PENALIZED TRIOGRAMS: TOTAL VARIATION REGULARIZATION FOR BIVARIATE SMOOTHING

ROGER KOENKER AND IVAN MIZERA

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 distinct advantages over competing tensor product methods that are more commonly used in statistical applications.

Triograms employ basis functions consisting of linear "tent functions" defined with respect to a triangulation of a given planar domain. As in knot selection for univariate splines, Hansen, *et al* adopt the regression spline approach of Stone (1994). Vertices of the triangulation are introduced or removed sequentially in an effort to balance fidelity to the data and parsimony.

In this paper we explore a smoothing spline variant of the triogram model based on a simple roughness penalty adapted to the piecewise linear structure of the triogram model. The proposed roughness penalty may be interpreted as a total variation penalty on the gradient of the fitted function.

"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 instrument, splines are a elegantly simple, yet eminently practical tool. In the statistical literature on splines there continues to be a vigorous debate over the relative merits of penalty methods for smoothing splines, versus regression splines relying on knot selection. Both computational tractibility and statistical efficiency play important roles in this debate, and the resulting rivalry has significantly broadened the scope of both approaches.

Version: August 8, 2001. A preliminary version of this paper was prepared for a conference in honor of George Judge held at the University of Illinois, May 1-2, 1999. The research was partially supported by NSF SBR-9617206 and SES-99-11184. The authors would like to thank Steve Portnoy and Xuming He for many helpful discussions, and Peter Cowell and Henry Munneke for providing the Chicago land value data.

In an innovative recent paper Hansen, Kooperberg, and Sardy (1998) have introduced a class of linear spline models for bivariate smoothing problems. These triogram models are defined on triangulations of polyhedral planar domains; knot selection strategies adapted from Stone, Hansen, Kooperberg, and Troung (1997) are employed to control the degree of smoothing of the estimates. The primary objective of the present paper is to begin to explore a smoothing spline approach to the estimation of triograms. The roughness penalty we employ may be viewed as an attempt to extend the total variation roughness penalty suggested in Koenker, Ng, and Portnoy (1994) to bivariate settings.

2. Roughness Penalties and Nonparametric Regression

In its classical univariate form the (cubic) smoothing spline solves the problem of finding a function g minimizing

$$\sum_{i=1}^{n}(y_i - g(x_i))^2 + \lambda \int (g''(x))^2 dx,$$

over a Sobolev space of continuous functions with absolutely continuous first derivative and square-integrable second derivative. The tuning parameter λ controls the smoothness of the fitted function. In this form the estimator $\hat{g}(\cdot)$ is a natural cubic spline with knots at the observed x_i's and may be interpreted as an estimate of the conditional mean function. The penalty term may be viewed as representing a prior belief that the L_2 norm of g'' is unlikely to exceed a specified bound controled by the choice of λ.

2.1. Total variation roughness in the one-dimensional case.

There is nothing sacred about the Gaussian, conditional mean, formulation of the smoothing spline problem and there have been numerous efforts to explore alternative forms of both the fidelity and roughness penalties to achieve modified objectives. One such effort is described in Koenker, Ng, and Portnoy (1994), where a non-parametric approach to estimating conditional quantile functions is suggested based on g minimizing

$$\sum_{i=1}^{n} \rho_\tau(y_i - g(x_i)) + \lambda J(g), \tag{2.1}$$

where $\rho_\tau(u) = u(\tau - I(u < 0))$ generates a fidelity term appropriate for conditional quantile estimation, and the roughness penalty $J(g)$ is taken to be total variation of the first derivative of g.

Recall that the total variation of a function f from $[a, b]$ to $\mathbb{R}$ is given by

$$V(f) = \sup \sum_{k=0}^{n-1} |f(x_{k+1}) - f(x_k)|, \tag{2.2}$$

where the sup is taken over all possible partitions, $a = x_0 < x_1 < ... < x_n = b$ For a continuous function $f : \mathbb{R} \to \mathbb{R}$, the celebrated Banach (1925) indicatrix theorem gives

$$V(f) = \int N(y)dy, \tag{2.3}$$

where $N(y) = \#\{x : f(x) = y\}$ is the Banach indicatrix of f, the function counting the number of roots for each value in the range of f; see *e.g.* Natanson (1974, Thm VIII.5.3). If f is absolutely continuous, we can also write, again see Natanson (1974, Thm IX.4.8),

$$V(f) = \int |f'(x)|dx, \tag{2.4}$$

which for $f = g'$ yields the roughness penalty

$$J(g) = V(g') = \int |g''(x)|dx. \tag{2.5}$$

This establishes a clear link of the total variation penalty of Koenker, Ng, and Portnoy to the classical L_2 roughness penalty.

Total variation proves to be a natural alternative penalty for quantile regression fidelity from a computational viewpoint since it preserves the piecewise linear form of the objective function and thus preserves the linear programming formulation of the optimization problem. Solutions to the problem (2.1) 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 imposes 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 more 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 questions about how to measure the roughness of surfaces. The thin plate smoothing splines of Harder and Desmarais (1972) whose theory was developed by Duchon (1976,1977), Meinguet (1979) and Wahba and Wendelberger (1980), and others, minimize

$$\sum_{i=1}^{n}(z_i - g(x_i, y_i))^2 + \lambda J(g, \Omega, \| \cdot \|_2^2) \tag{2.6}$$

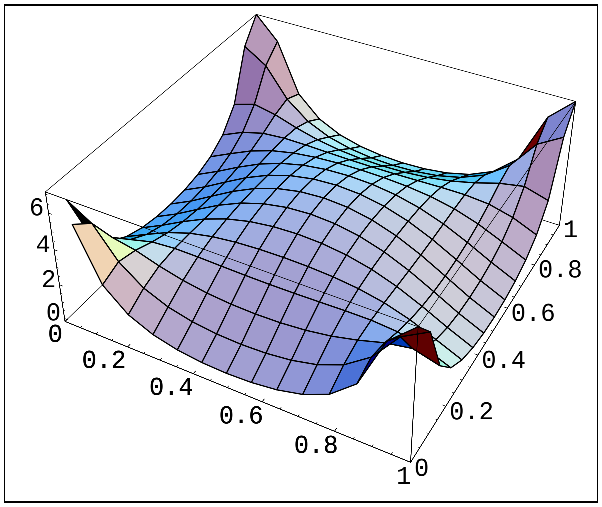

Figure 2.1. Thin plate penalty integrand for the He, Ng Portnoy tent function interpolant.

with the roughness penalty defined as,

$$J(g, \Omega, \| \cdot \|_2^2) = \iint_{\Omega} \| \nabla^2 g \|_2^2 dxdy = \iint_{\Omega} (g_{xx}^2 + 2g_{xy}^2 + g_{yy}^2) dxdy \tag{2.7}$$

The integrand of the thin plate penalty is the squared Hilbert-Schmidt (Frobenius) norm of the Hessian of g. This dependence on the norm is explicitly recognized in our penalty notation in anticipation of taxonomic challenges that lie ahead. In the classical thin-plate problem, Ω is taken to be all of $\mathbb{R}^2$, and this simplifies the computations considerably. However, as noted by Green and Silverman (1994) there can be considerable disparities between such solutions and solutions based on versions of the penalty defined over restricted domains.

This is illustrated in the canonical example of He, Ng, and Portnoy (1998) of interpolating the four points, $(0, 0, 0), (1, 0, 0), (1, 1, 1), (1, 0, 0)$ forming a tent on the unit square. Solving (2.6) with $\Omega = \mathbb{R}^2$ yields a function whose thin plate integrand is illustrated in Figure 2.1. When integration is restricted to the unit square the thin plate penalty is roughly 2.77; this corrects the computation reported in He, Ng, and Portnoy (1998). But simpler candidates can yield considerably smaller penalties, *e.g.* $g(x, y) = xy$, gives $J(g, [0, 1]^2, || \cdot ||_2^2 = 2)$.

I would be willing to conjecture that xy is optimal. Do you know somebody who could solve it as a homework?

If $g(x, y) = h(x)$ for some h, then a straightforward computation shows that, on rectangular $\Omega = \Omega_1 \times \Omega_2$,

$$J(g, \Omega, || \cdot ||_2^2) = J(h, \Omega_1, || \cdot ||_2^2)\ \mu(\Omega_2), \tag{2.8}$$

where $J(h, \Omega_1, || \cdot ||_2^2)$ specializes to the classical univariate penalty $\int (g''(x))^2 dx$, and $\mu(\Omega_2)$ denotes the Lebesgue measure of Ω_2. Thus, the thin plate penalty (2.6) may be viewed as a natural bivariate extension of the classical univariate roughness penalty.

This raises the following questions. Can we, by analogy with the univariate total variation penalty (2.5), define a bivariate roughness penalty? How should we define total variation of the gradient of a function of two variables? These questions require a brief mathematical detour.

2.2. **Total Variation in Higher Dimensions.** The quest for a satisfactory definition of total variation for functions from $\mathbb{R}^k$ to $\mathbb{R}^m$ has engaged the mathematical community for more than a century. Only for $k = 1$, and m arbitrary, does the classical univariate definition (2.2) of Jordan (1881) adapt in a straightforward way, see Dinculeanu (1967). Early definitions for $k \geq 2$ and $m = 1$ by Tonelli (1926, 1936), and others suffered from coordinate-dependence and attendent reliance on rectangular domains. In nonparametric regression this is a drawback, as we will argue below. The first orthogonally-invariant definitions were introduced by Kronrod (1949, 1950) in the spirit of the Banach indicatrix theorem (2.3). For a real function f on $\Omega \subseteq \mathbb{R}^2$, the Kronrod variation is equal to

$$V_K(f, \Omega) = \int_{\Omega} \mathcal{L}^1(f^{-1}(x)))\, dx, \tag{2.9}$$

where $\mathcal{L}^1(f^{-1}(x))$ is the length, defined as one-dimensional Hausdorff, or similar measure, of the preimage of x and dx denotes the integration over the one-dimensional Lebesgue measure so that we need not care if the length of the preimage is infinite for sets of zero measure. This line of development for arbitrary k and $m = 1$ was subsequently cultivated in Russian literature by Vitushkin (1955) and Ivanov (1975).

However, a definition for general k and $m > 1$, requires a reconsideration of (2.4). We can illustrate this approach in the simple univariate case. The definition of the total variation is given in two steps: for smooth f from $[a, b]$ to $\mathbb{R}$ via (2.4), and then for more general f via extension. The left panel of Figure 2.2 shows several smooth approximants, f^{ν}, to a simple univariate jump function f. The total variation of f can be obtained by a limit transition from total variations of its smooth approximants.

There are numerous pitfalls on this path. One is illustrated in the right panel of Figure 2.2. It is necessary to take liminf rather than simple limits, and the mode of approximation must be formalized properly. Rigorous develepoment inevitably invokes portions of the theory of Schwartz distributions. As in the theory of Sobolev spaces, the formalism of distributions is needed for differentation and limit transitions; the functions under consideration remain standard. This approach to multidimensional total variation dominates the recent mathematical literature; Ambrosio, Fusco, and Pallara (2000) give a recent account of the theory developed in the context of geometric measure theory and variational calculus, tracing its origins back to Fichera (1954) and De Giorgi (1954).

As an initial step, it is convenient to outline the functional domain in a qualitative way, without recourse to any particular total variation functional. Functions with bounded variation are defined to be those whose derivatives, in the sense of

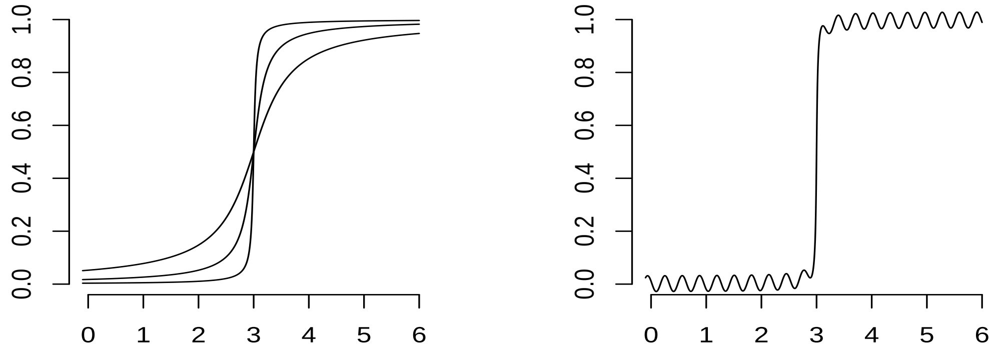

FIGURE 2.2. Mollification of the elementary jump function.

Schwartzian distributions, are measures. For a smooth function f from $\mathbb{R}^k$ to $\mathbb{R}^m$, we define, in the vein of (2.4),

$$V(f, \Omega, \|\cdot\|) = \int_\Omega \|\nabla f\|\, dx; \tag{2.10}$$

here dx denotes (multiple) integration with respect to k-dimensional Lebesgue measure.

A lower semicontinuous functional J initially defined for smooth functions can be extended to a broader domain using the approach of Serrin (1961). We define

$$J(g) = \liminf J(g^\nu), \tag{2.11}$$

where the right-hand side expression denotes the inf of $\liminf J(g^\nu)$ over all sequences g^ν approaching g in the sense of distributions. (The use of liminf is essential as shown in Figure 2.2.) Since smooth functions are dense, with respect to distributional convergence, in functions of bounded variation, and a total variation functional of the form (2.10) is lower semicontinuous – this property is related to its convexity – the extension step concludes the definition. We omit, however, the proof of the fact that the extension is finite for any functional of type (2.10). We only remark that this property is not automatic, as our next example shows.

Consider the thin-plate functional (2.7); it is lower semicontinuous, thus there is some hope that it can be extended beyond its traditional domain. It turns out, however, that such an extension assigns $+\infty$ to any g with discontinuous derivatives. In particular, any function with a spike or a sharp ridge is evaluated as infinitely rough. Serrin (1961) gives arguments why such an outcome is essentially unavoidable, in a sense, not depending on a particular extension scheme. The reader may verify the conclusion in dimension one and thus for any function for which formula (2.8) applies. A jump in derivative with magnitude 1 interpolated by a piecewise linear function increasing on an interval of length $2r$ results in the thin-plate functional of order $(2r)^{-1}$; letting $r \to 0$ makes this infinite. Another example is the cone $g(x,y) = (x^2+y^2)^{1/2}$ on the unit circle; a straightforward computation involving polar coordinates gives $J(g, \Omega, \|\cdot\|_2^2) = \int_0^1 r^{-1} dr = +\infty$. Here, we have neglected the contribution of the spike itself, which can be shown to be finite and nonzero.

2.3. **Roughness penalties based on total variation.** Given our definition of total variation, we are now prepared to define a total-variation based roughness penalty. For a function, g, from $\mathbb{R}^2$ to $\mathbb{R}$ we define

$$J(g, \Omega, \|\cdot\|) = V(\nabla g, \Omega, \|\cdot\|) = \iint_\Omega \|\nabla^2 g\| \, dx \, dy, \tag{2.12}$$

Any such penalty—regardless of the choice of the norm—can be considered an extension of the univariate penalty (2.5).

Theorem 2.1. *Suppose that g is a function from $\mathbb{R}^2$ to $\mathbb{R}$ such that $g(x,y) = h(x)$ for some h. There is a constant c depending only on the choice of the matrix norm in* (2.12)*, but not on g, such that for any $\Omega = \Omega_1 \times \Omega_2$,*

$$J(g, \Omega, \|\cdot\|) = c\, J(h, \Omega_1, \|\cdot\|)\ |\Omega_2|, \tag{2.13}$$

where $J(g, \Omega, \|\cdot\|) = \int_{\Omega_1} |h''(x)| \, dx$, and $|\Omega_2|$ denotes the Lebesgue measure of Ω_2.

Proof: Let c be the norm of the 2×2 matrix containing 1 in the upper left corner and zeros elsewhere. By the properties of the norm, the norm of the matrix containing u instead of 1 in the upper left corner and zeros otherwise is $c|u|$. Note that in the Hessian, all second-order partial derivatives are zero, except for $g_{xx}(x,y) = h''(x)$; thus

$$J(g, \Omega, \|\cdot\|) = c \iint_\Omega |h''(x)| \, dx \, dy$$

and (2.13) follows by the Fubini theorem for all smotth g and hence by extension for all g under consideration. ■

For denoising images with a view toward reconstructing discontinuities in derivatives, Scherzer (1998) proposed using the penalty corresponding to the ℓ_1 norm in (2.10); for smooth functions g from $\mathbb{R}^2$ to $\mathbb{R}$ this penalty is equal to

$$J(g, \Omega, \|\cdot\|_1) = \iint (|g_{xx}| + 2|g_{xy}| + |g_{yy}|) dx dy. \tag{2.14}$$

Related penalties have been recently proposed in the statistical literature by He, Ng, and Portnoy (1998), who introduced a bivariate form of the quantile smoothing spline using a roughness penalty that sums univariate total variation of the function along rectangular grid lines. Their roughness penalty may be viewed as a total variation of the gradient in the Tonelli-Cesari vein, (2.10), with the ℓ_1 norm applied to the diagonal of the Hessian,

$$J(g, \mathit{\Omega}, \|\cdot\|_{HNP}) = \iint (|g_{xx}| + |g_{yy}|)dxdy, \tag{2.15}$$

Their 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 potential disadvantage of the tensor product formulation in some applications is its lack of orthogonal equivariance. Functions well oriented with respect to the xy-axes may prove to be much more difficult to fit when the observations are rotated. Invariance considerations provide valuable guidance through the forest of potential definitions of total variation and roughness penalty functionals.

2.4. **Invariance and equivariance.** Given a function f from $\mathbb{R}^2$ to $\mathbb{R}$, with gradient vector, ∇f, after an orthogonal change of coordinates, $x = U\xi$, the new gradient of f is equal to $(\nabla f)^\tau U$. Imposing invariance of the total variation functional $V(f, \mathit{\Omega}, \|\cdot\|)$ for any f and U, we arrive at the requirement that $\|Ux\| = \|x\|$ for any $x \in \mathbb{R}^2$ and any orthogonal matrix U. The only norm satisfying this requirement, up to multiplication by a constant, is the Euclidean norm. Therefore, the only coordinate-independent total variation functional is, for $k = 2$, a constant multiple of

$$V(f, \mathit{\Omega}, \|\cdot\|_2) = \iint_{\mathit{\Omega}} \sqrt{f_x^2 + f_y^2}\, dx\, dy. \tag{2.16}$$

Transforming the integral reveals that (2.16) is nothing but the Kronrod (1949) variation (2.9).

If, however, $f = \nabla g$, as we require for our roughness penalties, the definition (2.10) applied to $\nabla g : \mathbb{R}^2 \to \mathbb{R}^2$, we obtain,

$$J(g, \mathit{\Omega}, \|\cdot\|) = V(\nabla g, \mathit{\Omega}, \|\cdot\|) = \iint_{\mathit{\Omega}} \|\nabla^2 g\|\, dx\, dy, \tag{2.17}$$

where $\nabla^2 g$ is the Hessian of g. Invariance with respect to translations comes for free, since we work with derivatives. The requirement of orthogonal invariance for the penalty J, leads to the requirement that

$$\|U^\tau H U\| = \|H\|, \tag{2.18}$$

for any orthogonal matrix U and any symmetric matrix H. There are many norms satisfying this property – apparently any norm which is a symmetric function of the eigenvalues satisfies (2.18). In fact, von Neumann (1937) proved that every norm satisfying $\|A\| = \|UA\| = \|AU\|$ for any A and any unitary matrix U must be a symmetric function of the singular values of A.

The leading example of such a norm is the Hilbert-Schmidt (Frobenius, Euclidean) norm of the matrix. The resulting penalty is, for sufficiently smooth g, given by,

$$J(g, \Omega, \|\cdot\|_2) = \int_\Omega \sqrt{g_{xx}^2 + 2g_{xy}^2 + g_{yy}^2}\, dx\, dy,$$

which brings us back through Pythagorean pathways to the thin-plate penalty. Only the squaring of the norm is at issue. Other possibilities include the spectral norm, the maximal absolute value of the eigenvalues, or absolute value of the trace; we refer the reader to Mizera (2001) for some other intriguing candidates.

Another attractive property of total variation roughness penalties, particularly when paired with absolute error fidelity, is their scale equivariance. If g minimizes

$$\sum_{i=1}^{n} |z_i - g(x_i, y_i)| + \lambda J(g, \Omega, \|\cdot\|) \tag{2.19}$$

then cg minimizes (2.19) with z_i replaced by cz_i, provided that $J(cg, \Omega, \|\cdot\|) = |c| J(g, \Omega, \|\cdot\|)$. This is clearly not the case for the thin-plate penalty. However, for Gaussian fidelity the thin plate penalty is well matched.

Efficient numerical solution of the variational problems arising from such general forms of roughness penalties based on total variation appears quite challenging. However, by restricting the domain of functions over which we are optimizing some progress can me made. One such restriction leads to penalized versions of the piecewise linear triograms of Hansen, Kooperberg, and Sardy (1998).

3. Triograms

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.

The continuous functions g on $\mathcal{U}$ that are linear when restricted to $\delta \in \Delta$ are called triograms. Their collection, $\mathcal{G}$, associated with the triangulation, Δ, is a finite-dimensional linear space space. 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 their simplicity. Hansen, Kooperberg and Sardy propose a regression spline approach to estimating triogram models 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." Selecting a good triangulation is clearly critical to success, and considerable attention needs to be devoted to stepwise addition and deletion strategies for vertices to achieve a "good" choice of Δ. Motivated by the success of penalty methods elsewhere in the spline literature we were encouraged to explore an alternative penalized triogram approach.

3.1. **A Roughness Penalty for Triograms.** Thin-plate penalties are inappropriate for triograms for the reasons described in the previous section: such penalties assign infinity to any function with a discontinuity in the gradient, and thus are inherently incapable of discriminating among triograms. Roughness penalties based on the total variation of the gradient offer a more straightforward solution. Fortuitiously, it also turns out that the troublesome choice of the norm disappears once we insist on a coordinate-independent penalty for triograms, all penalties reduce to a single one.

Theorem 3.1. *Suppose that $g : \mathbb{R}^2 \to \mathbb{R}$, is a piecewise-linear function on the triangulation, Δ. For any coordinate-independent penalty of the form* (2.17)*, there is a constant c dependent only on the choice of the norm such that*

$$J(g, \mathit{\Omega}, \|\cdot\|) = c \sum_e \|\nabla g_e^+ - \nabla g_e^-\| \, \|e\| \tag{3.1}$$

where e runs over all the interior edges of the triangulation $\|e\|$ is the Euclidean length of the edge e, and $\|\nabla g_e^+ - \nabla g_e^-\|$ is the Euclidean length of the difference between gradients of g on the triangles adjacent to e.

Proof: Evaluating J, we split the integration domain $\mathit{\Omega}$ to disjoint pieces whose contribution to J is determined separately. First, the contribution of all linear parts, the interiors of the triangles, is 0; the second derivatives vanish thereon.

The contribution of the edge e is the corresponding term in (3.1): consider the trapezoidal region consisting of two triangles adjacent to the edge. Extend the functions on the triangles linearly to have a rectangular domain—this should not alter the penalty. Coordinatewise independence then allows for rotating the rectangle so that its edges are parallel to xy-axes; the application of Theorem 2.1 then gives the desired result.

The final, and most technical part of the proof is to show that the contribution of any vertex of the triangulation is 0. This is done employing the definition (2.11); the sequence g^ν approximating g is obtained via mollification: g^ν is taken to be the convolution of g with $\nu^2\phi(\nu x, \nu y)$, where ϕ is a smooth function assigning 0 to all values outside the unit circle whose integral is equal to 1; for instance, $\phi(x, y) \propto \exp(-1/(1 - x^2 - y^2))$. When $\nu \to \infty$, g^ν approaches g in the distributional sense. The contribution of the vertex is bounded from above by

$$\liminf \iint_{B_\nu} \| \nabla^2 g^\nu \| \, dx \, dy, \tag{3.2}$$

where B_ν is the circle centered at the vertex with radius $1/\nu$. Since any two norms on a finite-dimensional vector space are equivalent, (3.2) is bounded from above by a constant multiple of (3.2) with the Hilbert-Schmidt norm, the constant depending only on the original norm. In what follows, C stands for a generic constant. Since the derivatives of g in the neighborhood of the vertex are piecewise constant, with finitely many pieces, we have

$$\begin{aligned}|g^{\nu}_{xx}(x,y)| &= \left|\iint \nu^3\phi_x(\nu u,\nu v)g_x(x-u,y-v)\,du\,dv\right|\\ &\le \nu\iint \left|\nu^2\phi_x(\nu u,\nu v)g_x(x-u,y-v)\right|\,du\,dv\\ &\le \nu\iint C\left|\nu^2\phi_x(\nu u,\nu v)\right|\,du\,dv \le C\nu\iint |\phi_x(u,v)|\,du\,dv = C\nu;\end{aligned}$$

the same inequalities hold for the other terms in the Hessian, $\nabla^2 g^{\nu}$, the constants being independent of x, y, and ν. By the properties of the Hilbert-Schmidt norm,

$$\iint_{B_\nu} \|\,\nabla^2 g^{\nu}\|\,dx\,dy \le \iint_{B_\nu} C\nu\,dx\,dy = C\nu^{-1}.$$

The last term goes to 0 when $\nu \to \infty$. Note that for the thin-plate penalty, the bound for the elements of the Hessian will be $C\nu^2$, and the contribution of the vertex does not vanish, though it is finite. ■

The triogram model is an elementary example of the class of ridge function models surveyed in a recent paper by Pinkus (1997). 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. See Candes (2000) for a unified statistical prespective.

4. Computation of Penalized Triograms

4.1. **A Basis for $\mathcal{G}$.** A basis for the linear space $\mathcal{G}$ consists of the linear "tent" functions, $\{B_i(u)\}_{i=1}^p$, that may be expressed in terms of the barycentric coordinates of points u represented by the vertices v_1, v_2, v_3 of the triangle δ containing u,

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

and satisfying the condition

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

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

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

where

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

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; thus they are linearly independent. They are also affine equivariant; that is, for any non-singular, 2×2 matrix A, and vector $b \in \mathbb{R}^2$,

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

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

Terminology here may need to be reconciled with Section 2.

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

4.2. **Computing the Fidelity.** Since any function, $g \in \mathcal{G}$, may be expressed in terms of the barycentric basis functions and the values, β_i, that it takes at the vertices of the triangulation as,

$$g(x, y) = \sum_{i=1}^{p} \beta_i B_i(x, y),$$

we can 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,

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

where the p-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 may wish to choose, $p < n$ and there would be a need to compute some barycentric coordinates for some elements of the matrix, A.

4.3. **Computing the Penalty.** Fix the triangulation, Δ, and consider the triogram, $g \in \mathcal{G}$ on a specified triangle $\delta \in \Delta$. Let $\{(x_i, y_i, z_i),\ i = 1, 2, 3\}$ denote the points at 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,

$$\begin{aligned} |c_k| &= |\eta_{ij}^{\top}(\nabla g_{\delta_i} - \nabla g_{\delta_j})| \ \cdot \|e_{k(i,j)}\|, \\ &= \|(\nabla g_{\delta_i} - \nabla g_{\delta_j})\| \ \cdot \|e_{k(i,j)}\|, \end{aligned}$$

where η_{ij} denotes the unit vector orthogonal to the edge. The second formulation follows from the fact that η_{ij} is just the gradient gap renormalized to have unit length; this can be easily seen by considering a canonical orientation in which the edge $k(i,j)$ runs from $(0,0)$ to $(1,0)$. The penalty is then computed by summing these contributions over all interior edges.

Since the gradient terms are linear in the parameters, β_i, determining the function at the vertices, the penalty may also be expressed as a piecewise linear function of these values,

$$\sum_k |c_k| = \sum_k |h_k^{\top}\hat{\beta}|,$$

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,

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

A family of conditional quantile triogram models can be estimated by simply replacing $|\cdot|$ by $\rho_\tau(\cdot)$, see Section 6.3.

4.4. **Penalized Triograms for Conditional Mean Models.** A corresponding penalized least squares problem may be formulated as,

$$\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} (z_i - a_i^{\top}\beta)^2 + \lambda \sum_{k=1}^{M} (h_k^{\top}\beta)^2.$$

Like the median regression problem this may be viewed as an augmented regression problem with response vector $(z^{\top}, 0^{\top}) \in \mathbb{R}^{n+M}$, and design matrix $[A^{\top} \vdots H^{\top}]^{\top}$ where $A = (a_i^{\top})$ and $H = (h_k^{\top})$.

4.5. **On Sparsity.** A crucial feature of the penalized triogram estimators described above is the sparsity of the augmented design matrices. In the fidelity component A, rows have at most three non-zero elements needed to represent the barycentric coordinates of the (x_i, y_i) points not included as basic vertices. For basic rows, the vector a_i is one in only one element and zero everywhere else. In the penalty matrix, H, each row has four nonzero entries and the remaining elements are zero.

To appreciate the consequences of this it may help to consider an example. Suppose we have $n = 1600$ observations and we introduce basic vertices (knots) at each of the points, $\{(x_i, y_i) : i = 1, ..., n\}$. The resulting matrix A is just the $n = 1600$ identity matrix. The number of interior edges of the Delaunay triangulation is given by

$$e = 3n - 2c - 3,$$

where c denotes the number of exterior edges, see Okabe, Boots, Sugihara, and Chiu (2000). So the matrix H is 4753 by 1600 in a typical example, and the augmented ℓ_1 regression problem is thus, 6353 by 1600. This may appear computationally intractable, and would be intractable on many machines using conventional statistical software, But recognizing the sparsity of the problem, that is noting that only 0.2 percent of the more than 10 million elements of the design matrix are nonzero, drastically reduces the memory requirement and computational complexity of the problem from about 80Mb to only 160Kb.

In our MATLAB implementation of the triogram software only the nonzero elements of the design matrix need to be stored, along with their identifying indices. This drastically reduces the memory requirements of the computations and improves efficiency. Much better performance would be possible if the computations were recoded a lower level language like C or fortran using one of several available sparse matrix libraries.

4.6. **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_λ denote the dimension of $\hat{g}_\lambda$ defined in this way, we may consider selecting λ to minimize,

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

It should be emphasized that that this is a purely *ad hoc* expedient at this stage and needs considerable further investigation.

5. 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.[1]

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 $\mathcal{V} = \{v_i \in \mathbb{R}^2 : i = 1, ...n\}$ in general position consists of all triangles whose circumscribing circle contains no $\mathcal{V}$-points in their interior. There is a vast literature on how to compute the Delaunay triangulation.

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 $\mathcal{V}$, 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 noted 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.[2]

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 $\hat{F}_h$ using the horizontal triangulation Δ_h with $\hat{F}_v$ using the vertical triangulation Δ_v. Rippa's performance measure is integrated squared error

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

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

[2] This is related to the following observation by Bern and Eppstein (1992). If you consider lifting the points in $\mathcal{V}$ 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 $\mathcal{V}$ from the point $(0, 0, -\infty)$. The projection of this lower portion of the hull onto the xy-plane is the Delaunay triangulation.

Rippa showed that the triangulation Δ_h is preferred to Δ_v in the sense of ISE iff

$$\frac{|H_{11}|}{|H_{22}|} < \left(\frac{\beta}{\alpha}\right)^2 . \tag{5.1}$$

One may ask whether this result is altered if one replaces integrated *squared* error by integrated *absolute* error.

$$IAE(\hat{F}_i, F, \mathcal{U}) = \int_{\mathcal{U}} |\hat{F}_i(x,y) - F(x,y)| dxdy$$

Proposition 5.1. *The triangulation Δ_h is preferred to Δ_v is the sense of IAE iff* (5.1) *holds.*

The proof of the second proposition appears in the appendix. The comparison of the two triangulations thus reduces to exactly the same expression as in the squared error case. The sensitivity of the approximation quality to the choice of the triangulation underlines the need for careful selection especially if a small number of vertices are employed. An advantage of penalty methods in this respect is that their reliance on a considerably larger set of vertices can compensate to some extent for deficiencies in the triangulation.

6. Examples

In this section we consider several examples to illustrate the performance of the proposed methods. We begin with a simple artificial data example.

6.1. Example 1.

In the first example, we consider estimating a noisy cone,

$$z_i = \max\{0, 1/3 - 1/2\sqrt{x_i^2 + y_i^2}\} + u_i.$$

The (x_i, y_i)'s are generated as independent uniforms on $[-1,1]^2$, and the u_i are iid Gaussian with standard deviation $\sigma = .02$. The sample size is taken to be $n = 400$, and the fits are based on the Delaunay triangulation of all n points. With $n = 400$ the number of Delaunay edges is roughly 1200, so the resulting ℓ_1 regression problems are roughly 1600 by 400. Even so, the fitting in MATLAB is quite quick, about 1.5 seconds per fit on a Sun Ultra 2.

In Figure 5.1 we illustrate four different triogram fits corresponding to various values of the smoothing parameter λ. In the first panel, the fit is essentially an interpolation of the observations and is evidently too rough. Above the panels we report the value of λ and the effective dimension of the fitted function as measured by the number of observations interpolated by the fitted function. In the first panel this number is nearly 400. In the second panel p_λ has been reduced to 36, and the fit is quite accurate. The flat area outside the region $A = \{x^2 + y^2 \leq 4/9\}$ is quite well represented and the cone is quite smooth. In the third panel p_λ has been reduced to 17, and the fitted function is already somewhat oversmoothed. In particular, the sharp edge on the boundary of the region A has been lost. In the last of the four

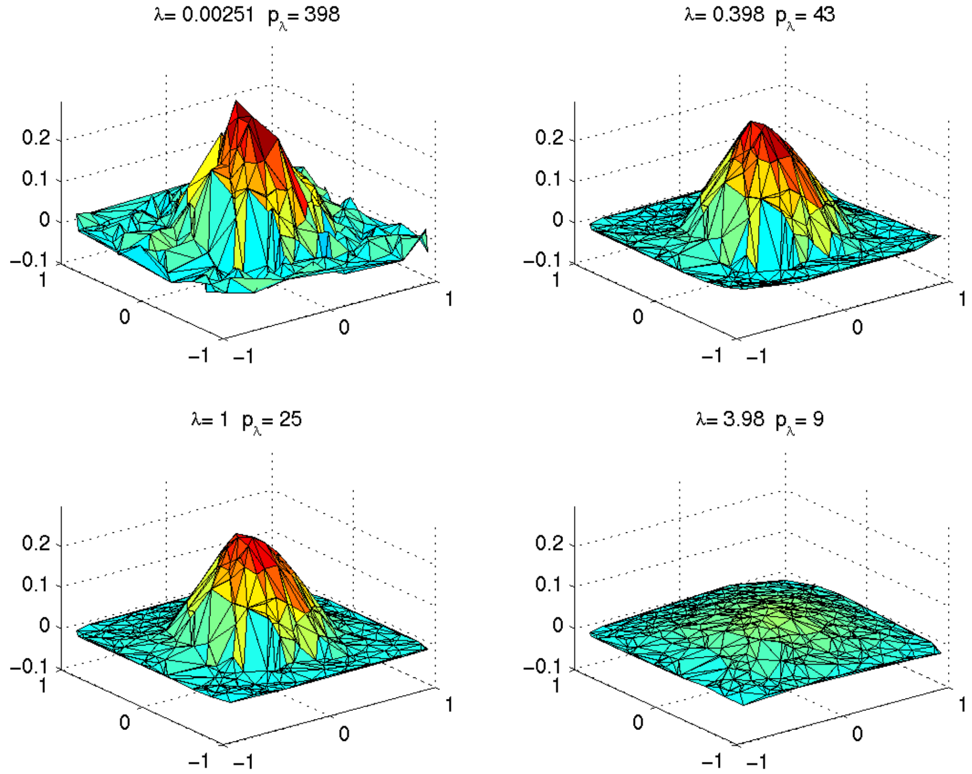

Figure 6.1. Four median triogram fits for the noisy cone example. The values of the smoothing parameter λ and the number of interpolated points in the fidelity component of the objective function, p_λ are indicated above each of the four plots.

panels, the structure has been lost entirely and we quite close to the limiting linear fit achieved as $\lambda \to \infty$. This last plot conveys some idea of the nature of the Delaunay triangulation of the domain of fitted functions.

In the next figure we illustrate four fits of the same data for the least squares version of the triogram estimator. In this case, following Hastie and Tibshirani (1990), we use the trace of the linear operator defining the least squares fit to measure the dimension of the fitted surface. This seems to yield a plausible estimate of the dimension, and is monotonically decreasing in λ, but further study is clearly warranted.

6.2. **Example 2.** In the second example we consider estimating the function

$$g_0(x, y) = \frac{40 \exp(8((x - .5)^2 + (y - .5)^2))}{(\exp(8((x - .2)^2 + (y - .7)^2)) + \exp(8((x - .7)^2 + (y - .2)^2)))}$$

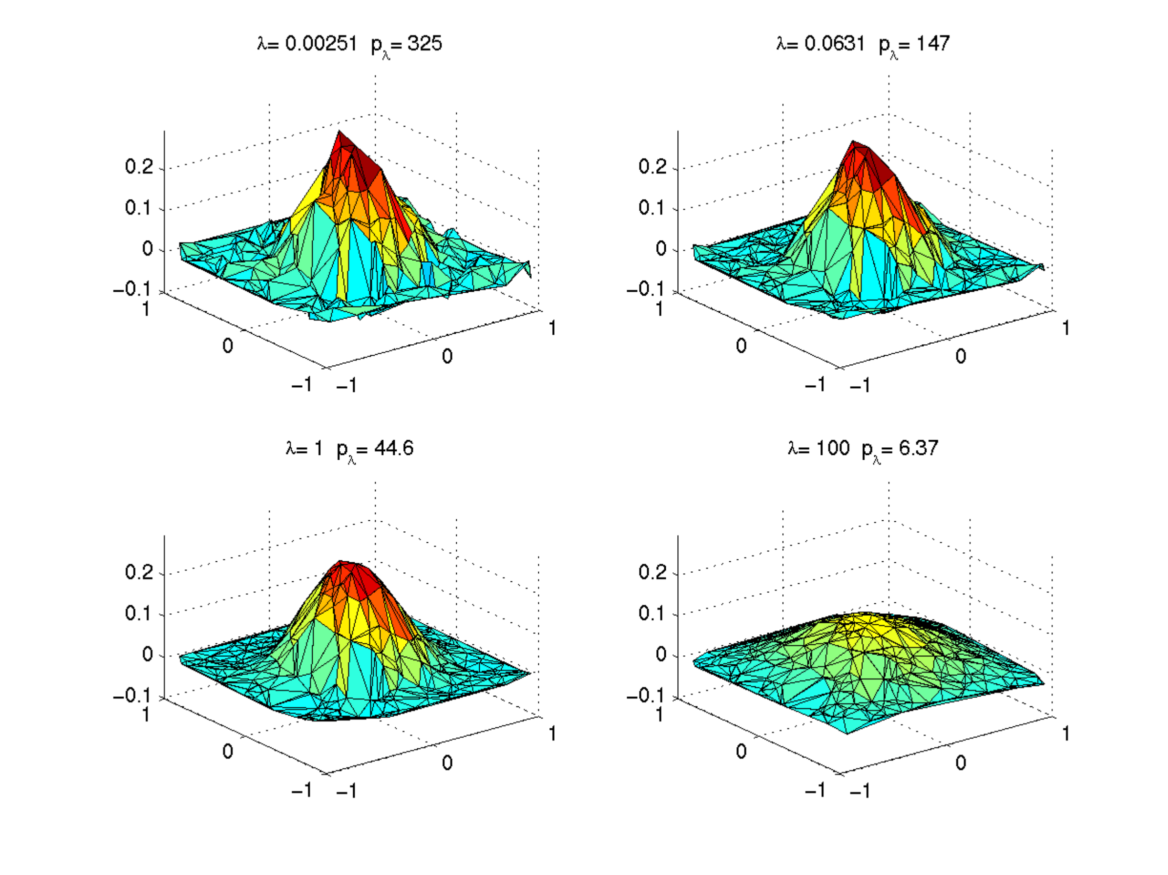

FIGURE 6.2. Four mean triogram fits for the noisy cone example. The values of the smoothing parameter λ and the trace of the linear operator defining the estimator, p_λ are indicated above each of the four plots.

The function has a ridge along the 45 degree line and therefore presents a challenge to tensor product methods. It has been previously considered by Gu, Bates, Chen, and Wahba (1989), Breiman (1991), Friedman (1991), He and Shi (1996), and Hansen, Kooperberg, and Sardy (1998), among others. Using the experimental design of He and Shi (1996) we compare their L_1 and L_2 tensor product regression spline estimators with the L_1 and L_2 versions of the penalized triogram. The observations $\{(x_i, y_i) : \; i = 1, ..., n\}$ are generated as independent uniforms on $[0,1]^2$ and we generate,

$$z_i = g_0(x_i, y_i) + u_i,$$

with iid $\{u_i\}$. Three distributions are considered: standard normal $\mathcal{N}(0,1)$; the normal mixture, $.95\mathcal{N}(0,1) + .05\mathcal{N}(0,25)$; and slash, $\mathcal{N}(0,1)/U[0,1]$. The sample

Distribution	L_1 tensor	L_1 triogram	L_2 tensor	L_2 triogram
Normal	0.609	0.442	0.544	0.3102
	(0.095)	(0.161)	(0.072)	(0.093)
Normal Mixture	0.691	0.515	0.747	0.602
	(0.233)	(0.245)	(0.327)	(0.187)
Slash	0.689	4.79	31.1	171.1
	(6.52)	(125.22)	(18135)	(4723)

Table 6.1. Comparative MISE for fitting the Gu, Bates, Chen and Wahba function.

size is $n = 100$. As a measure of performance we focus exclusively on,

$$\text{MISE} = \text{average}\{n^{-1} \sum (\hat{g}_n(x_i, y_i) - (g_0(x_i, y_i))^2\},$$

averaging over the $R = 1000$ replications.

In Table 5.1 we report He and Shi's results for their tensor product regression splines, and the corresponding results for the L_1 and L_2 penalized triogram approach. The selection of λ for the triogram fitting was made by minimizing $SIC(\lambda)$ over a grid of λ's from 0.1 to 1.0. The grid consisted of the points $\{\lambda = 10^{i/20} : \ i = -20, -19, ..., 0\}$. This procedure yielded a fit with median p_λ of 16.

The performance of the L_1 triogram estimator is quite good for the normal and normal mixture error distributions.[3] However, it appears that the L_1 triogram fails badly for the slash distribution. It is worth delving into this failure a bit further. The first observation to be made is that the failure is due entirely to two spectacular disasters out of the 1000 replications. If we drop the two worst replications, the slash entry in the table changes from 4.79 (125.22) to .486(3.25), and now appears quite competitive. What went wrong?

In each case the explanation lies in a single outlying z_i value that happened to occur on the convex hull of the observed (x_i, y_i) points. Since the boundary edges of the triangulation do not contribute to the penalty, the only consequence of over-zealous fitting of such points is the associated interior connecting edge effects. For sufficiently small values of λ this contribution is dominated by the gain in fidelity acheived by exact fitting of the outlying point. There seem to be two important lessons to be learned from this experience. First, one ignores the boundary effects of the fitting procedure *at one's peril*. And second, the SIC criterion is not to be trusted in cases in which it exhibits significant discontinuities.

If one had an informative boundary condition that could be explicitly built into the penalty function, this would help considerably. Or, if our λ selection had managed to

[3] He and Shi (1996) also report performance of MARS (Friedman (1991)) and PIMPLE (Breiman (1991)), which they find less satisfactory than their tensor product approach.

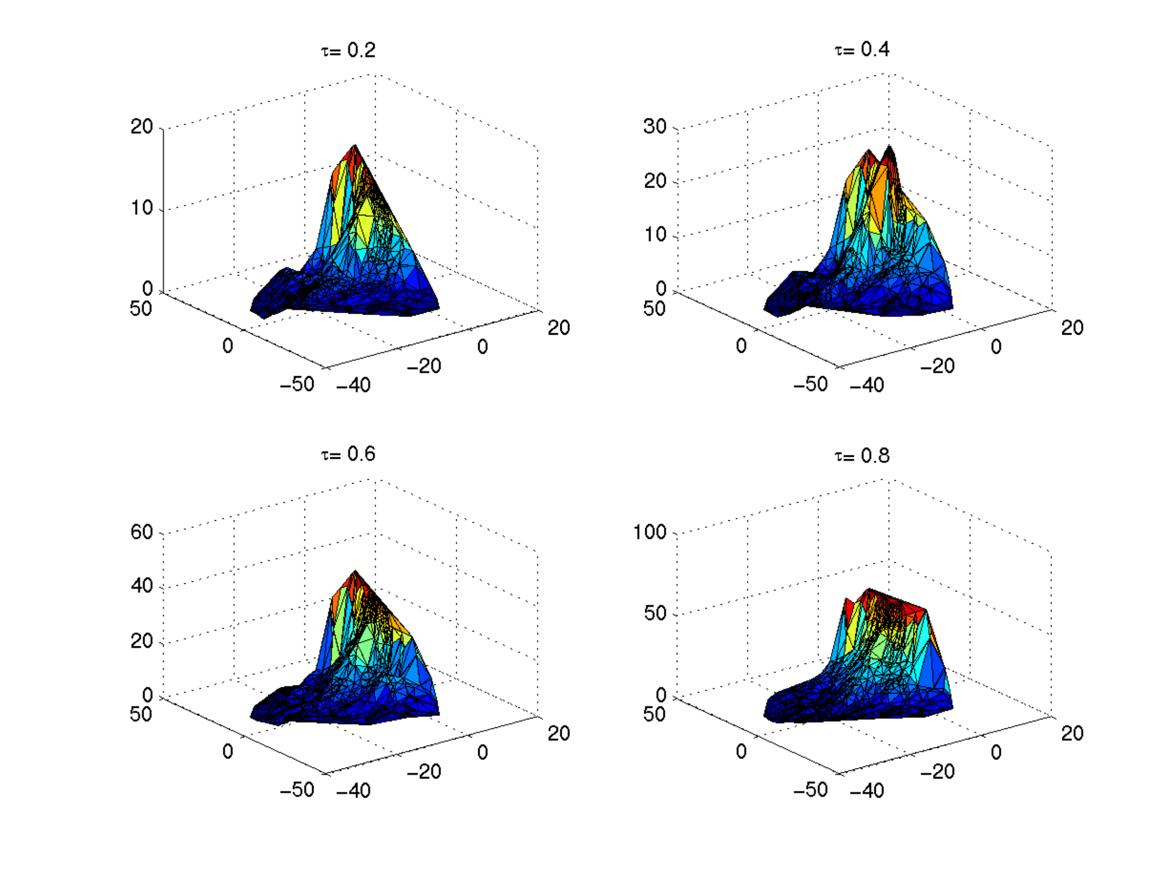

FIGURE 6.3. Four quintile triogram fits for Chicago land values.

select a more parsimonious model the situation would also have been salvaged. One could also place the blame on the slash specification, which is admittedly extreme, but this would be tantamount to a repudiation of the robustness objective.

6.3. **Chicago Land Values.** Our final example involves estimating a model for Chicago land values. The data consists of 761 land sales occuring during the period 1995-1997 in the Chicago metropolitan area. We take the sale price of the land in dollars per square foot as z_i and (x_i, y_i) pairs are measured in miles from the intersection of State and Madison. We illustrate four fits corresponding to the four quintiles of the land value distribution. In each case the smoothing parameter, λ, is chosen, rather arbitrarily, to be .5.

7. CONCLUSIONS

We believe that regularization, or shrinkage, methods offer a promising complementary approach to knot selection for triogram models. Roughness penalties based on total variation seem particularly well-suited to triogram applications. They satisfy natural equivariance requirements and are computationally very attractive. 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 pragmatic issues of algorithmic design. Further work is clearly necessary, but a strong *prima facie* case has been made for the attractive features of total variation penalties and their application to triogram estimation.

Appendix

Proof of Proposition 4.1 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 $(u,v) \in \mathbb{R}^2_+$. 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, $|\delta|$, 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 evaluating, we conclude that Δ_h is preferred to Δ_v, just as in the $\mathcal{L}_2$ case, iff

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

References

Ambrosio, L., N. Fusco, and D. Pallara (2000): *Functions of bounded variation and free discontinuity problems.* Clarendon Press, Oxford.

Banach, S. (1925): "Sur les lignes rectifiables et les surfaces dont l'aire est finie," *Fund. Math.*, 7, 225–236.

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.

Breiman, L. (1991): "The Π Method for Estimating Multivariate Functions From Noisy Data (Disc: P145-160)," *Technometrics*, 33, 125–143.

Candes, E. J. (2000): "Ridgelets: Estimating with Ridge Functions," preprint, Department of Statistics, Stanford University.

De Giorgi, E. (1954): "Su uns teoria generale della misura $(r-1)$-dimensionale in uno spazio a r dimensioni," *Ann. Math. Pura Appl. (4)*, 36, 191–213.

Dinculeanu, N. (1967): *Vector measures.* Pergamon Press, Oxford, New York.

Donoho, D., S. Chen, and M. Saunders (1998): "Atomic decomposition by basis pursuit," *SIAM J. of Scientific Computing*, 20, 33–61.

Duchon, J. (1976): "Interpolation des fonctions de deuix variables suivant le principe de la flexion des plaques minces," *R.A.I.R.O., Analyse numérique*, 10, 1–13.

——— (1977): "Splines minimizing rotation-invariant semi-norms in Sobolev spaces," in *Constructive Theory of Functions of Several Variables, Oberwolfach 1976*, Lecture Notes in Mathematics 571, pp. 85–100. Springer, Berlin.

Fichera, G. (1954): *Lezioni sulle transformazioni lineari.* Istituto matematico dell'Università di Trieste, vol I.

Friedman, J. H. (1991): "Multivariate Adaptive Regression Splines (Disc: P67-141)," *The Annals of Statistics*, 19, 1–67.

Green, P. J., and B. W. Silverman (1994): *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach.* Chapman-Hall.

Gu, C., D. M. Bates, Z. Chen, and G. Wahba (1989): "The Computation of Generalized Cross-validation Functions Through Householder Tridiagonalization With Applications to the Fitting of Interaction Spline Models," *SIAM Journal on Matrix Analysis and Application*, 10, 457–480.

Hansen, M., C. Kooperberg, and S. Sardy (1998): "Triogram Models," *J. of Am. Stat. Assoc.*, 93, 101–119.

Harder, R. L., and R. N. Desmarais (1972): "Interpolation using surface splines," *J. Aircraft*, 9, 189–191.

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

He, X., P. Ng, and S. Portnoy (1998): "Bivariate quantile smoothing splines," *J. Royal Stat. Soc. (B)*, 60, 537–550.

He, X., and P. Shi (1996): "Bivariate Tensor-product B-splines in a Partly Linear Model," *Journal of Multivariate Analysis*, 58, 162–181.

Ivanov, L. D. (1975): *Variations of sets and functions.* Nauka, Moskva, [in Russian].

Jordan, C. (1881): "Sur la série de Fourier," *C. R. Acad. Sci. Paris*, XCII, 228–230.

Koenker, R., P. Ng, and S. Portnoy (1994): "Quantile Smoothing Splines," *Biometrika*, 81, 673–680.

Kronrod, A. S. (1949): "On linear and planar variation of functions of several variables," *Doklady Akademii Nauk SSSR*, 66, 797–800, [in Russian].

——— (1950): "On functions of two variables," *Uspekhi matematicheskikh nauk*, 5, 24–134, [in Russian].

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

MIZERA, I. (2001): "Plastic Splines," preprint.

NATANSON, I. (1974): *Theory of Functions of a Real Variable*. Ungar.

OKABE, A., B. BOOTS, K. SUGIHARA, AND S. CHIU (2000): *Spatial Tessilations*. Wiley.

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.

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

SCHERZER, O. (1998): "Denoising with higher order derivatives of bounded variation and application to parameter estimation," *Computing*, 60, 1–27.

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

SERRIN, J. (1961): "On the definition and properties of certain variational integrals," *Trans. Amer. Math. Soc.*, 101, 139–167.

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

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

TIBSHIRANI, R. (1996): "Regression Shrinkage and Selection Via the Lasso," *J. Royal Stat. Soc. (B)*, 58, 267–288.

TONELLI, L. (1926): "Sulla quadratura delle superficie I, II, III," *Rend. Acc. Naz. Lincei*, 3, 357–362, 445–450, 633–638.

——— (1936): "Sulle funzioni di due variabili generalmente a variazione limitata," *Ann. Scuola Norm. Sup. Pisa Cl. Sci. (2)*, 5, 315–320.

VITUSHKIN, A. G. (1955): *On multidimensional variations*. GITTL, Moskva, [in Russian].

VON NEUMANN, J. (1937): "Some matrix inequalities and metrization of matrix space," *Izv. Nauchno-issledovatelskogo Inst. Mat. i Mekh. Tomsk. Gos. Univ.*, 1.

WAHBA, G., AND J. WENDELBERGER (1980): "Some new mathematical methods for variational objective analysis using splines and cross validation," *Monthly Weather Review*, 108, 1122–1143.

UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

UNIVERSITY OF ALBERTA, EDMONTON