Hindawi Publishing Corporation
Journal of Probability and Statistics
Volume 2016, Article ID 1285026, 9 pages
http://dx.doi.org/10.1155/2016/1285026

Research Article

Exploratory Methods for the Study of Incomplete and Intersecting Shape Boundaries from Landmark Data

Fathi M. O. Hamed[1] **and Robert G. Aykroyd**[2]

[1] *University of Benghazi, Benghazi, Libya*
[2] *University of Leeds, Leeds, UK*

Correspondence should be addressed to Robert G. Aykroyd; r.g.aykroyd@leeds.ac.uk

Received 25 July 2016; Revised 10 October 2016; Accepted 12 October 2016

Academic Editor: Z. D. Bai

Structured spatial point patterns appear in many applications within the natural sciences. The points often record the location of key features, called landmarks, on continuous object boundaries, such as anatomical features on a human face. In other situations, the points may simply be arbitrarily spaced marks along a smooth curve, such as on handwritten numbers. This paper proposes novel exploratory methods for the identification of structure within point datasets. In particular, points are linked together to form curves which estimate the original shape from which the points are the only recorded information. Nonparametric regression methods are applied to polar coordinate variables obtained from the point locations and periodic modelling allows closed curves to be fitted even when data are available on only part of the boundary. Further, the model allows discontinuities to be identified to describe rapid changes in the curves. These generalizations are particularly important when the points represent shapes which are occluded or are intersecting. A range of real-data examples is used to motivate the modelling and to illustrate the flexibility of the approach. The method successfully identifies the underlying structure and its output could also be used as the basis for further analysis.

1. Introduction

Many scientific investigations involve the recording of spatially located data. This data might summarize objects within an image as digitized versions of continuous curves. Once the data are collected often the original context is lost and the aim of the analysis is to identify which points are associated with each other and to link the points to reconstruct the original shape. These can then be seen as estimates of continuous curves and object outlines. If the original scene contains multiple structures, then the analysis must also divide the points into groups with separate curves used to describe the points in each group. It is important to note that this is likely to form only the first part of an analysis and hence can be seen as exploratory data analysis.

This paper looks at the use of smoothing splines to identify and describe geometric patterns in sets of points. It is assumed that the points lie on smooth curves but that a dataset may contain multiple intersecting curves. It is vital that this be done in a nonparametric way so that the widest possible range of patterns can be highlighted. In general, these are closed, or nearly closed, curves and so a transformation to polar coordinates is used to simplify the analysis. Intersecting curves are described by allowing discontinuities in the fitted curves. These procedures are illustrated using simulated data and varied real datasets describing human faces, gorilla skulls, handwritten number 3's, and an archaeological site. These provide a wide variety of point patterns and reinforce the general usefulness of the proposed methods. For mathematical detailed description and applications of shape-based analysis of points, refer to, for example, Batschelet [1], Bookstein [2], Dryden and Mardia [3], and Lele and Richtsmeier [4].

To allow for this wide variety of possible curves a nonparametric fitting approach, such as splines, can be used (see, e.g., [5, 6]). The flexibility is helpful in the exploratory statistical analysis of a dataset, and the results can be used to suggest parametric equations for later analysis. Nonparametric regression is the general name for a range of curve fitting techniques which make few a priori assumptions about the true shape. In nonparametric regression, several different families of basis functions can be used to describe curves; one of the common kinds of basis for smooth curves is the spline. Splines are generally defined as piecewise polynomials in which curve, or line, segments are joined together to form a continuous function. The spline smoothing approach to nonparametric regression is discussed, for example, by Silverman [7] and extended to deal with branching curves by defining a roughness penalty by Silverman and Wood [8]. For an introduction to natural cubic spline see Green and Silverman [9]. For more review of spline methods in statistics see Wegman and Wright [10], Silverman [11], Silverman [7], Nychka [12], and Wahba [13].

It is important to note that there are many existing general frameworks for performing spline-based regression. For example, multivariate adaptive regression splines (MARS) [14] or its more robust generalizations, RMARS [15] and RCMARS [16], with a good overview and comparison in [17]. These follow the general approach of general additive modelling [18] and give a formal framework for fitting and model selection.

A brief introduction to splines, along with the extension to circular data, is given in Section 2. The main results of this paper are given in Section 3 by considering modelling for single curves with occlusions and multiple intersecting curves. Although simulated examples are used to illustrate, the main real-data examples are given in Section 4. General discussion appears in Section 5.

2. Nonparametric Curve Estimation and Periodic Splines

A smoothing spline is a nonparametric curve estimator that is defined as the solution to a minimization problem. It provides a flexible smooth function for situations in which a simple polynomial or nonlinear regression model is not suitable. For a set of n observations $\chi = \{(x_i, y_i),\ i = 1, 2, \dots, n\}$ consider a regression problem where the observations are assumed to satisfy

$$y_i = f(x_i) + \epsilon_i, \quad i = 1, 2, \dots, n, \tag{1}$$

where the errors ϵ_i are uncorrelated with zero mean and constant variance, σ^2. Then the spline smoothing method uses the data to construct a curve f by minimizing the objective function

$$\begin{aligned} J(f; \chi, \lambda) &= \frac{1}{n}\sum_{i=1}^{n}\left(y_i - f(x_i)\right)^2 \\ &\quad + \lambda \int_{\infty}^{-\infty} \left(f^{(m)}(x)\right)^2 dx, \end{aligned} \tag{2}$$

where $f^{(m)}$ represents the mth derivative of f, with m being a positive integer, and λ is a smoothing parameter. For more details of smoothing splines see, for example, Eubank [19], Eubank [6], and Cantoni and Hastie [20]. An alternative definition of the level of smoothing is in terms of an *equivalent degrees of freedom*, `Df`, which describes the amount of information in the data needed to estimate the residuals. The function *smooth.spline* [21] allows λ or `Df` to be specified, but the degrees of freedom have been used in what follows as this gives a more intuitive interpretation.

The above objective function consists of two parts: the first measures the agreement of the function and the data and the second is a roughness penalty reflecting the total curvature—this can also be interpreted in a Bayesian setting as the likelihood and prior. Hence, for given `Df`, the estimate of f is given by

$$\widehat{f}(x, \mathtt{Df}) = \min_{f} J(f; \chi, \mathtt{Df}), \quad x \in \mathbf{R}. \tag{3}$$

If `Df` is large then the function is rough but closely fits the data, whereas when `Df` is small then the function is smooth but may not fit the data well. Here the choice of `Df` is made automatically using standard leave-one-out cross-validation [22]; that is,

$$\widehat{\mathtt{Df}} = \min_{\mathtt{Df}} \sum_{i=1}^{n} \left(y_i - \widehat{f}^{-i}(x_i, \mathtt{Df})\right)^2, \tag{4}$$

where $\widehat{f}^{-i}(\cdot, \mathtt{Df})$ is the fitted spline curve, for given parameter `Df`, and with the ith data point, (x_i, y_i), being removed. Then $\widehat{f}(\cdot, \widehat{\mathtt{Df}})$ is the fitted curve using the cross-validation estimate of the degrees of freedom.

Figure 1 shows fitted curves using splines with different degrees of freedom, `Df`. The true curve is a sine function with noise level $\sigma = 1/4$ which corresponds to a signal to noise ratio ($\mathtt{SNR} = \sigma_f/\sigma$) of about 2. In (a) `Df` is about half the value found using cross-validation which is used in (b), with (c) using double the cross-validation degrees of freedom. The small degrees-of-freedom value gives a smoother fitted curve that ignores many of the points in the data whereas a large value produces a rougher fit which more closely follows the data. The automatic choice was $\widehat{\mathtt{Df}} = 5.5$ which gives a very good fit to the data reproducing the sin curve well.

For this dataset, the periodic nature of the sin function has, so far, been ignored, and it is clear that the extreme left and right do not match exactly. For such datasets, made up of angles or directions, ignoring the periodic nature of the measurements when smoothing may produce unacceptable edge effects. A simple approach for dealing with this issue will now be considered.

Suppose that the dataset is made up of paired angles and distances which will be denoted as $\vartheta = \{(\theta_i, r_i) : i = 1, \dots, n\}$ for a sample of size n. A simple approach for periodic data measured in the interval $(0, 2\pi)$, say, is to repeat the data. That is, for each angle θ_i, the corresponding new angular values are $(\theta_i - p\pi, \dots, \theta_i - \pi,\ \psi,\ \theta_i + \pi, \dots, \theta_i + p\pi)$, where $p = 1, 2, \dots$, and similarly repeat the corresponding radial distances r_i to be $(r_i, \dots, r_i)$. This produces a dataset, $\vartheta^p = \{(\theta_i, r_i) : i = 1, \dots, n'\}$, with $n' = (2p + 1) \times n$ data values, and even for small p (e.g., $p = 1$ or 2) this gives a very good approximation to the full periodic spline. Cogburn and Davis [23] present the theory of periodic smoothing spline with application to the estimation of periodic functions and the R function `periodicSpline` from the package `splines` might provide an alternative computational approach.

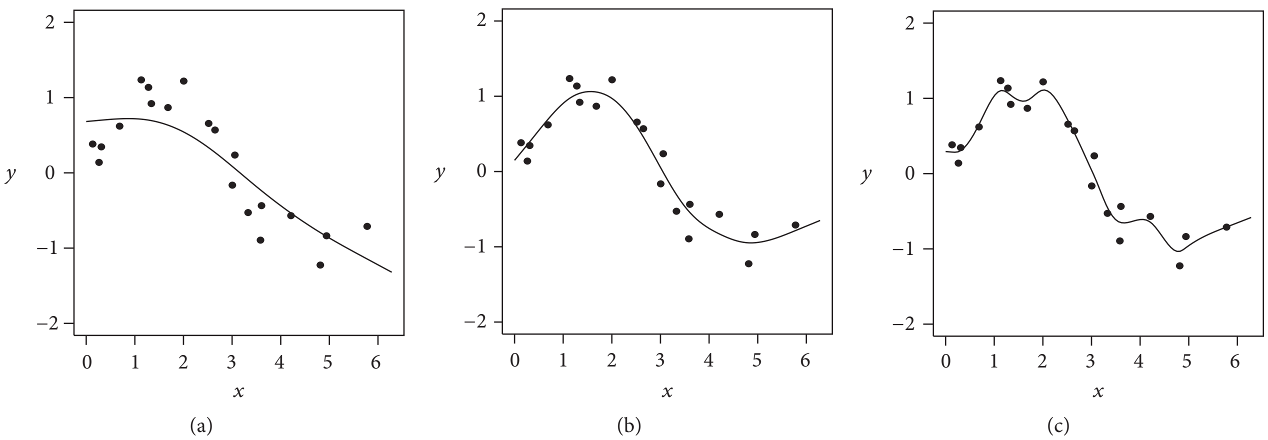


Figure 1: Smoothing spline fits to sin data: (a) $\mathtt{Df} = 3$; (b) $\widehat{\mathtt{Df}} = 5.5$; (c) $\mathtt{Df} = 11$.

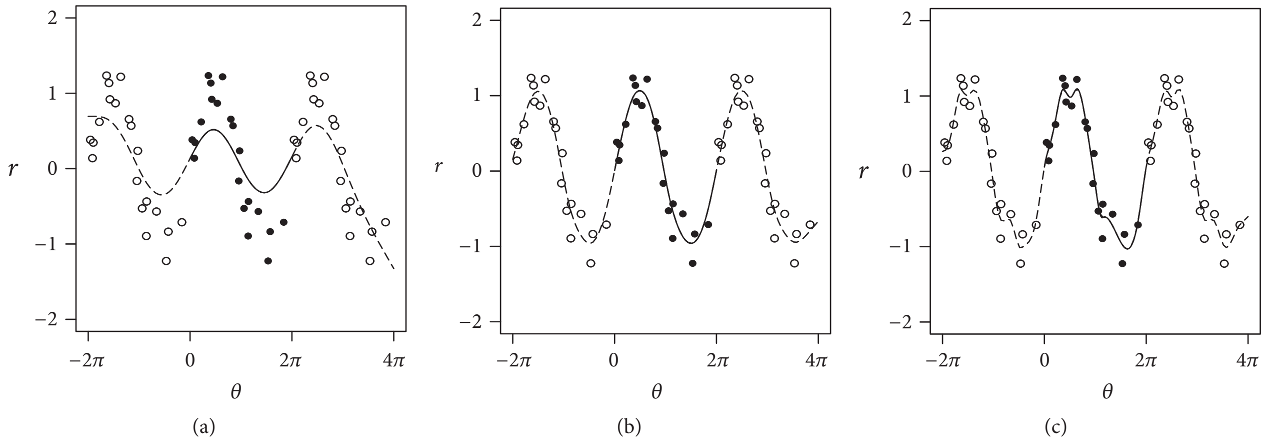


Figure 2: Periodic spline fits to sin data: (a) $\mathtt{Df} = 7$; (b) $\mathtt{Df}_{\mathrm{CV}} = 15$; (c) $\mathtt{Df} = 30$.

As illustration consider Figure 2 which shows fitted curves equivalent to those in Figure 1 but with $p = 1$. The solid circles are the original data with open circles representing the copied data points. Similarly, the solid line is the spline fitted curve over the original interval with the dashed line showing the fitted curve over the copied data points. In all cases the fit is better than in Figure 1, with the periodic nature, reproduced well, and as before the cross-validation choice of smoothing has produced an excellent reconstruction of the true sin curve.

Once fitted a residual sum of squares, RSS, calculated on the original data values, can be used as a measure of goodness-of-fit. Here this will be calculated using the radial distances with definition

$$\mathrm{RSS} = \sum_{i=1}^{n} \left(r_i - \hat{r}\left(\theta_i, \widehat{\mathtt{Df}}\right)\right)^2 \tag{5}$$

but other versions could be used, for example, the Euclidean distance between fitted and observed points.

Of course, the approach could lead to a poor fit if the data is not periodic, but to prevent this it is possible to allow for a discontinuity in the relationship. Here the approach of Gu [24], who considered discontinuities in cubic splines with a jump at a known location, will be extended to the periodic case and with an unknown discontinuity location.

Suppose that the points $\vartheta = \{(r_i, \theta_i) : i = 1, \ldots, n\}$ are partitioned into two groups with the first, ϑ_1, containing all the points with angles up to and including the change point and ϑ_2 those with angles above. Assuming that the points are ordered in increasing value of the angle, so that $\theta_1 \le \cdots \le \theta_n$, then let $\vartheta_1 = \{(\theta_i, r_i);\ i = 1, \ldots, k\}$ be the data before the change point and $\vartheta_2 = \{(\theta_i, r_i);\ i = k+1, \ldots, n\}$ the remaining data. For change point at θ_k two curves are fitted to the data such that

$$\hat{r}\left(\theta, \widehat{\mathtt{Df}}\right) = \begin{cases} \min\limits_{r} J\left(r; \vartheta_1, \widehat{\mathtt{Df}}_1\right), & \text{for } \theta \le \theta_k \\ \min\limits_{r} J\left(r; \vartheta_2, \widehat{\mathtt{Df}}_2\right), & \text{for } \theta > \theta_k, \end{cases} \tag{6}$$

where cross-validation is used separately on the two parts leading to two degrees of freedom, $\widehat{\mathtt{Df}} = (\widehat{\mathtt{Df}}_1, \widehat{\mathtt{Df}}_2)$. The significance of the change point could be assessed through a chi-squared test, but here a change point influence graph is considered based on the goodness-of-fit.

Consider the sin data shown in Figure 3(a) which has a change point of size about 1 introduced at $\theta = \theta_{[10]} \approx 2.3$.

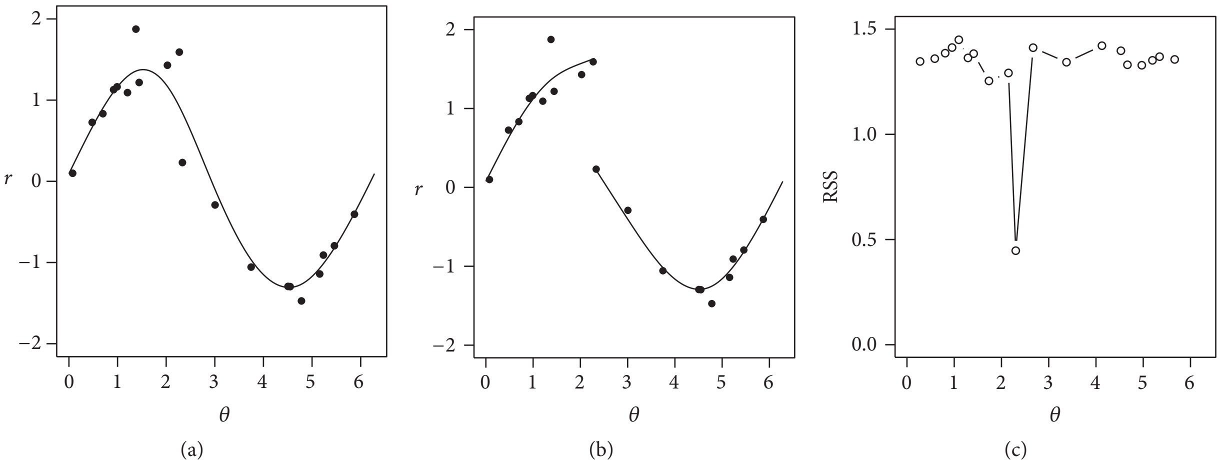


Figure 3: (a) Data from sin function with a discontinuity; (b) best two-part curve; (c) residual sum of square for two-part curves.

The curve in Figure 3(a) is fitted using a smoothing spline, but ignoring the change point, it is possible that the curves in Figure 3(b) are fitted using smoothing splines with change point at the estimated location. The automatically chosen value for the degrees of freedom, $\widehat{\mathtt{Df}}$, for the single curve in (a) is 14.5 whereas for the two-part curve the overall degrees of freedom are $\widehat{\mathtt{Df}}_1 + \widehat{\mathtt{Df}}_2 = 11$. Figure 3(c) shows the residual sum of squares, RSS, for each possible change point location with a very clear minimum. The RSS for the curve in Figure 3(a) is 1.3 while, in (b), it has reduced to 0.45, which is substantially smaller and provides a much better description of the data. Hence this approach provides an intuitive approach to finding change points in data automatically.

3. A Model for Multiple Overlapping Curves

3.1. Motivation. To motivate the modelling, consider an unobserved true scene containing a few objects of various shape and sizes, with possible overlap. However, instead of the scene being recorded faithfully, only partial information is taken and, in particular, only points along the edges of the objects are recorded. These points might be chosen to identify features with special significance or they might simply be at equal or random locations along the edge. Further, due to overlaps, points from the full edge may not be in the dataset. Once collected, there is no record of which points are from which object, and no record is kept of possible object shapes nor even the number of objects. Hence, let the dataset consists of a collection of n points, $\chi = \{(x_i, y_y) : i = 1, \dots, n\}$, recorded within some small region in 2D.

Figure 4 shows example datasets which will be analysed later. Panel (a) shows a human face profile with the forehead, eyes, nose, mouth, and chin clearly identifiable on the left—the points on the right locate the back of the neck and the hairline. Panel (b) shows points located along a handwritten number 3 at approximately equally spaced intervals.

3.2. Modelling a Single Curve with Occlusion. Before the periodic smoothing spline approach can be applied it is necessary for the data to be first transformed to polar coordinates. First define a *centre*, (ξ, ζ), which can be estimated using the data centroid $(\widehat{\xi}, \widehat{\zeta}) = (\overline{x}, \overline{y})$ and then use the one-to-one transformation

$$r_i = \left((x_i - \overline{x})^2 + (y_i - \overline{y})^2 \right)^{1/2},$$
$$\theta_i = \tan^{-1} \left(\frac{(y_i - \overline{y})}{(x_i - \overline{x})} \right), \qquad (7)$$
$$i = 1, \dots, n.$$

This gives rise to an alternative data representation via the centre $(\overline{x}, \overline{y})$ and polar coordinates $\varphi = \{(r_i, \theta_i) : i = 1, \dots, n\}$. Note that although this representation contains $n+2$ pieces of information, by construction, the polar coordinate variables are not independent. Of course, other estimates of centre could be considered, such as the point which minimizes the variance of the radii. In particular, this measure should be more robust to presence of occlusions.

To illustrate the transformation and the subsequent spline smoothing consider the simulated data in Figure 5. Panel (a) shows the given points along with the sample centre marked with a "+"; the points in (b) are the corresponding polar coordinates relative to this centre. Also shown in (b) are the nonperiodic smoothing spline (continuous black line) and the period smoothing spline (dashed red line). These are all closely aligned except at the extreme angles. Once transformed back into Cartesian coordinates, as shown in panel (c), the slight discrepancies between the fitted splines are more clearly visible. At the far right of the plot, the periodic spline curve is closed and more naturally represents a possible object, whereas the nonperiodic spline is not closed making it difficult to interpret if this were part of the edge of a real object.

Figure 6 shows a second elliptical dataset but where part of the ellipse is missing. The Cartesian data are shown in (a) and (c), with the polar transformed data in (b). Panel (b) shows the nonperiodic spline and the period spline with dramatic differences which are even more obvious when the

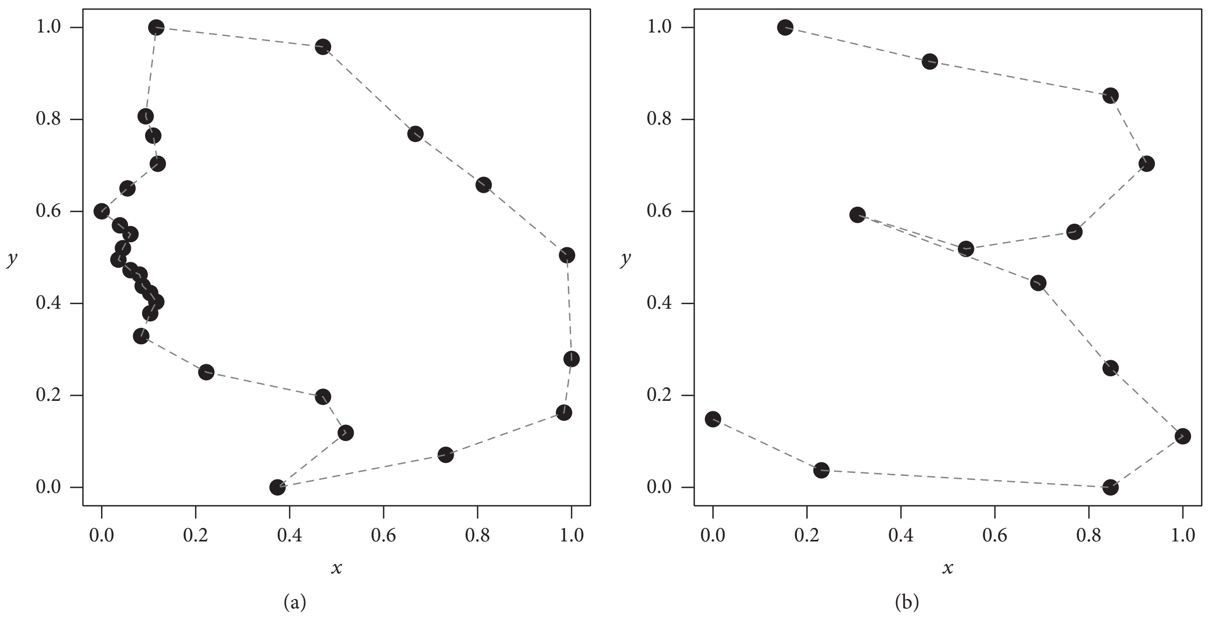


Figure 4: Real datasets: (a) human face and (b) handwritten number 3.

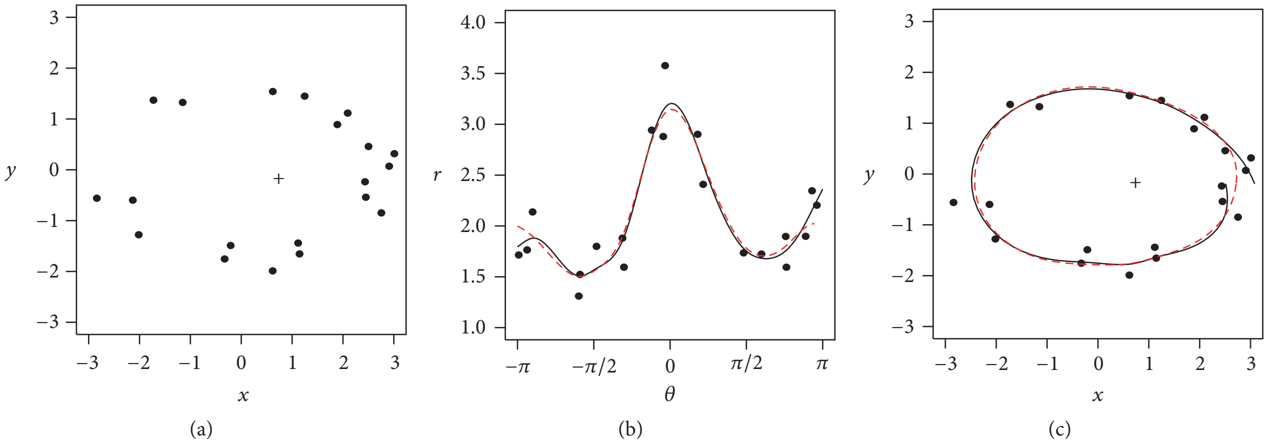


Figure 5: (a) Simulated data; (b) polar coordinate data with fitted spline curve; (c) data with back-transformed fitted curves. In (b) and (c) the solid curves use standard splines, whereas the dashed use the periodic spline.

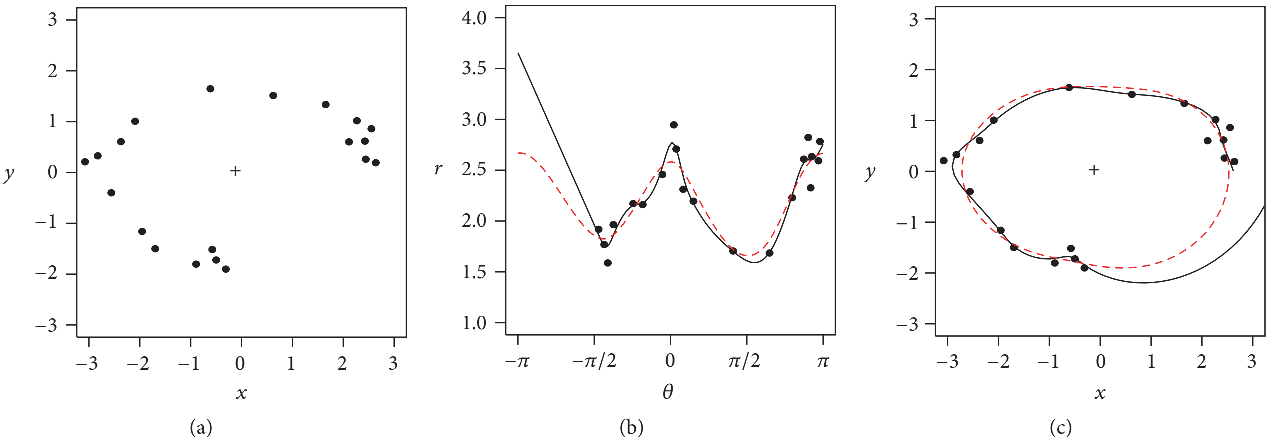


Figure 6: (a) Occluded data; (b) polar coordinate data with fitted spline curve; (c) data with back-transformed fitted curves. In (b) and (c) the solid curves use standard splines, whereas the dashed use the periodic spline.

fitted curves are transformed back into Cartesian coordinates, as shown in panel (c). The periodic smoothing spline has done a very good job of interpolating the missing part of the curve and the results can easily be relied upon in further analysis. In particular, slight changes in the position of a few critical point will lead to very different shapes for the nonperiodic spline.

To summarize, application of smoothing splines to periodic point data has proved very successful. The modification of the duplicated data is a simple, yet effective way to create closed curves and to interpolate where data are missing. The approach has provided a robust and informative reconstruction of the unknown curve from the data.

3.3. Modelling Multiple Intersecting Curves. To allow for intersecting and overlapping curves the points are partitioned into m groups, S_j, where $j = 1, 2, \ldots, m$. That is, $S_j \subseteq (1, \ldots, n)$ with $S_i \cap S_j = \emptyset$ when $i \neq j$ and $S_1 \cup \cdots \cup S_m = (1, \ldots, n)$. To record group membership a matrix $W_{n\times m} = (w_{ij})$ is defined, where $w_{ij} = 1$ if point i belongs to group j $(i \in S_j)$ and $w_{ij} = 0$ otherwise. Then, $\sum_j w_{ij} = 1$ and $\sum_i w_{ij} = n_j$, where n_j is the number of points in the jth group; that is, $n_j = |S_j|$. For each group, working in polar coordinates, there is a centre, (ξ_j, ζ_j), and coordinates relative to the centre, $\vartheta_j = \{(r_{ij}, \theta_{ij}) : i = 1, \ldots, n_j\}$, with the full set of parameters denoted as $\vartheta = \{\vartheta_j : j = 1, \ldots, m\}$. The corresponding Cartesian coordinates can be written as $\Gamma_j = \{(\mu_{ij}, \nu_{ij}) : i = 1, \ldots, n_j\}$, with

$$\begin{aligned} \mu_{ij} &= \xi_j + r_{ij}\cos\left(\theta_{ij}\right), \\ \nu_{ij} &= \zeta_j + r_{ij}\sin\left(\theta_{ij}\right), \\ &\qquad \text{for } i = 1, \ldots, n_j,\ \ j = 1, \ldots, m, \end{aligned} \tag{8}$$

and the full collection of data as $\Gamma = \{\Gamma_j : j = 1, \ldots, m\}$. Further, it is assumed that the point locations are recorded with error giving observed measurements

$$\begin{aligned} x_{ij} &= \mu_{ij} + \epsilon_{ij}, \\ y_{ij} &= \nu_{ij} + \varepsilon_{ij}, \\ &\qquad \text{for } i = 1, \ldots, n_j,\ \ j = 1, \ldots, m, \end{aligned} \tag{9}$$

where ϵ and ε are independent Gaussian random variables with zero mean and constant variance σ^2.

In what follows the full dataset will, without further explanation, be referred to using either $\chi = \{(x_{ij}, y_{ij}) : i = 1, \ldots, n_j,\ j = 1, \ldots, m\}$ and $\vartheta = \{(\theta_{ij}, r_{ij}) : i = 1, \ldots, n_j,\ j = 1, \ldots, m\}$ or equivalently, but without explicit reference to the group membership, $\chi = \{(x_i, y_i) : i = 1, \ldots, n\}$ and $\vartheta = \{(\theta_i, r_i) : i = 1, \ldots, n\}$ as is most convenient and intuitive.

3.4. Estimation with Multiple Intersecting Curves. Now consider estimation of the model unknowns from observed data. Start by supposing that a dataset is available but that the group membership information is intact; then the group centres could be estimated as

$$\begin{aligned} \hat{\xi}_j &= \overline{x}_j = \frac{\sum_{i=1}^{n} w_{ij} x_i}{\sum_i w_{ij}}, \\ \hat{\zeta}_j &= \overline{y}_j = \frac{\sum_{i=1}^{n} w_{ij} y_i}{\sum_i w_{ij}} \end{aligned} \tag{10}$$

and, although some of these are unimportant, corresponding polar coordinate representation of point i relative to group centre j is

$$\hat{r}_{ij} = \left(\left(x_i - \overline{x}_j\right)^2 + \left(y_i - \overline{y}_j\right)^2\right)^{1/2}, \tag{11}$$

$$\hat{\theta}_{ij} = \tan^{-1}\left(\frac{y_i - \overline{y}_j}{x_i - \overline{x}_j}\right), \tag{12}$$

where $i = 1, \ldots, n$ and $j = 1, \ldots, m$. The overall residual sum of squares is then the sum of the separate components

$$\text{RSS} = \sum_{j=1}^{m} \text{RSS}_j = \sum_{j=1}^{m}\sum_{i=1}^{n_j}\left(r_{ij} - \hat{r}\left(\theta_{ij}, \widehat{\texttt{Df}}_j\right)\right)^2. \tag{13}$$

Now consider the case when the group membership is unknown and must be inferred from the data. The aim is to find linked points by fitting curves. Some datasets have more than one curve and some have intersecting curves. Then classifying the points into groups may help to fit the correct curves that represent the data.

In general, this can be thought of as a change point problem, as already discussed, to address the lack of stationarity in the values. A change point occurs at some point in the data if all of the values up to and including it share a common curve while all those after the change point share another. This is exactly the same situation as the discussion in Section 2 and hence the same method of solution is applied.

4. Application to Real Data

4.1. General. The previous sections have illustrated the proposed exploratory data analysis tools on simulated example, whereas in this section the success of the approach is demonstrated on a varied range of real datasets. There is no wish to construct formal equations to define the shape but to stimulate further analyses.

4.2. Example 1: Face Data. The first experiment is conducted on data extracted from the human face [25] in a study looking at changes in shape due to growth in children. Figure 7(a) shows the data with points joining the points; then (b) shows the points transformed to polar coordinates along with fitted spline curves. Figure 7(c) shows the data set with back-transformation fitted values, and the solid curve shows those from the standard spline while the dotted curve shows those from the periodic spline. It is clear from the fitted curves that there is not much difference between the periodic and

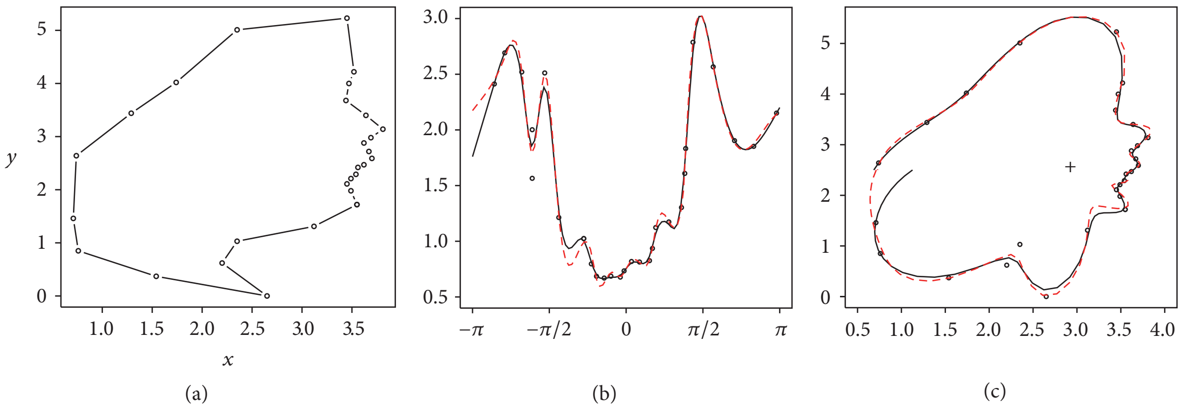


Figure 7: (a) Face data; (b) polar coordinate data with fitted spline curves; (c) back-transformed fitted curves. In (b) and (c) the solid curves use standard splines, whereas the dotted use periodic splines.

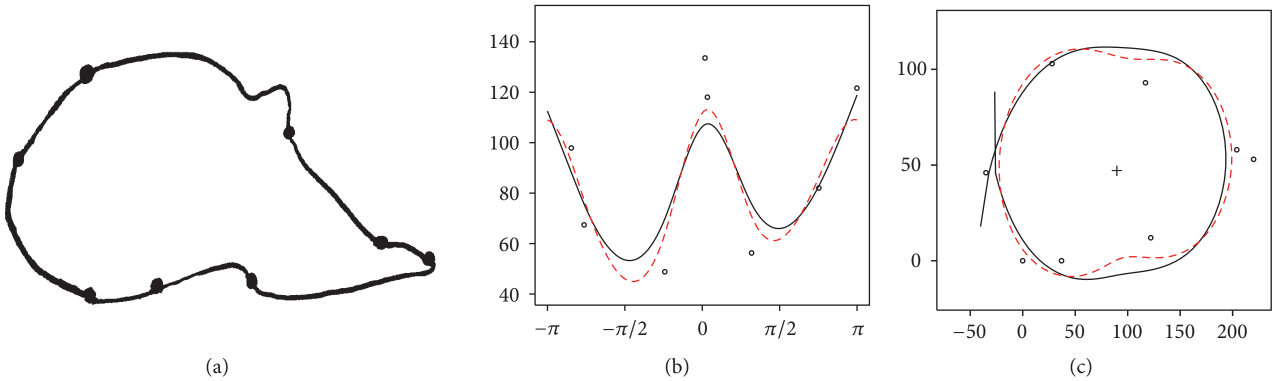


Figure 8: (a) Schematic diagram of a gorilla skull with anatomical landmarks for a male gorilla; (b) landmarks in polar coordinate and spline curves; (c) landmarks along with back-transformed fitted spline curves. In (b) and (c) the solid curves use standard splines, whereas the dotted use the periodic spline.

the standard smoothing splines. Both produce well fitted curves for the face. It is worth noting that the fitted curve can be evaluated at arbitrarily close locations, not only at the data points, and hence a smoothly interpolated curve can be drawn.

4.3. Example 2: Gorilla Skulls. This dataset, taken from Dryden and Mardia [3], is composed of 8 anatomical landmarks from the skulls of 29 male and 30 female gorillas. A landmark is defined as *a point of correspondence on each object that matches between and within populations* [3]. Figure 8(a) shows a schematic diagram of a typical skull with the landmarks indicated.

Figure 8(c) shows landmarks for one of the male gorillas and Figure 8(b) the corresponding points in polar coordinates along with spline fitting to the dataset and in Figure 8(c) after back-transforming. For both, the fits are good but at the expense of low smoothing in the spline. This fitting procedure was repeated for the other gorilla skulls and surprisingly the smoothed curves give good summaries allowing the skulls to be easily categorised into four main groups covering mainly male skulls which are rather elongated and two covering mainly female skulls which appear more rounded. The males lead to generally larger values of the degrees of freedom ($6 < \texttt{Df} < 8$) than the females ($\texttt{Df} \approx 2$). In fact, the automatic choice of the degrees-of-freedom parameter can be used as a simple discrimination variable giving only 8 out of 59 incorrectly classified skulls. It is important to note that this was not a preconceived discriminator but was identified by the exploratory analysis. This has highlighted the usefulness of simple and flexible tools as a preliminary step in a more wide-ranging investigation.

4.4. Example 3: The Number 3. Another dataset, again taken from Dryden and Mardia [3], is made up of 13 landmarks from 30 handwritten number 3's; see Figure 9(a). Suppose the data are divided into two subsets with n_1 and n_2 observations, respectively. The best partition is made according to the minimum value of the overall residual sum of squares, RSS, which is displayed in panel (c). Each subset is transformed to polar coordinates using the different centres marked "+" in panel (a). Each subset is indicated by different marks along with their fitted spline curves as plotted in panel (b) with the back-transformed fitted curves in panel (a). Clearly, this has described the two-part curves very well. Again, this demonstrates the flexibility of the procedure.

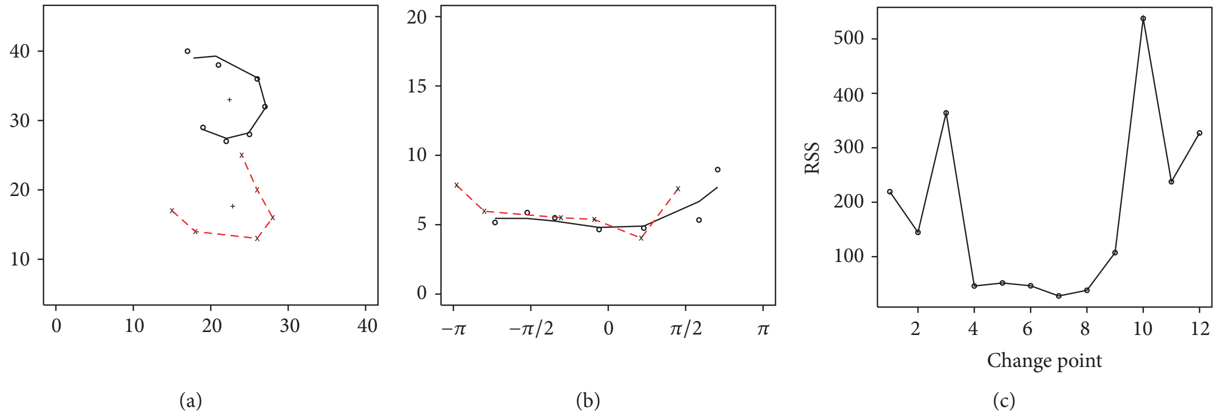


FIGURE 9: (a) A typical *number 3* dataset along with the back-transformed fitted two-part spline; (b) points in polar coordinates and two-part spline curve; (c) RSS plotted against change point position.

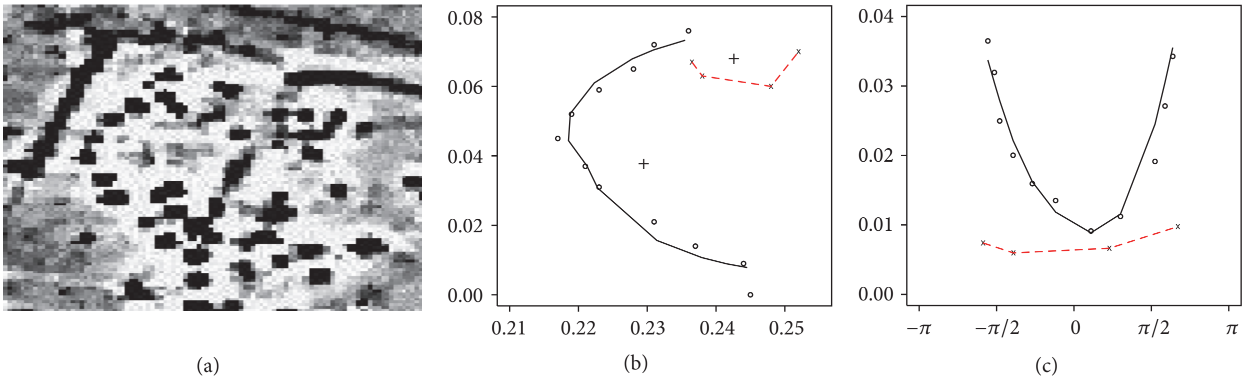


FIGURE 10: (a) Magnetic survey data for part of an iron-age archaeological site; (b) selected pits along with back-transformed fitted two-part spline curve; (c) polar coordinates and fitted two-part spline curve.

4.5. Example 4: Archaeological Site Data. The data in Figure 10(a) shows part of a typical image dataset (supplied by Alistair Marshall of the Guiting Power Amenity Trust; see Aykroyd et al. [26] for details) from a magnetic survey of an archaeological site. As well as linear features, which represent ditches, there are also several drifts of pits, but blurring and noise tend to camouflage the exact locations. Panel (b) shows the locations of some of these pits, appearing as small circles, and panel (c) shows the corresponding polar coordinates relative to the two data centres (marked "+" in (b)). According to the minimum value of the residual sum of squares, RSS, the observations can be classified automatically into two groups.

The data centres are calculated for each subset, the small circles are the data in the first subset, and "×" are the data in the second subset, with the fitted curves plotted in Figure 10(c). Then the fitted curves are back-transformed into Cartesian coordinate as shown in panel (b). The solid curve is for the first subset while the dotted curve is for the second subset. The aim of the analysis is to identify which points are associated with each other and to fit curves to the points, and this has been achieved well. The resulting linked points might then form part of further analysis or aid physical excavation.

5. Discussion

Making sense of clouds of points, apparently randomly placed across a 2D region, is a key task in many statistical investigations. When the points are recorded without additional information, the first task is to infer structure by linking points using a data-driven approach. This paper has proposed and investigated a simple, yet effective method based on change point identification and nonparametric spline smoothing. It provides an intuitive explanatory tool to identify patterns in the point locations. When it is assumed that the structures form lines and curves, the change points divide the data into subsets, with the splines providing a flexible method to infer the shape of the structures. The method has easily dealt with occlusions and intersections in scenes with multiple curves. Similar results might be achieved by applying more general modelling approaches, such as MARS, RARS, RCMARS; for details see, for example, [17], but we believe that a more straightforward and intuitive approach can have equal impact by bringing a range of easy-to-use tool to a wider audience. Further, for all users the methods considerations can be used to suggest further analyses based on more sophisticated approaches.

There is scope for extending the approach to include larger numbers of curves where it is not possible to divide the curves with a single change point. The nature of the problem is closely related to classification where the group membership is missing. This strongly suggests that a probabilistic approach might be considered based on statistical distribution models. This would then fit into the general framework where the EM algorithm has proven very useful. Also, there is a need to extend the approach to deal with unordered points and ones which are not star-shaped. These are areas of possible future work. Further, it is of interest to develop a similar procedure which would allow more formal modelling and model section, perhaps following the approach of general additive modelling [18].

The applications are various and varied with an illustrative example of the method when the data points are anatomical landmarks defined by geometrical features, equally spaced but blindly placed points along smooth curves and from extreme intensity points in grey-scale images. Further, the results of the analysis have provided new variables which could be the starting point for other analyses. Hence there is potential for this to be a valuable exploratory data analysis method in the tool-kit of applied statisticians and applied scientists.

Competing Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

References

[1] E. Batschelet, *Circular Statistics in Biology*, Academic Press, London, UK, 1981.

[2] F. L. Bookstein, *Morphometric Tools for Landmark Data: Geometry and Biology*, Cambridge Univesity Press, Cambridge, UK, 1991.

[3] I. L. Dryden and K. V. Mardia, *Statistical Shape Analysis*, Wiley Series in Probability and Statistics: Probability and Statistics, John Wiley & Sons, Chichester, UK, 1998.

[4] S. Lele and J. Richtsmeier, *An Invariant Approach to Statistical Analysis of Shapes*, Chapman & Hall/CRC, 2001.

[5] T. P. Ryan, *Modern Regression Methods*, Wiley Series in Probability and Statistics: Applied Probability and Statistics, John Wiley & Sons, New York, NY, USA, 1997.

[6] R. L. Eubank, *Nonparametric Regression and Spline Smoothing*, Marcel Dekker, New York, NY, USA, 1999.

[7] B. W. Silverman, "Some aspects of the spline smoothing approach to nonparametric regression curve fitting," *Journal of the Royal Statistical Society B*, vol. 47, no. 1, pp. 1–52, 1985.

[8] B. W. Silverman and J. T. Wood, "The nonparametric estimation of branching curves," *Journal of the American Statistical Association*, vol. 82, no. 398, pp. 551–558, 1987.

[9] P. J. Green and B. W. Silverman, *Non Parametric Regression and Generalized Linear Models: A Roughness Penalty Approach*, Chapman and Hall, 1994.

[10] E. J. Wegman and I. W. Wright, "Splines in statistics," *Journal of the American Statistical Association*, vol. 78, no. 382, pp. 351–365, 1983.

[11] B. W. Silverman, "A fast and efficient cross-validation method for smoothing parameter choice in spline regression," *Journal of the American Statistical Association*, vol. 79, no. 387, pp. 584–589, 1984.

[12] D. Nychka, "Splines as local smoothers," *The Annals of Statistics*, vol. 23, no. 4, pp. 1175–1197, 1995.

[13] G. Wahba, "Splines in nonparametric regression," in *Encyclopedia of Environmetrics*, John Wiley & Sons, New York, NY, USA, 2006.

[14] J. H. Friedman, "Multivariate adaptive regression splines," *The Annals of Statistics*, vol. 19, pp. 1–67, 1991.

[15] A. Özmen and G. W. Weber, "RMARS: robustification of multivariate adaptive regression spline under polyhedral uncertainty," *Journal of Computational and Applied Mathematics*, vol. 259, pp. 914–924, 2014.

[16] A. Özmen, G. W. Weber, I. Batmaz, and E. Kropat, "RCMARS: robustication of CMARS with different scenarios under polyhedral uncertainty set," *Communications in Nonlinear Science and Numerical Simulation*, vol. 16, pp. 1780–1787, 2011.

[17] A. Özmen, *Robust Optimization of Spline Models and Complex Regulatory Networks*, Contributions to Management Science, Springer, 2016.

[18] T. J. Hastie and R. J. Tibshirani, *Generalized Additive Models*, Chapman & Hall/CRC, 1990.

[19] R. L. Eubank, "Diagnostics for smoothing splines," *Journal of the Royal Statistical Society. Series B. Methodological*, vol. 47, no. 2, pp. 332–341, 1985.

[20] E. Cantoni and T. Hastie, "Degrees-of-freedom tests for smoothing splines," *Biometrika*, vol. 89, no. 2, pp. 251–263, 2002.

[21] R Core Team, *R: A Language and Environment for Statistical Computing,*, R Foundation for Statistical Computing, Vienna, Austria, 2016, http://www.R-project.org/.

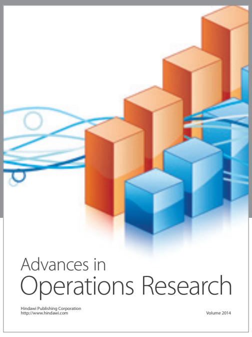
[22] P. Craven and G. Wahba, "Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation," *Numerical Mathematics*, vol. 31, pp. 377–403, 1979.

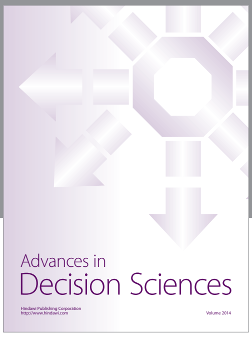
[23] R. Cogburn and H. T. Davis, "Periodic splines and spectral estimation," *The Annals of Statistics*, vol. 2, pp. 1108–1126, 1974.

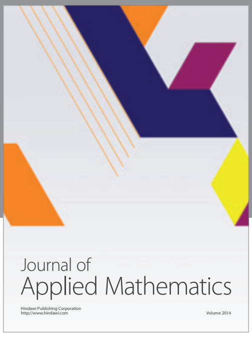
[24] C. Gu, "Multivariate spline regression," in *Smoothing and Regression: Approaches, Computation and Application*, M. G. Schimek, Ed., pp. 329–354, John Wiley & Sons, New York, NY, USA, 2000.

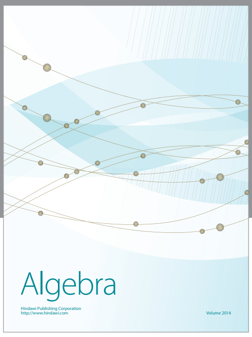
[25] R. J. Morris, J. T. Kent, K. V. Mardia, R. G. Aykroyd, M. Fidrich, and A. Linney, *Exploratory Analysis of Facial Growth*, Leeds University Press, 1999.

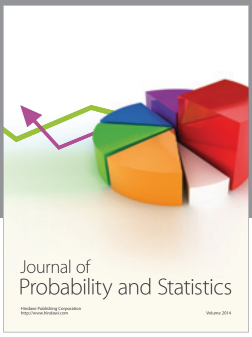
[26] R. G. Aykroyd, J. G. Haigh, and G. T. Allum, "Bayesian methods applied to survey data from archeological magnetometry," *Journal of the American Statistical Association*, vol. 96, no. 453, pp. 64–76, 2001.


Advances in
Operations Research

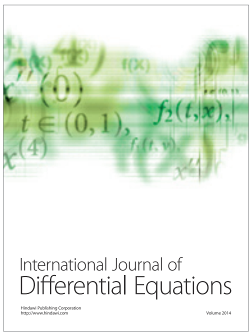

Advances in
Decision Sciences


Journal of
Applied Mathematics

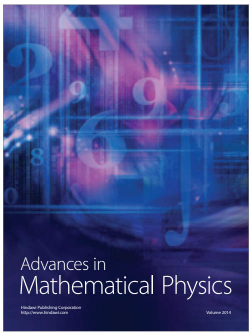

Algebra

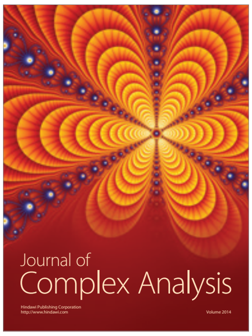

Journal of
Probability and Statistics

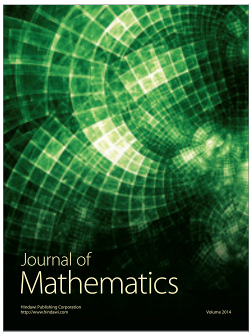
The Scientific
World Journal

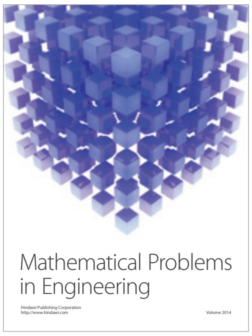

International Journal of
Differential Equations

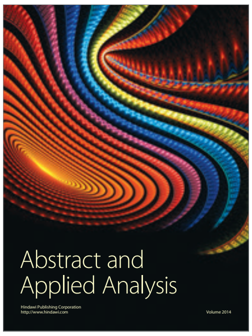
International Journal of
Combinatorics

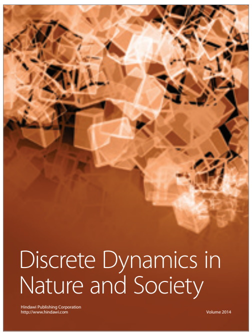

Advances in
Mathematical Physics

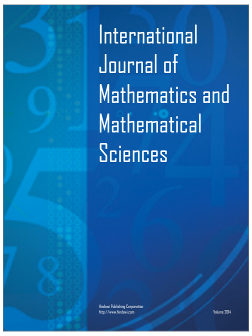

Journal of
Complex Analysis


Journal of
Mathematics


Mathematical Problems
in Engineering

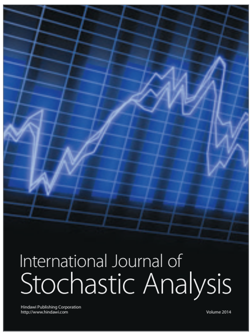

Abstract and
Applied Analysis


Discrete Dynamics in
Nature and Society


International
Journal of
Mathematics and
Mathematical
Sciences

Journal of
Discrete Mathematics

Journal of
Function Spaces


International Journal of
Stochastic Analysis

Journal of
Optimization