Supplementary materials for this article are available online.
Please click the JCGS link at http://pubs.amstat.org.

Fast TV Regularization for 2D Maximum Penalized Likelihood Estimation

George O. MOHLER, Andrea L. BERTOZZI,
Thomas A. GOLDSTEIN, and Stanley J. OSHER

Total Variation-based regularization, well established for image processing applications such as denoising, was recently introduced for Maximum Penalized Likelihood Estimation (MPLE) as an effective way to estimate nonsmooth probability densities. While the estimates show promise for a variety of applications, the nonlinearity of the regularization leads to computational challenges, especially in multidimensions. In this article we present a numerical methodology, based upon the Split Bregman L1 minimization technique, that overcomes these challenges, allowing for the fast and accurate computation of 2D TV-based MPLE. We test the methodology with several examples, including V-fold cross-validation with large 2D datasets, and highlight the application of TV-based MPLE to point process crime modeling. The proposed algorithm is implemented as the Matlab function TVMPLE. The Matlab (mex) code and datasets for examples and simulations are available as online supplements.

Key Words: Crime; Density estimation; Spatial point process; Split Bregman minimization; Total Variation.

1. INTRODUCTION

We consider the following problem, referred to as *density estimation*, in this article: given an iid sample $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_N \in \mathbb{R}^d$ with common probability density $u(\mathbf{x}), \mathbf{x} \in \mathbb{R}^d$, construct from the sample an estimate $\hat{u}$ of the unknown density u. Maximum Penalized Likelihood Estimation (MPLE) provides a general framework for constructing such an estimate, in which a regularized version of the log-likelihood function is maximized:

$$\hat{u}(\mathbf{x}) = \underset{v(\mathbf{x}) \geq 0, \|v\|_1 = 1}{\operatorname{argmax}} \left\{ \sum_{i=1}^{N} \log(v(\mathbf{y}_i)) - \alpha R(v) \right\}. \tag{1.1}$$

George O. Mohler is Assistant Adjunct Professor of Mathematics (E-mail: *gmohler@math.ucla.edu*), Andrea L. Bertozzi is Professor of Mathematics, Thomas A. Goldstein is Research Associate, and Stanley J. Osher is Professor of Mathematics, Department of Mathematics, University of California, Los Angeles, CA 90095-1555.

Journal of Computational and Graphical Statistics, Volume 20, Number 2, Pages 479–491
DOI: 10.1198/jcgs.2010.09048

The first term on the right of (1.1) is the log-likelihood function and the second term is introduced to enforce the probability density estimate to possess regularity (Eggermont and LaRiccia 2001). While a variety of penalty functionals $R(v)$ have appeared in the literature (for a review of MPLE see Eggermont and LaRiccia 2001), many standard methods of both MPLE and non-MPLE type perform poorly when the underlying probability density of the data has sharp gradients (Sardy and Tseng 2006).

To improve estimates in the case of nonsmooth densities, Koenker and Mizera (2006) and Sardy and Tseng (2006) proposed taking the penalty to be the Total Variation (TV) of the density,

$$\hat{u}(\mathbf{x}) = \underset{v(\mathbf{x})\geq 0, \|v\|_1=1}{\operatorname{argmax}} \left\{ \sum_{i=1}^{N} \log(v(\mathbf{y}_i)) - \alpha \int |\nabla v(\mathbf{x})| \, d\mathbf{x} \right\}. \tag{1.2}$$

In the work of Sardy and Tseng (2006), the estimate given by (1.2) was shown to outperform estimators such as the taut string (Davies and Kovac 2004), logspline (Kooperberg and Stone 2002), and rectangular kernel with global bandwidth (Sheather and Jones 1991) for a variety of nonsmooth target densities. However, the results in the work of Sardy and Tseng (2006) are for single-variable probability densities and, in the multidimensional setting, the efficient solution of the optimization problem on the right side of (1.2) is nontrivial.

Similar types of optimization problems arise in image processing and a number of computational methods have been developed for their solution. For example, the Rudin–Osher–Fatemi (ROF) model,

$$\hat{u}(\mathbf{x}) = \underset{v(\mathbf{x})}{\operatorname{argmin}} \left\{ \frac{\mu}{2} \int (f(\mathbf{x}) - v(\mathbf{x}))^2 \, d\mathbf{x} + \int |\nabla v(\mathbf{x})| \, d\mathbf{x} \right\}, \tag{1.3}$$

constructs the *denoised* estimate $\hat{u}(\mathbf{x})$ of $u(\mathbf{x})$ from an observed noisy image $f(\mathbf{x}) = u(\mathbf{x}) + \xi$ (Rudin, Osher, and Fatemi 1992). Here the noise ξ is assumed to be Gaussian, though similar models can be constructed for other types of noise (Le, Chartrand, and Asake 2007). There is a large body of literature on techniques for solving (1.3) and many of the techniques approach the problem by either solving a regularized form of (1.3) directly, or by attacking the differentiable *dual formulation* of the problem, which requires the enforcement of linear inequality constraints and may require the solution of nonlinear equations.

In this article we present a novel computational method for the fast solution of (1.2) in $d = 2$ spatial dimensions based upon a Split Bregman method developed for image processing applications in the article by Goldstein and Osher (2009). The method is straightforward to implement and solves (1.2) quickly, in $O(n^2)$ operations where n^2 is the number of grid points in the discretization of the spatial domain. Thus computationally intensive parameter selection techniques such as V-fold cross-validation are feasible, even for larger values of V when (1.2) must be solved hundreds or thousands of times. The organization of the article is as follows: In Section 2, we review the Split Bregman methodology as a general technique for L1 minimization. In Section 3, we formulate the methodology for TV-based MPLE, using a spatially discretized approximation in place of

(1.2). In Section 4, we illustrate the efficiency of the Split Bregman method, as well as the potential benefits of TV-based MPLE in the context of crime modeling.

1.1 NOTATION

In our discussion of discrete optimization problems, we employ the following "vector norm" notation to avoid cumbersome summation. Consider a grid function $v_{i,j}$ defined at grid points (i, j) in some rectangular domain Ω. Here we assume for simplicity that Ω has grid spacing $\Delta x = \Delta y = 1$. We shall use the following norm and inner product notations:

$$\|v\|_1 = \sum_{(i,j)\in\Omega} \left|v_{(i,j)}\right|, \qquad \|v\|_2^2 = \sum_{(i,j)\in\Omega} \left|v_{(i,j)}\right|^2.$$

We will also use "∇" to denote the first-order discrete gradient operator and BV norms as follows:

$$(\nabla v)_{i,j} = (v_{i+1,j} - v_{i,j}, v_{i,j+1} - v_{i,j}), \qquad \|\nabla v\|_1 = \sum_{(i,j)\in\Omega} |(\nabla v)_{i,j}|.$$

In some circumstances, we wish to consider grid functions that are vector-valued at each pixel. For the sake of clarity, we shall use the "arrow" superscript to denote such vector-valued quantities. For example, we may write $\vec{d} = \nabla v$ to emphasize that the value of $\vec{d}$ at each grid location is an ordered pair.

2. THE SPLIT BREGMAN METHOD: A GENERAL L1 MINIMIZATION TECHNIQUE

The Split Bregman method (Goldstein and Osher 2009) is a technique for solving general L1-regularized problems of the form

$$\underset{v}{\operatorname{argmin}}\{\|\Phi v\|_1 + H(v)\}, \tag{2.1}$$

where $v \in R^n$, $\Phi : R^n \to R^m$ is a linear operator, and $H(\cdot) : R^n \to R$ is convex. For example, choosing $\Phi = \nabla$ and $H(v) = \frac{\mu}{2}\|v - f\|_2^2$ yields the ROF model.

This Split Bregman method has the advantage that it does not require regularization, continuation, or the enforcement of inequality constraints (Goldstein and Osher 2009). Furthermore, the technique has been shown to be an extremely efficient solver for L1-regularized denoising problems, as well as a large class of problems from compressed sensing.

The Split Bregman method works by "decoupling" the L1 and L2 terms in (2.1), using a splitting originally introduced by Yin et al. (2008b). When we introduce the auxiliary variable $\vec{d} \leftarrow \Phi v$, the problem (2.1) becomes

$$\underset{v}{\operatorname{argmin}}\{\|\vec{d}\|_1 + H(v)\} \quad \text{such that } \vec{d} = \Phi v. \tag{2.2}$$

For example, if we choose $\Phi = \nabla$, where v is a two-dimensional array, then $\vec{d} = (d_x, d_y) = \nabla v$. To solve this constrained problem, we convert it to an unconstrained problem using a quadratic penalty function:

$$\underset{v,\vec{d}}{\operatorname{argmin}}\left\{\|\vec{d}\|_1 + H(v) + \frac{\lambda}{2}\|\vec{d} - \Phi v\|_2^2\right\}. \tag{2.3}$$

This formulation of the problem is advantageous because the unconstrained problem (2.3) can be solved using a simple alternating minimization scheme (Yin et al. 2008b; Goldstein and Osher 2009). The first step of this alternating scheme is to minimize with respect to v. When $H(\cdot)$ is differentiable, this can usually be done directly by solving a system of equations, or else an approximate solver (such as Gauss–Seidel) can be used to obtain an approximate solution. We next minimize (2.3) with respect to $\vec{d}$. This optimization problem is element-wise decoupled, and the solution can be written explicitly as

$$\vec{d}^* = \operatorname{shrink}(\Phi v, 1/\lambda), \tag{2.4}$$

where

$$\operatorname{shrink}(\vec{z}, \lambda)_i = \max\{\|z_i\|_2 - \lambda, 0\}\frac{\vec{z}_i}{\|z_i\|_2}.$$

Note that the quadratic penalty function in (2.3) only approximately enforces the constraint $\vec{d} = \Phi v$. We wish to enforce this constraint exactly. A standard approach to this problem is to use a continuation scheme: solve (2.3) with an increasing sequence of penalty parameters, $\lambda_1 < \lambda_2 < \cdots < \lambda_n$. Unfortunately, for large values of λ, minimization with respect to v in (2.3) becomes ill-conditioned and the alternating minimization scheme stalls.

To avoid these difficulties, the Split Bregman approach uses a fixed value for λ, and enforces the constraint $\vec{d} = \Phi v$ using a Bregman iteration technique (Goldstein and Osher 2009). For a detailed discussion of this approach, we refer the reader to the works of Chang, He, and Fang (2006), Osher et al. (2005), and Yin et al. (2008a). An in-depth description of the application of this technique to the Split Bregman method can be found in the article by Goldstein and Osher (2009).

To apply Bregman iteration to problem (2.3), we add a vector, $\vec{b}^k$, inside of the quadratic penalty function. We then solve a sequence of unconstrained problems defined by

$$(\hat{u}^k, \vec{d}^k) = \underset{v,\vec{d}}{\operatorname{argmin}}\left\{\|\vec{d}\|_1 + H(v) + \frac{\lambda}{2}\|\vec{d} - \Phi v - \vec{b}^{k-1}\|_2^2\right\}, \tag{2.5}$$

$$\vec{b}^k = \vec{b}^{k-1} + \Phi\hat{u}^k - \vec{d}^k. \tag{2.6}$$

After the alternating minimization scheme approximately solves each unconstrained problem, the Bregman vector is updated using the rule (2.6). This rule is the analog of the "adding back the noise" technique, which has been used to enhance image denoising (Osher et al. 2005). When the minimization (2.5) is (approximately) solved with one iteration of alternating minimization, this scheme becomes

$$\hat{u}^k = \underset{v}{\operatorname{argmin}}\left\{H(v) + \frac{\lambda}{2}\|\vec{d}^{k-1} - \Phi v - \vec{b}^{k-1}\|_2^2\right\}, \tag{2.7}$$

$$\vec{d}^k = \operatorname{shrink}(\Phi\hat{u}^k + \vec{b}^{k-1}, 1/\lambda), \tag{2.8}$$

$$\vec{b}^k = \vec{b}^{k-1} + \Phi\hat{u}^k - \vec{d}^k. \tag{2.9}$$

In the articles by Goldstein and Osher (2009) and Osher et al. (2005), it is shown that (under sufficient assumptions) this algorithm converges in the sense that, as $k \to \infty$, we have $\|\vec{d}^k - \Phi\hat{u}^k\|_2 \to 0$ and $\|\hat{u}^k - \hat{u}\|_2 \to 0$ where $\hat{u}$ is some solution to (2.1).

3. SPLIT BREGMAN IN THE CONTEXT OF TV-BASED MPLE

For our purposes, we wish to solve a discretized problem of the form

$$\operatorname*{argmin}_{v\geq 0}\{\|\nabla v\|_1 + \mu h(v)\} \quad \text{such that } \sum_{i,j} v_{i,j} = 1, \tag{3.1}$$

where $h(v) = -\sum_{i,j} w_{i,j}\log(v_{i,j})$ and w_{ij} is the point count in bin (i, j) of the spatial discretization. To apply the Split Bregman method, we begin by introducing the auxiliary variable $\vec{d} \leftarrow \nabla v$, and adding the corresponding quadratic penalty function as is done in (2.3). However, unlike the formulation (2.3), we have an additional equality constraint because the function v must integrate to unity. To eliminate this constraint, we add an additional quadratic penalty function to get

$$\operatorname*{argmin}_{v\geq 0,\vec{d}}\left\{\|\vec{d}\|_1 + \mu h(v) + \frac{\lambda}{2}\|\vec{d} - \nabla v\|_2^2 + \gamma\left(1 - \sum_{i,j} v_{i,j}\right)^2\right\}, \tag{3.2}$$

where λ and γ are positive constants.

To enforce the equality constraints exactly, we add “Bregman vectors” inside of the penalty functions. These vectors are updated after each unconstrained minimization problem is (approximately) solved. The resulting formulation is

$$(\hat{u}^k, \vec{d}^k) = \operatorname*{argmin}_{v\geq 0,\vec{d}}\left\{\|\vec{d}\|_1 + \mu h(v)\right. \tag{3.3}$$

$$\left. + \frac{\lambda}{2}\|\vec{d} - \nabla v - \vec{b}^{k-1}\|_2^2 + \gamma\left(1 - \sum_{i,j} v_{i,j} - b_1^{k-1}\right)^2\right\}, \tag{3.4}$$

$$\vec{b}^k = \vec{b}^{k-1} + \Phi\hat{u}^k - \vec{d}^k, \tag{3.5}$$

$$b_1^k = b_1^{k-1} + \sum_{i,j}\hat{u}_{i,j}^k - 1. \tag{3.6}$$

All that remains is to describe the solution of the unconstrained optimization problem (3.3)–(3.4). Note that only an approximate solution needs to be computed at each step. We approximately solve this minimization problem using one iteration of the alternating scheme described above. To minimize with respect to $\vec{d}$, we use the explicit formula (2.4). To compute an approximate minimizer with respect to v, we use one sweep of element-wise descent. To derive the element-wise descent formula, we begin by computing the first variation of (3.3) with respect to v. The resulting optimality condition for $v_{i,j}$ is

$$-\frac{\mu w_{i,j}}{v_{i,j}} - \lambda\Delta v_{i,j} + \lambda(\nabla^T\vec{b}_{i,j} - \nabla^T\vec{d}_{i,j}) + \gamma\left(\sum v_{i,j} + b_1 - 1\right) = 0. \tag{3.7}$$

This equation simplifies to a quadratic equation in $v_{i,j}$, which can be written componentwise as

$$(4\lambda + \gamma)v_{i,j}^2 - \alpha_{i,j} v_{i,j} - \mu w_{i,j} = 0, \tag{3.8}$$

where

$$\begin{aligned} \alpha_{i,j} &= \lambda(v_{i+1,j} + v_{i-1,j} + v_{i,j+1} + v_{i,j-1}) && (3.9)\\ &\quad + \lambda(d_{x,i-1,j} - d_{x,i-1,j} + d_{y,i,j-1} - d_{y,i,j}) && (3.10)\\ &\quad + \lambda(b_{x,i-1,j}^k - b_{x,i-1,j}^k - b_{y,i,j-1}^k + b_{y,i,j}^k) && (3.11)\\ &\quad + \gamma\Bigg(1 - b_1^k - \sum_{(i',j')\neq(i,j)} v_{i',j'}\Bigg). && (3.12) \end{aligned}$$

Element-wise minimization is performed by solving this equation at each grid point, and then selecting the positive root. Note that the energy (3.3)–(3.4) is convex with respect to $v_{i,j}$ for $v_{i,j} > 0$. It follows that (3.8) will always have a unique nonnegative root.

Minimization of (3.1) with Dirichlet boundary conditions is accomplished by applying the element-wise minimization formula (3.8) only to interior grid points. To achieve Neumann boundary conditions, we use a slight modification of (3.8). We first adopt the convention that $v_{i,j} = d_{i,j} = b_{i,j} = 0$ whenever the point (i, j) does not lie in the grid domain. We also replace the coefficient $(4\lambda + \gamma)$ in (3.8) with the coefficient $(\beta_{i,j}\lambda + \gamma)$ where $\beta_{i,j}$ is the number of grid points in the set $\{(i+1, j), (i-1, j), (i, j+1), (i, j-1)\}$ that lie in the grid domain.

The choice of parameters λ and γ has a significant impact on the convergence rate of the algorithm. It is desirable to choose large values for these parameters in order to strongly enforce the equality constraints. On the other hand, assigning large values to these parameters may make the optimization problem (3.7) ill-conditioned, and slows down the iterative solver. A simple rule for setting these parameters is derived by considering the linearization of (3.7):

$$\left(\frac{\mu w_{i,j}}{v_{i,j}^2} - \lambda\Delta\right)v_{i,j} + \gamma\Big(\sum v_{i,j} + b_1 - 1\Big) = \lambda(\nabla^T \vec{d}_{i,j} - \nabla^T \vec{b}_{i,j}). \tag{3.13}$$

Note that the first term in the linearization, $\mu w_{i,j}/v_{i,j}^2$, contributes only to the diagonal of the system. To guarantee that the problem (3.7) remains well-conditioned, we choose λ and γ such that the sum of the magnitudes of the off-diagonal terms in the system (3.13) remains comparable to this diagonal term. For an $n \times n$ problem, we expect elements of v to be $O(n^{-2})$ (because they sum to unity). We therefore expect the magnitude of the leftmost, diagonal term in the linear system (3.13) to be $O(n^4\mu)$. The second term in (3.13), $-\lambda\Delta v$, makes an $O(\lambda)$ contribution of the off-diagonal terms, and so we must choose $\lambda = O(\mu n^4)$. The third term involves a summation over all values of v, and makes an $O(\gamma n^2)$ contribution to the off-diagonal terms, and so we must choose $\gamma = O(\mu n^2)$. Empirically, we have found that choosing $\lambda = 2\mu n^4$ and $\gamma = 2\mu n^2$ works well.

4. RESULTS

We first test the Split Bregman method using the Weighted Uniform target density plotted in Figure 1. The target density $u(\mathbf{x}), \mathbf{x} \in [0,1] \times [0,1]$, takes on three values, $u = 2.6060$ (square region), $u = 0.7818$ (outer region), and $u = 0$ (circular region), and has jump discontinuities across the boundaries separating the three regions.

We discretize the 2D spatial region using a 128 × 128 resolution and estimate the target density using 10-fold cross-validation (Sardy and Tseng 2006) for sample sizes of 1000, 4000, and 16,000 points. Letting $\hat{u}^k$ denote the Split Bregman estimate of u at step k, we iterate the Split Bregman method until the stopping criterion,

$$\|\hat{u}^{k+1} - \hat{u}^k\|_2 + \|b^{k+1} - b^k\|_2 + \|b_1^{k+1} - b_1^k\|_2 \le tol, \tag{4.1}$$

is reached. During cross-validation, we use the strictly positive approximate estimate, $\hat{u}^\epsilon = (1-\epsilon)\hat{u} + \epsilon$, where ϵ is a small constant. The reason for this adjustment is that $\hat{u}$ takes on zero values in the circular region and we find that a small number of isolated points near the boundary of the region dominate the log-likelihood function and result in oversmoothed estimates. In this example we take $\epsilon = 10^{-12}$ and find the oversmoothing to be greatly reduced.

In Table 1, we list the Mean Integrated Squared Error (MISE), $E[\int (u(\mathbf{x}) - \hat{u}^\epsilon(\mathbf{x}))^2\, d\mathbf{x}]$, along with the average runtime per sample to complete cross-validation and the average runtime of the Split Bregman method. Using a simple bisection method to find the optimal value of the smoothing parameter, the cross-validation typically requires around 25 parameter values to be evaluated. Since we use 10-fold cross-validation, 10 Split Bregman calls are required per parameter value; however, cross-validation only takes 1–2 min per sample due to the efficiency of the Split Bregman method. Because the grid size is fixed, we actually observe a decrease in the runtime for larger sample sizes due to improved conditioning in the element-wise descent step given by (3.8). For all examples the parameter estimation routines are implemented in MATLAB and the Split Bregman routine is implemented in C (and called from MATLAB).

We point out that using a fast method such as Split Bregman is essential if cross-validation is to be feasible. For example, a classical optimization method such as gradient descent (Rudin, Osher, and Fatemi 1992) with the regularization $\|\nabla v\|_1 \leftarrow \sqrt{(\nabla v)^2 + \beta^2}$ (Acar and Vogel 1994) is appealing for the minimization of (1.2), as it is straightforward

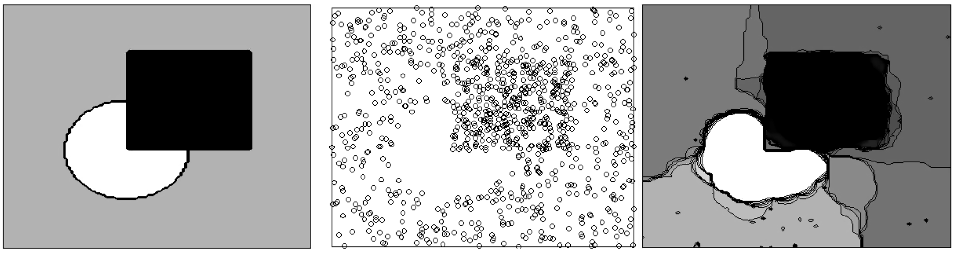

Figure 1. Left: Contour plot of 2D target density. Middle: Sample size of 1000 points from the target density. Right: Contour plot of the Split Bregman estimate with smoothing parameter selected using hand and eye.

Table 1. Computational results for the Split Bregman method.

# of points	MISE ± SE	Avg. runtime/CV	Avg. runtime/SB
1000	0.140 ± 0.004	113.86 sec	0.45 sec
4000	0.103 ± 0.003	63.98 sec	0.26 sec
16,000	0.057 ± 0.001	58.47 sec	0.23 sec

to implement. However, we find that the minimization requires several minutes using such an approach and thus cross-validation, where (1.2) needs to be solved hundreds of times, would take hours or days.

In Figure 2, we compare contour plots of the Split Bregman method, fixed bandwidth Gaussian kernel smoothing, and the average shifted histogram method (Scott 1992) applied to sample sizes of 1000, 4000, and 16,000 points from the target density. For the average shifted histogram we use a biweight kernel (Scott 1992) for the weight distribution. Kernel density estimation (Silverman 1986) is often used for spatial density estimation in fields such as seismology and ecology and has the advantage that it is straightforward to implement (though the method can have high computational cost for large datasets). Similar estimates are obtained using the average shifted histogram, which can be viewed as an approximation to kernel density estimation. The advantage of such a method, however, is that the computational cost scales linearly with the size of the data.

In Figures 1 and 2, the Split Bregman method is able to resolve the sharp gradients of the target density, whereas the kernel density estimate and average shifted histogram (also obtained through 10-fold CV) oversmooth in these regions. We note that the Split Bregman estimate is moderately noisy for low point counts, but this is due in part to the parameter selection process. For instance, we can obtain better qualitative results for lower point counts if the parameter is selected by hand and eye (see Figure 1). As the number of points in the sample increases, the noise disappears from the Split Bregman estimates and the method is able to capture both the sharp gradients and the flat regions quite well (see Figure 2).

TV-Based MPLE and Crime Modeling

Next we highlight an application of the Split Bregman method using residential burglary data collected by the Los Angeles Police Department for the years 2004–2005 within an 18 km × 18 km region of the San Fernando Valley in Los Angeles (see Figure 3). The data consist of the spatial location where the crime occurred (geocoded from the residential address) as well as a time window in which the crime occurred (typically a several-hour window, for instance the time a victim was at work and the house was unoccupied).

Criminological research suggests that victims of personal or property crimes are more likely to be victimized in the near future (see Farrell and Pease 2007; Short et al. 2009) and in the case of residential burglary, evidence indicates that this elevated risk spreads to neighboring houses as well (Johnson et al. 2007). One explanation of this phenomenon is that burglars will often return to the same house, or a neighboring house, shortly after a burglary and commit another offense.

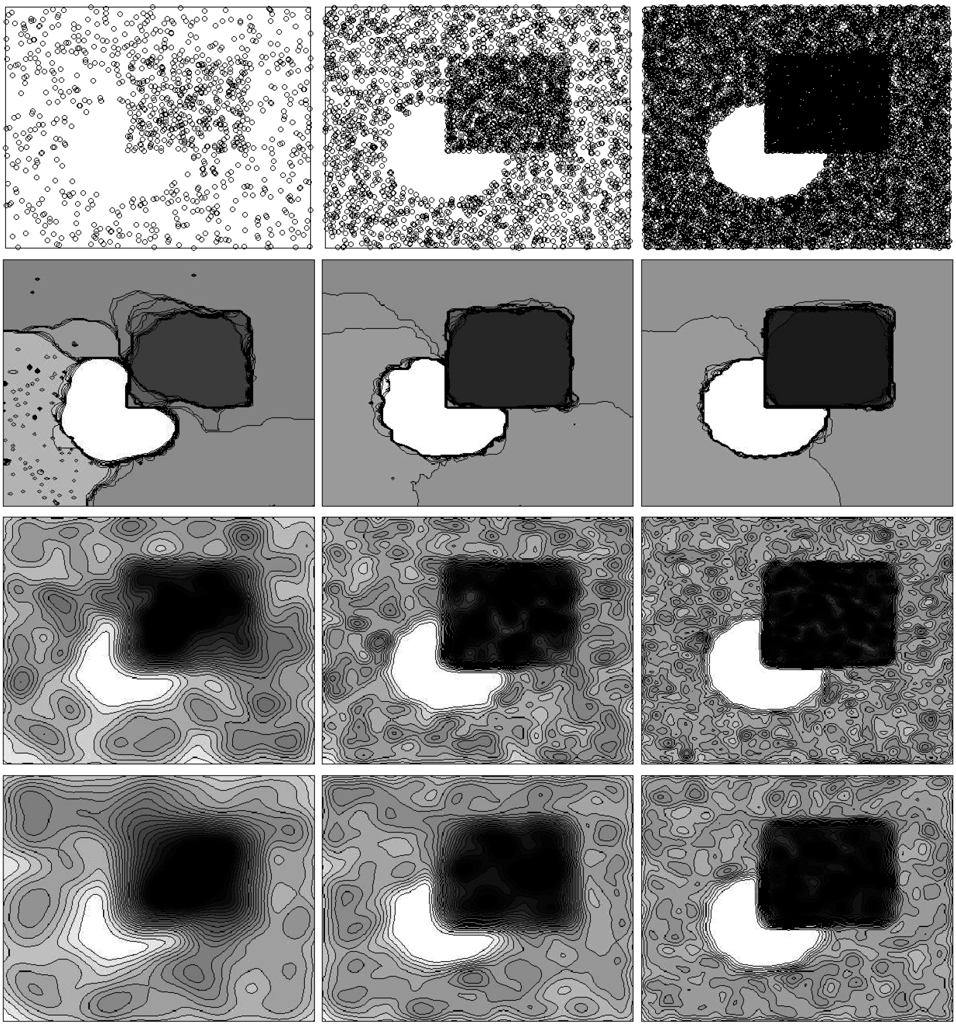

Figure 2. Top row: Sample sizes of 1000, 4000, and 16,000 points from the target density. Second row: Contour plots of the Split Bregman estimate obtained through 10-fold CV. Third row: Contour plots of the Gaussian kernel estimate obtained through 10-fold CV. Fourth row: Contour plots of the average shifted histogram estimate obtained through 10-fold CV.

In the work of Mohler et al. (2008), a 2D self-exciting point process $N(t, x, y)$ is used to model this type of behavior, where the conditional intensity of N is given by

$$\lambda(t, x, y) = \mu(x, y) + \int_{t_0}^{t} \int_{x'} \int_{y'} \nu(t - t', x - x', y - y')\, dN(t', x', y'). \tag{4.2}$$

The first term μ represents the intensity of background events, independent of previous events, and the second term models the intensity of offspring events triggered by either the background events or other offspring.

One method for estimating μ is to assume $\mu = \overline{\mu} \cdot u$ where u is a probability density estimated from the spatial coordinates of the data. The time interval of the data can then

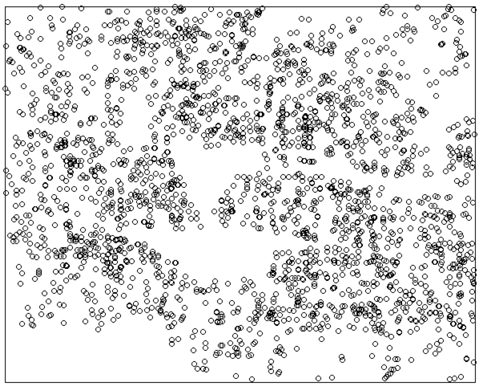

Figure 3. Spatial distribution of residential burglaries occurring during 2004 in an 18 km×18 km region of the San Fernando Valley in Los Angeles.

be split into two intervals and one can maximize the log-likelihood function over the more recent data while constructing the density from the older data (Peng, Schoenberg, and Woods 2005). We apply this methodology to the burglary data, fitting u to the data from 2004 and maximizing the log-likelihood function over the 2005 data in order to choose an optimal smoothing parameter. Similarly to the previous example, we take $\epsilon = 10^{-3}$ in order to prevent a small number of isolated points from dominating the likelihood function. We point out that μ is typically estimated concurrently with the kernel ν; however, for simplicity we take $\nu = 0$ in this example. In Figure 4, we display contour plots of the Split Bregman estimate of the density of burglaries, a Gaussian kernel estimate, and average shifted histogram estimate for comparison.

Similar types of self-exciting models are used in seismology to describe the distribution of earthquake aftershocks (Ogata 1998) and standard methods for estimating the background intensity μ include spline, kernel smoothing, and Voronoi estimation (Silverman 1986; Ogata and Katsura 1988; Okabe et al. 2000). In the case of crime, however, sharp gradients in μ need to be accounted for due to the specific structure of cities. For example, in Figure 3 sharp boundaries exist between residential areas (where the points are distributed), commercial areas (upper middle region), and public parks (lower middle and bottom regions). While a method such as kernel smoothing may provide a good fit according to a measure like the Akaike Information Criterion, if events are distributed in unrealistic regions then forecasts based on the method may be met with skepticism by police and other practitioners. Thus we believe TV-based MPLE may be a good alternative to standard methods for the purpose of point process crime modeling, as it can be seen in Figure 4 that the method is able to resolve the sharp boundaries in the crime data, while away from the boundaries undersmoothing is kept relatively low. The Gaussian kernel density estimate and average shifted histogram, by comparison, oversmooth the density into the middle and lower regions where crime cannot occur.

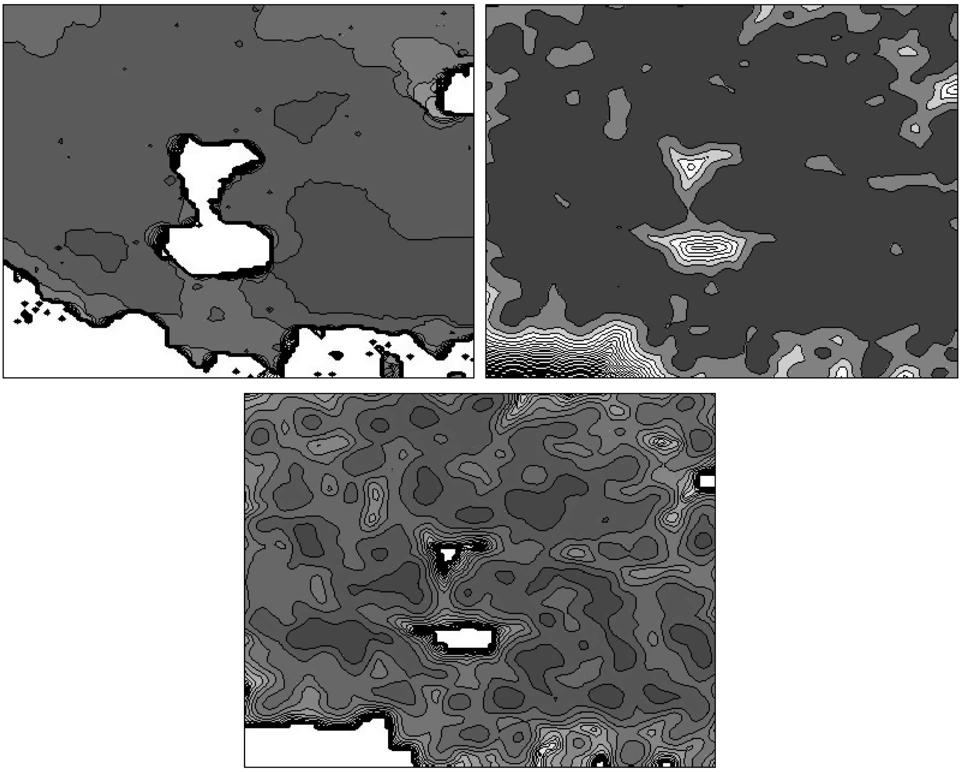

Figure 4. Top left: A contour plot of the Split Bregman estimate of the spatial density of residential burglaries on a logarithmic scale. Top right: A contour plot of a Gaussian kernel estimate of the spatial density of residential burglaries on a logarithmic scale. Bottom: A contour plot of an average shifted histogram estimate of the spatial density of residential burglaries on a logarithmic scale.

5. CONCLUDING REMARKS

We presented an efficient computational methodology for Maximum Penalized Likelihood Estimation when the penalty is chosen as the Total Variation of the estimate, with applications to the estimation of nonsmooth 2D probability densities. The method allows for the fast approximation of TV-based MPLE on standard 2D grids, even for large sample sizes and computationally intensive parameter selection procedures.

In the future it may be of interest to consider other regularizations for MPLE along the lines of the Total Variation penalty. A number of extensions to (1.3) have been proposed in the image processing literature, for example, choosing the smoothing parameter to be spatially dependent. Whereas variable bandwidth kernel estimates can improve upon fixed bandwidth estimates, this may also be the case for TV-based MPLE.

Other classical statistical problems, such as multivariate regression, may also be approached using techniques similar to those developed in this article. In cases where a given dataset exhibits sharp peaks or jump discontinuities, Maximum Penalized Likelihood Regression with a Total Variation penalty may outperform standard regression techniques.

Last, we believe that TV-based MPLE will find wide application in the point process modeling of crime. Further studies in this area will focus on the incorporation of background intensity estimates obtained through TV-based MPLE into the self-exciting point process framework.

SUPPLEMENTAL MATERIALS

Matlab TVMPLE routine: The Matlab (mex) code "TVMPLE.c" and datasets used in the first example are contained in the zip file TVMPLE.zip available online. Please refer to the readme file for a description of how to compile and run the code and of the datasets contained in the zip file. (TVMPLE.zip, zip archive)

ACKNOWLEDGMENTS

This work was supported in part by ARO MURI grant 50363-MA-MUR, NSF grant ACI-0321917, ONR grant N000140810363, ONR grant N000141010221, NSF grant BCS-0527388, and the Department of Defense. The authors also thank Frederic Schoenberg, Jeffrey Brantingham, and George Tita for helpful discussions, and George Tita and the LAPD for providing the residential burglary data used in this study.

[Received April 2009. Revised March 2010.]

REFERENCES

Acar, R., and Vogel, C. R. (1994), "Analysis of Bounded Variation Penalty Methods for Ill-Posed Problems," *Inverse Problems*, 10, 1217–1229. [485]

Chang, T. C., He, L., and Fang, T. (2006), "MR Image Reconstruction From Sparse Radial Samples by Using Bregman Iteration," in *Proceedings of the 13th Annual Meeting of ISMRM*, Seattle, 696. [482]

Davies, P. L., and Kovac, A. (2004), "Densities, Spectral Densities and Modality," *The Annals of Statistics*, 32, 1093–1136. [480]

Eggermont, P. P. B., and LaRiccia, V. N. (2001), *Maximum Penalized Likelihood Estimation: Volume I: Density Estimation*, New York: Springer. [480]

Farrell, G., and Pease, K. (eds.) (2007), *Repeat Victimization*, New York: Criminal Justice Press. [486]

Goldstein, T., and Osher, S. (2009), "The Split Bregman Method for L1 Regularized Problems," *SIAM Journal on Imaging Sciences*, 2 (2), 323–343. [480-483]

Johnson, S. D., Bernasco, W., Bowers, K. J., Elffers, H., Ratcliffe, J., Rengert, G., and Townsley, M. (2007), "Space-Time Patterns of Risk: A Cross National Assessment of Residential Burglary Victimization," *Journal of Quantitative Criminology* 23, 201–219. [486]

Koenker, R., and Mizera, I. (2006), "Density Estimation by Total Variation Regularization," preprint, University of Alberta. [480]

Kooperberg, C., and Stone, C. J. (2002), "Logspline Density Estimation With Free Knots," *Computational Statistics and Data Analysis*, 12, 327–347. [480]

Le, T., Chartrand, R., and Asake, T. J. (2007), "A Variational Approach to Reconstructing Images Corrupted by Poisson Noise," *Journal of Mathematical Imaging and Vision*, 27 (3), 257–263. [480]

Mohler, G., Short, M., Brantingham, P., Schoenberg, F., and Tita, G. (2008), "Self-Exciting Point Process Modeling of Crime," in review. [487]

Ogata, Y. (1998), “Space-Time Point Process Models for Earthquake Occurrences,” *Annals of the Institute of Statistical Mathematics*, 50 (2), 379–402. [488]

Ogata, Y., and Katsura, K. (1988), “Likelihood Analysis of Spatial Inhomogeneity for Marked Point Patterns,” *Annals of the Institute of Statistical Mathematics*, 40, 20–39. [488]

Okabe, A., Boots, B., Sugihara, K., and Chiu, S. (2000), *Spatial Tessellations* (2nd ed.), Chichester: Wiley. [488]

Osher, S., Burger, M., Goldfarb, D., Xu, J., and Yin, W. (2005), “An Iterative Regularization Method for Total Variation-Based Image Restoration,” *Multiscale Modeling and Simulation*, 4, 460–489. [482,483]

Peng, R. D., Schoenberg, F. P., and Woods, J. (2005), “A Space-Time Conditional Intensity Model for Evaluating a Wildfire Hazard Index,” *Journal of the American Statistical Association*, 100 (469), 26–35. [488]

Rudin, L., Osher, S., and Fatemi, E. (1992), “Nonlinear Total Variation Based Noise Removal Algorithms,” *Physica D*, 60, 259–268. [480,485]

Sardy, S., and Tseng, P. (2006), “Density Estimation by Total Variation Penalized Likelihood Driven by the L1 Information Criterion,” preprint, University of Washington. [480,485]

Scott, D. W. (1992), *Multivariate Density Estimation*, New York: Wiley. [486]

Sheather, S. J., and Jones, M. C. (1991), “A Reliable Data-Based Bandwidth Selection Method for Kernel Density Estimation,” *Journal of the Royal Statistical Society, Ser. B*, 53, 683–690. [480]

Short, M. B., D’Orsogna, M. R., Brantingham, P. J., and Tita, G. E. (2009), “Measuring and Modeling Repeat and Near-Repeat Burglary Effects,” *Journal of Quantitative Criminology*, 25 (3), 325–339. [486]

Silverman, B. W. (1986), *Density Estimation for Statistics and Data Analysis*, London: Chapman & Hall. [486, 488]

Yin, W., Osher, S., Goldfarb, D., and Darbon, J. (2008a), “Bregman Iterative Algorithms for L1-Minimization With Applications to Compressed Sensing,” *SIAM Journal on Imaging Science*, 1, 142–168. [482]

Yin, W., Wang, Y., Yang, J., and Zhang, Y. (2008b), “A New Alternating Minimization Algorithm for Total Variation Image Reconstruction,” *SIAM Journal on Imaging Science*, 1 (3), 248–272. [481,482]