OSE SEMINAR 2014

Ridge-Based Methods and Applications to Spatiotemporal Data

Seppo Pulkkinen

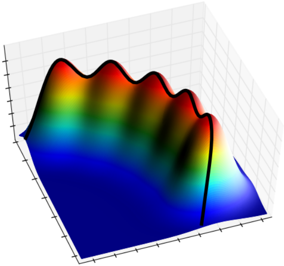
University of Turku, Department of Mathematics and Statistics

ÅBO, NOVEMBER 14 2014

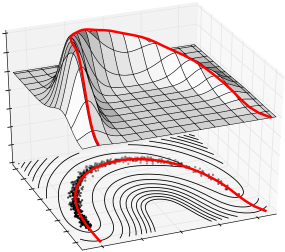
OPTIMIZATION AND SYSTEMS ENGINEERING

Åbo Akademi University

Function Ridges



(a) general function



(b) density of a point set

- A *ridge* is an elevated region of a function surface passing through its peaks.
- Density ridges correspond to the underlying structure of a point set when the observations follow a *generative model*.

Ridge Definition

- A ridge point is a local maximum in the *subspace* spanned by the Hessian eigenvectors $\{\boldsymbol{v}_i(\cdot)\}_{i=m+1}^d$ corresponding to the $d-m$ smallest eigenvalues $\{\lambda_i(\cdot)\}_{i=m+1}^d$.
- The eigenvectors $\{\boldsymbol{v}_i(\cdot)\}_{i=m+1}^d$ correspond to the directions of greatest negative curvature.

Definition

Let $f : \mathbb{R}^d \to \mathbb{R}$ be a twice differentiable function and let $0 \leq m < d$. A point $\boldsymbol{x} \in \mathbb{R}^d$ belongs to the m-dimensional *ridge set* $\mathcal{R}_f^m$ if

$$\begin{aligned} \nabla f(\boldsymbol{x})^T \boldsymbol{v}_i(\boldsymbol{x}) &= 0, \quad \text{for all } i > m, \\ \lambda_{m+1}(\boldsymbol{x}) &< 0, \\ \lambda_1(\boldsymbol{x}) > \lambda_2(\boldsymbol{x}) > \cdots > \lambda_{m+1}(\boldsymbol{x}), &\quad \text{if } m > 0, \end{aligned}$$

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 $\nabla^2 f(\boldsymbol{x})$, respectively.

Generative Model

- The observations are assumed to follow a generative model

$$\boldsymbol{X} \sim \boldsymbol{f}(\Theta) + \boldsymbol{\varepsilon},$$

  where
  - $\boldsymbol{f} : \mathbb{R}^m \rightarrow \mathbb{R}^d$ is a generating function, $m < d$,
  - Θ follows some distribution in $\mathcal{D} \subset \mathbb{R}^m$,
  - $\boldsymbol{\varepsilon} \sim \mathcal{N}_d(\mathbf{0}, \sigma^2)$.
- The above model induces the *marginal density*

$$p_{\boldsymbol{X}}(\boldsymbol{x}) = C_{\sigma,d} \int_{\mathcal{D}} p_{\boldsymbol{X}}(\boldsymbol{x} \,|\, \Theta = \boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta}$$

  with some constant $C_{\sigma,d}$.
- The model can be extended to contain multiple generating functions.
- Assuming the above model, ridges of the marginal density can be used as an estimate for the generating functions.

Kernel Density Estimation

- In practice, the marginal density $p_{\boldsymbol{X}}$ is not known a priori. However, it can be estimated *nonparametrically* from the observations.

Definition

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

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

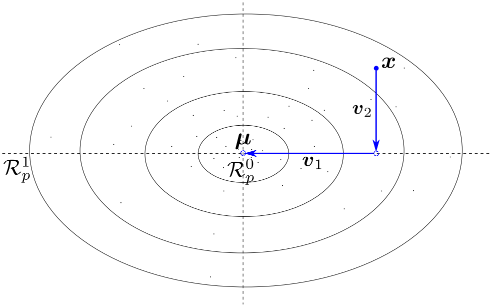
where the kernel $K_{\boldsymbol{H}} : \mathbb{R}^d \to]0, \infty[$ is the Gaussian function

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

with a symmetric and positive definite kernel bandwidth matrix $\boldsymbol{H} \in \mathbb{R}^{d \times d}$.

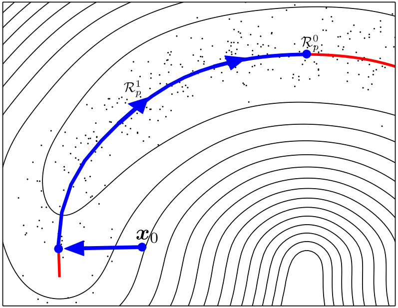
- Existing methods can be used for determining an optimal *bandwidth* matrix $\boldsymbol{H}$ (e.g. the ks package for R).

Successive Ridge Projections: the Linear Case and PCA



- When p is a normal density with mean $\boldsymbol{\mu}$ and symmetric and positive definite covariance matrix $\boldsymbol{\Sigma}$, we have
  - $\mathcal{R}_p^0 = \{\boldsymbol{\mu}\}$ and $\mathcal{R}_p^1 = \boldsymbol{\mu} + \mathrm{span}(\boldsymbol{v}_1)$,
  - $\nabla \log p(\boldsymbol{x}) = -\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})$ and $\nabla^2 \log p(\boldsymbol{x}) = -\boldsymbol{\Sigma}^{-1}$.
- The first step of the Newton iteration restricted to each subspace $\mathrm{span}(\boldsymbol{v}_{m+1}, \boldsymbol{v}_{m+2}, \ldots, \boldsymbol{v}_d)$ yields a ridge point $\boldsymbol{x}^* \in \mathcal{R}_p^m$.
- We obtain the *principal components* of a given point set by replacing the mean and covariance with their sample estimates.

The Nonlinear Case: Differential Equation Formulation

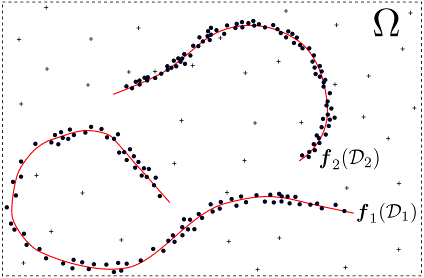


- As in the linear case, the principal component coordinates of a point can be obtained by successive projections onto lower-dimensional ridge sets of the underlying density p (or its estimate $\hat{p}_H$).
- This gives rise to a nonlinear extension of PCA that we call KDPCA (kernel density PCA).
- Ridge projections can be obtained by seeking for maxima along curves $\boldsymbol{\gamma}_m$, with $m = d-1, d-2, \ldots, 1$, that are solutions to

$$\frac{d}{dt}\left\{\left[\sum_{i=1}^{m} \boldsymbol{v}_i(\boldsymbol{\gamma}_m(t))\boldsymbol{v}_i(\boldsymbol{\gamma}_m(t))^T\right]\nabla \log \hat{p}_{\boldsymbol{H}}(\boldsymbol{\gamma}_m(t))\right\} = \boldsymbol{0}, \qquad t \geq 0,$$

$$\boldsymbol{\gamma}_m(0) = \boldsymbol{x}_0.$$

Multiple Generating Functions



- It is straightforward to extend the model to multiple generating functions.
- Difficulties arise in the presence of intersections.
- The conditions defining a boundary of a ridge set $\mathcal{R}_p^m$ are:
  - $\lambda_i(\boldsymbol{x}) = \lambda_j(\boldsymbol{x})$ for some $i \neq j$ such that $0 \leq i < j \leq m$
  - $\lambda_i \geq 0$ for some $i > m$.
- These conditions need to be tested in the ridge tracing algorithm (also third derivative conditions are needed because we are computing derivatives of eigenvectors).

Subspace-Constrained Trust Region Newton Method

- Ridge projections are done by using a *trust region* Newton method as the corrector in a predictor-corrector method.
- As in the classical trust region method (Moré and Sorensen), the idea is to maximize the quadratic model

$$Q_k(\boldsymbol{s}) = \log \hat{p}_{\boldsymbol{H}}(\boldsymbol{x}_k) + \nabla \log \hat{p}_{\boldsymbol{H}}(\boldsymbol{x}_k)^T \boldsymbol{s} + \frac{1}{2} \boldsymbol{s}^T \nabla^2 \log \hat{p}_{\boldsymbol{H}}(\boldsymbol{x}_k) \boldsymbol{s}.$$

- At each iteration, the method solves the *subspace-constrained* trust region subproblem

$$\max_{\boldsymbol{s}} Q_k(\boldsymbol{s}) \quad \text{s.t.} \quad \begin{cases} \|\boldsymbol{s}\| \leq \Delta_k, \\ \boldsymbol{s} \in S_m(\boldsymbol{x}_k), \end{cases}$$

  where

$$S_m(\boldsymbol{x}_k) = \operatorname{span}(\boldsymbol{v}_{m+1}(\boldsymbol{x}_k), \boldsymbol{v}_{m+2}(\boldsymbol{x}_k), \ldots, \boldsymbol{v}_d(\boldsymbol{x}_k)).$$

- In addition to finding maxima, the method finds m-dimensional ridge points. It does an approximate projection in a curvilinear coordinate system.

Comparison to the mean-shift method

- So far, the *mean-shift* method has been the standard approach to finding maxima and ridges of kernel densities.
- The mean-shift iteration is defined as

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \boldsymbol{s}_k, \quad \text{where } \boldsymbol{s}_k = \boldsymbol{f}_{\boldsymbol{H}}(\boldsymbol{x}_k) - \boldsymbol{x}_k$$

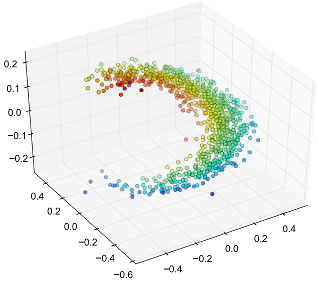
  and

$$\boldsymbol{f}_{\boldsymbol{H}}(\boldsymbol{x}) = \frac{\sum_{i=1}^{N} K_{\boldsymbol{H}}(\boldsymbol{x} - \boldsymbol{y}_i)\boldsymbol{y}_i}{\sum_{i=1}^{N} K_{\boldsymbol{H}}(\boldsymbol{x} - \boldsymbol{y}_i)}.$$

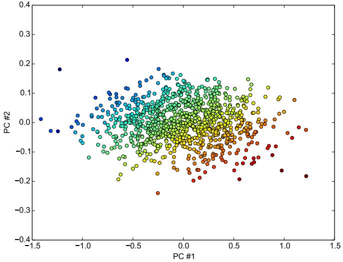
- This fixed-point iteration has (sub)linear convergence rate.
- The mean-shift method can also be constrained to an eigenvector subspace.
- On the other hand, the proposed Newton-based method:
  - has superlinear convergence rate.
  - can be proven to converge to a ridge point.

Dimensionality Reduction with KDPCA

- **Task:** Find a low-dimensional representation of a point set so that its structure is preserved.
- **Example:** a point set sampled from a two-dimensional manifold with noise and the coordinates recovered by using KDPCA.



(a) three-dimensional point set



(b) two-dimensional coordinates

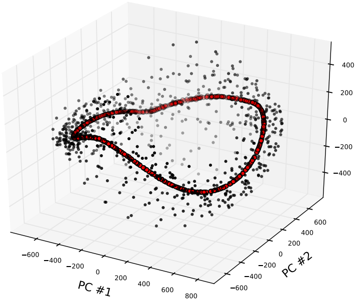
Application of KDPCA to Time Series Data (KDSSA)

- The *phase space* trajectory of a time series $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ is given by

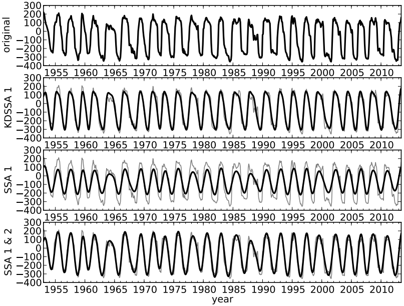
$$\boldsymbol{Y}_{x,L} = \begin{bmatrix} x_1 & x_2 & x_3 & \cdots & x_L \\ x_2 & x_3 & x_4 & \cdots & x_{L+1} \\ x_3 & x_4 & x_5 & \cdots & x_{L+2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{n-L+1} & x_{n-L+2} & x_{n-L+3} & \cdots & x_n \end{bmatrix},$$

  where L is a user-supplied time window length.
- In the classical *singular spectrum analysis)* (SSA), the linear PCA is applied to the trajectory matrix.
- KDPCA can be applied to the trajectory matrix as well. This gives rise to the KDSSA method (kernel density singular spectrum analysis).

Application of KDPCA to Time Series Data (KDSSA)



(a) phase space trajectory and its ridge projection

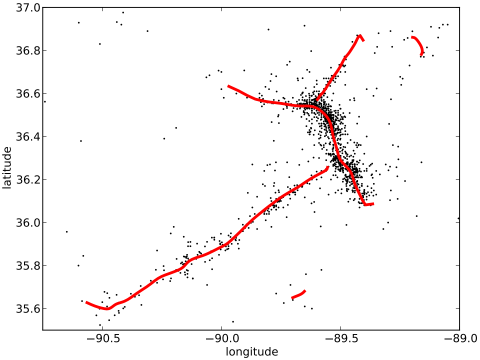


(b) the original time series and its first KDSSA and SSA components

- KDSSA can identify closed loops in phase space that are typical for *quasiperiodic* time series (periodic time series with noise).
- It can be used for extraction of periodic components from such time series, which is not possible by using the linear SSA.

Extraction of Curvilinear Structures from Spatial Data

- **Task:** Find the curvilinear structures from a low-dimensional but large spatial point set (> 10000 samples).
- **Example:** Identification of fault lines from an earthquake catalog.



Conclusions

Main contributions so far:

- A rapidly converging trust region Newton method for projecting a point onto a ridge of the underlying density.
- A robust and efficient method for finding curvilinear structures for noisy data.
- A novel nonlinear extension of the linear principal component analysis based on kernel density ridges.

Literature

- S. Pulkkinen and M.M. Mäkelä and N. Karmitsa (2014).
  A generative model and a generalized trust region Newton method for noise reduction.
  Computational Optimization and Applications, **57**(1):129-165
- S. Pulkkinen (2015).
  Ridge-based method for finding curvilinear structures from noisy data.
  Computational Statistics and Data Analysis, **82**:89-109
- S. Pulkkinen (2014).
  Nonlinear kernel density principal component analysis with application to climate data.
  Statistics and Computing, to appear

The end of the presentation

Thank you for listening!

The end of the presentation

Thank you for listening!

Questions?