Cass Business School
City of London

Faculty of Actuarial Science and Insurance

Geometrically Designed, Variable Knot Regression Splines: Variation Diminishing Optimality of Knots.

Vladimir K. Kaishev, Dimitrina S. Dimitrova, Steven Haberman and Richard Verrall.

Statistical Research Paper No. 29

October 2006

ISBN 1-905752-03-2

Cass Business School
106 Bunhill Row
London EC1Y 8TZ
T +44 (0)20 7040 8470
www.cass.city.ac.uk

“Any opinions expressed in this paper are my/our own and not necessarily those of my/our employer or anyone else I/we have discussed them with. You must not copy this paper or quote it without my/our permission”.

Geometrically designed, variable knot regression splines: Variation diminishing optimality of knots

by

Vladimir K. Kaishev*, Dimitrina S. Dimitrova, Steven Haberman and Richard Verrall

Cass Business School, City University, London

Summary

A new method for Computer Aided Geometric Design of variable knot regression splines, named GeDS, has recently been introduced by Kaishev et al. (2006). The method utilizes the close geometric relationship between a spline regression function and its control polygon, with vertices whose y-coordinates are the regression coefficients and whose x-coordinates are certain averages of the knots, known as the Greville sites. The method involves two stages, A and B. In stage A, a linear LS spline fit to the data is constructed, and viewed as the initial position of the control polygon of a higher order ($n > 2$) smooth spline curve. In stage B, the optimal set of knots of this higher order spline curve is found, so that its control polygon is as close to the initial polygon of stage A as possible, and finally the LS estimates of the regression coefficients of this curve are found. In Kaishev et al. (2006) the implementation of stage A has been thoroughly addressed and the pointwise asymptotic properties of the GeD spline estimator have been explored and used to construct asymptotic confidence intervals.

In this paper, the focus of the attention is at giving further insight into the optimality properties of the knots of the higher order spline curve, obtained in stage B so that it is nearly a variation diminishing (shape preserving) spline approximation to the linear fit of stage A. Error bounds for this approximation are derived. Extensive numerical examples are provided, illustrating the performance of GeDS and the quality of the resulting LS spline fits. The GeDS estimator is compared with other existing variable knot spline methods and smoothing techniques and is shown to perform very well, producing nearly optimal spline regression models. It is fast and numerically efficient, since no deterministic or stochastic knot insertion/deletion and relocation search strategies are involved.

Keywords: spline regression, B-splines, Greville abscissae, variable knot splines, control polygon, asymptotic confidence interval

**Corresponding author's address: Faculty of Actuarial Science and Statistics, Cass Business School, City University, 106 Bunhill Row, London EC1Y 8TZ, UK. E-mail address: V.Kaishev@city.ac.uk, tel: +44(0)2070408453*

1. Introduction.

Consider the problem of nonparametric spline regression estimation in which, a response variable y is related to an independent variable $x \in [a, b]$, through the functional relationship

$$y = f(x) + \epsilon, \tag{1}$$

where ϵ is a random error variable with zero mean and $f(\cdot)$ is an unknown function, approximated with a n-th order (degree $n-1$) polynomial spline $f(\boldsymbol{t}_{k,n}; x)$. The latter is defined on the set of knots

$$\boldsymbol{t}_{k,n} = \{t_1 = \ ... = t_n = a < t_{n+1} < ... < t_{n+k} < t_{n+k+1} = b = \ ... = t_{2\,n+k}\} \tag{2}$$

as

$$f(\boldsymbol{t}_{k,n}; x) = \boldsymbol{\theta}' \boldsymbol{N}_n(x) = \textstyle\sum_{i=1}^{p} \theta_i N_{i,n}(x), \tag{3}$$

where $\boldsymbol{\theta} = (\theta_1, \ ..., \theta_p)'$ is the vector of regression coefficients and $\boldsymbol{N}_n(x) = (N_{1,n}(x), \ ..., N_{p,n}(x))'$, $p = n + k$, are the B-splines of order n. B-splines are defined on $\boldsymbol{t}_{k,n}$ through the Mansfield-De Boor-Cox recurrence relation

$$N_{i,1}(t) = \begin{cases} 1 & \text{if} \quad t_i \le t < t_{i+1} \\ 0 & \text{otherwise} \end{cases},$$

$$N_{i,n}(t) = \tfrac{t - t_i}{t_{i+n-1} - t_i} N_{i,n-1}(t) + \tfrac{t_{i+n} - t}{t_{i+n} - t_{i+1}} N_{i+1,n-1}(t). \tag{4}$$

from which it can be seen that $N_{i,n}(t) = 0$ for $t \notin [t_i, t_{i+n}]$. In the sequel, where necessary, we will emphasize the dependence of the spline regression $f(\boldsymbol{t}_{k,n}; x)$ on $\boldsymbol{\theta}$ by using the alternative notation $f(\boldsymbol{t}_{k,n}, \boldsymbol{\theta}; x)$. The nonparametric spline regression problem is then to estimate the degree of the spline, n, the number of the knots, k, their location and the regression coefficients $\boldsymbol{\theta}$, based on a sample of observations $\{y_i\}_{i=1}^N$ at some design points $\{x_i\}_{i=1}^N$.

Several different nonparametric spline approximation methods can be outlined. Under the direct approach, n and k are considered fixed (but unknown), and the knots $\boldsymbol{t}_{k,n}$ are assumed to be unknown parameters which have to be estimated by solving a non-linear least squares optimization problem (see DeBoor and Rice (1968), Jupp (1978), Hu (1993) and Lindstrom (1999)), based on the sample $\{y_i, x_i\}_{i=1}^N$. There are a number of difficulties related to this approach which have been pointed out by Jupp (1978) and Lindstrom (1999). All these difficulties have been shortly summarized by Carl DeBoor , who writes, "...it is essentially impossible to characterize a best approximation, that is to give a computationally useful criteria by which a best approximation can be recognized and distinguished from other approximations" (see DeBoor 2001, page 239).

As an alternative to the non-linear approach, adaptive knot selection procedures, such as step-wise knot inclusion/deletion strategies, have been developed by Smith (1982),

Friedman and Silverman (1989), Friedman (1991), Stone et al. (1997) and more recently by Zhou and Shen (2001), where some drawbacks of this approach have been pointed out.

Another group of works applies reversible jump Markov chain Monte Carlo (RJMCMC) based methods to develop Bayesian adaptive splines, such as those of Smith and Kohn (1996), Denison et al. (1998) and Biller (2000), in the context of generalized linear models. These procedures simulate tens of thousands of spline models, which are then averaged point-wise, to produce a resulting estimate of f. These methods are thus associated with a high computational cost and the inconvenience of having the resulting model in a non-explicit form. A stochastic optimization algorithm for free-knot splines, called adaptive genetic splines (AGS), was recently proposed by Pittman (2002) but the related computational cost is also a concern, as noted by the author.

Smoothing spline fitting methods, involving a smoothing penalty in the objective function have also been proposed in the statistical literature. We will mention here the hybrid adaptive splines (HAS) of Luo and Wahba (1997) and the penalized splines, considered by Eubank (1988), Wahba (1990), Marx and Eilers (1996), Rupert and Carroll (2000), Rupert (2002) and Wood (2003). Some asymptotic results, related to spline regression estimation are due to Agarwal and Studden (1980) and more recently to Huang (2003), where other references can be found.

Recently, a geometrically motivated method of variable knot spline regression estimation, which is new and very different from the existing methods, has been proposed by Kaishev et al. (2006). It is based on the so called Schoenberg's variation diminishing spline (VDS) approximation scheme, applied to the knot selection problem. The VDS approximation has some nice geometric properties such as *shape preservation*, which have made it fundamental in developing the Computer Aided Geometric Design (CAGD) methodology. These properties have been essential in developing the new variable-knot spline regression estimation method of Kaishev et al. (2006), called Geometrically Designed (GeD) spline estimation or simply GeDS. The latter produces a spline fit which is a least squares estimate with respect to its regression coefficients, but whose knots are placed in such a way that the fit has also the characteristics of a VDS approximation.

The purpose of this paper is to give some further insight into the optimality properties of the knot placement proposed by Kaishev et al. (2006), to explore further the pointwise asymptotic properties and related confidence intervals and the numerical performance of the proposed GeD spline estimator and compare it with other existing spline estimators.

The paper is organized as follows. In Sections 2 we recall some important geometric properties of the B-spline regression which have motivated the introduction of GeD spline estimation and are related to the Schoenberg's variation diminishing splines. Thus, Section 2.2 summarizes and extends the discussion presented in Kaishev et al. (2006), of

the fact that a spline regression function has a control polygon, and by manipulating the position of its vertexes it is possible to estimate the location of the knots and the regression coefficients. Section 3 gives a brief outline of the two stages A and B of the GeD spline regression estimation method and provides further comments on the solution of the constrained minimization problem of stage B. The optimality properties of the knots of the higher order spline regression model, obtained in stage B are discussed and explored in Section 4. These knots are such that their related higher order spline curve is nearly a variation diminishing approximation to the control polygon of stage A. Bounds for its deviation from the variation diminishing approximation are established by Theorem 1 and its Corollaries 1.1 and 1.2, in Section 4. In Section 4.1 the averaging knot location method, proposed in Kaishev et al. (2006), which gives good approximate values of the optimal knots of stage B, is revisited. It is shown that it leads to bounds, given by Theorem 2 and Corollaries 2.1 and 2.2, which are sharper than those established by Theorem 1 and its corollaries. Section 5 gives a summary of the pointwise asymptotic properties of GeDS, including the construction of asymptotic confidence intervals. In Section 6, six numerical examples are presented, on which the GeDS method is thoroughly tested and compared with other existing spline approximation methods. Proofs of the theorems and their corollaries are given in the Appendix.

2. Geometric interpretation of the spline regression estimation.

Since our main purpose in this paper is to explore the optimality properties of the knots, placed according to the GeD spline regression method of Kaishev et al. (2006), we will first review its basic characteristics and give a short description of it. The method is motivated by the observation that the spline regression $f(\boldsymbol{t}_{k,n}, \boldsymbol{\theta}; x)$ introduced in (3) as a function of an independent variable $x \in [a, b]$ can be viewed as a special case of a parametric spline curve. A parametric spline curve $\boldsymbol{Q}(t)$ is given coordinate-wise as

$$\boldsymbol{Q}(t) = \{x(t),\ y(t)\} = \{\textstyle\sum_{i=1}^{p} \xi_i\, N_{i,n}(t),\ \sum_{i=1}^{p} \theta_i\, N_{i,n}(t)\},$$

where t is a parameter, and $x(t)$ and $y(t)$ are spline functions, defined on one and the same set of knots $\boldsymbol{t}_{k,n}$. In view of the identity

$$x(t) = \textstyle\sum_{i=1}^{p} \xi_i^*\, N_{i,n}(t) = t, \tag{5}$$

known as linear precision property, with ξ_i^* defined as the averages

$$\xi_i^* = (t_{i+1} + \ldots + t_{i+n-1})/(n-1),\ i = 1,\ \ldots,\ p. \tag{6}$$

of the $n-1$ consecutive knots $t_{i+1},\ \ldots,\ t_{i+n-1}$, we can express a spline regression function $f(\boldsymbol{t}_{k,n}, \boldsymbol{\theta}; t)$, $t \in [a, b]$, as

$$\boldsymbol{Q}^*(t) = \{t,\ f(t_{k,n}, \boldsymbol{\theta}; t)\} = \{\textstyle\sum_{i=1}^{p} \xi_i^*\, N_{i,n}(t),\ \sum_{i=1}^{p} \theta_i\, N_{i,n}(t)\}, \tag{7}$$

i.e., $f(t_{k,n}, \boldsymbol{\theta}; x)$, $x \in [a, b]$ can be equivalently expressed in a parametric form as a spline regression curve $\boldsymbol{Q}^*(x)$.

The values ξ_i^* given by (6) are known as the Greville abscissae. We will alternatively use the notation $\boldsymbol{\xi}^*(\boldsymbol{t}_{k,n})$, to indicate the dependence of the set of Greville sites $\boldsymbol{\xi}^* = \{\xi_1^*, \ ..., \ \xi_p^*\} \equiv \boldsymbol{\xi}^*(\boldsymbol{t}_{k,n})$ on the knots $\boldsymbol{t}_{k,n}$.

Based on this parametric interpretation, it has been noted by Kaishev et al. (2006) that $\boldsymbol{Q}^*(t)$ can be characterized by a polygon $\boldsymbol{C}_{\boldsymbol{Q}^*}$, which is closely related to it and is called its *control polygon*. The vertices of the control polygon, called *control points,* are the points, $\boldsymbol{c}_i$, whose x- and y-coordinates are correspondingly the Greville sites ξ_i^* and the B-spline regression coefficients θ_i, i.e., $\boldsymbol{c}_i = (\xi_i^*, \theta_i)$, $i = 1, \ ..., \ p$. This close relationship between the spline regression curve and its control points is discussed and illustrated in Section 2.2. Due to the partition of unity property of B-splines,

$$\sum_{i=j-n+1}^{j} N_{i,n}(t) = 1, \ \text{for any } t \in [t_j, \ t_{j+1}), \ j = n, \ ..., \ n+k,$$

every point of the spline regression curve $\boldsymbol{Q}^*(t)$ of order n is a convex combination of n control points $\boldsymbol{c}_i$, i.e., $\boldsymbol{Q}^*(t) = \sum_{i=j}^{n+j-1} \boldsymbol{c}_i \, N_{i,n}(t)$ for $t \in [t_{n+j-1}, \ t_{n+j}]$, $j = 1, \ ..., \ k+1$. This means that each polynomial segment of $\boldsymbol{Q}^*(t)$ lies within the convex hull of the n control points, $\boldsymbol{c}_j, \ ..., \ \boldsymbol{c}_{j+n-1}$, $j = 1, \ ..., \ k+1$, defining it (see Section 2.2). The convex hull of $\boldsymbol{c}_j, \ ..., \ \boldsymbol{c}_{j+n-1}$ is the smallest convex polygon, enclosing these points.

In fact, the control polygon $\boldsymbol{C}_{\boldsymbol{Q}^*}$ with vertices $\boldsymbol{c}_i = (\xi_i^*, \theta_i)$ is itself a linear spline function, and hence can be expressed as

$$\boldsymbol{C}_{\boldsymbol{Q}^*}(t) = \{\sum_{i=1}^{p} \xi_i^* \, N_{i,2}(t), \ \sum_{i=1}^{p} \theta_i \, N_{i,2}(t)\} = \{t, \ \sum_{i=1}^{p} \theta_i \, N_{i,2}(t)\} \equiv \sum_{i=1}^{p} \theta_i \, N_{i,2}(t) \ . \quad (8)$$

In (8), $\sum_{i=1}^{p} \xi_i^* \, N_{i,2}(t) = t$ since $N_{i,2}(t)$ are defined over the knots $\boldsymbol{t}_{p-2,2}$, where $t_1 \equiv \xi_1^*$, $t_{p+2} \equiv \xi_p^*$ and $t_{i+1} \equiv \xi_i^*$, $i = 1, \ ..., \ p$ and the linear precision property (5) applies.

Since $\boldsymbol{Q}^*(t)$ is a convex combination of its control points, its graph lies within the convex hull of its control polygon $\boldsymbol{C}_{\boldsymbol{Q}^*}$. Moreover, as has been pointed out by Kaishev et al. (2006), the spline regression curve $\boldsymbol{Q}^*(t)$ lies close to its control polygon $\boldsymbol{C}_{\boldsymbol{Q}^*}$ also because $\boldsymbol{Q}^*(t)$ is the *shape preserving,* Schoenberg's VDS approximation of $\boldsymbol{C}_{\boldsymbol{Q}^*}$. Since the concept of VDS approximation to a function g, defined on $[a, b]$ is central in deriving the optimality properties of the GeDS knots, we will recall its definition and basic properties.

2.1. Schoenberg's variation diminishing spline approximation.

Given a set of knots $\boldsymbol{t}_{k,n}$, a function g, defined on $[a, b]$, can be approximated by the spline function

$$V[g](x) = \sum_{i=1}^{p} g(\xi_i^*) \, N_{i,n}(x), \quad (9)$$

where ξ_i^*, $i = 1, 2, \ ..., \ p$ are the Greville abscissae, obtained from $\boldsymbol{t}_{k,n}$, using (6).

The spline $V[g]$ is known as the Schoenberg's variation diminishing spline approximation of order n to g, on the set of knots $\boldsymbol{t}_{k,n}$. It is constructed by simply evaluating g at the Greville sites (6) and taking the values $g(\xi_i^*)$ as the B-spline coefficients. The *variation diminishing* character of (9) is due to the fact that $V[g]$ crosses any straight line at most as many times as does the function g itself. The latter suggests the following properties, which justify the importance of the VDS approximation in CAGD applications.

Property 1 (Shape preservation). The VDS approximation is *shape preserving* since it preserves the shape of the function g it approximates. More precisely, if g is positive, then $V[g]$ is also positive; if g is monotone, then $V[g]$ is also monotone; and if g is convex, $V[g]$ is also convex.

Property 2 (Reproduction of straight lines). The VDS approximation reproduces any straight line $l(t)$, $t \in [a, b]$. In particular, $V[t] = t$, which follows from the linear precision property (5).

We will see in Section 3 that the way knots are found in stage B allows the GeD spline approximation to incorporate the features of a VDS approximation. Properties 1 and 2 are also used in the next section to show the closeness of a spline regression curve to its control polygon, a fact essentially used to motivate the GeDS estimation method. Further details on geometric modelling with splines and related results are to be found in Farin (2002).

2.2. The spline regression curve and its control points.

Since the graph of $\boldsymbol{Q}^*(t)$ lies within the convex hull of its control polygon $\boldsymbol{C}_{\boldsymbol{Q}^*}$ and since $\boldsymbol{Q}^*(t)$ is the *shape preserving*, Schoenberg's VDS approximation of $\boldsymbol{C}_{\boldsymbol{Q}^*}$, (as follows from Property 1, Section 2.1, taking $g \equiv \boldsymbol{C}_{\boldsymbol{Q}^*}$), the spline regression curve $\boldsymbol{Q}^*(t)$ closely follows the shape of $\boldsymbol{C}_{\boldsymbol{Q}^*}$. We illustrate the shape preserving and convex hull properties in Fig. 1 where functional spline regression curves, $\boldsymbol{Q}^*(t)$, of order $n = 3$ and $n = 4$ and their control polygons, $\boldsymbol{C}_{\boldsymbol{Q}^*}$, are plotted. The grey areas in Fig. 1 are the two convex hulls, formed by $\boldsymbol{c}_4, \boldsymbol{c}_5, \boldsymbol{c}_6$ for the quadratic curve (left panel) and $\boldsymbol{c}_3, \boldsymbol{c}_4, \boldsymbol{c}_5, \boldsymbol{c}_6$ for the cubic curve (right panel) within which the corresponding segment of $\boldsymbol{Q}^*(t)$ for $t \in [t_6, t_7]$ lie.

Note that a linear spline curve $\boldsymbol{Q}(t)$ (order $n = 2$) coincides with its control polygon $\boldsymbol{C}_{\boldsymbol{Q}}$. In the quadratic case ($n = 3$), the spline curve $\boldsymbol{Q}(t)$, evaluated at the knots $t_3, t_4, \ldots, t_{k+4}$, interpolates $\boldsymbol{C}_{\boldsymbol{Q}}$ and is tangential to each of its segments, $\boldsymbol{c}_i, \boldsymbol{c}_{i+1}$, dividing it in a proportion $(t_{i+2} - t_{i+1}) : (t_{i+3} - t_{i+2})$, $i = 1, \ldots, k + 2$. This is illustrated in the left panel of Fig. 1, for the case of $k = 3$, where $\Delta_j = t_{j+1} - t_j$, $j = 3, \ldots k + 3$. In the cubic case ($n = 4$), the spline curve evaluated at a knot, $\boldsymbol{Q}(t_{i+3})$ is somewhere within the triangle of points $\boldsymbol{c}_i, \boldsymbol{c}_{i+1}\, \boldsymbol{c}_{i+2}$, $i = 1, 2, \ldots, p$, as can be seen from the right panel of Fig. 1. Hence, the higher is the degree, the stronger is the curve's deviation from its control polygon $\boldsymbol{C}_{\boldsymbol{Q}}$,

but it still remains within the convex hull of $\boldsymbol{C_Q}$. This suggests that a quadratic B-spline curve is very well suited as a compromise between smoothness and shape preservation.

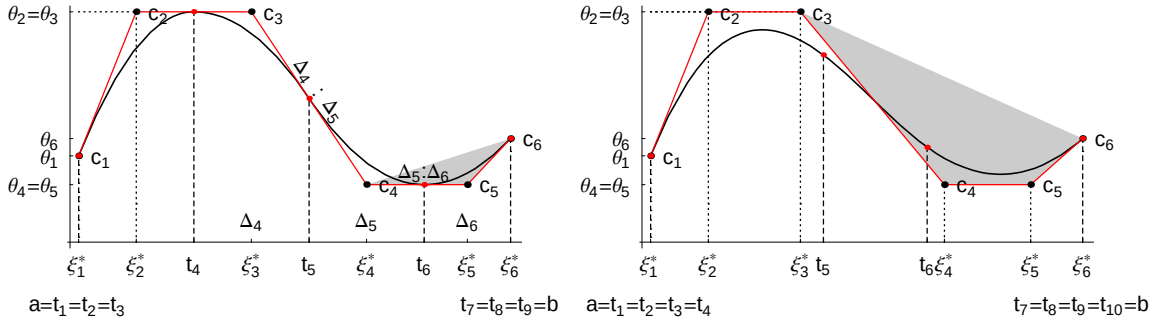

Fig. 1. Quadratic (left panel) and cubic (right panel) functional spline curves $\boldsymbol{Q}^*(t)$ and their control polygons $\boldsymbol{C_{Q^*}}$.

The close geometric relationship between the spline regression curve $\boldsymbol{Q}^*(x) = \{x,\ f(\boldsymbol{t}_{k,n}, \boldsymbol{\theta}; x)\}$, $x \in [a, b]$, and its control polygon $\boldsymbol{C}_{f(\boldsymbol{t}_{k,n},\boldsymbol{\theta};x)}$, is the foundation of the GeDS method, proposed in Kaishev et al. (2006). Here, we briefly summarize the logic behind this new geometrically motivated estimation approach. Since the x-coordinates of the vertices $\boldsymbol{c}_i = (\xi_i^*, \theta_i)$, $i = 1, \ldots p$, of $\boldsymbol{C}_{f(\boldsymbol{t}_{k,n},\boldsymbol{\theta};x)}$ are the Greville sites, ξ_i^*, obtained from $\boldsymbol{t}_{k,n}$, and the y-coordinates are the regression coefficients θ_i, estimation of $\boldsymbol{t}_{k,n}$ and $\boldsymbol{\theta}$, based on $\{y_i, x_i\}_{i=1}^N$, affects the geometric position of the control polygon $\boldsymbol{C}_{f(\boldsymbol{t}_{k,n},\boldsymbol{\theta};x)}$. On the other hand, due to the shape preserving and convex hull properties, $\boldsymbol{C}_{f(\boldsymbol{t}_{k,n},\boldsymbol{\theta};x)}$ defines the location and the shape of the spline curve $f(\boldsymbol{t}_{k,n}, \boldsymbol{\theta}; x)$. So, manipulating the vertices $\boldsymbol{c}_i$ of $\boldsymbol{C}_{f(\boldsymbol{t}_{k,n},\boldsymbol{\theta};x)}$, affects the knots $\boldsymbol{t}_{k,n}$, through (6), and the regression coefficients $\boldsymbol{\theta}$, which affects the position of the regression curve $f(\boldsymbol{t}_{k,n}, \boldsymbol{\theta}; x)$ itself. The latter conclusion has motivated the construction, in stage A of GeDS, of a control polygon as a linear least squares spline fit to the data, whose knots determine the knots $\boldsymbol{t}_{k,n}$, and whose B-spline coefficients, are viewed as initial estimate of $\boldsymbol{\theta}$, which is improved further in stage B (see Section 3). This is the basis of the approach which has been used by Kaishev et al. (2006) in constructing GeD variable knot spline approximation to the unknown function f in (1). The GeDS method is briefly described in the next Section 3.

3. The GeD spline regression estimation method.

In this section we will briefly outline the two stages of the GeD spline regression method, introduced in Kaishev et al. (2006), following the considerations of Sections 2. In stage A an appropriate control polygon in the form of a piece-wise linear LS fit which captures the shape of the data is constructed by starting with a straight line fit and adding knots where the current fit deviates most from the data. The rule for positioning the knots, the stopping rule for terminating this process and a complete description of the algorithm of stage A are given in Kaishev et al. (2006). The result of stage A is a piece-wise linear LS spline fit which is viewed as the initial position of the control polygon of

a smooth, higher order LS spline fit, obtained in stage B. In stage B a smooth LS spline fit to the data which closely follows the shape of the piece-wise linear fit from stage A is constructed. To achieve this, the knots of the latter linear fit are used to locate the knots of a functional spline curve, which is not an LS fit to the data, but which does follow the shape of the linear fit from stage A in the sense that it is nearly a VDS approximation to it. Then, its B-spline coefficients are adjusted in order to ensure that it is an LS fit. In Kaishev et al. (2006) stages A and B have been given the following more formal definition as certain optimization problems.

Stage A. Fix the order $n = 2$. Starting from a straight line fit and adding one knot at a time, find the least squares linear spline fit $\hat{f}(\boldsymbol{\delta}_{l,2}, \hat{\boldsymbol{\alpha}}; x) = \sum_{i=1}^{p} \hat{\alpha}_i N_{i,2}(x)$ with a number of internal knots l, number of B-splines $p = l + 2$ and with a set of knots $\boldsymbol{\delta}_{l,2} = \{\delta_1 = \delta_2 < \delta_3 < ... < \delta_{l+2} < \delta_{l+3} = \delta_{l+4}\}$, such that the ratio of the residual sums of squares

$$\mathrm{RSS}(l+q)/\mathrm{RSS}(l) = \sum_{j=1}^{N} (y_j - \hat{f}(\boldsymbol{\delta}_{l+q,2}; x_j))^2 \Big/ \sum_{j=1}^{N} (y_j - \hat{f}(\boldsymbol{\delta}_{l,2}; x_j))^2 \geq \alpha_{\mathrm{exit}}$$

where α_{exit} is a certain threshold level. This means that $\hat{f}(\boldsymbol{\delta}_{l,2}, \hat{\boldsymbol{\alpha}}; x)$ could not be significantly improved if q more knots are added, $q \geq 1$, and therefore $\hat{f}(\boldsymbol{\delta}_{l,2}, \hat{\boldsymbol{\alpha}}; x)$ adequately reproduces the "shape" of the unknown underlying function f. The resulting linear LS spline fit $\hat{f}(\boldsymbol{\delta}_{l,2}, \hat{\boldsymbol{\alpha}}; x)$ is viewed as a control polygon with vertices $(\xi_i, \hat{\alpha}_i)$, $i = 1, ..., p$, where $\xi_i \equiv \delta_{i+1}$, $i = 1, ..., p$. The fit $\hat{f}(\boldsymbol{\delta}_{l,2}, \hat{\boldsymbol{\alpha}}; x)$ is constructed following an algorithm described in Kaishev et al. (2006).

Stage B. For each of the values of $n = 3, ..., n_{\max}$, find the optimal position of the knots $\tilde{\boldsymbol{t}}_{l-(n-2),n}$, as a solution of the constrained minimization problem

$$\min_{\substack{\boldsymbol{t}_{l-(n-2),n},\\ \xi_{i+1} < t_{i+n} < \xi_{i+n-1},\\ i=1,...,k}} \left\| \hat{f}(\boldsymbol{\delta}_{l,2}, \hat{\boldsymbol{\alpha}}; x) - \boldsymbol{C}_{f(\boldsymbol{t}_{l-(n-2),n}, \hat{\boldsymbol{\alpha}}; x)} \right\|, \tag{10}$$

where $\|g\| := \max_{a \leq x \leq b} |g(x)|$ is the uniform (L_∞) norm of a function $g(x)$, and ξ_i, $i = 1, ..., p$ are the x-coordinates of the vertices of the control polygon $\hat{f}(\boldsymbol{\delta}_{l,2}, \hat{\boldsymbol{\alpha}}; x)$ obtained in stage A. In fact, minimization in (10) is over all polygons $\boldsymbol{C}_{f(\boldsymbol{t}_{l-(n-2),n}, \hat{\boldsymbol{\alpha}}; x)}$ which have vertices $(\xi_i^*, \hat{\alpha}_i)$, with x-coordinates which are the Greville sites $\boldsymbol{\xi}^*(\boldsymbol{t}_{l-(n-2),n})$, and y-coordinates, coincident with the y-coordinates $\hat{\alpha}_i$ of the vertices of the polygon $\hat{f}(\boldsymbol{\delta}_{l,2}, \hat{\boldsymbol{\alpha}}; x)$.

Our purpose here will be to comment on the possibility of solving problem (10) and to give some further insight into the optimality of the knots $\tilde{\boldsymbol{t}}_{l-(n-2),n}$ obtained as its solution. In order to do so, we first note that the two polygons $\hat{f}(\boldsymbol{\delta}_{l,2}, \hat{\boldsymbol{\alpha}}; x)$ and $\boldsymbol{C}_{f(\boldsymbol{t}_{l-(n-2),n}, \hat{\boldsymbol{\alpha}}; x)}$ have the same number of vertices $p = l + 2$, since the number of internal knots in $\boldsymbol{t}_{l-(n-2),n}$ is $l - (n - 2)$. Ideally, it will be desirable to find an optimal set of knots $\tilde{\boldsymbol{t}}_{l-(n-2),n}$ for which the minimum in (10) is zero, i.e., $\boldsymbol{C}_{f(\tilde{\boldsymbol{t}}_{l-(n-2),n}, \hat{\boldsymbol{\alpha}}; x)} \equiv \hat{f}(\boldsymbol{\delta}_{l,2}, \hat{\boldsymbol{\alpha}}; x)$. In other words,

one would require that $\tilde{\boldsymbol{t}}_{l-(n-2),n}$ be such that $f(\tilde{\boldsymbol{t}}_{l-(n-2),n}, \hat{\boldsymbol{\alpha}}; x)$ becomes the VDS approximation to $\hat{f}(\boldsymbol{\delta}_{l,2}, \hat{\boldsymbol{\alpha}}; x)$, or equivalently $\hat{f}(\boldsymbol{\delta}_{l,2}, \hat{\boldsymbol{\alpha}}; x)$ becomes the control polygon of the spline function $f(\tilde{\boldsymbol{t}}_{l-(n-2),n}, \hat{\boldsymbol{\alpha}}; x)$. In this way the knots $\tilde{\boldsymbol{t}}_{l-(n-2),n}$ match best the geometrical form of $\hat{f}(\boldsymbol{\delta}_{l,2}, \hat{\boldsymbol{\alpha}}; x)$ and as a consequence, the geometrical form of the data.

Since the two polygons in (10), $\boldsymbol{C}_{f(\boldsymbol{t}_{l-(n-2),n}, \hat{\boldsymbol{\alpha}}; x)}$ and $\hat{f}(\boldsymbol{\delta}_{l,2}, \hat{\boldsymbol{\alpha}}; x)$, have the same y-coordinates $\hat{\boldsymbol{\alpha}}$, they will coincide if their x-coordinates coincide, i.e., if $\xi_i^* \equiv \xi_i$, $i = 1, \ldots, p$. The latter would be fulfilled if, for given Greville sites $\xi_i^* = \xi_i$, it would be possible to solve the system (6) with respect to $\boldsymbol{t}_{l-(n-2),n}$.

However, to find $\tilde{\boldsymbol{t}}_{l-(n-2),n}$, so that equations (6) are fulfilled with respect to ξ_i, $i = 1, \ldots, p$ is, in general, impossible. This is easily seen from the fact that (6) represents an over-determined system of equations, with constraints on the knots, given by the definition (2) of $\boldsymbol{t}_{l-(n-2),n}$. Since $\xi_1 = a$ and $\xi_p = b$, the system (6) contains l equations and $l-(n-2)$ ordered, unknown knots, $(n > 2)$. Thus, it is in general impossible to place the knots $\tilde{\boldsymbol{t}}_{l-(n-2),n}$ in such a way that $\boldsymbol{C}_{f(\tilde{\boldsymbol{t}}_{l-(n-2),n}, \hat{\boldsymbol{\alpha}}; x)} \equiv \hat{f}(\boldsymbol{\delta}_{l,2}, \hat{\boldsymbol{\alpha}}; x)$, i.e., $\xi_i^* \equiv \xi_i$, for any fixed set $\{\xi_i\}$, $i = 1, \ldots, p$. Instead, what is achieved by solving (10) is that $\boldsymbol{C}_{f(\tilde{\boldsymbol{t}}_{l-(n-2),n}, \hat{\boldsymbol{\alpha}}; x)}$ gets as close to $\hat{f}(\boldsymbol{\delta}_{l,2}, \hat{\boldsymbol{\alpha}}; x)$ as possible, simultaneously with $\boldsymbol{\xi}^*$ getting as close to $\boldsymbol{\xi}$ as possible. Note that since we view the x-coordinates of the vertices of $\hat{f}(\boldsymbol{\delta}_{l,2}, \hat{\boldsymbol{\alpha}}; x)$, ξ_i, as Greville sites of a higher order spline curve $f(\boldsymbol{t}_{l-(n-2),n}, \hat{\boldsymbol{\alpha}}; x)$, the constraints $\xi_{i+1} < t_{i+n} < \xi_{i+n-1}$, $i = 1, \ldots, k$ in (10), follow from (6).

Since the resulting curve $f(\tilde{\boldsymbol{t}}_{l-(n-2),n}, \hat{\boldsymbol{\alpha}}; x)$ is the variation diminishing (i.e. shape preserving) spline approximation of its control polygon $\boldsymbol{C}_{f(\tilde{\boldsymbol{t}}_{l-(n-2),n}, \hat{\boldsymbol{\alpha}}; x)}$ (see Section 2), and since the latter is the best uniform (L_∞) approximation of $\hat{f}(\boldsymbol{\delta}_{l,2}, \hat{\boldsymbol{\alpha}}; x)$ in (10), $f(\tilde{\boldsymbol{t}}_{l-(n-2),n}, \hat{\boldsymbol{\alpha}}; x)$ will closely follow the shape of $\hat{f}(\boldsymbol{\delta}_{l,2}, \hat{\boldsymbol{\alpha}}; x)$. The fact that $f(\tilde{\boldsymbol{t}}_{l-(n-2),n}, \hat{\boldsymbol{\alpha}}; x)$ is nearly a VDS approximation to $\hat{f}(\boldsymbol{\delta}_{l,2}, \hat{\boldsymbol{\alpha}}; x)$ is proved in Section 4. However, as has been noted in Kaishev et al. (2006), $f(\tilde{\boldsymbol{t}}_{l-(n-2),n}, \hat{\boldsymbol{\alpha}}; x)$ is not a least squares approximation to the data set. In order to preserve the shape of $f(\tilde{\boldsymbol{t}}_{l-(n-2),n}, \hat{\boldsymbol{\alpha}}; x)$ and at the same time to make it an LS fit to the data, its optimal knots $\tilde{\boldsymbol{t}}_{l-(n-2),n}$ are preserved, whereas its B-spline coefficients $\hat{\alpha}_i$ are released, i.e., they are assumed to be unknown parameters, $\boldsymbol{\theta}$, which are estimated in the least squares sense, based on $\{y_i, x_i\}_{i=1}^N$. Thus, for a fixed $n = 3, \ldots, n_{\max}$, the least squares fit $\hat{f}\left(\tilde{\boldsymbol{t}}_{l-(n-2),n}, \hat{\boldsymbol{\theta}}; x\right)$ which solves

$$\min_{\boldsymbol{\theta}} \left[\sum_{j=1}^{N} \left(y_j - f(\tilde{\boldsymbol{t}}_{l-(n-2),n}, \boldsymbol{\theta}; x_j)\right)^2\right]$$

is found. Finally, the order $\tilde{n}$ whose fit $\hat{f}\left(\tilde{\boldsymbol{t}}_{l-(\tilde{n}-2),\tilde{n}}, \hat{\boldsymbol{\theta}}; x\right)$ has the minimum residual sum of squares is chosen.

In Section 4 we give results which shed some light on the optimality properties of the knots, chosen according to (10). Since (10) is a non-linear optimization problem, a method for its approximate solution, called the averaging knot location method has been

proposed in Kaishev et al. (2006). It comprises a very important part of GeDS and its properties are explored here in Section 4.1.

4. The optimal choice of the knots, $\tilde{\boldsymbol{t}}_{l-(n-2),n}$, in stage B of GeDS.

The optimal choice of the knots, $\tilde{\boldsymbol{t}}_{l-(n-2),n}$, in (10) can be given the following interpretation. Consider the n-th order parametric spline approximation $\boldsymbol{V}^a[\hat{f}]$ to the polygon $\hat{f}\,(\boldsymbol{\delta}_{l,2}, \hat{\boldsymbol{\alpha}}; x) = \sum_{i=1}^{p} \hat{\alpha}_i\, N_{i,2}(t)$ of stage A, given as

$$\boldsymbol{V}^a[\hat{f}]\,(t) = \{V_x^a[\hat{f}]\,(t),\, V_y^a[\hat{f}]\,(t)\} = \{\sum_{i=1}^{p} \xi_i\, N_{i,n}(t),\, \sum_{i=1}^{p}\ \hat{f}\,(\boldsymbol{\delta}_{l,2}, \hat{\boldsymbol{\alpha}}; \xi_i)\, N_{i,n}(t)\}$$

$$= \{\sum_{i=1}^{p} \xi_i\, N_{i,n}(t),\, \sum_{i=1}^{p} \hat{\alpha}_i\, N_{i,n}(t)\}\ , \qquad (11)$$

where the B-splines, $N_{i,n}(t)$, are defined on $\tilde{\boldsymbol{t}}_{l-(n-2),n}$. The approximation $\boldsymbol{V}^a[\hat{f}]$ is constructed coordinate-wise by defining the B-splines $N_{i,n}(t)$ on the set of knots $\tilde{\boldsymbol{t}}_{l-(n-2),n}$ and taking the x- and y-coordinates, $(\xi_i, \hat{\alpha}_i)$, of the vertices of $\hat{f}\,(\boldsymbol{\delta}_{l,2}, \hat{\boldsymbol{\alpha}}; x)$ as the B-spline coefficients of the splines $V_x^a[\hat{f}]\,(t)$ and $V_y^a[\hat{f}]\,(t)$. Hence, the control polygon, $\boldsymbol{C}_{\boldsymbol{V}^a[\hat{f}]}$, of the parametric spline approximation $\boldsymbol{V}^a[\hat{f}]\,(t)$, coincides with the control polygon, $\hat{f}\,(\boldsymbol{\delta}_{l,2}, \hat{\boldsymbol{\alpha}}; x)$, from stage A, i.e., following (8) we have

$$\boldsymbol{C}_{\boldsymbol{V}^a[\hat{f}]} = \{\sum_{i=1}^{p} \xi_i\, N_{i,2}(t),\, \sum_{i=1}^{p} \hat{\alpha}_i\, N_{i,2}(t)\} = \{t,\, \sum_{i=1}^{p} \hat{\alpha}_i\, N_{i,2}(t)\} \equiv \hat{f}\,(\boldsymbol{\delta}_{l,2}, \hat{\boldsymbol{\alpha}}; x),$$

where $\sum_{i=1}^{p} \xi_i\, N_{i,2}(t) = t$, since the B-splines $N_{i,2}(t)$ are defined on $\boldsymbol{\delta}_{l,2}$, where $\delta_1 \equiv \xi_1$, $\delta_{p+2} \equiv \xi_p$ and $\delta_{i+1} \equiv \xi_i$, $i = 1, \ldots, p$ and the linear precision property (5) applies. Note that $V_y^a[\hat{f}]\,(t) = \sum_{i=1}^{p} \hat{\alpha}_i\, N_{i,n}(t) \equiv f(\tilde{\boldsymbol{t}}_{l-(n-2),n}, \hat{\boldsymbol{\alpha}}; t)$ is the spline curve, whose control polygon $\boldsymbol{C}_{f\,(\tilde{\boldsymbol{t}}_{l-(n-2),n}, \hat{\boldsymbol{\alpha}}; x)}$ is the best uniform approximation to $\hat{f}\,(\boldsymbol{\delta}_{l,2}, \hat{\boldsymbol{\alpha}}; x)$ (see stage B, Section 3).

Following (9) and (7), the VDS approximation of $\hat{f}\,(\boldsymbol{\delta}_{l,2}, \hat{\boldsymbol{\alpha}}; x)$ on $\tilde{\boldsymbol{t}}_{l-(n-2),n}$ may be expressed in a parametric form as

$$\boldsymbol{V}[\hat{f}]\,(t) = \{V_x[\hat{f}]\,(t),\, V_y[\hat{f}]\,(t)\} = \{t,\, \sum_{i=1}^{p}\ \hat{f}\,(\boldsymbol{\delta}_{l,2}, \hat{\boldsymbol{\alpha}}; \xi_i^*)\, N_{i,n}(t)\}$$

$$= \{\sum_{i=1}^{p} \xi_i^*\, N_{i,n}(t),\, \sum_{i=1}^{p}\ \hat{f}\,(\boldsymbol{\delta}_{l,2}, \hat{\boldsymbol{\alpha}}; \xi_i^*)\, N_{i,n}(t)\}\ . \qquad (12)$$

As noted in stage B, Section 3, since the knots $\tilde{\boldsymbol{t}}_{l-(n-2),n}$ are the solution of the minimization problem (10), $\boldsymbol{\xi}^*(\tilde{\boldsymbol{t}}_{l-(n-2),n})$ are as close as possible to the x-coordinates, $\boldsymbol{\xi}$, of the vertices of $\hat{f}\,(\boldsymbol{\delta}_{l,2}, \hat{\boldsymbol{\alpha}}; x)$. Hence, $V_x^a[\hat{f}]\,(t) = \sum_{i=1}^{p} \xi_i\, N_{i,n}(t)$ in (11), is as close to the straight line $V_x[\hat{f}]\,(t) = t = \sum_{i=1}^{p} \xi_i^*\, N_{i,n}(t)$ in (12), as possible. In other words, $V_x^a[\hat{f}]\,(t) \approx t$ and one can conclude that $\boldsymbol{V}^a[\hat{f}]\,(t)$ is nearly a functional spline approximation to $\hat{f}\,(\boldsymbol{\delta}_{l,2}, \hat{\boldsymbol{\alpha}}; x)$, i.e., $V_y^a[\hat{f}]\,(t) = f(\tilde{\boldsymbol{t}}_{l-(n-2),n}, \hat{\boldsymbol{\alpha}}; t)$, is nearly a variation diminishing (shape preserving) spline approximation to $\hat{f}\,(\boldsymbol{\delta}_{l,2}, \hat{\boldsymbol{\alpha}}; x)$. This statement is made more precise by Corollary 1.1 of Theorem 1, which gives a bound for the error

$$\|V_x[\hat{f}]\,(t) - V_x^a[\hat{f}]\,(t)\| = \|t - \sum_{i=1}^{p} \xi_i\, N_{i,n}(x)\|\ ,$$

and by Corollary 1.2, which applied to $\hat{f}\,(\boldsymbol{\delta}_{l,2}, \hat{\boldsymbol{\alpha}}; x)$ gives a bound for the error

$$\|V_y[\hat{f}]\,(t) - V_y^a[\hat{f}]\,(t)\| = \|\textstyle\sum_{i=1}^{p}\ \hat{f}\,(\boldsymbol{\delta}_{l,2}, \hat{\boldsymbol{\alpha}}; \xi_i^*)\, N_{i,n}(t) - \sum_{i=1}^{p}\ \hat{f}\,(\boldsymbol{\delta}_{l,2}, \hat{\boldsymbol{\alpha}}; \xi_i)\, N_{i,n}(t)\|\ .$$

Theorem 1 establishes a bound for $\|V\,[g] - V^a[g]\|$ in the general case when g is any continuous function $g \in C[a, b]$, where $V\,[g]$ is the VDS approximation of g, defined in (9) and $V^a[g]$ is a non parametric (functional) version of (11), defined in (14).

Theorem 1. Let $\{\xi_i\}_{i=1}^{p}$ be an ordered set, $a = \xi_1 < \xi_2 < ... < \xi_p = b$, and let $\boldsymbol{t}_{k,n}$, $(p \geq n \geq 2, k = p - n)$, be a set of knots, defined as in (2), with

$$t_{i+n} = \xi_{i+1}, \qquad i = 1,\ ...,\ k, \quad \text{if } n = 2$$

$$\xi_{i+1} < t_{i+n} < \xi_{i+n-1}, \qquad i = 1,\ ...,\ k, \quad \text{if } n > 2\ . \tag{13}$$

Then, for the n-th order spline approximation $V^a[g]$, defined on $\boldsymbol{t}_{k,n}$, of a continuous function $g \in C[a, b]$, given by

$$V^a[g]\,(x) = \textstyle\sum_{i=1}^{p} g(\xi_i)\, N_{i,n}(x)\ , \tag{14}$$

we have

$$\|V[g] - V^a[g]\| \leq (n-2)\,\omega(g;\, \max_{j\in\{1,...,p-1\}}(\xi_{j+1} - \xi_j))\ , \tag{15}$$

where $V[g]$ is the Schoenberg's VDS approximation, defined on $\boldsymbol{t}_{k,n}$ following (9) and

$$\omega(g;\ h) := \max\{\,|\,g(x) - g(y)\,|\ :\ |\,x - y\,|\ \leq h,\ \ x, y \in [a, b]\}$$

is the modulus of continuity of the function g at h.

Corollary 1.1. Under the assumptions of Theorem 1 and if g is the straight line t, i.e., $g \equiv t$, we have

$$\|V[t] - V^a[t]\| = \|t - \textstyle\sum_{i=1}^{p} \xi_i\, N_{i,n}(t)\| \leq (n-2)\max_{j\in\{1,...,p-1\}}(\xi_{j+1} - \xi_j). \tag{16}$$

Corollary 1.2. Under the assumptions of Theorem 1 and assuming that g is a linear spline function $g(\boldsymbol{\delta}_{p-2,2}, \boldsymbol{\alpha}; t) = \sum_{i=1}^{p} \alpha_i\, N_{i,2}(t)$ with vertices (ξ_i, α_i), where $\alpha_i \in \mathbb{R}$ and $\boldsymbol{\delta}_{p-2,2}$ is such that $\delta_1 \equiv \xi_1$, $\delta_{p+2} \equiv \xi_p$, $\delta_{i+1} \equiv \xi_i$, $i = 1,\ ...,\ p$, we have

$$\|V[g] - V^a[g]\| = \|\textstyle\sum_{i=1}^{p}\ g(\boldsymbol{\delta}_{p-2,2}, \boldsymbol{\alpha}; \xi_i^*)\, N_{i,n}(x) - \sum_{i=1}^{p}\ g(\boldsymbol{\delta}_{p-2,2}, \boldsymbol{\alpha}; \xi_i)\, N_{i,n}(x)\|$$

$$\leq \max_{j\in\{1,...,p-(n-2)\}}(\max_{q\in\{j,...,j+(n-2)\}}\{\alpha_q\} - \min_{q\in\{j,...,j+(n-2)\}}\{\alpha_q\})\ . \tag{17}$$

Remark 1. Note that in the case when $V^a[g]$ is a quadratic spline approximation to $g(\boldsymbol{\delta}_{p-2,2}, \boldsymbol{\alpha}; t)$, i.e., when $n = 3$, the bound (17) simplifies to

$$\|V[g] - V^a[g]\| \leq \max_{j\in\{1,...,p-1\}} |\,\alpha_{j+1} - \alpha_j\,|\ . \tag{18}$$

Remark 2. It is worth mentioning that the spline approximation scheme $V^a[g]$, defined in (14), belongs to the class of the so called "quasi-interpolants" which have some nice approximation properties. For the latter, we refer to De Boor (2001).

4.1. The averaging knot location method.

The minimization problem (10), in stage B, is a constrained non-linear optimization problem with respect to the knots and although it is related to linear splines, it is still computationally involved. In addition, as with any other non-linear optimization problem, finding the globally optimal solution is not guaranteed. The knots $\tilde{\boldsymbol{t}}_{l-(n-2),n}$, which are the optimal solution, may also be (almost) coalescent and this may cause edges and corners in the final LS fit in stage B. In order to avoid these undesirable features, but to preserve the optimality properties of the knots, as described in stage B and Section 4, we propose to place the knots in stage B of GeDS according to (19), which we call the *averaging knot location method*.

Thus, the following method, giving an easy to evaluate, approximate solution to the minimization problem (10), is implemented in stage B, so that the final GeD spline fit is $\hat{f}\left(\bar{\boldsymbol{t}}_{l-(n-2),n}, \hat{\boldsymbol{\theta}}; x\right)$, where $\bar{\boldsymbol{t}}_{l-(n-2),n}$ is given by (19).

The averaging knot location method: Given the control polygon $\hat{f}\,(\boldsymbol{\delta}_{l,2}, \hat{\boldsymbol{\alpha}}; x)$ of stage A, for each of the values of $n = 3, \ldots, n_{\max}$, calculate the knot placement $\bar{\boldsymbol{t}}_{l-(n-2),n}$ with internal knots, defined as the averages of the x-coordinates of the vertices of $\hat{f}\,(\boldsymbol{\delta}_{l,2}, \hat{\boldsymbol{\alpha}}; x)$, i.e.,

$$\bar{t}_{i+n} = (\xi_{i+1} + \ldots + \xi_{i+n-1})/(n-1), \qquad i = 1, \ldots, k. \tag{19}$$

Note that $\xi_i = \delta_{i+1}$, $i = 1, \ldots, k+2$. The choice of the knots $\bar{\boldsymbol{t}}_{l-(n-2),n}$ according to (19) makes it possible to significantly improve the bounds, which hold for $\tilde{\boldsymbol{t}}_{l-(n-2),n}$ and are given by Corollaries 1.1 and 1.2. The improved bounds for the set of knots $\bar{\boldsymbol{t}}_{l-(n-2),n}$ are established by Corollaries 2.1 and 2.2 of Theorem 2 given next.

Theorem 2. Let $\{\xi_i\}_{i=1}^{p}$ be an ordered set, $a = \xi_1 < \xi_2 < \ldots < \xi_p = b$, and let $\boldsymbol{t}_{k,n}$, $(p \geq n \geq 2, k = p - n)$, be a set of knots, defined as in (2), with

$$t_{i+n} = (\xi_{i+1} + \ldots + \xi_{i+n-1})/(n-1), \qquad i = 1, \ldots, k$$

Then, for the n-th order spline approximation $V^a[g]$, defined on $\boldsymbol{t}_{k,n}$, of a continuous function $g \in C[a, b]$, given by

$$V^a[g]\,(x) = \textstyle\sum_{i=1}^{p} g(\xi_i)\, N_{i,n}(x)\ , \tag{20}$$

we have

$$\|V[g] - V^a[g]\| \leq \left\lceil \tfrac{(n-2)^2}{2\,(n-1)} \right\rceil \omega(g;\, \max_{j\in\{1,\ldots,p-1\}} (\xi_{j+1} - \xi_j))\ , \tag{21}$$

where $\lceil \nu \rceil := \min\{z \in \mathbb{Z} : \nu \leq z\}$, $V[g]$ is the Schoenberg's VDS approximation, defined on $\boldsymbol{t}_{k,n}$ and $\omega(g;\ h)$ is the modulus of continuity of the function g at h.

Corollary 2.1. Under the assumptions of Theorem 2 and if g coincides with the straight line t, i.e., $g \equiv t$, then

$$\|V[t] - V^a[t]\| = \|t - \sum_{i=1}^{p} \xi_i N_{i,n}(t)\| \le \frac{(n-2)^2}{2(n-1)} \max_{j\in\{1,\dots,p-1\}} (\xi_{j+1} - \xi_j). \quad (22)$$

Corollary 2.2. Under the assumptions of Theorem 2, with $n = 3$, and assuming that g is a linear spline function $g(\boldsymbol{\delta}_{p-2,2}, \boldsymbol{\alpha}; t) = \sum_{i=1}^{p} \alpha_i N_{i,2}(t)$ with vertices (ξ_i, α_i), where $\alpha_i \in \mathbb{R}$ and $\boldsymbol{\delta}_{p-2,2}$ is such that $\delta_1 \equiv \xi_1$, $\delta_{p+2} \equiv \xi_p$, $\delta_{i+1} \equiv \xi_i$, $i = 1, \dots, p$, we have

$$\|V[g] - V^a[g]\| = \|\sum_{i=1}^{p} g(\boldsymbol{\delta}_{p-2,2}, \boldsymbol{\alpha}; \xi_i^*) N_{i,3}(x) - \sum_{i=1}^{p} g(\boldsymbol{\delta}_{p-2,2}, \boldsymbol{\alpha}; \xi_i) N_{i,3}(x)\|$$

$$\le \frac{1}{4} \max_{j\in\{1,\dots,p-1\}} |\alpha_{j+1} - \alpha_j| . \quad (23)$$

In order to illustrate the bound (22) and how accurately the averaging knot location method (19) solves system (6) with respect to the knots for given Greville sites, we have randomly generated abscissa values ξ_j, $j = 1, \dots, p$ for three fixed numbers of vertices p, equal respectively to 6 ($k = 3$), 11 ($k = 8$) and 23 ($k = 20$). The number of simulations for each value of p is 1000. The corresponding thousand graphs of $\sum_{i=1}^{p} \xi_i N_{i,n}(t)$, $t \in [0, 1]$, in the quadratic case ($n = 3$), with knots defined by (19), are plotted in Fig. 2 (a), (b) and (c).

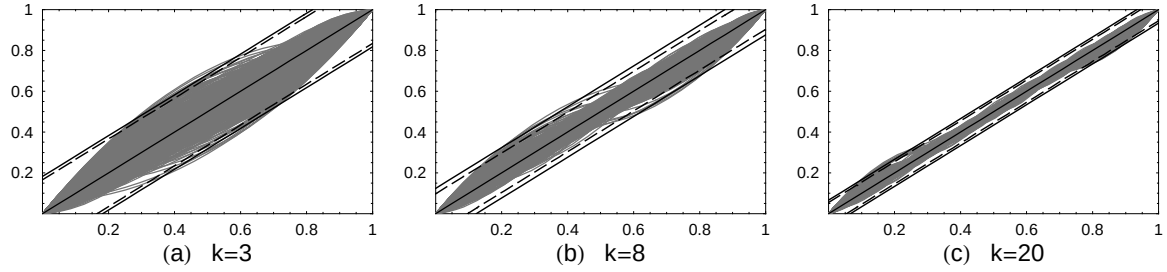

Fig. 2. Graphs of 1000 simulations of $\sum_{i=1}^{p} \xi_i N_{i,3}(t)$, with $\boldsymbol{t}_{k,3}$ according to (25) and estimates of $\hat{e}_{0.95}$ and $\hat{\varepsilon}_{0.95}$ for: (a) $p = 6$ $(k = 3)$, $\hat{e}_{0.95} = 0.17$, $\hat{\varepsilon}_{0.95} = 0.18$; (b) $p = 11$ $(k = 8)$, $\hat{e}_{0.95} = 0.10$, $\hat{\varepsilon}_{0.95} = 0.12$; (c) $p = 23$ $(k = 20)$, $\hat{e}_{0.95} = 0.05$, $\hat{\varepsilon}_{0.95} = 0.07$.

In Fig. 2, two corridors are also shown. The first, defined by the dashed lines, is based on the 95 sample percentile of $e = \|t - \sum_{i=1}^{p} \xi_i N_{i,3}(t)\|$, denoted by $\hat{e}_{0.95}$. The second corridor (the solid lines) is based on the 95 sample percentile $\hat{\varepsilon}_{0.95}$ of the bound in (22), denoted by ε. As can be seen from Fig. 2, the maximum deviation of $\sum_{i=1}^{p} \xi_i N_{i,3}(t)$ from the straight line t is reasonable, and rapidly decreases as the number of knots increases. Thus, the higher the number of knots, the more accurately the averaging knot location method (19) solves system (6). Similar conclusions are found to hold for the cubic case ($n = 4$), applying both $\hat{e}_{0.95}$ and $\hat{\varepsilon}_{0.95}$. As seen from Fig. 2, the solid line deviates insignificantly from the dashed line, so that the bound in (22) is nearly sharp for $n = 3$.

Remark 3. Note that, as seen from the bounds (16) and (22), the quality of the reconstruction of $\hat{f}(\boldsymbol{\delta}_{l,2}, \hat{\boldsymbol{\alpha}}; x)$, in stage B, using either $\boldsymbol{C}_{f(\tilde{\boldsymbol{t}}_{l-(n-2),n}, \hat{\boldsymbol{\alpha}}; x)}$ or $\boldsymbol{C}_{f(\bar{\boldsymbol{t}}_{l-(n-2),n}, \hat{\boldsymbol{\alpha}}; x)}$, depends on the maximal distance between the knots $\boldsymbol{\delta}_{l,2}$, obtained in stage A. By adding more knots at appropriate sites, the maximal distance may be decreased, which will make the bound (22) smaller. However, such an addition should be done in such a way that the geometry of $\hat{f}(\boldsymbol{\delta}_{l,2}, \hat{\boldsymbol{\alpha}}; x)$ is preserved. To achieve this, one may apply the Boehm's knot insertion

formula (see e.g., Farin 2002) and add a knot at the middle of the interval, where $\max_{j\in\{1,\ldots,p-1\}}(\xi_{j+1}-\xi_j)$ is attained. It is worth pointing out that, based on our experience with GeDS, the reconstruction in stage B is quite satisfactory and such knot insertion has not been implemented.

Remark 4. The choice of the knots $\bar{\boldsymbol{t}}_{l-(n-2),n}$ in (19) can also be given an interpretation, related to the problem of optimal recovery of a function g, by interpolating it at some fixed points, with an n-th order spline on a set of knots $\boldsymbol{t}_{k,n}$. The problem is to find the optimal set of knots, $\boldsymbol{t}_{k,n}^{\text{opt}}$ for which the bound on the interpolation error is minimized over all possible choices of $\boldsymbol{t}_{k,n}$. Such optimal interpolation has been considered by Michelli, Rivlin and Winograd (1976). An approximate solution to this optimal recovery problem has been proposed by De Boor (2001). In our case, if we apply this scheme to the polygon $\hat{f}\,(\boldsymbol{\delta}_{l,2}, \hat{\boldsymbol{\alpha}}; x)$ and view its vertices $(\xi_i, \hat{\alpha}_i)$ as given data points, then the approximate solution of this optimal interpolation problem, as proposed by De Boor (2001), is the set of knots $\bar{\boldsymbol{t}}_{l-(n-2),n}$ in (19).

5. Asymptotic properties of GeDS and related inference.

Pointwise asymptotic properties of the proposed GeD spline estimation method have been explored in Kaishev et al. (2006) where related large sample statistical inference has also been provided. To investigate the pointwise asymptotic behaviour of the GeDS estimation error $\hat{f}\big(\bar{\boldsymbol{t}}_{l-(n-2),n}, \hat{\boldsymbol{\theta}}; x\big) - f(x)$ its decomposition

$$\begin{aligned}&\hat{f}\big(\bar{\boldsymbol{t}}_{l-(n-2),n}, \hat{\boldsymbol{\theta}}; x\big) - f\,(x)\\ &= \big[\hat{f}\big(\bar{\boldsymbol{t}}_{l-(n-2),n}, \hat{\boldsymbol{\theta}}; x\big) - E\,\hat{f}\big(\bar{\boldsymbol{t}}_{l-(n-2),n}, \hat{\boldsymbol{\theta}}; x\big)\big] + \big[E\,\hat{f}\big(\bar{\boldsymbol{t}}_{l-(n-2),n}, \hat{\boldsymbol{\theta}}; x\big) - f(x)\big]\end{aligned}$$

has been considered, where the first and the second terms on the right-hand side are correspondingly referred to as the variance and the bias terms. In the asymptotic analysis, carried out in Kaishev et al. (2006), as the sample size, N_i, grows to infinity with $i = 1, 2, \ldots$, under some mild assumptions with respect to the sequences of design points $\{x_j\}_{j=1}^{N_i}$, it has been shown that the knots $\bar{\boldsymbol{t}}_{l_i-(n-2),n}$, $n \geq 2$, obtained by the GeDS estimation method, have global mesh ratios

$$M_{\bar{\boldsymbol{t}}_i}^{(r)} = \frac{\max_{n\leq j\leq l_i+1+n-r}(\bar{t}_{i,j+r}-\bar{t}_{i,j})}{\min_{n\leq j\leq l_i+1+n-r}(\bar{t}_{i,j+r}-\bar{t}_{i,j})}, \quad r \geq n$$

which form a sequence, bounded in probability by a constant $\gamma > 0$, i.e., $M_{\bar{\boldsymbol{t}}_i}^{(r)} \leq \gamma$, except on an event whose probability tends to zero as $N_i \to \infty$ (see Lemmas 2 and 3 of Kaishev et al. 2006).

Based on these results, and on a theorem from approximation theory establishing the stability of the L_∞ norm of the L_2 projections onto the linear space of splines $S_{\boldsymbol{t}_{k,n}}$, two asymptotic properties of the GeDS estimator have been established. Thus, Theorems 1 and 2 from Kaishev et al. (2006) give a bound for the bias term and a sufficient condi-

tion for it to be of negligible magnitude compared to the variance term. After its appropriate standardization, $\hat{f}\left(\bar{\boldsymbol{t}}_{l-(n-2),n}, \hat{\boldsymbol{\theta}}; x\right)$ has been shown (see Theorem 3 of Kaishev et al. 2006) to converge to a standard normal distribution, given that a suitable value of α_{exit} in the stopping rule of Stage A has been chosen. This characteristic of GeDS allows for the construction of $100\,(1-\alpha)\,\%$ asymptotic confidence intervals

$$\hat{f}\left(\bar{\boldsymbol{t}}_{l-(n-2),n}, \hat{\boldsymbol{\theta}}; x\right) \pm z_{1-\alpha/2} \sqrt{\text{Var}\left(\hat{f}\left(\bar{\boldsymbol{t}}_{l-(n-2),n}, \hat{\boldsymbol{\theta}}; x\right) \,\middle|\, \vec{\boldsymbol{x}}\right)}, \tag{24}$$

where $z_{1-\alpha/2} = \Phi^{-1}(1-\alpha/2)$, $n \geq 2$, $\vec{\boldsymbol{x}} = (x_1, \ ..., \ x_N)$,

$$\text{Var}\left(\hat{f}\left(\bar{\boldsymbol{t}}_{l-(n-2),n}, \hat{\boldsymbol{\theta}}; x\right) \,\middle|\, \vec{\boldsymbol{x}}\right) = \sigma^2 \, \boldsymbol{N}_n'(x) \left\{\langle \boldsymbol{F}'(\vec{\boldsymbol{x}}), \boldsymbol{F}(\vec{\boldsymbol{x}})\rangle\right\}^{-1} \boldsymbol{N}_n(x)\,(1 + o_P(1)),$$

and the matrix $\boldsymbol{F}(\vec{\boldsymbol{x}}) = (\boldsymbol{N}_n(x_1), \ ..., \ \boldsymbol{N}_n(x_N))$. In the next section, numerical tests of the proposed GeD spline estimator are performed and confidence intervals around the final fits are constructed, using the above results.

6. GeDS in action.

The proposed GeDS method has been implemented using *Mathematica* 5.0 and a standard PC (Pentium IV, 1.4 Ghz, 512 RAM) has been used for all test examples.

In order to obtain a GeDS estimate, most often it is necessary to input only the set of data $\{x_i, \, y_i\}_{i=1}^{N}$. The two parameters, $\alpha_{\text{exit}} \in (0, \, 1)$ and $\beta \in [0, \, 1]$, defined in steps 10 and 5 of stage A of GeDS (see Appendix A of Kaishev et al. 2006), by means of which the exit from GeDS can be controlled, have default preassigned values, which in general need not be re-set. The parameter α_{exit} is related to the stopping rule, which determines when to exit from stage A, i.e., it determines the number and location of the knots, $\boldsymbol{\delta}_{l,2}$, of $\hat{f}\,(\boldsymbol{\delta}_{l,2}, \, \hat{\boldsymbol{\alpha}}; \, x)$ and hence the number and location of the knots of the final higher order LS spline fit $\hat{f}\left(\bar{\boldsymbol{t}}_{l-(n-2),n}, \hat{\boldsymbol{\theta}}; x\right)$. The parameter β is related to the cluster weights of the clusters of residuals of same signs, as defined in step 5 of stage A of GeDS (see Appendix A of Kaishev et al. 2006). Its choice depends on the wiggliness of the recovered function f and the level of the noise ϵ. In the Normal case, $\epsilon \sim \mathcal{N}(0, \, \sigma_\epsilon^2)$, the noise level is defined by the variance σ_ϵ^2. As will be illustrated, for most of the examples GeDS gives very good results with the default values $\alpha_{\text{exit}} = 0.9$, $\beta = 0.5$. Our experience shows that choices of $\alpha_{\text{exit}} \in (0, \, 0.7)$ may cause exit after the first few steps which, for most functions, does not lead to an adequate resulting fit.

The choice of β depends on the level of the signal-to-noise ratio (SNR), $\text{SNR} = (\text{var}(f))^{0.5} / \sigma_\epsilon$ and on the degree of smoothness of f. As will be seen, in most of the numerical examples, the appropriate value of β was 0.5, which means that the within-cluster mean residual value and the cluster range can be considered equally important components of the weights w_j, $j = 1, \ ..., \ l$, (see Appendix A of Kaishev et al. 2006). However, based on our experience, when the SNR is high and f is smooth, recommended values are $\beta \in [0.5, \, 0.6]$, $\alpha_{\text{exit}} = 0.9$. If the SNR is high and f is a wiggly

function then the recommended choice is $\beta \in [0.5, 0.6]$, $\alpha_{\text{exit}} \in [0.99, 0.999]$, since otherwise underfitting may result. In the case when SNR is low and f is smooth, one may use $\beta \in [0.4, 0.5]$, $\alpha_{\text{exit}} \in [0.9, 0.99]$. It is known that, when the SNR is low and the underlying function is very unsmooth, recovering f is very difficult and different choices of β and α_{exit} may need to be attempted.

In order to facilitate comparison of GeDS with existing smoothing methods, we have simulated data using the functions given in Table 1, which have been widely used in testing other existing smoothing procedures.

Table 1. Summary of test functions.

Function	Specification
1	$f_1(x) = (4x-2) + 2e^{-16(4x-2)^2}$
2	$f_2(x) = \sin(8x-4) + 2e^{-16(4x-2)^2}$
HeaviSine	$f_3(x) = 4\sin(4\pi x) - \text{sgn}(x-0.3) - \text{sgn}(0.72-x)$
Doppler	$f_4(x) = \sqrt{x(1-x)}\,\sin\left(\frac{2\pi(1+\epsilon)}{(x+\epsilon)}\right),\ \epsilon = 0.05$
Bumps	$f_5(x) = \sum_j h_j\left(1+\left\lvert\frac{x-s_j}{w_j}\right\rvert\right)^{-4}$, $\{h_j\} = \{4, 5, 3, 4, 5, 4.2, 2.1, 4.3, 3.1, 5.1, 4.2\}$ $\{s_j\} = \{0.1, 0.13, 0.15, 0.23, 0.25, 0.40, 0.44, 0.65, 0.76, 0.78, 0.81\}$ $\{w_j\} = \{0.005, 0.005, 0.006, 0.01, 0.01, 0.03, 0.01, 0.01, 0.005, 0.008, 0.005\}$
Blocks	$f_6(x) = \sum_j h_j \frac{1+\text{sgn}(x-s_j)}{2}$, $\{h_j\} = \{4, -5, 3, -4, 5, -4.2, 2.1, 4.3, -3.1, 2.1, -4.2\}$ $\{s_j\} = \{0.1, 0.13, 0.15, 0.23, 0.25, 0.40, 0.44, 0.65, 0.76, 0.78, 0.81\}$

The data sets, used to test GeDS were simulated by adding noise, $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2)$, to each of the six functions, as given in Table 2.

Table 2. Summary of examples used to test GeDS.

Example No	***Function (data)***	***Interval***	***Sample size, N***	***Data x_i, $i = 1, \ldots, N$***	***Noice level, σ_ϵ***	***SNR***
1	$f_1(x)$	[0, 1]	256 150	$U(0, 1)$	0.6, 0.4, 0.25 0.25	2, 3, 5 5
2	$f_2(x)$	[0, 1]	256	$U(0, 1)$	0.3	3
3	HeaviSine	[0, 1]	2048	$x_i = (i-1)/2047$	1	7
4	Doppler	[0, 1]	2048	$x_i = (i-1)/2047$	1	7
5	Bumps	[0, 1]	2048	$x_i = (i-1)/2047$	1	7
6	Blocks	[0, 1]	2048	$x_i = (i-1)/2047$	1	7

As can be seen, we have included examples testing GeDS for different values of SNR, and for various characteristics of the data set: small and large sample sizes, x-values in a grid or uniformly generated within different intervals, $x \in [a, b]$. Note also that the test functions possess different smoothness properties: some of them are relatively smooth, while others are very wiggly.

In order to compare the quality of the fits produced by GeDS to those given by other authors, we use the mean square error (MSE), defined with respect to the true function f, rather than to the data, i.e.,

$$\text{MSE} = \left\{ \sum_{i=1}^{N} \left(f(x_i) - \hat{f}\left(\bar{\boldsymbol{t}}_{l-(n-2),n}, \hat{\boldsymbol{\theta}}; x_i\right)\right)^2 \right\} \Big/ N .$$

Note that, in practice, the underlying function is unknown and a set of observations is fitted. For this reason, we give also the L_2-error of approximation, defined as $\sqrt{\text{RSS}}$. However, for a fair comparison between the smoothing methods, one would need all model parameter values, such as, the number of knots (regression functions) and degree of the spline fits etc., which often are not reported in full. In order to compare the speed of computation on equal grounds, one would need to implement all of the available methods using the same hardware and software, and test them on entirely identical simulated data sets. Such a comparison is outside the scope of this paper.

Stage A of the GeD spline estimator has been thoroughly illustrated in Kaishev et al. (2006). Here we concentrate on the final GeD spline fit resulting from stage B.

We have run GeDS with 400 simulated data sets for Examples 1 and 2, and 31 data sets for Examples 3-6 as has been done by other authors in testing their methods (see, for example, Luo and Wahba, 1997). This allows us to compute the median of the MSE, obtained using GeDS, and compare it with the MSE medians given by other authors. However, in order to illustrate how GeDS performs, in each example we have used a single data set randomly chosen among the simulated data sets.

We compare most of our results with those of Luo and Wahba (1997) since, along with the median MSE values for their fits, they give also the order and the number of the basis functions. The Bumps and Blocks have been excluded from the comparison, since Luo and Wahba (1997) use versions of these functions which differ from ours, i.e., from those proposed by Donoho and Johnstone (1994). The GeD fits in Examples 1 and 2 are compared with the optimal spline fits, produced following the standard LS non-linear optimization approach and its penalized version, developed by Lindstrom (1999). The latter has been implemented, using the transformation of the knots, proposed by Jupp (1974) and the *Mathematica* function `NMinimize`, which attempts to find the global minimum. Due to the drawbacks of the non-linear optimization approach, it has not been feasible to produce optimal spline fits for the spatially inhomogeneous functions, recovered in Examples 3-6 from large data sets, using *Mathematica*, and a standard PC.

Example 1. This smooth function first appears as a test example in Fan and Gijbels (1995). It has been used later by Luo and Wahba (1997), Denison et al. (1998) and Zhou and Shen (2001) to test their fitting procedures. With this example, we illustrate that GeDS works well for data sets with different sample sizes and various noise levels, assuming ϵ is normally distributed. It takes between 0.89 sec and 1.66 sec to compute the GeDS fits, given in Table 3.

Table 3. (Example 1) Summary of fits produced by GeDS.

Fit No	***Graph***	N	σ_ϵ	***n***	***k***	***Internal knots***	α_{exit}, β	L_2 ***– error, MSE***
1	Fig. 3, (a)	150	0.25	3	4	{0.37, 0.46, 0.54, 0.62}	0.9, 0.5	2.87, 0.001282
2	Fig. 3, (b)	256	0.25	3	4	{0.38, 0.46, 0.54, 0.63}	0.9, 0.5	4.01, 0.001359
3	Fig. 3, (c)	256	0.4	3	4	{0.38, 0.46, 0.54, 0.60}	0.95, 0.5	6.17, 0.006573
4	Fig. 3, (d)	256	0.6	3	5	{0.26, 0.39, 0.51, 0.55, 0.62}	0.95, 0.5	9.03, 0.021918

The L_2-errors of all the fits are within the noise level and their visual quality is very good, as can be seen from Fig. 3. The 95% confidence intervals given in Fig. 3 have been calculated using (24) with the corresponding known ('oracle') σ_ϵ.

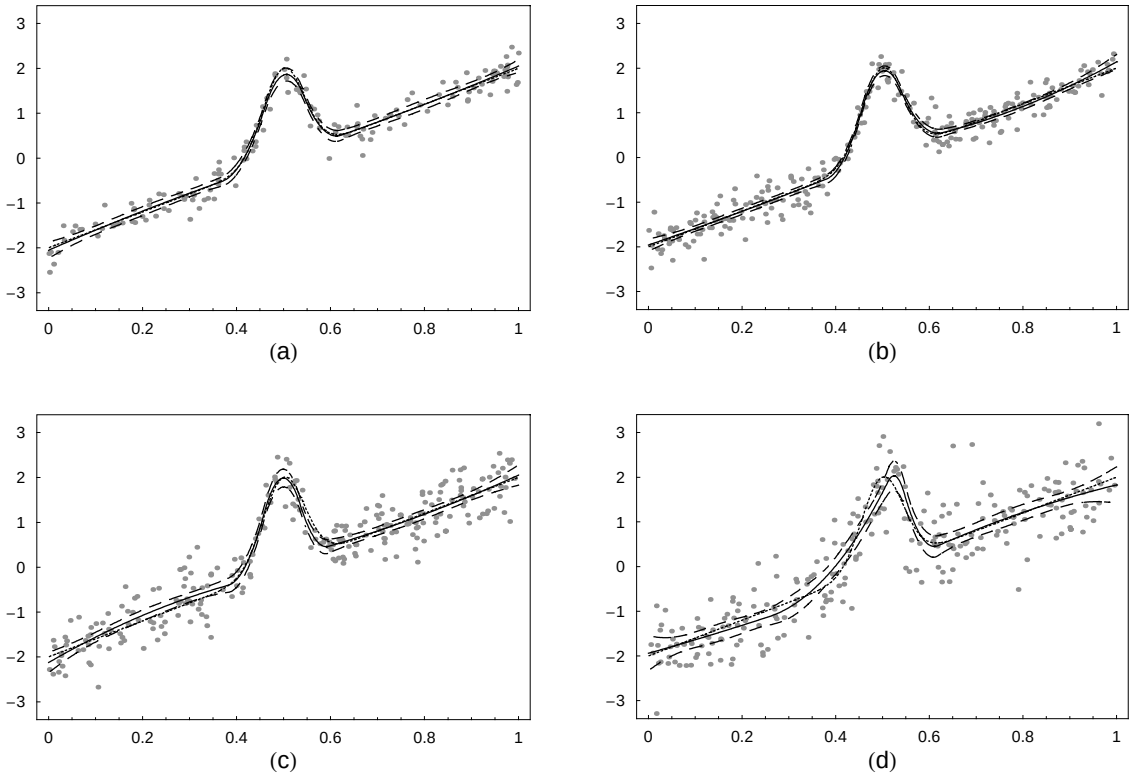

Fig. 3. (Example 1) Graphs of the final quadratic B-spline fits and confidence intervals, produced by GeDS: (a) $N = 150$, $\sigma = 0.25$; (b) $N = 256$, $\sigma = 0.25$; (c) $N = 256$, $\sigma = 0.4$; (d) $N = 256$, $\sigma = 0.6$; The dotted function is the true function.

Note that the first two fits in Table 3 are obtained with $\alpha_{\text{exit}} = 0.9$ and $\beta = 0.5$. Since the noise levels for fits No 3 and 4 are higher than for fits No 1 and 2, α_{exit} has been increased to 0.95, because, in the case of a smooth function and a high noise level, the relative improvements in RSS from one step to another would be smaller and more steps would be needed to recover the function.

In the case $\sigma_\epsilon = 0.4$, we have compared the quadratic GeD spline fit (No 3, Table 3) with the optimal quadratic spline fits obtained applying the LS non-linear optimization

method (NOM) and its penalized version (PNOM), due to Lindstrom (1999). The results are summarized in Table 4. As can be seen, the three fits are very close, comparing the L_2-errors and the location of the knots. However, the GeD fit recovers the original function significantly better than the fits NOM and PNOM, as indicated by the corresponding MSE values. The NOM optimal fit produces an edge at 0.425 and visually deviates stronger from the shape of the underlying function, which is one of the drawbacks noted by Lindstrom (1999). The computation time needed for GeDS is less then a second, and for PNOM and NOM it is respectively 11 and 20 minutes, using the *Mathematica* function `NMinimize`.

Table 4. (Example 1) The fits produced by GeDS, PNOM and NOM.

Fit No	***Method***	***n***	***k***	***Internal knots***	***L_2 – error, MSE***
1	GeDS	3	4	{0.38, 0.46, 0.53, 0.60}	6.17, 0.006573
2	PNOM	3	4	{0.40, 0.44, 0.52, 0.62}	6.16, 0.007364
3	NOM	3	4	{0.42, 0.43, 0.53, 0.60}	6.14, 0.010285

A frequency plot of the number of internal knots and box plots for the three linear GeD spline fits for data sets with $N = 150$, $\sigma_\epsilon = 0.25$, $N = 256$, $\sigma_\epsilon = 0.25$ and $N = 256$, $\sigma_\epsilon = 0.4$, over the 400 GeDS runs are presented in Fig. 4 (a) and (b).

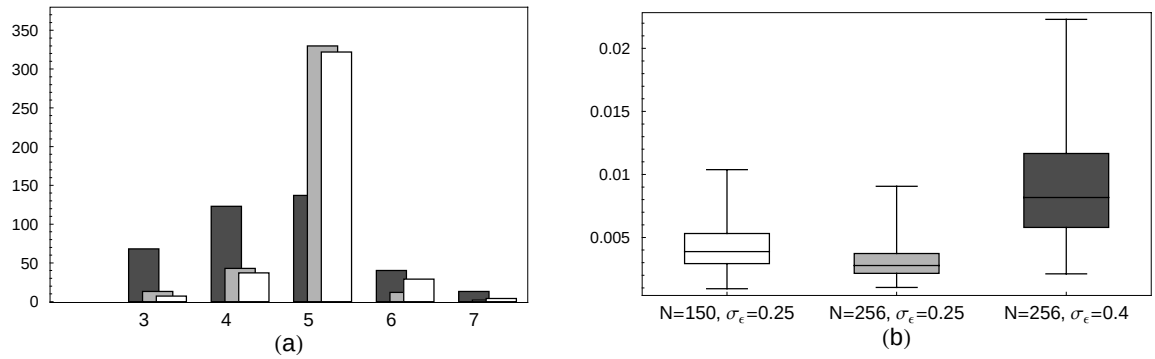

Fig. 4. (a): A frequency plot of the number of knots of the 400 linear GeD spline fits; (b): Box plots of the MSE values of the 400 linear GeD spline fits;

As can be seen from Fig. 4 (a), the number of knots of the GeD fits for higher noise level ($\sigma_\epsilon = 0.4$) is more dispersed over the range of values 3 to 7, than for the case of lower noise level ($\sigma_\epsilon = 0.25$) as is natural to expect. On the other hand, as can be seen from the box plots in Fig 4 (b), GeDS performs best in the case of larger sample size and lower noise level ($N = 256$, $\sigma_\epsilon = 0.25$). The median MSE value of the 400 linear fits, for $\sigma_\epsilon = 0.4$, with median number of internal knots $k = 5$, is 0.009. This is lower than the MSE value 0.012 of Luo and Wahba (1997), and is equal to that of Zhou and Shen (2001), both obtained using cubic splines with a higher number of regression functions (e.g., 13 for the fit of Luo and Wahba, 1997).

Example 2. The function f_2 (see Table 1) appears as a test example in Fan and Gijbels (1995), Luo and Wahba (1997), Denison et al. (1998) and Zhou and Shen (2001). Using the GeDS algorithm we have produced linear, quadratic and cubic fits which are illustrated in Fig. 5 and whose details are given in Table 5.

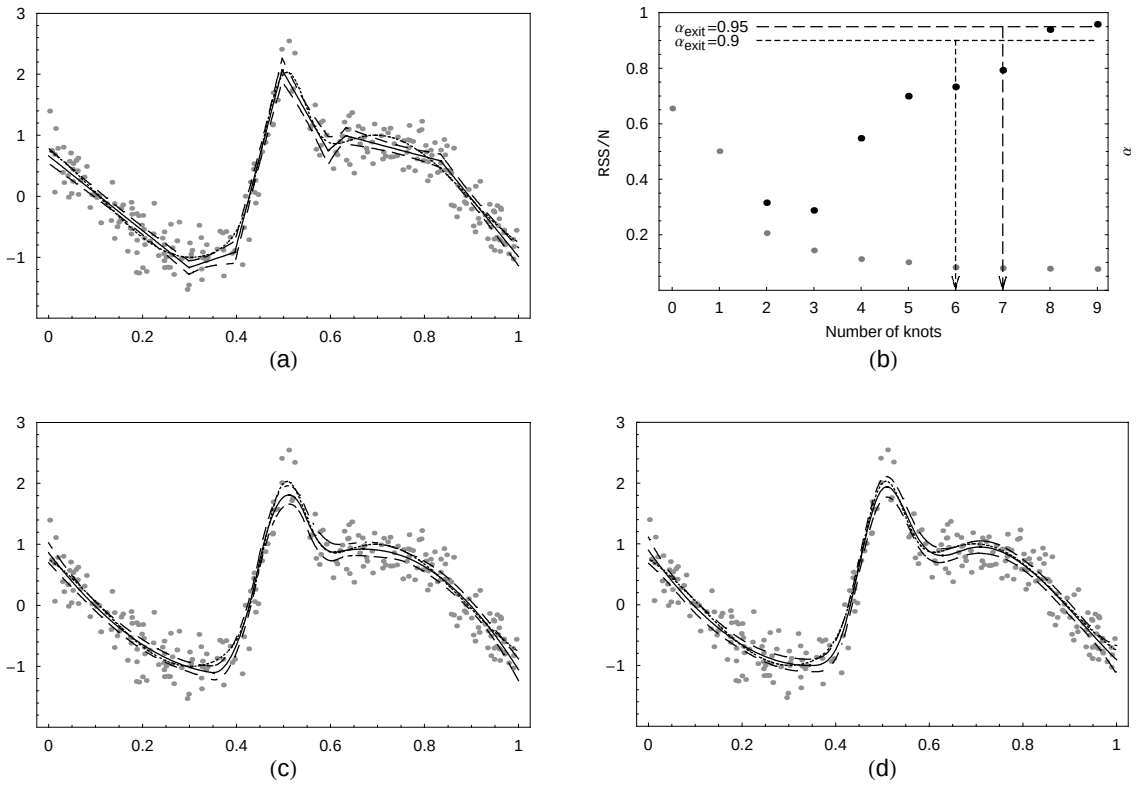

Fig. 5. (Example 2) Graphs of the final spline fits and confidence intervals, produced by GeDS: (a) linear; (c) quadratic; (d) cubic; (b) the values of the α-ratio - black dots and the values of RSS/N - grey dots, at each iteration in stage A; The dotted function in (a), (c), (d) is the true function.

Table 5. (Example 2) Summary of fits produced by GeDS.

Fit No	***Graph***	***n***	***k***	***Internal knots***	α_{exit}, β	L_2 ***– error, MSE***
1	Fig. 5, (a)	2	6	{0.30, 0.40, 0.50, 0.60, 0.63, 0.83}	0.9, 0.5	4.60, 0.009931
2	Fig. 5, (c)	3	5	{0.35, 0.45, 0.55, 0.61, 0.73}	0.9, 0.5	4.63, 0.005961
3	–	4	4	{0.40, 0.50, 0.57, 0.69}	0.9, 0.5	4.99, 0.019523
4	–	3	6	{0.33, 0.37, 0.45, 0.55, 0.61, 0.73}	0.95, 0.5	4.53, 0.006153
5	Fig. 5, (d)	4	5	{0.35, 0.42, 0.50, 0.57, 0.69}	0.95, 0.5	4.51, 0.004258

The SNR of the sample data is 3, as for fit No 3 of Example 1. Since f_2 is also relatively smooth we have used $\alpha_{\text{exit}} = 0.95$ and $\beta = 0.5$ in order to obtain the cubic fit in Fig. 5 (d), which has very good visual quality and low MSE value. The GeD spline fits No 1-3 of Table 5, with number of regression functions $k + n = 8$, are obtained with the default values $\alpha_{\text{exit}} = 0.9$ and $\beta = 0.5$. The cubic fit, No 3, with four knots, underfits the data while, as seen from Fig. 5 (a) and (c), the linear and quadratic fits are sufficiently accu-

rate. Adding one more knot by running GeDS with the higher value of $\alpha_{\text{exit}} = 0.95$ improves the cubic fit as illustrated by Fig. 5 (d). The 95% confidence intervals given in Fig 5 have been calculated using (24) with the known ('oracle') σ_ϵ. The behavior of the stopping rule of stage A, is illustrated in Fig. 5 (b). It can be seen that with $\alpha_{\text{exit}} = 0.9$ the algorithm exits with 6 internal knots for the linear fit and the RSS is 21.17. This means that the RSS of the linear fit with 8 knots is at least 90% of the value 21.17, i.e., the residual sum of squares has stabilized for three consecutive steps at which models with 6, 7 and 8 knots have been computed. If $\alpha_{\text{exit}} = 0.95$ the algorithm exits one step later, with 7 internal knots for the linear fit and RSS = 20.38 since the improvement in RSS for the next two consecutive steps is less than 5% of 20.38. So, we see that the stopping rule, based on the idea of exiting upon reaching a certain level of stabilization in RSS, tends to select models with the appropriate number of knots.

The median MSE value for the 400 linear and quadratic fits are equal to 0.0075 and 0.0095 respectively, and are comparable with those produced by other authors. For example, Luo and Wahba (1997) report MSE = 0.007 and number of basis functions equal to 13 for their HAS models. For all 400 linear fits the number of internal knots used by GeDS is between 5 and 7. It takes 1.58 seconds to compute fits No 1-3 and 1.88 seconds to compute fits No 4 and 5 of Table 5.

Based on the L_2-errors, given in Table 5, it can be seen that the best GeDS fit for this particular function is the cubic one, No 5 in Table 5. We have compared it with the optimal cubic spline fits PNOM and NOM with the same number of knots. The results are summarized in Table 6. As in Examples 1, the GeD fit is significantly better in terms of MSE and visual quality. The location of the knots is similar for GeDS and PNOM (fit No 2), both avoiding replicate knots. However, the optimal fit NOM (fit No 3) has 3 replicate knots at 0.5 and hence, produces an edge and visually deviates more strongly from the shape of the underlying function. The computation time needed, for GeDS is less then two seconds and for PNOM and NOM it is, respectively, 1.1 hour and 1.9 hour, using the *Mathematica* function `NMinimize`.

Table 6. (Example 2) The fits produced by GeDS, PNOM and NOM.

Fit No	***Method***	***n***	***k***	***Internal knots***	***L_2 – error, MSE***
1	GeDS	3	5	{0.35, 0.42, 0.50, 0.57, 0.69}	4.51, 0.004258
2	PNOM	3	5	{0.33, 0.44, 0.50, 0.55, 0.76}	4.47, 0.005216
3	NOM	3	5	{0.32, 0.50, 0.50, 0.50, 0.78}	4.43, 0.006598

Example 3. The HeaviSine function is one of the four functions introduced by Donoho and Johnstone (1994) and widely used as test examples by other authors, see for example Fan and Gijbels (1995), Luo and Wahba (1997), Denison et al. (1998), Zhou and Shen (2001), Lee (2000), Pittman (2002). It is a smooth function with two discontinuities at $x = 0.3$ and $x = 0.72$. It takes 55 seconds to obtain simultaneously the linear,

quadratic and cubic GeD spline fits, given in Table 7. In this and the following examples of spatially inhomogeneous curves, we have set the value for α_{exit} at 0.99, to prevent GeDS from producing a spline approximation which is too smooth for adequately representing the 'shape' of the data.

Table 7. (Example 3) Summary of fits produced by GeDS.

Fit No	***Graph***	***n***	***k***	***Internal knots***	α_{exit}, β	L_2 ***– error MSE***
1	–	2	18	{0.10, 0.13, 0.18, 0.29, 0.30, 0.30, 0.32, 0.38, 0.44, 0.57, 0.63, 0.71, 0.71, 0.72, 0.74, 0.83, 0.84, 0.99}	0.99, 0.5	46.56 0.2203
2	Fig. 6	3	17	{0.11, 0.16, 0.23, 0.29, 0.30, 0.31, 0.35, 0.41, 0.50, 0.60, 0.67, 0.71, 0.72, 0.73, 0.79, 0.84, 0.92}	0.99, 0.5	43.42 0.0482
3	–	4	16	{0.14, 0.20, 0.26, 0.30, 0.31, 0.33, 0.38, 0.46, 0.55, 0.64, 0.69, 0.72, 0.73, 0.77, 0.81, 0.89}	0.99, 0.5	44.82 0.0942

For the quadratic GeDS fit (No 2 in Table 7), illustrated in Fig. 6, the median number of regression functions $k+n$ is only 20 while the median MSE value 0.057, is comparable with 0.04 given by Luo and Wahba (1997) for their cubic spline model with 50 basis functions. Our GeDS algorithm uses between 17 and 21 internal knots to fit the 31 simulated data sets in the linear case. Based on the L_2-errors for the linear, quadratic and cubic fits given in Table 7, the best GeDS fit for this particular function is of degree 2.

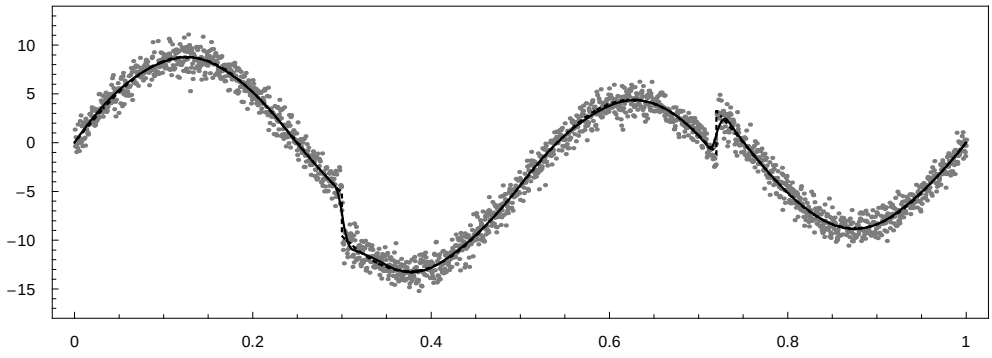

Fig. 6. (Example 3) Graph of the quadratic GeD spline fit. The dotted function is the true function.

Example 4. This function is known as the Doppler function. It is highly oscillating, especially near the origin, where most of the procedures fail to recover it. Using the GeDS algorithm we have obtained six different fits for the same data set with SNR equal to 7. Fits No 1-3, given in Table 8, are calculated simultaneously in 304 seconds with $\alpha_{\text{exit}} = 0.99$. The quadratic one (No 2) has 46 knots and MSE = 0.13. For comparison, the HAS cubic fit, produced by Luo and Wahba (1997) has MSE = 0.10 with 120 basis functions. Based on the quadratic GeD spline fits, obtained for 31 simulated data sets, the median MSE value is 0.089 and median number of knots is 62, using $\alpha_{\text{exit}} = 0.999$. The number of knots for the 31 quadratic fits is between 50 and 78.

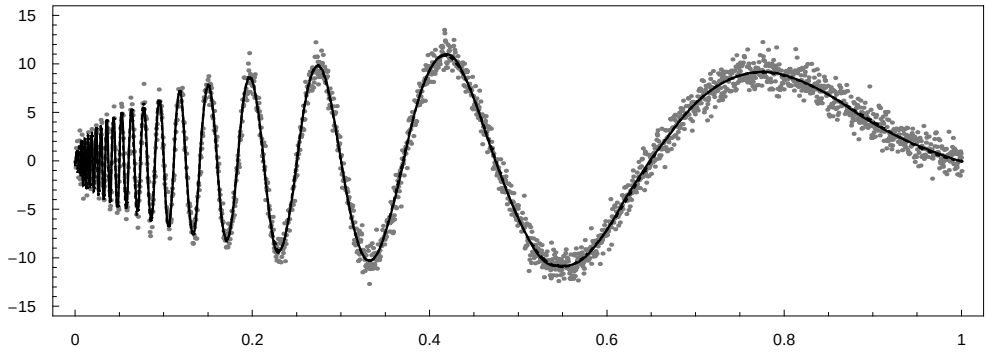

Fig. 7. (Example 4) Graph of the quadratic GeD spline fit. The dotted function is the true function.

Comparing the L_2-errors of the fits of degree 1, 2 and 3, summarized in Table 8 the best fit for the Doppler function is the quadratic one. The GeDS fit No 5, given in Fig. 7, is seen to fit very well the Doppler function near the origin, avoiding oversmoothing.

Table 8. (Example 4) Summary of fits produced by GeDS.

Fit No	***Graph***	***n***	***k***	α_{exit}, β	L_2 ***– error, MSE***
1	–	2	47	0.99, 0.5	48.24, 0.199802
2	–	3	46	0.99, 0.5	46.77, 0.125328
3	–	4	45	0.99, 0.5	49.04, 0.233945
4	–	2	74	0.999, 0.5	45.21, 0.114633
5	Fig. 7	3	73	0.999, 0.5	44.92, 0.060037
6	–	4	72	0.999, 0.5	46.10, 0.106811

Example 5. The Bumps function is very wiggly and also difficult to fit. Following the prescription for choosing α_{exit} in the case of fitting wiggly functions with high SNR, we have set $\alpha_{\text{exit}} = 0.99$ and have obtained the GeDS fits whose details are summarized in Table 9.

Table 9. (Example 5) Summary of fits produced by GeDS.

Fit No	***Graph***	***n***	***k***	α_{exit}, β	L_2 ***– error, MSE***
1	–	2	83	0.99, 0.5	48.59, 0.283631
2	–	3	82	0.99, 0.5	56.03, 0.631448
3	–	4	81	0.99, 0.5	66.44, 1.198390
4	Fig. 8	2	103	0.999, 0.5	44.51, 0.140580
5	–	3	102	0.999, 0.5	47.96, 0.264664
6	–	4	101	0.999, 0.5	52.29, 0.445403

Looking at the L_2-errors we see that the fit with the lowest L_2-error is the linear one, which is illustrated in Fig. 8. A linear fit for Bumps is given also by Lee (2000) whose MDL procedure automatically chooses the order of the fit within the range 1 to 4. Based on 31 simulated data sets the median MSE value for the linear fit is 0.22, for the qua-

dratic fit it is 0.51 and the median number of knots is 90. The GeDS estimator places between 79 and 102 knots for these 31 fits. For comparison, the median MSE value reported by Pittman (2002) for the cubic AGS fit is 0.4001, for a certain median number of knots, which is not reported.

As fits No 1-6 in Table 9 indicate, by increasing the α_{exit} parameter it is possible to improve the quality of the final fit, allowing GeDS to add more knots where necessary. Fits No 1-3 are obtained simultaneously in 795 seconds, whereas fits No 4-6 are computed in 1255 seconds.

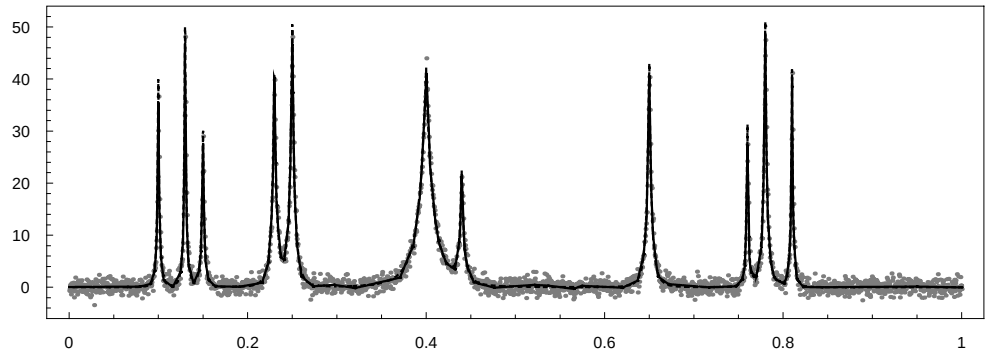

Fig. 8. (Example 5) Graph of the linear GeD spline fit. The dotted function is the true function.

Example 6. For the Blocks function, in order to obtain fits No 1-4 given in Table 10, we have run GeDS with $\alpha_{\text{exit}} = 0.99$ and $\alpha_{\text{exit}} = 0.999$. The details of the linear and quadratic fits for both values of α_{exit}, are presented in Table 10. The best fit, produced by GeDS is linear, No 3, and it is illustrated in Fig 9.

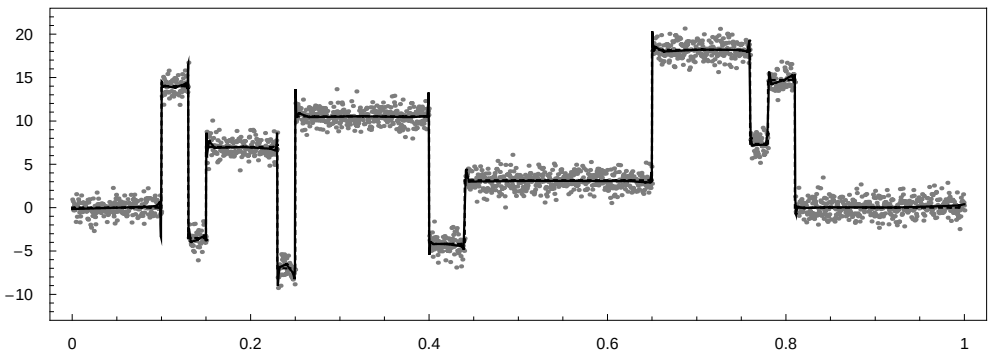

Fig. 9. (Example 6) Graph of the linear GeD spline fit. The dotted function is the true function.

Fits No 1-2 are obtained in 344 seconds and No 3-4 in 856 seconds. Our median MSE value, based on 31 runs with $\alpha_{\text{exit}} = 0.999$ is 0.12 with 83 median number of knots. For comparison, the median MSE value given by Zhou and Shen (2001) is 0.08, who do not report the number of knots of their SARS fit.

Table 10. (Example 6) Summary of fits produced by GeDS.

Fit No	***Graph***	***n***	***k***	α_{exit}, β	$L_2 - error$, ***MSE***
1	–	2	53	0.99, 0.5	55.63, 0.642906
2	–	3	52	0.99, 0.5	59.80, 0.860989
3	Fig. 9	2	85	0.999, 0.5	42.43, 0.082962
4	–	3	84	0.999, 0.5	43.68, 0.126953

7. Discussion and conclusions.

Based on the results of Section 4, we can conclude that the knots of GeDS, placed according to the knot averaging method, approximate very well the optimal variation diminishing knots of stage B (see Section 3). Thus, based on its variation diminishing (shape preserving) character, the GeD spline estimator has been shown in Section 6 to be successful in fitting both smooth and spatially inhomogeneous functions. Its large sample statistical properties, such as asymptotic normality, established in Kaishev et al. (2006) facilitates the construction of asymptotic confidence intervals with respect to the unknown function f, illustrated in Examples 1 and 2 of Section 6.

Based on the results presented in the present paper and also in Kaishev et al. (2006) we can conclude that the GeDS method is a fast, stable, automatic statistically viable estimation procedure with an appropriate geometric interpretation which allows to follow the entire fitting process. The existence of the two parameters α_{exit} and β combines automation with some flexibility in tuning GeDS to cope with the particular noise level, and smoothness characteristics of the underlying function. The numerical results of Section 6 show that the GeD spline regression models are comparable with those obtained with other methods, including the penalized non-linear optimization method of Lindstrom (1999). In particular, in Examples 1 and 2, GeDS managed to find knot placements which are nearly optimal but avoiding replicate knots, as seen from Tables 4 and 6.

Acknowledgements

The authors would like to acknowledge support received through a research grant from the UK Institute of Actuaries.

Appendix

Proof of Theorem 1. Note that, for $n = 2$, $\xi_i \equiv \xi_i^*$, $i = 1, \ldots, p$, hence $V^a[g] \equiv V[g]$ and the bound in (22), which is zero, is sharp. For $n > 2$, from (6) it follows that $\xi_1^* \equiv a \equiv \xi_1$ and $\xi_p^* \equiv b \equiv \xi_p$, and from the definitions of $V[g]$ and $V^a[g]$, (9) and (14) respectively, we have

$$\|V[g] - V^a[g]\| = \max_{t\in[a,b]} \mid \sum_{i=1}^{p} (g(\xi_i^*) - g(\xi_i))\, N_{i,n}(t) \mid$$

$$\leq \max_{t\in[a,b]} \sum_{i=1}^{p} \mid (g(\xi_i^*) - g(\xi_i)) \mid N_{i,n}(t)$$

$$\leq \max_{t\in[a,b]} \sum_{i=1}^{p} \{\max_{j\in\{2,\ldots,p-1\}} \mid g(\xi_j^*) - g(\xi_j) \mid\} N_{i,n}(t)$$

$$\leq \max_{j\in\{2,\ldots,p-1\}} \mid g(\xi_j^*) - g(\xi_j) \mid \max_{t\in[a,b]} \sum_{i=1}^{p} N_{i,n}(t)$$

$$= \max_{j\in\{2,\ldots,p-1\}} \mid g(\xi_j^*) - g(\xi_j) \mid, \qquad (25)$$

where the last equality follows from the partition of unity property of B-splines (See Section 2). Applying the definition of the modulus of continuity to (25) we have

$$\|V[g] - V^a[g]\| \leq \max_{j\in\{2,\ldots,p-1\}} \mid g(\xi_j^*) - g(\xi_j) \mid \qquad (26)$$

$$\leq \omega(g; \max_{j\in\{2,\ldots,p-1\}} \mid \xi_j^* - \xi_j \mid). \qquad (27)$$

From (13), it follows that $\xi_{j-(n-2)} < t_{j+1}$ and $t_{j+n-1} < \xi_{j+(n-2)}$, $j = 2, \ldots, p-1$. From the definition (6) of the Greville sites ξ_i^* we have $t_{j+1} < \xi_j^* < t_{j+n-1}$, $j = 2, \ldots, p-1$, where we define $\xi_{1-l} := a$ and $\xi_{p+l} := b$, $l = 1, 2, \ldots$, to avoid difficulties in notation. Hence, $\xi_{j-(n-2)} < \xi_j^* < \xi_{j+(n-2)}$, $j = 2, \ldots, p-1$. Applying the latter inequalities and assuming that the maximum in (27) is achieved for some $j = j^m$, in the case $\xi_{j^m}^* > \xi_{j^m}$, we have

$$\mid \xi_{j^m}^* - \xi_{j^m} \mid \leq \xi_{j^m+(n-2)} - \xi_{j^m} \leq (n-2) \max_{j\in\{j^m,\ldots,j^m+(n-2)-1\}} (\xi_{j+1} - \xi_j) \qquad (28)$$

and if $\xi_{j^m}^* < \xi_{j^m}$ we have

$$\mid \xi_{j^m}^* - \xi_{j^m} \mid \leq \xi_{j^m} - \xi_{j^m-(n-2)} \leq (n-2) \max_{j\in\{j^m-(n-2),\ldots,j^m-1\}} (\xi_{j+1} - \xi_j). \qquad (29)$$

It is not difficult to see that both maximums in (28) and (29) are bounded by $(n-2) \max_{j\in\{1,\ldots,p-1\}} (\xi_{j+1} - \xi_j)$, so, from (27), we obtain

$$\|V[g] - V^a[g]\| \leq \omega(g; \mid \xi_{j^m}^* - \xi_{j^m} \mid)$$

$$\leq \omega(g; (n-2) \max_{j\in\{1,\ldots,p-1\}} (\xi_{j+1} - \xi_j)). \qquad (30)$$

Using the fact that $\omega(g; h)$ is a monotone function in h and that it is also subadditive in h, i.e., $\omega(g; h + w) \leq \omega(g; h) + \omega(g; w)$, from (30) we finally obtain

$$\|V[g] - V^a[g]\| \leq (n-2)\, \omega(g; \max_{j\in\{1,\ldots,p-1\}} (\xi_{j+1} - \xi_j)) .$$

This completes the proof of Theorem 1.□

Proof of Corollary 1.1. This follows directly from (15) and from the definition of $\omega(g; h)$, i.e.,

$$\|V[t] - V^a[t]\| = \|t - \sum_{i=1}^{p} \xi_i\, N_{i,n}(t)\| \leq (n-2) \max_{j\in\{1,\ldots,p-1\}} (\xi_{j+1} - \xi_j).\square$$

Proof of Corollary 1.2. From (26) we have

$$\|V[g] - V^a[g]\| \leq \max_{j\in\{2,\ldots,p-1\}} \mid g(\xi_j^*) - g(\xi_j) \mid$$

$$= \max_{j\in\{2,\ldots,p-1\}} \mid \sum_{i=1}^{p} \alpha_i N_{i,2}(\xi_j^*) - \sum_{i=1}^{p} \alpha_i N_{i,2}(\xi_j) \mid$$

$$= \max_{j\in\{2,\ldots,p-1\}} \mid \sum_{i=1}^{p} \alpha_i N_{i,2}(\xi_j^*) - \alpha_j \mid$$

$$= \max_{j\in\{2,\ldots,p-1\}} \Big| \sum_{i=j-(n-2)}^{j+(n-2)} \alpha_i N_{i,2}(\xi_j^*) - \alpha_j \Big| \qquad (31)$$

since, as shown in the course of the proof of Theorem 1, $\xi_{j-(n-2)} < \xi_j^* < \xi_{j+(n-2)}$, $j = 2, \ldots, p-1$. In the last equality we have defined $\xi_{1-l} := a$ and $\xi_{p+l} := b$, $l = 1, 2, \ldots$. Since g is a linear spline, we know that if $\xi_q \le \xi_j^* \le \xi_{q+1}$, $j-(n-2) \le q < j+(n-2)$ then $\sum_{i=j-(n-2)}^{j+(n-2)} \alpha_i N_{i,2}(\xi_j^*) = \alpha_q N_{q,2}(\xi_j^*) + \alpha_{q+1} N_{q+1,2}(\xi_j^*)$, which is a convex combination of only two B-spline coefficients. Assuming that the maximum in (31) is achieved for some $j = j^m$, in the case when $\xi_{j^m} < \xi_q \le \xi_{j^m}^* \le \xi_{q+1}$, $j^m \le q < j^m + (n-2)$ we have

$$\max_{j\in\{2,\ldots,p-1\}} \Big| \sum_{i=j-(n-2)}^{j+(n-2)} \alpha_i N_{i,2}(\xi_j^*) - \alpha_j \Big| = \Big| \sum_{i=q}^{q+1} \alpha_i N_{i,2}(\xi_{j^m}^*) - \alpha_{j^m} \Big|$$

$$\le (\max_{q\in\{j^m,\ldots,j^m+(n-2)\}} \{\alpha_q\} - \min_{q\in\{j^m,\ldots,j^m+(n-2)\}} \{\alpha_q\}) \qquad (32)$$

and if $\xi_q \le \xi_{j^m}^* \le \xi_{q+1} \le \xi_{j^m}$, $j^m - (n-2) \le q < j^m$ we have

$$\max_{j\in\{2,\ldots,p-1\}} \Big| \sum_{i=j-(n-2)}^{j+(n-2)} \alpha_i N_{i,2}(\xi_j^*) - \alpha_j \Big| = \Big| \sum_{i=q}^{q+1} \alpha_i N_{i,2}(\xi_{j^m}^*) - \alpha_{j^m} \Big|$$

$$\le (\max_{q\in\{j^m-(n-2),\ldots,j^m\}} \{\alpha_q\} - \min_{q\in\{j^m-(n-2),\ldots,j^m\}} \{\alpha_q\}) \,. \qquad (33)$$

It is not difficult to see that both differences on the right-hand sides of the inequalities in (32) and (33) are bounded by

$$\max_{j\in\{1,\ldots,p-(n-2)\}} (\max_{q\in\{j,\ldots,j+(n-2)\}} \{\alpha_q\} - \min_{q\in\{j,\ldots,j+(n-2)\}} \{\alpha_q\}) \,.$$

Hence, from (31), we obtain

$$\|V[g] - V^a[g]\| \le$$
$$\max_{j\in\{1,\ldots,p-(n-2)\}} (\max_{q\in\{j,\ldots,j+(n-2)\}} \{\alpha_q\} - \min_{q\in\{j,\ldots,j+(n-2)\}} \{\alpha_q\})$$

This completes the proof of Corollary 2.1.□

Proof of Theorem 2. Consider the $\max_{j\in\{2,\ldots,p-1\}} \mid \xi_j - \xi_j^* \mid$ and assume it is achieved for some j^m, $n \le j^m < p-n$. We can express $\xi_{j^m}^*$ in terms of ξ_{j^m}, using the definitions (6) and (19). After some algebra it is not difficult to see that

$$\mid \xi_{j^m} - \xi_{j^m}^* \mid$$

$$= \frac{1}{(n-1)^2} \Big| \sum_{i=1}^{n-2} i\,(\xi_{j^m+(n-1-i)} + \xi_{j^m-(n-1-i)}) - (n-1)(n-2)\,\xi_{j^m} \Big| \qquad (34)$$

and if we now rearrange the terms in the sum in (34), we obtain

$$\mid \xi_{j^m} - \xi_{j^m}^* \mid = \frac{1}{(n-1)^2} \Big| \sum_{i=1}^{n-2} i\,((\xi_{j^m+(n-1-i)} - \xi_{j^m}) - (\xi_{j^m} - \xi_{j^m-(n-1-i)})) \Big| \,. \qquad (35)$$

Assume that $\sum_{i=1}^{n-2} i\,(\xi_{j^m+(n-1-i)} - \xi_{j^m}) > \sum_{i=1}^{n-2} i(\xi_{j^m} - \xi_{j^m-(n-1-i)})$. In this case, it is not difficult to see that (35) is bounded by

$$\Big| \xi_{j^m} - \xi_{j^m}^* \Big| \le \frac{1}{(n-1)^2} \sum_{i=1}^{n-2} i\,(\xi_{j^m+(n-1-i)} - \xi_{j^m})$$

$$\leq \frac{1}{(n-1)^2}\ \frac{(n-2)\,(n-1)}{2}\ (\xi_{j^m+(n-2)} - \xi_{j^m})$$

$$\leq \frac{(n-2)}{2\,(n-1)}\ (\xi_{j^m+(n-2)} - \xi_{j^m})$$

$$\leq \frac{(n-2)^2}{2\,(n-1)}\ \max_{j\in\{1,\ldots,p-1\}}\,(\xi_{j+1} - \xi_j)\ . \qquad (36)$$

Similarly, it can be shown that if $\sum_{i=1}^{n-2} i\,(\xi_{j^m+(n-1-i)} - \xi_{j^m}) \leq \sum_{i=1}^{n-2} i(\xi_{j^m} - \xi_{j^m-(n-1-i)})$ the bound in (36) also holds. Thus, from (36) and (27) we have

$$\|V[g] - V^a[g]\| \leq \omega\Big(g;\ \frac{(n-2)^2}{2\,(n-1)}\ \max_{j\in\{1,\ldots,p-1\}}\,(\xi_{j+1} - \xi_j)\Big) \qquad (37)$$

Using the monotonicity and subadditivity of $\omega(g;\,h)$ in h, from (37) we finally obtain

$$\|V[g] - V^a[g]\| \leq \Big\lceil \frac{(n-2)^2}{2\,(n-1)} \Big\rceil\ \omega(g;\ \max_{j\in\{1,\ldots,p-1\}}\,(\xi_{j+1} - \xi_j))$$

where $\lceil \nu \rceil := \min\{z \in \mathbb{Z} : \nu \leq z\}$. Applying similar reasoning, one can show that the bound (21) holds also in the case when $2 \leq j^m < n$ or $p - n \leq j^m < p - 1$. This completes the proof of Theorem 2.□

Proof of Corollary 2.1. This follows directly from (37) and from the definition of $\omega(g;\,h)$, i.e.,

$$\|V[t] - V^a[t]\| = \|t - \textstyle\sum_{i=1}^{p} \xi_i\, N_{i,n}(t)\| \leq \frac{(n-2)^2}{2\,(n-1)}\ \max_{j\in\{1,\ldots,p-1\}}\,(\xi_{j+1} - \xi_j)\ .\square$$

Proof of Corollary 2.2. From (31) we have

$$\|V[g] - V^a[g]\| \leq \max_{j\in\{2,\ldots,p-1\}}\ \Big|\ \textstyle\sum_{i=j-1}^{j+1} \alpha_i\, N_{i,2}(\xi_j^*) - \alpha_j\ \Big|\ . \qquad (38)$$

We need to consider the cases when $\xi_{j-1} < \xi_j^* \leq \xi_j,\ 2 \leq j \leq p$ and $\xi_j \leq \xi_j^* < \xi_{j+1},\ 1 \leq j \leq p-1$. In the first case, applying the Mansfield-De Boor-Cox recurrence formula (4) to express $N_{i,2}(\xi_j^*)$ in the maximum in (38) we have

$$\max_{j\in\{2,\ldots,p\}}\ |\ \alpha_{j-1}\, N_{j-1,2}(\xi_j^*) + \alpha_j\, N_{j,2}(\xi_j^*) - \alpha_j\ |$$

$$= \max_{j\in\{2,\ldots,p\}}\ \Big|\ \alpha_{j-1}\ \frac{\xi_j - \xi_j^*}{\xi_j - \xi_{j-1}} + \alpha_j\ \frac{\xi_j^* - \xi_{j-1}}{\xi_j - \xi_{j-1}} - \alpha_j\ \frac{\xi_j - \xi_{j-1}}{\xi_j - \xi_{j-1}}\ \Big|$$

$$= \max_{j\in\{2,\ldots,p\}}\ \Big|\ (\alpha_{j-1} - \alpha_j)\ \Big|\ \Big(\frac{\xi_j - \xi_j^*}{\xi_j - \xi_{j-1}}\Big)$$

$$< \max_{j\in\{2,\ldots,p\}}\ \Big|\ (\alpha_{j-1} - \alpha_j)\ \Big|\ \Big(\frac{\frac{1}{4}\,(\xi_j - \xi_{j-1})}{\xi_j - \xi_{j-1}}\Big)$$

$$= \frac{1}{4}\ \max_{j\in\{2,\ldots,p\}}\ |\ (\alpha_{j-1} - \alpha_j)\ |\ , \qquad (39)$$

where we have expressed ξ_j^* in terms of ξ_j, using (6) and the definition of t_{i+n}, $i = 1,\ \ldots,\ k$ in Theorem 2, and have used the fact that $\xi_j - \xi_{j-1} > \xi_{j+1} - \xi_j$ to arrive at the last inequality. Similarly, it is not difficult to see that the same bound as in (39) holds in the case when $\xi_j \leq \xi_j^* \leq \xi_{j+1}$. This completes the proof of Corollary 2.2.□

References

Agarwal, G. G. and Studden, W. J. (1980). Asymptotic integrated mean square error using least squares and bias minimizing splines. *Ann. Statist*., **8**, 1307-1325.

Biller, C. (2000). Adaptive Bayesian regression splines in semiparametric generalized linear models. *J. Comput. and Graph. Stat.*, **9**, 122-140.

De Boor, C. (2001). *A practical Guide to Splines*, Revised Edition, New York: Springer.

De Boor, C. and Rice, J. (1968). Least squares cubic spline approximation II. Variable knots. Comp. Sci.Dpt. *Technical Report* 21, Purdue Univrsity, West Laffayet, Indiana.

Denison, D., Mallick, B., and Smith, A. (1998). Automatic Bayesian curve fitting, *J. R. Statist. Soc.*, B, **60**, 333-350.

Donoho, D. and Johnstone, I (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, **81**, 425-455.

Eubank, R.(1988). *Spline smoothing and Nonparametric Regression.* Dekker, New York.

Fan, J. and Gijbels, I. (1995). Data-driven bandwidth selection in local polynomial fitting:Variable bandwidth and spatial adaptation. *J. R. Statist. Soc*. B, **57**, 371-394.

Farin, G. (2002). *Curves and Surfaces for CAGD*, Fifth Edition, San Francisco: Morgan Kaufmann.

Friedman, J. H. (1991). Multivariate adaptive regression splines (with discussion). *Ann. Statist*. **19**, 1-141.

Friedman, J. H. and Silverman, B.W. (1989). Flexible Parsimonious smoothing and additive modeling (with discussion).*Technometrics*., **31**, . 3-39.

Hu, Y. (1993). An algorithm for data reduction using splines with free knots. *IMA J. Numer. Anal.*, **13**, 328-343.

Huang, J. Z. (2003). Local assymptotics for polynomial spline regression. *Ann. Statist.*, **31**, 1600-1635.

Jupp, D. (1978). Approximation to data by splines with free knots. *SIAM J. Num. Analysis.*, **15**, 328-343.

Kaishev, V. K., Dimitrova, D. S., Haberman, S. and Verrall R. (2006). Geometrically designed, variable knot regression splines: Asymptotics and inference. Statistical Res. Paper 28, Cass Business School, City University, London.

Lee, T. C. M. (2000). Regression spline smoothing using the minimum description length principle. *Stat. & Prob. Letters*, **48**, 71-82.

Lindstrom, M. J. (1999). Penalized estimation of free-knot splines. *J. Comput. and Graph. Stat.*, **8**, 2, 333-352.

Luo, Z., and Wahba, G. (1997). Hybrid adaptive splines. *J. Am. Statist. Ass.*, **92**, 107-115.

Marx, B. D. and Eilers, P. H.C. (1996). Flexible Smoothing with B-splines and Penalties. *Stat. Science*, **11**, 2, 89-121.

Micchelli, C. A. Rivlin, T.J. and Winograd, S. (1976). The optimal recovery of smooth functions. *Numer. Math.*, **26**, 191-200.

Pittman, J. (2002). Adaptive Splines and Genetic Algorithms. *J. Comput. and Graph. Stat.*, **11**, 3, 1-24.

Rupert, D. (2002). Selecting the number of knots for penalized splines. *J. Comput. and Graph. Stat.*, **11**, 4, 735-757.

Rupert, D., and Carroll, R. J. (2000). Spatially-Adaptive penalties for spline fitting. *Australian and New Zealand Journal of Statistics*, **42**, 205-223.

Smith, P. L. (1982). Curve fitting and modeling with splines using statistical variable selection techniques. *Report NASA 166034*, Langley Research Center, Hampton, VA.

Smith, M. and Kohn, R. (1996). Nonparametric regression using Bayesian variable selection. *J. Econometrics, **75***, 317-344.

Stone, C. J., Hansen, M.H., Kooperberg, C. and Truong, Y. K. (1997). Polynomial Splines and their tensor products in extended linear modeling. *Ann. Statist*., **25**, 1371-1470.

Zhou, S. and Shen, X. (2001). Spatially adaptive regression splines and accurate knot selection schemes. *J. Am. Statist. Ass.*, **96**, 247-259.

Wahba, G. (1990). *Spline Models for Observational Data.* SIAM, Philadelphia

Wood, S. N. (2003). Thin plate regression splines. *J. R. Statist. Soc*. B, **65**, Part 1, 95-114.

Cass Business School
City of London

FACULTY OF ACTUARIAL SCIENCE AND INSURANCE

Actuarial Research Papers since 2001

Report Number	Date	Publication Title	Author
135.	February 2001.	On the Forecasting of Mortality Reduction Factors. ISBN 1 901615 56 1	Steven Haberman Arthur E. Renshaw
136.	February 2001.	Multiple State Models, Simulation and Insurer Insolvency. ISBN 1 901615 57 X	Steve Haberman Zoltan Butt Ben Rickayzen
137.	September 2001	A Cash-Flow Approach to Pension Funding. ISBN 1 901615 58 8	M. Zaki Khorasanee
138.	November 2001	Addendum to "Analytic and Bootstrap Estimates of Prediction Errors in Claims Reserving". ISBN 1 901615 59 6	Peter D. England
139.	November 2001	A Bayesian Generalised Linear Model for the Bornhuetter-Ferguson Method of Claims Reserving. ISBN 1 901615 62 6	Richard J. Verrall
140.	January 2002	Lee-Carter Mortality Forecasting, a Parallel GLM Approach, England and Wales Mortality Projections. ISBN 1 901615 63 4	Arthur E.Renshaw Steven Haberman.
141.	January 2002	Valuation of Guaranteed Annuity Conversion Options. ISBN 1 901615 64 2	Laura Ballotta Steven Haberman
142.	April 2002	Application of Frailty-Based Mortality Models to Insurance Data. ISBN 1 901615 65 0	Zoltan Butt Steven Haberman
143.	Available 2003	Optimal Premium Pricing in Motor Insurance: A Discrete Approximation.	Russell J. Gerrard Celia Glass
144.	December 2002	The Neighbourhood Health Economy. A Systematic Approach to the Examination of Health and Social Risks at Neighbourhood Level. ISBN 1 901615 66 9	Les Mayhew
145.	January 2003	The Fair Valuation Problem of Guaranteed Annuity Options : The Stochastic Mortality Environment Case. ISBN 1 901615 67 7	Laura Ballotta Steven Haberman
146.	February 2003	Modelling and Valuation of Guarantees in With-Profit and Unitised With-Profit Life Insurance Contracts. ISBN 1 901615 68 5	Steven Haberman Laura Ballotta Nan Want
147.	March 2003.	Optimal Retention Levels, Given the Joint Survival of Cedent and Reinsurer. ISBN 1 901615 69 3	Z. G. Ignatov Z.G., V.Kaishev R.S. Krachunov
148.	March 2003.	Efficient Asset Valuation Methods for Pension Plans. ISBN 1 901615707	M. Iqbal Owadally
149.	March 2003	Pension Funding and the Actuarial Assumption Concerning Investment Returns. ISBN 1 901615 71 5	M. Iqbal Owadally

150.	Available August 2004	Finite time Ruin Probabilities for Continuous Claims Severities	D. Dimitrova Z. Ignatov V. Kaishev
151.	August 2004	Application of Stochastic Methods in the Valuation of Social Security Pension Schemes. ISBN 1 901615 72 3	Subramaniam Iyer
152.	October 2003.	Guarantees in with-profit and Unitized with profit Life Insurance Contracts; Fair Valuation Problem in Presence of the Default Option[1]. ISBN 1-901615-73-1	Laura Ballotta Steven Haberman Nan Wang
153.	December 2003	Lee-Carter Mortality Forecasting Incorporating Bivariate Time Series. ISBN 1-901615-75-8	Arthur E. Renshaw Steven Haberman
154.	March 2004.	Operational Risk with Bayesian Networks Modelling. ISBN 1-901615-76-6	Robert G. Cowell Yuen Y, Khuen Richard J. Verrall
155.	March 2004.	The Income Drawdown Option: Quadratic Loss. ISBN 1 901615 7 4	Russell Gerrard Steven Haberman Bjorn Hojgarrd Elena Vigna
156.	April 2004	An International Comparison of Long-Term Care Arrangements. An Investigation into the Equity, Efficiency and sustainability of the Long-Term Care Systems in Germany, Japan, Sweden, the United Kingdom and the United States. ISBN 1 901615 78 2	Martin Karlsson Les Mayhew Robert Plumb Ben D. Rickayzen
157.	June 2004	Alternative Framework for the Fair Valuation of Participating Life Insurance Contracts. ISBN 1 901615-79-0	Laura Ballotta
158.	July 2004.	An Asset Allocation Strategy for a Risk Reserve considering both Risk and Profit. ISBN 1 901615-80-4	Nan Wang
159.	December 2004	Upper and Lower Bounds of Present Value Distributions of Life Insurance Contracts with Disability Related Benefits. ISBN 1 901615-83-9	Jaap Spreeuw
160.	January 2005	Mortality Reduction Factors Incorporating Cohort Effects. ISBN 1 90161584 7	Arthur E. Renshaw Steven Haberman
161.	February 2005	The Management of De-Cumulation Risks in a Defined Contribution Environment. ISBN 1 901615 85 5.	Russell J. Gerrard Steven Haberman Elena Vigna
162.	May 2005	The IASB Insurance Project for Life Insurance Contracts: Impart on Reserving Methods and Solvency Requirements. ISBN 1-901615 86 3.	Laura Ballotta Giorgia Esposito Steven Haberman
163.	September 2005	Asymptotic and Numerical Analysis of the Optimal Investment Strategy for an Insurer. ISBN 1-901615-88-X	Paul Emms Steven Haberman
164.	October 2005.	Modelling the Joint Distribution of Competing Risks Survival Times using Copula Functions. I SBN 1-901615-89-8	Vladimir Kaishev Dimitrina S, Dimitrova Steven Haberman
165.	November 2005.	Excess of Loss Reinsurance Under Joint Survival Optimality. ISBN1-901615-90-1	Vladimir K. Kaishev Dimitrina S. Dimitrova
166.	November 2005.	Lee-Carter Goes Risk-Neutral. An Application to the Italian Annuity Market. ISBN 1-901615-91-X	Enrico Biffis Michel Denuit

167.	November 2005	Lee-Carter Mortality Forecasting: Application to the Italian Population. ISBN 1-901615-93-6	Steven Haberman Maria Russolillo
168.	February 2006	The Probationary Period as a Screening Device: Competitive Markets. ISBN 1-901615-95-2	Jaap Spreeuw Martin Karlsson
169.	February 2006	Types of Dependence and Time-dependent Association between Two Lifetimes in Single Parameter Copula Models. ISBN 1-901615-96-0	Jaap Spreeuw
170.	April 2006	Modelling Stochastic Bivariate Mortality ISBN 1-901615-97-9	Elisa Luciano Jaap Spreeuw Elena Vigna.
171.	February 2006	Optimal Strategies for Pricing General Insurance. ISBN 1901615-98-7	Paul Emms Steve Haberman Irene Savoulli
172.	February 2006	Dynamic Pricing of General Insurance in a Competitive Market. ISBN1-901615-99-5	Paul Emms
173.	February 2006	Pricing General Insurance with Constraints. ISBN 1-905752-00-8	Paul Emms
174.	May 2006	Investigating the Market Potential for Customised Long Term Care Insurance Products. ISBN 1-905752-01-6	Martin Karlsson Les Mayhew Ben Rickayzen

Statistical Research Papers

Report Number	Date	Publication Title	Author
1.	December 1995.	Some Results on the Derivatives of Matrix Functions. ISBN 1 874 770 83 2	P. Sebastiani
2.	March 1996	Coherent Criteria for Optimal Experimental Design. ISBN 1 874 770 86 7	A.P. Dawid P. Sebastiani
3.	March 1996	Maximum Entropy Sampling and Optimal Bayesian Experimental Design. ISBN 1 874 770 87 5	P. Sebastiani H.P. Wynn
4.	May 1996	A Note on D-optimal Designs for a Logistic Regression Model. ISBN 1 874 770 92 1	P. Sebastiani R. Settimi
5.	August 1996	First-order Optimal Designs for Non Linear Models. ISBN 1 874 770 95 6	P. Sebastiani R. Settimi
6.	September 1996	A Business Process Approach to Maintenance: Measurement, Decision and Control. ISBN 1 874 770 96 4	Martin J. Newby
7.	September 1996.	Moments and Generating Functions for the Absorption Distribution and its Negative Binomial Analogue. ISBN 1 874 770 97 2	Martin J. Newby
8.	November 1996.	Mixture Reduction via Predictive Scores. ISBN 1 874 770 98 0	Robert G. Cowell.
9.	March 1997.	Robust Parameter Learning in Bayesian Networks with Missing Data. ISBN 1 901615 00 6	P.Sebastiani M. Ramoni
10.	March 1997.	Guidelines for Corrective Replacement Based on Low Stochastic Structure Assumptions. ISBN 1 901615 01 4.	M.J. Newby F.P.A. Coolen

11.	March 1997	Approximations for the Absorption Distribution and its Negative Binomial Analogue. ISBN 1 901615 02 2	Martin J. Newby
12.	June 1997	The Use of Exogenous Knowledge to Learn Bayesian Networks from Incomplete Databases. ISBN 1 901615 10 3	M. Ramoni P. Sebastiani
13.	June 1997	Learning Bayesian Networks from Incomplete Databases. ISBN 1 901615 11 1	M. Ramoni P.Sebastiani
14.	June 1997	Risk Based Optimal Designs. ISBN 1 901615 13 8	P.Sebastiani H.P. Wynn
15.	June 1997.	Sampling without Replacement in Junction Trees. ISBN 1 901615 14 6	Robert G. Cowell
16.	July 1997	Optimal Overhaul Intervals with Imperfect Inspection and Repair. ISBN 1 901615 15 4	Richard A. Dagg Martin J. Newby
17.	October 1997	Bayesian Experimental Design and Shannon Information. ISBN 1 901615 17 0	P. Sebastiani. H.P. Wynn
18.	November 1997.	A Characterisation of Phase Type Distributions. ISBN 1 901615 18 9	Linda C. Wolstenholme
19.	December 1997	A Comparison of Models for Probability of Detection (POD) Curves. ISBN 1 901615 21 9	Wolstenholme L.C
20.	February 1999.	Parameter Learning from Incomplete Data Using Maximum Entropy I: Principles. ISBN 1 901615 37 5	Robert G. Cowell
21.	November 1999	Parameter Learning from Incomplete Data Using Maximum Entropy II: Application to Bayesian Networks. ISBN 1 901615 40 5	Robert G. Cowell
22.	March 2001	FINEX : Forensic Identification by Network Expert Systems. ISBN 1 901615 60X	Robert G.Cowell
23.	March 2001.	Wren Learning Bayesian Networks from Data, using Conditional Independence Tests is Equivalant to a Scoring Metric ISBN 1 901615 61 8	Robert G Cowell
24.	August 2004	Automatic, Computer Aided Geometric Design of Free-Knot, Regression Splines. ISBN 1-901615-81-2	Vladimir K Kaishev, Dimitrina S.Dimitrova, Steven Haberman Richard J. Verrall
25.	December 2004	Identification and Separation of DNA Mixtures Using Peak Area Information. ISBN 1-901615-82-0	R.G.Cowell S.L.Lauritzen J Mortera,
26.	November 2005.	The Quest for a Donor : Probability Based Methods Offer Help. ISBN 1-90161592-8	P.F.Mostad T. Egeland., R.G. Cowell V. Bosnes Ø. Braaten
27.	February 2006	Identification and Separation of DNA Mixtures Using Peak Area Information. (Updated Version of Research Report Number 25). ISBN 1-901615-94-4	R.G.Cowell S.L.Lauritzen J Mortera,

28.	October 2006	Geometrically Designed, Variable Knot Regression Splines : Asymptotics and Inference. ISBN 1-905752-02-4	Vladimir K Kaishev, Dimitrina S.Dimitrova, Steven Haberman Richard J. Verrall
29.	October 2006	Geometrically Designed, Variable Knot Regression Splines : Variation Diminishing Optimality of Knots. ISBN 1-905752-03-2	Vladimir K Kaishev, Dimitrina S.Dimitrova, Steven Haberman Richard J. Verrall

Papers can be downloaded from

http://www.cass.city.ac.uk/arc/actuarialreports.html

Faculty of Actuarial Science and Insurance

Actuarial Research Club

The support of the corporate members

- CGNU Assurance
- English Matthews Brockman
- Government Actuary's Department

is gratefully acknowledged.

Cass Business School
106 Bunhill Row,
London EC1Y 8TZ.

ISBN 1-905752-03-2