Computational Statistics and Data Analysis 82 (2015) 89–109

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Ridge-based method for finding curvilinear structures from noisy data

Seppo Pulkkinen *

TUCS - Turku Centre for Computer Science, Turku, Finland
Department of Mathematics and Statistics, University of Turku, 20014 Turku, Finland

ARTICLE INFO

Article history:
Received 11 September 2013
Received in revised form 30 May 2014
Accepted 16 August 2014
Available online 26 August 2014

Keywords:
Principal curve
Filament
Generative model
Ridge curve
Density estimation
Predictor–corrector method

ABSTRACT

Extraction of curvilinear structures from noisy data is an essential task in many application fields such as data analysis, pattern recognition and machine vision. The proposed approach assumes a random process in which the samples are obtained from a generative model. The model specifies a set of generating functions describing curvilinear structures as well as sampling noise and background clutter. It is shown that ridge curves of the marginal density induced by the model can be used to estimate the generating functions. Given a Gaussian kernel density estimate for the marginal density, ridge curves of the density estimate are parametrized as the solution to a differential equation. Finally, a predictor–corrector algorithm for tracing the ridge curve set of such a density estimate is developed. Efficiency and robustness of the algorithm are demonstrated by numerical experiments on synthetic datasets as well as observational datasets from seismology and cosmology.

1. Introduction

Detection and extraction of curvilinear structures from noisy data is an essential task in many practical applications: extraction of blood vessels that form filament- and tree-like structures is an important task in medical imaging (e.g. Baş, 2011; Flasque et al., 2001; Hoover et al., 2000); in cosmological data, stars and galaxies form filament-like patterns (e.g. Jang, 2006; Novikov et al., 2006; Sousbie et al., 2008), and in astronomy, detection of solar flares involves finding filaments from solar images (e.g. Shih and Kowalski, 2003). Identification of curvilinear structures from noisy data with background clutter is a typical task in remote sensing (e.g. Banfield and Raftery, 1992; Martínez and Ludeña, 2011) and seismology (e.g. Dasgupta and Raftery, 1998; Stanford and Raftery, 2000). Other applications, where fitting curves to noisy data is an important task, include freeway traffic modeling (e.g. Chen et al., 2004; Einbeck and Dwyer, 2011), process monitoring (e.g. Dong and McAvoy, 1996), path estimation from GPS tracks (e.g. Brunsdon, 2007) and shape analysis in computer graphics (e.g. Su et al., 2013).

One of the most well-known approaches to extract curvilinear structures from noisy data is to use the so-called *principal curves*. This approach dates back to Hastie (1984) and Hastie and Stuetzle (1989). A principal curve is defined as a curve passing through the "middle" of the data in a certain sense. Further variations of the principal curve approach have been developed, for instance, by Kégl et al. (2000), Kégl and Krzyzak (2002) and Tibshirani (1992). All of these approaches, however, make rather restrictive assumptions. For instance, they attempt to fit a single curve with no self-intersections, or as the method of Kégl and Krzyzak (2002), require complicated parameter adjustments when self-intersecting or multiple curves are sought from the data.

* Correspondence to: Department of Mathematics and Statistics, University of Turku, 20014 Turku, Finland. Tel.: +358 2 3335686.
E-mail address: seppo.pulkkinen@utu.fi.

http://dx.doi.org/10.1016/j.csda.2014.08.007
0167-9473/

In order to overcome the limitations of the original principal curve definition, locally defined variants of a principal curve have been proposed (e.g. Delicado, 2001; Delicado and Huerta, 2003; Einbeck et al., 2005; Genovese et al., 2009, 2012; Ozertem and Erdogmus, 2011). This paper extends an earlier paper by Pulkkinen et al. (2014) refining the ideas presented by Ozertem and Erdogmus (2011). The key idea in these two papers is to estimate the probability density from given data and extract curvilinear structures from the data from *ridge curves* of the density estimate. Since the definition of a ridge is based only on local derivative information, this approach does not suffer from the limitations of the earlier approaches. For the projection of a sample point onto a ridge, a subspace-constrained variant of the standard mean-shift method (e.g. Cheng, 1995; Comaniciu and Meer, 2002) was proposed by Ozertem and Erdogmus (2011). An improved Newton-based method for this purpose was developed by Pulkkinen et al. (2014). Recently, some extensions of ridge-based methods have been made for the more difficult problem of parametrization of principal curves by iteratively tracing ridge curves of the density (e.g. Baş and Erdogmus, 2011; Baş, 2011; Baş et al., 2012).

A *generative model* for describing a random process that generates a noisy point set containing curvilinear structures was proposed by Pulkkinen et al. (2014). In the model, the data points are assumed to be sampled from a set of *generating functions* with additive noise. In this paper the model is extended to include background clutter being often present in practical applications. Furthermore, it is demonstrated by Pulkkinen et al. (2014) that ridge curves of the *marginal density* induced by the model can be used to estimate the underlying generating functions. Differently to the earlier local principal curve approaches, where no statistical assumptions are made about the data-generating process, the proposed model provides a more statistically oriented approach and a tool for assessing the bias of the principal curve estimate.

For a computational implementation of the ridge curve approach, we consider *nonparametric* estimation of the marginal density by using *Gaussian* kernels (e.g. Scott, 1992). This approach allows to estimate the density directly from the samples with no prior knowledge on the data-generating process, which is often the case in real-world tasks. Based on the recent developments in multivariate density estimation (e.g. Duong, 2007), we also discuss how to automatically choose the kernel *bandwidth* since this step is crucial for the practical applicability of the method.

A major contribution of this paper is the development of a computationally efficient and robust algorithm for tracing ridge curves of a Gaussian kernel density estimate. Adapting the theory of *gradient extremals* from theoretical chemistry (e.g. Hoffman et al., 1986), it is shown that a ridge curve can be parametrized by tracing a solution curve of a differential equation. A predictor–corrector algorithm is developed for this purpose. The algorithm first finds a set of modes (maxima) of the density, and starting from each mode iteratively traces the ridge curve passing through it. Since the choice of the mode-finding and corrector methods largely determines the performance of the algorithm, the trust region Newton method developed by Pulkkinen et al. (2014) is utilized for these purposes. This choice is motivated by the results of Pulkkinen et al. (2014) showing that the Newton-based method is not only more efficient than the mean-shift method and its subspace-constrained variant but also converges to a ridge point or mode under mild assumptions.

The main difficulty in tracing ridge curves is that they can have a very complex structure. Differently to the earlier ridge-based principal curve methods by Baş et al., where this issue was not considered in detail, a rigorous treatment for detection of different types of singular points along a ridge curve is given. The analysis is based on the theory of ridge curves from digital image processing (e.g. Eberly, 1996). In addition, we discuss some strategies for choosing the starting points. These considerations arise when the input data has multiple, possibly intersecting curvilinear structures. The proposed approach is also more robust than the one by Einbeck et al. (2005), where such issues were not rigorously treated, or the heuristic graph-based approach by Delicado and Huerta (2003).

The remaining part of this paper is organized as follows. In Section 2 we describe the generative model and discuss how to use ridge curves to estimate the generating functions. Sections 3 and 4 are devoted to the development of the ridge tracing algorithm. In Section 5 we demonstrate the performance and reliability of the proposed algorithm on synthetically generated point sets as well as two observational datasets from seismology and cosmology. Finally, Section 6 summarizes this paper with concluding remarks.

2. Probabilistic model and density estimation

In this section we recall the generative model from Pulkkinen et al. (2014) and extend it to include background clutter. The model describes a process for generating a noisy point set containing curvilinear structures (related models have also been considered by Genovese et al., 2009, 2012 and Tibshirani, 1992). Since our aim is to make as few parametric assumptions on the data-generating model as possible, we consider the marginal density induced by the model. For estimation of the curvilinear structures from the marginal density, we define the concept of a ridge curve. Finally, for a computational implementation of this approach we consider nonparametric estimation of the marginal density by using Gaussian kernels.

2.1. The model

In the model, the sample points fall into two distinct categories. A sample either belongs to some curvilinear structure, that we call a *filament*, or is background clutter. The type of a sample point is modeled by the random variable

$$T = \begin{cases} 1, & \text{if the sample belongs to a filament,} \\ 0, & \text{if the sample is background clutter} \end{cases}$$