Kaishev, V. K., Dimitrova, D. S., Haberman, S. & Verrall, R. J. (2006). Geometrically designed, variable know regression splines: asymptotics and inference (Report No. Statistical Research Paper No. 28). London, UK: Faculty of Actuarial Science & Insurance, City University London.

CITY UNIVERSITY LONDON

City Research Online

Cass Business School
City of London

Faculty of Actuarial Science and Insurance

Geometrically Designed, Variable Knot Regression Splines: Asymptotics and Inference.

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

Statistical Research Paper No. 28

October 2006

ISBN 1-905752-02-4

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: Asymptotics and inference

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 least squares (LS) splines with variable knots, named GeDS, is presented. It is based on the property that the spline regression function, viewed as a parametric curve, has a control polygon and, due to the shape preserving and convex hull properties, closely follows the shape of this control polygon. The latter has vertices, whose x-coordinates are certain knot averages, known as the Greville sites and whose y-coordinates are the regression coefficients. Thus, manipulation of the position of the control polygon and hence of the spline curve may be interpreted as estimation of its knots and coefficients. These geometric ideas are implemented in the two stages of the GeDS estimation method. 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. To implement stage A, an automatic adaptive knot location scheme for generating linear spline fits is developed. At each step of stage A, a knot is placed where a certain bias dominated measure is maximal. This stage is equipped with a novel stopping rule which serves as a model selector. The optimal knots defined in stage B ensure that the higher order spline curve is nearly a variation diminishing (i.e., shape preserving) spline approximation to the linear fit of stage A. Error bounds for this approximation are derived in Kaishev et al. (2006). The GeDS method produces simultaneously linear, quadratic, cubic (and possibly higher order) spline fits with one and the same number of B-spline regression functions.

Large sample properties of the GeDS estimator are also explored, and asymptotic normality is established. Asymptotic conditions on the rate of growth of the knots with the increase of the sample size, which ensure that the bias is of negligible magnitude compared to the variance of the GeD estimator, are given. Based on these results, pointwise asymptotic confidence intervals with GeDS are also constructed and shown to converge to the nominal coverage probability level for a reasonable number of knots and sample sizes.

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

**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 a response variable y and an independent variable x, taking values within an interval $[a, b]$ and assume there is a relationship between x and y of the form

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

where $f(\cdot)$ is an unknown function and ϵ is a random error variable with zero mean and variance $E\,\epsilon^2 = \sigma^2 > 0$. We will consider the regression problem of estimating $f(\cdot)$, based on a sample of observations $\{x_i, y_i\}_{i=1}^N$. The design points $\{x_i\}_{i=1}^N$ may be either deterministic or random.

Different methods for the solution of this regression problem have been proposed and the related literature is extensive. One popular approach is to approximate f with an n-th order (degree $n-1$) spline function defined on $[a, b]$. As is well known, n-th order splines, on a set of k internal knots, form a linear functional space, an element of which is represented as a linear combination of appropriate spline basis functions. Thus, a spline function is defined by its order n, by the number and location of its k internal knots and by the coefficients in front of the basis functions.

It is also well known that least squares fitting with splines of a fixed degree is a linear optimization problem, if the number of knots and their location are fixed. However, since the latter are in general unknown and need to be defined, several approaches to constructing free-knot regression splines have been developed. The direct approach is to assume that n and k are fixed (but unknown), and to find the knot locations which minimize the (non-linear) least squares criterion (see e.g. Jupp, 1978) or an appropriately penalized version of it (see Lindstrom, 1999). For an extensive discussion of the (dis)advantages of non-linear free-knot spline estimation, we refer to Lindstrom (1999).

In order to circumvent the difficulties related to the non-linear optimization approach, a number of authors have developed adaptive knot selection procedures, such as step-wise knot inclusion/deletion strategies. Among the latter are the early work of Smith (1982), the TURBO spline modelling technique of Friedman and Silverman (1989), the MARS method proposed by Friedman (1991), the POLYMARS of Stone et al. (1997) and the spatially adaptive regression splines (SARS) of Zhou and Shen (2001). Other methods, such as the knot removal algorithm of Lytch and Mørken (1993) and the minimum description length (MDL) regression splines of Lee (2000), have been proposed as well. Multivariate spline regression and knot location has also been considered by Kaishev (1984). Further references to alternative spline fitting methods, such as smoothing spline techniques are to be found in Kaishev et al. (2006).

Asymptotic properties of least squares spline regression estimators have been considered by Agarwal and Studden (1980), Zhou et al. (1998), and more recently by Huang (2003) and by Wang and Yang (2006). Assuming $f \in C^q[0, 1]$, Agarwal and Studden (1980) give expressions in terms of the derivative $f^{(q)}(x)$, for the design density, the

knot placement density and the number of knots which minimize the asymptotic integrated mean square error. More recently, Zhou et al. (1998) and Huang (2003), have studied local (pointwise) asymptotic properties of least squares regression splines. Under the assumption of asymptotic uniformity of the knot placement, Zhou et al. (1998) provide explicit expressions for the asymptotic pointwise bias and variance of regression splines. A less stringent assumption on the knot mesh has been considered by Huang (2003), who establishes some asymptotic results for general estimation spaces. These asymptotic results shed some light on the large sample properties of least squares splines under some conditions on the joint asymptotic behaviour of the number and position of the knots and the sample size.

In this paper, we present a new variable knot spline regression estimation method which is very different from the existing methods and includes two stages. In stage A, a least squares linear spline regression fit to the data is constructed, following a novel knot location method. The latter places knots sequentially, one at a time, at sites where a certain bias dominated measure is maximal. Stage A is equipped with an appropriate stopping rule which serves as a model selector (see Section 3, and Appendix A for the complete description of stage A). In stage B, an optimal set of knots of a smoother, higher order ($n > 2$) least squares spline approximation is found so that the latter has also the characteristics of a Schoenberg's variation diminishing spline approximation of the linear spline fit from stage A. We show that this new spline regression estimation method has a direct Geometric Design interpretation. It stems from the fact that Schoenberg's variation diminishing spline approximation scheme is the fundamental concept underlying parametric B-spline curve and surface modelling in Computer Aided Geometric Design (see e.g. Farin, 2002). For this reason we will call our new method the GeD spline estimation method and will refer to the related estimator as a GeD spline estimator or simply GeDS. Optimality properties of the knots of the GeD spline estimator are established in Kaishev et al. (2006) where further algorithmic details, related to GeDS, are also to be found. The numerical performance of GeDS, compared to other existing spline estimators is addressed more thoroughly in Kaishev et al. (2006). In the present paper, the focus of our attention is on introducing this new spline estimator and on its statistical properties. Under some mild conditions on the design points $\{x_i\}_{i=1}^{N}$, we establish its asymptotic normality and give conditions under which the bias term of the approximation error becomes asymptotically negligible compared to the variance term. The construction of pointwise asymptotic confidence intervals is also considered and illustrated numerically.

The paper is organized as follows. In Section 2, it is shown that the spline regression function can be interpreted as a special case of a parametric spline curve, with a (control) polygon closely related to it. This geometric relationship is based on the established convex hull and variation diminishing properties of the spline regression curve. Since the vertices of its control polygon are defined in terms of its regression coeffi-

cients and knots, it is proposed that their estimation is performed through positioning the control polygon of the spline regression curve so that it follows the noise-perturbed shape of the underlying function. This geometric characterization of the regression problem is used in Section 3 to develop the GeD spline regression estimation method and in particular, to formulate its two stages, A and B, as optimization problems. In Section 4, some pointwise asymptotic properties of the GeDS estimator are established. The GeDS method and its properties are illustrated numerically in Section 5, based on a simulated example. It is shown how the large sample results of Section 4 can be used to construct asymptotic pointwise confidence intervals with a required coverage probability. In Section 6 we provide some discussion and conclusions. Detailed description of part A of GeDS are given in Appendix A and proofs of the results of Section 4 are given in Appendix B.

2. The B-spline regression and its control polygon.

Denote by $S_{\boldsymbol{t}_{k,n}}$ the linear space of all n-th order spline functions defined on a set of non-decreasing knots $\boldsymbol{t}_{k,n} = \{t_i\}_{i=1}^{2n+k}$, where $t_n = a$, $t_{n+k+1} = b$. In this paper we will use splines with simple knots, except for the n left and right most knots which will be assumed coalescent, i.e., $\boldsymbol{t}_{k,n} = \{t_1 = \ ... = t_n < t_{n+1} < ... < t_{n+k} < t_{n+k+1} = \ ... = t_{2n+k}\}$.

Following the Curry-Schoenberg theorem, a spline regression function $f \in S_{\boldsymbol{t}_{k,n}}$, can be expressed as

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

where $\boldsymbol{\theta} = (\theta_1, ..., \theta_p)'$ is a vector of real valued regression coefficients and $\boldsymbol{N}_n(x) = (N_{1,n}(x), \ ..., N_{p,n}(x))'$, $p = n + k$, are B-splines of order n, defined on $\boldsymbol{t}_{k,n}$. It is well known that $\sum_{i=j-n+1}^{j} N_{i,n}(t) = 1$, for any $t \in [t_j, t_{j+1})$, $j = n, ..., n + k$, and $N_{i,n}(t) = 0$ for $t \notin [t_i, t_{i+n}]$.

In the sequel, where necessary, we will emphasize the dependence of the spline $f(\boldsymbol{t}_{k,n}; x)$ on $\boldsymbol{\theta}$ by using the alternative notation $f(\boldsymbol{t}_{k,n}, \boldsymbol{\theta}; x)$.

The spline regression problem of Section 1 can now be more precisely stated as follows. For a fixed order of the spline n, given a sample of observations $\{y_i, x_i\}_{i=1}^{N}$, estimate the number of knots k, their locations $\boldsymbol{t}_{k,n}$ and the regression coefficients, $\boldsymbol{\theta}$.

In order to solve this estimation problem and develop the GeD spline estimator, our purpose in this section will be first to introduce an alternative way of expressing the spline regression $f(\boldsymbol{t}_{k,n}, \boldsymbol{\theta}; x)$. Recall that the standard way is to consider it as a function of the independent variable $x \in [a, b]$, following the expression $f(\boldsymbol{t}_{k,n}, \boldsymbol{\theta}; x) = \sum_{i=1}^{p} \theta_i N_{i,n}(x)$. Alternatively, $f(\boldsymbol{t}_{k,n}, \boldsymbol{\theta}; x)$ can be viewed as a special case of a parametric spline curve $\boldsymbol{Q}(t)$, $t \in [a, b]$. A parametric spline curve $\boldsymbol{Q}(t)$ is given coordinate-wise as

$$\boldsymbol{Q}(t) = \{x(t),\ y(t)\} = \{\textstyle\sum_{i=1}^{p} \xi_i\, N_{i,n}(t),\ \sum_{i=1}^{p} \theta_i\, N_{i,n}(t)\}\,, \tag{2}$$

where t is a parameter, and $x(t)$ and $y(t)$ are spline functions, defined on one and the same set of knots $\boldsymbol{t}_{k,n}$, with coefficients ξ_i and θ_i, $i = 1,\ ...,\ p$, respectively. If the coefficients ξ_i in (2) are chosen to be the knot averages

$$\xi_i^* = (t_{i+1} + ... + t_{i+n-1})/(n-1),\ i = 1,\ ...,\ p, \tag{3}$$

then it is possible to show that the identity

$$x(t) = \textstyle\sum_{i=1}^{p} \xi_i^*\, N_{i,n}(t) = t\ , \tag{4}$$

referred to as the linear precision property of B-splines, holds. In view of (2) and (4), the spline regression function $f(\boldsymbol{t}_{k,n}, \boldsymbol{\theta}; x)$ can be expressed as a parametric spline curve as

$$\boldsymbol{Q}^*(t) = \{t,\ f(t_{k,n}, \boldsymbol{\theta}; t)\} = \{\textstyle\sum_{i=1}^{p} \xi_i^*\, N_{i,n}(t),\ \sum_{i=1}^{p} \theta_i\, N_{i,n}(t)\}\ , \tag{5}$$

where $t \in [a, b]$ and ξ_i^* is the average of the $n-1$ consecutive knots $t_{i+1},\ ...,\ t_{i+n-1}$ given by (3). In what follows, it will be convenient to use $\boldsymbol{Q}^*(t)$ and $f(t_{k,n}, \boldsymbol{\theta}; t)$ interchangeably to denote a functional spline regression curve.

The values ξ_i^* given by (3) 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 $\boldsymbol{t}_{k,n}$.

The interpretation (5) of the regression function $f(\boldsymbol{t}_{k,n}, \boldsymbol{\theta}; x)$ as a parametric spline curve $\boldsymbol{Q}^*(t)$ is fundamental for our aim of developing a geometrically motivated least squares, variable knot, spline regression smoother. It allows us to characterize the spline regression curve $\boldsymbol{Q}^*(t)$ by a polygon, which is closely related to $\boldsymbol{Q}^*(t)$, and is called the control polygon of $\boldsymbol{Q}^*(t)$, denoted by $\boldsymbol{C}_{\boldsymbol{Q}^*}(t)$. It is constructed by connecting the points $\boldsymbol{c}_i = (\xi_i^*, \theta_i)$, $i = 1,\ ...,\ p$, called control points, by straight lines. So, $\boldsymbol{C}_{\boldsymbol{Q}^*}(\xi_i^*) = \theta_i$, $i = 1,\ ...,\ p$. In Fig.1, the geometric relationship between a spline regression curve and its control polygon $\boldsymbol{C}_{\boldsymbol{Q}^*}(t)$ is illustrated. This relationship is due to the fact that both the x and y coordinates of the control points $\boldsymbol{c}_i$, $i = 1,\ ...,\ p$, are related to the spline regression curve $\boldsymbol{Q}^*(t)$. More precisely, the x-coordinates, ξ_i^*, are the Greville sites (3), obtained from the knots $\boldsymbol{t}_{k,n}$, and the y-coordinates, θ_i, are simply the spline regression coefficients. Since, $\sum_{i=j-n+1}^{j} N_{i,n}(t) = 1$, for any $t \in [t_j, t_{j+1})$, $j = n,\ ...\ , n+k$, the curve $\boldsymbol{Q}^*(t)$ is a convex combination of its control points, and its graph lies within the convex hull of its control polygon $\boldsymbol{C}_{\boldsymbol{Q}^*}$. The convex hull of $\boldsymbol{c}_1,\ ...,\ \boldsymbol{c}_p$ is the smallest convex polygon, enclosing these points. Due to the convex hull property, the curve is in a close vicinity of its control polygon which is illustrated in Fig. 1 with respect to two adjacent polynomial segments of $\boldsymbol{Q}^*(t)$. The grey areas in Fig. 1 are the two convex hulls, formed by $\boldsymbol{c}_3, \boldsymbol{c}_4, \boldsymbol{c}_5$ and $\boldsymbol{c}_4, \boldsymbol{c}_5, \boldsymbol{c}_6$, within which the two segments of $\boldsymbol{Q}^*(t)$, for $t \in [t_5, t_6]$ and $t \in [t_6, t_7]$, lie. For further details related to geometric modelling with splines we refer to Cohen et al. (2001) and the classic book by Farin (2002).

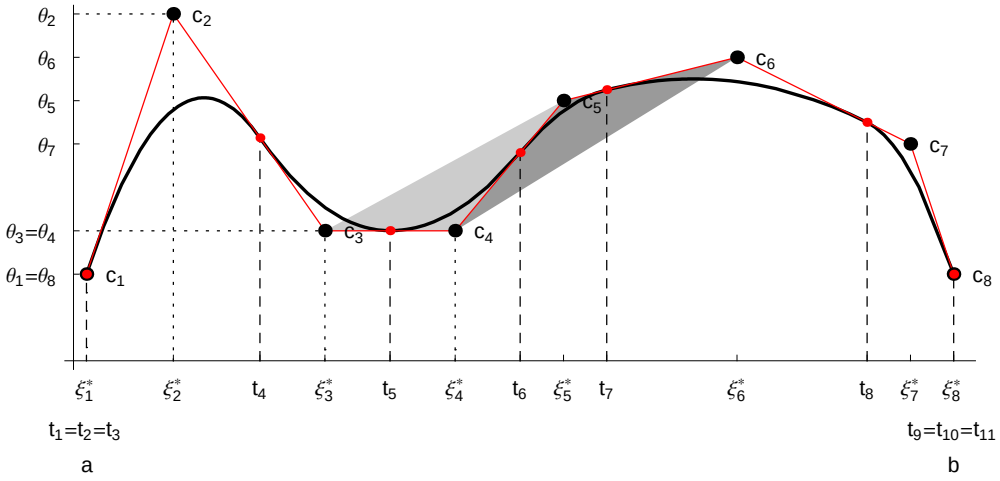

Fig. 1. A quadratic, functional spline regression curve $\boldsymbol{Q}^*(t)$ and its control polygon $\boldsymbol{C}_{\boldsymbol{Q}^*}$.

Another reason for the spline regression curve $\boldsymbol{Q}^*(t)$ to be close to its control polygon $\boldsymbol{C}_{\boldsymbol{Q}^*}(t)$ is that $\boldsymbol{Q}^*(t)$ is the Schoenberg's variation diminishing spline approximation of $\boldsymbol{C}_{\boldsymbol{Q}^*}(t)$, i.e.,

$$V[\boldsymbol{C}_{\boldsymbol{Q}^*}](t) = \sum_{i=1}^{p} \boldsymbol{C}_{\boldsymbol{Q}^*}(\xi_i^*)\, N_{i,n}(t) = \sum_{i=1}^{p} \theta_i\, N_{i,n}(t) \equiv \boldsymbol{Q}^*(t), \qquad (6)$$

where ξ_i^*, $i = 1, 2, \ldots, p$, are the Greville abscissae, obtained from $\boldsymbol{t}_{k,n}$ and $\boldsymbol{C}_{\boldsymbol{Q}^*}(\xi_i^*) = \theta_i$ by the definition of the control polygon.

Given a set of knots, $\boldsymbol{t}_{k,n}$, the spline approximation $V[g](x) = \sum_{i=1}^{p} g(\xi_i^*)\, N_{i,n}(x)$ of any function g, defined on $[a, b]$, is known as the Schoenberg's variation diminishing spline (VDS) 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 (3) and taking the values $g(\xi_i^*)$ as the B-spline coefficients of the VDS approximation.

It is important to recall a property of $V[g]$, which is crucial for developing the GeD estimator. That is, the VDS approximation, $V[g]$ 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. The *variation diminishing* character of $V[g]$ is due to the fact that it crosses any straight line at most as many times as does the function g itself. In view of the convex hull property and the shape preserving property of (6) it is more clear why $\boldsymbol{Q}^*(t)$ lies so close to its control polygon $\boldsymbol{C}_{\boldsymbol{Q}^*}(t)$.

In summary, it has been established that the spline regression function $f(\boldsymbol{t}_{k,n}, \boldsymbol{\theta}; x)$, (which we alternatively denoted as $\boldsymbol{Q}^*(x)$, $x \in [a, b]$), can be expressed in the form (5) and that its control polygon, $\boldsymbol{C}_{f(\boldsymbol{t}_{k,n},\boldsymbol{\theta};x)}$, has vertices $\boldsymbol{c}_i = (\xi_i^*, \theta_i)$, $i = 1, \ldots p$, where ξ_i^* are the Greville sites (3), obtained from $\boldsymbol{t}_{k,n}$. The latter suggests that, given n and k, locating the knots $\boldsymbol{t}_{k,n}$ and finding the regression coefficients $\boldsymbol{\theta}$ of $f(\boldsymbol{t}_{k,n}, \boldsymbol{\theta}; x)$, based on the set of observations $\{y_i, x_i\}_{i=1}^{N}$, is equivalent to finding the location of the x- and y-coordinates of the vertices of $\boldsymbol{C}_{f(\boldsymbol{t}_{k,n},\boldsymbol{\theta};x)}$. This establishes the important fact that estimation of $\boldsymbol{t}_{k,n}$ and

$\boldsymbol{\theta}$ affects the geometrical position of the control polygon $\boldsymbol{C}_{f(\boldsymbol{t}_{k,n},\boldsymbol{\theta};x)}$, which, due to the shape preserving and convex hull properties, defines the location of the spline curve $f(\boldsymbol{t}_{k,n}, \boldsymbol{\theta}; x)$. Inversely, locating the vertices $\boldsymbol{c}_i$ of $\boldsymbol{C}_{f(\boldsymbol{t}_{k,n},\boldsymbol{\theta};x)}$ affects the knots $\boldsymbol{t}_{k,n}$, through (3), and the values of $\boldsymbol{\theta}$, and hence affects the position of the regression curve $f(\boldsymbol{t}_{k,n}, \boldsymbol{\theta}; x)$. The latter conclusion motivates 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 an initial estimate of $\boldsymbol{\theta}$, which is improved further in stage B (see Section 3). This is the basis of our approach to constructing the GeD variable knot spline approximation to the unknown function f in (1), and this is developed in the next section.

3. Geometrically designed spline regression.

In this section we introduce the GeD spline regression method which is motivated by the ideas, outlined in Section 2. The method "positions" first an initial control polygon, which reproduces the "shape" of the data, applying least squares approximation. Secondly, an optimal set of knots of a higher order ($n > 2$) smooth spline curve is found, so that it preserves the shape of the initial control polygon and then this curve is fitted to the data, to adjust its position in the LS sense. In this way, it is ensured that the n-th order smooth LS fit follows the shape of the initial control polygon, and hence the shape of the data. This procedure simultaneously produces quadratic, cubic, or higher order splines and the LS fit with the minimum residual sum of squares is chosen as the final fit which recovers best the underlying unknown function f. The two stages of this approach may be given a formal interpretation as certain optimization problems with respect to the variables k, $\boldsymbol{t}_{k,n}$, $\boldsymbol{\theta}$ and n. Hence, the approach produces a solution which does not necessarily coincide with the globally optimal solution under the free-knot non-linear optimization approach. As illustrated by the numerical examples presented here and also in Kaishev et al. (2006), it produces LS spline fits which are characterized by a small number of non-coalescent knots and very low mean square errors. Thus, GeD spline fits are shown to be nearly optimal (see the example in Section 5 and also examples 1 and 2 of Kaishev et al., 2006) and to enjoy some very good large sample properties, such as asymptotic normality, established in Section 4. The latter allow for the construction of asymptotic confidence intervals illustrated in Section 5. The GeD spline estimation involves the following two stages:

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

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

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 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 Appendix A.

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

$$\min_{\substack{\boldsymbol{t}_{l-(n-2),n},\\ \xi_{i+1}<t_{i+n}<\xi_{i+n-1},\\ i=1,\ldots,k}} \left\| \hat{f}\,(\boldsymbol{\delta}_{l,2},\, \hat{\boldsymbol{\alpha}};\, x) - \boldsymbol{C}_{f(\boldsymbol{t}_{l-(n-2),n},\hat{\boldsymbol{\alpha}};x)} \right\|_{\infty}, \tag{7}$$

where $\|g\|_{\infty} := \max_{a\leq x\leq b} |\, g(x)\,|$ defines the uniform (L_{∞}) norm of a given 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 (7) is over all polygons $\boldsymbol{C}_{f(\boldsymbol{t}_{l-(n-2),n},\hat{\boldsymbol{\alpha}};x)}$ with vertices $(\xi_i^*,\, \hat{\alpha}_i)$, whose x-coordinates coincide with the Greville sites $\boldsymbol{\xi}^*(\boldsymbol{t}_{l-(n-2),n})$, and whose y-coordinates, coincide with the y-coordinates $\hat{\alpha}_i$ of the vertices of the polygon $\hat{f}\,(\boldsymbol{\delta}_{l,2},\, \hat{\boldsymbol{\alpha}};\, x)$. Clearly, 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 $l - (n - 2)$.

As shown in Kaishev et al. (2006), the optimization problem (7) has no optimal solution such that the minimum in (7) is zero, i.e., for which $\boldsymbol{C}_{f(\tilde{\boldsymbol{t}}_{l-(n-2),n},\hat{\boldsymbol{\alpha}};x)} \equiv \hat{f}\,(\boldsymbol{\delta}_{l,2},\, \hat{\boldsymbol{\alpha}};\, x)$. Instead, the objective of stage B, (i.e. of the minimization in (7)) is to produce a set of optimal knots $\tilde{\boldsymbol{t}}_{l-(n-2),n}$, which ensures that $\hat{f}\,(\boldsymbol{\delta}_{l,2},\, \hat{\boldsymbol{\alpha}};\, x)$ becomes (nearly) the control polygon of the spline regression function $f(\tilde{\boldsymbol{t}}_{l-(n-2),n},\, \hat{\boldsymbol{\alpha}};\, x)$, i.e., that $\boldsymbol{C}_{f(\tilde{\boldsymbol{t}}_{l-(n-2),n},\hat{\boldsymbol{\alpha}};x)} \simeq \hat{f}\,(\boldsymbol{\delta}_{l,2},\, \hat{\boldsymbol{\alpha}};\, x)$. In this way, $\tilde{\boldsymbol{t}}_{l-(n-2),n}$ is placed so that $f(\tilde{\boldsymbol{t}}_{l-(n-2),n},\, \hat{\boldsymbol{\alpha}};\, x)$ becomes (nearly) the Schoenberg's variation diminishing spline approximation of $\hat{f}\,(\boldsymbol{\delta}_{l,2},\, \hat{\boldsymbol{\alpha}};\, x)$ and hence, due to its convex hull and shape preserving properties (see Section 2), $f(\tilde{\boldsymbol{t}}_{l-(n-2),n},\, \hat{\boldsymbol{\alpha}};\, x)$ lies very close to $\hat{f}\,(\boldsymbol{\delta}_{l,2},\, \hat{\boldsymbol{\alpha}};\, x)$, and hence to the "shape" of the data for which the linear LS approximation is $\hat{f}\,(\boldsymbol{\delta}_{l,2},\, \hat{\boldsymbol{\alpha}};\, x)$ (according to stage A). This is the fundamental concept of optimal knot placement in GeDS. For a proof of the fact that $f(\tilde{\boldsymbol{t}}_{l-(n-2),n},\, \hat{\boldsymbol{\alpha}};\, x)$ is nearly a VDS approximation of $\hat{f}\,(\boldsymbol{\delta}_{l,2},\, \hat{\boldsymbol{\alpha}};\, x)$ with appropriate error bounds, we refer to Kaishev et al. (2006) (see Theorem 1 and Corollaries 1.1 and 1.2).

However, we note that the fit $f(\tilde{\boldsymbol{t}}_{l-(n-2),n},\, \hat{\boldsymbol{\alpha}};\, x)$ will not be a least squares approximation to the data set. In order to obtain an LS fit to the data and at the same time to preserve the shape of $f(\tilde{\boldsymbol{t}}_{l-(n-2),n},\, \hat{\boldsymbol{\alpha}};\, x)$, its optimal knots $\tilde{\boldsymbol{t}}_{l-(n-2),n}$ are preserved, whereas its B-spline coefficients $\hat{\alpha}_i$ are released and are assumed to be unknown parameters, $\boldsymbol{\theta}$, which are estimated in the least squares sense, using $\{y_i,\, x_i\}_{i=1}^{N}$. Thus, for a fixed $n = 3,\, ...,\, n_{\max}$, we find the least squares fit $\hat{f}\big(\tilde{\boldsymbol{t}}_{l-(n-2),n},\, \hat{\boldsymbol{\theta}};\, x\big)$ which solves

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

Finally, we choose 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. In this way, along with the number of knots and their locations, the degree of the spline is also estimated. This is an important feature of the proposed estimation method which is rarely offered by other spline estimation procedures. One alternative that we are aware of is the MDL method of Lee (2000). Of course, any of the produced final fits of order $n \neq \tilde{n}$ could be used, if other features were more desirable, for example if better smoothness were required.

Since (7) is a constrained non-linear optimization problem, and although for linear splines, our experience shows that it is still difficult to solve. As with other nonlinear optimization problems, finding the global optimum 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. To avoid these complications, the following simple knot placement method, called the *averaging knot location method*, is shown in Kaishev et al. (2006) to produce an approximation, $\bar{\boldsymbol{t}}_{l-(n-2),n}$, given by (8), to the optimal solution $\tilde{\boldsymbol{t}}_{l-(n-2),n}$ of (7). Bounds of this approximation are also established in Kaishev et al. (2006).

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 $n = 3, \ldots, n_{\max}$, select 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.

$$\bar{t}_{i+n} = (\delta_{i+2} + \ldots + \delta_{i+n})/(n-1), \qquad i = 1, \ldots, l-(n-2). \tag{8}$$

The choice of the knots $\bar{\boldsymbol{t}}_{l-(n-2),n}$ according to (8) leads to an improvement in the bounds established by Theorem 1 in Kaishev et al. (2006), which hold for $\tilde{\boldsymbol{t}}_{l-(n-2),n}$. The improved bounds for the set of knots $\bar{\boldsymbol{t}}_{l-(n-2),n}$ are given by Theorem 2 and its Corollaries 2.1 and 2.2 in Kaishev et al. (2006).

In summary, using the averaging knot location method (8), the GeDS estimation method simultaneously produces LS spline fits $\hat{f}\left(\bar{\boldsymbol{t}}_{l-(n-2),n}, \hat{\boldsymbol{\theta}}; x\right)$ of order $n = 2, 3, \ldots$ with the same number of basis functions $p = l-(n-2)+n = l+2$. Hence, the spline estimation spaces $S_{\bar{\boldsymbol{t}}_{l-(n-2),n}}$, $n = 2, 3, \ldots$ are of one and the same dimension. Note that from (8) when $n = 2$, $\bar{\boldsymbol{t}}_{l,2} \equiv \boldsymbol{\delta}_{l,2}$ and therefore, $\hat{\boldsymbol{\theta}} \equiv \hat{\boldsymbol{\alpha}}$ and $\hat{f}\left(\bar{\boldsymbol{t}}_{l,2}, \hat{\boldsymbol{\theta}}; x\right) \equiv \hat{f}(\boldsymbol{\delta}_{l,2}, \hat{\boldsymbol{\alpha}}; x)$. For further approximation theoretic results and related algorithmic details, we refer to Kaishev et al. (2006).

4. Local asymptotic properties of the GeD spline estimator.

The purpose of this section is to explore the local asymptotic properties of the proposed GeD spline estimation method and provide some large sample statistical inference.

Local asymptotic properties of least squares spline regression estimators are useful in constructing asymptotic confidence intervals and have been considered by Zhou et al. (1998), and more recently by Huang (2003). To investigate the pointwise asymptotic behaviour of the GeDS estimation error $\hat{f}\big(\bar{\boldsymbol{t}}_{l-(n-2),n}, \hat{\boldsymbol{\theta}}; x\big) - f(x)$ we will consider its decomposition

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

where the first and the second terms on the right-hand side are correspondingly referred to as the variance and the bias terms. As was noted in Section 1, the design points $\{x_i\}_{i=1}^N$ can either be deterministic or random. Without loss of generality, we will consider here the case of random design points under which $\{x_i, y_i\}_{i=1}^N$ is a random sample from the joint distribution (X, Y) of the predictor variable X and the response variable Y. It will be convenient to use the notation $\vec{\boldsymbol{x}} = (x_1, \ldots, x_N)$. In addition, we assume that the errors are homoscedastic, so that $\sigma_\epsilon^2(x) = E(\epsilon^2 \mid X = x) = \sigma^2$ is a constant. The results easily carry over to the heteroscedastic errors and fixed design case.

Thus, in our asymptotic analysis, 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 $\{x_j\}_{j=1}^{N_i}$ (see Assumption 1), we show that the GeDS estimation method produces estimates of the knots $\bar{\boldsymbol{t}}_{l_i-(n-2),n}$, $n \geq 2$, whose global mesh ratios form a sequence bounded in probability (see Lemmas 2 and 3). 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 $S_{\boldsymbol{t}_{k,n}}$, we will establish two asymptotic properties of the GeDS estimator. Thus, Theorems 1 and 2 below give a bound for the bias term and a sufficient condition for it to be of negligible magnitude compared to the variance term.

We also study in this section the asymptotic distribution of the GeD spline estimator $\hat{f}\big(\bar{\boldsymbol{t}}_{l-(n-2),n}, \hat{\boldsymbol{\theta}}; x\big)$. After its appropriate standardization, $\hat{f}\big(\bar{\boldsymbol{t}}_{l-(n-2),n}, \hat{\boldsymbol{\theta}}; x\big)$ is shown (see Theorem 3) 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 asymptotic confidence intervals, illustrated in Section 5. Proofs of the results of this section are given in Appendix B.

In what follows, we will rely on the sufficient asymptotic conditions for the least squares spline estimate to be well defined, established by Huang (2003). As is well known, the least squares estimate is an orthogonal projection relative to an appropriate inner product. The latter can be defined relative to a finite sample as $\langle f_1, f_2\rangle_N = \frac{1}{N}\sum_{i=1}^N f_1(x_i)\, f_2(x_i)$ and its theoretical version is given as $\langle f_1, f_2\rangle = \int_a^b f_1(x)\, f_2(x)\, p(x)\, dx$, for any square integrable functions f_1 and f_2 on $[a, b]$. Denote by $\|f_1\|_N = \langle f_1, f_1\rangle_N^{1/2}$ and $\|f_1\| = \langle f_1, f_1\rangle^{1/2}$. It will be required that as the sample size N goes to infinity, $\langle f_1, f_2\rangle_N$ converges to $\langle f_1, f_2\rangle$, (see Remark 1). In order to

investigate the asymptotic properties of the GeD spline estimator we will need the following assumption.

Assumption 1. Let ζ_i, $i = 1, 2, \ldots$ be a sequence of designs on $[a, b]$ with spectrum points $\zeta_i = \{a \le x_{i,1} < \ldots < x_{i,N_i} \le b\}$, $N_{i-1} < N_i$, at which the unknown function f is observed. Assume the spectrum points of ζ_i are randomly generated according to a density $0 < p(x) < \infty$, with respect to Lebesgue measure, such that the sequence of global mesh ratios

$$M_{\zeta_i}^{(1)} = \frac{\max_{1 \le j \le N_i - 1}(x_{i,j+1} - x_{i,j})}{\min_{1 \le j \le N_i - 1}(x_{i,j+1} - x_{i,j})}$$

is bounded in probability, i.e., $M_{\zeta_i}^{(1)} = O_P(1)$.

Note that this assumption requires the design points to be asymptotically quasi-uniformly distributed. Our asymptotic setting is such that for each random sample $\{x_{i,j}, y_{i,j}\}_{j=1}^{N_i}$, the GeD spline regression estimation method produces a linear least squares spline fit $\hat{f}(\boldsymbol{\delta}_{l_i,2}, \hat{\boldsymbol{\alpha}}; x)$ with knots $\boldsymbol{\delta}_{l_i,2}$ and higher order fits $\hat{f}\left(\bar{\boldsymbol{t}}_{l_i-(n-2),n}, \hat{\boldsymbol{\theta}}; x\right)$, $n > 2$, with knots $\bar{\boldsymbol{t}}_{l_i-(n-2),n}$, where l_i is determined by the choice of the parameter α_{exit}^i for each i. Recall that the latter parameter controls the exit from GeDS by the stopping rule given in stage A and that the spline estimation spaces $S_{\bar{\boldsymbol{t}}_{l_i-(n-2),n}}$, $n = 2, 3, \ldots$ are of one and the same dimension, $p = l_i + 2$. Next, we give a result which relates the rate of growth of l_i to that of the sample size N_i, which is used in proving the main theorems of this section.

Lemma 1. Given a sequence of random samples $\{x_{i,j}, y_{i,j}\}_{j=1}^{N_i}$ from (X, Y), there exists a sequence of α_{exit}^i, such that, for a fixed $n \ge 2$, $\lim_{i\to\infty} l_i \,/\, N_i^{1/(2n+1)} = \infty$ and $\lim_{i\to\infty} l_i \log N_i / N_i = 0$.

It is clear that the sequence of α_{exit}^i from Lemma 1 could be different for different values of n. Unfortunately, one can not specify general conditions which will determine a required sequence α_{exit}^i since the latter depends on the variability of the unknown function f and the noise level σ. The following two lemmas establish other characteristics of the knot meshes generated in stage A of GeDS which are important for the asymptotic analysis.

Lemma 2. If Assumption 1 holds then, for any sequence α_{exit}^i, the sequence of global mesh ratios, $M_{\boldsymbol{\delta}_i}^{(r)}$,

$$M_{\boldsymbol{\delta}_i}^{(r)} = \frac{\max_{2 \le j \le l_i - r + 3}(\delta_{i,j+r} - \delta_{i,j})}{\min_{2 \le j \le l_i - r + 3}(\delta_{i,j+r} - \delta_{i,j})}, \quad r \ge 2$$

of the knot sets $\boldsymbol{\delta}_{l_i,2} = \{\delta_{i,1} = \delta_{i,2} < \ldots < \delta_{i,l_i+3} = \delta_{i,l_i+4}\}$, generated according to stage A of GeDS, is bounded in probability. In other words, there exists a constant $\gamma > 0$ such that, except on an event whose probability tends to zero as $N_i \to \infty$, $M_{\boldsymbol{\delta}_i}^{(r)} \le \gamma$.

Given the result of Lemma 1, we next show that under Assumption 1 the knot sequences of the higher order fit $\hat{f}\left(\bar{\boldsymbol{t}}_{l_i-(n-2),n}, \hat{\boldsymbol{\theta}}; x\right)$, $n \ge 2$ of Stage B also have bounded mesh ratios.

Lemma 3. If the sequence of global mesh ratios, $M_{\boldsymbol{\delta}_i}^{(r)}$, $r \geq 2$, of the knot sets $\boldsymbol{\delta}_{l_i,2}$, generated in stage A, is bounded in probability by a constant $\gamma > 0$, then the global mesh ratio, $M_{\bar{\boldsymbol{t}}_i}^{(r)}$, $r \geq n$, of the knot sequence $\bar{\boldsymbol{t}}_{l_i-(n-2),n}$, $n \geq 2$, generated in stage B, is also bounded by γ, i.e.,

$$M_{\bar{\boldsymbol{t}}_i}^{(r)} = \frac{\max_{n \leq j \leq l_i+1+n-r}(\bar{t}_{i,j+r} - \bar{t}_{i,j})}{\min_{n \leq j \leq l_i+1+n-r}(\bar{t}_{i,j+r} - \bar{t}_{i,j})} \leq \gamma, \quad r \geq n \tag{9}$$

except on an event whose probability tends to zero as $N_i \to \infty$.

Remark 1. Under Assumption 1, the assertions of Lemmas 1 and 3 imply that, for some appropriate α_{exit}^i, $\lim_{i\to\infty} l_i \log N_i / N_i = 0$, and that the global mesh ratio of the knots $\bar{\boldsymbol{t}}_{l_i-(n-2),n}$ is bounded. Since $\hat{f}\left(\bar{\boldsymbol{t}}_{l_i-(n-2),n}, \hat{\boldsymbol{\theta}}; x\right)$, $n \geq 2$, is an LS estimator, one can apply Lemma 2.3 of Huang (2003) to establish that the latter are sufficient conditions for the theoretical norm to be close to the empirical norm, uniformly over $S_{\bar{\boldsymbol{t}}_{l_i-(n-2),n}}$, i.e.,

$$\sup_s \left| \|s\|_{N_i} / \|s\| - 1 \right| = o_P(1), \tag{10}$$

where $s \in S_{\bar{\boldsymbol{t}}_{l_i-(n-2),n}}$. This is essential for our asymptotic analysis since it ensures that the problem of least squares GeD spline estimation is well defined.

We are now in a position to establish the asymptotic properties of the GeD spline estimator, which are used later in constructing asymptotic pointwise confidence intervals. We start with the following theorem.

Theorem 1. Under Assumption 1, there exist a sequence α_{exit}^i and an absolute constant C such that, except on an event whose probability tends to zero as $N_i \to \infty$,

$$\left\| E\left(\hat{f}\left(\bar{\boldsymbol{t}}_{l_i-(n-2),n}, \hat{\boldsymbol{\theta}}; x\right) \middle| \vec{\mathbf{x}} \right) - f(x) \right\|_\infty \leq C\, \rho_{N_i},$$

where $\rho_{N_i} = \inf \left\{ \|f - s\|_\infty : s \in S_{\bar{\boldsymbol{t}}_{l_i-(n-2),n}} \right\}$, $n \geq 2$.

The bound in Theorem 1 can be specified, imposing a certain smoothness condition on the unknown function f. Thus, if $f \in C^q[a, b]$ and $n \geq q$ then it can be shown (see e.g. Schumaker, 1981) that $\rho_{N_i} = O(l_i^{-q})$.

Next, we consider the asymptotic behaviour of the bias term, compared to the variance term in the GeDS estimation error decomposition. We state the following theorem.

Theorem 2. Under Assumption 1, if $f \in C^q[a, b]$, then there exists a sequence α_{exit}^i such that, for $n \geq q$,

$$\sup_{x\in[a,b]} \left| \frac{E\left(\hat{f}\left(\bar{\boldsymbol{t}}_{l_i-(n-2),n}, \hat{\boldsymbol{\theta}}; x\right) \middle| \vec{\mathbf{x}} \right) - f(x)}{\sqrt{\text{Var}\left(\hat{f}\left(\bar{\boldsymbol{t}}_{l_i-(n-2),n}, \hat{\boldsymbol{\theta}}; x\right) \middle| \vec{\mathbf{x}} \right)}} \right| = o_P(1).$$

The above theorem together with the following result facilitates the construction of pointwise confidence intervals for $f(x)$, using the proposed GeD spline estimator.

Theorem 3. Under Assumption 1, suppose $\lim_{\lambda\to\infty} E[\epsilon^2 \mathbb{1}_{\{|\epsilon|>\lambda\}} \mid X = x] = 0$. Then, there exists a sequence α^i_{exit} such that

$$P\Big[\hat{f}\big(\bar{\boldsymbol{t}}_{l_i-(n-2),n}, \hat{\boldsymbol{\theta}}; x\big) - E\Big(\hat{f}\big(\bar{\boldsymbol{t}}_{l_i-(n-2),n}, \hat{\boldsymbol{\theta}}; x\big) \,\Big|\, \vec{\boldsymbol{x}}\Big) \le t\sqrt{\operatorname{Var}\Big(\hat{f}\big(\bar{\boldsymbol{t}}_{l_i-(n-2),n}, \hat{\boldsymbol{\theta}}; x\big) \,\Big|\, \vec{\boldsymbol{x}}\Big)} \,\Big|\, \vec{\boldsymbol{x}}\Big] - \Phi(t)$$
$$= o_P(1)$$

for $x \in [a, b]$ and $t \in \mathbb{R}$. Hence,

$$\frac{\hat{f}\big(\bar{\boldsymbol{t}}_{l_i-(n-2),n}, \hat{\boldsymbol{\theta}}; x\big) - E\big(\hat{f}\big(\bar{\boldsymbol{t}}_{l_i-(n-2),n}, \hat{\boldsymbol{\theta}}; x\big) \mid \vec{\boldsymbol{x}}\big)}{\sqrt{\operatorname{Var}\big(\hat{f}\big(\bar{\boldsymbol{t}}_{l_i-(n-2),n}, \hat{\boldsymbol{\theta}}; x\big) \mid \vec{\boldsymbol{x}}\big)}} \xrightarrow{d} \mathcal{N}(0, 1), \quad N_i \underset{i\to\infty}{\longrightarrow} \infty.$$

Theorem 3 establishes asymptotic normality of the variance term in the error decomposition of the GeDS estimator and enables us to construct asymptotic pointwise confidence intervals for $E\big(\hat{f}\big(\bar{\boldsymbol{t}}_{l_i-(n-2),n}, \hat{\boldsymbol{\theta}}; x\big) \,\big|\, \vec{\boldsymbol{x}}\big)$, $n \ge 2$. Furthermore, combining Theorem 3 with Theorem 2 allows for the construction of asymptotically valid confidence intervals for the unknown function f.

As known from the standard regression theory, in the finite sample case,

$$\operatorname{Var}\big(\hat{f}\big(\bar{\boldsymbol{t}}_{l-(n-2),n}, \hat{\boldsymbol{\theta}}; x\big) \,\big|\, \vec{\boldsymbol{x}}\big) = \sigma^2 \boldsymbol{N}_n'(x) \{\langle \boldsymbol{F}'(\vec{\boldsymbol{x}}), \boldsymbol{F}(\vec{\boldsymbol{x}})\rangle_N\}^{-1} \boldsymbol{N}_n(x),$$

where the matrix $\boldsymbol{F}(\vec{\boldsymbol{x}}) = (\boldsymbol{N}_n(x_1), \ ..., \boldsymbol{N}_n(x_N))$. Thus, a $100(1-\alpha)\%$ confidence interval can be constructed as

$$\hat{f}\big(\bar{\boldsymbol{t}}_{l-(n-2),n}, \hat{\boldsymbol{\theta}}; x\big) \pm z_{1-\alpha/2} \sqrt{\operatorname{Var}\big(\hat{f}\big(\bar{\boldsymbol{t}}_{l-(n-2),n}, \hat{\boldsymbol{\theta}}; x\big) \,\big|\, \vec{\boldsymbol{x}}\big)}, \tag{11}$$

where $z_{1-\alpha/2} = \Phi^{-1}(1-\alpha/2)$, $n \ge 2$. Following Theorem 6.1 of Huang (2003), in view of Remark 1,

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

which means that (11) is an asymptotically valid confidence interval.

To conclude this section, recall that Theorem 2 gives the conditions on the order of the number of knots under which the bias term of the GeDS estimation error is of negligible magnitude compared to the variance term. The condition $\lim_{i\to\infty} l_i \,/\, N_i^{1/(2q+1)} = \infty$ implies that, in order not to consider the bias asymptotically in constructing a confidence interval, one needs to use higher number of knots than what is needed for achieving the optimal rate of convergence, $N_i^{-2q/(2q+1)}$. The latter is obtained by balancing the rate of convergence of the squared bias and the variance terms (see Stone, 1982). Therefore, as was noted by Zhou et al. (1998), the knots obtained by using the generalized cross validation (GCV) as a model selector lead to undersmoothed $\hat{f}(x)$ and can not be used for the construction of asymptotic confidence intervals for $f(x)$.

Using the proposed GeDS estimator, the optimal rate of convergence is achievable (see Section 6). However, in order for the conditions of Theorem 3 to be fulfilled and to have asymptotic normality, one should use values of α^i_{exit} higher than what is needed for

matching the optimal rate. The latter is possible and one has a considerable degree of flexibility since the condition $\lim_{i\to\infty} l_i \log N_i / N_i = 0$ on the rate of growth of the number of knots is much weaker than the condition $\lim_{i\to\infty} l_i^2 / N_i = 0$ used in Zhou et al. (1998). The construction of asymptotic confidence intervals and appropriate choices of α^i_{exit} are illustrated in the next section, where further comments are provided.

5. Simulation studies.

The GeDS method has been thoroughly tested numerically and compared with other spline methods and the results of this comparison are given in Kaishev et al. (2006). Examples include different values of signal-to-noise ratio, small and large sample sizes, x-values in a grid or uniformly generated within $[a, b]$. The overall conclusion is that GeDS has performed very well both in terms of efficiency and quality of the fit. Here we will illustrate briefly the GeDS method using the following test example, which appears in Schwetlick and Schütze (1995).

Table 1. Example used to test GeDS.

Test function	***Interval***	***Sample size, N***	x_i, $i = 1, \ldots, N$	***Noise level,*** σ_ϵ
$f(x) = \frac{10x}{1+100x^2}$	$[-2, 2]$	90	$x_i = -2 + \frac{(2-(-2))}{89}(i-1)$	$U(-0.05, 0.05)$

For a simulated data set, graphs of the linear spline fits, produced at each consecutive iteration in stage A of GeDS, preceding the final one, are given in Fig. 2. As can be seen, the initial straight line fit, presented in Fig. 2 (a), is sequentially improved by adding knots, one at each iteration, to reach the fit $\hat{f}(\boldsymbol{\delta}_{8,2}; x)$, plotted in Fig. 3 (a), which can not be further significantly improved by adding more knots. Applying the averaging knot location (8) to the knots $\boldsymbol{\delta}_{8,2}$ of the linear fit $\hat{f}(\boldsymbol{\delta}_{8,2}; x)$, the set of knots $\bar{\boldsymbol{t}}_{8-(n-2),n}$ of the quadratic, $n = 3$, and cubic, $n = 4$, fits, $\hat{f}(\bar{\boldsymbol{t}}_{8-(n-2),n}; x)$, are defined. The LS spline fits $\hat{f}\big(\bar{\boldsymbol{t}}_{8-(n-2),n}, \hat{\boldsymbol{\theta}}; x\big)$, resulting from stage B of GeDS, are plotted in Fig. 3 (b) and (c) for $n = 3$ and $n = 4$, respectively. The polygons $\boldsymbol{C}_{f(\tilde{\boldsymbol{t}}_{7,3},\hat{\boldsymbol{\alpha}};x)}$ and $\boldsymbol{C}_{f(\bar{\boldsymbol{t}}_{7,3},\hat{\boldsymbol{\alpha}};x)}$, plotted in Fig. 3 (d), using dot-dashed and dashed lines, and obtained with $\tilde{\boldsymbol{t}}_{7,3}$ as the solution of (7) and with $\bar{\boldsymbol{t}}_{7,3}$, calculated using (8), are seen to be very close to each other and also close to the initial control polygon $\hat{f}(\boldsymbol{\delta}_{8,2}, \hat{\boldsymbol{\alpha}}; x)$. The final LS fits, $\hat{f}(\tilde{\boldsymbol{t}}_{7,3}; x)$ and $\hat{f}(\bar{\boldsymbol{t}}_{7,3}; x)$, obtained with the optimal knots $\tilde{\boldsymbol{t}}_{7,3}$ and with the knots $\bar{\boldsymbol{t}}_{7,3}$, according to the averaging knot location method (8), have close L_2-errors, respectively 0.2798 and 0.2944, which confirms that $\bar{\boldsymbol{t}}_{7,3}$ approximates very well the optimal set of knots $\tilde{\boldsymbol{t}}_{7,3}$.

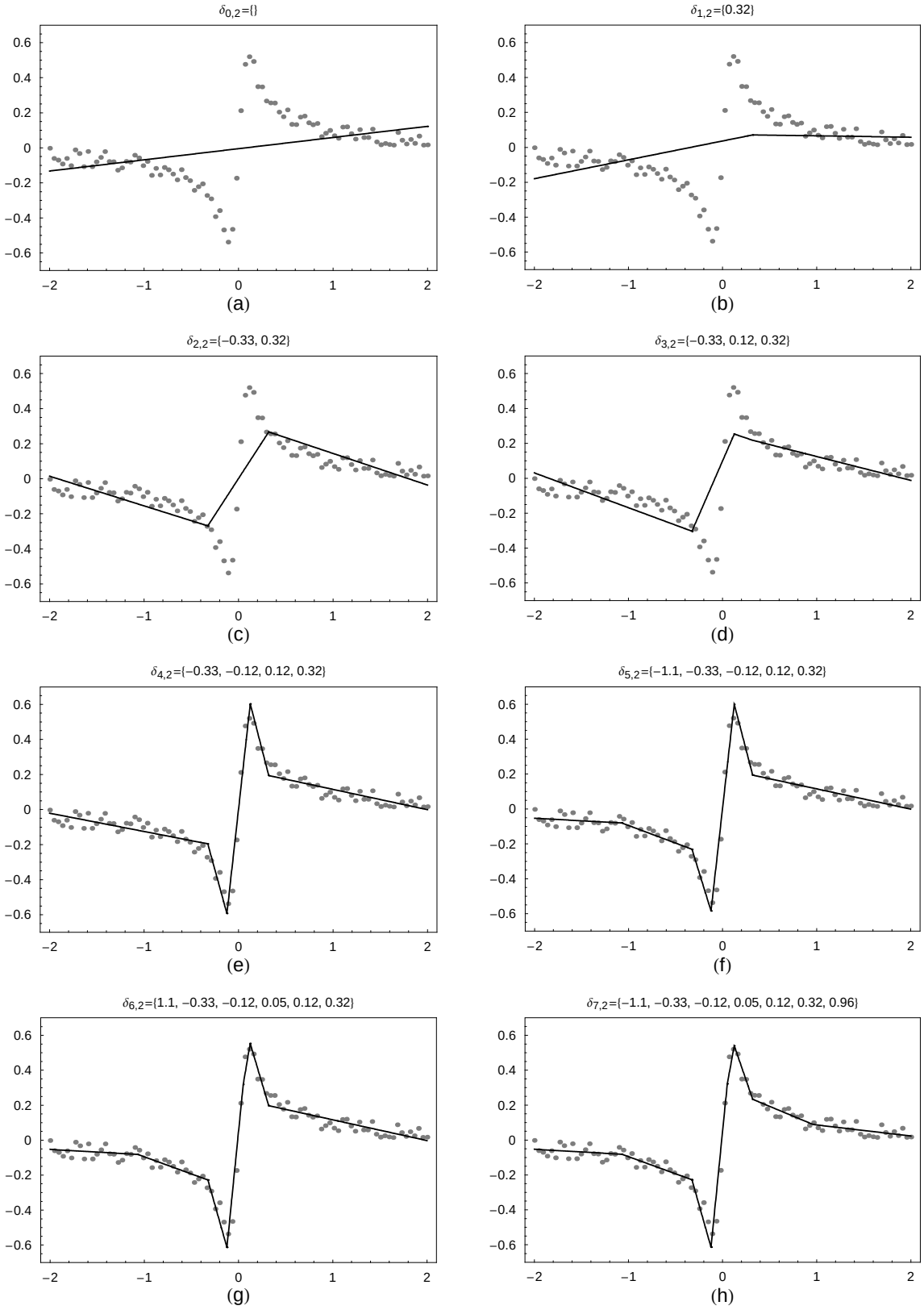

Fig. 2. The linear spline fits, obtained at each consecutive iteration in stage A, except the final one (given in Fig. 3 (a)).

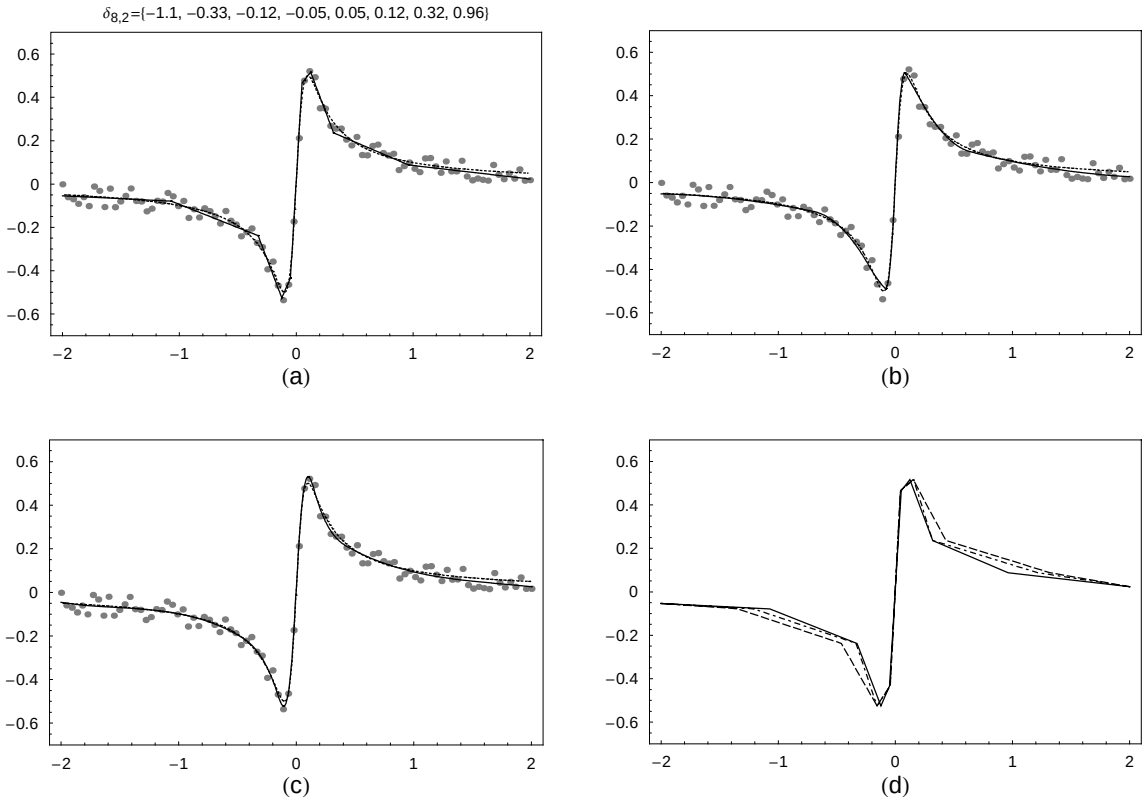

Fig. 3. The final GeD spline fits: (a) linear; (b) quadratic; (c) cubic; (d) graphs of $\hat{f}\,(\boldsymbol{\delta}_{8,2}, \hat{\boldsymbol{\alpha}}; x)$ - the solid line, $\boldsymbol{C}_{f(\tilde{\boldsymbol{t}}_{8-(n-2),n}, \hat{\boldsymbol{\alpha}}; x)}$ - the dot-dashed line and $\boldsymbol{C}_{f(\tilde{\boldsymbol{t}}_{8-(n-2),n}, \hat{\boldsymbol{\alpha}}; x)}$ - the dashed line; The dotted curve in (a), (b), (c) is the true function.

The details of the final linear fit, and its corresponding quadratic and cubic spline fits are presented in Table 2. The computation time for the three fits is less then a second (0.89 sec. on a PC, Pentium IV, 1.4 Ghz, 512 RAM). Note that the values for the parameters α_{exit} and β are the default preassigned values 0.9, 0.5 (see Kaishev et al. 2006 for more detailed comments on the effect of the choice of α_{exit} and β). The parameter β is defined and discussed in step 5 of Stage A (see Appendix A). As can be seen, the function f is symmetric and GeDS places, symmetrically around the origin, 8, 7 and 6 knots, respectively for the linear, quadratic and cubic LS fits. As can be seen from Table 2, all the fits are of a very good quality with respect to the MSE, defined as $\text{MSE} = \sum_{i=1}^{N} (y_i - \hat{f}(x_i))^2 \,/\, N$.

Table 2. Summary of GeD spline fits.

Fit No	***Graph***	***n***	***k***	***Internal knots***	α_{exit}, β	L_2 **–** ***error, MSE***
1	Fig. 3, (a)	2	8	{−1.1, −0.33, −0.12, −0.05, 0.05, 0.12, 0.32, 0.96}	0.9, 0.5	0.2699, 0.000189
2	Fig. 3, (b)	3	7	{−0.69, −0.22, −0.09, 0.00, 0.09, 0.22, 0.64}	0.9, 0.5	0.2944, 0.000127
3	Fig. 3, (c)	4	6	{−0.51, −0.17, −0.04, 0.04, 0.16, 0.47}	0.9, 0.5	0.2631, 0.000119

Based on the L_2-errors for the linear, quadratic and cubic fits given in Table 2, the best GeDS fit for this particular data set is the cubic one. We have compared it (No 3, Table 2) with the optimal cubic 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 3. As can be seen, the three fits are very close, comparing the L_2-errors, the MSE and the location of the knots. However, the GeD fit recovers best the original function as indicated by the corresponding MSE values. The computation time needed for GeDS is less then a second (0.89 sec.) whereas for PNOM and NOM it is respectively, 4.5 hours and 1.4 hours, using the *Mathematica* function `NMinimize`.

Table 3. The fits obtained by GeDS, PNOM and NOM.

Fit No	***Method***	***n***	***k***	***Internal knots***	***L_2 – error, MSE***
1	GeDS	4	6	{−0.51, −0.17, −0.04, 0.04, 0.16, 0.47}	0.2631, 0.000119
2	PNOM	4	6	{−0.53, −0.16, −0.06, 0.05, 0.17, 0.51}	0.2623, 0.000131
3	NOM	4	6	{−0.48, −0.15, −0.07, 0.05, 0.18, 0.40}	0.2614, 0.000154

We test GeDS by fitting 1000 simulated data sets from the function f, given in Table 1. A frequency plot of the number of internal knots of the 1000 linear GeD spline fits and box plots of the MSE values of the linear, quadratic and cubic GeDS fits are presented in Fig. 4 (a) and (b).

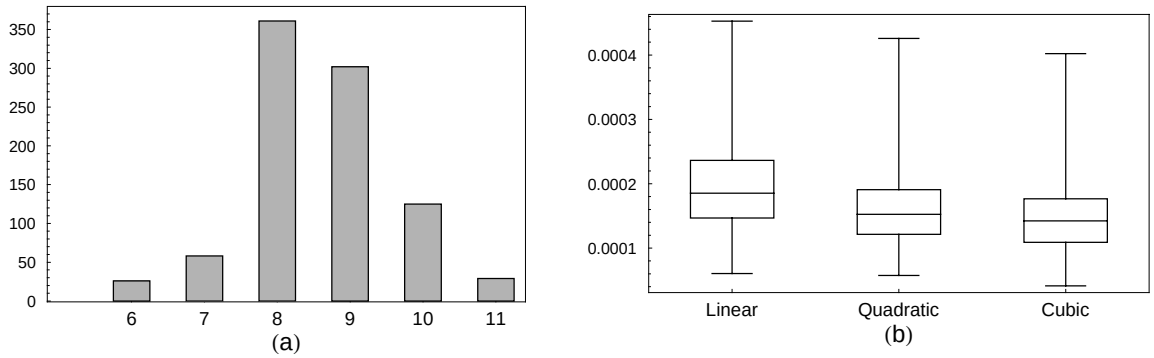

Fig. 4. (a): A frequency plot of the number of knots of the 1000 linear GeD spline fits; (b): Box plots of the MSE values of the 1000 linear, quadratic and cubic GeD spline fits.

The box plots presented in Fig. 4 (b) confirm that the best GeDS fit for this particular function is the cubic one. Since the number of internal knots, k, of a quadratic (cubic) GeDS fit is always one (two) less than that of the corresponding linear fit, the frequency plot for the 1000 quadratic (cubic) GeDS fits is identical to the one in Fig. 4 (a) but over the range $k = 5, 6, 7, 8, 9, 10$ ($k = 4, 5, 6, 7, 8, 9$). The 1000 linear, quadratic and cubic GeD spline fits, with median number of regression functions $n + k = 10$, have median L_2-errors 0.260, 0.267, 0.264 respectively, which are lower than 0.277, obtained by Schwetlick and Schütze (1995) for a quartic (of order 5) fit with the same number of regression parameters and optimally located knots. The optimal quartic fit of Schwetlick

and Schütze (1995) is obtained starting from 15 knots and after three time-consuming knot generation, removal and relocation stages.

Constructing confidence intervals

The second part of our simulation study is devoted to the implementation of the asymptotic results for the developed GeD spline estimator, given in Section 4. We illustrate the practical construction of finite-sample pointwise confidence intervals for the unknown function $f(x)$ and test their performance in achieving the required nominal coverage probability levels. For this purpose, we again use the test function given in Table 1 but with normally distributed error term, $\epsilon_i \sim \mathcal{N}(0, 0.015)$, and for equally spaced design points $\{x_j\}_{j=1}^{N_i}$, $i = 1, 2, 3$ with sample sizes $N_1 = 100$, $N_2 = 500$, $N_3 = 1000$. To assess the finite-sample performance of the constructed confidence intervals we evaluate their empirical coverage probabilities. The latter are calculated as the percentage of coverage of the true value $f(x)$ by the $100\,(1-\alpha)\%$ confidence interval defined in (11), based on 1000 replications for each sample size, N_i. Confidence intervals are obtained using both the true σ^2 ("oracle" value) and its estimate, $\hat{\sigma}^2$, proposed by Hall et al. (1990) but only the results for the oracle values are presented. This is because our simulated results show that the estimate $\hat{\sigma}^2$ exhibits positive bias for small samples (see also Zhou et al., 1998) and hence, unjustifiably increases the empirical coverage probability values.

In order to compute the GeD spline estimator $\hat{f}\big(\bar{\boldsymbol{t}}_{l-(n-2),n}, \hat{\boldsymbol{\theta}}; x\big)$ and its variance, needed in (11), we have selected the sequence $\alpha^1_{\text{exit}} = 0.9$, $\alpha^2_{\text{exit}} = 0.99$, $\alpha^3_{\text{exit}} = 0.999$ for the stopping rule (see step 10 of Stage A), which determines the number of knots l at exit of stage A. These values of α_{exit} have been chosen so that the requirement of Theorems 2 and 3 with respect to the rate of growth of l with the sample size are met. Thus, the median number of knots selected by GeDS for each α^i_{exit}, $i = 1, 2, 3$, is $l_1 = 10$, $l_2 = 16$, $l_3 = 25$ correspondingly.

In Fig. 5 we have plotted the empirical coverage probabilities for 95% level pointwise confidence intervals as a function of x. The empirical average coverage probability (EACP) over all $\{x_j\}_{j=1}^{N_i}$ are also presented under each of the plots.

As can be seen from Fig. 5, for all the linear, quadratic and cubic fits, the coverage probability converges to its nominal level which is reached for the cubic fit for the sample size of 1000 and median number of knots 25. The convergence for the linear and quadratic fits are slower, as expected, since the dimension of the spline spaces is the same for all the fits of different order. However, the coverage probability improves with the order of the spline for all sample sizes. It has to be noted that the test function is symmetrical with strong and sharp variation around the origin which makes it difficult to fit, based on noisy data. This is reflected in the spikes in the coverage probability observed in the neighborhood of the origin which almost disappear for the cubic fit with 1000 data points.

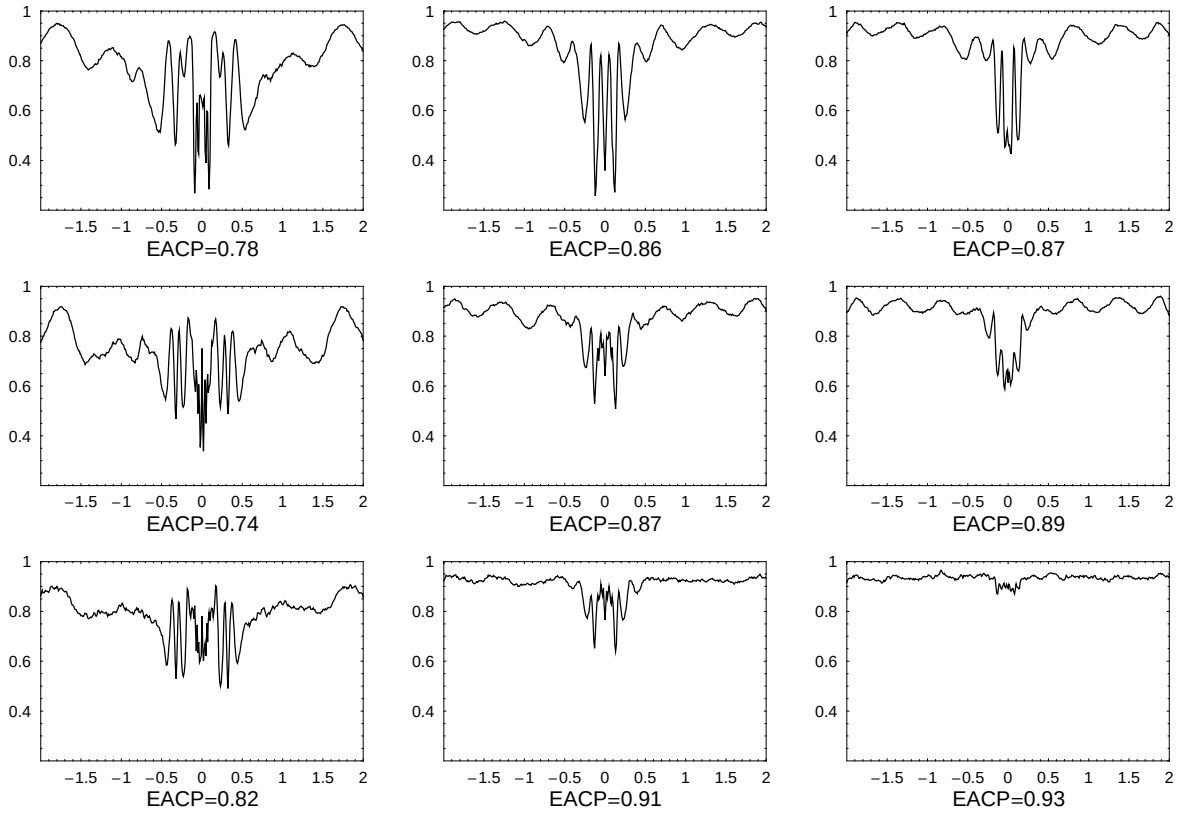

Fig. 5. Empirical coverage probabilities of 95% pointwise confidence intervals, obtained by GeDS: linear fit - first column; quadratic fit - second column; cubic fit - third column. Sample sizes: $N_1 = 100$ - first row; $N_2 = 500$ - second row; $N_3 = 1000$ - third row;

It is essential to mention here that, for such finite samples, it is important not only to assess the appropriate rate of growth of the number of knots but also to determine their absolute number and location. We believe that our number of knots is close to being minimal and because they are optimally located using GeDS, the 95% nominal level of the coverage probability is achieved already for sample size $N = 1000$. Whereas, for example, if we fit a cubic spline to the data, using the same number and rate of growth of the knots but placing them uniformly, the EACP is much worse, i.e., EACP = 0.20 for $N = 100$, EACP = 0.34 for $N = 500$, EACP = 0.63 for $N = 1000$.

Analyzing the rate of decrease of the bias term for the linear fit (not presented), compared to its rate of decrease for the cubic fit over the sample sizes (see Fig 6, first row), shows that this rate is much stronger for the cubic fit than in the linear case. This is not surprising since the test function is smooth but oscillates around the origin and, using a linear spline, it is difficult to achieve the quality of the approximation (improve the bias) obtained by the smoother cubic fit with same number of regression coefficients. At the same time the variance term decreases with the sample size at a similar rate in both cases. This explains the slight deterioration in the EACP (from 0.78 to 0.74) for the linear spline fit.

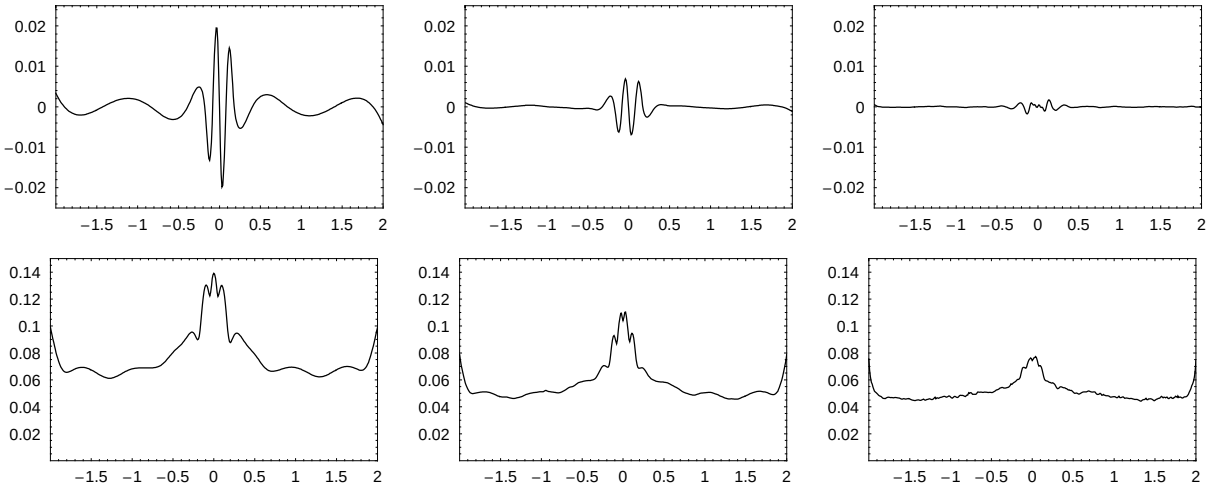

Fig. 6. Empirical bias (first row) and standard deviation (second row) of the cubic GeDS fits over 1000 replicates. Sample sizes: $N_1 = 100$ - first column; $N_2 = 500$ - second column; $N_3 = 1000$ - third column;

In order to illustrate the local adaptability of the GeDS estimators (see stage A in Appendix A and Theorems 1 and 2 of Kaishev et al., 2006), in Fig. 6 we have plotted the empirical bias and standard deviation of $\hat{f}\left(\bar{\boldsymbol{t}}_{l-(n-2),n}, \hat{\boldsymbol{\theta}}; x\right)$, $n = 4$, as a function of x for $N = 100$, $N = 500$, $N = 1000$. One can see that the bias term becomes negligible compared to the variance term, in fact for $N = 1000$ it is on average 300 times smaller, which corroborates the result of Theorem 2. As with the coverage probability, the bias and variance also exhibit rough behaviour around the origin which smooths out with the sample size. For brevity, we omit here the corresponding plots for the linear and quadratic fits.

6. Discussion.

One of the important characteristics of the GeDS estimation procedure is that it gives simultaneously linear, quadratic, cubic, etc. fits because once the LS linear spline fit in stage A is found, using the averaging knot location method (8), the knots for the higher order LS spline fits of stage B are immediately obtained. As far as we have been able to establish, no other spline fitting procedure is capable of doing this. Hence, one has the flexibility to choose the degree of the final fit providing best compromise between smoothness and accuracy.

As an alternative to the stopping rule, described in step 10 of stage A (see Appendix A), we have implemented two additional stopping criteria according to which our algorithm exits with number of knots which minimizes Stein's unbiased risk estimate (SURE) (see Stein, 1981)

$$R(\hat{f}) = \sum_{i=1}^{N} \left(y_i - \hat{f}\left(\bar{\boldsymbol{t}}_{k,n}, \hat{\boldsymbol{\theta}}; x_i\right)\right)^2 \Big/ N + D\,\tfrac{(k+n-1)}{N}\,\sigma^2 \qquad (12)$$

or the generalized cross validation (GCV) (see e.g., Craven and Wahba, 1979)

$$\mathrm{GCV}(\hat{f}) = \left(\frac{\sum_{i=1}^{N} \left(y_i - \hat{f}\left(\bar{t}_{k,n}, \hat{\boldsymbol{\theta}}; x_i\right)\right)^2}{N} \right) \Big/ \left(1 - \frac{d(k)}{N}\right)^2 \qquad (13)$$

criterion. We have assumed that the minimum is attained when SURE or GCV do not decrease in two consecutive iterations in stage A. Rules (12) and (13) depend on the choice of the parameters D and $d(k)$, and when $D = 2$ and $d(k) = k + 1$ they behave roughly as our stopping rule. The choice $D = 3$ and $d(k) = 3k + 1$, as noted by Zhou and Shen (2001) tends to yield a smaller model, underfitting the underlying function f. For a comparative study of different model selection methods, we refer to Lee (2002). Applying (12) or (13), GeDS becomes entirely automatic and can be applied if such a feature is preferred to the flexibility of controlling the output provided by our stopping rule.

We have addressed here pointwise large sample properties of this new spline estimator, such as the asymptotic behaviour of the variance and bias components of the estimation error and the construction of confidence intervals, based on the established asymptotic normality. The results of the simulation study corroborate well with the theoretical findings and support the strong practical appeal of the proposed GeD spline estimator. In conclusion, we believe that the proposed GeDS method provides a novel solution to the spline regression problem and in particular, to the problem of estimating the number and position of the knots. It is motivated by geometric arguments and can be extended to multivariate non-parametric smoothing as well as to generalized linear models. Details of how this may be done are outside the scope of this paper and are the subject of ongoing research.

Acknowledgements

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

Appendix A

Stage A. A knot insertion scheme for variable knot, LS linear B-spline regression.

In order to implement stage A of GeDS, an automatic knot insertion scheme is proposed to construct a variable knot, least squares, linear spline which reproduces the shape of the noise perturbed underlying function f, based on the data set $\{x_i, y_i\}_{i=1}^{N}$. The algorithm may be given the following geometric interpretation. It starts from an LS fit, in the form of a straight line segment, as described in step 1 below. The latter is then sequentially "broken" into a piece-wise linear LS fit, by adding knots, one at a time, at those

points, where the fit deviates most from the shape of the underlying function, (see Fig. 2), according to a bias driven measure of appropriately defined clusters of residuals (see steps 2 - 8). A stopping rule is introduced, which serves as a model selector and allows us to determine the appropriate number and location of the knots of the linear spline fit $\hat{f}(\boldsymbol{\delta}_{l,2}, \hat{\boldsymbol{\alpha}}; x)$ (see steps 9 - 10). So, as is illustrated in Section 5, the linear GeD spline fit $\hat{f}(\boldsymbol{\delta}_{l,2}, \hat{\boldsymbol{\alpha}}; x)$ is a sufficiently accurate reconstruction of f, given that no further smoothness is required. If a smoother fit is needed, a higher order GeD spline is constructed in stage B of the estimation procedure, based on the geometrical form of $\hat{f}(\boldsymbol{\delta}_{l,2}, \hat{\boldsymbol{\alpha}}; x)$.

The knot insertion scheme in stage A can be described as a "greedy" one (see Hastie, 1989), since at each iteration it places a knot, δ^*, where a within-cluster bias dominated measure is maximal (see steps 3 and 5), which is very near to the site where placing a knot gives the largest reduction in the residual sum of squares. This can be quantified using the fact that given an LS fit $\hat{f}(\boldsymbol{\delta}_{k,2}; x)$, with $0 < k < l$ internal knots, if a knot, δ^*, is added in the interval $[\delta_{j^*}, \delta_{j^*+1}]$, $2 \le j^* < k+2$, then the updated LS fit $\hat{f}(\boldsymbol{\delta}^*_{k+1,2}; x)$ adjusts best to the data in $[\delta_{j^*}, \delta_{j^*+1}]$, since $|\hat{f}(\boldsymbol{\delta}_{k,2}; x) - \hat{f}(\boldsymbol{\delta}^*_{k+1,2}; x)|$, $x \in [\delta_j, \delta_{j+1}]$ decreases exponentially in $|j^* - j|$, which is the number of knots between x and δ^*. This follows from Theorem 1 of Zhou and Shen (2001). A formal description of the algorithm of stage A is given next.

Step 1. Set $n = 2$ and $k = 0$. The starting set of knots is $\boldsymbol{\delta}_{0,2} = \{\delta_i\}_{i=1}^4$ with $a = \delta_1 = \delta_2 < \delta_3 = \delta_4 = b$. Find the LS spline fit in the form of the straight line

$$\hat{f}(\boldsymbol{\delta}_{0,2}, \hat{\boldsymbol{\alpha}}; x) = \hat{\alpha}_1 N_{1,2}(x) + \hat{\alpha}_2 N_{2,2}(x) \ .$$

Calculate the residuals $r_i \equiv r(x_i) = y_i - \hat{f}(\boldsymbol{\delta}_{0,2}, \hat{\boldsymbol{\alpha}}; x_i)$, $i = 1, \ ..., N$ and the residual sum of squares $\mathrm{RSS}(k) = \sum_{i=1}^N r_i^2$ of the fit with k internal knots. Since the i-th residual $r(x_i)$, is a function of x_i, $i = 1, 2, \ ..., N$ we will refer to x_i as the x-value of the i-th residual.

Step 2. Group the consecutive residuals r_i, $i = 1, \ ..., N$ into clusters by their sign, i.e., find a number u, $1 \le u \le N$ and a set of integer values $d_j > 0$, $j = 1, \ ..., u$ such that

$$\begin{aligned}&\mathrm{sign}(r_1) = \ ... = \mathrm{sign}(r_{d_1}) \ne \mathrm{sign}(r_{d_1+1}) = \mathrm{sign}(r_{d_1+2}) = \ ... = \mathrm{sign}(r_{d_1+d_2}) \ne \\ &... \ne \mathrm{sign}(r_{d_1+d_2+...+d_{u-1}+1}) = \mathrm{sign}(r_{d_1+d_2+...+d_{u-1}+2}) = \ ... = \mathrm{sign}(r_{d_1+d_2+...+d_u}),\end{aligned}$$

and $\sum_{j=1}^u d_j = N$. Note that the clusters are formed and numbered consecutively, following the order of the residuals, i.e., the order of their x-values $x_1 < x_2 < ... < x_N$.

Step 3. For each of the u clusters of residuals of identical signs, calculate the within-cluster mean residual value

$$\begin{aligned}m_j &= \left(\sum_{i=1}^{d_j} r_{d(j)+i}\right) / d_j \\ &= \left(\sum_{i=1}^{d_j} (\hat{f}_{d(j)+i} - E\,\hat{f}_{d(j)+i}) + (E\,\hat{f}_{d(j)+i} - f_{d(j)+i}) + \epsilon_{d(j)+i}\right) / d_j \\ &= \sum_{i=1}^{d_j} (\hat{f}_{d(j)+i} - E\,\hat{f}_{d(j)+i}) / d_j + \sum_{i=1}^{d_j} (E\,\hat{f}_{d(j)+i} - f_{d(j)+i}) / d_j + \sum_{i=1}^{d_j} \epsilon_{d(j)+i} / d_j \ , \\ &j = 1, \ ..., u,\end{aligned}$$

where $d(j) = d_1 + d_2 + ... + d_{j-1}$ and the three terms in the last decomposition can be interpreted as the within-cluster average variance, bias, and error, respectively. Calculate also the within-cluster range η_j, defined as the difference between the right-most and the left-most x-value of the residuals belonging to the j-th cluster, i.e.,

$$\eta_j = x_{d(j+1)} - x_{d(j)+1},\ j = 1,\ ...,\ u.$$

Throughout the sequel we will refer to $[x_{d(j)+1},\ x_{d(j+1)}]$ as the within-cluster interval.

Step 4. Find $m_{\max} = \max_{1 \le j \le u}(m_j)$ and $\eta_{\max} = \max_{1 \le j \le u}(\eta_j)$ and calculate, correspondingly, the normalized within-cluster mean and range values $m'_j = m_j / m_{\max}$ and $\eta'_j = \eta_j / \eta_{\max}$, so that $0 < m'_j \le 1$, $0 < \eta'_j \le 1$.

Step 5. Calculate the cluster weights

$$w_j = \beta\, m'_j + (1 - \beta)\, \eta'_j,\ \ j = 1,\ ...,\ u, \tag{14}$$

where, β is a real valued parameter, $0 \le \beta \le 1$. The value w_j serves as a measure, attached to the j-th cluster of residuals of identical sign, which measures the deviation of the current least squares linear spline fit $\hat{f}(\boldsymbol{\delta}_{k,2}, \hat{\boldsymbol{\alpha}}; x)$ from f in the j-th cluster. Obviously, the weight w_j itself is a weighted sum of the normalized, within-cluster mean and within-cluster range values and the weight β is one of the parameters whose value will need to be chosen at the start of stage A.

Step 6. Order the clusters in descending order of their weights w_j, $j = 1,\ ...,\ u$, i.e., create a list of corresponding cluster indices $\{j_1, j_2, ..., j_u\}$ such that $w_{j_1} \ge w_{j_2} \ge ... \ge w_{j_u}$. Thus, in order to improve $\hat{f}(\boldsymbol{\delta}_{k,2}, \hat{\boldsymbol{\alpha}}; x)$, in the next step a new knot is inserted, at an appropriate location, in the within-cluster interval of x-values, corresponding to the j_1-th cluster.

Step 7. Check whether there is already a knot in the within-cluster interval of the j_1-th cluster, which is the cluster with the highest rank, according to the ordering in step 6, i.e., check whether

$$\delta_i \in [x_{d(j_1)+1},\ x_{d(j_1)+d_{j_1}}],$$

for each internal knot $\delta_i \in \boldsymbol{\delta}_{k,2}$, $i = 3,\ ...,\ k+2$. If there is already a knot in the within-cluster interval of the j_1-th cluster, the check is repeated for the cluster with index j_2, and so on until the first cluster, with index j_s, say, is found, whose within-cluster interval does not contain a knot. Then, insert a new knot δ^* at the site

$$\delta^* = \left(\sum_{i=d(j_s)+1}^{d(j_s)+d_{j_s}} r_i\, x_i\right) \Big/ \left(\sum_{i=d(j_s)+1}^{d(j_s)+d_{j_s}} r_i\right). \tag{15}$$

Note that (15) is a convex combination of the x-values of the residuals in the j_s-th cluster, hence $\delta^* \in [x_{d(j_s)+1},\ x_{d(j_s)+d_{j_s}}]$. The new knot position (15) can be viewed as a weighted average of the x-values of the residuals in the j_s-th cluster, the weights being

the normalized values of the residuals. The set of knots $\boldsymbol{\delta}_{k,2}$ is being updated as $\boldsymbol{\delta}^*_{k+1,2} := \boldsymbol{\delta}_{k,2} \bigcup \{\delta^*\}$.

So, a new knot is placed where the cluster weight (14) is maximal. In view of the decomposition in step 3, the cluster weight (14) can be referred to as a bias dominated measure since the bias component is dominant in this cluster compared to the variance and error terms (at least at the initial iterations when there are small number of knots in the linear fit and the approximation error is large).

Step 8. Find the least squares linear spline fit

$$\hat{f}\,(\boldsymbol{\delta}^*_{k+1,2},\, \hat{\boldsymbol{\alpha}};\, x) = \sum_{i=1}^{p} \hat{\alpha}_i\, N_{i,2}(x)\ .$$

Since $\boldsymbol{\delta}^*_{k+1,2}$ contains the new knot, the number of B-splines, p, will increase by one.

Step 9. Calculate the residuals r_i, $i = 1,\ ...,\ N$ and the RSS$(k+1)$ for $\hat{f}\,(\boldsymbol{\delta}^*_{k+1,2},\, \hat{\boldsymbol{\alpha}};\, x)$. Note that $\boldsymbol{\delta}_{k,2} \subset \boldsymbol{\delta}^*_{k+1,2}$ implies that $S_{\boldsymbol{\delta}_{k,2}} \subset S_{\boldsymbol{\delta}^*_{k+1,2}}$. Hence $\hat{f}\,(\boldsymbol{\delta}_{k,2},\, \hat{\boldsymbol{\alpha}};\, x) \in S_{\boldsymbol{\delta}^*_{k+1,2}}$ and applying the orthogonality property of least squares estimation it is easy to see that

$$\sum_{i=1}^{N} \left(y_i - \hat{f}\,(\boldsymbol{\delta}_{k,2},\, \hat{\boldsymbol{\alpha}};\, x_i)\right)^2 = \sum_{i=1}^{N} \left(y_i - \hat{f}\,(\boldsymbol{\delta}^*_{k+1,2},\, \hat{\boldsymbol{\alpha}};\, x_i)\right)^2 + \sum_{i=1}^{N} \left(\hat{f}\,(\boldsymbol{\delta}^*_{k+1,2},\, \hat{\boldsymbol{\alpha}};\, x_i) - \hat{f}\,(\boldsymbol{\delta}_{k,2},\, \hat{\boldsymbol{\alpha}};\, x_i)\right)^2 . \quad (16)$$

Equation (16) implies that RSS$(k+1) <$ RSS(k). It is obvious also that RSS(k) will converge to zero as $k + n \longrightarrow N$ since, when $k + n = N$ the fit interpolates the data. The greedy fashion of the new knot placement (15), combined with equation (16), gives rise to the rule for exit from stage A of the GeDS method, and this is given next.

Step 10. Let $q \geq 1$ be a fixed integer, chosen at the beginning of stage A. If the set of knots, $\boldsymbol{\delta}^*_{k+1,2}$, contains less than q internal knots, then the algorithm goes back to step 2. If this is not the case and $\boldsymbol{\delta}^*_{k+1,2}$ contains q or more internal knots then the ratio

$$\alpha = \text{RSS}(k+1)\,/\,\text{RSS}(k+1-q)$$

is calculated. Note that from (16) it follows that $0 < \alpha < 1$. If $\alpha \geq \alpha_{\text{exit}}$, an exit from stage A is performed with the spline fit $\hat{f}\,(\boldsymbol{\delta}_{l,2},\, \hat{\boldsymbol{\alpha}};\, x)$, $l = k + 1 - q$. If $\alpha < \alpha_{\text{exit}}$ then $\hat{f}\,(\boldsymbol{\delta}^*_{k+1,2},\, \hat{\boldsymbol{\alpha}};\, x)$ is taken as the current fit and the algorithm goes back to step 2. The value α_{exit} is chosen ex ante to be close to 1. This is because the ratio α will be close to zero if the fit has improved significantly by adding δ^* and will tend to 1 if no improvement has been achieved in the last $q+1$ consecutive iterations, i.e, the corresponding values of the RSS have stabilized. Our experience has shown that this rule of exit works well as a model selector with $q = 2$, i.e., stabilization of RSS in three consecutive iterations is sufficient to exit from stage A with the appropriate number of knots. Hence, in the implementation of GeDS, q has been fixed equal to 2 by default.

This completes the description of stage A of GeDS.

Appendix B

Proof of Lemma 1. The condition $\lim_{i\to\infty} l_i \,/\, N_i^{1/(2n+1)} = \infty$, implies that, for a fixed $n \geq 2$, the number of knots in the set $\bar{\boldsymbol{t}}_{l_i-(n-2),n}$ should grow at a rate higher than $N_i^{1/(2n+1)}$. At the same time the condition $\lim_{i\to\infty} l_i \log N_i / N_i = 0$ requires the number of knots l_i to be of order smaller than $N_i / \log N_i$. From the definition of the stopping rule in step 10 of Stage A (see Appendix A), it can be seen that $0 < \alpha^i_{\text{exit}} < 1$ and that $\alpha^i_{\text{exit}} = 1$ corresponds to $l_i = N_i$, $\alpha^i_{\text{exit}} = 0$ corresponds to $l_i = \text{const}$. Hence, α^i_{exit} can be chosen (close to 1) so that l_i is of order between $N_i^{1/(2n+1)}$ and $N_i / \log N_i$, as required. □

Proof of Lemma 2. Given Assumption 1, in order to prove that $M^{(r)}_{\boldsymbol{\delta}_i} \leq \gamma$, we need to investigate the knot meshes $\boldsymbol{\delta}_{l_i,2}$ resulting from stage A of GeDS, as $i \to \infty$. Following step 7 of stage A (see Appendix A), it can be seen that the knots $\boldsymbol{\delta}_{l_i,2}$ are non-replicate and are obtained as weighted averages of design points within clusters (see (15)). Consider two consecutive design points, $x_{i,j}$, $x_{i,j+1}$. These may either be part of a common cluster of residuals, in which case $x_{i,j}$ and/or $x_{i,j+1}$ could be the end points of the cluster interval, or fall in two separate (consecutive) clusters. Analyzing all possible ways in which two consecutive design points, $x_{i,j}$, $x_{i,j+1}$, may become part of residual clusters (see step 1 and 2 of Stage A in Appendix A), one can conclude that a knot can be placed at each of the design points $x_{i,j}$, $x_{i,j+1}$ and no more than one knot can be located in the interval $(x_{i,j}, x_{i,j+1})$. Hence, irrespective of α^i_{exit}, the knots $\boldsymbol{\delta}_{l_i,2}$ disperse between the design points without concentration as $i \to \infty$. Therefore, for any sequence α^i_{exit} and for any $\varepsilon > 0$, there will exist $\gamma^* > 0$ such that $P(|M^{(r)}_{\boldsymbol{\delta}_i}| > \gamma^*) \leq \varepsilon$, $r \geq 2$, since the global mesh ratio of the meshes of data points, $M^{(1)}_{\zeta_i}$, is bounded in probability and according to (15) the knots $\boldsymbol{\delta}_{l_i,2}$ are weighted averages of the data points within a cluster. Therefore, there should exist a constant $\gamma > 0$ such that $M^{(r)}_{\boldsymbol{\delta}_i} \leq \gamma$ all i, except on an event whose probability is zero as $N_i \to \infty$. □

Proof of Lemma 3. Consider first the case $n = 2$. Then, $\bar{\boldsymbol{t}}_{l_i,2} \equiv \boldsymbol{\delta}_{l_i,2}$ and (9) is fulfilled for any $r \geq 2$ by assumption. In the case $n > 2$, from (8) we have $\bar{t}_{i,j+n} = (\delta_{i,j+2} + \ldots + \delta_{i,j+n}) / (n-1)$ and for $r = n$, (9) is fulfilled since,

$$M^{(n)}_{\bar{\boldsymbol{t}}_i} = \frac{\max_{n\leq j\leq l_i+1}(\bar{t}_{i,j+n} - \bar{t}_{i,j})}{\min_{n\leq j\leq l_i+1}(\bar{t}_{i,j+n} - \bar{t}_{i,j})} = \frac{\max_{n\leq j\leq l_i+1}(\delta_{i,j+2}+\ldots+\delta_{i,j+n}-\delta_{i,j+2-n}-\ldots-\delta_{i,j})/(n-1)}{\min_{n\leq j\leq l_i+1}(\delta_{i,j+2}+\ldots+\delta_{i,j+n}-\delta_{i,j+2-n}-\ldots-\delta_{i,j})/(n-1)}$$

$$= \frac{\max_{n\leq j\leq l_i+1}\{(\delta_{i,j+2}-\delta_{i,j+2-n})+\ldots+(\delta_{i,j+n}-\delta_{i,j})\}}{\min_{n\leq j\leq l_i+1}\{(\delta_{i,j+2}-\delta_{i,j+2-n})+\ldots+(\delta_{i,j+n}-\delta_{i,j})\}} \leq \frac{(n-1)\max_{2\leq j\leq l_i-n+3}(\delta_{i,j+n}-\delta_{i,j})}{(n-1)\min_{2\leq j\leq l_i-n+3}(\delta_{i,j+n}-\delta_{i,j})} \leq \gamma$$

except on an event whose probability tends to zero as $N_i \to \infty$. In fact, it is also fulfilled for $r > n$ since, $M^{(r)}_{\bar{\boldsymbol{t}}_i} \leq \gamma$ leads to $M^{(r+1)}_{\bar{\boldsymbol{t}}_i} \leq \gamma$, which completes the proof. □

Proof of Theorem 1. According to Lemma 1, there exists a sequence α^i_{exit} such that $\lim_{i\to\infty} l_i \log N_i / N_i = 0$. In addition, under Assumption 1, from Lemma 3 we have that the global mesh ratio, $M^{(n)}_{\bar{\boldsymbol{t}}_i}$, of $\bar{\boldsymbol{t}}_{l_i-(n-2),n}$ is bounded. The result now follows from Theorem 5.1 of Huang (2003) and Theorem 1 of De Boor (1976), establishing a bound in

terms of the global mesh ratio $M_{\bar{\boldsymbol{t}}_i}^{(n)}$ on the L_∞ norm of the orthogonal least squares projection onto the space of splines of order n with knots $\bar{\boldsymbol{t}}_{l_i-(n-2),n}$.□

Proof of Theorem 2. According to Lemma 1, there exists a sequence α_{exit}^i such that $\lim_{i\to\infty} l_i \,/\, N_i^{1/(2\,q+1)} = \infty$ and $\lim_{i\to\infty} l_i \log N_i / N_i = 0$. It follows from Theorem 1 that

$$\sup_{x\in[a,b]} \left| E\left(\hat{f}\left(\bar{\boldsymbol{t}}_{l_i-(n-2),n}, \hat{\boldsymbol{\theta}}; x\right) \,\middle|\, \vec{\mathbf{x}}\right) - f(x) \right| = O_P(l_i^{-q}). \tag{17}$$

Under the above conditions and in view of Remark 1, it can be shown, following Theorem 5.2 of Huang (2003), that

$$\mathrm{Var}\left(\hat{f}\left(\bar{\boldsymbol{t}}_{l_i-(n-2),n}, \hat{\boldsymbol{\theta}}; x\right) \,\middle|\, \vec{\mathbf{x}}\right) \geq C\,(l_i / N_i)\,(1 + o_P(1)). \tag{18}$$

where C is a constant independent of x. The desired result now follows combining (17) and (18).□

Proof of Theorem 3. It was established that under Assumption 1, Lemma 3 holds and the sets of knots $\bar{\boldsymbol{t}}_{l_i-(n-2),n}$ have bounded mesh ratios. From Lemma 1 it follows that for some appropriate sequence α_{exit}^i, $\lim_{i\to\infty} l_i \log N_i / N_i = 0$, and hence (10) holds (see Remark 1). The assertion of the theorem now follows from Theorem 3.1 of Huang (2003) since $\lim_{i\to\infty} l_i \log N_i / N_i = 0$ implies the condition $\lim_{i\to\infty} l_i / N_i = 0$ required there. □

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.

Cohen, E., Riesenfeld, R. F. and Elber, G. (2001). *Geometric Modelling with Splines: An Introduction,* Natick, Massachusetts: A K Peters.

Craven, P. and Wahba, G. (1979). Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the generalized cross validation. *Numerische Mathematik*, **31**, 337-403.

De Boor, C. (1976). A bound on the L_∞ norm of L_2 approximation by splines in terms of a global mesh ratio. *Math. Comp.* 30, 765-771.

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.

Hall, P., Kay, J.W. and Titterington, D.M. (1990). Asymptotically optimal difference-based estimation of variance in non-parametric regression. *Biometrika*, **77**, 521-528.

Hastie, T. (1989). Discussion of "Flexible Parsimonious smoothing and additive modeling (with discussion)" by Friedman, J. H. and Silverman, B.W. *Technometrics*, **31**, 23-29.

Huang, J. Z. (2003). Local asymptotics 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. (1984). A computer program package for solving spline regression problems, In: *Proceedings in Computational Statistics*, *COMPSTAT* (eds T. Havranek, Z. Sidak and M. Novak), pp. 409-414, Wien: Physica-Verlag.

Kaishev, V. K., Dimitrova, D. S., Haberman, S. and Verrall R. (2006). Geometrically designed, variable knot regression splines: Variation diminishing optimality of knots. Statistical Res. Paper 29, 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.

Lee, T. C. M. (2002). On algorithms for ordinary least squares regression spline fitting: A comparable study. *J. Statist. Comput. Simul.*, **72(8)**, 647-663.

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

Lytch, T. and Mørken, K. (1993). A data reduction strategy for splines with application to the approximation of functions and data. *IMA J. Numer. Anal.*, **8**, 185-208.

Schumaker, L. (1981). *Spline Functions: Basic Theory.* Wiley, New York.

Schwetlick, H. and Schütze, T. (1995). Least squares approximation by splines with free knots. *BIT. Numerical Math.*, ***35***, 854-866.

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

Stein, C. (1981). Estimation of the mean of a multivariate normal. *The Ann. Statist*., **9**, 1135-1151.

Stone, C. J. (1982). Optimal global rates of convergence for non parametric regression. *Ann. Statist*., **10**, 1040-1053.

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.

Wang, J. and Yang, L. (2006). Polynomial spline confidence bands for regression curves. downloadable at http://www.msu.edu/~yangli/bandfull.pdf

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

Zhou, S., Shen, X. and Wolfe, D.A. (1998). Local asymptotics for regression splines and confidence regions. *Ann. Statist.*, **26**, 1760-1782.

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

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 1-905752-02-4