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 roughness penalty adapted to the piecewise linear structure of the triogram model. We show that the proposed roughness penalty may be interpreted as a total variation penalty on the gradient of the fitted function. The methods are illustrated with both real and artificial examples, including an application to estimated quantile surfaces of land value in the Chicago metropolitan area.

"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 an 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 tractability and statistical efficiency play important roles in this debate, and the resulting rivalry has significantly broadened the scope of both approaches.

Key words and phrases. Regularization, penalty methods, total variation, spline smoothing.

Version: December 6, 2002. Corresponding author: Roger Koenker, Department of Economics, University of Illinois, Champaign, IL, 61820, USA; (from January 7 to June 30, 2003, Department of Economics, University College London, Drayton House, 30 Gordon Street, London, WC1H 0AX), Email: roger@ysidro.econ.uiuc.edu.

Hansen, Kooperberg, and Sardy (1998) have recently 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 the regression spline methods of Stone, Hansen, Kooperberg, and Troung (1997) are proposed to control the degree of smoothness of the estimates. The primary objective of the present paper is to explore a smoothing spline approach to the estimation of triograms. The roughness penalty we employ may be interpreted as an extension to bivariate settings of the total variation roughness penalty suggested in Koenker, Ng, and Portnoy (1994) for univariate smoothing problems.

2. On Roughness Penalties

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 $\hat{g}$ 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 interpreted as a prior belief that the L_2 norm of g'' is unlikely to exceed a specified bound controlled by the choice of λ.

2.1. Total Variation Roughness in the One-dimensional Case

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 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 the total variation of the first derivative of g. If g' is absolutely continuous, we can also write, Natanson (1974, Theorem IX.4.8),

$$J(g) = V(g') = \int |g''(x)| dx. \tag{2.2}$$

This establishes a clear link of the total variation roughness penalty to the classical L_2 penalty. For smooth functions, the total variation penalty is simply the L_1 analogue of the classical L_2 smoothing spline penalty. Mammen and van de Geer (1997) have studied total variation penalties with L_2 fidelity in the univariate setting;

together with Portnoy (1997), they have explored the asymptotic behavior of related estimators.

From a theoretical perspective, the use of total variation penalties enables the fit to mimic sharp bends, spikes and edges more easily than the conventional L_2 penalties. For this reason, total variation penalties have received considerable attention in the image processing literature dating back to Rudin, Osher, and Fatemi (1992), see also Rudin and Osher (1994). In contrast to the penalty (2.2), which penalizes variation in the derivative of $\hat{g}$, the imaging literature has focused primarily on variation of the function itself. See Tasdizen, Whitaker, Burchard, and Osher (2002) and Scherzer (1998) for exceptions. Total variation penalties applied on the function itself have been recently considered also in the statistical literature, by Nicholls (1998) and Davies and Kovac (2001). This approach permits discontinuities in the fitted functions, and thus helps to illuminate statistical aspects of related work on total variation penalties for bivariate functions used for edge detection and image segmentation.

From a more pragmatic viewpoint, total variation penalties are well matched for the quantile regression fidelity since they preserve the linear programming formulation of the optimization problem defining the estimator. Solutions to the problem (2.1) take the form of piecewise linear functions with jumps in their derivative at a few of 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. See Section 4.7 below and Tibshirani (1996) and Donoho, Chen, and Saunders (1998) for related discussion of the model-selection, shrinkage effects of L_1 type penalties.

2.2. **Thin Plate Penalties for Bivariate Smoothing**

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), Wahba and Wendelberger (1980), and others, minimize

$$\sum_{i=1}^{n}(z_i - g(x_i, y_i))^2 + \lambda J(g, \Omega, \|\cdot\|_2^2), \tag{2.3}$$

with the roughness penalty defined as

$$J(g, \Omega, \|\cdot\|_2^2) = \iint_\Omega \| \nabla^2 g\|_2^2 dxdy = \iint_\Omega (g_{xx}^2 + 2g_{xy}^2 + g_{yy}^2)dxdy. \tag{2.4}$$

A considerable computational simplification is achieved in the classical form of thin-plate setting, when Ω is taken to be all of $\mathbb{R}^2$; 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.

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

$$J(g, \Omega, ||\cdot||_2^2) = J(h, \Omega_1, ||\cdot||_2^2)\ \mu(\Omega_2), \tag{2.5}$$

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 (2.3) 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 (2.2), define a bivariate roughness penalty? How should we define total variation of the gradient of a function of two variables?

2.3. **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 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 attendant reliance on rectangular domains—a drawback in nonparametric regression, as we argue below. The first orthogonally-invariant definitions were introduced by Kronrod (1949, 1950) in the spirit of the Banach indicatrix theorem. However, eventually the approach based on the Vitali formula interpreted in the language of Schwartzian distribution prevailed; 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 in the theory of Sobolev spaces, the formalism of distributions is needed only for differentiation and limit transitions; the functions under consideration remain standard. A convenient initial step is 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 distributions, are measures. The defined functions encompass not only smooth functions with bounded variation, but also, for instance, piecewise linear or piecewise constant (that is, discontinuous) functions (with bounded variation). For a smooth function f from $\mathbb{R}^k$ to $\mathbb{R}^m$, we define

$$V(f, \Omega, \|\cdot\|) = \int_\Omega \|\nabla f\|\, dx; \tag{2.6}$$

here dx denotes (multiple) integration with respect to k-dimensional Lebesgue measure. The functional V, initially defined for smooth functions, is lower semicontinuous,

hence it can be extended to a broader domain via the approach of Serrin (1961):

$$V(f) = \liminf V(f^\nu), \tag{2.7}$$

where the right-hand side expression denotes the inf of $\liminf V(f^\nu)$ over all sequences f^ν approaching f in the sense of distributions. It can be shown that the extended V assigns a finite value to every function with bounded variation; that is, the extension is nontrivial.

For a function g from $\mathbb{R}^2$ to $\mathbb{R}$ we thus define a penalty

$$J(g, \Omega, \|\cdot\|) = V(\nabla g, \Omega, \|\cdot\|) = \iint_\Omega \|\nabla^2 g\| \, dx\, dy. \tag{2.8}$$

This penalty assigns a finite value to every function whose gradient has bounded variation—in particular, to every piecewise-linear continuous function on bounded domains with finite number of linear pieces. This is in contrast with the behavior of the thin-plate functional (2.4). Since the latter is also lower semicontinuous, one may contemplate an extension analogous to (2.7) also for this functional. However, as shown by Serrin (1961), any such an extension assigns $+\infty$ to any f with discontinuous derivatives; in particular, any function with a spike or a sharp ridge is evaluated as infinitely rough.

Any penalty of the form (2.8) can be considered an extension of the univariate penalty (2.2), regardless of the choice of the norm.

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 (2.8), but not on g, such that for any $\Omega = \Omega_1 \times \Omega_2$,*

$$J(g, \Omega, \|\cdot\|) = c\, J(h, \Omega_1, \|\cdot\|)\ |\Omega_2|, \tag{2.9}$$

where $J(h, \Omega_1, \|\cdot\|) = \int_{\Omega_1} |h''(x)|\, dx$, 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 (2.9) follows by the Fubini theorem for all smooth g and hence by extension for all g under consideration. ∎

2.4. **The Choice of the Norm: Invariance and Equivariance**

For denoising images with a view toward reconstructing discontinuities in derivatives, Scherzer (1998) proposed using the penalty corresponding to the ℓ_1 norm in (2.6); for

smooth functions g from $\mathbb{R}^2$ to $\mathbb{R}$ this penalty is equal to

$$J(g, \Omega, \|\cdot\|_1) = \iint (|g_{xx}| + 2|g_{xy}| + |g_{yy}|)dxdy. \tag{2.10}$$

A related penalty has 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 of (2.6), with the ℓ_1 norm applied to the diagonal of the Hessian,

$$J(g, \Omega, \|\cdot\|_{HNP}) = \iint (|g_{xx}| + |g_{yy}|)dxdy. \tag{2.11}$$

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 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, as stressed by Green and Silverman (1994), provide valuable guidance through the forest of potential definitions of roughness penalty functionals.

The requirement of orthogonal invariance for the penalty J leads to the condition

$$\|U^T H U\| = \|H\|, \tag{2.12}$$

satisfied by 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 (2.12). 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

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

Other possibilities include the spectral norm, the maximal absolute value of the eigenvalues, or absolute value of the trace.

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

$$\sum_{i=1}^{n} |z_i - g(x_i, y_i)| + \lambda J(g, \Omega, \|\cdot\|) \tag{2.13}$$

then cg minimizes (2.13) 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, but for Gaussian fidelity the thin plate penalty *is* well matched in this sense.

3. Triograms

Efficient numerical solution of the variational problems arising from 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).

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.

3.1. A Roughness Penalty for Triograms

As already pointed out, thin-plate penalties are inappropriate for triograms, since their penalties assign infinity to any function with a discontinuity in the gradient, and thus the thin plate penalty is inherently incapable of discriminating among triograms. Roughness penalties based on the total variation of the gradient are more promising. Fortuitously, it also turns out that the troublesome choice of a norm disappears; once we insist on a coordinate-independent penalty for triograms, all penalties coalesce.

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* (2.8)*, there is a constant c dependent only on the choice of the norm such that,*

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

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 Ω 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 an edge e is the corresponding term in (3.1): consider the quadrilateral 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 only 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 (2.7). 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. A common example of such a $\phi(x, y)$ is a multiple of $\exp(-1/(1 - x^2 - y^2))$ on the unit circle and 0 elsewhere. When $\nu \to \infty$, g^ν approaches g in the distributional sense. The contribution of the vertex is bounded from above by

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

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 would be $C\nu^2$, and the contribution of the vertex would not vanish, though it would be finite. ∎

The crucial consequence of Theorem 3.1 is that, from the variational point of view, there is only one viable form of the total variation penalty penalizing the gradient of triograms: that given by formula (3.1). Penalized quantile triograms, functions minimizing

$$\sum_{i=1}^{n} \rho_\tau(z_i - g(x_i, y_i)) + \lambda J(g, \Omega, \| \cdot \|) \tag{3.3}$$

over the space of triograms $\mathcal{G}$ can be then found as a solution of a linear programming problem as we show in the next section.

4. **Computation of Penalized Triograms**

Modern linear programming methods provide an extremely efficient means of solving for penalized triogram estimators. We will briefly describe how to express elements

of the linear space $\mathcal{G}$ in terms of a finite set of basis functions, and then show how the fidelity and penalty terms can be written as piecewise linear functions of the coefficients of this expansion. This leads to a linear programming formulation of (3.3) for which interior point, log-barrier methods exploiting the sparsity of the linear algebra offer an efficient solution strategy.

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 $0 \le B_i(u) \le 1$ with

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

4.2. **Computing the Fidelity**

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:

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

Thus, we can express the fidelity of the function, $\hat{g}(x,y)$, fitted 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. However, one may wish to choose $p < n$ and there would be a need to compute some nontrivial 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. We have

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

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,

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

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, i.e.,

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

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 an augmented linear ℓ_1 problem minimizing,

$$\sum_{i=1}^n |z_i - a_i^T \beta| + \lambda \sum_{k=1}^M |h_k^T \beta|. \tag{4.1}$$

This approach to estimating penalized median surfaces may be immediately extended to estimation of penalized quantile surfaces by minimizing,

$$\sum_{i=1}^n \rho_\tau(z_i - a_i^T \beta) + \lambda \sum_{k=1}^M |h_k^T \beta|. \tag{4.2}$$

4.4. **Penalized Triograms for Conditional Mean Models**

A corresponding penalized least squares problem may be formulated to minimize

$$\sum_{i=1}^n (z_i - a_i^T \beta)^2 + \lambda \sum_{k=1}^M (h_k^T \beta)^2. \tag{4.3}$$

Like the median regression problem this may be viewed as an augmented regression problem with response vector $(z^T, 0^T) \in \mathbb{R}^{n+M}$, and design matrix $[A^T : H^T]^T$ where $A = (a_i^T)$ and $H = (h_k^T)$. This L_2 variant of penalized triogram has been explored further by Hansen and Kooperberg (2002). It may be noted that the primal-dual interior point methods used to mininize (4.1) begin their iterations by solving (4.3), and continue to take iteratively reweighted least squares steps until a solution is reached with at least p zero residuals.

4.5. **Penalized Triograms as Linear Programs**

The problem (4.1) is piecewise linear in β, so it is straightforward to reformulate it as a linear program. Let

$$X = \begin{bmatrix} A \\ \lambda H \end{bmatrix},$$

where A denotes the matrix with rows $(a_i^T)_{i=1}^n$ and $H = (h_i^T)_{i=1}^M$, and denote the augmented response vector as $\zeta = (z^T, 0^T)^T$. The problem (4.1) of minimizing the ℓ_1 norm of the vector $\zeta - X\beta$ may be expressed as

$$\min_{(\beta,u,v)} \{1^T u + 1^T v | \zeta = X\beta + u - v,\ \beta \in \mathbb{R}^p, u \geq 0, v \geq 0\}.$$

So we are minimizing a linear function subject to linear equality and inequality constraints. The simplex algorithm provides an efficient solution method for moderate size problems of this type, but recent development of interior point methods provide an effective strategy even for extremely large problems. In effect, the interior point approach replaces the inequality constraints by a logarithmic barrier penalty function, and rather than traversing outside edges of the constraint set, it takes Newton type steps from the interior of the constraint set toward the boundary solution. See Portnoy and Koenker (1997) for a detailed discussion.

A salient feature of the polyhedral structure of the underlying optimization is that solutions can be characterized by a set of "basic variables". In the present case this basic set, h, consists of p elements of the first $n + M$ integers. Solutions may be written as,

$$\hat{\beta}(h) = X(h)^{-1}\zeta(h),$$

where $X(h)$ denotes the p by p submatrix of X with rows indexed by h, and $\zeta(h)$ denotes the p corresponding elements of ζ. (When multiple solutions exist they constitute a convex polyhedral set with solutions of the form $\hat{\beta}(h)$ as extreme points.) Some of the rows of $X(h)$ will be drawn from the upper A part of the X matrix, and some will come from the lower, λH part. If we now evaluate the fit at the n observed points we have

$$\hat{\gamma} = (\hat{g}(x_i))_{i=1}^n = AX(h)^{-1}\zeta(h).$$

For any element $i \in h$, the product $x_i^T X(h)^{-1}$ is a unit basis vector e_j such that $x_i^T X(h)^{-1}\zeta(h) = e_j^T\zeta(h) = \zeta_i$. When i comes the first n integers, so x_i is a row of A, this implies that the ith data point is interpolated, since for these points $\zeta_i = z_i$. When i comes from the set $\{n + 1, ..., n + M\}$ the ith fitted value is determined by the contribution of the penalty part of the X matrix, in conjunction with the interpolated points. Thus, only a subset of the n observations, say p_λ of them, are needed to describe the fit. Of course *all* of the observations are needed to determine *which* observations are interpolated, but once the fit is found, it can be described completely if one knows the p_λ points that are interpolated, the triangulation, and the value of λ. We will interpret p_λ as an effective dimension of the fitted model. As λ increases, p_λ decreases; near $\lambda = 0$ all of the observed points are interpolated, and for sufficiently large λ only the 3 points necessary to determine the best fitting "median plane" need to be interpolated.

4.6. **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 vertices of the triangulation. For observations whose (x_i, y_i) points are included as vertices of the triangulation, 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 corresponding to the contribution of the four vertices of the quadrilateral corresponding to each edge. The remaining elements of each h_k row are zero. This sparsity of the design matrix contrasts sharply with the situation for thin plate splines where the corresponding matrices for the penalty component are dense.

To appreciate the consequences of this it may help to consider an example. Suppose we have $n = 1600$ observations and we introduce vertices 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 at first sight appear computationally intractable, and would be intractable on most 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, reduces the memory requirement and computational complexity of the problem from about 80Mb to only 160Kb.

In our Matlab and R implementations only the nonzero elements of the design matrix are stored, along with their identifying indices. This drastically reduces the memory requirements of the computations and improves efficiency. As we note more explicitly below, fitting our penalized quantile triograms requires only a few seconds for our smaller examples, and up to about a minute on the larger ones in our Matlab implementation. Much better performance is attained in R where the computations are recoded in fortran using widely available sparse matrix libraries. See Koenker and Ng (2002) for details.

4.7. **Automatic λ Selection**

Despite numerous reservations expressed in the literature about over reliance on automatic methods for selecting smoothing parameters, regularization methods not confronting the problem of selecting λ would be considered incomplete. A full discussion of this deep and delicate subject is beyond the our present scope; instead, we mention several approaches that we have considered in the course of our investigations.

Rudin, Osher, and Fatemi (1992) propose a discrepancy method, choosing λ to match a preselected fidelity value. We employ this method in our first example in the next section where we choose λ to match the fidelity achieved by other smoothing method. See Wahba (1990) for further comments on this approach.

One can also base a λ selection criterion on a reasonable measure for the effective dimension of a fit $\hat{g}_\lambda$ for given λ. As with other ℓ_1 type estimation methods, such a measure is provided by simply counting the number p_λ of interpolated observations in the fidelity component of objective function. The recent work of Meyer and Woodroofe (2000) gives an additional support for this choice—they consider the

divergence,

$$\mathrm{div}(\hat{g}) = \sum \frac{\partial}{\partial y_i} \hat{g}(x_i)$$

as a general measure of effective dimension for nonparametric regression. For linear estimators this measure yields the trace of the corresponding linear operator, which has been suggested for classical smoothing splines by several other authors. The complementary quantity, n minus the effective dimension, is sometimes referred to as "equivalent degrees of freedom." For monotone regression estimation, Meyer and Woodroofe (2000) show that the pool adjacent violators algorithm yields a piecewise constant estimate, $\hat{g}$ with $\mathrm{div}(\hat{g})$ equal to the number of distinct values taken by $\hat{g}$. For our total variation penalized quantile triograms $\mathrm{div}(\hat{g}) = p_\lambda$ the number of points interpolated by $\hat{g}$, a fact that follows from the concluding comments of Section 4.5.

Once the effective dimension, p_λ, of the fit is defined, it can be plugged into any of the well-known model selection criterion. Koenker, Ng, and Portnoy (1994) suggest using the Schwarz (1978) criterion in their one-dimensional context. In our setting SIC selects λ minimizing

$$SIC(\lambda) = \log(n^{-1} \sum \rho_\tau(z_i - \hat{g}_\lambda(x_i, y_i))) + .5 n^{-1} p_\lambda \log n.$$

Analogously, we may define a version of the Akaike criterion and there are certainly other possible uses for p_λ in this vein.

Rather than simply counting the number of observations in the fidelity component with absolute residuals smaller than a specified tolerance, p_λ can be obtained by computing the trace of the hat matrix in the final weighted least squares step of the interior point algorithm. This approach has the added advantage that it suggests a linearized form of the generalized cross-validation criterion that can be used for λ selection. Finally, classical cross validation is also entirely feasible in small to moderate size problems, and further work on computational shortcuts may be able to extend the range of applicability of these methods.

4.8. **On Triangulations**

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 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.

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.

Another way to characterize the Delaunay triangulation is that it maximizes, the minimum angle occurring in the triangulation. This maxmin property was long considered a major virtue of the Delaunay method for reasons of numerical stability. Relatively recently, however, it has been noted by Rippa (1992) that the benefits of this prejudice against long, thin triangles depend upon 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.

The sensitivity of the approximation quality to the choice of the triangulations suggests 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 degree for deficiencies in the triangulation. One may restrict attention to Delaunay triangulations based on the observed (x, y) points, but it is straightforward to incorporate "dummy vertices" at other points in the plane, vertices that contribute to the penalty term, but not to the fidelity. By so doing one can ameliorate the effect of the initial triangulation and refine the fit to achieve more flexibility. This approach is illustrated in the treatment of the examples.

4.9. **Boundary and Qualitative Constraints**

A triogram is convex if and only if it is convex on all pairs of adjacent triangles. This condition is easily checked for each quadrilateral since it reduces to checking a linear inequality on the values taken by the function at the four vertices of the quadrilateral. Imposing convexity on penalized triogram fitting thus amounts to adding m linear inequality constraints to the problems already introduced, where m denotes the number of interior edges of the triangulation. This is particularly straightforward in the case of the quantile fidelity given the linear programming formulation of the optimization problem. Similarly, it is straight forward to impose constraints on the boundary of the fitted surface if prior information about how to treat these edges is available.

5. **Examples**

5.1. **Cobar mining data**

The first example employs data consisting of 38 measurements on the "true width" of an ore-bearing rock layer from a mine in Cobar, Australia. This data has been analyzed with two types of thin-plate penalty methods in Green and Silverman (1994); hence we may refer interested readers to their analyses and figures for comparisons. In accord with Green and Silverman, we do not employ an automatic method for λ selection, but rather provide two exploratory fits for different λ, to illustrate the

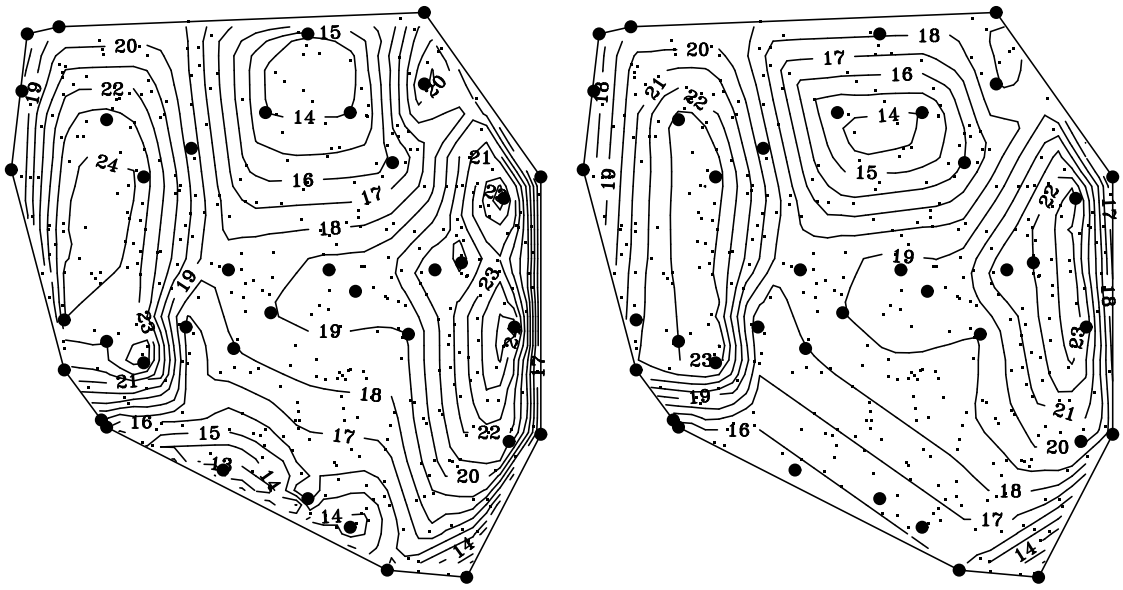


FIGURE 5.1. Median Contour Plots of the Cobar Ore Data

virtues of the smoothing method. We selected two λ's yielding the same (ℓ_2) fidelity for our fits as that achieved by Green and Silverman for their two reported fits, an interpolatory and a smoothing one. The contour plots of the two fits are illustrated in Figure 5.1. The plot on the left interpolates all 38 of the observed points and corresponds to $\lambda = .001$, while the right plot interpolates 23 of the observed points and corresponds to $\lambda = .088$. The observed xy points are indicated by the solid circles, dummy vertices by the points.

We restrict the domain of our analyses to the convex hull of the observed data; Green and Silverman used a slightly larger domain. To increase the flexibility of the fitted surface we have added 388 dummy vertices (xy-points) that have no influence on the fidelity term, but are treated just as the observed xy points in the basis expansion and penalty term. The dummy vertices are generated randomly from a uniform distribution on the specified domain.

5.2. **A Monte-Carlo Experiment**

In our second example we consider estimating the function

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

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

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

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 (x_i, y_i)'s are generated as independent uniforms on $[0,1]^2$, and we generate

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

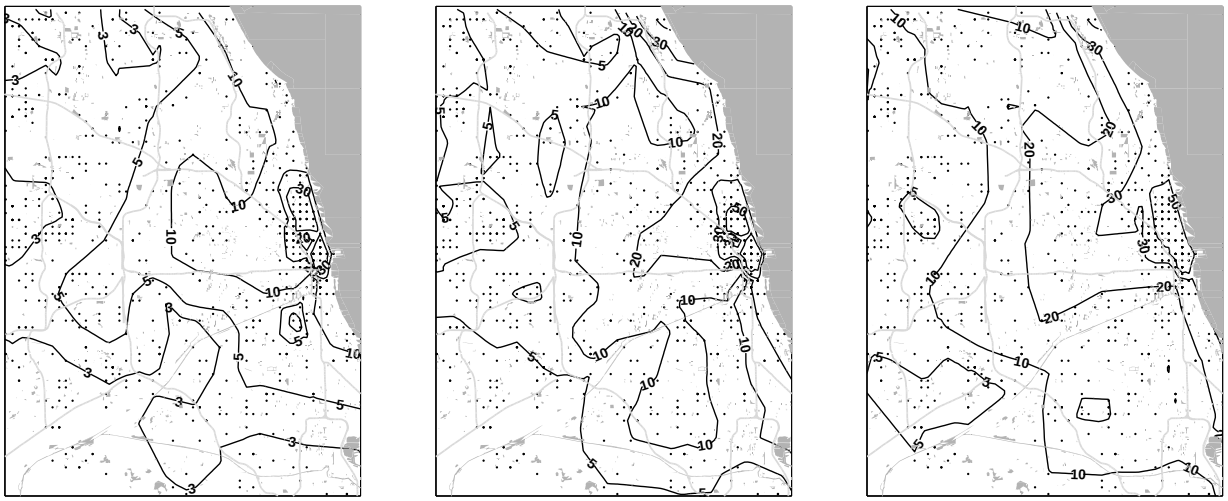
with iid u_i. Three distributions for the u_i are considered: standard normal $\mathcal{N}(0,1)$; the normal mixture, $.95\mathcal{N}(0,1) + .05\mathcal{N}(0,25)$; and slash, $\mathcal{N}(0,1)/U[0,1]$. The sample size is $n = 100$. As a measure of performance we focus exclusively on

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

averaging over the $R = 1000$ replications.

In Table 6.1 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 $\lambda = 10^{i/20}$ with $i = -20, -19, ..., 0$. 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. 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. 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 achieved by exact fitting of the outlying point.



First quartile Median Third quartile

FIGURE 5.2. Contours of Quartile Surfaces of Chicago Land Values.

5.3. **Chicago Land Values**

Our final example involves estimating a model for Chicago land values. The data consist of 1194 vacant land sales occurring at 758 distinct sites in the Chicago metropolitan area during the period 1995-1997. We take the sale price of the land in dollars per square foot as z_i and (x_i, y_i). In Figures 5.2 we illustrate three contour plots corresponding to fitted surfaces estimated by solving the problem (4.2) for the three quartiles $\tau \in \{.25, .50, .75\}$ of the land value distribution. In these plots Lake Michigan appears in the upper right corner and the Interstate highways are indicated in grey to provide landmarks for the metropolitan area. The central business district appears in each of the plots as a closed contour of high land value near the lake. The scattered points in the plots indicate the locations of the observed land sales used to fit the land value surfaces. In each case the smoothing parameter λ is chosen, somewhat arbitrarily, to be .25. This value is intermediate between the value selected by the SIC and AIC criteria mentioned in Section 4.7 In our judgment SIC yields a somewhat oversmoothed fit, with $p_\lambda = 57$, while AIC yields a a somewhat undersmoothed fit, with $p_\lambda = 220$, for the median model. Contours are labeled in dollars per square foot. We would like to emphasize that it may be advantageous in many smoothing problems to estimate a family of conditional quantile curves, or surfaces, since the commonly assumed iid error model is rarely really plausible. For land values, or annual temperatures, or snowfalls, it seems useful to have some estimate of the way variability varies over the domain.

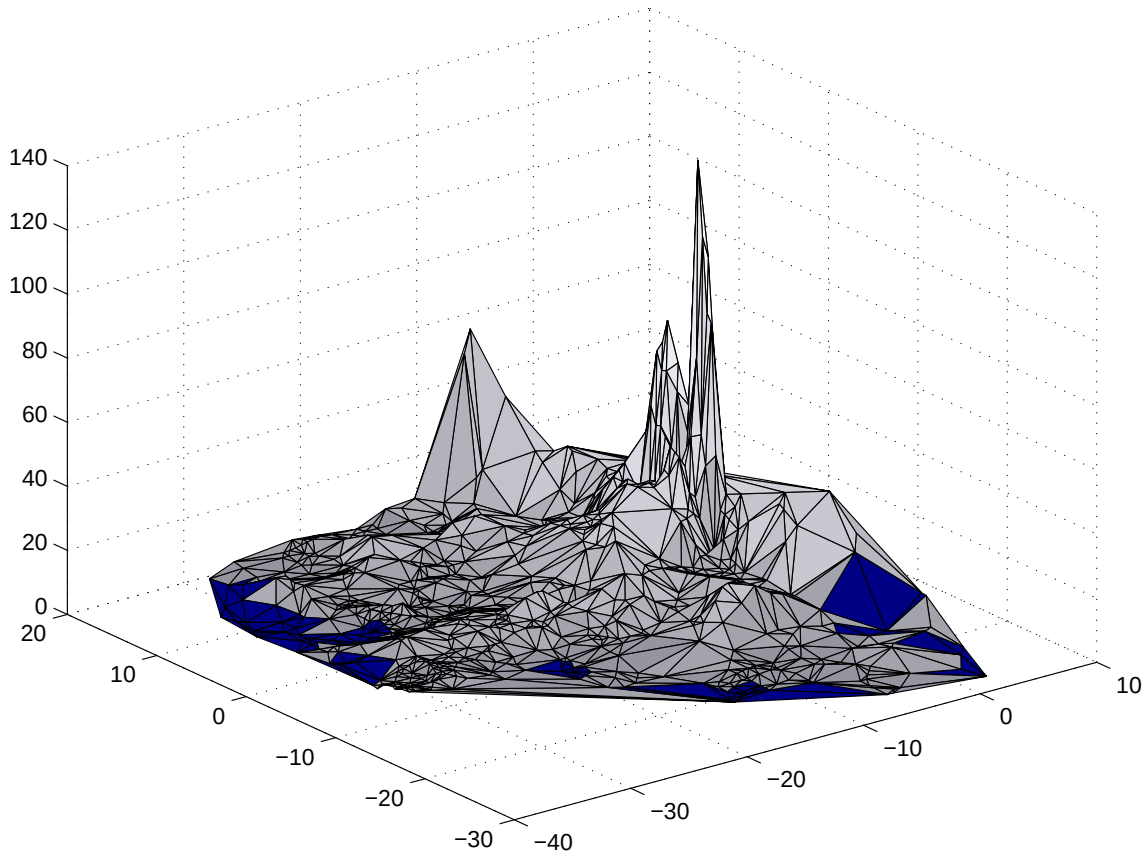


Figure 5.3. Perspective Plot of Median Chicago Land Value.

We illustrate a perspective plot of a median fit of the Chicago land value distribution in Figure 5.3. This fit again corresponds to the $\lambda = .25$. It is possible to recognize the peak corresponding to the central business district, and another mode further north along the lake, The perspective plot is somewhat difficult to interpret without further geographical reference points like those of Figure 5.2, but it does illustrate the ability of the penalized triogram to capture the sharp peaks of the land value distribution.

Among several possible refinements of this simple model for land values, we note that it is quite straightforward to add other covariates like the parcel size in a partially linear model formulation. See Koenker and Mizera (2002) for further details on this approach.

6. **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. 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.

7. Acknowlegments

A preliminary version of this paper was presented at a conference in honor of George Judge held at the University of Illinois, May 1-2, 1999. The research was partially supported by the NSF grant SES-99-11184, and by the Natural Sciences and Engineering Research Council of Canada. The authors would like express their appreciation to Jana Jurečková for the hospitality afforded them by the Department of Probability and Statistics of Charles University, to Steve Portnoy, Xuming He, and Hee-Seok Oh for many helpful discussions, and to Peter Colwell and Henry Munneke for providing the Chicago land value data.

References

AMBROSIO, L., N. FUSCO, AND D. PALLARA (2000): *Functions of bounded variation and free discontinuity problems*. Clarendon Press, Oxford.

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

DE GIORGI, E. (1954): "Su una 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 deux 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., AND C. KOOPERBERG (2002): "Spline Adaptation in Extended Linear Models," *Statistical Science*, 17, 2–51.

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.

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.

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

KOENKER, R., AND I. MIZERA (2002): "Comment on Hansen and Kooperberg: Spline Adaptation in Extended Linear Models," *Statistical Science*, 17, 30–31.

KOENKER, R., AND P. NG (2002): "SparseM: A Sparse Matrix Package for R," `http://cran.r-project.org/src/contrib/PACKAGES.html#SparseM`.

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].

MAMMEN, E., AND S. VAN DE GEER (1997): "Locally Adaptive Regression Splines," *The Annals of Statistics*, 25, 387–413.

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

MEYER, M., AND M. WOODROOFE (2000): "On the degrees of freedom in shape-restricted regression," *Annals*, 28, 1083–1104.

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

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

PORTNOY, S. (1997): "Local Asymptotics for Quantile Smoothing Splines," *The Annals of Statistics*, 25, 414–434.

PORTNOY, S., AND R. KOENKER (1997): "The Gaussian Hare and the Laplacian Tortoise: Computability of squared-error versus absolute-error estimators, with discusssion," *Stat. Science*, 12, 279–300.

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

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

RUDIN, L., S. OSHER, AND E. FATEMI (1992): "Nonlinear total variation based noise removal algorithms," *Physica D*, 60, 259–268.

RUDIN, L. I., AND S. OSHER (1994): "Total variation based image restoration with free local constraints," in *Proceeedings ICIP-94*, vol. I. IEEE Computer Society Press.

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.

TASDIZEN, T., R. WHITAKER, P. BURCHARD, AND S. OSHER (2002): "Geometric Surface Processing via Normal Maps," preprint.

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.

WAHBA, G. (1990): *Spline models for observational data*, CBMS-NSF Regional Conference Series in Applied Mathematics. SIAM, Philadelphia.

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