PROCEEDINGS OF THE ROYAL SOCIETY A

rspa.royalsocietypublishing.org

Research

Article submitted to journal

Subject Areas:

Applied mathematics, Computational mathematics, Differential equations, Graph Theory, Mathematical modelling, Statistics

Keywords:

Crime Hotspots, Density Estimation, Graph Laplacian, Maximum Penalized Likelihood Estimation, Nonlocal means, Nyström's extension

Author for correspondence:

J. T. Woodworth

e-mail: jwoodworth@math.ucla.edu

Nonlocal Crime Density Estimation Incorporating Housing Information

J. T. Woodworth [1], G. O. Mohler [2], A. L. Bertozzi[1] and P. J. Brantingham[3]

[1]Department of Mathematics, University of California, Los Angeles

[2]Department of Mathematics & Computer Science, Santa Clara University

[3]Department of Anthropology, University of California, Los Angeles

Given a discrete sample of event locations, we wish to produce a probability density that models the relative probability of events occurring in a spatial domain. Standard density estimation techniques do not incorporate priors informed by spatial data. Such methods can result in assigning significant positive probability to locations where events cannot realistically occur. In particular, when modeling residential burglaries, standard density estimation can predict residential burglaries occurring where there are no residences. Incorporating the spatial data can inform the valid region for the density. When modeling very few events, additional priors can help to correctly fill in the gaps. Learning and enforcing correlation between spatial data and event data can yield better estimates from fewer events. We propose a nonlocal version of Maximum Penalized Likelihood Estimation based on the H^1 Sobolev seminorm regularizer that computes nonlocal weights from spatial data to obtain more spatially accurate density estimates. We evaluate this method in application to a residential burglary data set from San Fernando Valley with the nonlocal weights informed by housing data or a satellite image.

1. Introduction

In real-world applications, satellite images, housing data, census data, and other types of geographical data become highly relevant for modeling the probability of a certain type of event. The methodology presented here provides a general framework paired with fast algorithms for incorporating external information in density estimation computations.

Royal Society Publishing
Informing the science of the future

In density estimation, one is given a discrete sample of event locations, drawn from some unknown density u on the spatial domain, and tries to approximately recover u [1]. Relating the events to the additional data allows one to search over a smaller space of densities, which can yield more accurate results with fewer events. We refer to the additional data source as the function $g(x)$ defined over the spatial domain Ω. We may typically assume two things about the relationship between g and u : 1) g informs the support of u via $g(x)=0 \Rightarrow u(x)=0$ and 2) u varies smoothly with g in a nonlocal way (explained below). This method allows the additional information in g to significantly improve the recovery of u.

(a) Maximum Penalized Likelihood Estimation

Although there are other classes of methods in the density estimation literature which are quite popular (such as average shifted histogram and kernel density estimation [2]), in this work we shall focus on Maximum Penalized Likelihood Estimation (MPLE). MPLE provides a general framework for finding an approximate density from sampled events. The likelihood of events occurring at the locations $\{x_i\}_{i=1}^n$ according to a proposed probability u is the product of the probability evaluated at each of those locations:

$$\mathcal{L}\left(u, \{x_i\}_{i=1}^n\right) = \prod_{i=1}^n u(x_i).$$

MPLE approximates u as the maximizer of a log-likelihood term combined with a penalty term, typically enforcing smoothness [3],

$$\hat{u} = \underset{u\geq 0, \int_\Omega u dx=1}{\operatorname{argmax}} \sum_{i=1}^n \log(u(x_i)) - P(u).$$

Without some kind of penalty term, the solution is just a weighted sum of Dirac deltas located at the training samples. Typical choices of $P(u)$ include the TV-norm, $P(u)=\lambda\int_\Omega |\nabla u| dx$, and the H^1 Sobolev seminorm $P(u)=\frac{\lambda}{2}\int_\Omega |\nabla u|^2 dx$. λ is the parameter which controls the amount of regularization. This is typically chosen via cross-validation, when it is computationally feasible.

(b) MPLE applied to Crime

In recent years there have been several studies on the application of MPLE to crime data [4–6]. This body of work emphasizes the fact that crime density is not spatially smooth. Rather, it can have boundaries corresponding to the local geography.

Mohler et al. and Kostic et al. model this by choosing penalty functions that are edge-preserving, TV and Ginzburg-Landau respectively [4,6]. Smith et al. more closely follows the idea presented here. That work introduces a modified H^1 MPLE, which based the penalty term on an additional component of the data [5]. The method assumes that the the valid region of the probability density estimate is known a priori. In their application to residential burglary the valid region was the approximate support of the housing density in the region. If we denote the valid region by D, then the modified penalty term is just a standard H^1 MPLE with a factor z_ϵ^2 in the integral, where z_ϵ is a smooth Ambrosio-Tortorelli approximation of $(1-\delta(\partial D))$:

$$\hat{u} = \underset{u\geq 0, \int_\Omega u=1}{\operatorname{argmin}} \frac{1}{2}\int_\Omega z_\epsilon^2 |\nabla u|^2 dx - \mu \sum_{i=1}^n \log(u(x_i)),$$

$$z_\epsilon(x) = \begin{cases} 1 & \text{if } d(x,\partial D) > \epsilon, \\ 0 & \text{if } x \in \partial D. \end{cases}$$

(c) Graph-based methods

In spectral graph theory, data is represented as nodes of a weighted graph, where the weight on each edge indicates the similarity between the two nodes. Such data structures have been very successfully applied to data clustering problems and image segmentation [7–9]. The standard theory behind this is described in [10,11] and a tutorial on spectral clustering is given in [12]. A theory of nonlocal calculus was developed first by Zhou and Schölkopf in 2004 [13] and put in a continuous setting by Gilboa and Osher in 2008 [14]. Such methods were originally used for image denoising [14,15], but the general framework led to methods for inpainting, reconstruction, and deblurring [16–20]. Compared with local methods, nonlocal methods are generally better able to handle images with patterns and texture. Further, by choosing an appropriate affinity function, the methods can be made suitable for a wide variety of different of data sets : not just images.

In this article we present nonlocal H^1 MPLE (NL H^1 MPLE), which modifies the standard H^1 MPLE energy to account for spatial inhomogeneities, but unlike Smith et al. [5], we do so in a nonlocal way, which has the benefit of leveraging recent fast algorithms and the potential to generalize to other applications.

The organization of this article is as follows: In Sec. 2, we introduce the NL H^1 MPLE method and review the nonlocal calculus and numerical methods on which it is based. In Sec. 3 we demonstrate the advantages of NL H^1 MPLE by comparing it with standard H^1 MPLE when applied to modeling residential burglary. In Sec. 4 we summarize our conclusions and discuss directions for future research.

2. Nonlocal Crime Density Estimation

We propose replacing the H^1 seminorm regularizer of H^1 MPLE with a linear combination of an H^1 regularizer and a nonlocal smoothing term $\iint_{\Omega\times\Omega}\left(\nabla_{w,s}u(x,u)\right)^2 dxdy$ where $\nabla_{w,s}$ denotes the nonlocal symmetric-normalized gradient depending on an affinity function w derived from the spatial data, g. More details are found in Sec. (b). The energy we optimize is thus

$$\hat{u} = \underset{u\geq 0,\int_\Omega u=1}{\operatorname{argmax}} \sum_{i=1}^{n} \log\left(u(x_i)\right) - \alpha \iint_{\Omega\times\Omega} \left(\nabla_{w,s}u(x,u)\right)^2 dxdy - \frac{\beta}{2}\int_\Omega |\nabla u(x)|^2 dx. \tag{2.1}$$

The nonlocal term in equation (2.1) is tolerant of sharp changes in the probability density estimate, as long as they coincide with sharp nonlocal changes in the spatial data. This follows from the definitions presented in the following sections and is discussed in more detail in the appendix.

(a) Nonlocal means

Nonlocal means was originally developed for the application of image denoising, but can also be interpreted as an affinity function. The formula for the nonlocal means affinity, $w_{\mathbf{Im}}$, is given by [15]

$$w_{\mathbf{Im}}(x,y) = \exp\left(-\frac{\left(K_r * |\mathbf{Im}(x+\cdot) - \mathbf{Im}(y+\cdot)|^2\right)(0)}{\sigma^2}\right). \tag{2.2}$$

Here **Im** is the image the nonlocal means weights are based on, K_r is a nonnegative weight kernel of size $(2r+1)\times(2r+1)$, and σ is a scaling parameter. This function measures similarity between two pixels based on a weighted ℓ_2 difference between patches surrounding them in the image. In our experiments, the image **Im** is either a housing image or a satellite image. In practical settings, computing and storing all function values of w is a very computationally intensive task, so we use the fast approximation : Nyström's extension (see Sec. 2(d)).

(b) Nonlocal calculus and graphs

Nonlocal calculus was introduced in its discrete form by Zhou and Schölkopf [13] and put in a continuous framework by Gilboa and Osher [14]. In these definitions, $w(x,y)$ is a general nonnegative symmetric affinity function which generally measures similarity between the points x and y.

Let $\Omega\subset\mathbb{R}^n$, and $u(x)$ be a function $u:\Omega\to\mathbb{R}$. Then the nonlocal gradient of u at the point $x\in\Omega$ in the direction of $y\in\Omega$ is given by

$$\left(\nabla_w u\right)(x,y) = \left(u(y)-u(x)\right)\sqrt{w(x,y)}.$$

This suggests an analogous generalization of divergence, which in turn leads to the following definition of the nonlocal Laplacian.

$$\Delta_w u(x) = \int_\Omega (u(y) - u(x))\, w(x,y) dy \tag{2.3}$$

Now let $\{p_i\}_{i=1}^n$ be a discrete subset of Ω and let $w_{ij} = w(p_i, p_j)$ if $i \neq j$ and $w_{ii} = 0$. We then let $\{p_i\}_{i=1}^n$ be vertices and w_{ij} the edge weights on a weighted graph. Let $d_i = \sum_{j=1}^n w_{ij}$ be the weighted degree of the ith node. Then the graph Laplacian applied to the function on the graph, u, is given by Lu where

$$L_{ij} = \begin{cases} d_i & \text{if } i = j \\ -w_{ij} & \text{otherwise} \end{cases}, \quad \text{and so } (Lu)_i = \sum_{j=1}^n \left(u_i - u_j\right) w_{ij}.$$

To keep the spectrum of the graph Laplacian in a fixed range as the the number of samples in increased and thus to guarantee consistency, we must normalize the graph Laplacian. See Bertozzi and Flenner 2012 [21] for a more in depth discussion of this. We use the symmetric normalization.

$$L_{sym} := D^{-1/2} L D^{-1/2}, \quad D_{ij} = \begin{cases} d_i & \text{if } i = j \\ 0 & \text{otherwise} \end{cases}$$

Because we express our energy as applied to functions over continuous domains, we also introduce the following notation for the symmetric-normalized nonlocal gradient.

$$\nabla_{w,s} u(x,y) := \frac{\nabla_w u(x,y)}{\left(\int_\Omega w(x,z)dz \int_\Omega w(y,z)dz\right)^{1/4}}$$

(c) Numerical optimization

We must numerically find an approximate solution. The unconstrained energy has gradient flow

$$u_t = \alpha \Delta_{w,s} u + \beta \Delta u + \frac{1}{u} \sum_{i=1}^n \delta(x - x_i).$$

We evolve this equation, projecting onto the space of probability densities after each step. We discretize the equation as

$$\frac{u^{k+1} - u^k}{\delta t} = -\alpha L_{sym} u^{k+1} + \beta \Delta_h u^{k+1} + \frac{1}{u^k} \sum_{i=1}^n \delta(x - x_i).$$

Here Δ_h denotes the discrete Laplacian from the 5-point finite difference stencil with mesh size $h = 1$. Solving for u^{k+1} yields

$$u^{k+1} = (I + \alpha \delta t L_{sym} - \beta \delta t \Delta_h)^{-1} \left(\frac{\delta t}{u^k} \sum_{i=1}^n \delta(x - x_i) + u^k \right).$$

To approximate this, we use a split-time method

$$u^{k+1/2} = \left(I + \alpha \frac{\delta t}{2} L_{sym} \right)^{-1} \left(\frac{\delta t}{u^k} \sum_{i=1}^n \delta(x - x_i) + u^k \right),$$

$$u^{k+1} = \left(I - \beta \frac{\delta t}{2} \Delta_h \right)^{-1} \left(\frac{\delta t}{u^{k+1/2}} \sum_{i=1}^n \delta(x - x_i) + u^{k+1/2} \right).$$

To apply these operators, we use a spectral method. This has two advantages over forming and multiplying the matrices. First, we can approximate the projection onto the constraint by using the spectral decomposition of the discrete Laplacian (shown in Table 1). Second, the computation required to form and apply the entire symmetric graph Laplacian is too intensive. Fortunately, we can apply Nyström's extension (discussed in Sec. (d)), which is a popular method for approximating a portion of the eigenvectors and

Nyström ($\mathbf{Im}_g$) $\rightarrow \Phi, \Lambda : L_{sym} \approx \Phi \Lambda \Phi^T$. Initialize $u^0 \equiv 1/\lvert\Omega\rvert$, $succDiff = \infty$, $k = 0$. while $succDiff > 10^{-7}$ and $k < maxSteps = 800$ • $k = k + 1$ • $\vec{b} = \Phi^T \left[u^{k-1} + \frac{\delta t}{u^{k-1}} \sum_{i=1}^n \delta(x - x_i) \right]$ • $a_i = \frac{b_i}{1 + \alpha \frac{\delta t}{2} \lambda_i}$ • $\vec{u}^{k-1/2} = \Phi \vec{a}$ • $\vec{b} = \text{fft2} \left[u^{k-1/2} + \frac{\delta t}{u^{k-1/2}} \sum_{i=1}^n \delta(x - x_i) \right]$ • $a_i = \frac{b_i}{1 + 2\beta \delta t \pi^2 (m^2 + n^2)}$, $i \sim (m, n)^{th}$ Fourier mode, $a_1 = 1$ (guarantees integral 1 constraint) • $\vec{u}^k = \text{ifft2}\,(\vec{a})$ • $\vec{u}^k = \max\left(\vec{u}^k, 0\right)$ • $succDiff = \lVert u^k - u^{k-1} \rVert_2^2 / \lVert u^k \rVert_2^2$

Table 1: Nonlocal H^1 MPLE Algorithm

eigenvalues which approximate the operator well. To project onto the eigenvectors of Δ_h we apply the 2D Fast Fourier Transform.

In both the case of applying $(I + \alpha \frac{\delta t}{2} L_{sym})^{-1}$ and $(I - \beta \frac{\delta t}{2} \Delta_h)^{-1}$ we are applying operators of the form $(I + \delta t P)^{-1}$ where P is symmetric and positive semidefinite. In general, if P has spectral decomposition $P = \Phi \Lambda \Phi^T$ then we apply $(I + \delta t P)^{-1}$ to $\vec{w}$ by first projecting onto the eigenvectors : $\vec{a} = \Phi^T \vec{w}$, updating the coefficients $\tilde{a}_m = a_m/(1 + \delta t \lambda_m)$, and finally transforming back to the standard basis : $(I + \delta t P)^{-1} \vec{w} = \Phi \vec{\tilde{a}}$. We summarize the steps of our algorithm in Table 1.

(d) Nyström's extension

To apply the spectral method described in the previous section we need to approximate the eigenvectors and eigenvalues of the symmetric graph Laplacian. Here we present the Nyström's extension method and refer the reader to [16,21,22] for further justification and analysis. Nyström's extension is a technique for performing matrix completion well-known within the spectral graph theory community. In this setting, Nyström's extension is applied to the normalized affinity matrix $D^{-1/2} W D^{-1/2}$ where $W = (w_{ij})$. For simplicity, in this discussion we disregard the normalization factors. If we let N denote the set of nodes in our complete weighted graph, then we take X to be a small random sample from N and Y its complement. Then up to a permutation of the nodes we can write the affinity matrix as

$$W = \begin{pmatrix} W_{XX} & W_{XY} \\ W_{YX} & W_{YY} \end{pmatrix}.$$

The matrix W_{XY} contains the weights between nodes in X and nodes in Y and W_{XX}, W_{YX}, W_{YY} are defined similarly. Nyström's extension approximates the eigenvalues and eigenvectors of the affinity matrix by implicitly making the approximation:

$$W \approx \hat{W} = \begin{pmatrix} W_{XX} \\ W_{YX} \end{pmatrix} W_{XX}^{-1} \begin{pmatrix} W_{XX} & W_{XY} \end{pmatrix}.$$

This approximates $W_{YY} \approx W_{YX} W_{XX}^{-1} W_{XY}$. The error due to this approximation is determined by how well the rows of W_{XY} span the rows of W_{YY}. If the affinity matrix W is positive semidefinite then we can write it as a matrix transpose times itself : $W = V^T V$. In [23] the authors show that the Nyström extension thus approximates the unknown part of V (corresponding to W_{YY}) by orthogonally projecting it onto the range of the known part (corresponding to W_{XY}). In this setting it is clear that as the size of X grows, the approximation improves. Further our random choice of X is likely to yield W_{XY} full-rank if the rank of the rank of W is large.

In practice, one uses a diagonal decomposition of this formula (to avoid forming and applying the full matrix). It follows from analysis discussed in [22] that if W_{XX} is positive definite, the diagonal decomposition of the approximation is given by $\hat{W} = V \Lambda_S V^T$, where

$$S = W_{XX} + W_{XX}^{-1/2} W_{XY} W_{YX} W_{XX}^{-1/2}$$

which has diagonal decomposition $S = U_S \Lambda_S U_S^T$, and

$$V = \begin{bmatrix} W_{XX} \\ W_{YX} \end{bmatrix} W_{XX}^{-1/2} U_S \Lambda_S^{-1/2}.$$

For more detailed analysis and an explanation of how to incorporate the normalization factors, see [16, 21,22]. Ultimately, Nyström's extension approximates $|X|$ eigenvectors and eigenvalues of an $|N| \times |N|$ matrix, while only performing computations on matrices no bigger than $|N| \times |X|$.

(e) Cross-validation

Because our method consists primarily of simple coefficient updates after mapping to different eigenspaces, it is fast relative to methods with similar goals ([5] for instance). This speed increase allows us to perform V-fold cross-validation, which requires many evaluations of a density estimation method. Cross-validation is a methodology for choosing the smoothing parameter λ which yields probability densities that are predictive of small amounts of missing data [24]. The objective function is an application of Kullback-Leibler divergence [25], which results in maximizing a log-likelihood expression.

3. Numerical experiments

In this section, we demonstrate the advantage NL H^1 MPLE method over standard H^1 MPLE by evaluating its performance on residential burglary data from San Fernando Valley in Los Angeles, California, using of corresponding housing data and a satellite image to inform the nonlocal weights.

(a) Residential burglary

We perform experiments on residential burglary data from San Fernando Valley in 2005-2013, getting substantially different results than those shown in [4–6]. In Fig. 1 we show the data used (locations of residential burglaries in Fig. 1(a), housing in Fig. 1(b), satellite image in Fig. 1(c)), H^1 MPLE (Fig. 1(d)), housing-based NL H^1 MPLE (Fig. 1(c)), and satellite-based NL H^1 MPLE (Fig. 1(d)) density estimates on increasing subsets of data from 2005-2008. To evaluate performance, we compute the log-likelihood of each density on the residential burglaries from 2009-2013 (shown in Table 2).

As one would predict, the locations of residential burglaries in Fig. 1(a) are primarily restricted to the support of the housing density image Fig. 1(b). There are some locations in the burglary data set that correspond to locations with no residences (4,173 events out of 23,725 total), which we attribute to imprecision in the burglary data. Most such misplaced events occur on streets, suggesting that the actual event took place at a residence facing that street. Because of this inconsistency between the data sets, for the experiments which use the housing data, we adjust the residential burglary data for training and testing (for both H^1 and NL H^1), moving each event to the nearest house if it is within 2 pixels, and dropping the event otherwise. This results in 603 dropped events. For the experiments which do not use housing data, we work with the raw burglary data for training and testing.

We implement H^1 MPLE by applying our algorithm, described in Table 1 with $\alpha = 0$ and $\Phi = Id$. We choose the value of the regularization parameter β for each training data set by performing 10-fold log-likelihood cross-validation, searching over $\beta = $ `[0,10.^(-2:8)]`. We apply H^1 MPLE to both the raw and corrected burglary data.

For housing-based NL H^1 MPLE , we perform Nyström's extension with nonlocal means applied to g, the housing density image shown in Fig. 1(c). We use 400 random samples for Nyström's extension. We use the first 300 eigenvectors and eigenvalues in our computations. The nonlocal means weights are based on differences between patches of size 11×11 and $\sigma = 1 \cdot std(g)$, the standard deviation of the housing image. The weight kernel K_r, $r = 5$, is given as follows.

$$K_r(1+r+i, 1+r+j) = \frac{1}{r} \sum_{d=\max(|i|,|j|,1)}^{r} \frac{1}{(2d+1)^2}, \quad i,j = -r, \ldots, r$$

To choose the regularization parameters α, β, we perform 10-fold log-likelihood cross-validation, searching over $\alpha = $ `[0,10.^(-2:12)]`, $\beta = $ `[0,10.^(-2:8)]`. We apply housing NL H^1 MPLE to the corrected burglary data.

For satellite-based NL H^1 MPLE, we perform Nyström's extension with nonlocal means applied to g, the Google Maps image shown in Fig. 1(c). In applying nonlocal means to a color image, we interpret the image as a vector valued function with 3 components (one for each color channel) and so in equation (2.2) the expression $|\mathbf{Im}(x+\cdot) - \mathbf{Im}(y+\cdot)|^2$ is size $(2r+1) \times (2r+1) \times 3$. We use 800 random samples for Nyström's extension. We use the first 600 eigenvectors and eigenvalues in our computations. The nonlocal means weights are based on differences between patches of size 11×11 and $\sigma = 1 \cdot std(g)$, the standard deviation of the Google Maps image. The weight kernel is as in the previous case, but repeated on each color channel. To choose the regularization parameters α, β for each training set, we perform 10-fold log-likelihood cross-validation, searching over $\alpha = $ `[0,10.^(-2:12)]`, $\beta = $ `[0,10.^(-2:8)]`. We apply satellite NL H^1 MPLE to the raw burglary data.

The H^1 MPLE results transition from a completely smooth uniform density to a probability density with more apparent structure as the amount of training data increases. The NL H^1 MPLE housing and satellite results exhibit a similar trend, but are able to better approximate the correct support of the density with many fewer data points. The measurable benefit of nonlocal smoothing is shown by the log-likelihood values in Table 2. NL H^1 generally gets higher log-likelihood than H^1. This means the densities estimated by housing NL H^1 on corrected 2005-2008 data are more congruous with the corrected 2009-2013 data than the H^1 densities, and the densities estimated by satellite NL H^1 on raw 2005-2008 data are more congruous with the raw 2009-2013 data than the H^1 densities.

The added complexity of our algorithm results in an increase in run time from the standard H^1 MPLE, but the difference is not too substantial. We compare run times on a laptop with one Intel Core i7 processor that has two cores with processor speed 2.67GHz and 4GB of memory. The run time for Nyström applied to the housing image is typically about 17 seconds. The run time for Nyström applied to the satellite image is typically about 36 seconds. For cross-validation purposes, Nyström can be run once outside of the loop and the results used for all combinations of data sets and parameters. The run time for H^1 MPLE with parameters as chosen by cross-validation on the residential burglaries from 2005-2008 is typically about half a second. The run time for housing NL H^1 MPLE with parameters as chosen by cross-validation on the the residential burglaries from 2005-2008 is typically about 2.3 seconds. The run time for satellite NL H^1 MPLE with parameters as chosen by cross-validation on the the residential burglaries from 2005-2008 is typically about 1.5 seconds. The cross-validation run times depend on what range of parameters are being tested, but can easily be run in parallel across several computing nodes.

(b) Synthetic Density

To test the method's ability to recover a density, we must draw events from a known density and try to recover it. Because the method assumes a relationship between the spatial data g and the density u, we generate a synthetic density which is closely related to the housing data, shown in the bottom left of Fig. 2. This density is given by taking a random combination of the first 5 approximated eigenvectors of the graph

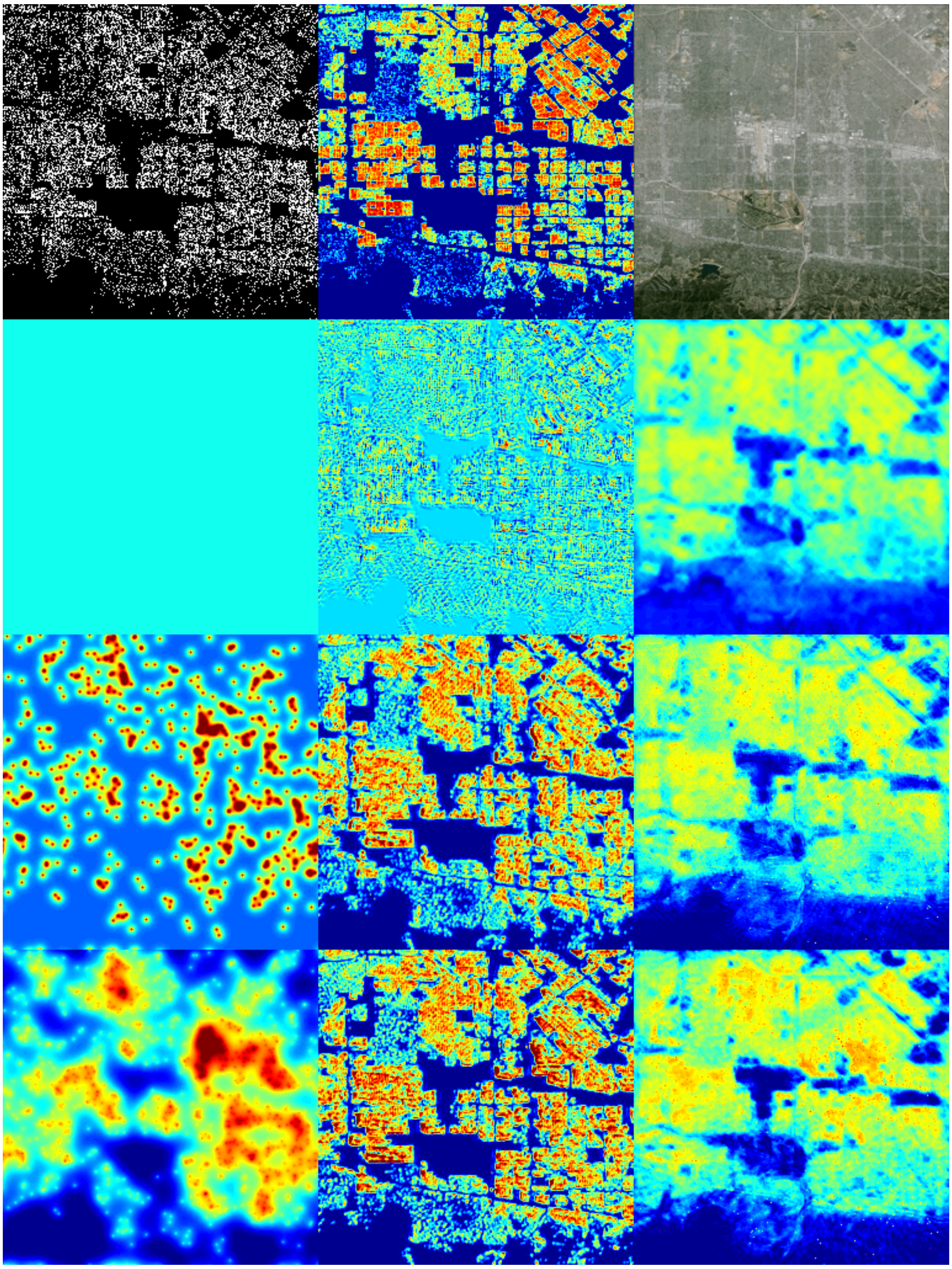

Figure 1:
Top row: data
(a) 2005-2013 Residential burglaries in San Fernando Valley (from LAPD)
(b) San Fernando Valley $\log(\min(\#$ housing units$, 7) + 1)$ (from LA County Tax Assessor)
(c) Satellite image of San Fernando Valley (from Google Maps)
Bottom three rows : MPLE of 50, 500, and 1000 random samples from '08 residential burglaries
(d) Column 1 : H^1 MPLE
(e) Column 2 : Housing NL H^1 MPLE
(f) Column 3 : Satellite NL H^1 MPLE

Training Data Set (corrected)	scaled Histogram	H^1	Housing NL H^1
50 random from 2008	-3.6039×10^{05}	**-1.3386** $\times 10^{05}$	-1.3396×10^{05}
100 random from 2008	-3.5991×10^{05}	**-1.3369** $\times 10^{05}$	-1.3369×10^{05}
500 random from 2008	-3.5197×10^{05}	-1.3282×10^{05}	**-1.3004** $\times 10^{05}$
1000 random from 2008	-3.4350×10^{05}	-1.3246×10^{05}	**-1.2953** $\times 10^{05}$
2008	-3.1905×10^{05}	-1.3189×10^{05}	**-1.2888** $\times 10^{05}$
2007-2008	-2.9846×10^{05}	-1.3174×10^{05}	**-1.2850** $\times 10^{05}$
2006-2008	-2.8152×10^{05}	-1.3136×10^{05}	**-1.2815** $\times 10^{05}$
2005-2008	-2.6847×10^{05}	-1.3121×10^{05}	**-1.2774** $\times 10^{05}$
Traing Data Set (raw)	scaled Histogram	H^1	Satellite NL H^1
50 random from 2008	-3.6959×10^5	-1.3733×10^5	**-1.3553** $\times 10^5$
100 random from 2008	-3.6822×10^5	-1.3732×10^5	**-1.3553** $\times 10^5$
500 random from 2008	-3.6342×10^5	-1.3583×10^5	**-1.3524** $\times 10^5$
1000 random from 2008	-3.5733×10^5	-1.3598×10^5	**-1.3525** $\times 10^5$
2008	-3.3313×10^5	-1.3535×10^5	**-1.3494** $\times 10^5$
2007-2008	-3.1326×10^5	-1.3525×10^5	**-1.3482** $\times 10^5$
2006-2008	-2.9630×10^5	-1.3496×10^5	**-1.3449** $\times 10^5$
2005-2008	-2.8334×10^5	-1.3488×10^5	**-1.3431** $\times 10^5$

Table 2: Log-likelihood of densities on residential burglaries from 2009-2013 (corrected & raw)

Laplacian (with weights based on the housing image) and then shifting and normalizing the result to yield a probability density. The coefficients are chosen uniformly at random in $[0,1]$ and the nonlocal weights are based on the housing data as they were in the previous section. This randomly generated density was chosen over others because it looks like a potential probability density for residential burglary.

We sample events according to this density by generating numbers uniformly at random in $[0,1]$ and inverting the cumulative distribution function associated with the density. In the top row of Fig. 2 we show the H^1 MPLE result on the 400 events ($\beta = 5 \times 10^4$), the housing NL H^1 MPLE result on the 400 events ($\alpha = 100, \beta = 0$), and the NL H^1 MPLE result on 400 events restricted to the first 5 eigenvectors. In the bottom row of Fig. 2 we show the synthetic density, the H^1 MPLE result on the 4,000 events ($\beta = 10^5$), the housing NL H^1 MPLE result on the 4,000 events ($\alpha = 10^8, \beta = 0$), and the NL H^1 MPLE result on 4,000 events restricted to the first 5 eigenvectors. In all cases, smoothing parameters were chosen to minimize mean absolute error of the probability density. The NL H^1 results and the restricted NL H^1 results do a substantially better job at recovering the probability density than H^1 MPLE. This suggests that if the correct density is well approximated by a combination of eigenvectors of the graph Laplacian, enforcing nonlocal smoothness can substantially improve recovery of the density.

4. Conclusions and Future work

In this paper we have looked at the problem of obtaining spatially accurate probability density estimates. The need for new approaches is demonstrated by the inadequate performance of standard techniques such as H^1 MPLE.

Our proposed solution accomplishes this by incorporating a nonlocal regularity term based on the H^1 regularizer and nonlocal means which fuses geographical information into the density estimate. Our experiments with the San Fernando Valley residential burglary data set demonstrate that this method does yield a probability density estimate with the correct support which also gives favorable log-likelihood results. Further, our results based on the Google Maps image suggest we can apply NL H^1 MPLE to a wide variety of geographic regions without obtaining specialized geographic data.

There are several others aspects of this and related problems to explore. In general, testing the method on other datasets would be interesting. This may present the added challenge of dealing with other types of geographical information since high-resolution housing density data may not be readily available. In

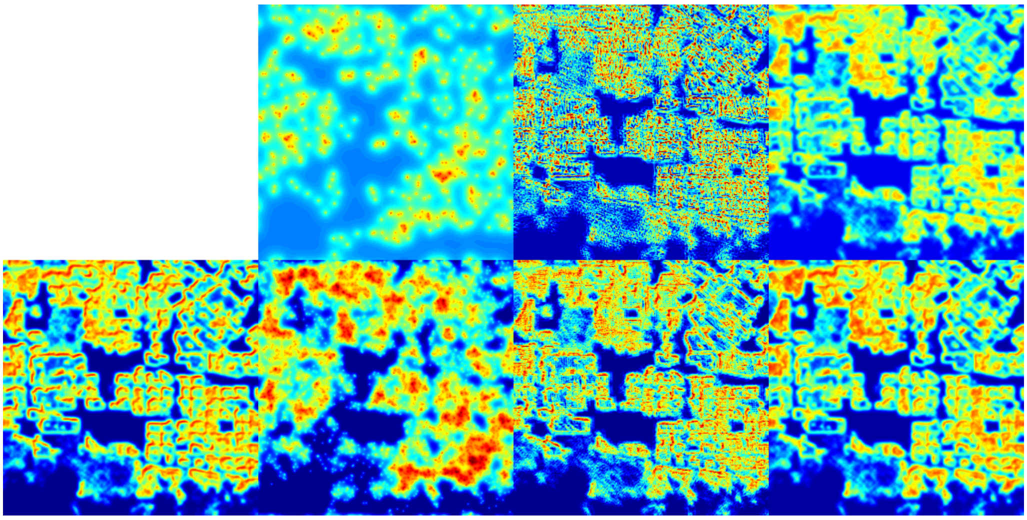

Figure 2: Synthetic density recovery (see Sec. 3(b))
Top row : density estimates based on 400 samples from synthetic density
$\overline{|\text{error}|}$: H^1 7.12473×10^{-6}, NL H^1 5.26617×10^{-6} , NL H^1 restricted 2.55042×10^{-6}
Bottom row : synthetic density and density estimates on 4,000 samples
$\overline{|\text{error}|}$: H^1 5.05662×10^{-6}, NL H^1 2.52831×10^{-6} , NL H^1 restricted 1.36416×10^{-6}

modeling the density of other types of events, the geographical data may not be related to housing at all. As the problem dictates, the nonlocal weights can be replaced with whatever weights seem appropriate for the data at hand. We have yet to incorporate time, leading indicators of crime, or census data into model. Any of these could further improve results and allow one to use density estimation in place of risk terrain modeling.

Finally, our method need not stand alone. Several sophisticated spatio-temporal models for probabilistic events make use of density estimation, typically using the standard methods [26–28]. By replacing the standard density estimation techniques with a nonlocally regularized MPLE such as ours, the density estimates in these models could improve, thus improving the overall result of the resulting simulation.

Acknowledgements

This work was supported by NSF grant DMS-0968309, W. M. Keck Foundation, ONR grant N000141210040, ONR grant N000141210838, AFOSR MURI grant FA9550-10-1-0569, ARO grant W911NF1010472, and NSF grant DGE-1144087. The authors would like to thank the LAPD for the residential burglary dataset, and the NSF Human Social Dynamics Program (BCS-0527388) for purchasing the housing data from the LA County Tax Assessor. The authors obtained the satellite image from Google Maps.

Data accessibility

The crime data cannot be shared because it contains human subject data.
The housing data is uploaded as online supplemental material.
The satellite image is uploaded as online supplemental material.
The synthetic density is uploaded as online supplemental material.

Appendix

To examine the effect of the nonlocal regularization term, we compute an alternate formulation of the NL H^1 MPLE problem and derive an inequality that solutions must satisfy. Recall from equation (2.1) that NL H^1 MPLE applied to the event samples $X = \{x_i\}_{i=1}^n$ with parameters $\alpha, \beta \geq 0$ is given by the following optimization.

$$u_{\alpha,\beta,X} := \underset{u\geq 0, \int_\Omega u=1}{\text{argmax}} \sum_{i=1}^n \log\left(u(x_i)\right) - \alpha \iint_{\Omega\times\Omega} \left(\nabla_{w,s} u(x,u)\right)^2 dxdy - \frac{\beta}{2}\int_\Omega |\nabla u(x)|^2 dx$$

For every such X, α, β one can show there exists nonnegative constants C_1, C_2 such that $u_{\alpha,\beta,X}$ is also the solution to a more constrained optimization.

$$u_{\alpha,\beta,X} = \text{argmax} \sum_{i=1}^n \log\left(u(x_i)\right) \text{ subject to}$$
$$\left\{ u \geq 0, \int_\Omega u = 1, \iint_{\Omega\times\Omega} \left(\nabla_{w,s} u(x,y)\right)^2 dxdy \leq C_1, \frac{1}{2}\int_\Omega |\nabla u(x)|^2 dx \leq C_2 \right\} \quad \text{(A 1)}$$

It can further be shown that for X and $\beta \geq 0$ fixed, C_1 is a non-increasing function of $\alpha \geq 0$ and for X and $\alpha \geq 0$ fixed, C_2 is a non-increasing function of $\beta \geq 0$.

Any solution of equation (A 1) satisfies $\iint_{\Omega\times\Omega} (\nabla_{w,s} u(x,y))^2 dxdy \leq C_1$, and likewise in the discrete setting we have the following.

$$\sum_{i,j\in\Omega} (u_i - u_j)^2 \frac{w_{ij}}{\sqrt{d_i d_j}} \leq C_1$$

Thus for some nonnegative discrete function $f : \Omega \times \Omega \to \mathbb{R}^{\geq 0}$ with $\sum_{i,j\in\Omega} f_{ij} \leq C_1$ we have the following.

$$\forall i, j \in \Omega, \quad (u_i - u_j)^2 \leq f_{ij} \frac{\sqrt{d_i d_j}}{w_{ij}} \quad \text{(A 2)}$$

Recalling that in our application, we set the weights w_{ij} to be nonlocal means applied to a housing image, $g : \Omega \to \mathbb{R}$, we can interpret what this means. Up to some factors constrained by the parameter C_1, the squared difference between the density at pixels i and j is bounded by $\sqrt{d_i d_j}/w_{ij}$. Thus the bound is made restrictive when : d_i and d_j are small, which means the patches of g around pixels i and j are very different from the rest of the image; and when w_{ij} is large, which means the neighborhoods of g around pixels i and j are similar to each other.

It is also worth noting that by constraint, the left-hand side of (A 2) is always smaller than or equal to 1. Thus for the inequality to be nontrivial, we must have $f_{ij} < w_{ij}/\sqrt{d_i d_j}$ for some pair $i, j \in \Omega$. Thus C_1 must be sufficiently small (or α sufficiently large) in order to guarantee that the nonlocal smoothing will have any effect on u.

References

1. Silverman BW.
Density estimation for statistics and data analysis. vol. 26.
CRC press; 1986.
2. Scott DW.
Multivariate density estimation.
Wiley; 1992.
3. Eggermont PPB, LaRiccia VN.
Maximum Penalized Likelihood Estimation: Regression. vol. 2.
Springer; 2001.
4. Mohler GO, Bertozzi AL, Goldstein TA, Osher SJ.
Fast TV regularization for 2D maximum penalized likelihood estimation.
Journal of Computational and Graphical Statistics. 2011;20(2):479–491.

5. Smith LM, Keegan MS, Wittman T, Mohler GO, Bertozzi AL.
Improving density estimation by incorporating spatial information.
EURASIP Journal on Advances in Signal Processing. 2010;2010:7.
6. Kostić T, Bertozzi AL.
Statistical Density Estimation Using Threshold Dynamics for Geometric Motion.
Journal of Scientific Computing. 2013;54(2-3):513–530.
7. Cour T, Benezit F, Shi J.
Spectral segmentation with multiscale graph decomposition.
In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition. vol. 2; 2005. p. 1124–1131 vol. 2.
8. Grady L, Schwartz EL.
Isoperimetric graph partitioning for image segmentation.
IEEE Transactions on Pattern Analysis and Machine Intelligence. 2006;28(3):469–475.
9. Shi J, Malik J.
Normalized cuts and image segmentation.
IEEE Transactions on Pattern Analysis and Machine Intelligence. 2000;22(8):888–905.
10. Chung FR.
Spectral graph theory. vol. 92.
AMS Bookstore; 1997.
11. Mohar B.
The Laplacian spectrum of graphs.
Graph theory, combinatorics, and applications. 1991;2:871–898.
12. Von Luxburg U.
A tutorial on spectral clustering.
Statistics and computing. 2007;17(4):395–416.
13. Zhou D, Schölkopf B.
A Regularization Framework for Learning from Graph Data.
In: ICML 2004 Workshop on Statistical Relational Learning and its Connections to Other Fields. Citeseer; 2004. p. 132.
14. Gilboa G, Osher SJ.
Nonlocal operators with applications to image processing.
Multiscale Modeling & Simulation. 2008;7(3):1005–1028.
15. Buades A, Coll B, Morel JM.
A review of image denoising algorithms, with a new one.
Multiscale Modeling & Simulation. 2005;4(2):490–530.
16. Merkurjev E, Kostic T, Bertozzi AL.
An MBO Scheme on Graphs for Classification and Image Processing.
SIAM Journal on Imaging Sciences. 2013;6(4):1903–1930.
17. Gilboa G, Osher SJ.
Nonlocal linear image regularization and supervised segmentation.
Multiscale Modeling & Simulation. 2007;6(2):595–630.
18. Zhang X, Chan TF.
Wavelet inpainting by nonlocal total variation.
Inverse problems and Imaging. 2010;4(1):191–210.
19. Peyré G, Bougleux S, Cohen L.
Non-local regularization of inverse problems.
In: Computer Vision–ECCV 2008. Springer; 2008. p. 57–68.
20. Lou Y, Zhang X, Osher SJ, Bertozzi AL.
Image recovery via nonlocal operators.
Journal of Scientific Computing. 2010;42(2):185–197.
21. Bertozzi AL, Flenner A.
Diffuse interface models on graphs for classification of high dimensional data.
Multiscale Modeling & Simulation. 2012;10(3):1090–1118.
22. Fowlkes C, Belongie S, Chung F, Malik J.
Spectral grouping using the Nystrom method.
IEEE Transactions on Pattern Analysis and Machine Intelligence. 2004;26(2):214–225.
23. Belongie S, Fowlkes C, Chung F, Malik J.
Spectral partitioning with indefinite kernels using the Nyström extension.
In: Computer Vision-ECCV 2002. Springer; 2002. p. 531–542.

24. Sardy S, Tseng P.
Density Estimation by Total Variation Penalized Likelihood Driven by the Sparsity ℓ_1 Information Criterion.
Scandinavian Journal of Statistics. 2010;37(2):321–337.
25. Kullback S, Leibler RA.
On information and sufficiency.
The Annals of Mathematical Statistics. 1951;22(1):79–86.
26. Mohler GO, Short MB, Brantingham PJ, Schoenberg FP, Tita GE.
Self-exciting point process modeling of crime.
Journal of the American Statistical Association. 2011;106(493).
27. Lewis E, Mohler GO, Brantingham PJ, Bertozzi AL.
Self-exciting point process models of civilian deaths in Iraq.
Security Journal. 2011;25(3):244–264.
28. Wang X, Brown DE.
The spatio-temporal modeling for criminal incidents.
Security Informatics. 2012;1(1):1–17.