Total Variation Regularization for Bivariate Density Estimation

Roger Koenker

University of Illinois, Urbana-Champaign

Oberwolfach: November, 2006

Joint work with Ivan Mizera, University of Alberta

Outline

The Dirac Catastrophe

Naïve application of maximum likelihood:

$$\max_{f \in \mathcal{F}} \sum_{i=1}^{n} \log f(X_i)$$

for any sufficiently rich class of densities $\mathcal{F}$ yields

$$\hat{f}(x) = dF_n = n^{-1} \sum_{i=1}^{n} \delta_{X_i}(x)$$

Vapnik (1998) Density estimation as a stochastic ill-posed problem.

Regularization as Good Thinking

I.J. Good's (1971) "Bayesian in mufti" penalized MLE:

$$\max_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \log f(x_i) - \int f(x)dx - \lambda \int ((\sqrt{f(x)})')^2 dx$$

or, with $u = \sqrt{f}$,

$$\max_{u \in \mathcal{U}} 2 \int \log u(x) dF_n(x) - \| u \|^2 - \lambda \| u' \|^2$$

Penalty Interpretation: Shrinking toward minimal Fisher information.
Euler Condition: $\frac{dF_n}{u} = u - \lambda u''$
Boundary Condition: $\lim_{x \to \pm\infty} u(x) = 0$.

Good Thinking 2.0

Good and Gaskins (1971) observed that the penalty,

$$J(f) = \int \left((\sqrt{f})' \right)^2 dx$$

produced $\hat{f}$'s that "looked too straight," and suggested the penalty,

$$J(f) = \int \left((\sqrt{f})' \right)^2 dx + \int \left((\sqrt{f})'' \right)^2 dx$$

focusing more on curvature as a measure of roughness.

Penalizing Log Density

Silverman (1982) proposed penalizing the third derivative of $\log f$,

$$J(f) = \int \left((\log f)'''\right)^2 dx$$

thereby shrinking toward the Gaussian density: $J(\phi) = 0$. Subsequent authors have emphasized the classical smoothing spline penalty:

$$J(f) = \int \left((\log f)''\right)^2 dx$$

Gu (2002) has an R implementation. This logspline approach has obvious roots in exponential family theory, Stone et al (1997) and Barron and Sheu (1991).

One, too, many Regularizations

Another early approach was the histosplines of Boneva, Kendall and Stefanov (1971)

$$\min_f\{\int (f(x) - f_n(x))^2 dx + \lambda \int (f^{(k)}(x))^2 dx$$

where f_n is a preliminary, undersmoothed histogram.
Closely related are the more recent proposals of Vapnik (1998)

$$\min_F\{\int (F(x) - F_n(x))^2 dx + \lambda \int (F^{(k)}(x))^2 dx$$

$$\min_F\{(\int |F(x) - F_n(x)| dx)^2 + \lambda \int (F^{(k)}(x))^2 dx$$

where F_n is the usual empirical cdf,

Regularization for Bivariate Density Estimation

For bivariate densities there are fewer proposals. The only implemented proposal seems to be the thin-plate log density estimator:

$$\min_{g} \sum_{i=1}^{n} g(z_i) + \lambda J(g)$$

with roughness penalty:

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

Equivariant to translation and rotation
Domain $\Omega = \mathbb{R}^2$ is convenient, but not universally appropriate.
Meinguet(1979), Wahba(1990), Green and Silverman (1998), Gu (2002).

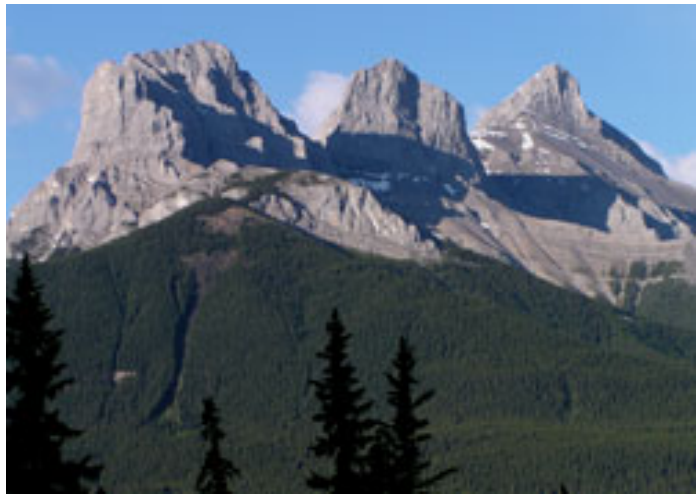
Total Variation Regularization

Replacing $\mathcal{L}_2$ by $\mathcal{L}_1$ penalties leads to total variation regularization.
For regression problems such penalties have been considered by:

- Rudin, Osher, and Fatemi (1992)
- Koenker, Ng, and Portnoy (1994)
- Mammen and van de Geer (1997)
- Davies and Kovac (2001)
- Koenker and Mizera (2004)
- Sardy and Tseng (2005).

Motivation for Total Variation Roughness

- L_2 penalties abhor sharp bends, good for gently rolling hills.
- L_1 (total variation) penalties better tolerate sharp peaks.



Three Variations on Total Variation for $f : [a, b] \to \mathbb{R}$

- Jordan(1881)

$$\bigvee f = \sup_{\pi} \sum_{k=0}^{n-1} |f(x_{k+1}) - f(x_k)|$$

  where π denotes partitions: $a = x_0 < x_1 < \ldots < x_n = b$.

- Banach (1925)

$$\bigvee f = \int N(y)dy$$

  where $N(y) = \text{card}\{x : f(x) = y\}$ is the Banach indicatrix.

- Vitali (1905)

$$\bigvee f = \int |f'(x)|dx$$

  for absolutely continuous f.

Total Variation for $f: \mathbb{R}^k \to \mathbb{R}^m$

A convoluted history from Vitali (1905) to de Giorgi (1954)
For smooth $f: \mathbb{R} \to \mathbb{R}$

$$\bigvee_{\Omega} f = \int_{\Omega} |f'(x)| dx$$

For smooth $f: \mathbb{R}^k \to \mathbb{R}^m$

$$\bigvee_{\Omega} f = \int_{\Omega} \| \nabla f(x) \| dx$$

Extension to nondifferentiable f via theory of distributions.

$$\bigvee_{\Omega} f = \int_{\Omega} \| \nabla f(x) * \varphi_\epsilon \| dx$$

Total Variation Penalties

Univariate: $\Omega \subset \mathbb{R}^1$

$$J_0(f) = \bigvee_{\Omega}(\log f) = \int_{\Omega} |(\log f)'| dx$$

$$J_1(f) = \bigvee_{\Omega}(\log f)' = \int_{\Omega} |(\log f)''| dx$$

Bivariate: $\Omega \subset \mathbb{R}^2$

$$J_0(f) = \bigvee_{\Omega} \log f = \int_{\Omega} \|\nabla \log f\| dx$$

$$J_1(f) = \bigvee_{\Omega} \nabla \log f = \int_{\Omega} \|\nabla^2 \log f\| dx$$

TV Penalized Maximum Likelihood Log Density Estimation

We have the generic problem, optimizing over a set of densities $\mathcal{F}$

$$\max_{f \in \mathcal{F}} \{ \sum \log f(X_i) \mid J(f) \leqslant K \}$$

We started by focusing on penalizing total variation of $g = \log f$, and considering Lagrangean expressions like,

$$\max_{g} \{ \int g dF_n - \int e^g - \lambda \| Dg \| \}$$

where D is some linear (differential) operator, e.g. $Dg = g''$, or $Dg = \nabla^2 g$, and $\| \cdot \|$ is an appropriate norm. For total variation we would have, for example,

$$\bigvee \nabla g = \int \| \nabla^2 g(x) \|_1 dx.$$

Duality

These are a convex optimization problems with interesting dual problems. The primal penalized maximum likelihood problem

$$\max_{g}\{\int g\,dF_n - \int e^g - \lambda\|Dg\|\}$$

has equivalent dual formulation,

$$\max_{h}\{-\int f\log f \mid f = d(F_n + D^*h)/dx \geqslant 0, \|h\|^* \leqslant \lambda\}$$

where D^* is the adjoint of D, and $\|\cdot\|^*$ is the dual of the norm $\|\cdot\|$. So we are (really) maximizing (Shannon) entropy!

Our Favorite Univariate Example

For univariate densities solving the primal problem,

$$\max_{g}\{\int g\,dF_n - \int e^g - \lambda \int |g''|\}$$

is equivalent to solving the dual problem,

$$\max_{h}\{-\int f \log f \mid f = d(F_n + h'')/dx \geqslant 0,\ \|h\|_\infty \leqslant \lambda\}$$

- Dual of L_1 norm is the L_∞ norm.
- Solution log densities are piecewise linear.
- Sup constraint on h is a (generalized) tube restriction.
- There are many variations on this construction.

Arcana of Implementation

- Optimization
  - Interior Point Methods
  - Sparse Algebra
- Discretization
  - Delone/Voronoi Tesselations
  - Total Variation for Simple Functions
- Domains
  - Undata Analysis
  - Boundary Constraints

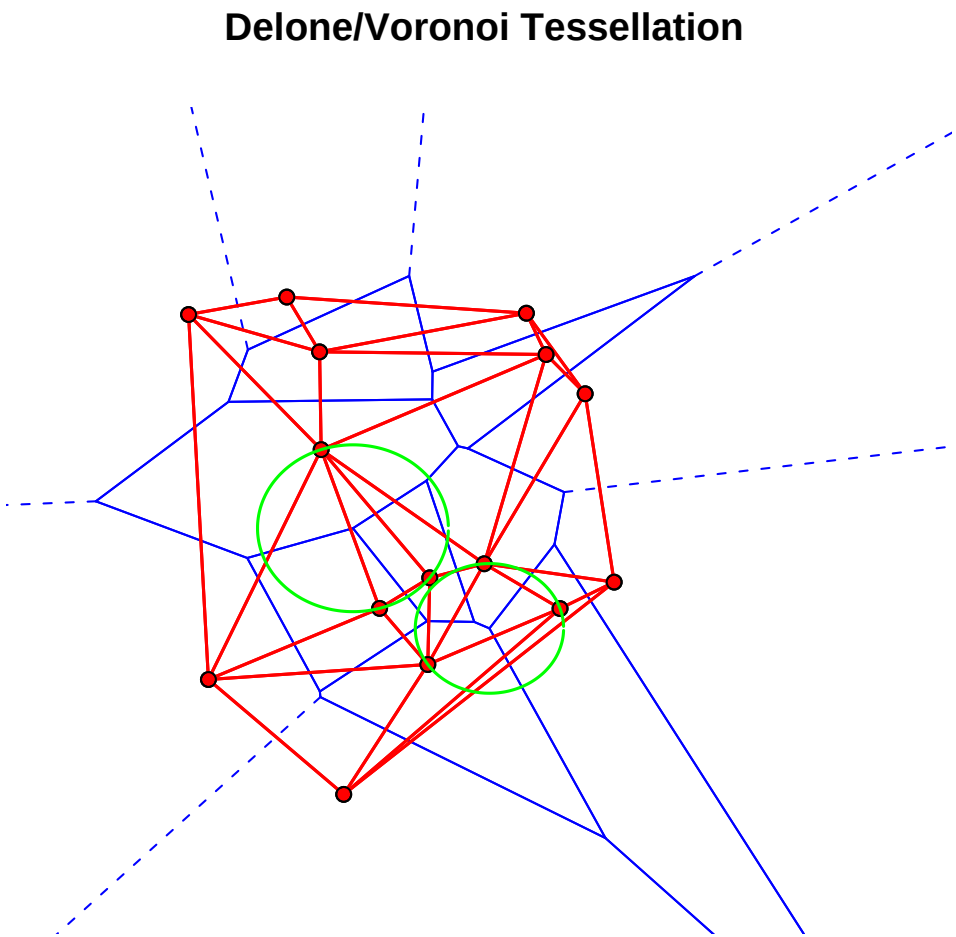
Interior Point Optimization

There has been a revolution in convex optimization:

- Frisch (1954), Fiacco and McCormick (1968) Karmarker (1984), ...
- Inequality constraints are replaced by logarithmic barrier functions enabling Newton-type steps.
- Sparse linear algebra is essential element for large probems.
- We are using a Danish commercial implementation called Mosek by Erling Anderson, and open source code by Michael Saunders.

Discretization

For scattered data one natural discretization employs piecewise linear functions on Delone triangulations:



Total Variation for Triograms

Theorem: For functions $f : \mathbb{R}^2 \to \mathbb{R}$ continuous and piecewise linear on a triangulation, $\mathcal{T}$, and any rotationally invariant norm, $\|\cdot\|$, chosen to define:

$$\bigvee \nabla f = \int \|\nabla^2 f\| dx$$

we have

$$\bigvee \nabla f = \sum_{e \in \mathcal{E}} \|\nabla_e^+ - \nabla_e^-\|_2 \, \|e\|_2,$$

where summation is over all the edges of the triangulation, ∇_e^+, ∇_e^- are gradients on the two triangles adjacent to the edge e, and $\|e\|_2$ denotes the length of the edge e.

Remark: Parameterizing f by its vector of function values $\xi = (f(x_i))$ at the vertices of $\mathcal{T}$ we can write,

$$\bigvee \nabla f = \|A\xi\|_1,$$

that is, as an ℓ_1-norm of a linear transformation of ξ. The matrix A is extremely sparse.

Undata Analysis

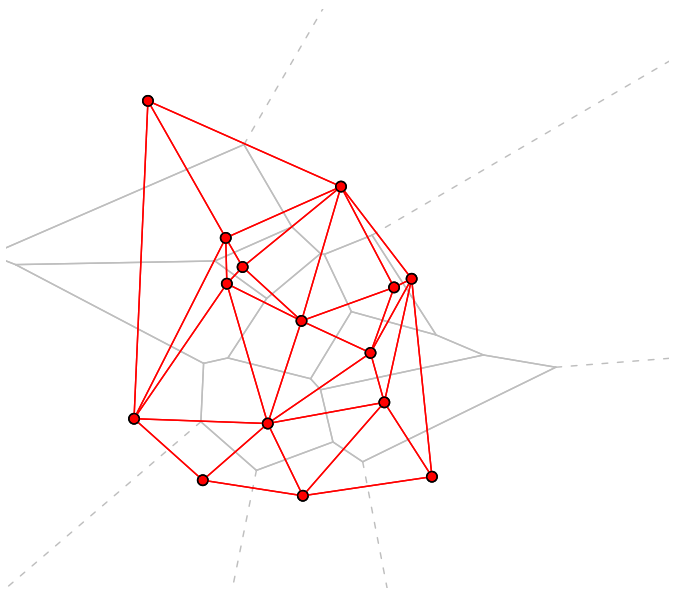
An intrinsic difficulty of density estimation involves boundary conditions and tail behavior. This draws us into the shadowy realm of undata analysis.

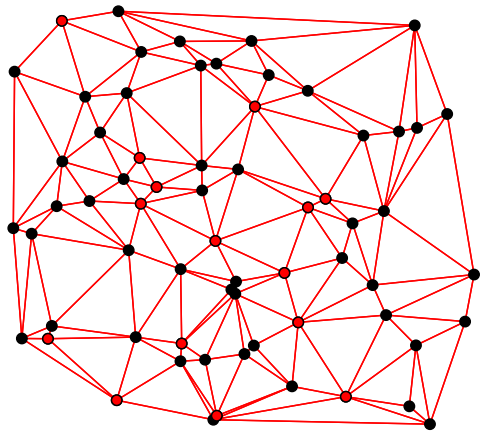


In the Linnean idiom, undata are living things with wavy edges.

Undata Analysis

In our context, undata are vertices of the Delone triangulation that aren't observed – points that do appear in the "prior," so to speak, but do not contribute to the fidelity/likelihood. They enable us to extend the domain beyond the observed points and increase the flexibility of the triangulation in the interior of the domain.





Chicago Land Values I

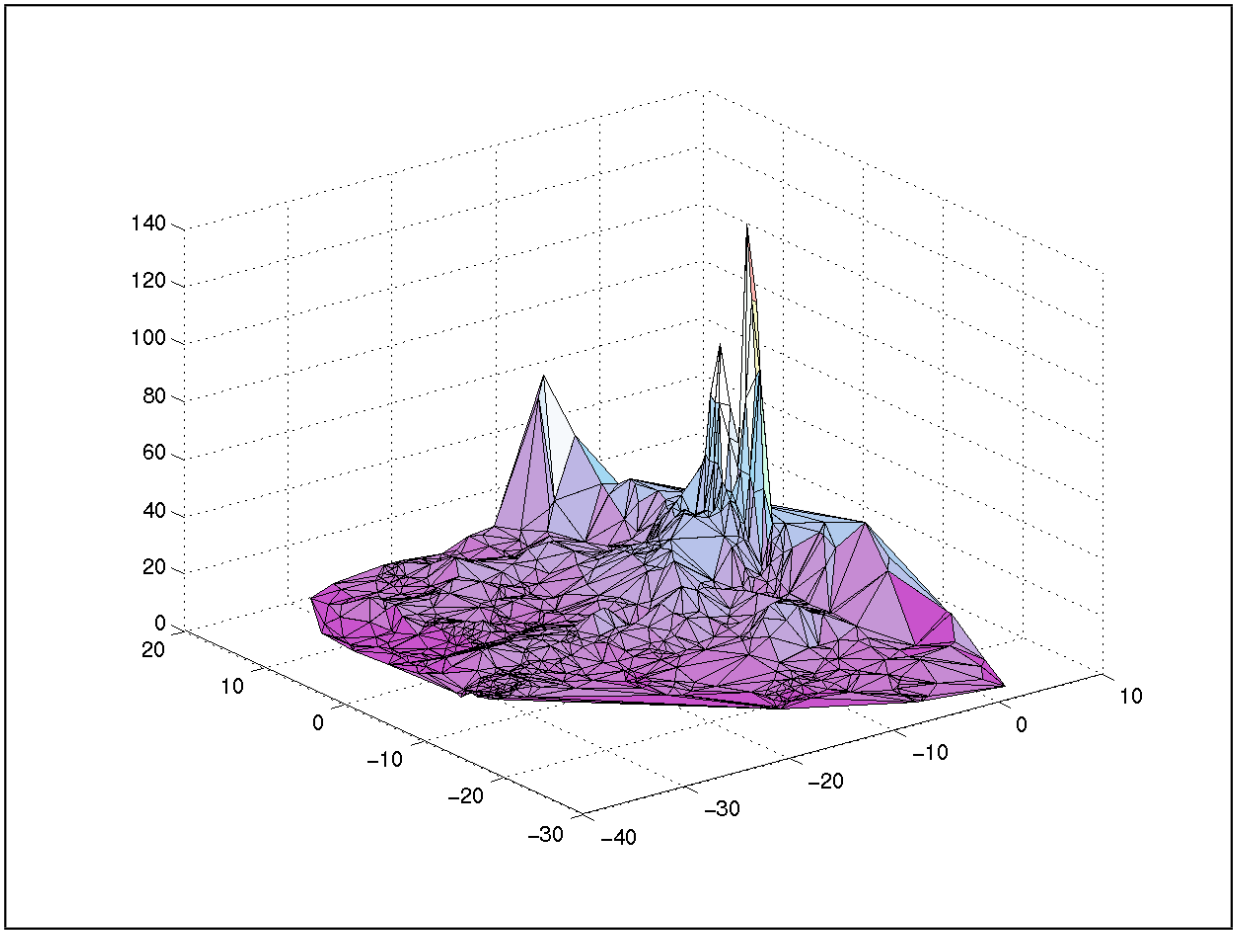


Figure: Perspective Plot of Median Regression Model for Chicago Land Values. Based on 1194 vacant land sales in Chicago Metropolitan Area in 1995-97, prices in dollars per square foot.

Chicago Land Values II

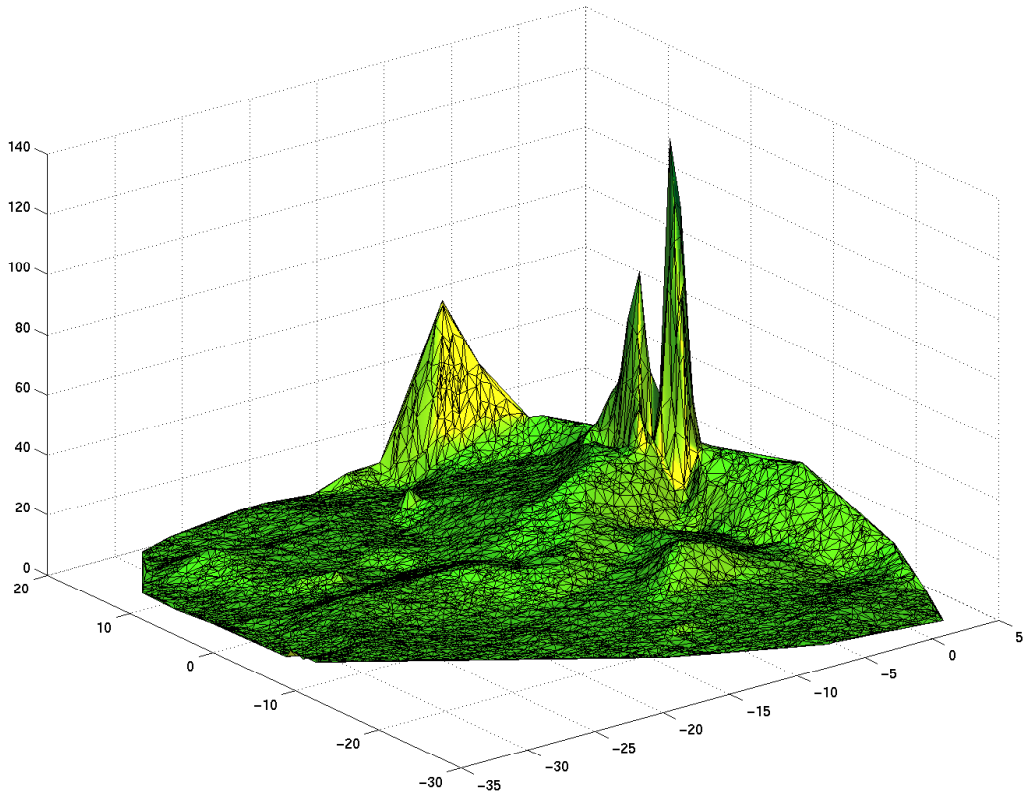
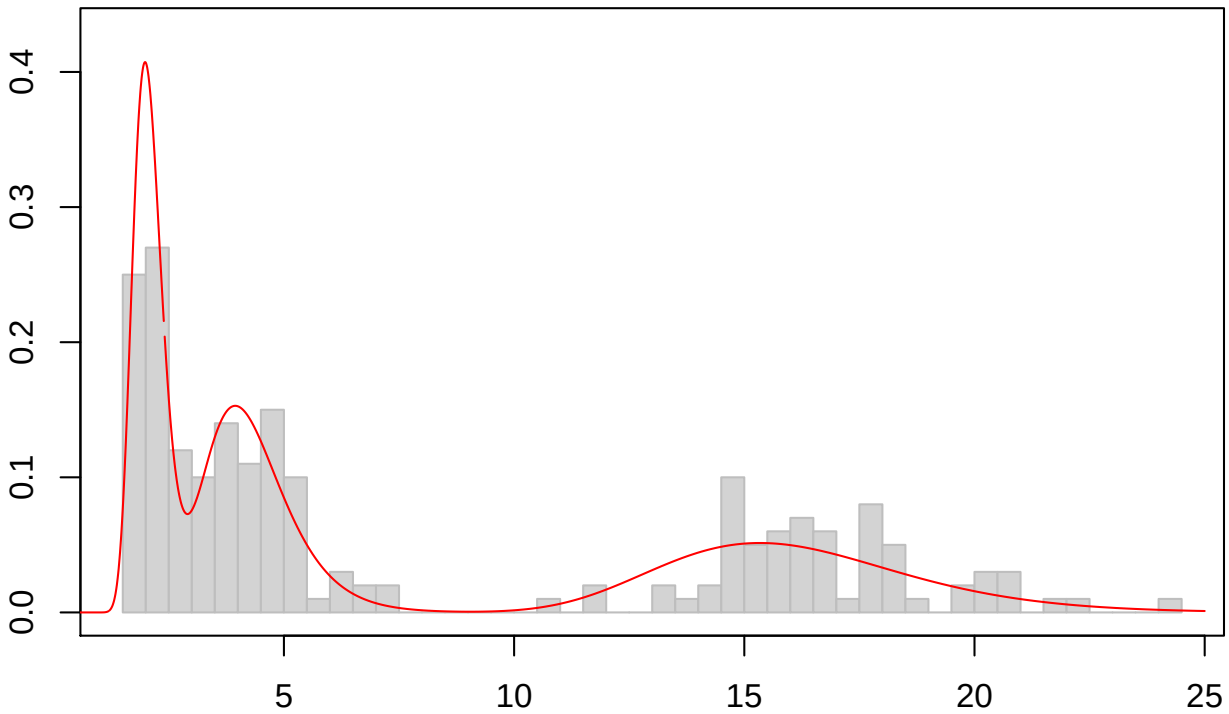


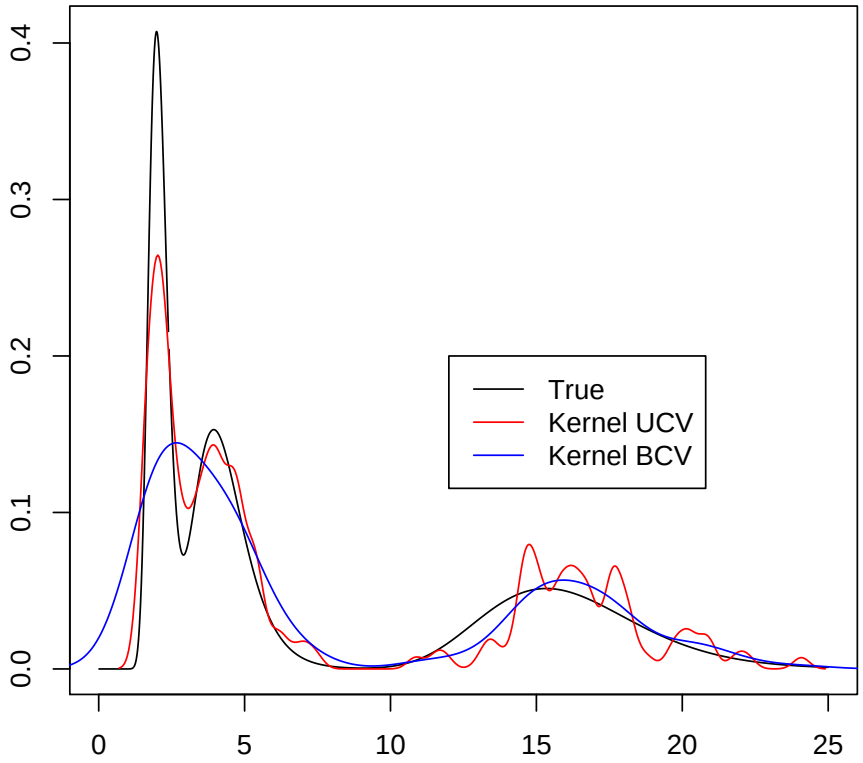
Figure: Chicago Land Values: Based on 1194 vacant land sales and 7505 "virtual" sales (undata) introduced to increase the flexibility of surface.

Density Estimation: Beyond the Histogram



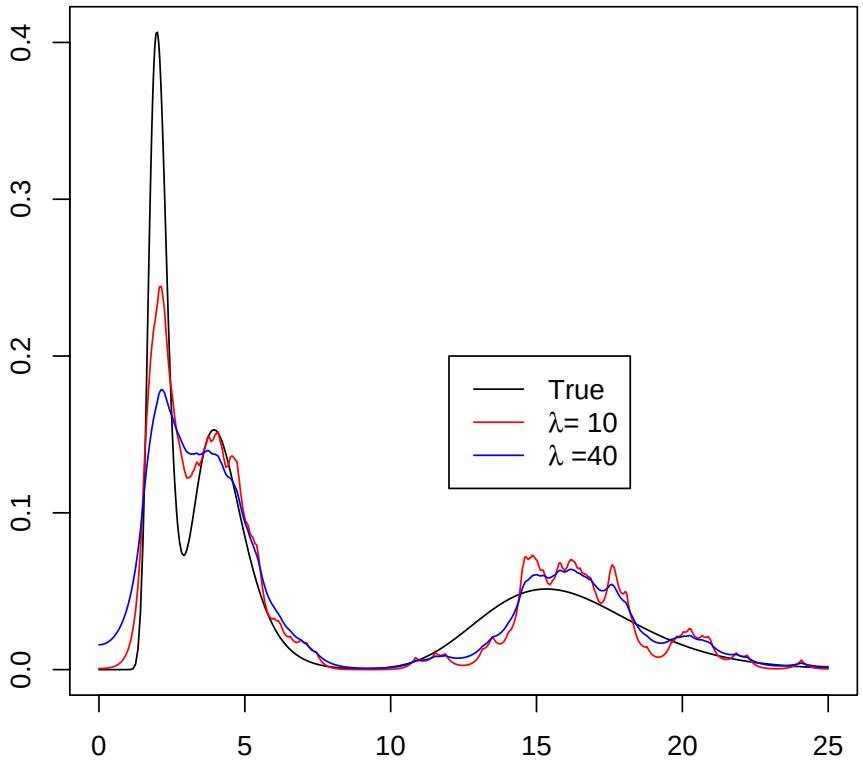
How do various methods perform for this test problem?

Kernel Density Estimation



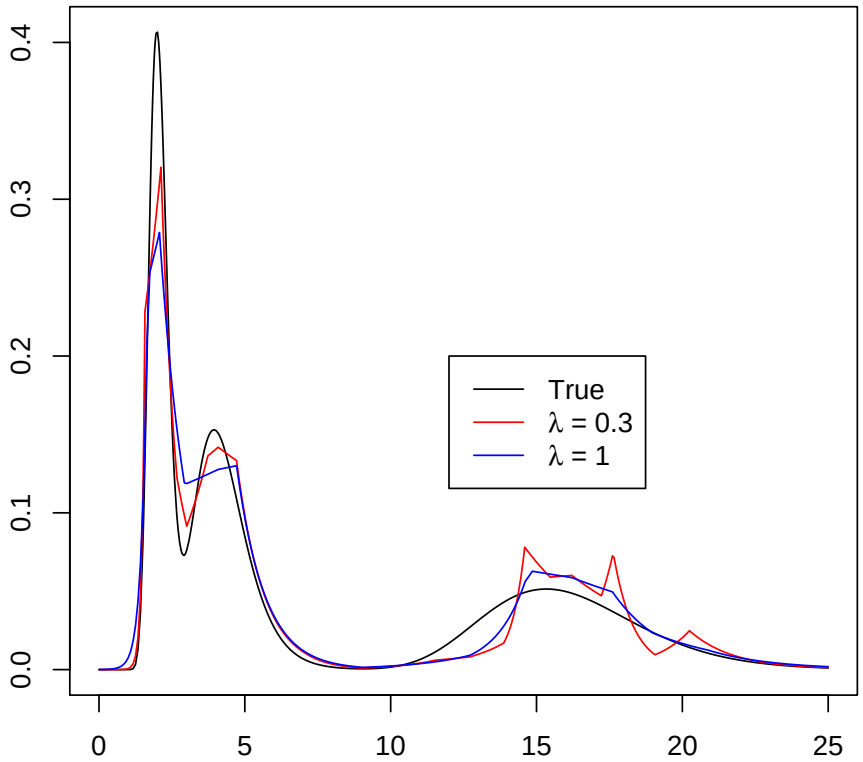
Two kernel estimates using Scott's cross-validation bandwidth selection.

Good's Penalized Density Estimation



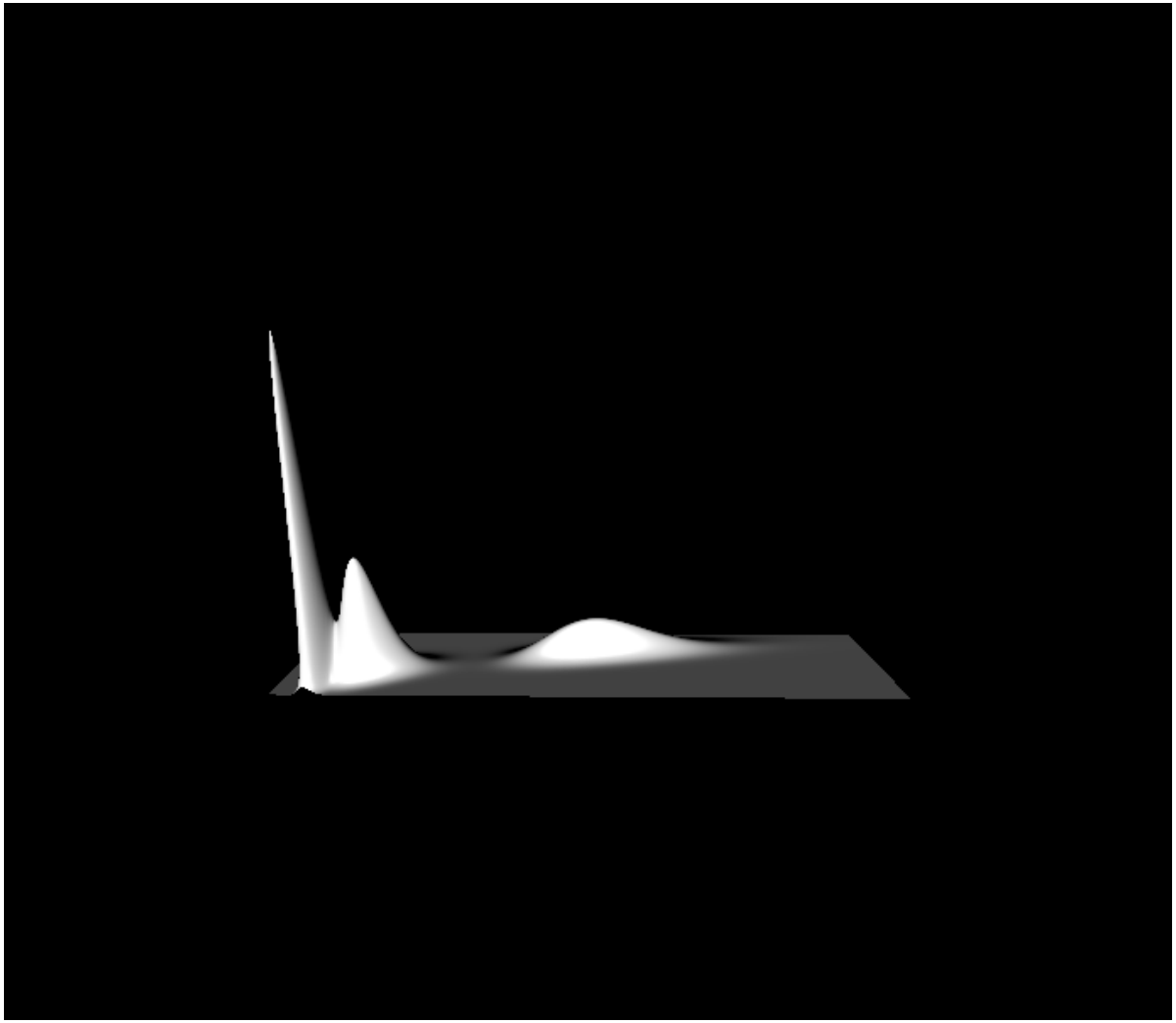
Good's (1971) original (Fisher information) L_2 penalty.

Total Variation Penalized Density Estimation



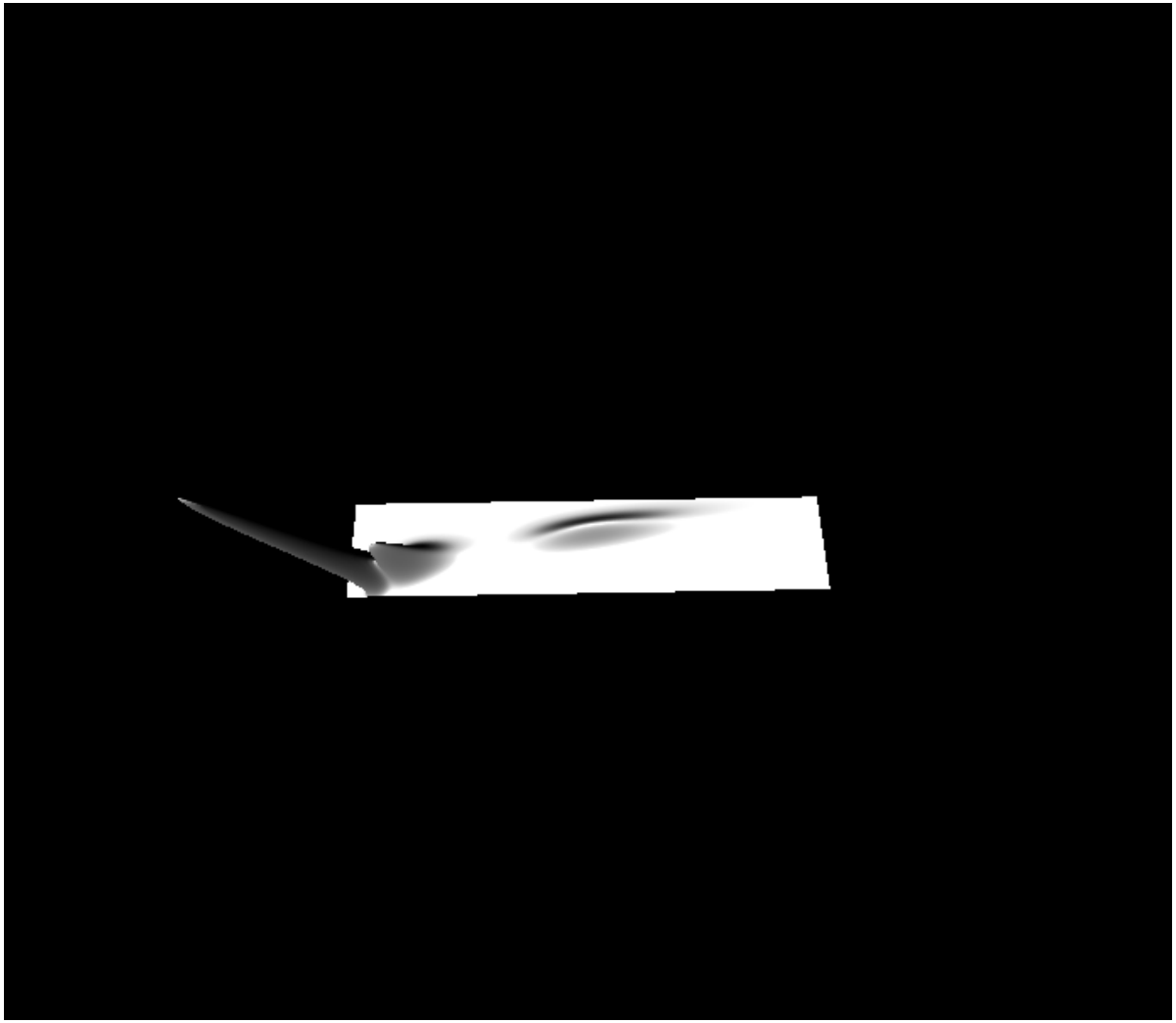
Total Variation penalty on $(\log f)'$.

A Bivariate Target Density



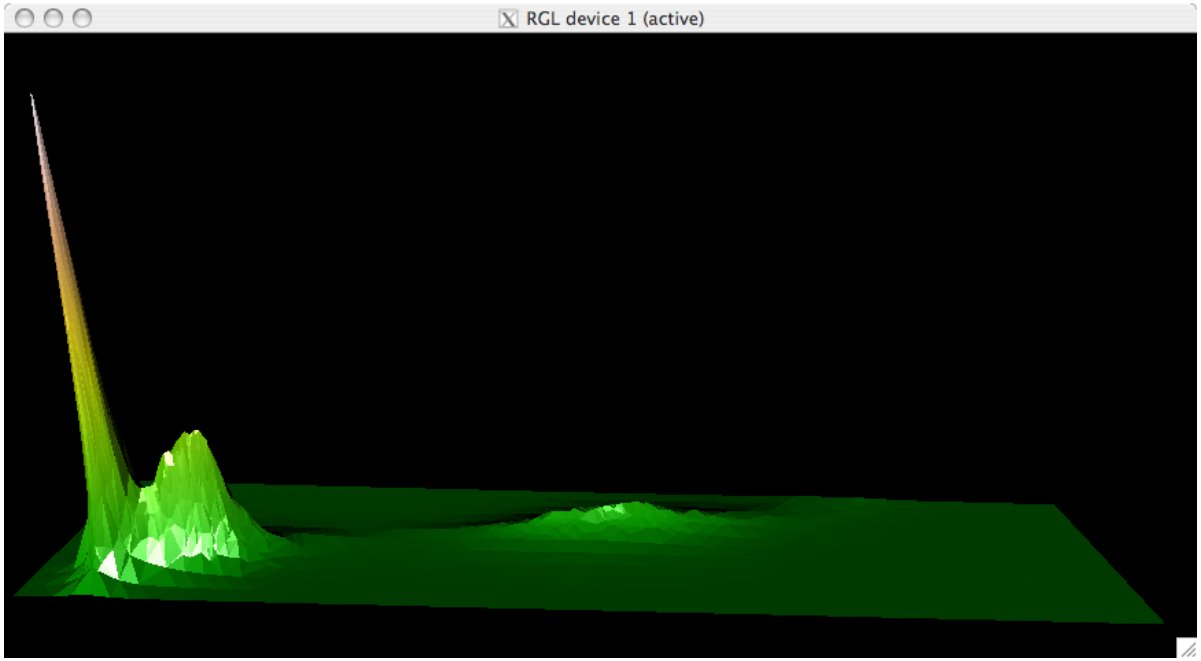
A bivariate version of the mixture of lognormals example.

A Bivariate Target Density



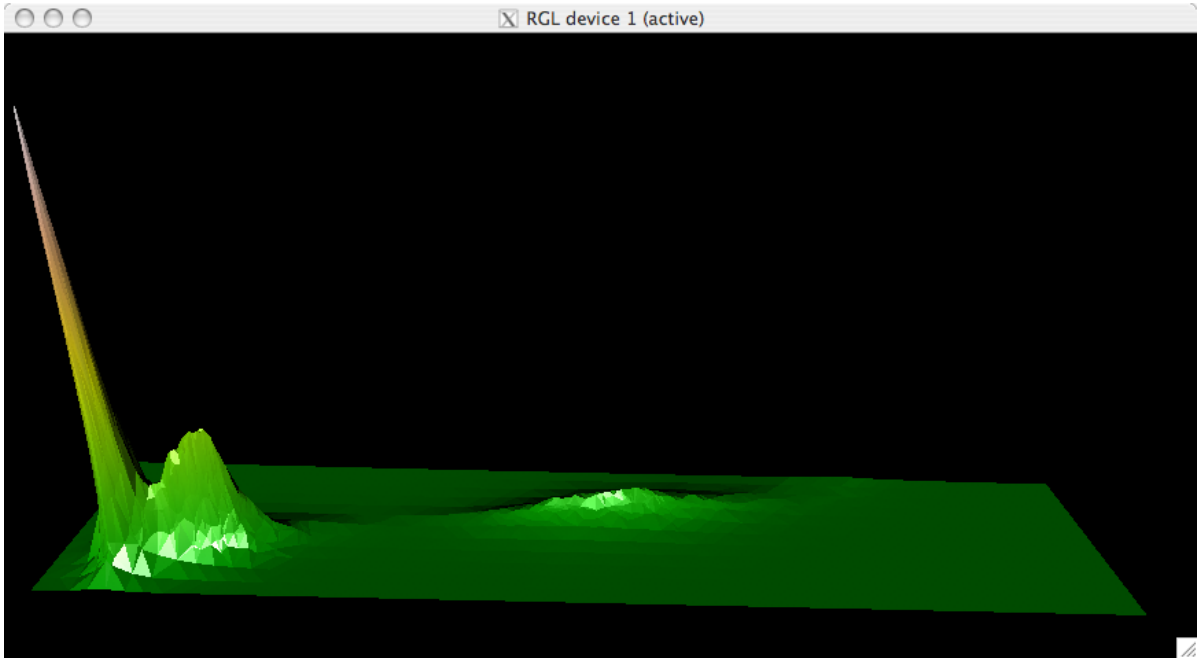
A bivariate version of the mixture of lognormals example.

Bivariate Total Variation Penalized Density Estimate



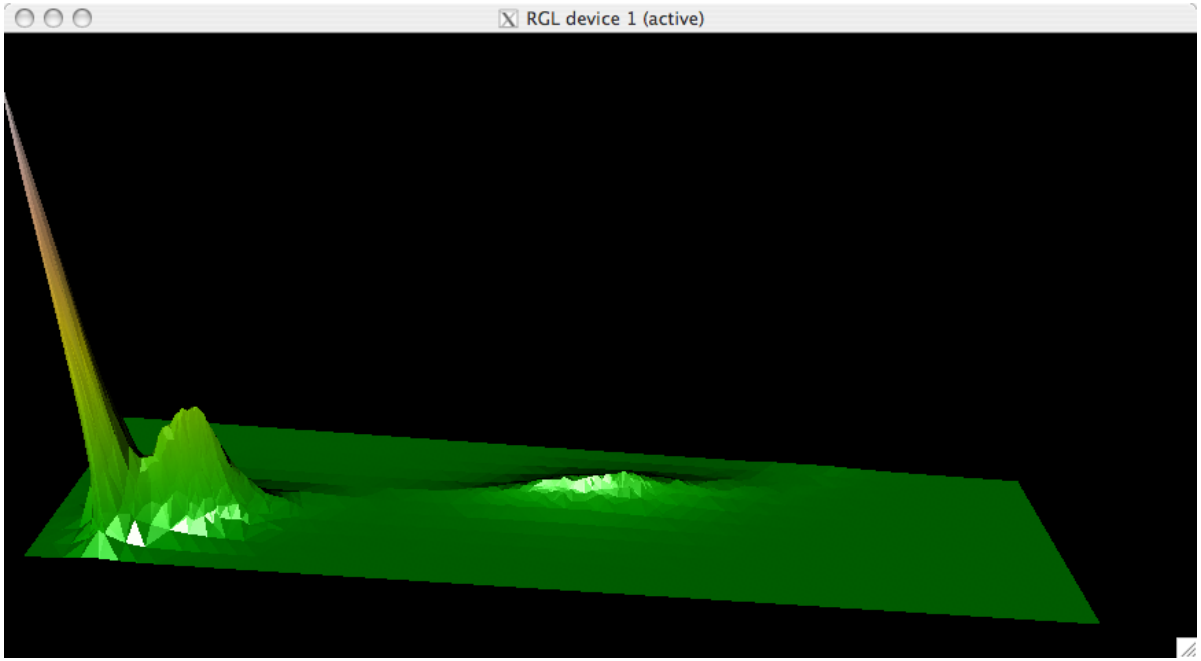
$\bigvee_\Omega \nabla \log f$ penalty, $\lambda = 2$.

Bivariate Total Variation Penalized Density Estimate



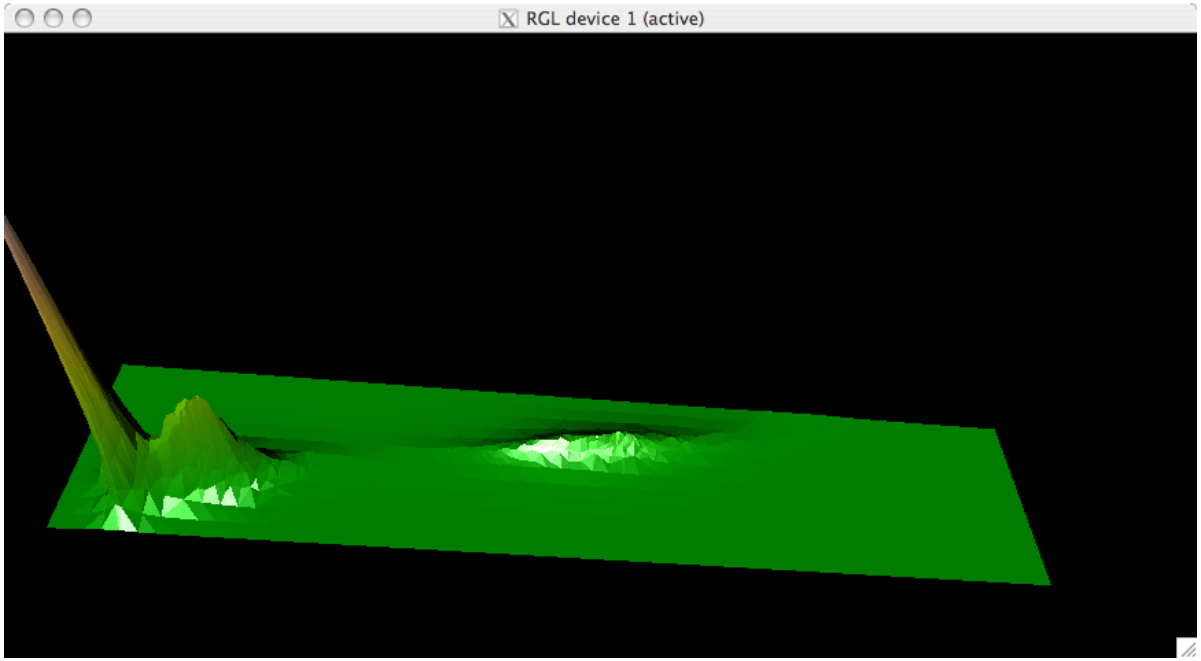
$\bigvee_\Omega \nabla \log f$ penalty, $\lambda = 2$.

Bivariate Total Variation Penalized Density Estimate



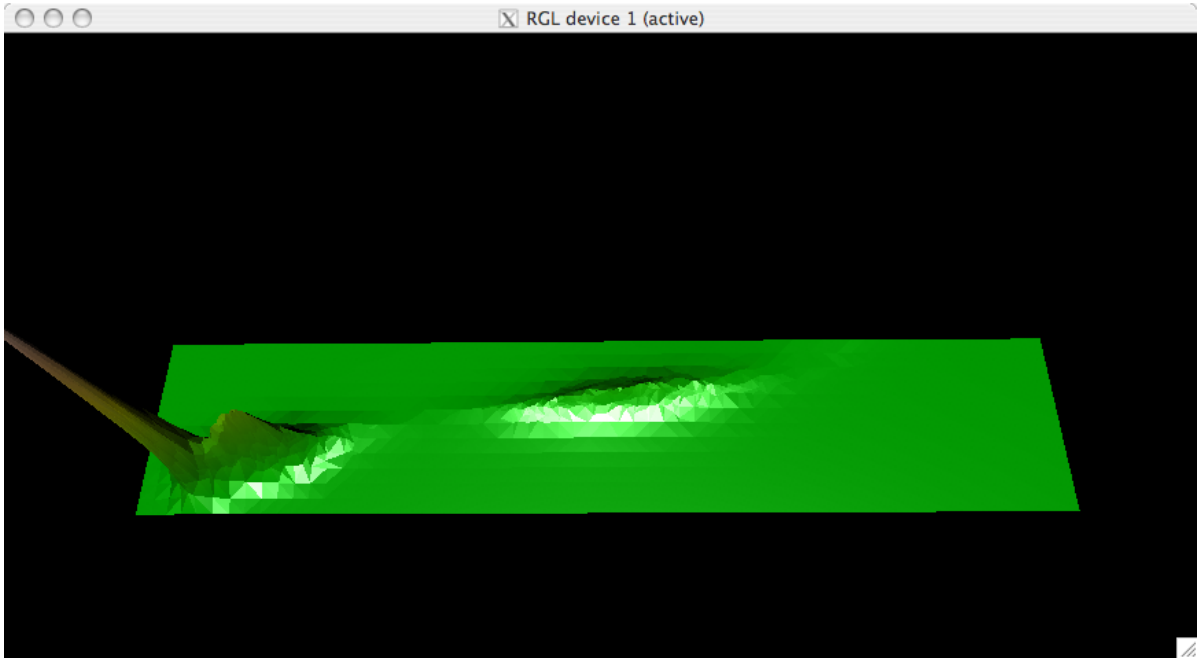
$\bigvee_\Omega \nabla \log f$ penalty, $\lambda = 2$.

Bivariate Total Variation Penalized Density Estimate



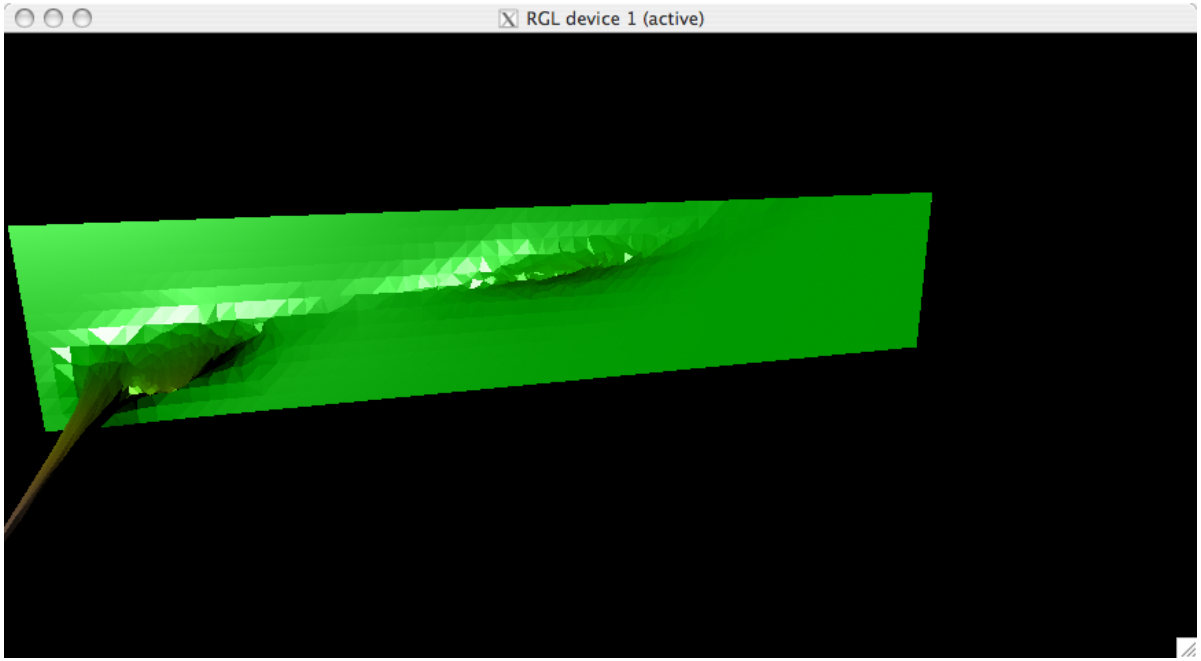
$\bigvee_\Omega \nabla \log f$ penalty, $\lambda = 2$.

Bivariate Total Variation Penalized Density Estimate



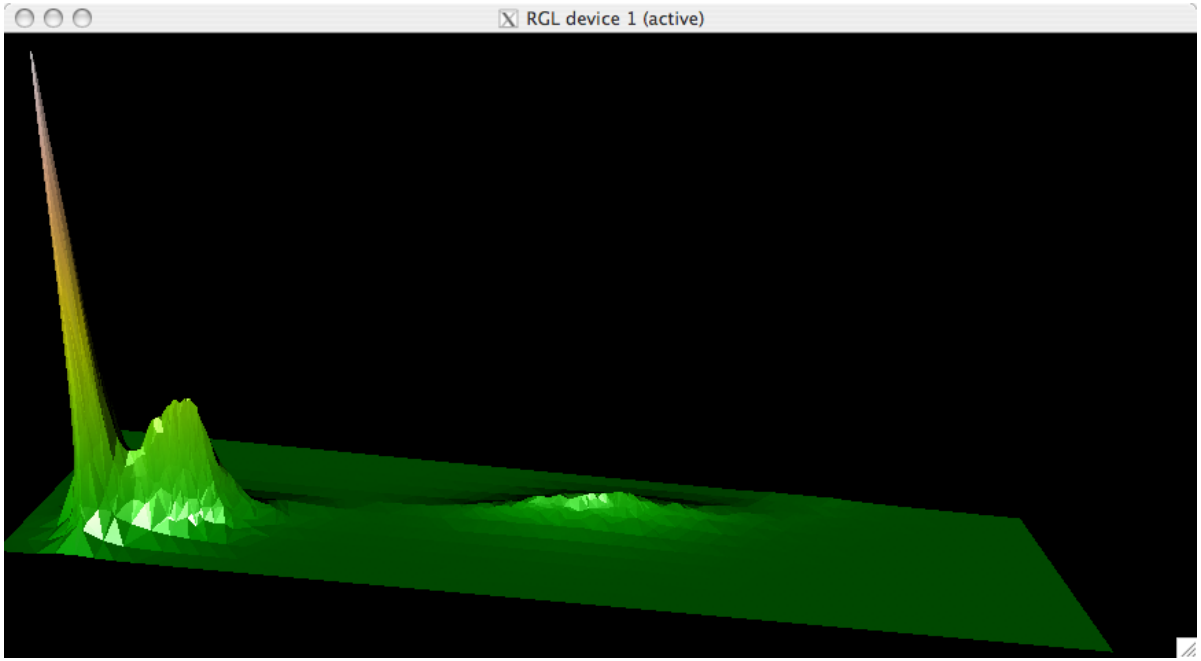
$\bigvee_\Omega \nabla \log f$ penalty, $\lambda = 2$.

Bivariate Total Variation Penalized Density Estimate



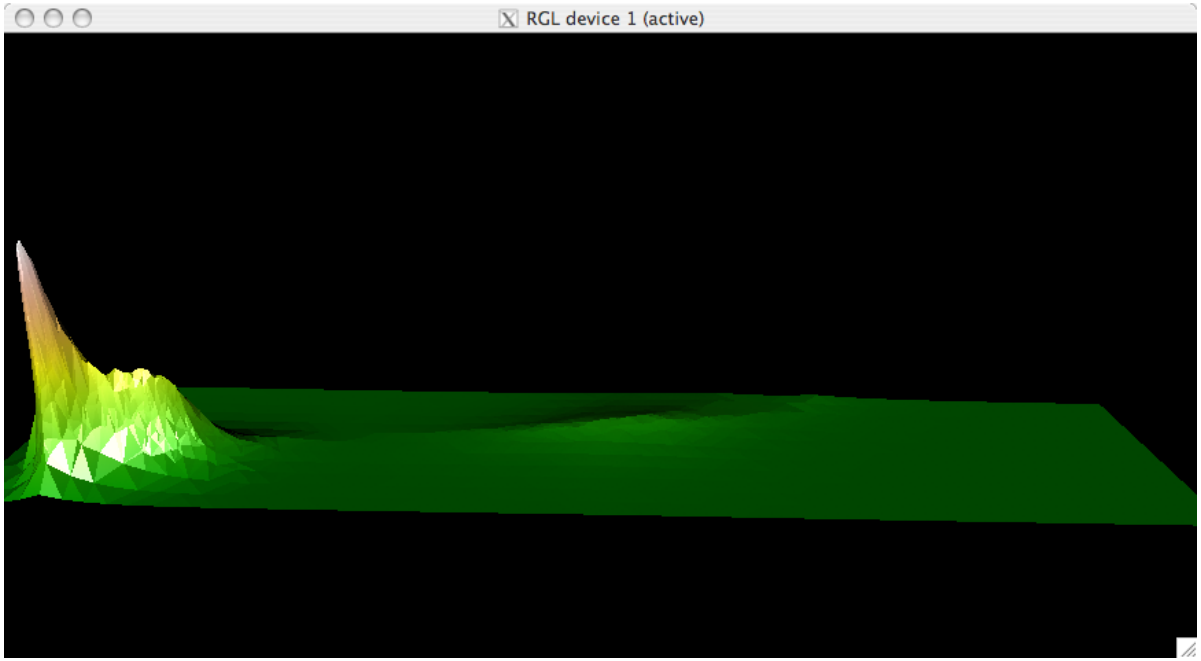
$\bigvee_\Omega \nabla \log f$ penalty, $\lambda = 2$.

Bivariate Total Variation Penalized Density Estimate



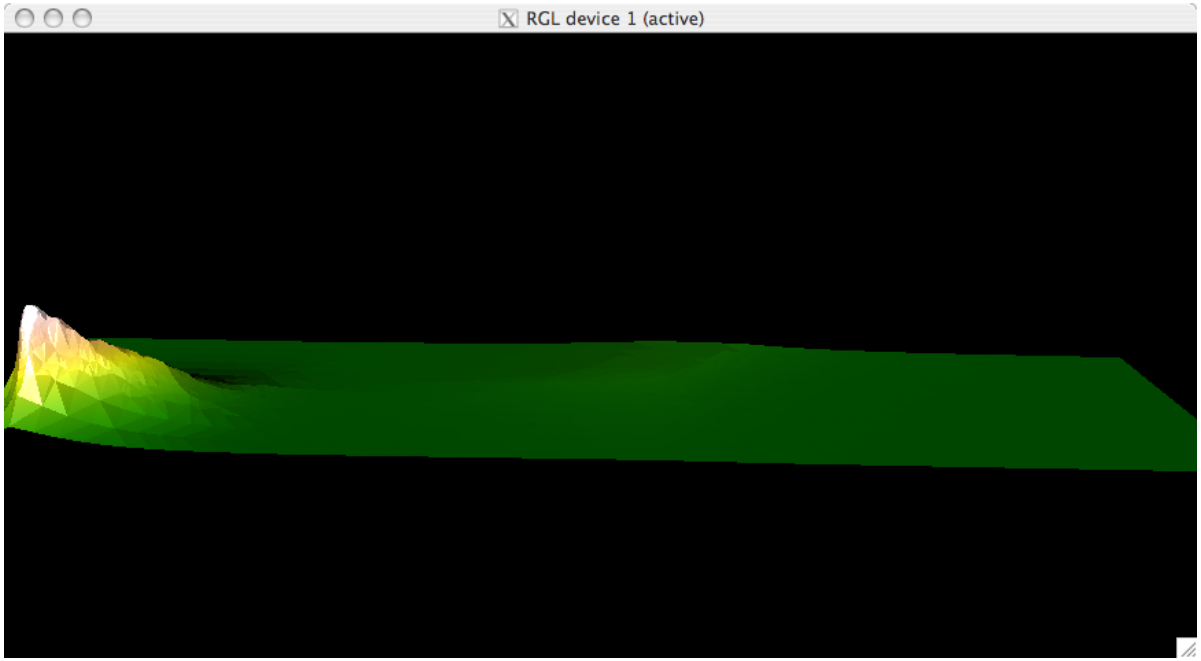
$\bigvee_\Omega \nabla \log f$ penalty, $\lambda = 2$.

Larger λ's Flatten the Estimate



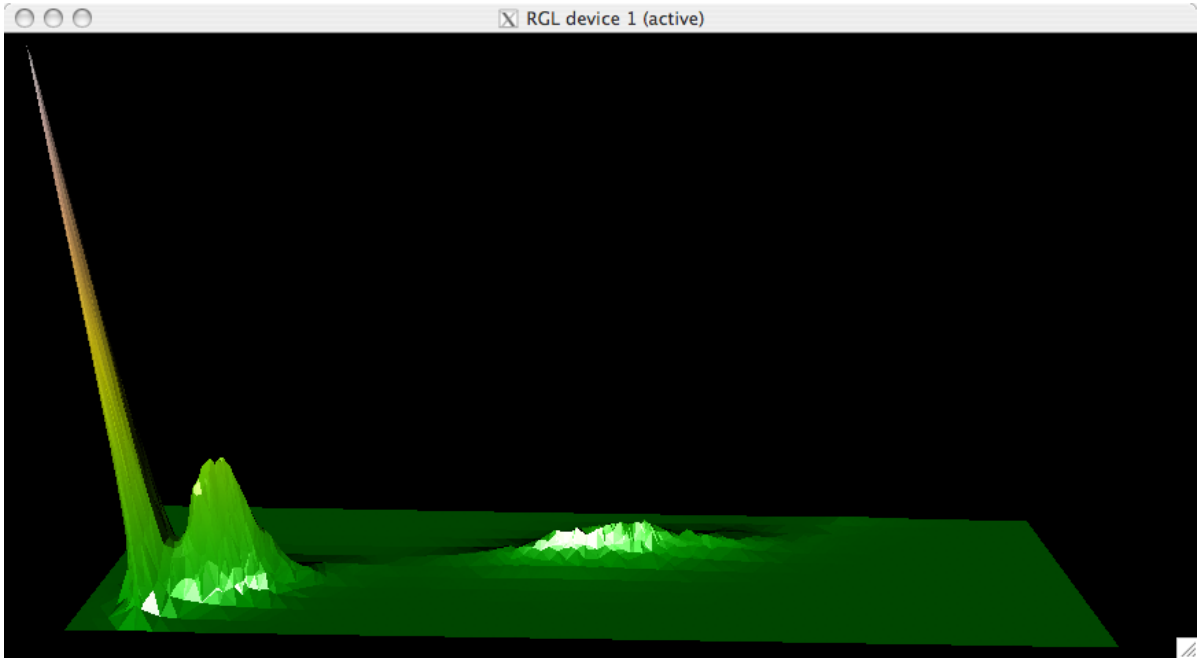
$\bigvee_{\Omega} \nabla \log f$ penalty, $\lambda = 5$.

Larger λ's Flatten the Estimate



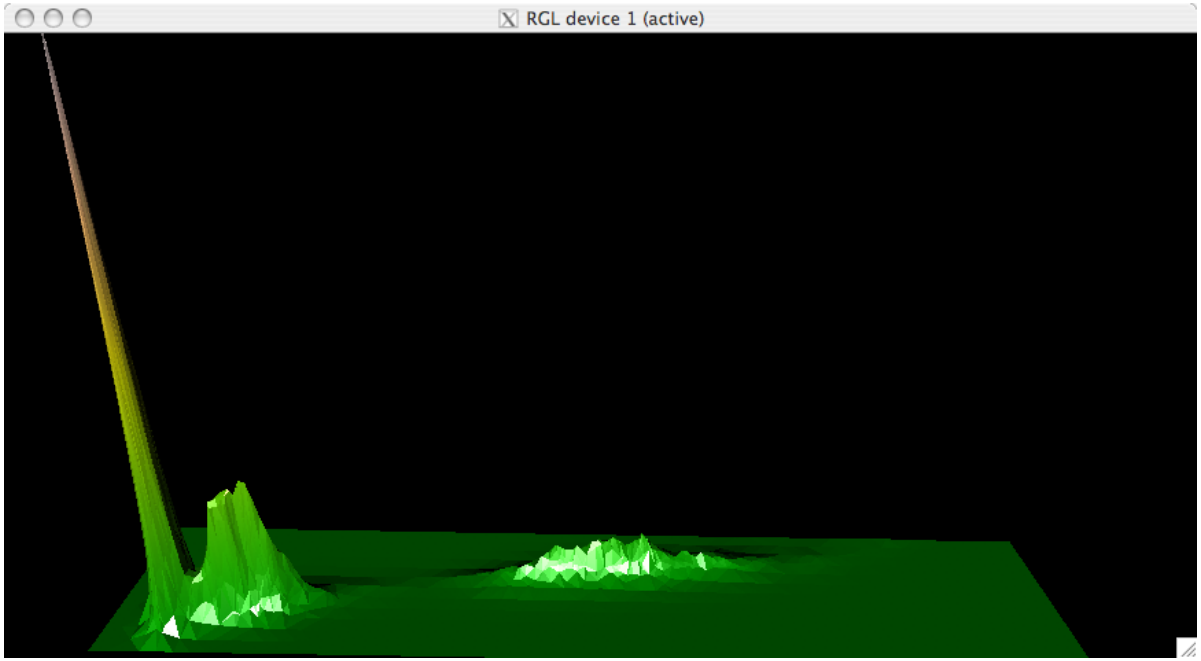
$\bigvee_\Omega \nabla \log f$ penalty, $\lambda = 10$.

Smaller λ's Roughen the Estimate



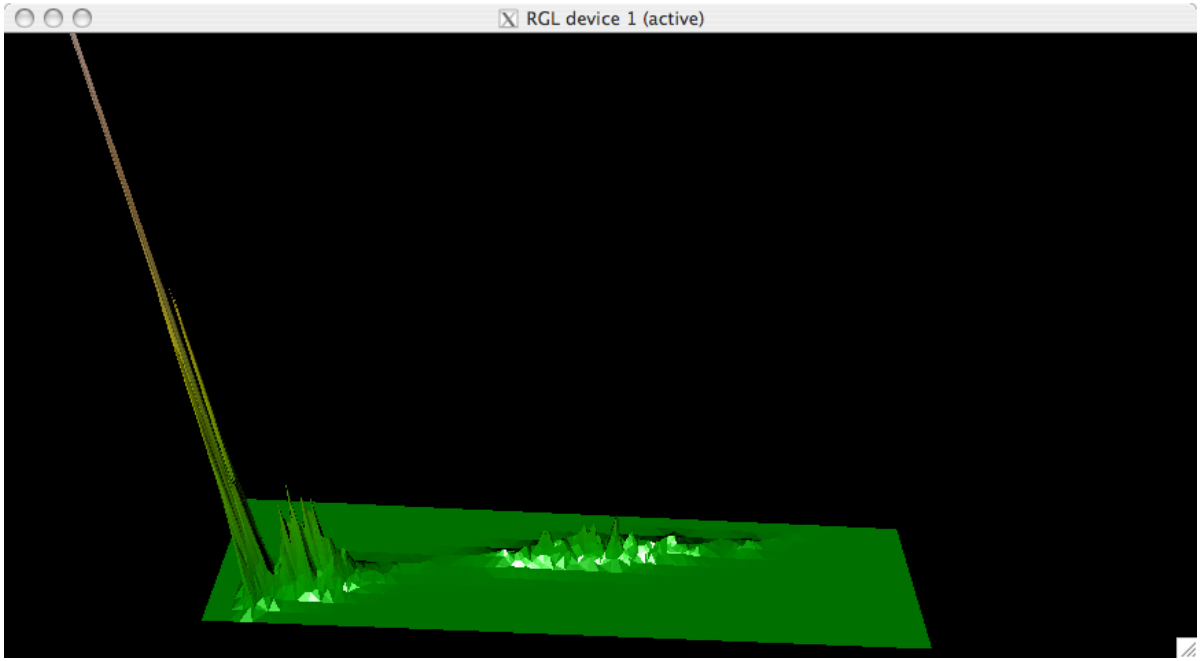
$\bigvee_\Omega \nabla \log f$ penalty, $\lambda = 1$.

Smaller λ's Roughen the Estimate



$\bigvee_\Omega \nabla \log f$ penalty, $\lambda = 0.5$.

Smaller λ's Roughen the Estimate



$\bigvee_\Omega \nabla \log f$ penalty, $\lambda = 0.1$.

Dogma of Goniolatry (Revised)

"Goniolatry, or the worship of angles, ..."
Thomas Pynchon (*Mason and Dixon*, 1997).

- Regularization provides a unified framework for density estimation.
- Duality leads to interesting connections to maximum entropy estimation.
- Total variation is a natural roughness penalty for some density estimation problems, particularly when the target density is edgy.
- Finite element methods and sparse linear algebra are computationally very crucial.
- Qualitative constraints and extensions to semi-parametric models are possible.

For Further Details

- Koenker and Mizera (2006) Density Estimation by TV Regularization, *Doksum Festschrift*.
- Koenker and Mizera (2006) The Alter Egos of the Regularized Likelihood Density Estimators: Deregularized maximum entropy Shannon, Rényi, Simpson, Gini and Stretched Strings, *Proceedings of the 7th Prague Symposium*.
- Koenker and Mizera (2006) Primal and Dual Formulations Relevant for the Numerical Estimation of a Probability Density via Regularization, *Tatra Mountains Math. Pub.*

Available from `http://www.econ.uiuc.edu/~roger`