# Soap film smoothing

Simon N. Wood[1], Mark V. Bravington[2] and Sharon L. Hedley[3]

[1]Mathematical Sciences, University of Bath, Bath BA2 7AY U.K.

[2]CSIRO Center for Mathematics and Information Science, Hobart, Tasmania, Australia.

[3]CREEM, University of St Andrews, UK*

s.wood@bath.ac.uk

February 13, 2008

**Abstract**

Conventional smoothing methods sometimes perform badly when used to smooth data over complex domains, by smoothing inappropriately across boundary features, such as peninsulas. Solutions to this smoothing problem tend to be computationally complex, and not to provide model smooth functions which are appropriate for incorporating as components of other models, such as generalized additive models, or mixed additive models. In this paper we propose a class of smoothers appropriate for smoothing over difficult regions of $\mathbb{R}^2$, which can be represented in terms of a low rank basis and one or two quadratic penalties. The key features of these smoothers are (i) that they do not 'smooth across' boundary features; (ii) that their representation in terms of a basis and penalties allows straightforward incorporation as components of GAMs, mixed models and other non-standard models; (iii) that smoothness selection for these model components is straightforward to accomplish in a computationally efficient manner via GCV, AIC or REML, for example; (iv) that their low rank means that their use is computationally efficient.

**Keywords:** Basis penalty smooth, differential equation smoothing, FELSPLINE, finite window smoothing, known boundary smoothing, spline.

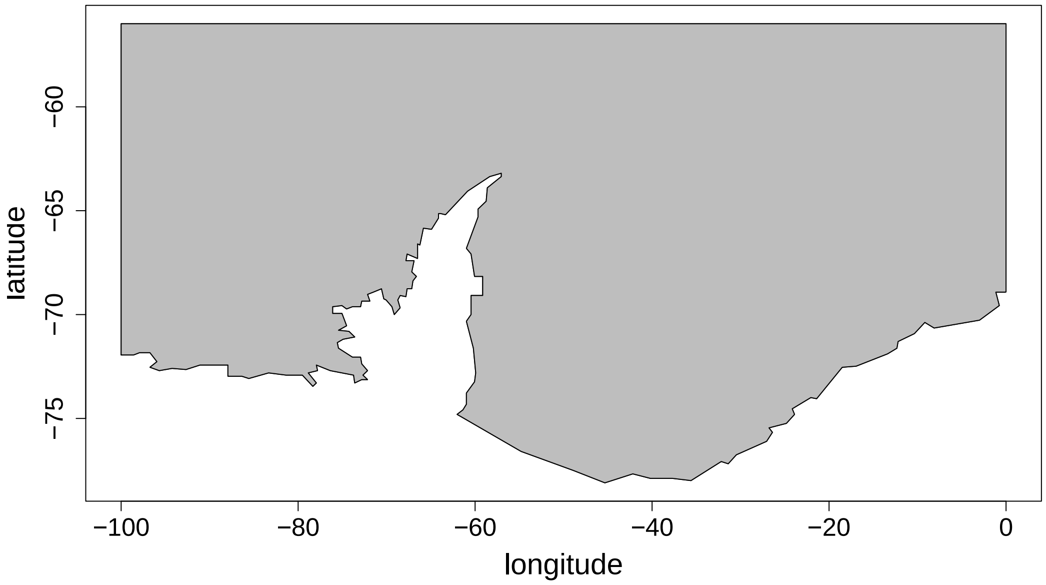

Figure 1: The Southern Ocean around the Antarctic Peninsula. The grey area represents a typical region over which spatially resolved whale abundance estimates are required. When modelling such abundances statistically, it is important not to 'smooth across' the Antarctic Peninsula: for whales, distances across land are very different to distances in water.

# 1 Introduction

This paper is about smoothing, primarily over two dimensional domains, when the boundary of the domain matters. We are interested in two issues: how to estimate a smooth function from scattered noisy data and known boundary values, and how to smooth in a way that respects complex boundary features (including when the boundary values are unknown). Figure 1 illustrates a domain that is problematic for conventional methods: the ocean around the Antarctic peninsula. Conventional smoothing tends to display 'leakage' across the peninsula, with estimates on either side of the peninsula linked inappropriately. To illustrate this issue, Ramsay (2002), provided a simple test example which perfectly demonstrates the problem, a modification of which is shown in figure 3: most smoothing methods have considerable difficulty in reconstructing this test function from samples, because of leakage across the domain boundary.

In assessing what smooth models might be appropriate, we are motivated by four basic considerations.

1. Given the spatial nature of this sort of smoothing problem, the smooths should be rotationally invariant.
2. The notion of smoothness adopted should be such that a 'completely smooth' function can be constructed which exactly meets (almost) any boundary condition on function values.
3. Many domains feature holes or islands, and the method should be general enough to deal with these.
4. Smoothers must be capable of straightforward incorporation as components of other models and must be computationally cheap enough for routine use.

Point 2 requires some further explanation. Most smoothing methods imply a measure of function wiggliness, which acts as a penalty on overly wiggly functions when estimating smooth functions from data. If this penalty is zero then the function is 'completely smooth' (e.g. straight lines are completely smooth according to a cubic spline penalty). Our requirement is that it should be possible to produce a function over the smoothing domain which can meet any arbitrary boundary condition (subject to some continuity restrictions) and have a zero penalty. Failure to meet this condition implies situations in which the smoothest possible model has a penalty greater than zero. This is problematic when using mixed model or (many) Bayesian approaches to smoothing. Such methods put a negative exponential prior on function wiggliness, and if the penalty can not be zeroed while meeting the boundary conditions then this prior will ascribe highest probability to model coefficient values that are actually impossible. The property of being able to zero the penalty is also desirable when boundary conditions are unknown. Otherwise, as we approach the completely smooth/zero penalty limit, the smooth will tend to show the sort of boundary-leakage problems we are seeking to avoid. For example, if we use a thin plate spline penalty around the

[*] Current address: The Schoolhouse, Denhead, St Andrews, KY16 8PA UK

Antarctic peninsula, then the model zeroing the penalty is a plane over the spatial domain, *even if the penalty is only integrated over the ocean*. This inevitably suggests a degree of correspondence between the two sides of the peninsula that is unreasonable in the biological situations motivating our work.

To meet objective 4 in practice, we aim to produce smoothing methods that can be implemented using a basis of relatively low rank, with a quadratic penalty measuring departure from smoothness (wiggliness). Such basis-penalty smoothers are easily incorporated as components of a wide variety of statistical models, and have available relatively reliable and computationally efficient methods for smoothness selection. For example such smooths are easily incorporated as components of generalized linear models to yield 'semi-parametric models' and 'generalized additive models', which can be estimated by penalized likelihood maximization with smoothness selected by GCV, AIC or similar criteria (see, for example, Kim and Gu, 2004 or Wood, 2006a). Similarly, the basis-penalty representation allows straightforward incorporation of such smooths as mixed model components, with the quadratic penalty matrix playing the role of a random effects inverse covariance matrix (see Ruppert et al., 2003 for an overview, and Fahrmeir et al. 2004, or Wood 2004 for general strategies). An important motivation for us is the need to embed penalized models of smooth functions over awkward domains within a somewhat complex model of Minke whale sightings around the Antarctic peninsula: for this application smoothing methods that can not be represented using a basis and penalties are quite intractable.

The literature on smoothing is large (see, for example, Wahba, 1990; Hastie and Tibshirani 1990; Green and Silverman, 1994; Wand and Jones, 1995; Bowman and Azzalini, 1997; Gu, 2002; Ruppert et al. 2003 and Wood, 2006a, for overviews of just some of the approaches), but the methods of Ramsay (1999, 2002) and Stone (1988) provide obvious starting points when attempting to build models incorporating finite window smoothers as components. Stone developed a finite window version of thin plate splines (Duchon, 1977), in which the thin plate spline penalty is evaluated only over the domain of interest. While preferable to a conventional thin plate spline for smoothing over complicated domains, this does not meet our objective 2, would require extension to deal with holes/islands and does not deal with the known boundary case. Ramsay's approach is demonstrably a huge improvement on conventional smoothers, such as thin plate splines, over difficult domains. It uses a smoothing objective that is compatible with our aims 1 and 3, but the method involves a quite complex computational strategy incompatible with our aim 4. To arrive at the computational approach, Ramsay uses a smoothing objective which would be compatible with our aim 2, but only under the undesirable assumption that, at the boundary, the gradient of the estimated function is zero along normals to the boundary. This implies that contours of the estimated function must meet the boundary at right angles, which is a strong assumption (see section 5.2).

The rest of this paper is structured as follows. We first provide an intuitive motivation of our proposed soap film smoothers, then we prove some useful results for the known boundary case which motivate the practical soap-film smoothing methods discussed subsequently. We then demonstrate the smoothers' performance on simulated problems, before providing two example applications.

## 2 The motivating physical model

The basic physical motivation for the proposed smoothers is as follows. Consider a loop of wire following the boundary of the region, $\Omega$, of the $x-y$ plane over which we are interested in smoothing. The vertical displacement of the loop above the plane gives the known function values at the boundary. An appropriate definition of a 'completely smooth' function over this domain can be obtained by considering a soap film supported by this boundary wire (in zero gravity). If we make the assumption that the vertical displacement of the wire is small then the height of the soap film, inside the boundary, is given by the function $f$ satisfying

$$\frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} = 0$$

and the boundary conditions: i.e. the soap film adopts a minimum surface tension configuration. Now, in order to smooth data within the domain, the soap film should distort smoothly from its minimum energy configuration by moving vertically towards the data. An appropriate measure of the total degree of distortion would be:

$$J_\Omega(f) = \int_\Omega \left( \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} \right)^2 dxdy.$$

Notice that, although isotropic, this differs from the usual thin plate spline penalty functional (Duchon, 1977, or see e.g. Green and Silverman, 1994) in three respects: it is only integrated over $\Omega$, rather than the whole $x, y$ plane, there is no mixed second derivative term, and the sum of the second derivative terms is squared, rather than the terms being separately squared. The latter feature allows the second derivatives with respect to $x$ and $y$ to be traded off against each other, so that the space of functions for which $J_\Omega(f)$ is zero is infinite dimensional: this is what allows functions with zero penalty to be curved enough to meet any boundary condition (aim 2 in the introduction).

So, if we have $n$ data points, $z_k$, which are noisy observations of $h(x_k, y_k)$ where $h$ is a smooth function over the domain, we might seek to estimate $h$ by minimizing (subject to the known boundary conditions)

$$\sum_{i=1}^{n} \{z_i - f(x_i, y_i)\}^2 + \lambda J_\Omega(f) \tag{1}$$

w.r.t. $f$. Here $\lambda$ is a tuneable smoothing parameter. As we will see, there is no particular need to employ a least squares loss function in the objective: any loss measure which depends on $f$ only via evaluations of $f$ at a finite number of points will do. In the next section we provide a characterization of the minimizer of (1) which allows it to be computed to arbitrary accuracy by relatively simple numerical methods.

# 3 Theory of soap film smoothers

**Theorem 1 (Soap film interpolation).** Consider a smooth function $f^*(x,y)$ over the $x,y$ plane. Let $B$ be a collection of closed loops in the $x,y$ plane, such that no two loops intersect and one 'outer' loop encloses all the others. Let $\Omega$ be the region made up of all $x,y$ points which are interior to an odd total number these loops. Suppose that $f^*(x,y)$ is known exactly on $B$, and that $z_k = f^*(x_k,y_k)$ are observations of $f^*$ at $n$ locations $x_k,y_k$ within $\Omega$. Let $f(x,y)$ be the function which

1. interpolates the known $f^*$ values on $B$ and the $z_k$'s at the $n$ points $(x_k,y_k)$;
2. satisfies $\partial^2 f/\partial x^2 + \partial^2 f/\partial y^2 = 0$ on $B$; and
3. minimizes
$$J_\Omega(f) = \int_\Omega \left( \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} \right)^2 dxdy.$$

$f$ is the function, meeting condition 1, such that

$$\frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} = \rho, \tag{2}$$

where

$$\frac{\partial^2 \rho}{\partial x^2} + \frac{\partial^2 \rho}{\partial y^2} = 0 \tag{3}$$

except at the $x_k,y_k$ points, and $\rho = 0$ on $B$. Equations (2) and (3) are the Poisson and Laplace equations, respectively.

*Proof.* Various proofs are possible. The following gives some insight into *why* the interpolant has the given form.

Let $R$ be a rectangle in the $x,y$ plane, which encloses $B$ and the boundary of which never touches $B$. Assume w.l.o.g. that each side of $R$ is aligned with either the $x$ or $y$ axis. Now impose a regular rectangular grid on $x,y$, aligned with the sides of $R$ and with grid spacing of $\Delta$ (in both $x$ and $y$ directions). Let the grid points be indexed $i$ in the $y$ direction and $j$ in the $x$ direction. The derivatives in the penalty functional $J_\Omega$ can now be approximated by symmetric finite differences on the grid. For example

$$\frac{\partial^2 f_{i,j}}{\partial x^2} \simeq \frac{f_{i,j+1} - 2f_{i,j} + f_{i,j-1}}{\Delta^2}$$

with equality in the limit as $\Delta \to 0$, where $f_{i,j}$ denotes $f(x,y)$ at grid point $i,j$.

Now let $\mathbf{f} = [f_{1,1}, f_{1,2}, \ldots, f_{2,1}, f_{2,2}, \ldots]^{\mathrm{T}}$ be the vector of $f(x,y)$ evaluated at all grid points. Suppose that $\mathbf{f}_{xx}$ and $\mathbf{f}_{yy}$ are the corresponding second derivative vectors of $f$ w.r.t. $x$ and $y$ respectively. The finite differencing approximation of the derivatives can be written as $\mathbf{f}_{xx} = \mathbf{D}_x\mathbf{f}$ and $\mathbf{f}_{yy} = \mathbf{D}_y\mathbf{f}$, where the non-zero elements of matrices $\mathbf{D}_x$ and $\mathbf{D}_y$ are the coefficients involved in finite differencing (plus constants determined by the assumptions made at the boundary of $R$, which turn out to be immaterial).

Consider the extended penalty functional

$$J_R(f) = \int_R \left( \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} \right)^2 dxdy.$$

It is easy to see that,

$$J_R(f) \simeq \mathbf{f}^{\mathrm{T}} (\mathbf{D}_x + \mathbf{D}_y)^{\mathrm{T}} (\mathbf{D}_x + \mathbf{D}_y) \mathbf{f} \Delta^2,$$

with equality in the limit as $\Delta \to 0$.

Now we seek the values of $\mathbf{f}$ minimizing $J_R$, subject to the interpolation constraints. Clearly the minimizing $\mathbf{f}$ must satisfy each of the system of equations

$$(\mathbf{D}_x + \mathbf{D}_y)^{\mathrm{T}} (\mathbf{D}_x + \mathbf{D}_y) \mathbf{f} = \mathbf{0} \tag{4}$$

which does not correspond to a point on $B$ or an $x_k, y_k$ point. Correspondence between grid points and boundary points can be established as follows. A $\Delta \times \Delta$ square with sides aligned with the $x, y$ co-ordinate system is centred on each gridpoint: any grid point whose square includes part of $B$ or an $x_k, y_k$ is treated as boundary point. The error associated with this gridding tends to 0 as $\Delta \to 0$.

Further progress relies on the fact that, away from boundaries of $R$, partial symmetry of $\mathbf{D}_x$ means that $\mathbf{D}_x^{\mathrm{T}}$ acts exactly as $\mathbf{D}_x$ in producing approximate second derivatives of any quantity gridded in the same way as $\mathbf{f}$, i.e. if $\mathbf{g}$ is a vector of any smooth function's values, gridded in the same way as $\mathbf{f}$, then both $\mathbf{D}_x^{\mathrm{T}}\mathbf{g}$ and $\mathbf{D}_x\mathbf{g}$ will yield the corresponding vector of finite difference approximations to the functions's second derivatives w.r.t. $x$, for points sufficiently far from the boundary of $R$. The same holds for $\mathbf{D}_y^{\mathrm{T}}$. Hence in the limit as $\Delta \to 0$, in which the approximation to $J_R$ is exact, (4) becomes:

$$\frac{\partial^4 f}{\partial x^4} + \frac{\partial^4 f}{\partial x^2 y^2} + \frac{\partial^4 f}{\partial y^2 x^2} + \frac{\partial^4 f}{\partial y^4} = 0$$

except at the $x_k, y_k$ (the boundary condition ensures that it also holds on $B$), and this can be re-written as

$$\frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} = \rho, \text{ where } \frac{\partial^2 \rho}{\partial x^2} + \frac{\partial^2 \rho}{\partial y^2} = 0$$

except at the $x_k, y_k$ points. From the boundary condition, $\rho = 0$ on $B$. Existence, uniqueness and continuity of solutions $f$ and $\rho$ over $\Omega$, subject to the given boundary conditions, follow from known properties of the Laplace and Poisson equations (see, e.g. Strauss, 1992 or Evans, 1998 section 2.2).

It remains to establish that the minimizer of $J_R$ minimizes $J_\Omega$ when considered over $\Omega$ alone, but this is straightforward. If $f$ were not the minimizer of $J_\Omega$ then we could replace $f$ by the minimizer of $J_\Omega$ over $\Omega$ and thereby reduce $J_R$, since $J_R$ is simply $J_\Omega$ integrated over a larger domain, and since $\left(\partial^2 f/\partial x^2 + \partial^2 f/\partial y^2\right)^2 = 0$

on $B$, so that the function inside $\Omega$ has no influence on $\left(\partial^2 f/\partial x^2 + \partial^2 f/\partial y^2\right)^2$ outside $\Omega$. But this reduction in $J_R$ leads to a contradiction unless the same $f(x, y)$ over $\Omega$ minimizes both $J_\Omega$ and $J_R$. □

**Remarks.**

1. The Laplace and Poisson equations are among the most well studied in the theory of PDEs, due in part to their frequent occurrence in physics. See Evans, (1998 section 2.2) or Strauss (1992 chapter 6) for theoretical introductions, and Press et al. (1992 chapter 19) or Strauss (1992 chapter 8) for an introduction to numerical solution methods.

2. Essentially the same construction can be used to derive interpolants over smoothly bounded domains in $\mathbb{R}^k$, where $k$ is any positive integer. However as $k$ increases above 2 it is necessary to increase the order, $m$ say, of the derivatives in the wiggliness measure, if the defining PDEs and boundary conditions are to constitute a well posed problem. For example, the singularities in the Laplace equation in 3 dimensions are not integrable, so a 3 dimensional soap film interpolant using a second derivative based $J_\Omega$ is not practical. A third derivative based penalty would not cause a problem, but the corresponding PDEs would be much less well studied, and might be harder to solve numerically, as well as requiring extra boundary conditions. A formal theory for the general, $k > 2$, case could be constructed using general Sobolev space theory (see Evans, 1998, Chapter 5), but since we know of no applications for $k > 2$ we do not pursue this further.

3. The structure of the theorem's proof immediately suggests the form of the defining equations for a 'soap film' smoother in 2 dimensions but using a third derivative based $J_\Omega$. A referee has pointed out that this might be helpful for establishing convergence rates for the soap film smoother.

4. Over a one dimensional region the soap film penalty would become the cubic spline penalty $\int f''(x)^2 dx$. In that case the soap film interpolant of $x_k, z_k$ data on $[a, b]$ is defined by $\partial^2 f/\partial x^2 = \rho$ where $\partial^2 \rho/\partial x^2 = 0$, except at the $x_k$, $f(x_k) = z_k\ \forall k$ and there are boundary conditions that $f(a)$ and $f(b)$ are known while $f''(a) = f''(b) = 0$. So $\rho$, the second derivative of $f$, is piecewise linear with derivative discontinuities at the $x_k$, but in that case $f$ is a natural cubic spline interpolant. See Wahba (1990) for more on splines.

5. An alternative proof can be constructed by writing any alternative interpolant as $\tilde{f} = f + h$ where $f$ is the soap film interpolant. Writing $g_{ab}$ for $\partial^2 g/\partial a \partial b$, we have $J_\Omega(\tilde{f}) = \int\int_\Omega (f_{xx} + f_{yy})^2 dxdy + \int\int_\Omega (h_{xx} + h_{yy})^2 dxdy + 2\int\int_\Omega h_{xx}(f_{xx} + f_{yy}) + h_{yy}(f_{xx} + f_{yy}) dxdy$. Repeated integration by parts establishes that, under the boundary conditions given in the theorem, the final integral is zero, from which the theorem follows easily. This is somewhat similar to the integration by parts proof of the cubic spline's smoothest interpolation property.

6. The theorem's characterization can be generalized to the case where the $z_k$ are observed with noise and we wish to smooth rather than interpolate. This is the subject of the following lemma.

**Lemma 1 (Soap film smoothing).** Let the setup be exactly as for Theorem 1, except that the $z_k$ are now measured with error, and $\mathbf{f}$ is now the vector $\mathbf{f} = [f(x_1, y_1), f(x_2, y_2), \dots, f(x_n, y_n)]^{\mathrm{T}}$. The function, $f(x, y)$, minimizing

$$\|\mathbf{z} - \mathbf{f}\|^2 + \lambda_f J_\Omega(f) \tag{5}$$

subject to the known conditions on $B$ must satisfy (2) and (3).

*Proof.* If the minimizing function were $\tilde{f}$, different to $f$, then we could interpolate the values in $\mathbf{f}$ using a soap film interpolant, thereby leaving $\|\mathbf{z} - \mathbf{f}\|^2$ unchanged but reducing $J_\Omega$, by the minimum $J_\Omega$ interpolant property of the soap film interpolant. i.e. there is a contradiction unless $\tilde{f} = f$. □

The proof generalizes trivially to the case where $\|\mathbf{z} - \mathbf{f}\|^2$ is replaced by any measure of lack of fit, $D(\mathbf{z}, \mathbf{f})$, which depends on $f$ only through $\mathbf{f}$.

# 4 Actual smoother construction

The characterization of the minimizer of (5) in terms of solutions to (2) and (3) is the key to convenient computation of soap film smoothers *and* to deriving the basis-penalty representation that is needed if these smoothers are to be useful as components of mixed models, additive models and other more general models. Here we provide the main results, while the appendix fills in some of the computational details required for practical computation.

Let $\rho_k(x, y)$ denote the function which is zero on $B$, satisfies (3) everywhere in $\Omega$ except at the single point $x_k, y_k$ (the $k^{\text{th}}$ data location) and satisfies $\int_\Omega \rho_k(x, y) dxdy = 1$. Then any function which satisfies (3) everywhere in $\Omega$ except at the set of points $\{x_k, y_k : k = 1, \dots, n\}$, can be written as

$$\rho(x, y) = \sum_{k=1}^{n} \gamma_k \rho_k(x, y).$$

where the $\gamma_k$ are coefficients, by the linearity of (3).

It is then straightforward to confirm that

$$J_\Omega(f) = \boldsymbol{\gamma}^{\mathrm{T}} \mathbf{S} \boldsymbol{\gamma},$$

where $\mathbf{S}$ is a matrix of fixed coefficients given by

$$\mathbf{S}_{i,j} = \int_\Omega \rho_i(x, y) \rho_j(x, y) dxdy.$$

The function $f$ can also be written in terms of the $\gamma_k$. Let $a(x, y)$ be the solution to (2) with $\rho(x, y) = 0 \; \forall \; x, y$, and subject to the boundary condition that $f(x, y)$ is known on $B$. Now define $g_i(x, y)$ as the solution to (2) with

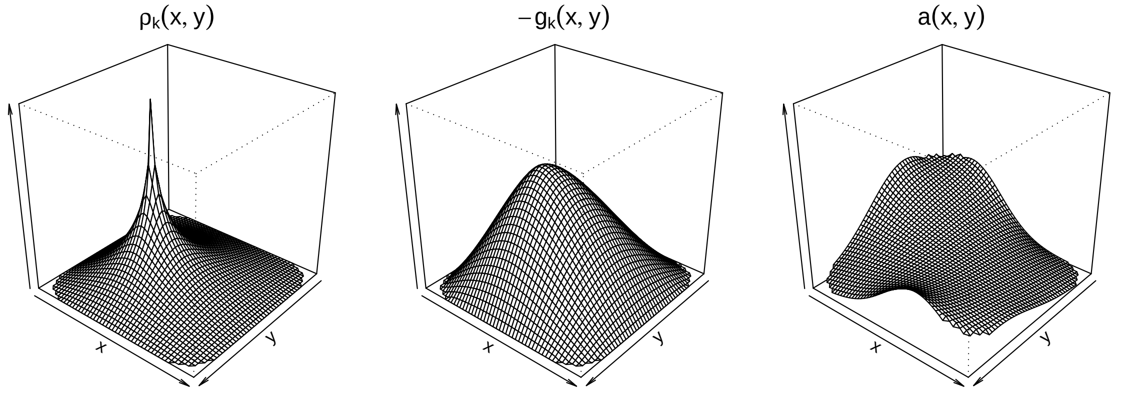

Figure 2: Example basis functions for a soap film smoother (known boundary case), as introduced in section 4. The basis functions relate to the smoothing problem considered in section 5.3, of reconstructing functions over a smoothly bounded subregion of the unit square. In all cases the basis functions have been evaluated by solving their defining PDEs on a $50 \times 50$ grid which is also the plotting grid. The left panel shows an example $\rho_k(x, y)$ for the case where $x_k^* = 0.4$ and $y_k^* = 0.6$. Notice the spike corresponding to $\rho_k(x, y)$'s integrable singularity at $x_k^*, y_k^*$. The $\rho_k(x, y)$s are not basis functions of the soap film itself, but are used to define the actual basis functions $g_k(x, y)$, and the penalty matrix $\mathbf{S}$. The middle panel shows the negative of the basis function $g_k(x, y)$ corresponding to the $\rho_k(x, y)$ shown in the left panel. The right panel shows $a(x, y)$ for the known boundary example in section 5.3.

$\rho(x, y) = \rho_i(x, y)$ and the boundary condition that $f$ is zero on $B$; linearity of (2) implies that the soap film smoother can be written as

$$f(x, y) = a(x, y) + \sum_{k=1}^{n} \gamma_k g_k(x, y). \tag{6}$$

Efficient numerical solution of the Poisson (2) and Laplace (3) equations is an exceedingly well studied problem. As a result, evaluation of the $a$, $g_k$, $\rho_k$ and hence $\mathbf{S}$ is computationally straightforward: the appendix provides further details, while figure 2 plots some evaluated basis functions.

Given the basis and penalty, (5) becomes the standard penalized regression problem of minimizing

$$\|\mathbf{z} - \mathbf{a} - \mathbf{X}\boldsymbol{\gamma}\|^2 + \lambda_f \boldsymbol{\gamma}^{\mathrm{T}} \mathbf{S} \boldsymbol{\gamma}$$

w.r.t. $\boldsymbol{\gamma}$, where $\mathbf{a}^{\mathrm{T}} = [a(x_1, y_1), a(x_2, y_2), \ldots, a(x_n, y_n)]$ and $X_{i,j} = g_j(x_i, y_i)$. Several computationally efficient $\lambda_f$ selection methods for such problems are also routinely available (Wood, 2006a, is one reference among many). The evaluated basis functions and penalty make it easy to incorporate soap film smooths as components of other models.

For large data sets it is unlikely to be practically worthwhile to work with the full soap film basis, and a penalized regression smoothing approach is computationally preferable. In this case one would select a moderate sized set of $x, y$ values, $\{x_k^*, y_k^* : k = 1, \ldots, K\}$, spread 'nicely' among the $x_k, y_k$ values ($K$ is substantially less than the number of data, $n$, of course). The basis functions and penalty matrix are then evaluated as if the $x_k^*, y_k^*$ were the data locations. If $g_k^*(x, y)$ is the $k^{\text{th}}$ resulting (regression) basis function, then the only substantive change in the fitting problem is that $\mathbf{X}$ is now an $n \times K$ matrix where $X_{i,j} = g_j^*(x_i, y_i)$. Given that, for any practically useful model, the degrees of freedom will be suppressed to well below $n$ by the penalty term, the reduction in basis dimension typically has very limited effect on the fitted model, provided only that we do not make $K$ too small (see the appendix). In what follows we will always use $x_k^*, y_k^*$ to denote the points at which (3) does not hold (the 'knots'), even when this is all the observation points (but we will not 'star' the corresponding basis functions).

## 4.1 Unknown boundaries

While the known boundary case is useful, there are many applications in which data lie within a problematic boundary region, but we do not have special knowledge about the function's value on that boundary. In this case it is natural to model the function values on the boundary. Purely for simplicity of presentation, assume that $B$ consists of a single closed loop, parameterized in terms of $r$, the distance along the loop from some arbitrary fixed starting point on $B$: hence the co-ordinates of B are given by $\{x_B(r), y_B(r)\}$, say. Now define the 'boundary function' $f_b(r) = f(x_B(r), y_B(r))$. One approach is to model $f_b(r)$ using a cyclic penalized regression spline smoother in $r$ (e.g. Wood, 2006a section 4.1.3). Suppose that such a smoother has basis expansion

$$f_b(r) = \sum_{j=1}^{J} \alpha_j \delta_j(r), \tag{7}$$

where the $\alpha_j$ are parameters and the $\delta_j(r)$ are known basis functions. Associated with the smoother is some penalty functional $J_b(f_b) = \boldsymbol{\alpha}^{\text{T}} \mathbf{S}_b \boldsymbol{\alpha}$, where $\mathbf{S}_b$ is a matrix of known coefficients. Typically the functional would be something like $\int_B f_b''(r)^2 dr$.

It can then be shown that the boundary condition induced function $a(x, y)$ has a basis function representation

$$a(x, y) = \sum_{j=1}^{J} \alpha_j a_j(x, y)$$

where $a_j(x, y)$ is the solution to (2) with $\rho(x, y) \equiv 0$ and the boundary condition that results from setting $\alpha_j = 1$ and $\alpha_i = 0 \; \forall \; i \neq j$ in (7). Again, the $a_j(x, y)$ are easily evaluated, using the same numerical method as for the $g_k(x, y)$ (see the appendix for details).

Given the basis functions and penalty for the boundary model, and defining $A_{i,j} = a_j(x_i, y_i)$, then a suitable fitting objective for this smoother would be to minimize

$$\|\mathbf{z} - \mathbf{A}\boldsymbol{\alpha} - \mathbf{X}\boldsymbol{\gamma}\|^2 + \lambda_f \boldsymbol{\gamma}^{\text{T}} \mathbf{S} \boldsymbol{\gamma} + \lambda_b \boldsymbol{\alpha}^{\text{T}} \mathbf{S}_b \boldsymbol{\alpha}$$

w.r.t. $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$. Here $\lambda_f$ and $\lambda_b$ are smoothing parameters (which can be selected by GCV, REML or similar). Although not essential, it would be usual to use a regression basis rather than a full basis for construction of $\mathbf{X}$, to avoid rank deficiency problems at low $\lambda_{f/b}$.

The presentational restriction to the case where $B$ is a single closed loop can be relaxed without introducing any extra technical difficulties: a cyclic smooth can be used for each loop of $B$ (each with an associated penalty). Similarly, dealing with the case in which $f$ is known along only part of $B$ involves extra computer coding, but requires no additional theory.

## 4.2 Variance estimation

Once expressed in terms of a basis and quadratic penalties, soap film smoothers can be treated just like any other such smooth. For example, writing $\boldsymbol{\beta}$ for the vector of all smooth coefficients and $\mathbf{S}_T$ as the 'total penalty matrix' so that $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{S}_T\boldsymbol{\beta} \equiv \lambda_f\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{S}\boldsymbol{\gamma} + \lambda_b\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{S}_b\boldsymbol{\alpha}$, then the Bayesian covariance matrix (Silverman, 1985) for $\boldsymbol{\beta}$ is given by

$$\mathbf{V}_{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X} + \mathbf{S}_T)^{-1}\sigma^2$$

where $\sigma^2$ is the response variable variance. The effective number of degrees of freedom for the smooth is given by

$$\mathrm{tr}\big\{(\mathbf{X}^{\mathrm{T}}\mathbf{X} + \mathbf{S}_T)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{X}\big\}.$$

Wood (2006a) is one source for the derivation of these results, and the generalizations beyond penalized least squares estimation.

Given a ready means for evaluating the basis functions of the soap film smoothers it is also very easy to derive variances of any quantity that can be computed from the fitted model. For convenience write $b_i(x, y)$ for the $i^{\mathrm{th}}$ basis function (each $b_i$ being one of the $a_k$'s or $g_k$'s). Now consider the example of the predicted value of $f$ at a new data point $\mathcal{X}, \mathcal{Y}$. The predicted value is given by

$$\mathbf{b}^{\mathrm{T}}\boldsymbol{\beta}$$

where $\mathbf{b} = [b_1(\mathcal{X}, \mathcal{Y}), b_2(\mathcal{X}, \mathcal{Y}), \ldots]^{\mathrm{T}}$. The posterior variance of the prediction is simply

$$\mathbf{b}^{\mathrm{T}}\mathbf{V}_{\boldsymbol{\beta}}\mathbf{b}.$$

The generalizations to vector predictions or other linear functionals of $f$ are obvious.

The preceding expressions are usually employed in a smoothing parameter conditional manner, by plugging the estimated smoothing parameters in to the expression for $\mathbf{S}_T$. This neglects smoothing parameter uncertainty, thereby potentially underestimating variances. See Wood (2006b) for a possible fix.

## 4.3 Comparison with alternative approaches

With the replacement of the known boundary values by the rather strong boundary condition that the normal derivative of $f$ must be zero on the boundary of $\Omega$, (1) is the objective used to motivate the FELSPLINE (finite element spline) approach: see Ramsay (2002) expression (3). Ramsay seeks an approximation to the minimiser in a finite element space (see Strauss, 1992, section 8.5 for a brief introduction) of functions in which the second derivatives in the penalty $J_\Omega$ are not well defined. To do this he finds a condition that must be satisfied by the minimizer of (1), subject to his boundary condition, but which only involves first derivatives. Specifically, writing $a_x$ for $\partial a/\partial x$ and $a_y$ for $\partial a/\partial y$, he seeks functions $f$ and $g$ such that

$$\sum_i f(x_i,y_i)u(x_i,y_i) - \lambda\int_\Omega g_x u_x + g_y u_y dxdy = \sum_i u(x_i,y_i)z_i, \text{ and } \int_\Omega gvdxdy + \int_\Omega f_x v_x + f_y v_y dxdy = 0$$

for all functions $u, v$ in the space. $f$ and $u$ are required to meet the normal derivative zero boundary conditions (this is Ramsay's expression 11). This new problem can conveniently be solved by finite element methods, and the resulting $f$ is the FELSPLINE. The objective reformulation trick is a very neat way of enabling an approximation to the minimizer of (1) to be found in a convenient finite element space, without the need to obtain derivatives undefined in that space, but it embeds a strong boundary condition into the method: contours of the FELSPLINE must meet the boundary orthogonally. In consequence FELSPLINE smooths towards the zero function, in contrast to the soap film smoother which smooths towards the, much less restrictive, null space of functions of the form of the $a(x, y)$ term in (6). As we show in section 5.2, this difference has quite big practical implications.

The other major difference between FELSPLINE and soap film smooths is the basis-penalty representation of the latter, which allows convenient incorporation of soap films as components of other models, ready access to computationally efficient smoothing parameter selection methods, and straightforward variance calculations. Purely practically, successive over relaxation (SOR) on a grid (see the appendix) is very much easier to code than the finite element methods required for FELSPLINE.

In place of (1), Stone's (1988) method used the objective function

$$\sum_{i=1}^n \{z_i - f(x_i,y_i)\}^2 + \lambda\int_\Omega \left(\frac{\partial^2 f}{\partial x^2}\right)^2 + 2\left(\frac{\partial^2 f}{\partial x\partial y}\right)^2 + \left(\frac{\partial^2 f}{\partial y^2}\right)^2 dxdy.$$

Lacking any more direct characterization of the minimizing $f$, he sought an approximate minimizer directly in a space spanned by tensor products of B-splines. Again this lacks the relative computational simplicity of the soap film, but a more serious objection is that the penalty still shrinks $f$ towards a plane over the $x, y$ domain, and this is not an appropriate smoothest model for many situations, including those discussed in sections 5.2, 5.1 and 6.2.

Very recently Eilers (2006) has proposed what might be viewed as an approximate version of Stone's method based on tensor products of P-splines (Eilers and Marx, 1996). It involves manually modifying the P-spline smoothing penalty to avoid enforcing smoothness across boundary features. This seems quite effective, but again smooths

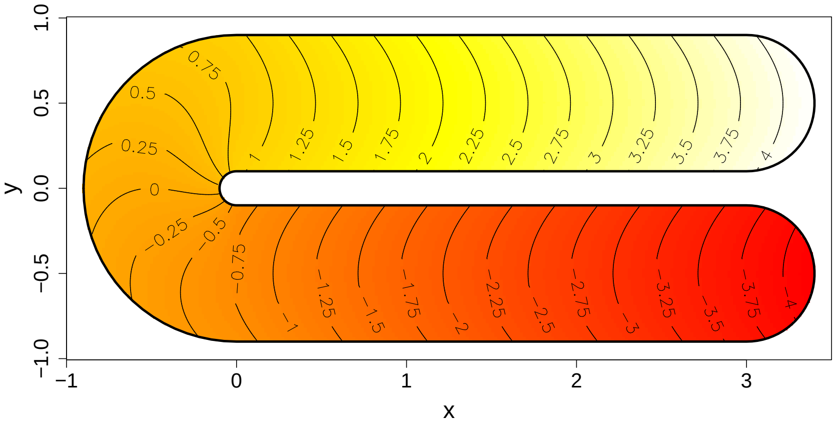

Figure 3: The horseshoe shaped test region used in section 5.1, with the test function shown as a colour map and contour plot over the region.

towards a plane over the domain. Eilers also examines a conformal mapping approach in which the original domain is mapped onto a rectangle, before smoothing. This solves some problems, but means that it is no longer clear what the smoothness penalty is measuring (it will be different for each domain). Wang and Ranalli (2007), take a somewhat similar approach, by modifying the measure of distance employed in the radial basis function components of a thin plate regression spline. Rather than employ conventional Euclidian distances between points, they employ shortest distances within the domain. For many applications this is intuitively appealing, but it is again not clear what the resulting smoothness penalty means in terms of function shape. Wang and Ranalli do not modify the linear basis components of the thin plate spline, so this method again smooths towards a plane over the domain.

## 5 Some simulated examples

This section compares soap film smoothers, thin plate splines and Ramsay's (2002) FELSPLINE method, using simulated datasets constructed to highlight the relative strengths and weaknesses of the different approaches.

### 5.1 A modified Ramsay horseshoe

This example modifies the simulation test presented in Ramsay (2002, section 5.2). Figure 3 shows the test function, which is similar to Ramsay's except that it bulges across the test region. The modification is a slightly more interesting test problem, since the true function can not be well approximated by $a(x, y)$ from (6), alone.

Around 600 function values were sampled from $x, y$ points randomly located within the domain of interest, and were perturbed by Gaussian noise with a standard deviation of 0.1, 1 or 10 (recall, from figure 3, that the test

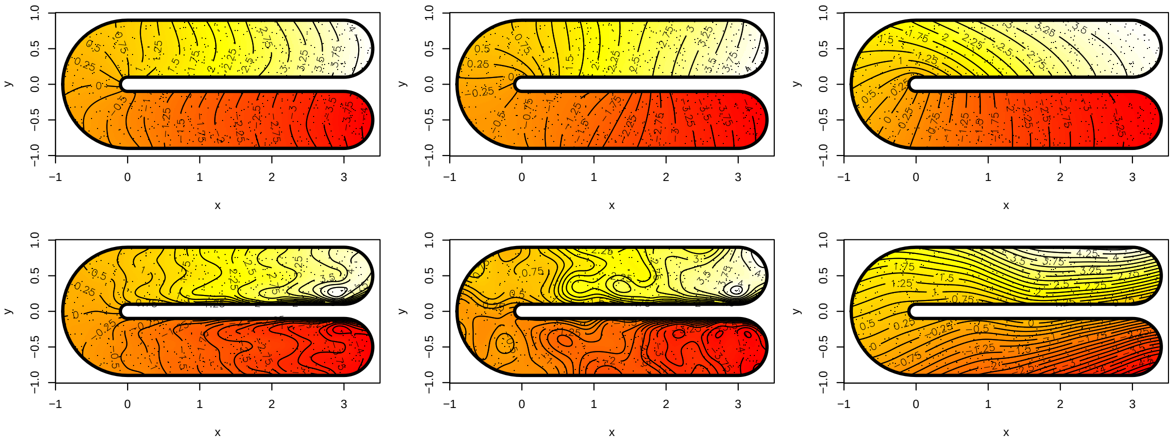

Figure 4: Example reconstructions of the test function from figure 3. On the top row are reconstructions using soap film smooths, while the bottom row shows the equivalent using thin plate splines. The columns relate to the standard deviation of the i.i.d. $N(0, \sigma^2)$ noise, added to the data sampled from the test function. The left column is for $\sigma = 0.1$, the middle column for $\sigma = 1$ and the right column for $\sigma = 10$. The dots on the plots show the data locations in the $x, y$ plane. For each column, both reconstructions are from the same data set. All smoothing parameters were selected by GCV.

function ranges from -4 to 4, approximately), with 200 replicates performed at each noise level. For each replicate we attempted to reproduce the test function using a thin plate spline (actually a thin plate regression spline, Wood, 2003, with basis dimension 100), and with a soap-film smooth with 32 interior knots $(x_k^*, y_k^*)$ and a rank 39 (40 knot) cyclic penalized cubic regression spline as the boundary curve. For each replicate, smoothing parameters were selected by GCV and the mean squared error in reconstructing the true function at the sampled points was calculated.

Figure 4 shows randomly selected (but reasonably typical) reconstructions for the thin plate spline and soap film smoothers at each noise level. Note how the thin plate splines' poor performance is clearly related to the need to smooth across the central gap in the domain. Also at high noise, the thin plate spline (lower right) is tending towards a wholly inappropriate plane, as a result of appropriately high penalization: this reflects a fundamental problem with the thin plate spline penalty which can not be wholly removed by simply evaluating that penalty over the domain of interest.

Figure 5 summarizes the mean square error performance of the two types of smoothing. As expected, the soap film smoother shows substantially better performance, with the relative difference declining with increasing noise.

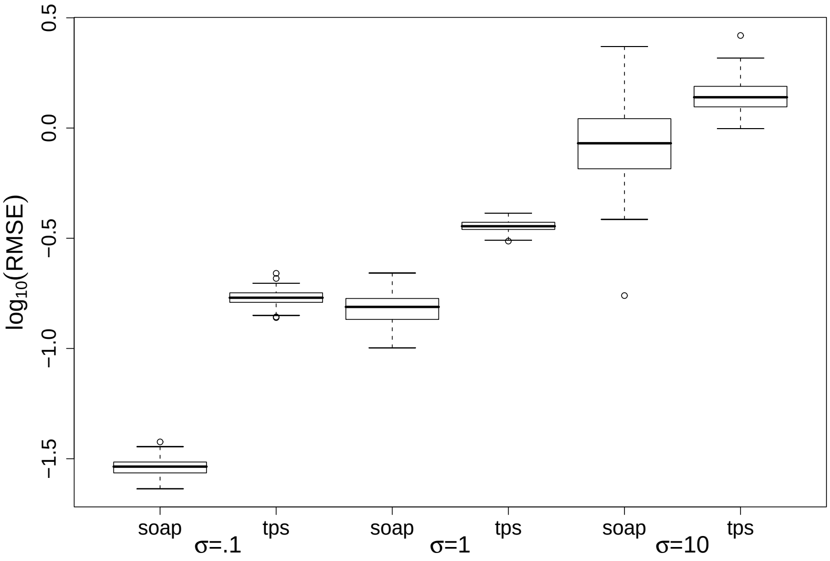

Figure 5: Boxplots of $\log_{10}$ of the root mean square error (RMSE) over 200 replicates for soap film ('soap') and thin plate spline ('tps') smoothers, when reconstructing the test function shown in figure 3 from noisy data. This example presents an obvious difficulty for a thin plate spline, which is clearly reflected in the soap film smoother's comparatively good performance. Note that in this case the boundary of the soap film smooth was treated as unknown.

## 5.2 Comparison with FELSPLINE

We also compared soap film smooths with Ramsay's (2002) FELSPLINE method, using code kindly provided by Tim Ramsay. For the test function illustrated in figure 3, soap film reconstructions look visually better than FELSPLINE reconstructions, because the FELSPLINE boundary conditions force contours to meet the boundary at right angles, but the MSE performance of both methods is very similar (soap films do a little better for high signal to noise ratio, whereas the FELSPLINE boundary condition actually provides a helpful constraint in low signal to noise ratio settings, where FELSPLINE tends to achieve slightly better MSE performance). However, Tim Ramsay pointed out to us that if a linear trend in $y$ is added to the test function then the soap film performance is virtually unchanged, while the FELSPLINE MSE performance deteriorates markedly as the strength of the trend is increased. This is because increasing the trend term makes the FELSPLINE boundary condition ever less tenable. The obvious alternative of smoothing using a linear regression plane plus FELSPLINE helps for extreme linear trends, but leads to a marked MSE deterioration relative to soap film smoothing on the figure 3 test function itself (or indeed on Ramsay's original 2002 test function).

The problems with the FELSPLINE boundary condition are perhaps best illustrated by changing the test function over the horseshoe domain, to the one shown at top left in figure 6. Following Ramsay (2002), the test function

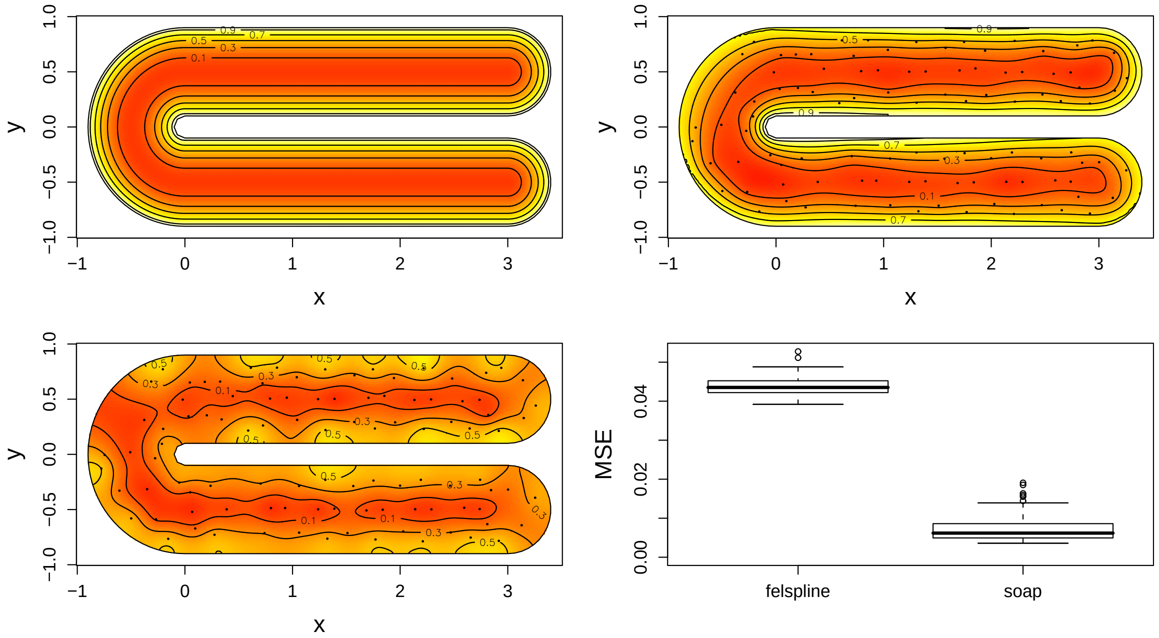

Figure 6: Results from the comparison of soap film smoothing and Ramsay's (2002) FELSPLINE, reported in section 5.2. At top left is a modified test function on a horseshoe domain. The top right shows a typical soap film reconstruction, while the lower left shows the FELSPLINE reconstruction from the same data (also typical). All smoothing parameters were chosen by GCV. The black dots show the 100 sample locations. At lower right are boxplots, over 100 replicates, of mean square error in reconstructing the truth (evaluated over a fine regular grid of points covering the whole horseshoe domain). The relatively poor performance of FELSPLINE appears to relate to its boundary condition, which states that contours must meet the boundary at right angles.

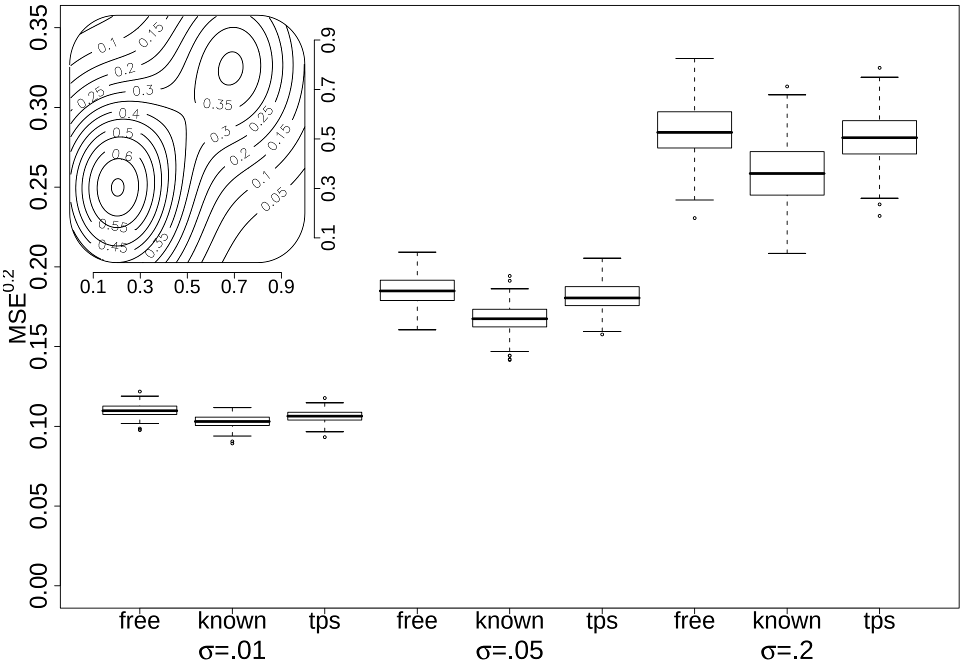

Figure 7: Simulation results from section 5.3, with the true function inset at top left. Each boxplot summarizes the MSE performance of one smoothing method at one noise level over 200 replicates. 'free' is a free boundary soap film. 'known' is a known boundary soap film and 'tps' is a penalized thin plate regression spline.

was sampled with added i.i.d. $N(0, 0.05^2)$ noise from 100 locations in the test domain (shown as black dots on the top right and bottom left panels of figure 6). These data were used to reconstruct the true test function shown at top left of figure 6, using FELSPLINE with GCV selected smoothing parameter, and using a free boundary soap film smoother (rank 39 boundary smooth, 40 $\gamma_i$ parameters) with smoothing parameters selected by GCV. 100 replicates were generated (each using the same $x, y$ locations, in order to reduce the otherwise prohibitive computational expense of evaluating the effective degrees of freedom of the FELSPLINE, which is needed for GCV). For each replicate of each method the MSE in reconstructing the true function across the domain was calculated, using a fine regularly spaced grid of around 3000 points. The lower right panel of figure 6 shows the MSE results.

The relatively poor FELSPLINE MSE performance is related to the boundary condition as is clearly illustrated by the typical FELSPLINE reconstruction shown at lower left of figure 6. Forcing contours to meet the boundary at right angles is not appropriate for this test function, and results in reconstructions that are visually much worse than the equivalent soap film reconstruction (top right of figure 6).

### 5.3 Smooth surface over a simple domain

The section 5.1 example illustrates that the soap film smoother is a considerable improvement in cases where the thin plate spline is expected to perform poorly. The simulations in this section are designed to investigate what can potentially be lost if a soap film smoother is used when it is not needed.

The panel inset at top left of figure 7 shows a smooth true function over an uncomplicated region (boundary shown), where a thin plate spline smooth should have no difficulty. We attempted to reconstruct this function from evaluations of the function at, on average, 382 random locations in the domain of interest. The function values were perturbed by i.i.d. $N(0, \sigma^2)$ random deviates where $\sigma$ was 0.01, 0.05 or 0.2, with 200 replicates performed at each noise level. To each replicate we fitted (i) a thin plate (regression) spline (basis dimension 100); (ii) a soap film smoother with a free boundary ($\dim(\boldsymbol{\gamma}) = 100$ and a rank 39 cyclic penalized cubic regression spline for the boundary) and (iii) a soap film smoother with known boundary condition (again $\dim(\boldsymbol{\gamma}) = 100$). All smoothing parameters were estimated by GCV. Typically, reconstructions appear to differ little between methods when plotted.

The mean squared error in reconstruction was evaluated at the sample points for each method applied to each replicate and the results are summarized in figure 7. The differences in MSE performance are modest relative to the average MSE, but the known boundary soap film smoother always has the best performance, followed by the thin plate spline, followed by the free boundary soap film smoother (these differences are all highly statistically 'significant' at all noise levels using pairwise tests).

The fact that the known boundary soap film has the best MSE performance is expected, given the amount of extra information that knowledge of the boundary provides. Similarly, it is unsurprising that there is some performance cost associated with treating the boundary curve as 'special' when there is no reason to do so: but it is encouraging that, for these examples at least, this cost appears to be relatively modest.

Timings were also recorded for this example. By far the most computer intensive part of the smoothing exercise is the soap film basis and penalty setup. Using a 2.1Ghz Pentium M processor this took approximately 40 seconds with a $150 \times 150$ PDE solution grid. It should be possible to substantially reduce this by using a more efficient PDE solver than SOR (see the appendix).

## 6 Real examples

Completing the motivating whale example of figure 1 would double the length of this paper. Instead we present two simple applications of soap film smoothers, in order to illustrate their practical utility.

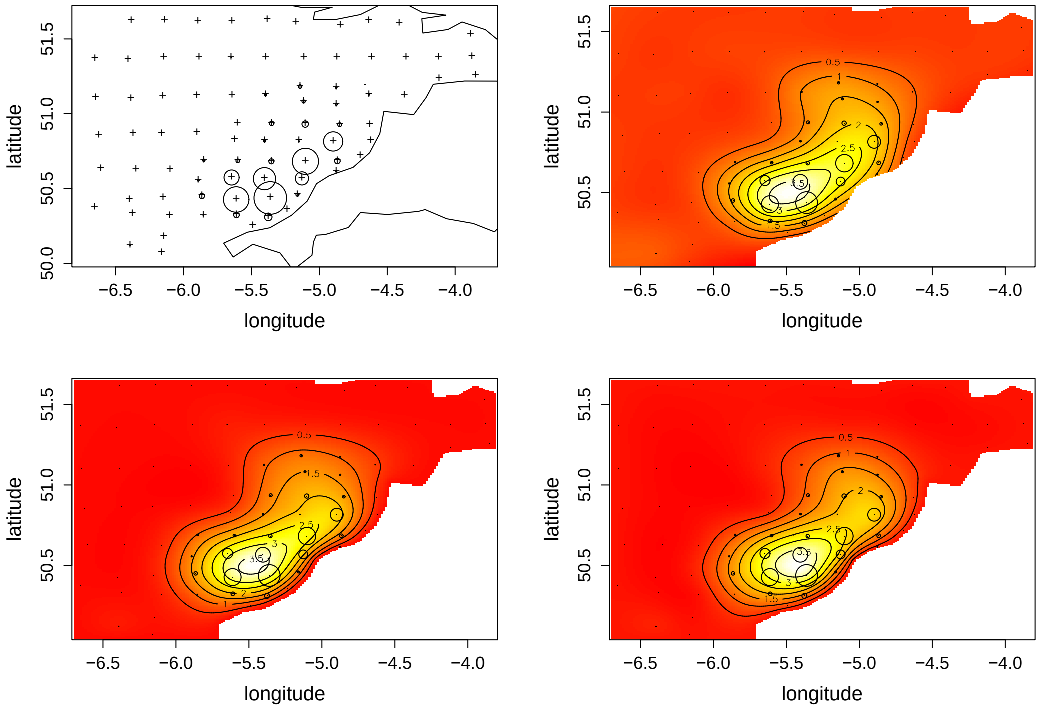

Figure 8: Illustration of the Sole egg data and models discussed in section 6.1. The top left panel shows the region of interest in the Bristol Channel, United Kingdom with the sampling station locations shown as + symbols and raw egg density data shown as open circles with area proportional to egg density (per $m^2$ sea surface). The top right panel shows the thin plate spline fit to the density data: notice the unrealistically high densities at the coast. The lower left panel shows a full soap film smooth fit to the density data, with the density set to zero at the boundary. The lower right panel shows a soap film smooth fit with a free boundary. The three panels showing the model estimates also show the sampling stations as open circles proportional to the observed egg density (with a dot showing the sample location, even at zero density).

## 6.1 Egg survey data

Marine fish egg surveys are undertaken for the purposes of stock assessment. It is difficult to directly survey adult fish in a way that allows good inference of abundance (or total mass), but eggs are more easily sampled, and from egg abundance it is possible to infer the total mass of adults required to produce such a number. Given the expense of gathering the egg data, it is worth being careful when constructing models to use in abundance estimation. The top left panel of figure 8 shows egg data collected from one short research cruise in the Bristol Channel, UK (see figure caption for symbol meanings). The data relate to the first egg developmental stage and are in the form of effective densities per $m^2$ of sea surface. Dixon (2003), Horwood (1993) and Horwood and Greer Walker (1990) provide further details.

Three alternative models were estimated from the egg density data. All had the basic form:

$$\sqrt{y_i} = f(\texttt{km.e}_i, \texttt{km.n}_i) + \epsilon_i$$

where $y_i$ is the observed egg density at the sampling station located at $\texttt{km.e}_i$ kilometres east of the longitudinal line through longitude -5.5 and $\texttt{km.n}_i$ kilometres north of the latitudinal line through latitude 51. The $\epsilon_i$ are independent zero mean random variables with constant variance. (The nearly square co-ordinate system overstates the extent to which the situation is really isotropic: north-south is not really the same as east-west here, and in fact smoothers based on longitude and latitude give barely distinguishable results.) The square root transform stabilizes variances. In all cases the smoothness of $f$ was estimated using GCV, but the details of $f$'s representation differed.

The top right panel of figure 8 shows $f$ estimated using a thin plate spline. Notice how the egg density is estimated to be quite substantial right up to the north coast of the Cornish peninsula (although suppressed in the plots, the model even estimates non-negligible egg densities *south* of the peninsula, where no eggs are found). In fact egg densities are zero at the coast, and the model is quite unrealistic in this respect. The lower left panel of figure 8 shows the $f$ estimate that results from using a soap film smooth with a zero boundary condition at the edge of the coloured area. This boundary condition is realistic: the survey is designed to completely cover the spatial distribution of the eggs, and egg densities are zero 'at the beach'. The resulting $f$ estimate is much more reasonable than the thin plate spline estimate. Note that in this case the fully optimal soap film smoother has been used with one parameter per datum. At the lower right of figure 8 is the $f$ estimate that results from using a soap film smoother with an estimated boundary condition. In this case the smooth had 40 $\gamma_i$ parameters and employed a rank 39 cyclic penalized cubic regression spline as the boundary smooth.

Notice how much more closely the soap film smooth reflects what is known about the sole egg density, and also how the free boundary version of the soap-film smoother suggests that the data really support a sharp decline in density towards the coast, rather than the rather high coastal density suggested by the thin plate spline (which has to try and ensure smoothness of the estimated function beyond the region of interest).

## 6.2 Aral sea chlorophyll

The final example concerns remote sensed chlorophyll data from the Aral sea. Chlorophyll estimates from satellite sensors tend to be somewhat noisy, and some smoothing of the available data can help to clarify the spatial pattern of chlorophyll density. The top left panel of figure 9 shows chlorophyll data from the SeaWifs satellite for the 38th 8 day observation period of the year in the Aral sea, which has a somewhat complicated shape. The data are in fact averages for this period over the years 1998 to 2002, but even so are rather noisy. See `http://seawifs.gsfc.nasa.gov/SEAWIFS` for further information about the NASA SeaWifs program.

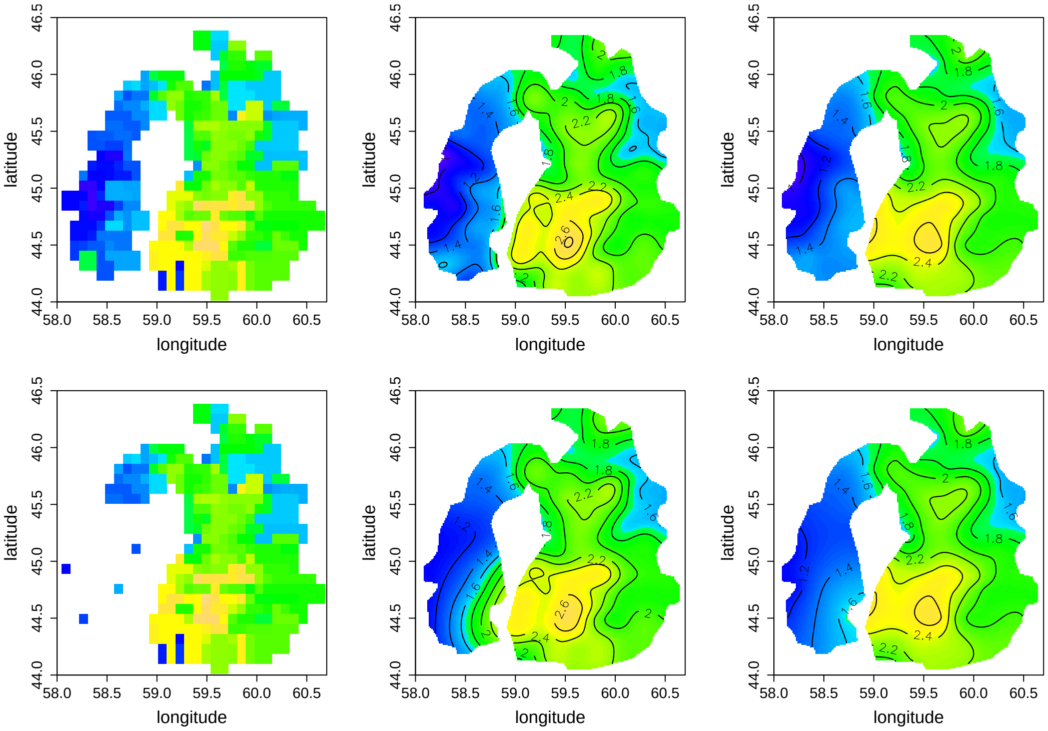

Figure 9: Modelling chlorophyll in the Aral sea. The top row relates to the full set of data available for the Aral sea, while the lower row simulates very uneven data coverage, by deletion of a high proportion of the data in the western arm of the sea. The first column illustrates the (logged) raw data. The second column shows thin plate spline estimates of the log chlorophyll concentrations. The rightmost column shows soap-film estimates of the log chlorophyll concentrations. The colour scale is the same in all plots. Notice the leakage across the central isthmus, which is apparent in the thin plate spline smooths, but not in the soap film smooths. Smoothing was performed using a nearly square co-ordinate system, but results based simply on longitude and latitude are almost indistinguishable.

Aral sea chlorophyll concentrations are expected to vary smoothly within the water of the sea, but there is no reason to expect smooth variation across the isthmus of the peninsula that projects northwards into the sea at longitude 59, for example. Two versions of the model:

$$\log\{E(\texttt{chl}_i)\} = f(\texttt{km.e}_i, \texttt{km.n}_i), \;\; \texttt{chl}_i \sim \text{Gamma}$$

were fitted to the raw satellite measurements. In one a thin plate spline was used to represent $f$, and in the second a soap film smoother was used (with 150 $\gamma$ parameters and a rank 49 cyclic cubic penalized regression spline used to represent the boundary). The $\texttt{km.e}_i, \texttt{km.n}_i$ nearly square co-ordinate system is as for the sole egg example, but with reference lines passing through location (59.5, 45).

Estimation and GCV smoothness selection was by the method of Wood (2008). The first row of figure 9 plots, from left to right, the raw data, the estimate of $f$ using a thin plate spline and finally the estimate of $f$ using a soap film smoother. Notice the way in which the thin plate spline smooths, inappropriately, across the isthmus of the central peninsula, so that relatively high chlorophyll densities are estimated for the southern part of the eastern shore of the western basin of the sea. These elevated densities have no support from the observed data. Similarly, the thin plate spline estimates a decline in chlorophyll abundance towards the southern half of the western shore of the eastern basin: again this is unsupported by data. The soap film avoids such artefacts.

We have exceptionally even data coverage in this example: in many applications much less even coverage is the norm. As a quick illustration of how uneven coverage can exaggerate artefacts when smoothing over difficult domains, we randomly removed most of the data south of latitude 45.5 in the western basin of the Aral sea. We then fitted the same models to these thinned data as had been fitted to the full data set. The lower row of figure 9 illustrates the result. The sparsity of data in the western basin causes the thin plate spline to extrapolate across the isthmus from the data rich eastern basin. As a result very high densities are estimated for large parts of the western basin, and in this case we know that these densities are wrong. The soap film smoother, on the other hand, produces an estimate that is very similar to the estimate from the full data.

## 7 Discussion

The soap film smoothers proposed here meet the four objectives listed in the introduction. Given that we are writing the paper after completing the work the reader may find this unsurprising, but in any case the performance of the method, evident from the examples in sections 5 and 6, is encouraging. As important, in practice, are the computational efficiency and convenience of the approach. These rest on the fact that we are able to evaluate basis functions and a quadratic wiggliness penalty for the smoothers, which allows us to use all the computational and theoretical machinery available for such basis-penalty smoothers, and to incorporate the smoothers as components of a wide variety of statistical models. The characterization of the smoothers provided by section 3 is the key to the basis-penalty representation and turns out to make the soap film basis computation rather straightforward: reliable solvers for the basis defining PDEs are readily available and easily coded. The results of section 3 also provide a nice example of the link between smoothing and differential equations highlighted by Ramsay (2000).

In the introduction we discussed why new methods were required beyond the work of Stone (1988) and Ramsay (2002), and it is worth revisiting this topic in the light of the examples presented in sections 5 and 6. Stone's approach uses a thin plate spline penalty integrated only over the domain of interest. Although this does avoid the un-modified thin plate spline's tendency to be influenced by the need to meet smoothness objectives completely outside the region of interests, the penalty still shrinks the smooth towards a plane over the domain of interest. This is not appropriate in the Aral sea example, where the straight line distances between points in the model co-ordinate system only relate to 'biological distance' if the straight line does not cross land. Similarly a plane over the $x$, $y$ plane is not a good 'smoothest model' for the figure 3 test function. The original test function of Ramsay (2002, figure 2) was completely flat on shortest paths across the domain (i.e. contours were almost all straight lines orthogonal to the boundary). It provides an even clearer example of the difficulty with Stone's method, which can not simultaneously allow this flatness within the domain of interest, while also allowing the function to follow the correct curve on the boundary. In contrast the separation of boundary and interior smoothness of the soap film smoother makes this easy to accommodate.

Ramsay's FELSPLINE is a substantial improvement over thin plate spline based smoothing, for smoothing over complex domains, but it requires a quite complex computational approach, making it difficult to incorporate a FELSPLINE as a component of another model. Section 5.2 also suggests that FELSPLINE can perform substantially less well than soap film smoothing when the FELSPLINE boundary condition is inappropriate. Forcing contours of the smooth to meet the boundary curve at right angles is a quite strong restriction.

In the known boundary case we are able to obtain the (numerically) exact minimizer of (5), and in that sense have the optimal solution to the problem posed. For the unknown boundary case our solution is not so elegant, in that we model the boundary condition and estimate it as part of fitting. This leaves open the question of whether an alternative formulation of the problem might enable the boundary model to be eliminated while still meeting our basic objectives. Our attempts to do this using (5) or close relatives, without our current boundary condition, proved fruitless. Some regularization at the boundary is surely necessary given that the null space of $J_\Omega$ is not finite dimensional, but the FELSPLINE boundary artefacts evident in figure 6 are not encouraging with regard to finding simpler alternatives to boundary model based regularization. In any case the soap film boundary model can have the practical advantage of controlling otherwise poor boundary behaviour.

To summarize, we believe that good finite domain smoothing relies on using a penalty like the soap film penalty

$J_\Omega$, with an infinite dimensional null space, otherwise smoothing is penalizing towards a low dimensional function which tends to inappropriately smooth across boundary features. When using $J_\Omega$, its infinite dimensional null space means that *some* boundary condition is needed to regularize the solution to the smoothing problem. The improved MSE performance of soap films relative to FELSPLINE (which both use $J_\Omega$, at least approximately) rests on the former's relatively flexible boundary condition, which in effect regularizes the penalty null space in a data adaptive manner.

**Note**: The code used here is freely available as an R package.

**Acknowledgments**

We are especially grateful to Tim Ramsay for providing us with the FELSPLINE Matlab code used in Ramsay (2002) and much associated advice, which allowed us to perform the comparisons reported in section 5.2, and to John Toland for patiently explaining the issues surrounding generalization of the theorem and for help placing the results in a wider PDE theory context. We wish to thank two anonymous referees for helping to improve the paper through a large number of constructive suggestions. Thanks also to Len Thomas for providing us with insightful comments and support and to Joe Horwood, for introducing us to the Sole egg problem. A substantial part of the work was funded by ONR project *Development of new methods and software for distance sampling surveys of wildlife populations* and CSIRO provided the authors with part of the travel money that made the work possible.

## Appendix. Computational details

This appendix fills in some of the computational details that are glossed over in the main paper, but are important for practical implementation.

**PDE solution method.** In the work reported here we used a successive over-relaxation (SOR) method, with Chebyshev acceleration, adapted from Press et al. (1992, section 19.5). This method solves a discretized version of the PDE concerned on a rectangular grid, at a computational cost of $O(N^3)$ for an $N \times N$ grid. One alternative is to use a multi-grid method: these are more complicated to code than SOR, but have a computational cost of $O(N^2)$ for an $N \times N$ grid (e.g. Press et al. 1992, section 19.6). Given that we typically use $N$ around 100-200, there is scope for substantial improvement in computational efficiency by using a multi-grid method.

**PDE solution grid**. The grid spacing, $\Delta$, used to solve the defining PDEs has to be chosen, and affects the numerical accuracy with which the soap film smoother can be computed. The discretization of the PDEs involves $O(\Delta^2)$ truncation errors, but in the absence of convergence rates for soap films, it is difficult to say much that is useful for setting up coarsest solution grids. However, the adequacy of any particular $\Delta$ choice can be checked by halving $\Delta$ and examining the effect on the computed soap film smoother. In the examples presented in section 5.1, doubling the resolution of the solution grid, from 200 to 400 in the $x$ direction, leads to a mean absolute change in fitted model predictions across the domain of $< 10^{-3}$. In any case, to avoid possible basis degeneracy, $\Delta$ must generally be smaller than the minimum distance between the $x_k^*, y_k^*$ points used to generate the basis.

**Basis dimension choice**. The other practical choice to make is the number of $x_k^*, y_k^*$ values, $K$. Provided $K$ is not overly restrictive the choice is not generally critical, since the smoothing parameter determines the effective degrees of freedom of the soap film, rather than $K$. However a small $K$ is more computationally efficient than a large one. In practice a simple check that $K$ is not overly restrictive is provided by refitting a smooth with, e.g. basis dimension $2K$ to the residuals of the original model, in order to check for missed pattern with respect to $x$ and $y$. Note the difference in type between the choice of the basis dimension, $K$, and the solution grid spacing $\Delta$. $K$ determines how many basis functions there will be, and hence how many pairs of PDEs must be solved, while $\Delta$ determines the accuracy with which those PDEs will be solved.

**The $\boldsymbol{\rho_k(x, y)}$ singularity**. An important issue is the handling of the singularity in the solution of (3) to find the basis functions $\rho_k(x, y)$. When discretizing (3) onto a grid with spacing $\Delta$ it is necessary to view the gridded values of the solution in the interior of $\Omega$ as representing averages of $\rho$ over the $\Delta \times \Delta$ square centered on each grid node. This quantity is finite everywhere within $\Omega$, since the fundamental solution of Laplace's equation (see e.g. Evans, 1998, section 2.2.1) has finite integral over any small region of the $x - y$ plane. In that case the $\rho_k$ can be evaluated by solving (3) subject to the zero boundary condition on $B$ and the internal boundary condition that the gridded value associated with the 'knot' $x_k^*, y_k^*$ is fixed at 1. The result is then normalized to obtain the numerical estimate of $\rho_k$, and this estimate will converge to $\rho_k$ as $\Delta \to 0$. In fact, for practical computation, nothing is gained by actually normalizing the $\rho_k$ — the normalizing constant may as well be absorbed into the $\gamma_k$. Note that for 3 dimensional domains the singularity would present an insurmountable problem: the solution in the vicinity of a 'knot' does not have a finite integral over any volume enclosing the knot.

**Defining the boundary**. The most tedious part of the soap film smoother construction to implement is the handling of the boundary $B$ (this is not unique to soap films). In the work reported here the boundary was discretized into a continuous sequence of line segments. Which solution grid points should count as lying 'on' the boundary is then determined by exactly the method outlined after equation (4), in section 3. Whether an $x, y$ point lies within $\Omega$ is determined by how many boundary line segments have to be crossed along any straight line from $x, y$ to beyond the edge of the solution grid: if the number is odd then the point is within $\Omega$.

**Prediction**. Predictions at the observation points requires only the model matrix, and hence no further PDEs need be solved. For prediction away from the observation points there are several alternatives. If the points at which predictions will be required are known from the outset, then a 'prediction matrix' can be set up at very little cost at the same time as the model matrix is produced. Another option is that the model basis functions are stored 'whole' in gridded form, so that

$$\hat{f}(x,y) = \sum_J \hat{\alpha}_j a_j(x,y) + \sum_k \hat{\gamma}_k g_k(x,y)$$

can be evaluated by bi-linear interpolation of the gridded functions, but this can be expensive in terms of storage. Alternatively the basis functions are re-computed one at a time, just as in the initial model set up, and the previous expression for $\hat{f}(x,y)$ accumulated. Finally, if the un-normalized grid dependent parameterization suggested earlier in this section is employed, and variance estimates are not needed, then (2) and (3) can be solved directly for $f$, by solving (3) with internal boundary points at every $x_k^*, y_k^*$. The average values of the approximate solution in the side $\Delta$ square surrounding the $k^{\text{th}}$ such internal boundary point are given by $\gamma_k'$, where $\boldsymbol{\gamma}' = \mathbf{E}\boldsymbol{\gamma}$ and $E_{ij} = \rho_j(x_i^*, y_i^*)$ for $i \neq j$ and $E_{ii}$ = 1. **E** is computed when the initial basis function computations are performed.

## References

Bowman, A.W. and A. Azzalini (1997) *Applied Smoothing Techniques for Data Analysis*. Oxford: Oxford University Press.

Dixon, C.E. (2003) *Multi-dimensional modelling of physiologically and temporally structured populations*. PhD Thesis, University of St Andrews, UK.

Duchon, J. (1977) Splines minimizing rotation invariant semi-norms in Solobev spaces. In W. Schemp and K. Zeller (eds), *Construction Theory of Functions of Several Variables*, pp 85-100. Berlin: Springer.

Eilers, P.H.C. and B.D. Marx (1996) Flexible smoothing with B-splines and penalties. *Statistical Science* 11(2), 89-121.

Eilers, P.H.C. (2006) P-spline smoothing on difficult domains. Seminar at the University of Munich.

Evans, L.C. (1998) *Partial Differential Equations*. American Mathematical Society, Providence, Rhode Island.

Fahrmeir, L., T. Kneib and S. Lang (2004) Penalized structured additive regression for space-time data: a Bayesian perspective. *Statistica Sinica* 14(3), 731-761

Green, P.J. and B.W. Silverman (1994) *Nonparametric Regression and Generalized Linear Models*. London: Chapman & Hall.

Gu, C. (2002) *Smoothing Spline ANOVA Models*. New York: Springer.

Hastie, T.J. and R.J. Tibshirani (1990) *Generalized Additive Models*. London: Chapman & Hall.

Horwood, J. (1993) The Bristol Channel Sole (*Solea solea (l.)*): A fisheries case study. *Advances in Marine Biology*, 29:215-367.

Horwood, J. and M. Greer Walker (1990). Determinacy of fecundity in Sole (*Solea solea*) from the Bristol Channel. *Journal of the Marine Biological Association of the United Kingdom* 70, 803-813.

Kim, Y.J. and C. Gu (2004) Smoothing spline gaussian regression: more scalable computation via efficient approximation. *Journal of the Royal Statistical Society, Series B* 66, 337-356.

Press, W.H., S.A. Teukolsky, W.T. Vetterling and B.P. Flannery (1992) *Numerical Recipes in C* (2nd edition). Cambridge: Cambridge University Press.

Ramsay, T. (1999) *A Bivariate Finite Element Smoothing Spline Applied to Image Restoration.* PhD Thesis, Queen's University, Kingston, Ontario, Canada.

Ramsay, J.O. (2000) Differential equation models for statistical functions. *Canadian Journal of Statistics* 28, 224-240.

Ramsay, T. (2002) Spline smoothing over difficult regions. *Journal of the Royal Statistical Society, Series B* 64, 307-319.

Ruppert, D., M.P. Wand, and R.J. Carroll (2003) *Semiparametric Regression.* Cambridge: Cambridge University Press.

Silverman, B.W. (1985) Some aspects of the spline smoothing approach to nonparametric regression curve fitting. *Journal of the Royal Statistical Society, Series B* 47, 1-53

Stone, G. (1988) *Bivariate splines.* PhD Thesis, University of Bath, UK.

Strauss, W.A. (1992) *Partial Differential Equations: An Introduction.* John Wiley & Sons.

Wahba, G. (1990) *Spline models fore observational data.* Philadelphia: SIAM.

Wand, M.P. and M.C. Jones *Kernel Smoothing.* London: Chapman & Hall.

Wang H and M.G. Ranalli (2007) Low-rank smoothing splines on complicated domains. *Biometrics* 63, 209-217.

Wood, S.N. (2003) Thin plate regression splines *Journal of the Royal Statistical Society, Series B* 65, 95-114

Wood, S.N. (2004) Stable and efficient multiple smoothing parameter estimation for generalized additive models. *Journal of the American Statistical Association* 99, 673-686.

Wood, S.N. (2006a) *Generalized Additive Models: An Introduction with R.* Boca Raton: CRC/Chapman & Hall.

Wood, S.N. (2006b) On confidence intervals for generalized additive models based on penalized regression splines. *Australian and New Zealand Journal of Statistics.* 48(4): 445-464.

Wood, S.N. (2008) Fast stable direct fitting and smoothness selection for generalized additive models. *Journal of the Royal Statistical Society, Series B* 70(2):