymbol{x} - \boldsymbol{y}_i)} - \frac{\boldsymbol{I}}{h^2} \end{aligned}$$

and

$$\frac{\nabla \hat{p}_h(\boldsymbol{x}) \nabla \hat{p}_h(\boldsymbol{x})^T}{\hat{p}_h(\boldsymbol{x})^2} = \frac{1}{h^4} \frac{\left[\sum_{i=1}^n (\boldsymbol{x} - \boldsymbol{y}_i) K_{h^2 \boldsymbol{I}}(\boldsymbol{x} - \boldsymbol{y}_i)\right] \left[\sum_{i=1}^n (\boldsymbol{x} - \boldsymbol{y}_i) K_{h^2 \boldsymbol{I}}(\boldsymbol{x} - \boldsymbol{y}_i)\right]^T}{\left[\sum_{i=1}^n K_{h^2 \boldsymbol{I}}(\boldsymbol{x} - \boldsymbol{y}_i)\right]^2}.$$

Thus, by the limit (34) we obtain

$$\lim_{h \to \infty} \left[h^4 \frac{\nabla^2 \hat{p}_h(\boldsymbol{x})}{\hat{p}_h(\boldsymbol{x})} + h^2 \boldsymbol{I}\right] = \frac{1}{n} \sum_{i=1}^n (\boldsymbol{x} - \boldsymbol{y}_i)(\boldsymbol{x} - \boldsymbol{y}_i)^T \tag{35}$$

and

$$\lim_{h \to \infty} h^4 \frac{\nabla \hat{p}_h(\boldsymbol{x}) \nabla \hat{p}_h(\boldsymbol{x})^T}{\hat{p}_h(\boldsymbol{x})^2} = \frac{1}{n^2} \left[\sum_{i=1}^n (\boldsymbol{x} - \boldsymbol{y}_i)\right] \left[\sum_{i=1}^n (\boldsymbol{x} - \boldsymbol{y}_i)\right]^T. \tag{36}$$

It follows from Assumption A.1 that

$$\frac{1}{n} \sum_{i=1}^n (\boldsymbol{x} - \boldsymbol{y}_i)(\boldsymbol{x} - \boldsymbol{y}_i)^T = \boldsymbol{x}\boldsymbol{x}^T + \frac{1}{n} \sum_{i=1}^n \boldsymbol{y}_i \boldsymbol{y}_i^T \tag{37}$$

and

$$\frac{1}{n^2} \left[\sum_{i=1}^n (\boldsymbol{x} - \boldsymbol{y}_i)\right] \left[\sum_{i=1}^n (\boldsymbol{x} - \boldsymbol{y}_i)\right]^T = \boldsymbol{x}\boldsymbol{x}^T. \tag{38}$$

The limit (33) then follows by substituting (37) and (38) into (35) and (36) and using equation (30). Finally, uniform convergence to the limits (31)–(33) follows from Proposition A.1 and Theorem A.1. □

The following two lemmata facilitate the proof of Theorem 3.2.

Lemma A.2. *Let $\hat{p}_h : \mathbb{R}^d \to \mathbb{R}$ be a Gaussian kernel density estimate and let Assumptions 3.1 and A.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*

$$\lambda_1(\boldsymbol{x}; h) < 0, \tag{39}$$

$$\lambda_i(\boldsymbol{x}; h) \neq \lambda_j(\boldsymbol{x}; h) \tag{40}$$

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 equation (3) corresponding to its r greatest eigenvalues, then for all $\varepsilon > 0$ there exists $h_0 > 0$ such that

$$\|\boldsymbol{W}(\boldsymbol{x}; h) - \boldsymbol{V}\| < \varepsilon \quad \text{for all } \boldsymbol{x} \in U \text{ and } h \geq h_0. \tag{41}$$

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 (33) under Assumption A.1 and continuity of eigenvalues of a matrix as a function of its elements (e.g. [26], Theorem 3.1.2), for all $\varepsilon > 0$ there exists k_0 such that

$$|h_k^4 \lambda_i(\boldsymbol{x}; h_k) + h_k^2 - \frac{n-1}{n}\tilde{\lambda}_i| < \varepsilon \tag{42}$$

for all $i = 1, 2, \ldots, r+1$, $\boldsymbol{x} \in U$ and $k \geq k_0$. Consequently, condition (39) holds for all $\boldsymbol{x} \in U$ for any sufficiently large h by Assumption 3.1. It also follows from Assumption 3.1, condition (42) 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

$$h_k^4 |\lambda_i(\boldsymbol{x}; h_k) - \lambda_j(\boldsymbol{x}; h_k)| \geq \big||h_k^4 \lambda_i(\boldsymbol{x}; h_k) - h_k^2| - |h_k^4 \lambda_j(\boldsymbol{x}; h_k) - h_k^2|\big| > \frac{n-1}{n}\varepsilon$$

for all $\boldsymbol{x} \in U$ and $k \geq k_1$. This implies condition (40). Similarly, condition (41) follows from uniform convergence to the limit (33) under Assumption A.1, condition (40) and continuity of eigenvectors as a function of matrix elements when the eigenvalues are distinct (e.g. [26], Theorem 3.1.3). □

Lemma A.3. *Assume 3.1 and A.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 A.2, and the set S_∞^r as in Theorem 3.2. Then the limit

$$\lim_{h\to\infty} h^2 \|\tilde{\boldsymbol{W}}(\boldsymbol{x}; h) \nabla \log \hat{p}_h(\boldsymbol{x})\| \tag{43}$$

exists for all $\boldsymbol{x} \in \mathbb{R}^d$. Furthermore, $\boldsymbol{x} \in S_\infty^r$ if and only if the limit (43) is zero.

Proof. By the limits (32) and (41) the limit (43) exists for all $\boldsymbol{x} \in \mathbb{R}^d$. Furthermore, for any $\boldsymbol{x} \in \mathbb{R}^d$, the condition that the limit (43) 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 A.2. By the orthogonality of the vectors $\boldsymbol{v}_i$, the definition of the set S_∞^r and Assumption A.1, this condition is equivalent to the condition that $\boldsymbol{x} \in S_\infty^r$. □

For the proof of Theorem 3.2, we define the set

$$S_h^r = \{\boldsymbol{x} \in \mathbb{R}^d \mid \|\tilde{\boldsymbol{W}}(\boldsymbol{x}; h) \nabla \log \hat{p}_h(\boldsymbol{x})\| = 0\}, \tag{44}$$

where the function $\tilde{\boldsymbol{W}}$ is defined as in Lemma A.3. Under Assumption A.1, we prove both claims of Theorem 3.2 by using the following two lemmata.

Lemma A.4. *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 3.1 and A.1 are satisfied, then for all $\varepsilon > 0$ there exists $h_0 > 0$ such that*

$$\sup_{\boldsymbol{x} \in S_h^r \cap U} \inf_{\boldsymbol{y} \in S_\infty^r} \|\boldsymbol{x} - \boldsymbol{y}\| < \varepsilon \quad \textit{for all } h \geq h_0. \tag{45}$$

Proof. The proof is by contradiction. Let $0 \leq r < d$ and let $U \subset \mathbb{R}^d$ be a compact set such that $U \cap S_\infty^r \neq \emptyset$. Assume that there exists $\varepsilon_1 > 0$ such that for all $h_0 > 0$ there exists $h \geq h_0$ such that condition (45) is not satisfied. This implies that for all $h_0 > 0$ there exists $h \geq h_0$ such that

$$\inf_{\boldsymbol{y} \in S_\infty^r} \|\boldsymbol{x} - \boldsymbol{y}\| \geq \varepsilon_1 \quad \text{for some } \boldsymbol{x} \in S_h^r \cap U. \tag{46}$$

Let $\{\boldsymbol{x}_k\}$ denote a sequence of such points $\boldsymbol{x}$ with the corresponding sequence h_k. Since the set $S_h^r \cap U$ is compact by the compactness of U and the continuity of $\tilde{\boldsymbol{W}}(\cdot, h)$ 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 $\boldsymbol{z}^* \notin S_\infty^r$ by condition (46). Thus, by Lemma A.3 we deduce that for some $c > 0$,

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

In view of the definition (44), the above limit implies that there exists $\varepsilon_2 > 0$ and k_0 such that for all $k \geq k_0$,

$$\|\boldsymbol{F}(\boldsymbol{z}^*; h_k) - \boldsymbol{F}(\boldsymbol{y}; h_k)\| \geq \varepsilon_2 \quad \text{for all } \boldsymbol{y} \in S_{h_k}^r \cap U. \tag{47}$$

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

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

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 A.1 and A.2), we deduce from (47) that for all $\varepsilon_2 > \varepsilon_3 > 0$ there exists k_1 such that

$$\|\boldsymbol{z}^* - \boldsymbol{y}\| + \varepsilon_3 \geq \|\boldsymbol{F}(\boldsymbol{z}^*; h_k) - \boldsymbol{F}(\boldsymbol{y}; h_k)\| \geq \varepsilon_2 \tag{48}$$

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

Condition (48) implies that for all $0 < \varepsilon_3 < \varepsilon_2$ there exists k_1 such that

$$\inf_{\boldsymbol{y} \in S^r_{h_k} \cap U} \|\boldsymbol{z}^* - \boldsymbol{y}\| \geq \varepsilon_2 - \varepsilon_3 \quad \text{for all } k \geq k_1. \tag{49}$$

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 (49). 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 A.5. *Let $\hat{p}_h$ be a Gaussian kernel density estimate, let $0 \leq r < d$, let Assumptions 3.1 and A.1 be satisfied and define the set S^r_∞ as in Theorem 3.2. 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*

$$\sup_{\boldsymbol{x} \in S^r_\infty \cap U} \inf_{\boldsymbol{y} \in \mathcal{R}^r_{\log \hat{p}_h}} \|\boldsymbol{x} - \boldsymbol{y}\| < \varepsilon \quad \textit{for all } h \geq h_0. \tag{50}$$

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 3.1. 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 $\text{span}(\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_r)$.

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

$$\boldsymbol{U}(\boldsymbol{x};h) = \arg\min_{\boldsymbol{U}' \in Q_{\boldsymbol{x},h}} \|\boldsymbol{U}' - \bar{\boldsymbol{V}}\|_F. \tag{51}$$

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

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

$O(d, d-r)$ denotes a $d \times (d-r)$ matrix having orthonormal columns and the matrix $\boldsymbol{W}(\boldsymbol{x};h)$ is defined as in Lemma A.2. It can be shown that the function $\boldsymbol{U}(\cdot;h)$ is well-defined for any $h > 0$.[1] 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 (40) is satisfied in such a set by Lemma A.2.

The above definitions and condition (41) 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 (32) as $h \to \infty$ by Lemma A.3 together with the property that

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

following from Assumption A.1 implies that for all $\varepsilon_1, \varepsilon_2 > 0$ there exists $h_0 > 0$ such that

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

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 $\boldsymbol{x}_0 \in S^r_\infty \cap U$ 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. [41]) 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 $\boldsymbol{x}_0 \in S^r_\infty \cap U$ 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 $\boldsymbol{x}_0 \in S^r_\infty \cap U$ condition (5a) 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 (5b) and (5c) are satisfied in the compact set D_ε for all sufficiently large h by conditions (39) and (40). Hence, we have proven that for all $\varepsilon > 0$ condition (50) holds for all sufficiently large h. □

[1] Problem (51) can be equivalently formulated as an *orthogonal Procrustes problem*. With the matrices defined above, this problem has a unique solution (e.g. [14]).

***Proof of Theorem 3.2** (page 7).* Follows directly from Lemmata A.4 and A.5 by the property that $\mathcal{R}^r_{\hat{p}_h} = \mathcal{R}^r_{\log \hat{p}_h} \subseteq S^r_h$ for all $0 \leq r < d$ and $h > 0$ by Lemma 3.1 and Definition 3.1. □

Next, we prove Theorem 3.3 under Assumption A.1 by using the following lemma.

Lemma A.6. *Let $\hat{p}_h$ be a Gaussian kernel density estimate, assume A.1 and define the set*

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

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

Proof. The proof is by contradiction. Assume that for all $r > r_0 = \max_{i=1,2,\dots,n} \|\boldsymbol{y}_i\|$ and $h_0 > 0$ there exists $h \geq h_0$ such that $\boldsymbol{x} \in U_h \setminus B(\mathbf{0}; r)$. 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

$$\|\boldsymbol{x}_k\| > r_k > r_0 = \max_{i=1,2,\dots,n} \|\boldsymbol{y}_i\| \quad \text{for all } k \geq k_0 \tag{52}$$

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

$$\log \hat{p}_{h_k}(\boldsymbol{x}_k) \geq \log \hat{p}_{h_k}(\boldsymbol{y}_j) \quad \text{for some } j \in \{1, 2, \dots, n\}. \tag{53}$$

By Assumption A.1 and condition (52) we have that $\|\boldsymbol{x}_k - \boldsymbol{y}_i\| \geq \|\boldsymbol{x}_k\| - r_0$ for all $k \geq k_0$ and $i = 1, 2, \dots, n$. 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 (27), this implies that

$$h_k^2 [\log \hat{p}_{h_k}(\boldsymbol{x}_k) + \log [(2\pi)^{\frac{d}{2}} h_k^d]] \leq -\frac{(\|\boldsymbol{x}_k\| - r_0)^2}{2} \tag{54}$$

for all $k \geq k_0$. On the other hand, by the limit (31), Assumption A.1 and the choice of r_0 we have

$$\begin{aligned} \lim_{k \to \infty} h_k^2 [\log \hat{p}_{h_k}(\boldsymbol{y}_j) + \log [(2\pi)^{\frac{d}{2}} h_k^d]] &= -\frac{1}{2n} \sum_{i=1}^{n} \|\boldsymbol{y}_j - \boldsymbol{y}_i\|^2 \\ = -\frac{1}{2n} \left(\sum_{i=1}^{n} \|\boldsymbol{y}_j\|^2 - 2 \sum_{i=1}^{n} \boldsymbol{y}_j^T \boldsymbol{y}_i + \sum_{i=1}^{n} \|\boldsymbol{y}_i\|^2 \right) &\geq -r_0^2 \end{aligned} \tag{55}$$

for all $j = 1, 2, \dots, n$. Plugging the limits (54) and (55) into inequality (53) then leads to a contradiction for any sufficiently large k since $\lim_{k \to \infty} \|\boldsymbol{x}_k\| = \infty$ by condition (52) and the assumption that the sequence $\{r_k\}$ is monotoneusly increasing. □

***Proof of Theorem 3.3** (page 11).* By Lemma A.6 there exists $r > \max_{i=1,2,\dots,n} \|\boldsymbol{y}_i\|$ such that $U_h \subseteq B(\mathbf{0}; r)$ for all sufficiently large h. Thus, condition (20) for all $\boldsymbol{x} \in U_h$ and such h follows from Assumption 3.1, compactness of the set $B(\mathbf{0}; r)$ and Lemma A.2. Finally, 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 (39). □

Turku
Centre *for*
Computer
Science

Joukahaisenkatu 3-5 A, 20520 TURKU, Finland | www.tucs.fi

University of Turku
Faculty of Mathematics and Natural Sciences
- Department of Information Technology
- Department of Mathematics

Turku School of Economics
- Institute of Information Systems Sciences

Åbo Akademi University
- Department of Computer Science
- Institute for Advanced Management Systems Research

ISBN 978-952-12-2970-1
ISSN 1239-1891