Math Geol (2007) 39: 419–434
DOI 10.1007/s11004-007-9104-x

Constrained Smoothing of Noisy Data Using Splines in Tension

Nicholas A. Teanby

Received: 29 June 2006 / Accepted: 1 February 2007 / Published online: 9 August 2007

Abstract A method of fitting a smooth cubic spline curve through noisy data points is presented. Overshoots of the spline curve between data points were prevented by applying tension to the fit using a quadratic spring approximation, which allowed a linear inverse theory approach to be adopted. Error-bars in the measured data were mapped through the inversion process to give the covariance of the fitted curve. This is an improvement over previous methods, which largely neglect the effect of data errors on the fit. Another improvement is to impose fixed constraints on the fit by simultaneously applying the method of Lagrange multipliers. The effect of these constraints on the covariance of the fitted curve is quantified using results from linear algebra. Example applications to synthetic data and a record of magnetic inclination from Hawaii are given.

Keywords Curve fitting · Data analysis · Lagrange multipliers

Introduction

A problem often encountered in the physical sciences is fitting a smooth curve through a series of (x, y) data points—either to allow the value of y to be determined at an arbitrary value of x, to increase the signal-to-noise ratio, or both. The solution to this problem has two sub-types: (1) interpolation, where the curve is required to pass exactly through each data point; and (2) smoothing, where we only wish the curve to fit the data to within measurement error. Piecewise cubic spline fits have proved popular for both cases because they have the property that the second derivative (curvature) of the fitted curve is minimized. Background material on

N.A. Teanby (✉)
Atmospheric, Oceanic & Planetary Physics, Clarendon Laboratory, Department of Physics, University of Oxford, Parks Road, Oxford OX1 3PU, UK
e-mail: teanby@atm.ox.ac.uk

spline properties and curve fitting can be found in standard texts by de Boor (1978), Lancaster and Šalkauskas (1986), Schumaker (1993), and Wahba (1990). Hutchinson and Gessler (1994) discuss the benefits of using splines when fitting curves including their ability to recover underlying trends from noisy data. Constraints can be imposed on fitted curves using Lagrange multipliers (e.g., Menke 1989; Talmi and Gilat 1977).

An undesirable property of the minimum curvature property of cubic splines is that the fitted curve can suffer from unphysical oscillations between data points (e.g., Schweikert 1966; Smith and Wessel 1990). These oscillations can be suppressed by relaxing the minimum curvature condition and applying tension to the curve. Cline (1974) discussed the problem of interpolation using splines in tension and provided algorithms for determining a smooth interpolating curve. Renka (1993) used a modified version of the iterative procedure presented in Reinsch (1967, 1971) to fit a smooth curve through a set of data points. More recent approaches have included minimizing the strain energy in an elastic bar (in 1D) or plate (in 2D) constrained to pass through the data points. Inoue (1986) obtained a smooth cubic spline fit by simultaneously minimizing the data misfit and roughness of the fitted curve, which is analogous to a bar or plate in tension. Mitášová and Mitáš (1993) used the radial splines of Talmi and Gilat (1977) and applied tension to obtain a smooth fit. They also allowed for application of anisotropic tension when fitting multi-dimensional data to preserve data trends.

Smith and Wessel (1990) present a method to interpolate a set of data points, which used an iterative finite difference method to determine the shape of a bar in tension passing through the data points. Their method forms the basis for the interpolation routine surface in the widely used GMT software package (Wessel and Smith 1998). Sandwell (1987) introduces Green's functions to solve for the displacement of the bar, which linearize the problem and remove the need for iteration, although no tension is applied to the fitted curve. Subsequently, the Green's function method of Sandwell (1987) has been generalized by Wessel and Bercovici (1998) to include a tension term. Wessel and Bercovici (1998) also discussed how smoothing instead of interpolation could be achieved by zeroing some of the smallest eigenvalues.

In geological and geophysical datasets, the data points often have large errors, so interpolation is not always appropriate and it is often better to fit a smooth curve through the data to within the error-bars. If the fitted curve is to form the basis of subsequent analysis, it is essential that the errors in the fit are quantified. However, the previous work with splines in tension has primarily focused on interpolation, sometimes with extension to smoothing, and often does not explicitly take the errors in the data points into account when calculating the fitted curve. Provision for simultaneously imposing fixed constraints on the fit has also been neglected.

Therefore, in this paper, a linear inverse theory approach is developed to fit a smooth curve through noisy data with an arbitrary spline function under tension. We extend upon previous work by mapping the errors in the measured data into the covariance of the fitted curve. This is made possible by our linear inversion approach. Fixed constraints on the fitted curve are imposed with the method of Lagrange multipliers if required. The effect of applying constraints is included in the error analysis. A systematic approach to choosing the tension parameter and number of fitting parameters is also discussed and examples are given.

Method

Representing a Curve as a Sum of Basis Functions

A convenient way to represent a curve $f(x)$ is as a sum of simpler curves, referred to as basis functions or splines. Basis functions can either be global or local, depending on whether each basis function is defined over the entire range of the dataset or is finite in extent. A Fourier series is an example of a global basis. We adopt a local basis here, as it does not introduce correlations between widely spaced data points and allows large datasets to be fitted in segments (e.g., Mitášová and Mitáš 1993). The curve parameterization consists of a set of M splines $B_j(x)$, each scaled by a spline coefficient m_j and centred on knot point K_j along the x-axis, for $j = 1, \dots, M$. The contribution to $f(x)$ by the jth spline is $m_j B_j(x)$, and the curve $f(x)$ is defined by the sum of all the basis functions

$$f(x) = \sum_{j=1}^{M} m_j B_j(x). \tag{1}$$

Figure 1 shows a graphical explanation of how the curve is constructed using (1). The derivatives of $f(x)$ can be calculated by simply summing the derivatives of the splines. For example, the gradient is given by

$$f'(x) = \sum_{j=1}^{M} m_j B'_j(x), \tag{2}$$

where the prime represents differentiation with respect to x.

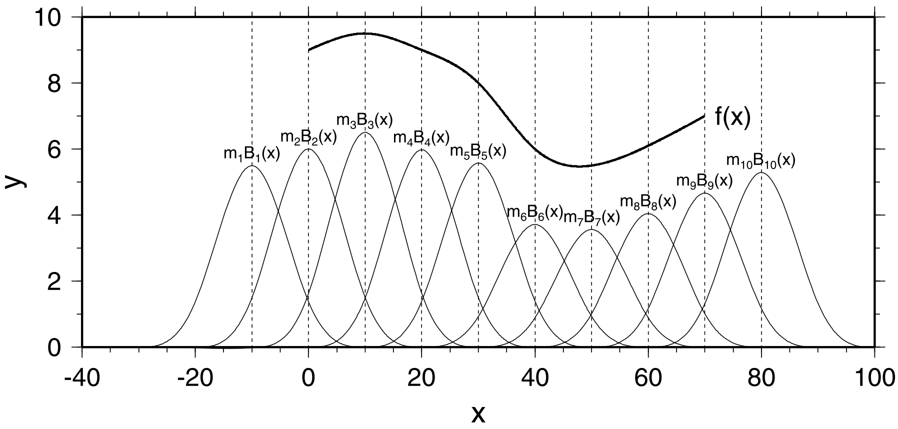

Fig. 1 Illustration of how a smooth curve $f(x)$ is constructed by summing basis functions $B_j(x)$. In this case $f(x)$ is the sum of 10 basis functions with knot spacing $K = 10$. The basis functions are identical, except that they have different scalings and are centred on different knot points. Note that the basis functions B_1 and B_{10} beyond the ends of $f(x)$ still influence the curve and must be included in the sum

Usually all the splines $B_j(x)$ are identical, but are centred on different knot points. Therefore, we only need define one parametric spline in terms of the normalized distance t_j from each knot point, such that $B_j(x) = B(t_j)$. The normalization of t_j is such that $t_j = 0$ is at the spline centre and $t_j = \pm 1$ is the distance to adjacent knot points.

The parametric spline $B(t_j)$ can be any smooth function. However, when fitting a curve $f(x)$ to measured data, piecewise cubic polynomials are often used as they maintain continuous first and second derivatives and have the property of minimizing the curvature (second derivative) of $f(x)$ (Lancaster and Šalkauskas 1986). Throughout the rest of this paper we adopt the parametric spline from Constable and Parker (1992), which is defined by four cubic polynomial segments

$$B(t) = \begin{cases} (2+t)^3, & -2 \le t < -1, \\ 4 - 6t^2 - 3t^3, & -1 \le t < 0, \\ 4 - 6t^2 + 3t^3, & 0 \le t < 1, \\ (2-t)^3, & 1 \le t < 2. \end{cases} \tag{3}$$

This spline is not unique, and any suitable basis function could be used for fitting the curve. However, if we require the derivatives of $f(x)$ to be continuous, then the derivatives of $B(t)$ must also be continuous across the segment boundaries and equal to zero at each end. This condition is satisfied by the above function, as shown in Fig. 2.

All splines which overlap with the data must be included in the curve parameterization. For the example shown in Fig. 1, which has equal knot spacing $K = 10$ and splines of width $4K = 40$, the first and last basis functions occur one knot spacing before and after the ends of $f(x)$, respectively. Splines $j = 2$ and $j = M - 1$ are centred on the ends of $f(x)$. In this case $K = (x_N - x_1)/(M - 3)$ and the normalized distance from the knot point is $t = (x - x_1)/K - (j - 2)$ for the jth spline.

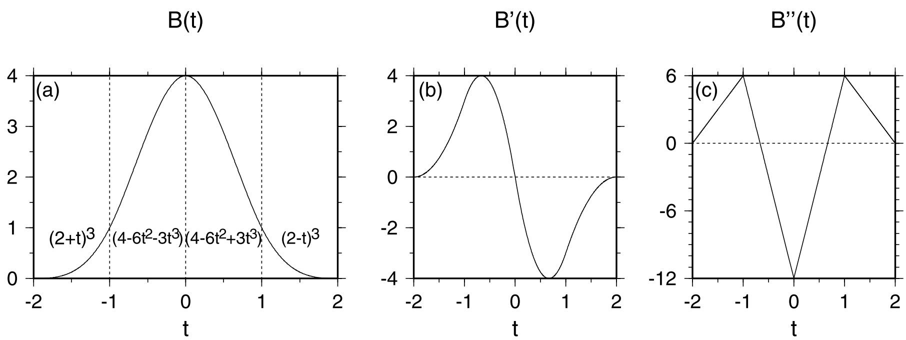

Fig. 2 **a** shows an example of a cubic spline basis function composed of four cubic polynomial segments. **b** and **c** show that the gradient and curvature of this spline are continuous across segment boundaries and zero at the end points. These are necessary conditions if the resulting fitted curve is also to have continuous gradient and curvature

Fitting a Smooth Curve to Noisy Data

Consider N measurements (x_i, y_i), $i = 1, \dots, N$ with errors σ_i through which we wish to fit a smooth curve $f(x)$. If the functional form of the curve is known, then a least squares fit to the function parameters is the most appropriate way to define $f(x)$. Otherwise it is appropriate to use a sum of basis functions to represent the curve.

The curve $f(x)$ is defined by the spline coefficients m_j in (1), which we will determine using linear inverse theory (Gubbins 2004; Menke 1989; Tarantola 1987). In this and subsequent sections, summations are represented by matrix multiplications where possible. This notation has the advantage that all variables can be included in the expressions, and we can use techniques from linear algebra to solve the resulting equations. Vectors are expressed with bold lower-case letters and matrices with bold upper-case letters. Equation (1) is applied at each of the N measurement points x_i and the resulting set of N simultaneous equations can then be expressed in matrix form as

$$\mathbf{f} = \mathbf{A}\mathbf{m}, \tag{4}$$

where $\mathbf{f}$ is an N-dimensional vector of curve evaluations at points x_i, $\mathbf{A}$ is an $N \times M$ matrix with elements $A_{ij} = B_j(x_i)$, and $\mathbf{m}$ is an M-dimensional vector of model parameters.

The problem is to find the fitting parameters m_j that minimise the misfit between the fitted curve $f(x)$ and the data points (x_i, y_i). Therefore, we define an objective function ϕ as the sum of the squared error weighted misfit

$$\phi = \sum_{i=1}^{N} \left(\frac{y_i - f(x_i)}{\sigma_i} \right)^2. \tag{5}$$

Defining $\mathbf{C}_e$ as the data covariance matrix with diagonal elements $C_{e_{ii}} = \sigma_i^2$ and writing in terms of matrices gives

$$\phi = (\mathbf{y} - \mathbf{A}\mathbf{m})^T \mathbf{C}_e^{-1} (\mathbf{y} - \mathbf{A}\mathbf{m}), \tag{6}$$

$$\phi = \mathbf{y}^T \mathbf{C}_e^{-1} \mathbf{y} - 2\mathbf{y}^T \mathbf{C}_e^{-1} \mathbf{A}\mathbf{m} + \mathbf{m}^T \mathbf{A}^T \mathbf{C}_e^{-1} \mathbf{A}\mathbf{m}, \tag{7}$$

where $\mathbf{C}_e^{-1}$ is the inverse of the covariance matrix with diagonal elements $C_{e_{ii}}^{-1} = 1/\sigma_i^2$. Note that using a diagonal covariance matrix $\mathbf{C}_e$ assumes no correlation between measurements. If correlations do exist, the covariance matrix $\mathbf{C}_e$ in the following equations should contain off-diagonal elements (Gubbins 2004; Menke 1989).

The minimum of ϕ gives the least squares solution and is found by differentiating (7) with respect to each of the fitting parameters m_k

$$\frac{\partial \phi}{\partial m_k} = -2\mathbf{y}^T \mathbf{C}_e^{-1} \mathbf{A} + 2\mathbf{A}^T \mathbf{C}_e^{-1} \mathbf{A}\mathbf{m}. \tag{8}$$

Setting these derivatives to zero minimises ϕ and can be re-arranged to give the standard linear inverse theory result for a weighted least squares solution

$$\mathbf{m} = \left(\mathbf{A}^T \mathbf{C}_e^{-1} \mathbf{A}\right)^{-1} \mathbf{A}^T \mathbf{C}_e^{-1} \mathbf{y}. \tag{9}$$

This represents the solution of a set of N simultaneous equations for the M fitting parameters m_j. The problem of finding $\mathbf{m}$ is over-determined for $N > M$, provided there is at least one data point between each knot point. Each additional knot point introduces one more basis function and hence at least one more data point is required for there to be at least as many equations as unknowns. If there are more parameters than data points then $\mathbf{A}^T\mathbf{C}_e^{-1}\mathbf{A}$ does not have an inverse and the problem is under-determined. For the case of smoothing a noisy dataset $N \gg M$ and the problem is over-determined. However, the distribution of the data points is important and some of the parameters will be better determined than others.

To calculate the errors on the fitting parameters $\mathbf{m}$, we use the results from Gubbins (2004) that if two vectors $\mathbf{a}$ and $\mathbf{b}$ are related by the linear transformation $\mathbf{G}$

$$\mathbf{b} = \mathbf{G}\mathbf{a}, \tag{10}$$

then the covariance of $\mathbf{b}$, $\mathbf{C}_b$, is related to the covariance of $\mathbf{a}$, $\mathbf{C}_a$, by

$$\mathbf{C}_b = \mathbf{G}\mathbf{C}_a\mathbf{G}^T. \tag{11}$$

Therefore, comparison with (9) gives $\mathbf{G} = (\mathbf{A}^T\mathbf{C}_e^{-1}\mathbf{A})^{-1}\mathbf{A}^T\mathbf{C}_e^{-1}$ and the error on the spline coefficients $\mathbf{m}$ reduces to

$$\mathbf{C}_m = \left(\mathbf{A}^T\mathbf{C}_e^{-1}\mathbf{A}\right)^{-1}. \tag{12}$$

The fitted curve $f(x)$ can be calculated at any set of x values x_q by defining a matrix $\mathbf{D}$ with elements $D_{qj} = B_j(x_q), q = 1, \dots, Q$,

$$\mathbf{f} = \mathbf{D}\mathbf{m}. \tag{13}$$

The errors in $\mathbf{m}$ can be mapped into the errors in $\mathbf{f}$ by

$$\mathbf{C}_f = \mathbf{D}\mathbf{C}_m\mathbf{D}^T. \tag{14}$$

The standard error $\sigma_{f(x_q)}$ of each fitted point $f(x_q)$ is given by the square roots of the diagonal elements of $\mathbf{C}_f$

$$\sigma_{f(x_q)} = \sqrt{C_{f_{qq}}}, \tag{15}$$

which can be used to give the 1-sigma error envelope of the fitted curve.

Constraints on the Fit

We often have additional constraints to apply to the fitted curve $f(x)$ that can be considered to have zero error. These constraints may be based on physical arguments or desirable properties of the fitted curve. For example, we may wish $f(x)$ to pass through the origin or constrain the gradient or curvature at the ends of the curves. These constraints should not be included in (9) as they have zero error which would lead to a singular inverse covariance matrix.

Instead, constraints should be imposed by using the method of Lagrange multipliers (Menke 1989; Boas 1983; Talmi and Gilat 1977), which has the advantage that any number of constraints can be imposed on the data, provided that the constraints can be expressed as a linear mapping of the model parameters.

First, a matrix $\mathbf{F}$ is defined such that $\mathbf{Fm}$ has the value of the constraints vector $\mathbf{h}$. We then have the equation of constraints

$$\mathbf{Fm} = \mathbf{h}, \tag{16}$$

where $\mathbf{m}$ is the vector of the model parameters, as before. For example, if we wish to constrain the curvature at the ends of $f(x)$ to be zero, $\mathbf{F}$ and $\mathbf{h}$ are given by

$$\mathbf{F} = \begin{pmatrix} B_1''(x_1) & B_2''(x_1) & \cdots & B_M''(x_1) \\ B_1''(x_N) & B_2''(x_N) & \cdots & B_M''(x_N) \end{pmatrix}, \tag{17}$$

$$\mathbf{h} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}. \tag{18}$$

Alternatively, if we require the dataset to pass through the point (x_0, y_0), then

$$\mathbf{F} = \begin{pmatrix} B_1(x_0) & B_2(x_0) & \cdots & B_M(x_0) \end{pmatrix}, \tag{19}$$

$$\mathbf{h} = (y_0). \tag{20}$$

Once the constraints are defined, we follow the approach of Menke (1989) and solve

$$\begin{bmatrix} \mathbf{A}^T\mathbf{C}_e^{-1}\mathbf{A} & \mathbf{F}^\mathbf{T} \\ \mathbf{F} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{m} \\ \lambda \end{bmatrix} = \begin{bmatrix} \mathbf{A}^T\mathbf{C}_e^{-1}\mathbf{y} \\ \mathbf{h} \end{bmatrix}, \tag{21}$$

where λ are dummy parameters known as the Lagrange multipliers and the large square brackets denote a partitioned compound matrix. The first matrix on the left-hand side of (21) is square, so the solution can be found by simple inversion

$$\begin{bmatrix} \mathbf{m} \\ \lambda \end{bmatrix} = \begin{bmatrix} \mathbf{A}^T\mathbf{C}_e^{-1}\mathbf{A} & \mathbf{F}^\mathbf{T} \\ \mathbf{F} & \mathbf{0} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{A}^T\mathbf{C}_e^{-1}\mathbf{y} \\ \mathbf{h} \end{bmatrix}. \tag{22}$$

However, to map the data covariance $\mathbf{C}_e$ into the covariance of the model parameters $\mathbf{C}_m$ using (11), we require an expression for $\mathbf{m}$ in terms of $\mathbf{y}$. We use the result that the inverse of a partitioned matrix is given by (e.g., Faliva and Zoia 2002)

$$\begin{bmatrix} \mathbf{E} & \mathbf{F}^\mathbf{T} \\ \mathbf{F} & \mathbf{0} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{E}^{-1} - \mathbf{E}^{-1}\mathbf{F}^T(\mathbf{F}\mathbf{E}^{-1}\mathbf{F}^T)^{-1}\mathbf{F}\mathbf{E}^{-1} & \mathbf{E}^{-1}\mathbf{F}^T(\mathbf{F}\mathbf{E}^{-1}\mathbf{F}^T)^{-1} \\ (\mathbf{F}\mathbf{E}^{-1}\mathbf{F}^T)^{-1}\mathbf{F}\mathbf{E}^{-1} & -(\mathbf{F}\mathbf{E}^{-1}\mathbf{F}^T)^{-1} \end{bmatrix}, \tag{23}$$

where $\mathbf{E}$ is a square $M \times M$ matrix, $\mathbf{F}$ is a rectangular $L \times M$ matrix, and $\mathbf{0}$ is the $L \times L$ null matrix. Setting $\mathbf{E} = \mathbf{A}^T\mathbf{C}_e^{-1}\mathbf{A}$ by comparison of (22) and (23) allows us to multiply out (22) and obtain an expression for $\mathbf{m}$

$$\mathbf{m} = \left(\mathbf{E}^{-1} - \mathbf{E}^{-1}\mathbf{F}^T\left(\mathbf{F}\mathbf{E}^{-1}\mathbf{F}^T\right)^{-1}\mathbf{F}\mathbf{E}^{-1}\right)\mathbf{A}^T\mathbf{C}_e^{-1}\mathbf{y} + \mathbf{E}^{-1}\mathbf{F}^T\left(\mathbf{F}\mathbf{E}^{-1}\mathbf{F}^T\right)^{-1}\mathbf{h}. \tag{24}$$

The constraint vector $\mathbf{h}$ is constant with zero covariance, so the second term does not contribute to the model covariance. Therefore, substituting $\mathbf{G} = (\mathbf{E}^{-1} - \mathbf{E}^{-1}\mathbf{F}^T(\mathbf{F}\mathbf{E}^{-1}\mathbf{F}^T)^{-1}\mathbf{F}\mathbf{E}^{-1})\mathbf{A}^T\mathbf{C}_e^{-1}$ from the first term into (11) gives the covariance of the model parameters

$$\mathbf{C}_\mathbf{m} = \left(\mathbf{E}^{-1} - \mathbf{E}^{-1}\mathbf{F}^T\left(\mathbf{F}\mathbf{E}^{-1}\mathbf{F}^T\right)^{-1}\mathbf{F}\mathbf{E}^{-1}\right)\mathbf{A}^T\mathbf{C}_e^{-1}\mathbf{C}_e \times \left(\left(\mathbf{E}^{-1} - \mathbf{E}^{-1}\mathbf{F}^T\left(\mathbf{F}\mathbf{E}^{-1}\mathbf{F}^T\right)^{-1}\mathbf{F}\mathbf{E}^{-1}\right)\mathbf{A}^T\mathbf{C}_e^{-1}\right)^T, \tag{25}$$

which simplifies to

$$\mathbf{C}_\mathbf{m} = \left(\mathbf{A}^T\mathbf{C}_e^{-1}\mathbf{A}\right)^{-1} - \left(\mathbf{A}^T\mathbf{C}_e^{-1}\mathbf{A}\right)^{-1}\mathbf{F}^T\left(\mathbf{F}\left(\mathbf{A}^T\mathbf{C}_e^{-1}\mathbf{A}\right)^{-1}\mathbf{F}^T\right)^{-1}\mathbf{F}\left(\mathbf{A}^T\mathbf{C}_e^{-1}\mathbf{A}\right)^{-1}. \tag{26}$$

Note that all the diagonal elements of the second term are positive, so adding constraints always reduces the errors in the parameters.

Applying Tension

The least squares fit of cubic spline basis functions in the previous sections implicitly gives the minimum curvature solution. If the distribution of data points is unfavourable, the resulting curve may overshoot the data considerably between data points in order to preserve the minimum curvature property (see Fig. 3).

Overshoots of $f(x)$ can be avoided by relaxing the minimum curvature condition and simultaneously minimizing the misfit and curve length. Solutions with overshoots will then be penalised due to their large curve length. Minimizing the curve length is analogous to applying tension to the spline fit. We redefine the objective function ϕ from (6) as

$$\phi = (1-\alpha)(\mathbf{y} - \mathbf{A}\mathbf{m})^T\mathbf{C}_e^{-1}(\mathbf{y} - \mathbf{A}\mathbf{m}) + \alpha T, \tag{27}$$

where T is the tension and α is the weight given to minimizing the tension compared to the misfit to the data. The value of α lies in the range $0 \le \alpha < 1$, with $\alpha = 0$ equivalent to the no tension case.

Differentiating ϕ with respect to the model parameters m_k as before gives

$$\frac{\partial \phi}{\partial m_k} = 2(1-\alpha)\left[-\mathbf{y}^T\mathbf{C}_e^{-1}\mathbf{A} + \mathbf{A}^T\mathbf{C}_e^{-1}\mathbf{A}\mathbf{m}\right] + \alpha\frac{\partial T}{\partial m_k}. \tag{28}$$

Therefore, to solve the problem we need to define T in terms of $\mathbf{m}$ so that the derivative of T can be calculated. To this end, we model our curve $f(x)$ as an elastic spring

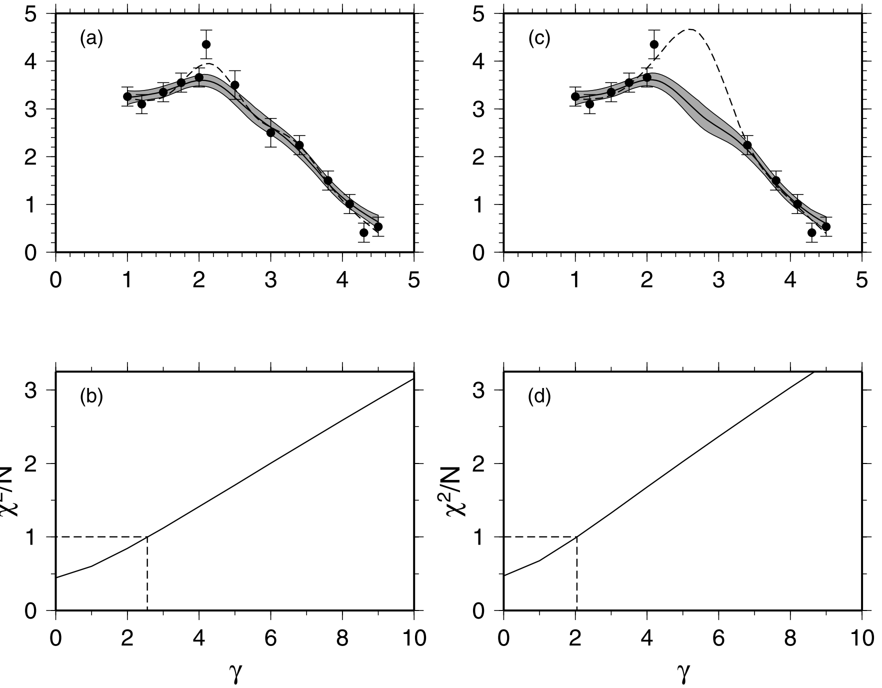

Fig. 3 Example application of spline fitting to a synthetic dataset (*solid circles* with error-bars). **a** shows the fit obtained without tension (*dashed line*) and with tension (*solid line* with *grey* error envelope). The tension parameter γ was determined from the trade-off curve in **b** and chosen such that $\chi^2/N = 1$. **c** and **d** are the same as **a** and **b**, except that the central data points have been removed. This causes the no-tension fit to have large overshoots between data points. The number of parameters $M = 9$ in each case

that obeys Hooke's law. Therefore, the tension T will be proportional to the length of the curve given by integrating an element of arc ds along the curve

$$T = \int_{x_1}^{x_N} \mathrm{d}s \tag{29}$$

$$= \int_{x_1}^{x_N} \left[\mathrm{d}x^2 + \mathrm{d}y^2\right]^{\frac{1}{2}} \tag{30}$$

$$= \int_{x_1}^{x_N} \left[1 + \left(\frac{\mathrm{d}y}{\mathrm{d}x}\right)^2\right]^{\frac{1}{2}} \mathrm{d}x. \tag{31}$$

We can approximate this integral with a simple quadrature scheme evaluated on a grid of x values x_q, spaced by Δx, for $q = 1, \ldots, Q$. The tension is then

$$T = \sum_{q=1}^{Q} \left[1 + \left(\sum_{j=1}^{M} m_j B_j'(x_q)\right)^2\right]^{\frac{1}{2}} \Delta x, \tag{32}$$

and Δx must be small enough for the quadrature scheme to give a reasonable approximation. The spline functions are smooth, so $\Delta x = K/5$ is sufficient to give an accurate approximation.

Differentiating with respect to m_k gives

$$\frac{\partial T}{\partial m_k} = \Delta x \sum_{q=1}^{Q} \left[B_k'(x_q) \left[\sum_{j=1}^{M} m_j B_j'(x_q) \right] \left[1 + \left(\sum_{j=1}^{M} m_j B'(x_q) \right)^2 \right]^{-\frac{1}{2}} \right]. \quad (33)$$

The square root means that this expression cannot be written in terms of linear matrix multiplications and when substituted into (28), the resulting formula cannot be rearranged to give $\mathbf{m}$ as the subject. Therefore, the problem is non-linear and solving for $\mathbf{m}$ requires iteration with a method such as conjugate gradients (Press et al. 1992) using the gradient $\partial\phi/\partial m_k$ to guide the iteration to a minimum in ϕ. Care must be taken to avoid converging on local minima in ϕ and techniques such as those in Acton (1990) or Sambridge (1999) could be employed to give a more robust minimization.

However, the problem can be linearized if we introduce a further approximation for the tension. Assuming that $\mathrm{d}y/\mathrm{d}x < 1$, we can use the binomial expansion $\sqrt{1+a} = 1 + a/2 + \cdots$ to re-write (32) as

$$T = \Delta x \sum_{q=1}^{Q} \left[1 + \frac{1}{2} \left(\sum_{j=1}^{M} m_j B_j'(x_q) \right)^2 + \cdots \right]. \quad (34)$$

Ignoring terms of higher order than the quadratic and differentiating with respect to m_k as before gives

$$\frac{\partial T}{\partial m_k} = \Delta x \sum_{q=1}^{Q} \left[B_k'(x_q) \sum_{j=1}^{M} m_j B_j'(x_q) \right]. \quad (35)$$

Introducing the matrix of spline gradients evaluated at the quadrature points, $\mathbf{D}'_{qj} = B_j'(x_q)$, and rewriting this in terms of matrices gives

$$\frac{\partial T}{\partial m_k} = \Delta x \mathbf{D}'^T \mathbf{D}' \mathbf{m}. \quad (36)$$

Substituting into (28) gives

$$\frac{\partial \phi}{\partial m_k} = 2(1-\alpha) w \left[-\mathbf{y}^T \mathbf{C}_e^{-1} \mathbf{A} + \mathbf{A}^T \mathbf{C}_e^{-1} \mathbf{A} \mathbf{m} \right] + \alpha \Delta x \mathbf{D}'^T \mathbf{D}' \mathbf{m}. \quad (37)$$

Setting $\partial\phi/\partial m_k = 0$ to minimize ϕ with respect to m_k then results in

$$0 = -2(1-\alpha)\mathbf{y}^T \mathbf{C}_e^{-1} \mathbf{A} + \left(2(1-\alpha) \mathbf{A}^T \mathbf{C}_e^{-1} \mathbf{A} + \alpha \Delta x \mathbf{D}'^T \mathbf{D}' \right) \mathbf{m}. \quad (38)$$

This is approximately equivalent to minimising the misfit and curve length if $\mathrm{d}y/\mathrm{d}x < 1$ everywhere and terms higher than quadratics can be neglected. If $\mathrm{d}y/\mathrm{d}x \not< 1$, the dataset can be rescaled so that it is. If the data is not rescaled, minimizing (34) will be equivalent to minimizing the tension in a set of quadratic springs (tension $\propto$ length^2) placed at intervals of Δx along the curve. These quadratic springs will still be effective at penalizing overshoots even if the data is not rescaled. Equation (38) can be re-arranged to give $\mathbf{m}$ as the subject and results in the linear solution

$$\mathbf{m} = \left[\mathbf{A}^T \mathbf{C}_e^{-1}\mathbf{A} + \frac{\alpha}{2(1-\alpha)}\Delta x \mathbf{D}'^T \mathbf{D}'\right]^{-1} \mathbf{A}^T \mathbf{C}_e^{-1}\mathbf{y}. \tag{39}$$

Combining the tension factor into a single term $\gamma = \Delta x \alpha / 2(1-\alpha)$ gives

$$\mathbf{m} = \left[\mathbf{A}^T \mathbf{C}_e^{-1}\mathbf{A} + \gamma \mathbf{D}'^T \mathbf{D}'\right]^{-1} \mathbf{A}^T \mathbf{C}_e^{-1}\mathbf{y}. \tag{40}$$

Therefore, using a non-linear spring analogy prevents overshooting of the data, while also allowing a linear solution to be obtained. This may be preferable to the linear spring as iteration is not required and problems with local minima are avoided.

Constraints can be added using the method of Lagrange multipliers as before by replacing $\mathbf{A}^T \mathbf{C}_e^{-1}\mathbf{A}$ in (22) with $\mathbf{A}^T \mathbf{C}_e^{-1}\mathbf{A} + \gamma \mathbf{D}'^T \mathbf{D}'$.

If the matrix $\mathbf{A}^T \mathbf{C}_e^{-1}\mathbf{A}$ has small eigenvalues, then the inverse can become unstable or singular if the spread of data points is such that some of the spline coefficients are not well constrained. The $\gamma \mathbf{D}'^T \mathbf{D}'$ term prevents this by providing damping of this matrix. The covariance of the model parameters $\mathbf{m}$ is given by (26), as before, which can be mapped to the covariance and error envelope of the fitted curve using (14) and (15), respectively.

Practical Considerations

For smoothly varying datasets with low noise the choice of fitting parameters is not critical because the data will constrain the fit very well. However, for noisy datasets containing regions of sparse data coverage and (possibly) spurious outliers the choice becomes more important and not so obvious. The choice of fitting parameters for a particular dataset can be aided by consideration of the χ^2 statistic (e.g., Bevington and Robinson 1992) defined by

$$\chi^2 = \sum_{i=1}^{N} \left(\frac{f(x_i) - y(x_i)}{\sigma_i}\right)^2. \tag{41}$$

If the model $f(x)$ fits the data points $y(x_i)$ to a level consistent with the data errors σ_i, then $\chi^2/N \approx 1$. If $\chi^2/N > 1$, the model provides a poor fit to the data, whereas if $\chi^2/N < 1$ the model fits the data points too precisely given the errors in the measurements.

Number of Basis Functions M

Probably the most fundamental parameter for the fit is the number of model parameters M, as this defines the amount of detail that $f(x)$ can have. If we choose M too small, then $f(x)$ will be too smooth and not be representative of the trends that we wish to model. On the other hand, if M is too large, then $f(x)$ will have too much detail, will be heavily influenced by the noise, and have spurious features.

Consideration of the physical processes that created the data can sometimes be used to guide the choice of M, such that the knot spacing K is comparable to the expected level of detail. For example, if we are oversampling a time series that has been low pass filtered with a cut off period of P, then variations with periods smaller than P should not be present in the dataset, so we should choose M such that $K \geq P$.

Consider an experiment in which we are sampling a sediment core at 1 cm intervals, from which we will measure the remanent magnetization to obtain a record of past magnetic field variations. It is known that magnetization of sediments is "locked-in" over a depth of ∼8 cm (Kent 1973; Teanby and Gubbins 2000)—effectively low pass filtering the record. Therefore, to fit a smooth curve to the resulting dataset, we should choose M such that $K \geq 8$ cm.

If we do not know apriori what level of detail we expect in our data, then the χ^2 statistic can be used to determine an appropriate choice of M. If measurement errors are well determined, then we should choose M such that $\chi^2/N \approx 1$. However, if the error-bars are underestimated, then we may choose M to be too high and introduce spurious features into $f(x)$.

A more robust alternative would be to start at a low value of M and steadily increase M, calculating χ^2 each time. Convergence in χ^2 (say $\Delta\chi^2 < 1\%$) then implies that increasing M gives no significant improvement to the fit and gives us an appropriate value of M.

Last but not least is the requirement that M should be less than the number of data points N. Otherwise, the inverse problem will be under-determined. It is preferable that $M \ll N$, as the lower the value of M, the better determined the spline coefficients will be. Wold (1974) discusses some alternative strategies for determining the knot positions.

Tension Factor γ

Ideally we would not need to apply any tension and would like to fit the data as well as possible. In practice we would like to fit the data but also stop overshoots between data points so that the fitted curve is useful for subsequent analysis. In this case just enough tension should be applied to stop the overshoots, but not so much so that the fit to the data is degraded. A good criteria for selecting an appropriate value of γ is to plot a trade-off curve of χ^2 as a function of γ and select a value of γ that gives $\chi^2/N \approx 1$.

In some cases the data may contain so much variability or outliers that the model is not able to fit the data to within errors, even with no tension applied. Therefore, χ^2/N will always be greater than unity and it will not be possible to select γ such that $\chi^2/N \approx 1$. In these cases our criteria for selecting γ must be modified, so that applying tension introduces a misfit of similar magnitude to that inherent in fitting the data points. Choosing γ such that $\Delta\chi^2 = \chi^2(\gamma) - \chi^2(\gamma = 0) \approx \chi^2(\gamma = 0)/N$ will apply a suitable amount of tension without significantly degrading the fit to the data points.

Dataset Scaling

The choice of tension factor γ is dependent on the range of the dataset. Datasets with similar shapes but different x and y ranges therefore require different values of γ. This is not a problem when fitting a single dataset, but if many datasets with similar characteristics are to be fitted, it may be desirable to scale the datasets to a normalized coordinate system prior to fitting so that the same value of γ can be used in each case, for example, by scaling each dataset so the x and y ranges are unity. This will avoid the need to plot separate trade-off curves for each dataset.

If the x_i and y_i data points are multiplied by scale factors s_x and s_y, respectively, then the error-bars σ_i must also be multiplied by s_y to preserve the properties of the covariance matrix during the inversion. The fitted curve must then be scaled back by dividing x_q by s_x, f_q by s_y, and $\sigma_{f(x_q)}$ by s_y. The same procedure is necessary if the dataset has been scaled to make the linear spring approximation valid.

Un-modelled Error Sources

If the dataset contains outliers or the data error-bars have been underestimated, then χ^2/N will be greater than unity. If this is the case, then mapping the data errors into the parameter errors will underestimate the error in $f(x)$. Extra scatter in the data can be incorporated into the error envelope of $f(x)$ by scaling the error-bars of $f(x)$ by the observed scatter divided by the expected scatter within one knot point of x. The uncorrected error-bar of $f(x)$, $\sigma_{f(x)}$, is then scaled to give the corrected error-bar $\sigma'_{f(x)}$

$$\sigma'_{f(x)} = \sigma_{f(x)} \sqrt{\frac{1}{n}\sum_{i=1}^{n}\frac{(y_i - f(x_i))^2}{\sigma_i^2}} \quad \forall x - K < x_i < x + K, \tag{42}$$

where n is the number of data points within one knot point of x. This correction should only be applied if the correction factor is greater than unity, i.e. the fitted error-bars should not be decreased. Incidentally, the root mean square (rms) deviation σ_{rms} of the datapoints from the fitted curve within one knot point of x can be defined in a similar way

$$\sigma_{\mathrm{rms}_{f(x)}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\big(y_i - f(x_i)\big)^2} \quad \forall x - K < x_i < x + K. \tag{43}$$

This is an indication of the datapoint scatter and will normally be much greater than the standard error of the fitted curve ($\sigma_{f(x)}$).

Example Applications

Synthetic Data

Figure 3 shows application of the above method to a smoothly varying synthetic dataset. In the first case, where the data points are fairly evenly spread, there are enough constraints to give a reasonable fit, both with and without applying any tension. However, when the central data points are removed, the no-tension fit has a large departure from the overall data trend in order to preserve the minimum curvature condition. Applying tension such that $\chi^2/N = 1$, effectively removes this overshoot, while preserving the fit to the data points to within the error-bars.

Hawaiian Palaeomagnetic Inclination Record

As a more realistic example we use the record of inclination of the Earth's magnetic field from Hawaiian lava samples compiled by Teanby et al. (2002) shown in Fig. 4. This dataset provides a good test of the method as it has numerous outliers and a region of sparse data coverage between 25 and 34 kyr (thousands of years before present). It is not possible to fit a smooth curve through all the data simultaneously because of disagreement between data points with the same age. Therefore, the no-tension case has $\chi^2/N = 15.44$ (much greater than unity). With no tension the sparse data region suffers from large overshoots, but outside of this region the fit is good. A trade-off curve was plotted and a value of $\gamma = 0.0018$ was selected such that $\chi^2 \approx \chi^2(\gamma = 0) + \chi^2(\gamma = 0)/N$. This removed the problems of the overshoots in the sparse data region, while preserving the quality of the fit elsewhere. The fits outside the sparse data region are very similar in the tension and no-tension case, and are well constrained by the data points.

Conclusions

A method has been developed for fitting a smooth curve under tension through noisy data points. The choice of basis function is arbitrary, although we used the piecewise cubic polynomial of Constable and Parker (1992). Tension was applied by using a quadratic spring approximation, and the resulting set of linear equations were solved using inverse theory. The improvement over previous work is that the errors in the data are taken into account when fitting the curve and these errors are mapped through into errors in the fit. This is particularly important if the fitted curve is to form the basis of further analysis. Provision was made for increasing the error-bars on the fitted curve if there were significant numbers of outliers or discrepancies in the dataset. Also, application of fixed constraints to the fitted curve were made possible by incorporating the method of Lagrange multipliers into the inversion scheme. This allowed the value of the curve, gradient, or curvature to be fixed at known control points. An objective way of selecting the tension parameter was also presented.

Matlab routines for fitting constrained splines in tension using this method are available from http://www.atm.ox.ac.uk/user/teanby/software.html.

Acknowledgements This research was funded by the UK Particle Physics and Astronomy Research Council. The author thanks two anonymous reviewers for constructive comments that helped improve the manuscript.

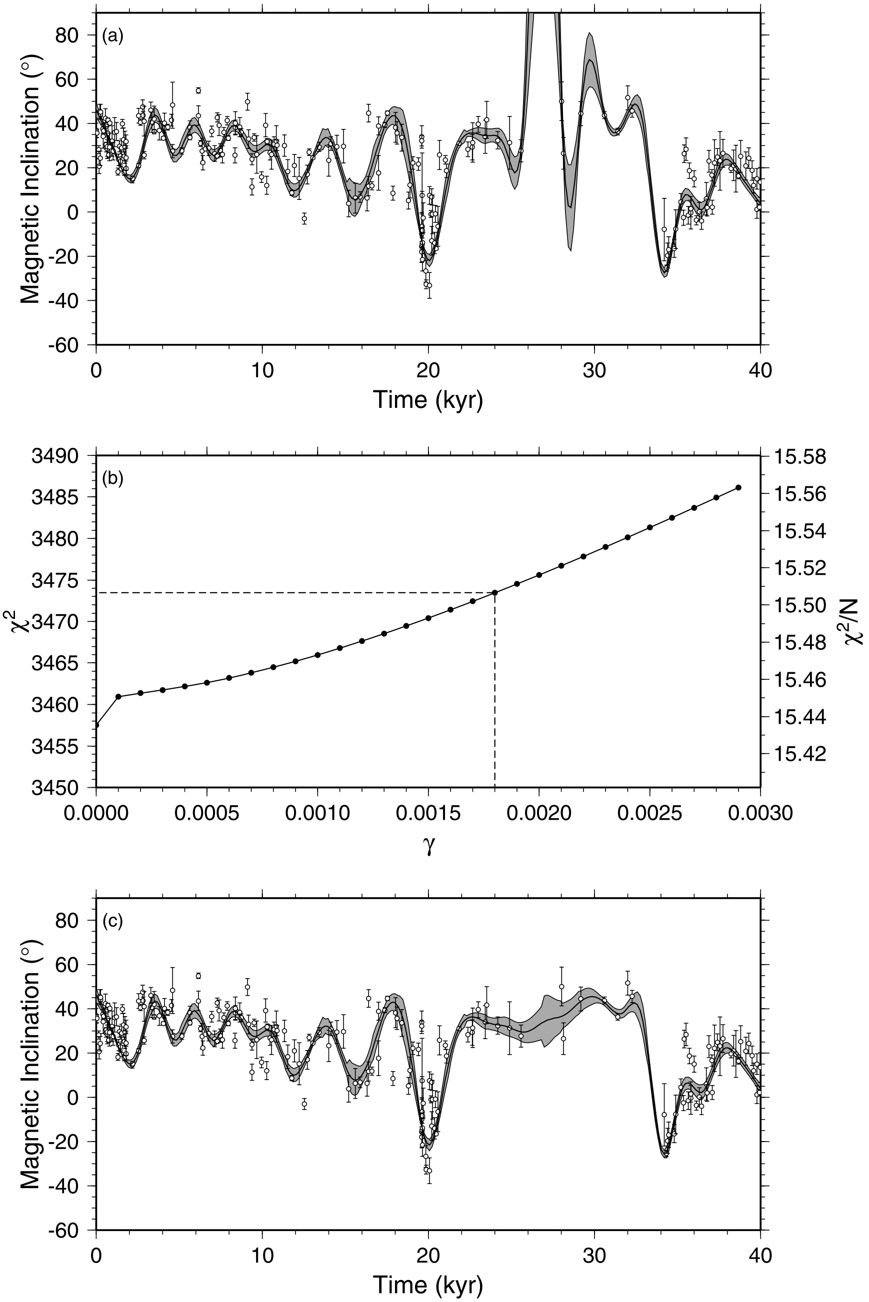

Fig. 4 **a** shows a cubic spline fit of the Hawaiian magnetic inclination record from Teanby et al. (2002) with zero tension. The curve fits the data very well, except from 25–34 kyr, where there are large overshoots caused by the sparse data coverage. **b** shows the effect of increasing tension γ on the χ^2 statistic. In **c** a tension of $\gamma = 0.0018$ has been applied (see the text for criteria), which has corrected the overshoots and preserved the fit elsewhere. The number of parameters $M = 37$ in each case, and the grey error envelope has been scaled to account for un-modelled error sources

References

Acton FS (1990) Numerical methods that work. Mathematical Association of America, Washington
Bevington PR, Robinson DK (1992) Data reduction and error analysis for the physical sciences, 2nd edn. WBC/McGraw-Hill, New York
Boas ML (1983) Mathematical methods in the physical sciences, 2nd edn. Wiley, New York
Cline AK (1974) Scalar- and planar- valued curve fitting using splines under tension. Commun Assoc Comput Mach 17(4):218–220
Constable C, Parker R (1992) Deconvolution of long-core palaeomagnetic measurements—spline therapy for the linear problem. Geophys J Int 104(3):453–468
de Boor C (1978) A practical guide to splines, revised edn. Springer, New York
Faliva M, Zoia MG (2002) On a partitioned inversion formula having useful applications in econometrics. Econom Theory 18(2):525–530
Gubbins D (2004) Time series analysis and inverse theory for geophysicists. Cambridge University Press, Cambridge
Hutchinson MF, Gessler PE (1994) Splines—more than just a smooth interpolator. Geoderma 62(1):45–67
Inoue H (1986) A least-squares smooth fitting for irregularly spaced data: finite-element approach using the cubic B-spline basis. Geophysics 51(11):2051–2066
Kent DV (1973) Post-depositional remanent magnetisation in deep-sea sediment. Nature 246(1):32–34
Lancaster P, Šalkauskas K (1986) Curve and surface fitting: an introduction. Academic Press, San Diego
Menke W (1989) Geophysical data analysis. In: Discrete inverse theory. International geophysics series, vol 45. Academic Press, San Diego, p 289
Mitášová H, Mitáš L (1993) Interpolation by regularized spline with tension: I. Theory and implementation. Math Geol 25(6):641–655
Press WH, Flannery BP, Teukolsky SA, Vetterling WT (1992) Numerical recipes, 2nd edn. Cambridge University Press, Cambridge
Reinsch CH (1967) Smoothing by spline functions. Numer Math 10(3):177–183
Reinsch CH (1971) Smoothing by spline functions. II. Numer Math 16(5):451–454
Renka RJ (1993) Algorithm-716 TSPACK—tension spline curve-fitting package. ACM Trans Math Softw 19(1):81–94
Sambridge M (1999) Geophysical inversion with a neighbourhood algorithm–I, searching a parameter space. Geophys J Int 138(2):479–494
Sandwell DT (1987) Biharmonic spline interpolation of GEOS-3 and SEASAT altimeter data. Geophys Res Lett 14(2):139–142
Schumaker LL (1993) Spline functions: basic theory. Krieger, Melbourne
Schweikert DG (1966) An interpolating curve using a spline in tension. J Math Phys 45:312–317
Smith WHF, Wessel P (1990) Gridding with continuous curvature splines in tension. Geophysics 55(3):293–305
Talmi A, Gilat G (1977) Method for smooth approximation of data. J Comput Phys 23(2):93–123
Tarantola A (1987) Inverse problem theory. Elsevier, New York
Teanby N, Gubbins D (2000) The effects of aliasing and lock-in processes on palaeosecular variation records from sediments. Geophys J Int 142(2):563–570
Teanby N, Laj C, Gubbins D, Pringle M (2002) A detailed palaeointensity and inclination record from drill core SOH1 on Hawaii. Phys Earth Planet Int 131(2):101–140
Wahba G (1990) Spline models for observational data. CBMS-NSF regional conference series in applied mathematics, vol 59. SIAM, Philadelphia
Wessel P, Bercovici D (1998) Interpolation with splines in tension: a Green's function approach. Math Geol 30(1):77–93
Wessel P, Smith WHF (1998) New, improved version of the generic mapping tools released. EOS Trans AGU 79(47):579
Wold S (1974) Spline functions in data analysis. Technometrics 16(1):1–11