ELSEVIER

Computers & Geosciences 32 (2006) 834–845

COMPUTERS & GEOSCIENCES

www.elsevier.com/locate/cageo

Automatic computation of hierarchical biquadratic smoothing splines with minimum GCV

P.A. Hancock[a,*], M.F. Hutchinson[b]

[a]Imperial College London, Silwood Park Campus, Ascot, Berkshire SL5 7PY, UK
[b]Centre for Resource and Environmental Studies, The Australian National University, A.C.T., 0200, Australia

Received 6 February 2005; accepted 21 October 2005

Abstract

A computationally efficient numerical strategy for fitting approximate minimum GCV bivariate thin plate smoothing splines to large noisy data sets was developed. The procedure discretises the bivariate thin plate smoothing spline equations using biquadratic B-splines and uses a nested grid SOR iterative strategy to solve the discretised system. For efficient optimisation, the process incorporates a double iteration that simultaneously updates both the discretised solution and the estimate of the minimum GCV smoothing parameter. The GCV was estimated using a minimum variance stochastic estimator of the trace of the influence matrix associated with the fitted spline surface. A Taylor series expansion was used to estimate the smoothing parameter that minimises the GCV estimate. The computational cost of the procedure is optimal in the sense that it is proportional to the number of grid points supporting the fitted biquadratic spline. Convergence was improved by adding a first order correction to the solution estimate after each smoothing parameter update. The algorithm was tested on several simulated data sets with varying spatial complexity and noise level. An accurate approximation to the analytic minimum GCV thin plate smoothing spline was obtained in all cases.

Keywords: Finite element; Smoothing spline; Multilevel method; Generalised cross validation; Randomized trace

1. Introduction

Thin plate smoothing splines are commonly used to fit smooth surfaces to noisy data. In environmental modelling applications, they have been used to construct surfaces representing surface climate processes (Hutchinson, 1995; Zheng and Basher, 1995), topography (Hutchinson, 1989a), remotely sensed data (Berman, 1994), pollutant dispersion (Ionescu et al., 2000) and plankton distributions (Wood and Horwood, 1995). They are also popular in other fields such as image analysis (Bhaskaran and Konstantinides, 1996), medical research (Lapeer and Prager, 2000) and data mining (Hegland et al., 1998a,b). For practical spatial interpolation problems, surface smoothness is a central issue given that the data observations contain a significant noise component. The data model underlying splines is therefore a statistical decomposition of the observed data into a spatially coherent signal and spatially discontinuous noise. The smoothing spline model is given by

$$z_i = g(\mathbf{x}_i) + \varepsilon_i, \quad \mathbf{x}_i \in \mathbb{R}^d, \quad i = 1, \ldots, n, \tag{1}$$

*Corresponding author. Tel.: +44 207 594 2493; fax: +44 207 594 2531.

E-mail addresses: p.hancock@imperial.ac.uk (P.A. Hancock), hutch@cres.anu.edu.au (M.F. Hutchinson).

0098-3004/$ - see front matter
doi:10.1016/j.cageo.2005.10.010

where z_i are a set of n observations, g is a suitably continuous function of d predictor variables, and ε_i are realisations of a random variable ε. The errors ε_i are generally assumed to be independent with mean zero and variance σ^2. They are assumed to be due to measurement error, and short range microscale variation that occurs over a range below the resolution of the data set (Hutchinson, 1993).

Thin plate smoothing spline functions are designed to approximate the spatially coherent signal, g, by effectively smoothing the data. A thin plate smoothing spline is defined to be the minimiser of

$$\frac{1}{n}\sum_{i=1}^{n}(z_i - f(\mathbf{x}_i))^2 + \lambda J_m^d(f) \tag{2}$$

over functions $f \in \mathscr{X}$, where $\mathscr{X}$ is a space of functions whose partial derivatives of total order m are in $\mathscr{L}^2(\mathbb{R}^d)$, λ is a positive smoothing parameter, and $J_m^d(f)$ is a measure of the roughness of the function f in terms of mth order partial derivatives. Calculation of $J_m^d(f)$ depends on m, the order of the partial derivatives, and the number of independent variables d. For $m = 2$ and $d = 2$, the case considered here

$$J_2^2(f) = \int_{-\infty}^{\infty}(f_{x_1x_1}^2 + 2f_{x_1x_2}^2 + f_{x_2x_2}^2)\,\mathrm{d}x_1\,\mathrm{d}x_2 \tag{3}$$

(Wahba, 1990).

Minimising expression (2) represents a trade-off between fitting the data as closely as possible whilst maintaining surface smoothness. The smoothing parameter λ controls the separation of signal and noise. If $\lambda = 0$ the function f exactly interpolates the data, implying zero noise, whereas if λ is very large the function approaches a plane.

Minimising the generalised cross-validation (GCV) has been shown by Craven and Wahba (1979) to be an accurate method of estimating the λ corresponding to the spline function f that best represents the underlying process g. The GCV is a measure of predictive error, that is defined by removing each data point in turn and summing, with appropriate weighting, the square of the discrepancy of each omitted data point from a surface fitted to all other data points. The GCV is actually calculated implicitly by the formula

$$GCV = n\frac{R}{T^2}, \tag{4}$$

where R is the residual sum of squares $\sum_{i=1}^{n}(z_i - f(x_i))^2$, and $T = \mathrm{tr}(I - A(\lambda))$, where $A(\lambda)$ is the influence matrix associated with the fitted spline (Craven and Wahba, 1979; Wahba, 1990). The quantity $\mathrm{tr}(A(\lambda))$, termed the signal by Wahba (1990), is the effective number of parameters of the fitted thin plate spline. Case studies by Hutchinson (1993); Hutchinson and Gessler (1994) have shown the signal to be a useful statistic in its own right. A signal exceeding $n/2$ can indicate insufficient data or short range correlation in the data (Hutchinson, 1998). Optimising λ by minimising the GCV gives an estimate of the unknown error variance σ^2, stated in Wahba (1990) as

$$\hat{\sigma}^2 = \frac{R}{T}. \tag{5}$$

Minimum GCV thin plate smoothing splines have been widely used in spatial interpolation applications e.g. (Hutchinson and Bischof, 1983; Hutchinson, 1998, 1995; Zheng and Basher, 1995; Price et al., 2000; Jeffrey et al., 2001). However, analytic procedures for calculating thin plate smoothing splines are $O(n^3)$, making them computationally impractical for large data sets. A number of strategies for improving the computational efficiency of analytic thin plate smoothing spline calculation exist (Beatson and Newman, 1992; Beatson and Powell, 1994; Beatson et al., 1996). Finite element techniques have also been developed (Hutchinson, 1989a; Hegland et al., 1998a; Ramsay, 2002). These methods tend to focus on the numeric-analytic properties of thin plate smoothing splines rather than their statistical properties, and do not incorporate an automatic mechanism for optimising smoothness. The method presented in this study is designed to efficiently compute accurate finite element approximations to minimum GCV bivariate thin plate smoothing splines. Another option for obtaining a computationally efficient approximation to minimum GCV thin plate splines is to use the method developed by Bates and Wahba (1982), which approximates thin plate splines for large data sets using knots.

Hutchinson (1989a, 2000) developed a simple multigrid based strategy based on SOR iteration that calculates finite element approximations to bivariate thin plate smoothing splines for elevation data in $O(N)$ operations, where N is the number of grid points. This method uses a simple Newton iteration to optimise the smoothing parameter to yield a specified residual sum of squares. This criterion is appropriate in the context of interpolating topography, where an estimate of the amount of noise is available (Hutchinson, 1989a). The