PENALIZED TRIOGRAMS, OR PRAGMATIC GONIOLATRY

ROGER KOENKER
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

Abstract. Hansen, Kooperberg, and Sardy (1998) introduced a family of continuous, piecewise linear functions defined over adaptively selected triangulations of the plane as a general approach to statistical modeling of bivariate densities, regression and hazard functions. These *triograms* enjoy a natural affine equivariance that offers several advantages over competing tensor product methods that are more commonly used in statistical applications.

Hansen, Kooperberg and Sardy adopt the regression spline approach of Stone (1994) and consider basis functions consisting of so-called "tent functions" defined with respect to a triangulation of a given planar domain. As in knot selection for more conventional univariate splines, vertices of the triangulation are introduced, or removed, sequentially in an effort to balance fidelity to the data and parsimony of the model using a form of generalized cross validation.

In this paper we explore some smoothing spline variants of the triogram model which are based on rather simple roughness penalties adapted to the piecewise linear structure of the triogram model. While these penalties could be used for density estimation and other applications we choose to emphasize the simplest nonparametric regression applications in this initial report.

"Goniolatry, or the worship of angles, ..."
Pynchon (1997)

1. Introduction

Piecewise polynomial functions, or splines, have proven to be an extremely powerful concept throughout approximation theory and the statistical literature on smoothing. Like the eponymous drafting tool, splines offer a marvelously flexible, yet eminently practical mathematical tool. I was introduced to splines by Dale Poirier twenty-five years ago when we were, momentarily, both new assistant professors at the University of Illinois. The ideas of smoothing splines that had been recently developed by Reinsch (1967) and Wahba (1975) seemed particularly appealing to me . This appeal was undoubtedly due in large part to the thorough indoctrination I was receiving at the

Version: May 7, 1999. This paper was prepared for a conference in honor of George Judge held at the University of Illinois, May 1-2, 1999. This is an exceedingly preliminary version, so readers' comments would be especially welcome. The research was partially supported by NSF SBR-9617206. I would like to thank Steve Portnoy and Ivan Mizera for many helpful discussions. I would also like to express my appreciation to Rolf Turner for his elegantly expedient Splus package for Delaunay triangulation.

time from George Judge regarding the virtues of shrinkage estimators, and the vices of pre-testing. In the statistical literature on splines there continues to be a vigorous debate about the relative merits of shrinkage versus pre-testing, or model selection. The latter approach has been championed by Charles Stone and his students under the rubric "regression splines", and they have suggested ingenious schemes for stepwise addition and deletion of spline knots. Others, notably Grace Wahba and more recently Green and Silverman (1994), have advocated the shrinkage approach.

In its simplest univariate form the smoothing spline solves the problem,

$$\min_{g\in\mathcal{G}} \sum_{i=1}^{n}(y_i - g(x_i))^2 + \lambda \int_a^b (g''(x))^2 dx, \tag{1.1}$$

where $\mathcal{G}$ is a Sobolev space of continuous functions on $[a,b]$ with absolutely continuous first derivative and square-integrable second derivative, and λ controls the smoothness of the fitted function. In this form the estimator $\hat{g}(\cdot)$ may be interpreted as an estimate of the conditional mean function

$$g_0(x) = E(y|x)$$

And the penalty term

$$\lambda \int_a^b (g''(x))^2 dx,$$

may be viewed as representing a prior belief that the L_2 norm of g'' is unlikely to exceed a specified bound.

Of course there is nothing sacred about the Gaussian formulation of the smoothing spline problem and there have been numerous efforts to modify both the fidelity and roughness penalty to achieve alternative objectives. One such effort is Koenker, Ng, and Portnoy (1994), which describes a non-parametric approach to estimating conditional quantile functions by solving problems of the form,

$$\min_{g\in\mathcal{G}} \sum_{i=1}^{n} \rho_\tau(y_i - g(x_i)) + \lambda TV(g'),$$

where $\rho_\tau(u) = u(\tau - I(u < 0))$ generates a fidelity term appropriate for quantile estimation, and the roughness penalty becomes total variation of the first derivative of g. If g' is absolutely continuous, we can write,

$$TV(g') = \int |g''(x)| dx,$$

thus establishing a clear link to the classical L_2 roughness penalty. Total variation proves to be a natural alternative penalty for quantile regression fidelity since it preserves the piecewise linear form of the objective function and therefore preserves the linear programming formulation of the optimization problem. Solutions to the problem (1.1) take the form of continuous, piecewise linear functions with jumps in their derivative at the observed x_i's. The L_1 nature of the total variation penalty has a rather different shrinkage effect than the classical L_2 penalty. Just as ordinary ℓ_1 regression seeks to identify p basic observations whose exact fit characterizes the p-dimensional parameter estimate, the L_1 penalty acts like a model selection device by identifying a small number of critical x_i points at which $\hat{g}'$ will be allowed to jump. The number of these selected jump points is controlled by the parameter λ, and provides a natural measure of the dimensionality of the fitted function. See Tibshirani (1996) and Donoho, Chen, and Saunders (1998) for related discussion of the model-selection/shrinkage effects of L_1 type penalties.

The extension of univariate smoothing splines to bivariate situations, and beyond, raises new question about how to measure the roughness of surfaces. The thin plate smoothing splines of Meinguet (1979) solve,

$$\min_g \sum_{i=1}^{n} (z_i - g(x_i, y_i))^2 + \lambda J(g) \tag{1.2}$$

with the roughness penalty defined as,

$$J(g) = \int\int ((\partial^2 g/\partial x^2)^2 + 2(\partial^2 g/\partial x \partial y)^2 + (\partial^2 g/\partial y^2)^2) dx dy$$

and seem to provide a nice extension of the natural cubic splines solving (1.1). Again, only linear functions have penalty zero. An important feature of the thin-plate penalty is its invariance to translations and rotations of the xy-coordinates. See Wahba (1990) and Green and Silverman (1994) for further details.

He, Ng, and Portnoy (1998) have recently proposed a bivariate form of the quantile smoothing spline using a roughness penalty that sums the univariate total variation of the function along rectangular grid lines. This formulation gives rise to bilinear tensor product splines that are continuous and piecewise linear on the grid lines, and bilinear on the rectangular patches between grid lines. Similar tensor product splines have also been widely used in the least-squares regression spline literature.

One way to escape the rather rigid demands of the rectangular grid formulation inherent in tensor product splines is to define piecewise polynomials on more flexible tesselations of the planar domain. Triangles prove particularly convenient and have been widely employed in the approximation theory literature as well as in the applied world of computer aided design. In an innovative recent paper, Hansen, Kooperberg, and Sardy (1998) study linear splines on triangulations from a regression spline standpoint. Their functions are dubbed triograms.[1]

The primary objective of this paper is to begin to explore a smoothing spline variant these triograms. In the next Section we describe some basic features of triograms. In Section 3 the proposed roughness penalty is introduced and some implementation issues are discussed. Section 4 describes some applications of the method. And the

[1] Not to be confused with the (perhaps) better known trigrams of the *I Ching*, and arising more recently in the study of the statistical approach to language processing.

final Section contains some concluding remarks and comments on prospects for further research.

2. Triograms: Some Background

Following Hansen, Kooperberg and Sardy, let $\mathcal{U}$ be a compact region of the plane, and let Δ denote a collection of sets $\delta_i : i = 1, \ldots, n$ with disjoint interiors such that $\mathcal{U} = \cup_{\delta \in \Delta} \delta$. In general the collection, Δ, is called a tessellation of $\mathcal{U}$. We will be concerned only with the case that the $\delta \in \Delta$ are planar triangles, in which case Δ is called a triangulation. We will further restrict attention to Δ whose vertices are not collinear. This is considerably stronger than the notion of "conforming" triangulations used by Hansen, Kooperberg and Sardy, but is natural if we consider vertices $V = \{v_i = (x_i, y_i)\}_{i=1}^n$, generated at random from a process with a (Lebesgue) density.

The space, $\mathcal{G}$, of continuous, piecewise linear functions on $\mathcal{U}$ associated with the triangulation, Δ consists of all continuous functions, g, which are linear when restricted to $\delta \in \Delta$. That $\mathcal{G}$ is a finite dimensional linear space can be easily seen by considering the basis functions $\{B_i(u)\}_{i=1}^n$ that span $\mathcal{G}$. Hansen, Kooperberg and Sardy show that these $B_j(u)$ may be expressed in terms of the barycentric coordinates of the point u represented by the vertices v_1, v_2, v_3 of the triangle, δ, containing u,

$$u_j = \sum_{i=1}^{3} B_i(u) v_i \quad j = 1, 2,$$

and satisfying the condition,

$$1 = \sum_{i=1}^{3} B_i(u)$$

Solving for the $B_i(u)$'s we have, provided the vertices aren't collinear, by Cramer's rule,

$$B_1(u) = \frac{\text{SignedArea } (u, v_2, v_3)}{\text{SignedArea } (v_1, v_2, v_3)}$$

where

$$\text{SignedArea } (v_1, v_2, v_3) = \frac{1}{2} \begin{vmatrix} v_{11} & v_{21} & v_{31} \\ v_{12} & v_{22} & v_{32} \\ 1 & 1 & 1 \end{vmatrix}$$

is the signed area of the triangle, δ. The remaining $B_i(u)$ are defined analogously by replacing the vertex v_i by u. Clearly, the $\{B_i(u)\}$ are linear in u on δ, and satisfy the interpolation conditions that $B_i(v_j) = 1$ for $i = j$ and $= 0$, otherwise, and thus are linearly independent. They are also affine equivariant, that is for any non-singular, 2×2 matrix A, and vector $b \in \Re^2$,

$$B_i(u) = B_i^*(Au + b) \qquad u \in \mathcal{U}$$

where $\{B_i(u)\}$ are formed from the vertices $\{v_i\}_{i=1}^n$ and $\{B_i^*\}$ are formed from the vertices $\{Av_i + b\}_{i=1}^n$. Thus, in particular, the basis is equivariant to rotations of the coordinate axes, a property notably missing in many other bivariate smoothing methods. Like univariate B-spline basis functions, $0 \leq B_i(u) \leq 1$ with

$$\sum_{i=1}^n B_i(u) = 1 \qquad u \in \mathcal{U}.$$

The piecewise linear functions in $\mathcal{G}$ defined over the triangulation Δ are called triograms. The piecewise linearity of the functions $g \in \mathcal{G}$ is obviously a stringent requirement, but there are persuasive arguments for the advantages offered by its simplicity. In the spirit of Stone, Hansen, Kooperberg, and Troung (1997), Hansen, Kooperberg and Sardy consider a regression spline approach to triograms in which vertices are sequentially added and deleted in an effort to find a parsimonious fit. The approach is remarkably flexible and can be used for density estimation, regression and other "extended linear models." In general, we may take $U_1, \ldots, U_n$ as a random sample, and let $\ell(g, U)$ denote the quasi-log-likelihood linking the common distribution of the U_i's to the functions of $g \in \mathcal{G}$. For a given triangulation Δ, the triogram estimate is then

$$\hat{g}_n = \arg\max_{g \in \mathcal{G}} \sum_{i=1}^n \ell(g, U_i)$$

Since $g(u) = \sum_{i=1}^m \beta_i B_i(u)$ we can also regard the problem of finding $\hat{g}_n$ as choosing the coefficients β_i to maximize the quasi-likelihood.

In the case of density estimation, it is convenient to model the log density by g, so the proposed model for the bivariate density becomes

$$f(u, \beta) = \exp\{\sum_{i=1}^m \beta_i B_i(u) - C(\beta)\}$$

where

$$C(\beta) = \int_{\mathcal{U}} \exp\{\sum \beta_i B_i(u)\} du$$

is the required normalizing constant. The parameters $\hat{\beta}_i$, which may be interpreted as heights of the unnormalized log-density at the vertices of the triangulation maximize

$$\ell_n(\beta) = \sum \log f(U_i, \beta).$$

This proves it to be advantageous from the point of view of numerical analysis since the likelihood equations have the relatively simple form

$$\int_{\mathcal{U}} B_j(u) f(u, \beta) du = n^{-1} \sum_{i=1}^n B_j(U_i), \qquad j = 1, \ldots, m.$$

Solution by Newton's method requires evaluating the Hessian with typical element

$$\int B_i(u)B_j(u)f(u,\beta)du,$$

but due to the piecewise linear structure of the $\{B_i\}$ the integrals can be evaluated analytically and the problem is globally concave as can be seen from the form of the Hessian.

In regression the situation is in some respects even simpler. Partition $U = (X, Y)$ with design points $X \in \Re^2$ and response $Y \in \Re$, and write

$$\ell(g, U) = -(Y - g(X))^2.$$

Then maximizing $\ell_n(\beta) = \sum_{i=1}^n \ell(g, U_i)$, yields the normal equations,

$$(B'B)\beta = B'Y$$

where $B'B = (n^{-1}\sum_{i=1}^n B_j(X_i)B_k(X_i))$ and $B'Y = (n^{-1}\sum_{i=1}^n B_j(X_i)Y_i)$. Noting that elements b_j of the matrix $B'B$ are nonzero iff the vertices v_i and v_j are joined by an edge in the triangulation Δ, one can – by careful ordering of the vertices – reduce the computational burden by using algorithms designed for banded linear equations.

Selecting a good triangulation is clearly critical to a successful estimate, $\hat{g}_n$. Hansen, Kooperberg and Sardy denote considerable attention to their stepwise addition and deletion of vertices designed to achieve a "good" choice of Δ. Motivated by penalty methods in the theory of univariate smoothing splines we explore an alternative penalized triogram approach in the next Section.

3. A Roughness Penalty for Triograms

Our initial motivation for the roughness penalty proposed below for triograms comes from prior work on univariate quantile smoothing splines. In Koenker, Ng, and Portnoy (1994) we considered the problem of estimating a function $g : [a, b] \to \Re$ given a sample $\{(x_i, y_i) : i = 1, \ldots, n\}$ by minimizing

$$R(g) = \sum_{i=1}^n \rho_\tau(y_i - g(x_i)) - \lambda P(g)$$

where $\rho_\tau(u) = u(\tau - I(u < 0))$, $P(g)$ denotes a measure of roughness of the function g, and λ is a tuning parameter controlling the smoothness of the fitted function. The first term of $R(g)$ represents fidelity to the data; minimizing its expectation yields the τ^{th} conditional quantile function of y_i given x_i. The conventional penalty function for univariate $\mathcal{L}_2$ smoothing splines is

$$P_2(g) = \int_a^b (g''(x))^2 dx.$$

Minimizing $R(g)$ with penalty P_2 is a quadratic programming problem and is discussed in detail in Bosch, Ye, and Woodworth (1995). A natural alternative to P_2, which maintains the piecewise linear form of the fidelity term is

$$P_1(g) = \int_a^b |g''(x)|dx.$$

But it proves convenient to extend the domain of P_1 somewhat. Recalling that the total variation of an absolutely continuous function $f : [a, b] \to \Re$ may be expressed as,

$$TV(g) = \int_a^b |g'(x)|dx,$$

we can extend $P_1(g)$, by employing instead

$$P_0(g) = TV(g').$$

Minimizing $R(g)$ over an appropriately extended Sobolev space $\mathcal{G}$ including the piecewise linear functions on $[a, b]$ yields a solution of the piecewise linear form. This is the essential content of Theorem 1 of Koenker, Ng, and Portnoy (1994), and yields an efficient ℓ_1-type algorithm for computing the estimator.

Total variation of the first derivative, thus emerges as a natural measure of roughness of the function g. By making λ sufficiently large we can force the fitted function to be linear, so $P_0(g) = 0$, while at the opposite extreme a sufficiently small λ ensures an interpolatory fit with fidelity equal zero. A careful choice of λ between these two extremes yields an estimated $\hat{g}_n$ representing a reasonable compromise between the competing demands of fidelity to the data and parsimony/smoothness of the fitted function.

There are a number of important qualitative features of $P_0(g)$ that distinguish it from its more familiar $P_2(g)$ competitor. In particular, note that while for any convex function

$$P_1(g) = g'(b) - g'(a),$$

$P_2(g)$ can be very sensitive to the way $g'(x)$ behaves on (a, b).

Example: Let $g(x) = (\alpha + 1)^{-1} x^{\alpha+1}$ on $[0, 1]$, with $\alpha > 0$, so $P_0(g) = 1$, irrespective of α, but $P_2(g) = \alpha^2/(2\alpha - 1)$ for $\alpha > 1/2$, and is undefined for $0 < \alpha \leq 1/2$.

So, $P_2(g)$ reflects a strong preference for quadratic behavior of g on (a, b), while $P_0(g)$ is indifferent to this intermediate behavior. The piecewise linear $\hat{g}$, whose slope jumps from $g'(a)$ to $g'(b)$ at some point $x_0 \in (a, b)$ is fine for $P_0(g)$, but it is anathema for $P_2(g)$. This illustrates tendency of the $P_0(g)$ penalty to be accepting of the piecewise linear solution, while the $P_2(g)$ penalty favors solutions with locally quadratic curvature.

3.1. **Image Restoration and the Tinker-Toy Model.** Another motivation for some form of total variation as a roughness penalty for triograms comes from recent work by Blomgren and Chan (1998) who have argued that in addition to important invariance attributes, methods of image reconstruction based on minimizing total variation have desirable edge preserving properties. They begin by defining total variation for real-valued functions $g : \mathbb{R}^m \to \mathbb{R}$ as,

$$TV_{m,1}(g) = \int_{\mathcal{U}} ||\nabla g|| dx.$$

Some caution is obviously required to interpret this integral for discontinuous g. Extending the definition to vector-valued functions, Blomgren and Chan propose, for $g : \mathbb{R}^m \to \mathbb{R}^n$,

$$TV_{m,n}(g) = (\sum_{i=1}^{n} TV_{m,1}^2(g^i))^{1/2}.$$

We should emphasize that this approach is quite distinct from the thin-plate and iterated Laplacian penalties described, for example, in Wahba (1990) for L_2 smoothing splines. In imaging applications the (range) coordinates of g are typically color intensities, e.g. red, green and blue. However, for our purposes it is natural to view these coordinates as components of the gradient, ∇g. Thus, for $g : \mathbb{R}^m \to \mathbb{R}$, we may consider the roughness penalty,

$$P_{1,m}(g) = TV_{2,2}(\nabla g),$$

so for $m = 2$,

$$P_{1,2}(g) = [(\int (g_{11}^2 + g_{12}^2)^{1/2} dx)^2 + (\int (g_{21}^2 + g_{22}^2)^{1/2} dx)^2]^{1/2}$$

This expression looks somewhat formidable, but it turns out to have a relatively benign interpretation in the specialized context of triograms. Consider the trapezoidal region consisting of two adjacent triangles. We would like to evaluate the roughness penalty on this region, by which we mean on the interior of this region, We are free to orient the region so the common edge of the two triangles lies along the x-axis and we will suppose, without loss of generality that our function takes the value 1 at the right end of this edge and takes the value 0 at all 3 other vertices. By definition our triogram is just the function that linearly interpolates on the interior of this set, but this causes some apparent difficulties for our definition since it suggests that $g_{ij}(x) = 0$ almost everywhere and is undefined on the crucial common edge.

Fortunately, these difficulties can be smoothed over by replacing the sharp edge by something rounded. Imagine a small cylinder fitting just inside the angle made by the edge, so a smoothed version of the function may be defined by replacing the sharp Λ of the original by the rounded upper arc of the cylinder and continuing in the usual linear fashion after the points of tangency. For those of a certain age, the mental image of a structure composed of interlocking triangular surfaces connected by those famous Tinker-Toy dowels may induce a Proustean epiphany. This Tinker-Toy model is a simplistic example of the class of ridge function models described in an important recent paper by Pinkus (1997). These ridge functions bring together a number of important ideas in the approximation theory literature as well as important statistical ideas involving projection pursuit and neural networks.

Now, reconsidering $P_{1,2}(g)$ we see that $g_{11} = g_{12} = g_{21} = 0$, and we have,

$$P_{1,2}(g) = \int |g_{22}| dx = |g_2(x, y+\epsilon) - g_2(x, y-\epsilon)||\text{length of edge}|.$$

The first term is independent of x, and also independent of ϵ as long as ϵ is greater than the diameter of the approximating cylinder. Thus, the roughness of the piecewise linear function on the trapezoid is simply the absolute difference in the slope of the function on the two triangles in the direction orthogonal to the edge multiplied by the length of edge. Note that the diameter of the approximating cylinder is irrelevant. In the next subsection we elaborate on this form of the roughness penalty and begin to analyze its suitability for triogram applications.

3.2. **Penalized Triograms.** It would be desirable at this stage to build a rigorous foundation for the roughness penalty $P_{1,2}$ introduced above from the existing materials of geometric measure theory. This may indeed yet prove to be feasible; Mizera (1998) constitutes an impressive beginning. But for the moment a more pragmatic mood must prevail, and I will merely describe in somewhat more detail how $P_{1,2}$ may be used in conjunction with triograms as an alternative to the model selection approach of Hansen, Kooperberg and Sardy.

Fix the triangulation, Δ, and consider the triogram, $g \in \mathcal{G}$ on a specified $\delta \in \Delta$. Let $\{(x_i, y_i, z_i),\ i = 1, 2, 3\}$ denote the three vertices of δ, so,

$$z_i = \theta_0 + \theta_1 x_i + \theta_2 y_i \qquad i = 1, 2, 3,$$

where θ denotes a vector normal to the plane representing the triogram restricted to δ. Solving the linear system, we obtain the gradient vector,

$$\nabla g_\delta = \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} = [\det(D)]^{-1} \begin{pmatrix} (y_2 - y_3) & (y_3 - y_1) & (y_1 - y_2) \\ (x_3 - x_2) & (x_1 - x_3) & (x_2 - x_1) \end{pmatrix} \begin{pmatrix} z_1 \\ z_2 \\ z_3 \end{pmatrix}$$

where D is the 3 by 3 matrix with columns $[1, x, y]$. This gradient is obviously constant on δ and linear in the values of the function at the vertices. Thus, for any pair of triangles, δ_i, δ_j, with common edge, $e_{k(i,j)}$ we have the constant gradients, $\nabla g_{\delta_i}, \nabla g_{\delta_j}$, and we can define the contribution of the edge to the total roughness of the function as,

$$|c_k| = |\eta'_{ij}(\nabla g_{\delta_i} - \nabla g_{\delta_j})| \; ||\text{length } e_{k(i,j)}||,$$

where η_{ij} denotes the unit vector orthogonal to the edge. The penalty is then computed by summing these contributions over all edges.

We have already seen that any function, $g \in \mathcal{G}$, may be expressed in terms of the barycentric basis functions and the values, γ_i, it takes at the vertices of the triangulation as,

$$g(x, y) = \sum_{i=1}^{m} \gamma_i B_i(x, y).$$

We will express the fidelity of the fitted function, $\hat{g}(x, y)$ to the observed sample, $\{(x_i, y_i, z_i), i = 1, ..., n\}$ in ℓ_1 terms as,

$$\sum_{i=1}^{n} |z_i - \hat{g}(x_i, y_i)| = \sum_{i=1}^{n} |z_i - a_i' \hat{\gamma}|$$

where the m-vectors, a_i denote the "design" vectors with elements, $a_{ij} = (B_j(x_i, y_i))$. In the simplest case there is a vertex at every point (x_i, y_i) and the matrix, $A = (a_{ij})$ is just the n-dimensional identity. Typically, however, one would like to choose, $m < n$ and there would be a need to compute some barycentric coordinates for the matrix, A.

Similarly, since the gradient terms are linear in the parameters, $\hat{\gamma}_i$, determining the function at the vertices, the penalty may also be expressed as a piecewise linear function of these values,

$$\sum_k |c_k| = \sum_k |h_k' \hat{\gamma}|,$$

where the index k runs over all of the edges formed by the triangulation, Δ.

The problem of optimizing the fidelity of the fitted function subject to a constraint on the roughness of the function, may thus be formulated as the augmented linear ℓ_1 problem,

$$\min_{\gamma \in \mathbb{R}^m} \sum_{i=1}^{n} |z_i - a_i' \gamma| + \lambda \sum_{k=1}^{M} |h_k' \gamma|.$$

3.3. **On Triangulation.** Up to this point we have taken the form of the triangulation, Δ, as fixed, it is now time to consider how to determine Δ given the observations, $\{(x_i, y_i, z_i),\ i = 1, ..., n\}$. In full generality, as we have already suggested, this is an extremely challenging problem that involves a delicate consideration of the function being estimated. This draws us back into the vertex insertion/deletion schemes like those described by Hansen, Kooperberg, and Sardy. Since it was our intention from the beginning to circumvent these aspects of the problem, replacing such model selection strategies by shrinkage governed by our proposed roughness penalty, we will focus on the classical triangulation method of Delaunay.[2]

[2]Б.Н. Делоне (1890-1973) a leading Russian authority on the theory of numbers; not to be confused with Robert Delaunay (1885-1941), the French painter and proponent of Orphism, a technique

A simple, direct characterization of the Delaunay triangulation may be stated for points in general position in the plane. We will say that points in $\mathbb{R}^2$ are in general position if no three points lie on a line, and no four points lie on a circle. The Delaunay triangulation of a set of points $V = \{v_i \in \mathbb{R}^2 : i = 1, ...n\}$ in general position consists of all triangles whose circumscribing circle contains no V-points in their interior. There is a vast literature on how to compute the Delaunay triangulation. In this paper I employ the Splus implementation of Turner (1996) based on the algorithm by Lee and Schacter (1980).

Delaunay triangulation abhors long, thin triangles. Indeed, another way to characterize the Delaunay triangulation is that it maximizes,

$$\alpha(\Delta) = \min_{\delta \in \Delta}\{a(\delta)\}$$

over all possible triangulations, Δ of the set V, where the scalar, $a(\delta)$ denotes the smallest angle of the triangle, δ. This maxmin property was long considered a major virtue of the Delaunay method. However, relatively recently it has been pointed out by Rippa (1992) that the benefits of this phobia about thinness are strongly linked to the eventual application of the triangulation. If, for example, the objective is to find a good interpolant for a function whose curvature happens to be very large in one direction and small in the other, then long thin triangles may be very advantageous.[3]

To see this consider the example suggested by Rippa. We have a quadrilateral $\mathcal{U}$, with vertices at $(\pm\alpha, o)$ and $(0, \pm\beta)$. Suppose we want to fit the quadratic function,

$$F(x, y) = \frac{1}{2}(x, y)H(x, y)'$$

and we would like to compare the performance of the piecewise linear interpolant $\hat{F}_n$ using the horizontal triangulation Δ_h with $\hat{F}_v$ using the vertical triangulation Δ_v. Rippa's performance measure is integrated squared error

$$ISE(\hat{F}_i, F, \mathcal{U}) = \int_{\mathcal{U}} (\hat{F}_i(x, y) - F(x, y))^2 dxdy$$

On a single triangle, δ, with vertices v_1, v_2, v_3 a result of Nadler gives

$$\int_{\delta} (F - \hat{F}_\delta)^2 dxdy = \frac{|\delta|}{180}[(d_1 + d_2 + d_3)^2 + d_1^2 + d_2^2 + d_3^2]$$

of isolating regions of pure colors in painting that occasionally achieves the appearance of Delaunay triangulation.

[3]This is related to the following observation by Bern and Eppstein (1992). If you consider lifting the points in V onto the paraboloid mapping (x, y) to $(x, y, x^2 + y^2)$, the convex hull of the lifted points can be split into an upper portion and a lower portion. A face of the convex hull belongs to the lower portion if it is supported by a plane that separates the points in V from the point $(0, 0, -\infty)$. The projection of this lower portion of the hull onto the xy-plane is the Delaunay triangulation.

where $|\delta|$ denotes the area of δ and

$$d_i = \frac{1}{2}(v_{i+1} - v_i)' H (v_{i+1} - v_i) \qquad i = 1, 2, 3$$

and $v_4 \equiv v_1$. Rippa then writes

$$ISE(\hat{F}_h, F, \mathcal{U}) = ISE(\hat{F}_v, F, \mathcal{U}) + \frac{2\alpha\beta}{15}[\beta^4 H_{22} - \alpha^4 H_{11}]$$

and concludes that Δ_h is better than Δ_v iff

$$\frac{|H_{11}|}{|H_{22}|} < \left(\frac{\beta}{\alpha}\right)^2.$$

It seemed interesting to explore how this result might change if one replaced integrated *squared* error by integrated *absolute* error. In Appendix A it is shown that one may write

$$\int_\delta |F - \hat{F}_\delta| dx dy = \frac{|\delta|}{6} |d_1 + d_2 + d_3|$$

and that the comparison of the two triangulations reduces to exactly the same expression as in the $\mathcal{L}_2$ case.

4. Examples

In the first set of examples I have generated (x_i, y_i)'s as independent uniforms on $[-1, 1]^2$, and the z_i as,

$$z_i = \max\{1/4 - 1/2(x_i^2 + y_i^2), 0\} + u_i,$$

where the u_i are iid Gaussian. The sample size is taken to be $n = 100$, and the fits are based on the Delaunay triangulation of all n points. With $n = 100$ the number of Delaunay edges is roughly 300, so the resulting ℓ_1 regression problems are roughly 400 by 100. Even so, the fitting is quite quick.

In Figure 4.1 we illustrate the Delaunay triangulation and 5 different fits corresponding to various values of the smoothing parameter λ. The first panel depicts the triangulation; the arrows emanating from each edge represent the gradient gap along that edge, so their direction indicates the direction of steepest ascent. Each fit in the remaining five panels has an indicated value of λ and the value of the fidelity term and roughness measure. The value $\lambda = .01$ corresponds to a fit that (almost) interpolates the data, so the fidelity in this case is almost zero, indicating a perfect fit. The penalty is, of course, very large in this case. As λ increases fidelity deteriorates, but roughness improves. By the time we get to $\lambda = 1$, the fit is clearly too smooth, and we have lost the basic features of the data.

In Figure 4.2 we illustrate the same example, but this time the fitted function is depicted as a contour plot. Again we wee that low values of the parameter λ yield very rough contours, while $\lambda = 1$ appears over-smoothed, and $\lambda = .5$ looks about right.

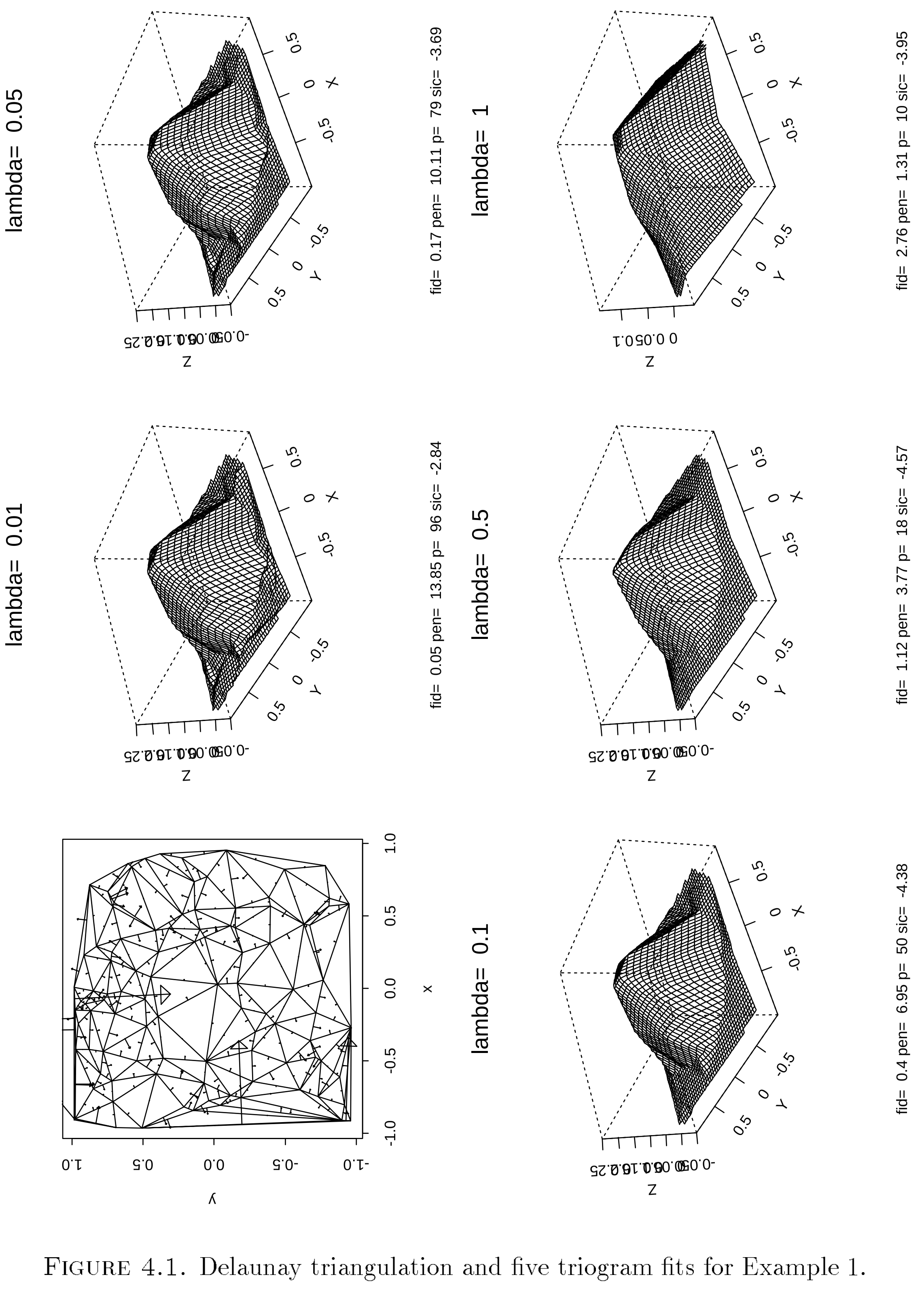

FIGURE 4.1. Delaunay triangulation and five triogram fits for Example 1.

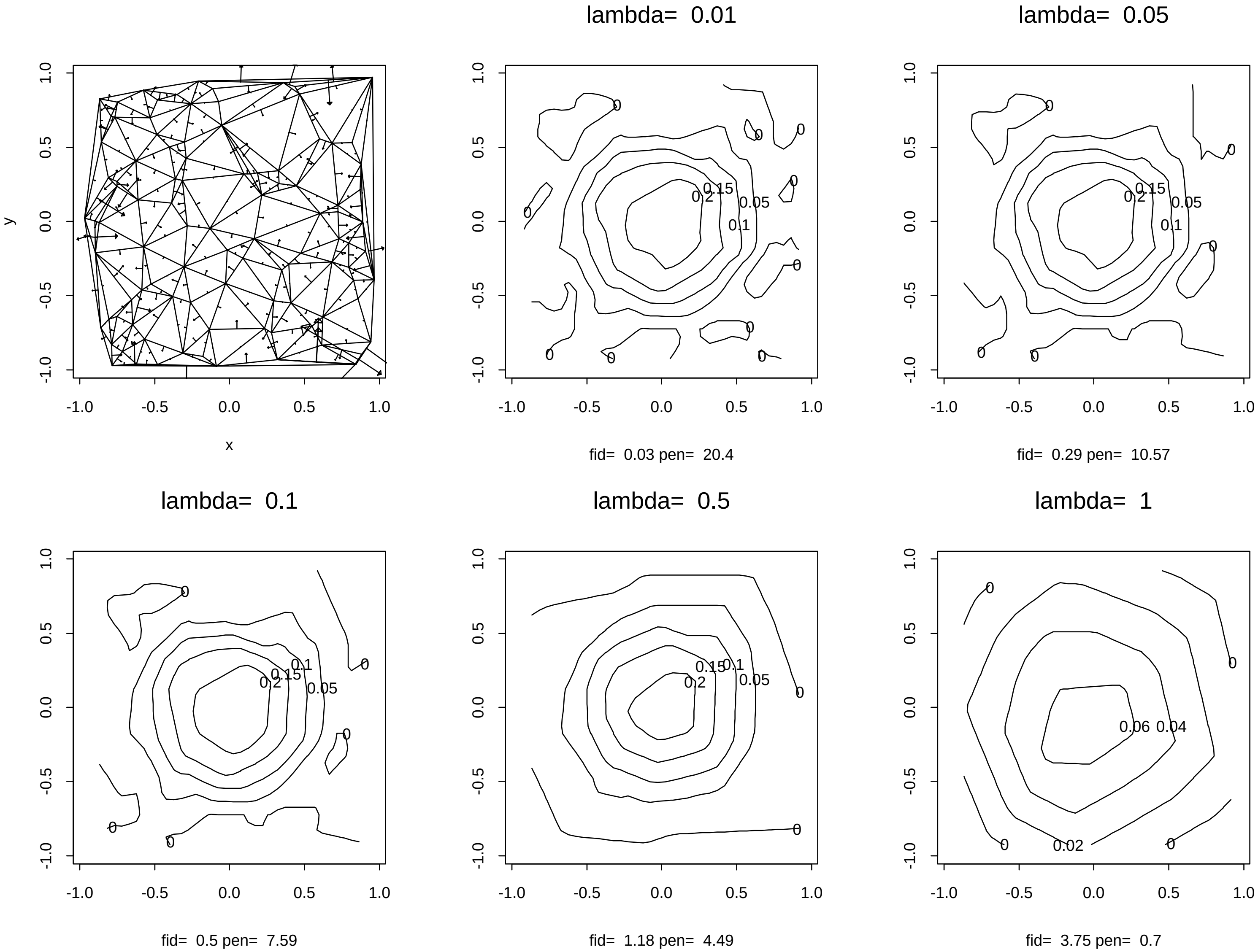

FIGURE 4.2. Delaunay triangulation and five triogram fits for Example 1.

4.1. **Automatic λ Selection.** In Koenker, Ng and Portnoy it was suggested that a variant of the well-known Schwarz (1978) model selection criterion could be used to automatically select λ. This suggestion can also be adapted to the present context in the following way. Given a fit $\hat{g}_\lambda(\cdot,\cdot)$ for a specified λ, we would like to have a reasonable measure of the dimension of the fitted function. As with other ℓ_1 type estimation methods, this is provided by simply counting the number of interpolated observations in the fidelity component of objective function. Letting $p(\lambda)$ denote the dimension of $\hat{g}_\lambda$ defined in this way, then we will consider selecting λ to minimize,

$$\log(n^{-1}\sum \rho_\tau(z_i - \hat{g}_\lambda(x_i, y_i))) + \tfrac{1}{2}n^{-1}p(\lambda)\log n.$$

Unfortunately, this standard form of the Schwarz criterion seems to have problems in the present instance, at least when the signal to noise ratio is high. In applications like our previous example, the standard form of the Schwarz criterion is often monotone increasing in λ indicating that the fit should be selected "as rough as possible". Similar behavior can be seen in fitting cubic smoothing splines and selecting λ using GCV. In this case I have explored the possibility of an ad-hoc "small-sample adjustment" of the first term, so the criterion takes the form,

$$\log((n - p(\lambda))^{-1}\sum \rho_\tau(z_i - \hat{g}_\lambda(x_i, y_i))) + \tfrac{1}{2}n^{-1}p(\lambda)\log n.$$

In this form, the criterion consistently selects a λ in the range, $[0.2, 0.5]$, values that yield quite a reasonable looking fit.

5. Conclusions

There are many possible lines of development for penalized triograms: from fundamental questions about the geometric measure theory of total variation of vector valued functions, to problems of evaluating asymptotic performance, to issues of algorithmic design. Leaving such matters to subsequent investigation, I will conclude – in the spirit of pragmatic goniolatry – by metaphorically nailing the following manifesto for penalized triograms to the provisional conclusion of this draft of the paper.

9.5 Theses on Goniolatry

1. Triograms constitute an useful tool for 3D nonparametric statistical modeling.
2. Adaptive model selection over competing triograms is hard.
3. Shrinkage methods based on global measures of function roughness is easier, because of convexity.
4. Penalized triograms may be useful for model selection, as suggested by Tibshirani's work on the lasso.
5. For functions locally characterized as "ridge functions" the penalty, $P_{1,2}(\nabla g)$ is natural and easily implemented.

6. Penalized triograms are easily adapted for use as bivariate regression quantile smoothing splines.
7. Similar meethods can be adapted for density and hazard estimation using the penalized likelihood methods.
8. Extensions to higher dimensions are possible by replacing triangles by simplices.
9. Automatic selection of the λ can be implemented along the lines suggested by Schwarz(1978)

9.5 The penalty, $P_{1,2}(\nabla g)$ has some claim as a natural *statistical* measure of roughness.

Appendix A

Consider the problem of interpolating the quadratic function

$$F(x,y) = \frac{1}{2}(x,y)H(x,y)'$$

by a piecewise linear function on the triangle, δ, with vertices $\{(0,0),(1,0),(u,v)\}$. We will assume the matrix H is positive definite and $(u,v) \in \Re^2_+$. Since the approximating function, $\hat{F}$, must agree with F at the vertices, we have

$$\hat{F}(x,y) = \frac{1}{2}H_{11}x + v^{-1}(f(u,v) - uH_{11}/2).$$

Then

$$\begin{aligned} I &= \int\int_\delta |\hat{F} - F| dxdy \\ &= \int_0^u \int_0^{xv/a} (\hat{F} - F)dydx + \int_u^1 \int_0^{v(1-x)/(1-u)} (\hat{F} - F)dydx \end{aligned}$$

since $\hat{F}(x,y) \geq F(x,y)$ on δ. Direct evaluation yields,

$$I = \frac{v}{24}(H_{11}(1 - u + u^2) + v(H_{12}(2u-1) + vH_{22}).$$

Now let $d_i = f(x_{i+1} - x_i, y_{i+1} - y_i) \quad i = 1,2,3$ where (x_i, y_i) denotes the i^{th} vertex of δ, and $(x_4, y_4) = (x_1, y_1)$. Then,

$$\begin{aligned} \sum_{i=1}^{3} d_i &= f(1,0) + f(u-1,v) + f(-u,-v) \\ &= H_{11}(1 - u + u^2) + vH_{12}(2u-1) + H_{22}v \end{aligned}$$

The area of δ is $v/2$ so we may express I as

$$I = \frac{|\delta|}{12}\sum_{i=1}^{3} d_i.$$

Finally, we would like to compare the integrated absolute error on the competing triangulations $\Delta_1 = \{(\pm\alpha, 0), (0,\beta)\}$ and $\Delta_2 = \{(\alpha,0),(0,\pm\beta)\}$. This comes down to comparing

$$I(\delta_1) = f(2\alpha, 0) + f(-\alpha,\beta) + f(-\alpha,-\beta)$$

and

$$I(\delta_2) = f(\alpha,\beta) + f(-\alpha,\beta) + f(0,-2\beta)$$

and we conclude that Δ_h is preferred to Δ_v, just as in the $\mathcal{L}_2$ case, iff

$$\frac{|H_{11}|}{|H_{12}|} < \left(\frac{\beta}{\alpha}\right)^2$$

REFERENCES

BERN, M., AND D. EPPSTEIN (1992): "Mesh generation and optimal triangulation," in *Computing in Euclidean Geometry*, ed. by D. Du, and F. Hwang, pp. 23–90. World Scientific Publishing.

BLOMGREN, P., AND T. CHAN (1998): "Color TV: Total variation methods for restoration of vector valued images," *IEEE Trans. Image Proc.*, 7, 304–309.

BOSCH, R. J., Y. YE, AND G. G. WOODWORTH (1995): "A Convergent Algorithm for Quantile Regression With Smoothing Splines," *CmpStDA*, 19, 613–630.

DONOHO, D., S. CHEN, AND M. SAUNDERS (1998): "Atomic decomposition by basis pursuit," *SIAM J. of Scientific Computing*, 20, 33–61.

GREEN, P. J., AND B. W. SILVERMAN (1994): *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach.* ChapmnHall:Lnd.

HANSEN, M., C. KOOPERBERG, AND S. SARDY (1998): "Triogram Models," *JASA*, 93, 101–119.

HE, X., P. NG, AND S. PORTNOY (1998): "Bivariate quantile smoothing splines," *JRSS(B)*, 60, 537–550.

KOENKER, R., P. NG, AND S. PORTNOY (1994): "Quantile Smoothing Splines," *Biomtrka*, 81, 673–680.

LEE, D., AND B. SCHACTER (1980): "Two algorithms for constructing a Delaunay triangulation," *International Journal of Computer and Information Sciences*, 9, 219 – 242.

MEINGUET, J. (1979): "Multivariate interpolation at arbitrary points made simple," *ZAMP*, 30, 292–304.

MIZERA, I. (1998): "Variations on Variations," personal communication.

PINKUS, A. (1997): "Approximating by Ridge Functions," in *Surface fitting and multiresolution methods*, ed. by A. Méhauté, C. Rabut, and L. Schumaker, pp. 1– 14.

PYNCHON, T. (1997): *Mason and Dixon.* Henry Holt.

REINSCH, C. (1967): "Smoothing by spline functions," *Numerische Matematik*, pp. 177–183.

RIPPA, S. (1992): "Long and thin triangles acn be good for linear interpolation," *SIAM J. Numer. Anal.*, 29, 257–270.

SCHWARZ, G. (1978): "Estimating the Dimension of a Model," *AnlsStat*, 6, 461–464.

STONE, C., M. HANSEN, C. KOOPERBERG, AND Y. TROUNG (1997): "Polynomial splines and their tensor products in extended linear modeling," *Annals*, 25, 1371–1470.

STONE, C. J. (1994): "The Use of Polynomial Splines and Their Tensor Products in Multivariate Function Estimation (Disc: P171-184)," *AnlsStat*, 22, 118–171.

TIBSHIRANI, R. (1996): "Regression Shrinkage and Selection Via the Lasso," *JRSS-B*, 58, 267–288.

TURNER, R. (1996): "Splus Functions for Delaunay triangulation," `http://www.stat.cmu.edu/S`.

WAHBA, G. (1975): "Smoothing Noisy Data by Spline Functions," *Numerische Matematik*, 24, 383–393.

WAHBA, G. (1990): *Spline Models for Observational Data.* SIAM.