ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Statistical Methodology
Series B

*J. R. Statist. Soc.* B (2019)
**81**, *Part* 5, *pp.* 839–860

# Fused density estimation: theory and methods

Robert Bassett

*Naval Postgraduate School, Monterey, USA*

and James Sharpnack

*University of California at Davis, USA*

[Received May 2018. Revised August 2019]

**Summary.** We introduce a method for non-parametric density estimation on geometric networks. We define *fused density estimators* as solutions to a total variation regularized maximum likelihood density estimation problem. We provide theoretical support for fused density estimation by proving that the squared Hellinger rate of convergence for the estimator achieves the minimax bound over univariate densities of log-bounded variation. We reduce the original variational formulation to transform it into a tractable, finite dimensional quadratic program. Because random variables on geometric networks are simple generalizations of the univariate case, this method also provides a useful tool for univariate density estimation. Lastly, we apply this method and assess its performance on examples in the univariate and geometric network setting. We compare the performance of various optimization techniques to solve the problem and use these results to inform recommendations for the computation of fused density estimators.

*Keywords*: Density estimation; Fusion penalty; Minimax rates; Non-parametrics; Total variation

## 1. Introduction

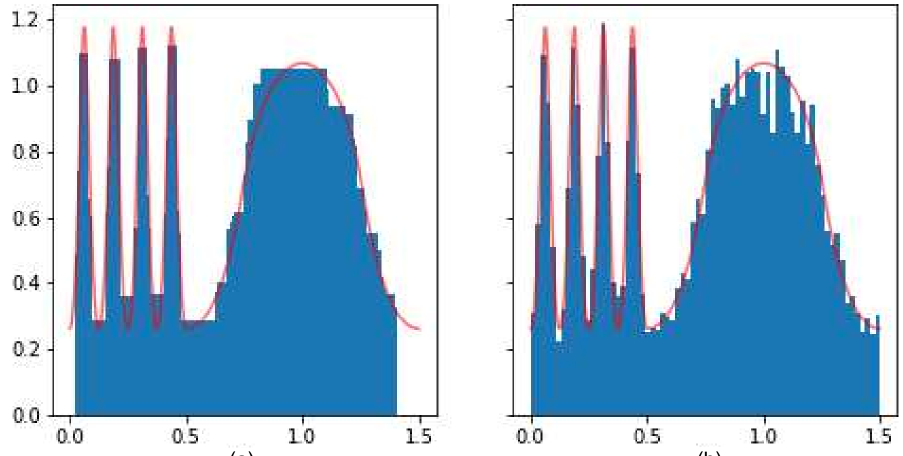
In the pantheon of statistical tools, the histogram remains the primary way to explore univariate empirical distributions. Since its introduction by Karl Pearson in the late 19th century, the form of the histogram has remained largely unchanged. In practice, the regular histogram, with its equal bin widths chosen by simple heuristic formulae, remains one of the most ubiquitous statistical methods. Most methodological improvements on the regular histogram have come from the selection of bin widths—this includes varying bin widths to construct irregular histograms—motivated by thinking of the histogram as a piecewise constant density estimate. In this work, we study a piecewise constant density estimation technique based on total variation penalized maximum likelihood. We call this method fused density estimation (FDE). We extend FDE from irregular histogram selection to density estimation over geometric networks, which can be used to model observations on infrastructure networks like road systems and water supply networks. The use of fusion penalties for density estimation is inspired by recent advances in theory and algorithms for the fused lasso over graphs (Padilla *et al.*, 2017; Wang *et al.*, 2016). Our thesis, that FDE is an important algorithmic primitive for statistical modelling, compression and exploration of stochastic processes, is supported by our development of fast implementations, minimax statistical theory and experimental results.

*Address for correspondence*: Robert Bassett, Department of Operations Research, Naval Postgraduate School, 1 University Circle, Monterey, CA 93943-5755, USA.
E-mail: robert.bassett@nps.edu

1369–7412/19/81839

Sturges (1926) provided a heuristic for regular histogram selection where, naturally, the bin width increases with the range and decreases with the number of points. The regular histogram is an efficient density estimate when the underlying density is uniformly smooth, but irregular histograms can 'zoom in' on regions where there are more data and better capture the local smoothness of the density. A simple irregular histogram, which is known as the equal area histogram, is constructed by partitioning the domain so that each bin has the same number of points. Denby and Mallows (2009) noted that the equal area histogram can often split bins unnecessarily when the density is smooth and merge bins when the density is variable, and proposed a heuristic method to correct this oversight. Recently, Li *et al.* (2016) proposed the essential histogram, which is an irregular histogram constructed such that it has the fewest number of bins and lies within a confidence band of the empirical distribution. Although theoretically attractive, in practice its complex formulation is intractable and requires approximation. If the underlying density is nearly constant over a region, then the empirical distribution is well approximated locally by a constant, and hence the essential histogram will tend not to split this region into multiple bins. Such a method is called *locally adaptive*, because it adapts to the local smoothness of the underlying density.

In Fig. 1, we compare FDE with the regular histogram, both of which have 70 bins. Because FDE can be thought of as a bin selection procedure, in this example we recompute the restricted maximum likelihood estimate after bin selection, which is common practice for model selection with lasso-type methods. We see that with 70 bins the regular histogram can capture the variability in the leftmost region of the domain but undersmooths in the rightmost region. We can compare this with FDE which adapts to the local smoothness of the true density. As a natural extension of one-dimensional data, we also consider distributions that lie on geometric networks—graphs where the edges are continuous line segments—such as is common in many infrastructure networks. Another motivation to use total variation penalties is that they are easily defined over any geometric network, in contrast with other methods, such as the essential histogram and multiscale methods. Fig. 2 depicts FDE for data in downtown San Diego. The geometric network is generated from the road network in the area, and observations on the geometric network are the locations of eateries (data extracted from the OpenStreetMap database: `https://www.openstreetmap.org`).



(a) (b)

**Fig. 1.** Comparison of (a) FDE and (b) the regular histogram of 10000 data points from a density (———) with varying smoothness—both have 70 bins

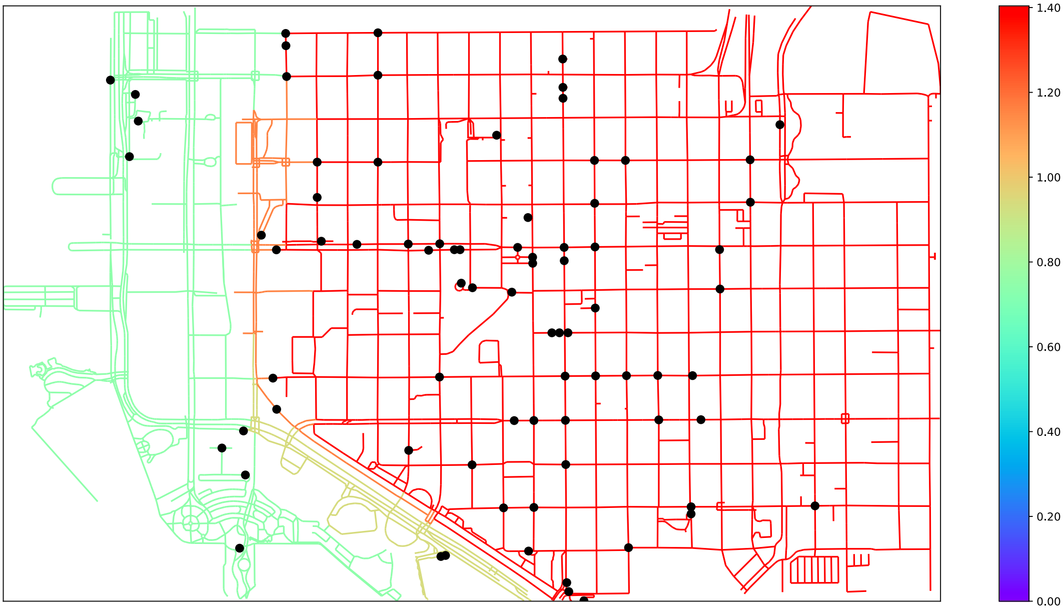


**Fig. 2.** FDE for the location of eateries in downtown San Diego

Without any constraints, maximum likelihood will select histograms that have high variation (as in Fig. 1). To regularize the problem, we bias the solution to have low total variation. Total variation penalization is a popular method for denoising images, time series and signals over the vertices of a graph, with many modern methods available for computation, such as the alternating direction method of multipliers (ADMM), projected Newton methods and split Bregman iteration (Tibshirani *et al.*, 2005; Rudin *et al.*, 1992; Beck and Teboulle, 2009; Wang *et al.*, 2016). Distributions over geometric networks, which we consider here, are distinguished from this literature by the fact that observations can occur at any point along an edge of the network. This leads to a variational density estimation problem, which we reduce to a finite dimensional formulation.

### *1.1. Contribution 1*

We show that variational FDE is equivalent to a total variation penalized weighted least squares problem enabling fast optimization.

To justify the use of FDE, we shall analyse the statistical performance of FDE for densities of log-bounded variation over geometric networks. The majority of statistical guarantees for density estimates control some notion of divergence between the estimate and the true underlying density. Several researchers have used the $L_2$-loss (the mean integrated square error) to evaluate their methods for tuning the bin width for the regular histogram (Scott, 1979; Freedman and Diaconis, 1981; Birgé and Massart, 1997). Although it is appealing to use $L_2$-loss, it is not invariant to the choice of base measure, and divergence measures such as $L_1$, Hellinger loss and the Kullback–Leibler divergence are preferred for maximum likelihood—an idea pioneered by Le Cam (Le Cam and Yang, 2012) and furthered by Devroye and Györfi (1985) and Hall and Wand (1988). By appealing to Hellinger loss, Birgé and Rozenholc (2006) proposed a method for optimal choice of the number of bins in a regular histogram, and we shall similarly focus on Hellinger loss.

### 1.2. Contribution 2

We provide a minimax non-parametric Hellinger distance rate guarantee for FDE in the univariate case, over densities of log-bounded variation.

When the log-density lies in a Sobolev space, an appropriate non-parametric approach to density estimation is maximum likelihood with a smoothing splines penalty (Silverman, 1982). The smoothing spline method is not locally adaptive because it does not adjust to the local smoothness of the density or log-density. Epi-splines (Royset and Wets, 2013) are density estimates that are formed by maximizing the likelihood such that the density, or log-density, has a representation in a local basis and lies in a prespecified constraint set. Donoho (1996) and Koo and Kim (1996) studied wavelet thresholding for density estimation and proved $L_p$-rate and Kullback–Leibler divergence guarantees respectively. In a related work, Koo and Kooperberg (2000) considered log-spline density estimation from binned data with stepwise knot selection. Willett and Nowak (2007) used a recursive partitioning approach to form adaptive polynomial estimates of the density, which is a similar approach to wavelet decomposition.

Total variation penalization has previously been proposed as a histogram regularization technique in Koenker and Mizera (2007), Sardy and Tseng (2010) and Padilla and Scott (2015). In particular, Padilla and Scott (2015) separately studied the variational form of the fused density estimate and a discrete variant, and provided theoretical guarantees for Lipschitz classes. Our computational results improve on these works by minimizing a variational objective directly, instead of separately proposing discrete approximations to the variational problem. Our theoretical analysis improves on previous work by studying total variation classes directly and showing that FDE achieves the minimax rate for Hellinger risk over all densities of log-bounded variation. Moreover, we consider density estimation on geometric networks and extend our Hellinger rate guarantees to this novel setting.

### 1.3. Contribution 3

We prove that the same Hellinger distance rate guarantee for the univariate case also holds for any connected geometric network.

### 1.4. Problem statement

When considering road systems and water networks, we observe that individual roads or pipes can be modelled as line segments, and the entire network constructed by joining these segments at nodes of intersection. Mathematically, we model this as a *geometric network* $G$, a finite collection of nodes $V$ and edges $E$, where each edge is identified with a closed and bounded interval of the real line (Fig. 3). Each edge in the network has a well-defined notion of length, inherited from the length of the closed interval. We fix an orientation of $G$ by assigning, for each edge $e=\{v_i, v_j\}$, a bijection between $\{v_i, v_j\}$ and the end points of the closed interval that is associated with $e$. This corresponds to the intuitive notion of 'gluing' edges together to form a geometric network. Because we discuss only geometric networks in this paper, we often refer to them as networks.

A *point* in a geometric network $G$ is an element of one of the closed intervals identified with edges in $G$, modulo the equivalence of end points corresponding to the same node. After assigning an orientation to the network, a point can be viewed as a pair $(e, t)$, where $e$ is an edge and $t$ is a real number in the interval that is identified with $e$. However, because we wish to emphasize the network as a geometric object in its own right, we shall use this notation only when our use of univariate theory makes it necessary.

A real-valued function $g$, defined on a geometric network $G$, is a collection of univariate

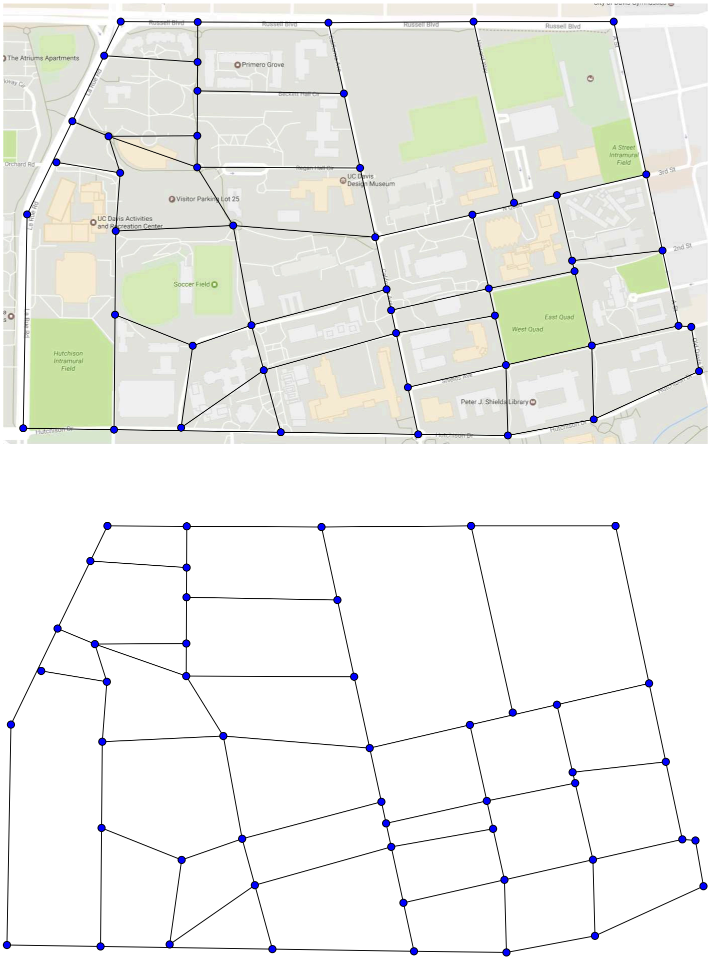


**Fig. 3.** Example of a geometric network: this network is formed from bike paths on a university campus; nodes identify intersections of path; edges are paths connecting these intersections, and the length of each edge is the length of the corresponding path; in forming the geometric network, we discard all information related to its embedding in $\mathbb{R}^2$, only preserving the network structure and path lengths

functions $\{g_e\}_{e \in E}$, defined on the edges of $G$. We require that the function respects the network structure, by which we mean that, for any two edges $e_1$ and $e_2$ which are incident at a node $v$, $g_{e_1}(v) = g_{e_2}(v)$. We abuse the notation slightly—by referring to $g_e(v)$, we mean $g$ evaluated at the end point of the interval identified with $v$. A geometric network $G$ inherits a measure from its univariate segments in a natural way, as the sum of the Lebesgue measure along each segment. With this measure we have a straightforward extension of the Lebesgue measure to $G$, making $G$ a measurable space.

For any random variable taking values on the network $G$, we assume that the measure that is induced by the random variable is absolutely continuous with respect to the base measure d$x$ and so has density $f$. We abuse the notation by using d$x$ to refer to both the Lebesgue measure and the base measure on a geometric graph; which of these we mean will be clear from its context. Furthermore, we assume that the density is non-zero everywhere, so that its logarithm is well defined. Moreover, we assume that the log-density is not arbitrarily variable, and for this we use the notion of total variation. Let $B \subseteq \mathbb{R}$. The *total variation* of a function $g : B \to \mathbb{R}$ is defined as

$$\mathrm{TV}(g) = \sup_{P \subset B} \sum_{z_i \in P} |g(z_i) - g(z_{i+1})|.$$

The supremum is over all partitions, or finite ordered point subsets $P$, of $B$. For a real-valued function $g$ defined on a network $G$, we extend the univariate definition to

$$\mathrm{TV}(g) = \sum_{e \in E} \mathrm{TV}(g_e).$$

One advantage of the use of the total variation penalty is that it is invariant to the choice of the segment length in the geometric network, so scaling the edge by a constant multiplier leaves the total variation unchanged. As a consequence, FDE will be invariant to the choice of edge length.

Let $f_0$ be a density on a geometric network $G$, and $x_1, \ldots, x_n$ an independent sample identically distributed according to $f_0$. Let $P_n = (1/n)\Sigma\delta_{x_i}$ be the empirical measure associated with the sample. We let $P_n$ act on a function, by which we mean that we take the expectation of that function with respect to $P_n$. So, for any function $f$,

$$P_n(f) = \int f \, \mathrm{d}P_n = \frac{1}{n} \sum_{i=1}^{n} f(x_i).$$

We also use $P(f)$ to denote $\int f \, \mathrm{d}P$ for non-empirical measures $P$.

Fix $\lambda \in \mathbb{R}^+$. A *fused density estimator* of $f_0$ is a density $\hat{f} = \exp(\hat{g})$, such that the log-density $\hat{g}$ is the minimizer of the following program:

$$\min -P_n(g) + \lambda \mathrm{TV}(g) \qquad \text{subject to} \int \exp(g) \, \mathrm{d}x = 1 \tag{1}$$

where the minimum is taken over all functions $g : G \to \mathbb{R}$ for which the expression is finite and the resulting $f$ is a valid density, i.e. $f \in \mathcal{F}$ and $g \in \mathcal{G}$ where

$$\mathcal{F} = \{\exp(g) : g \in \mathcal{G}\}, \qquad \mathcal{G} = \left\{ g : \mathrm{TV}(g) < \infty, \int_G \exp(g) \mathrm{d}x = 1 \right\}.$$

The set $\mathcal{F}$ will be referred to as the set of densities with *log-bounded variation*. Indeed, the integration constraint on elements of $\mathcal{G}$ makes them log-densities. Note that densities in $\mathcal{F}$ are necessarily bounded above and away from zero, as a result of the total variation condition.

Program (1) is variational, because it is a minimizer over an infinite dimensional function space. It is quite common for variational problems in non-parametric statistics to involve a reproducing kernel Hilbert space penalty, as opposed to a total variation penalty (Wahba, 1990). In the reproducing kernel Hilbert space setting, the Hilbert space enables us to establish representer theorems, which reduce the variational program to an equivalent finite dimensional program, so that it can be solved numerically. The space of functions of bounded variation, in contrast, is an example of a more general Banach space, so reproducing kernel Hilbert space results cannot be applied to this setting. In the next section, we discuss representer theorems for program (1) and further show that it can be solved by using a sparse quadratic program.

## 2. Computation

In this section we provide results towards the computation of fused density estimators. The key challenge is the variational formulation of the fused density estimator (1). For this, we prove that solutions to the variational program can be finitely parameterized. Moreover, we show that, after applying this representer theorem, the finite dimensional analogue of program (1) has an equivalent formulation as a total variation penalized least squares problem. Our main theorem of this section, which reduces the computation of a fused density estimator to a weighted fused lasso problem, follows.

> *Theorem 1* (informal). For $\lambda > 1/(2n)$ the fused density estimator exists almost surely. It can be computed as the minimizer to a finite dimensional convex, sparse and total variation penalized quadratic program, i.e. fused density estimators are solutions to an optimization of the form
>
> $$\min_{z \in \mathbb{R}^d} \tfrac{1}{2} z^{\mathrm{T}} P z + a^{\mathrm{T}} z + \|Dz\|_1. \tag{2}$$
>
> The details of this theorem, by which we mean the constructions of $P$, $D$, $a$ and the connection between the minimizer $\hat{z}$ of problem (2) and the fused density estimator $\hat{f}$, are given later in this section.

Theorem 1 demonstrates that the fused density estimator (1) can be computed as a specific incarnation of the generalized lasso, for which there are well-known fast implementations (Arnold and Tibshirani, 2016). In practice we solve the dual to this problem, which we discuss in theorem 3 in Section 2.1. Theorem 2 is a precise restatement of theorem 1. To prove it, we proceed through a series of important lemmas. Lemma 1 in Section 2.1 proves that minimizers of program (1) exist almost surely for $\lambda > 1/(2n)$, below which the solution degenerates to Dirac masses at observations. The almost surely qualification pertains to the maximum number of observations which occur at any single point of the network; if observations occur simultaneously then $\lambda$ must be increased to overcome degeneracy. Lemma 1 further transforms the FDE problem from constrained to unconstrained by removing the integration constraint. From this new formulation, lemma 2 shows that the search space for the fused density estimator problem can be reduced from functions of bounded variation to an equivalent, finite dimensional version. Theorem 2 performs the final step in the proof—demonstrating that the previously derived finite dimensional problem can be solved using an $l_1$ penalized quadratic program. The last subsection in this section is tangential but sheds further light on the structure of fused density estimators. Proposition 1, which we refer to as the ordering property, qualifies the local adaptivity of fused density estimators by describing their local structure. Omitted proofs can be found in the on-line appendix A.

### *2.1. Main computational results*

Our first lemma reduces the FDE formulation (1) to an unconstrained program where the integral constraint is incorporated in the objective. This result is originally due to Silverman (1982), who proved the result in the context of univariate density estimation and Sobolev norm penalties. Minor modifications enable us to extend it to geometric networks and the non-Sobolev total variation penalty.

> *Lemma 1.* The problem
>
> $$\min_{g \in \mathcal{G}} -P_n(g) + \lambda \mathrm{TV}(g) + \int_G \exp(g) \mathrm{d}x \tag{3}$$
>
> gives an equivalent formulation of program (1), because minimizers $\hat{g}$ of problem (3) satisfy $\int_G \exp(\hat{g}) \mathrm{d}x = 1$.

We remark that the objective in lemma 1 is equivalent to a total variation penalized Poisson process likelihood, where the log-intensity is $g$, so our computations also apply to that setting.

Lemma 1 gives that the fused density estimator definition (1) can instead be solved by the unconstrained problem (3) over all functions $g$ on $G$ of bounded variation. An alternative interpretation of the lemma is that the Lagrange multiplier that is associated with the constraint in program (1) is 1. The next lemma reduces the unconstrained problem (3) to an equivalent finite dimensional version. The proof technique is analogous to similar results in Mammen and van de Geer (1997). In the context of reproducing kernel Hilbert spaces, results that reduce variational problem formulations to finite dimensional analogues are referred to as *representer theorems*, e.g. Wahba (1990). We shall also use this language to describe our result, even though we are in a more general Banach space setting. The result demonstrates that fused density estimators have large, piecewise constant regions, which is a well-known property of fusion penalties (Tibshirani *et al.*, 2005; Kim *et al.*, 2009; Wang *et al.*, 2016).

> *Lemma 2* (representer theorem). A fused density estimator $\hat{f}$ must be piecewise constant along each edge. All discontinuities are contained in the set $\{x_1, \ldots, x_n\} \cup V$, the observations and the nodes of $G$.

Using lemma 2, we can parameterize fused density estimators with three finite dimensional vectors: the fused density estimator at the observation points $p$, the fused density estimator at the vertices of $G$, $k$, and the piecewise constant values of the fused density estimator, $c$. For simplicity, we assume that no two observations occur at the same location, which is a condition that we can and will relax in the remark following theorem 2.

Let $n_e$ denote the number of observations along edge $e$. We denote by $p_{e,i}$ the value, in the vector $p$, of the $i$th ordered observation along edge $e$. For a fused density estimator $f = \exp(g)$, $c_{e,i}$ is the value that is taken by $g$ between the $(i-1)$th and $i$th observation in the interval that is associated with $e$, where the zeroth and $(n_e+1)$th observations are set to be the end points of that interval. Similarly, $s_{e,i} = x_{e,i} - x_{e,i-1}$ with the convention that $x_{e,0}$ and $x_{e,n_e+1}$ are the end points of the interval. This gives the length of the segment between two observations, over which the fused density estimator is piecewise constant. We denote by $k_v$ the value in $k$ at the vertex $v$. For a given node $v$, let inc$(v)$ denote the set of edges which are incident to $v$ and denote by $c_{e,v}$ the segment in $c$ which is incident to $v$. Problem (3) becomes

$$\min_{p,c,k} \sum_{e\in E} \left\{ -\frac{1}{n} \sum_{i=1}^{n_e} p_{e,i} + \lambda \sum_{i=1}^{n_e} |p_{e,i} - c_{e,i}| + |p_{e,i} - c_{e,i+1}| + \sum_{i=1}^{n_e+1} s_{e,i} \exp(c_{e,i}) \right\}$$
$$+ \lambda \sum_{v\in V} \sum_{e\in \mathrm{inc}(v)} |k_v - c_{e,v}|.$$

The first summand, over the edges in $E$, gives the log-likelihood term, the total variation along an edge, and the integration term. The second summand gives the total variation at nodes of the geometric graph.

Let $F$ denote the objective function above. We show that this problem can be further reduced by removing the $p_{e,i}$-variables. Indeed, for any vectors $\hat{c}$ and $\hat{k}$, let $\tilde{F}(p) = F(p, \hat{c}, \hat{k})$. A necessary condition for any $\hat{p}$, $\hat{c}$ and $\hat{k}$ to minimize $F$ is $\hat{p} \in \arg\min_p \tilde{F}(p)$. But $\tilde{F}(p)$ does not have a lower bound when $\lambda < 1/(2n)$. Furthermore, the set $\arg\min_p \tilde{F}(p)$ is unbounded when $\lambda = 1/(2n)$, and $\tilde{F}(p)$ has a unique minimum when $\lambda > 1/(2n)$. These facts are clear from the graph of the functions $p_{e,i} \rightarrow -p_{e,i}/n + \lambda(|p_{e,i} - c_{e,i}| + |p_{e,i} - c_{e,i+1}|)$, which occur as summands in $F$. The function is given in Fig. 4. It is the maximum of three affine functions, from which we conclude that the minimum of $\tilde{F}$ is attained uniquely at $p_{e,i} = \max\{c_{e,i}, c_{e,i+1}\}$ when $\lambda > 1/(2n)$. We have shown that $\lambda = 1/(2n)$ is a critical point for the existence of fused density estimators below which the total variation penalty is not sufficiently strong to prevent degenerate solutions to program (1). For $\lambda > 1/(2n)$, the value of a fused density estimator at observations is well behaved

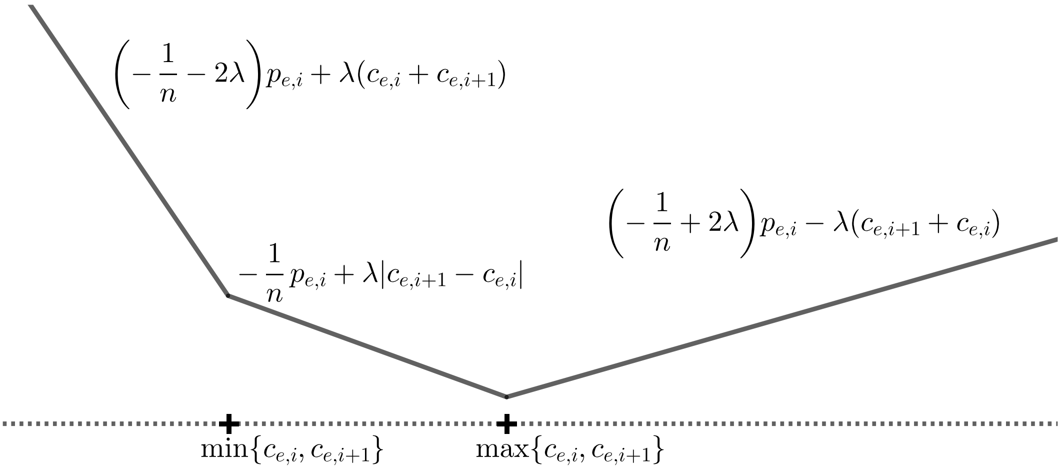


**Fig. 4.** The function $p_{e,i}\rightarrow -p_{e,i}/n+\lambda(|p_{e,i}-c_{e,i}|+|p_{e,i}-c_{e,i+1}|)$ as the maximum of three affine functions: it attains a unique minimum at $p_{e,i}=\max\{c_{e,i},c_{e,i+1}\}$ when $\lambda>1/(2n)$

and we can reduce $F(p,c,k)$ to

$$\min_{c,k}\sum_{e\in E}\left\{-\frac{1}{n}\sum_{i=1}^{n_e}\max\{c_{e,i},c_{e,i+1}\}+\lambda\sum_{i=1}^{n_e}|c_{e,i}-c_{e,i+1}|+\sum_{i=1}^{n_e+1}s_{e,i}\exp(c_{e,i})\right\}$$
$$+\lambda\sum_{v\in V}\sum_{e\in\text{inc}(v)}|k_v-c_{e,v}|.$$

Because $2\max\{c_{e,i},c_{e,i+1}\}=c_{e,i}+c_{e,i+1}+|c_{e,i}-c_{e,i+1}|$, we have the further equivalence

$$\min_{c,k}\sum_{e\in E}\left\{-\frac{1}{2n}\sum_{i=1}^{n_e}(c_{e,i}+c_{e,i+1})+\left(\lambda-\frac{1}{2n}\right)\sum_{i=1}^{n_e}|c_{e,i}-c_{e,i+1}|+\sum_{i=1}^{n_e+1}s_{e,i}\exp(c_{e,i})\right\}$$
$$+\lambda\sum_{v\in V}\sum_{e\in\text{inc}(v)}|k_v-c_{e,v}|.$$

By Rockafellar (2015), theorem 23.8, a necessary and sufficient condition for $\hat{c}$ and $\hat{k}$ to solve this problem is

$$0\in\sum_{e\in E}\left[-\frac{1}{2n}\sum_{i=1}^{n_e}\partial(c_{e,i}+c_{e,i+1})+\left(\lambda-\frac{1}{2n}\right)\sum_{i=1}^{n_e}\partial(|c_{e,i}-c_{e,i+1}|)+\sum_{i=1}^{n_e+1}\partial\{s_{e,i}\exp(c_{e,i})\}\right]$$
$$+\lambda\sum_{v\in V}\sum_{e\in\text{inc}(v)}\partial(|k_v-c_{e,v}|).$$

Here we make an important point. The subdifferential of each $(c_{e,i}+c_{e,i+1})$-term is constant, and the subdifferential of $|c_{e,i}-c_{e,i+1}|$ is piecewise constant, *depending only on the ordering* of the terms $c_{e,i}$ and $c_{e,i+1}$. Similarly, the subdifferential of the $|k_v-c_{e,v}|$-term is piecewise constant and again depends only on the ordering of its terms. Lastly, the subdifferential of $s_{e,i}\exp(c_{e,i})$ is given by its gradient: the $(e,i)$th co-ordinate of the subdifferential is $s_{e,i}\exp(c_{e,i})$.

Consider the transformation $z=\exp(c)$, $h=\exp(k)$. This transformation preserves the ordering of elements of $\hat{c}$ and $\hat{k}$, so the subdifferential of each absolute value term is invariant under this transformation. Pursuing this line of reasoning gives the following theorem. To facilitate its statement, we briefly establish some notation.

The total variation of piecewise constant function $f$ on $G$, which has been parameterized into vectors $z$ and $h$, can be expressed as a sum of pairwise distances between values in $z$ and $h$, i.e. there are sets $J_1$ and $J_2$ of index pairs such that

$$\mathrm{TV}(f)=\sum_{(i,j)\in J_1}|z_i-z_j|+\sum_{(i,j)\in J_2}|z_i-h_j|.$$

This formulation depends on the underlying graph structure and the locations of the observations. The right-hand side of this expression can be written as the $l_1$-norm of a vector $C_1z+C_2h$, where $C_1$ and $C_2$ are matrices with elements in $\{-1,0,1\}$, each having $|J_1|+|J_2|$ rows. We shall use the matrices $C_1$ and $C_2$, which satisfy $\mathrm{TV}(f)=\|C_1z+C_2h\|_1$, and $C_2$ is 0 in its first $|J_1|$ rows. Let $n_i=|J_i|$ for $i\in\{1,2\}$. Let

$$B=\begin{pmatrix}\{\lambda-1/(2n)\}I_{n_1\times n_1} & \mathbf{0}_{n_1\times n_2}\\ \mathbf{0}_{n_2\times n_1} & \lambda I_{n_2\times n_2}\end{pmatrix}$$

and let $D_1$ and $D_2$ be the matrices $BC_1$ and $BC_2$ respectively. We denote by $x_1,\ldots,x_n$ the locations of observation on $G$, and we further partition these into ordered observations along each edge, so that $x_{e,i}$ denotes the $i$th observation along edge $e$. Recall the definition of $s_{e,i}=x_{e,i}-x_{e,i+1}$, and let $S$ be a diagonal matrix with $s$ on its diagonal. Lastly, define the vector $w$ such that

$$w_{e,i}=\begin{cases}-1/(2n) & i=1 \text{ or } i=n_e+1,\\ -1/n & \text{otherwise.}\end{cases}$$

*Theorem 2.* Let $\lambda>1/(2n)$. Then the fused density estimator exists almost surely. It can be computed as follows. Let $z$ be a vector with indices enumerating the constant portions of the fused density estimator $\hat{f}$, such that $z_{e,i}$ denotes the value of the fused density estimator on the open interval between $x_{e,i}$ and $x_{e,i-1}$, or between an observation and the end of the edge if $i=1$ or $n_e+1$. Let $h$ be a vector with indices enumerating the nodes in $G$, such that $h_v$ denotes the value of $\hat{f}$ at node $v$. Then the fused density estimator $\hat{f}$ for this sample is the minimizer of

$$\min_{z,h}\tfrac{1}{2}z^{\mathrm{T}}Sz+w^{\mathrm{T}}z+\|D_1z+D_2h\|_1. \tag{4}$$

*Proof.* The proof follows directly from the line of reasoning before the theorem's statement. Details can be found in the on-line appendix A.

*Remark 1.* The condition on $\lambda$ is an important condition. As discussed previously, when $\lambda<1/(2n)$ the total variation penalty is not sufficiently strong to balance the likelihood term and minimizers of program (1) are degenerate. The almost surely condition is simply a requirement that no two observations occur at the same location, and no observations occur at nodes of the geometric network. With a slight modification of the assumption on $\lambda$, theorem 2 can be extended to the setting where multiple observations are allowed at a single location. This extension also allows observations to occur at nodes of the geometric network. In practice, this extension may be useful when dealing with imperfect data, though we do not focus on it here because it is a measure zero event in the density estimation paradigm. For completeness, we include the extension as theorem A.1 of the on-line appendix A.

Methods for computing solutions to the problem in theorem 2—a total variation regularized quadratic program—are well established. As in Kim *et al.* (2009), we rely on solving the dual quadratic program. For convenience, we write the dual problem as a minimization instead of its typical maximum formulation.

*Proposition 1.* The dual problem to problem (4) is

$$\begin{aligned} &\min_y \tfrac{1}{2} y^{\mathrm{T}} D_1 S^{-1} D_1^{\mathrm{T}} y + w^{\mathrm{T}} S^{-1} D_1^{\mathrm{T}} y, \\ &\|y\|_\infty \leqslant 1, \\ &D_2^{\mathrm{T}} y = 0. \end{aligned} \tag{5}$$

The primal solution $\hat{z}$ can be recovered from the dual $\hat{y}$ through the expression

$$\hat{z} = -S^{-1}(D_1^{\mathrm{T}} \hat{y} + w).$$

A more general statement of proposition 1, which suits the more general statement of theorem 2, can be found in the on-line appendix A. It is worth noting that strong duality between the primal and dual problems (4) and (5) follows immediately. Indeed, both are extended linear–quadratic programs in the sense of Rockafellar and Wets (2009). By theorem 11.42 in Rockafellar and Wets (2009), strong duality holds, and in addition both the primal and the dual problem attain their minimum values if and only if problem (4) is bounded. This is guaranteed by the assumption on $\lambda$ in theorem 2. Furthermore, the fact that the minimum of expression (4) is attained gives the existence of fused density estimators as asserted in theorem 2.

### 2.2. Additional properties of fused density estimators

In this section, we state a result on the local structure of a fused density estimator and provide additional comments on its implementational details. The result is intuitive: along an edge, the value of piecewise constant segments is inversely related to the length of the segment, relative to adjacent segments. Since smaller segments suggest higher probability in the corresponding region, this property demonstrates local structure of the estimator which aligns with essential global behaviour.

*Proposition 2* (ordering property). Let $s_{e,i}$ and $s_{e,i+1}$ be the lengths of two segments interior to an edge $e$, in the sense that $2 \leqslant i \leqslant n_e - 1$. Assume further that no two observations occur at the same location. Then $s_{e,i} \leqslant s_{e,i+1}$ implies that $\hat{z}_{e,i} \geqslant \hat{z}_{e,i+1}$. Similarly, $s_{e,i} \geqslant s_{e,i+1}$ implies that $\hat{z}_{e,i} \leqslant \hat{z}_{e,i+1}$.

Up to this point in our analysis, we have discussed the computation of the fused density estimator without consideration for preprocessing the data or post-processing our resulting fused density estimator. Since the computation and rates of convergence of FDE represent the bulk of our contribution, we maintain this perspective in the remainder of the paper. It is worth mentioning, however, that FDE is amenable to preprocessing and post-processing. Handling multiple observations at a single location in theorem 2 makes initial binning or minor discretizations of data (such as projecting observations *onto* a geometric network) straightforward. Moreover, FDE can be viewed exclusively as a method for generating adaptive bin widths, where the resulting bins can then be fitted to the data as in a regular histogram. This approach performs model selection (via FDE) and model fit (via a post-selection maximum likelihood estimate) of the histogram separately, and is common practice in model selection using lasso and related methods (Meir and Drton, 2017; Fithian *et al.*, 2014). When FDE is used exclusively to find bins, it becomes a change point localization method, instead of a non-parametric density estimator as in its original formulation. Though FDE is amenable to these examples of preprocessing and post-processing, we shall examine the fused density estimator as a density estimator in the remaining sections.

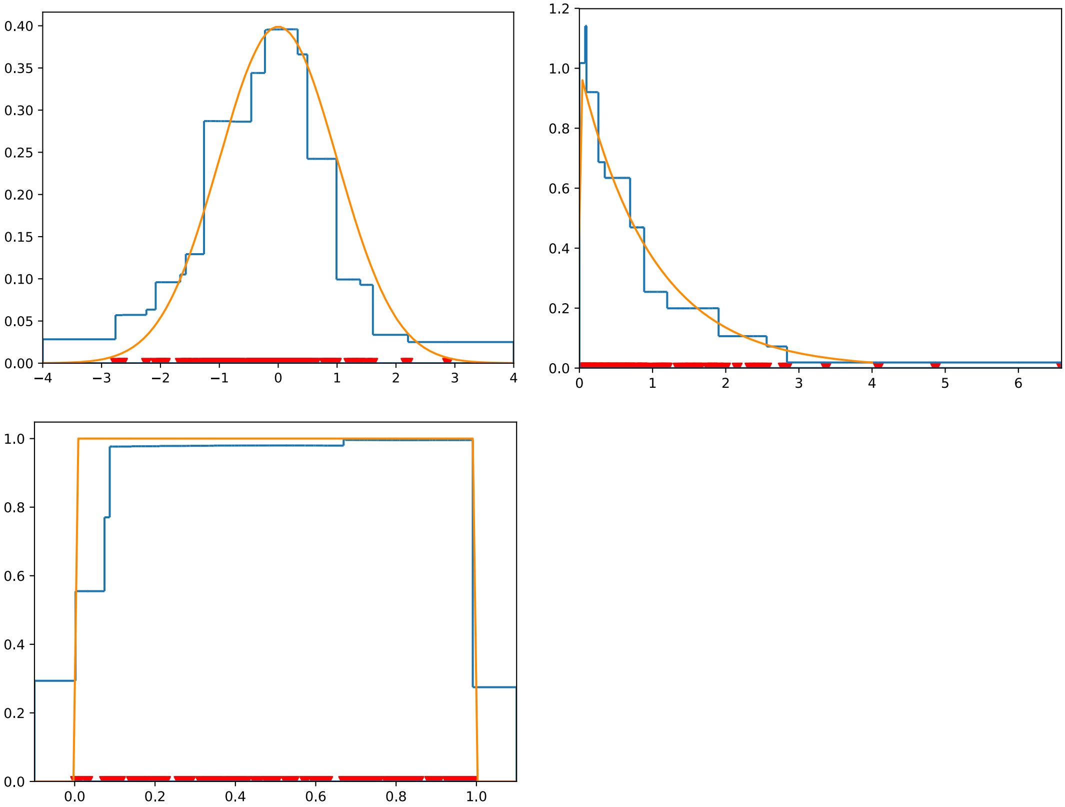


**Fig. 5.** Univariate densities and fused density estimators

We also make some suggestions on the selection of $\lambda$. The choice of $\lambda$ leads to a fixed number of piecewise constant portions of the fused density estimator. In this sense, the choice of the $\lambda$-parameter is analogous to choosing the number of bins in histogram estimation. One can tune this selection with an information criterion such as the Akaike information criterion AIC or Bayesian information criteria BIC by selecting the FDE over a grid of $\lambda$-values that minimizes the information criterion. Each of these information criteria requires the specification of the degrees of freedom, which can be set to the number of selected piecewise constant regions in the graph, as is done in the Gaussian case (Wang *et al.* 2016; Tibshirani *et al.*, 2012). Alternatively, one could use cross-validation as a selection criterion. Implementing cross-validation is often practical for large problems because the sparse quadratic program (5) can be solved very quickly, as we see in the next section.

## 3. Experiments

We have established a tractable formulation of the fused density estimator (5). Quadratic programming is a mature technology, so computing fused density estimators via quadratic programming dramatically improves its computation. Solvers that are designed to leverage sparsity in the $D_1$- and $D_2$-matrices allow the optimization portion of FDE to scale to large networks and many observations.

In this section we compute fused density estimators on synthetic and real world examples.

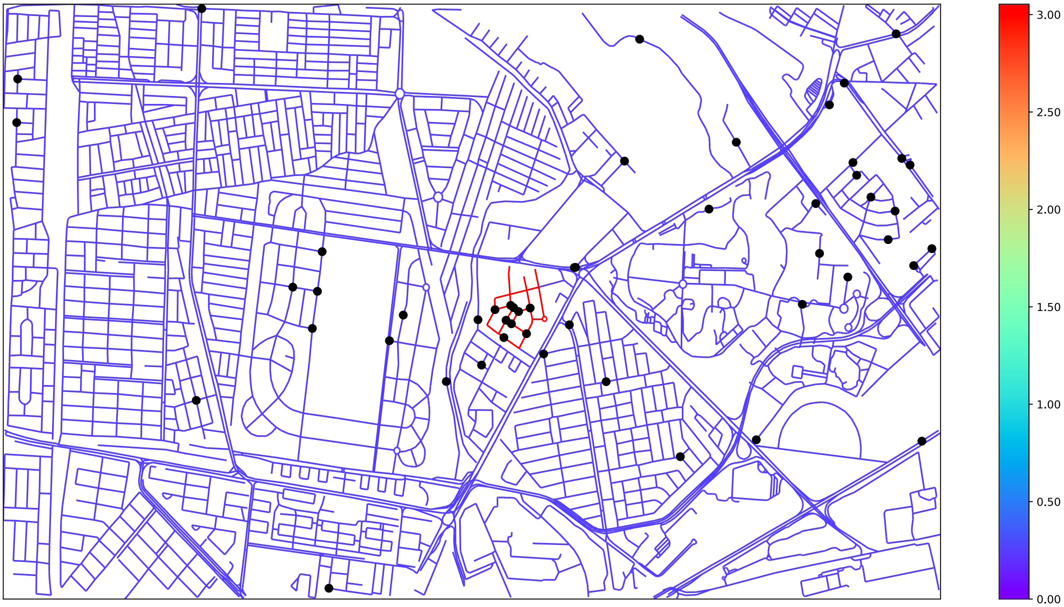


**Fig. 6.** A fused density estimator for the location of terrorist attacks in a neighbourhood of Baghdad: the detected hot spot contains the streets and alleys near a hospital

(These examples can be found at `github.com/rbassett3/FDE-Tools`, which also includes a Python package for FDE on geometric networks.) We evaluate the performance of various optimization methods and provide recommendations for solvers which implement those methods. To facilitate accessibility and customization of these tools, each of the solvers that we consider is open source and they compare favourably with commercial alternatives.

### *3.1. Univariate examples*

We first evaluate fused density estimators in the context of univariate density estimation—where the geometric network $G$ is simply a single edge connecting two nodes. The operator $D_1 + D_2$ is especially simple in this setting, corresponding to an oriented edge incidence matrix of a chain graph. Fig. 5 contains fused density estimators of the standard normal, exponential and uniform densities, each derived from 100 sample points. The $\lambda$-parameter in these experiments was selected by 20-fold cross-validation.

### *3.2. Geometric network examples*

We next evaluate fused density estimators on geometric networks. For each of these examples, the underlying geometric network is extracted from the OpenStreetMap database (`https://www.openstreetmap.org`). Fig. 6 is a fused density estimator with domain taken to be the road network in a region of the city of Baghdad. Observations are the locations of terrorist incidents which occurred in this region from 2013 to 2016, according to the global terrorism database (LaFree and Dugan, 2007). The density that we attempt to infer is the distribution for the location of terrorist attacks in this region of the city.

Fig. 7 is a fused density estimator on the road network of Monterey, California. The observations were generated according to a multivariate normal distribution and projected onto the nearest way point in the OpenStreetMap data set. These examples of fused density estimators

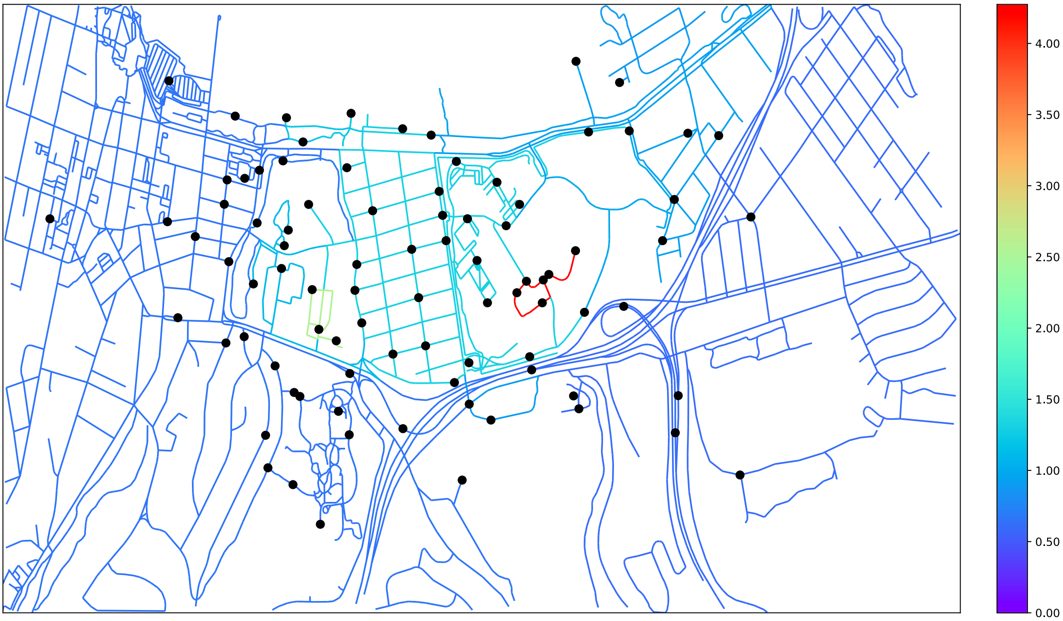


**Fig. 7.** A fused density estimator for artificial observations on Monterey's road network

on geometric networks illustrate some important properties of the estimator. The fused density estimators clearly respect the network topology. This is most obviously demonstrated in the Monterey example, where the red and light green regions, which correspond to elevated portions of the density, are chosen to be sparsely connected regions of the network. This is intuitive because the sparsely connected regions impact the fusion penalty less severely than a highly connected region, but it is one way that FDE reflects the underlying network structure. The Baghdad and Davis examples demonstrate that FDE can also be used for hot spot localization, and especially in low data circumstances. Lastly, we note that fused density estimators partition the geometric network into level sets, thereby forming various regions of the network into clusters. This clustering is an interesting aspect of FDE and suggests that it could be used to classify regions into areas of high and low priority.

### *3.3. Algorithmic concerns*

The two most prevalent methods for solving sparse quadratic programs are interior point algorithms and the ADMM. Interior point methods to solve problems of the form (4) were introduced by Kim *et al.* (2009). Interior point approaches have the benefit of requiring few iterations for convergence. The cost per iteration, however, depends crucially on the structure of $D_1$ and $D_2$ when performing a Newton step on the relaxed Karush–Kuhn–Tucker system. In the case of univariate fused density estimators, the Newton step requires inversion of a banded matrix: one which has its non-zero elements concentrated along the diagonal. Leveraging the banded structure allows inversion to be performed in linear time, which is crucial to the performance of the algorithm. For further details of interior point methods, we refer the reader to Wright (1997), Boyd and Vandenberghe (2004) and Nesterov and Nemirovskii (1994).

The ADMM proceeds by forming an augmented Lagrangian function and updating the primal and dual variables sequentially. More details can be found in Boyd *et al.* (2011) and

**Table 1.** Mean and standard deviation of run time for univariate OSQP experiments

| *Density* | *Results (s) for the following values of λ:* | | |
|---|---|---|---|
| | *0.006* | *0.05* | *0.1* |
| Exponential | 0.0361 ± 0.1310 | 0.0051 ± 0.0043 | 0.0045 ± 0.0045 |
| Normal | 0.0209 ± 0.0912 | 0.0112 ± 0.0569 | 0.0052 ± 0.0046 |
| Uniform | 0.0269 ± 0.1077 | 0.0769 ± 0.0565 | 0.0074 ± 0.0412 |

**Table 2.** Mean and standard deviation of run time for univariate CVXOPT experiments

| *Density* | *Results (s) for the following values of λ:* | | |
|---|---|---|---|
| | *0.006* | *0.05* | *0.1* |
| Exponential | 0.0087 ± 0.0012 | 0.0069 ± 0.0008 | 0.0078 ± 0.0011 |
| Normal | 0.0086 ± 0.0012 | 0.0071 ± 0.0009 | 0.0076 ± 0.0010 |
| Uniform | 0.0087 ± 0.0010 | 0.0065 ± 0.0008 | 0.0061 ± 0.0008 |

Bertsekas (2014). Compared with interior point methods, convergence of the ADMM usually requires more iterations of a less expensive update, whereas interior point methods converge in fewer iterations but require a more expensive update. In this section, we compare the performance of these algorithms on FDE problems. A comparison between the methods on the related problem of trend filtering can be found in Wang *et al.* (2016), where the algorithmic preferences pertained only to the 2×2 grid graph setting. Their results favour the ADMM approach, though the regularity of this graph structure makes generalizing to general graphs difficult.

For software, we use the operator splitting quadratic programming (OSQP) solver and CVXOPT. For the results, see Tables 1–4. These are mature sparse quadratic programming solvers that use ADMM and interior point algorithms respectively. They are both open source and compare favourably with commercial solvers (Stellato *et al.*, 2017; Caron, 2018). Our choice to use these solvers instead of custom implementations reflects that

(a) these tools are representative of what is available in practice,
(b) outsourcing this portion to other solvers reduces the ability for subtle differences in implementation to favour one method over another and
(c) these projects are production quality, so their implementations are likely to be of higher quality than custom implementations.

We first compare ADMM and interior point methods on univariate FDE problems. We perform 200 simulations, sampling 100 data points from each distribution. We let $\lambda$ range from 0.006 to 0.1. These choices correspond to the lower bound on the $\lambda$-parameter in theorem 2 and an upper bound which selects a constant or nearly constant density. We report in-solver time, in seconds, and do not include the time that is required to convert to the sparse formats that are required for each solver.

**Table 3.** OSQP run times for geometric network examples

| *Example* | *Results (s) for* the *following λ-parameters:* | | |
|---|---|---|---|
| | *Overfit* | *Middle* | *Underfit* |
| Baghdad | 0.1086 | 0.0686 | 0.0639 |
| San Diego | 0.0920 | 0.0961 | 0.0628 |
| Downtown Davis | 0.0269 | 0.0769 | 0.0074 |
| Davis | 12.0698 | 0.8539 | 0.6052 |

**Table 4.** CVXOPT run times for geometric network examples

| *Example* | *Results (s) for the following λ-parameters:* | | |
|---|---|---|---|
| | *Overfit* | *Middle* | *Underfit* |
| Baghdad | 1.5493 | 1.2813 | 1.1268 |
| San Diego | 0.5507 | 0.4956 | 0.3256 |
| Downtown Davis | 0.0812 | 0.5615 | 0.4864 |
| Davis | — | 13.4456 | 13.3911 |

In these experiments, the interior point terminated in around 10 iterations. The number of iterations in the ADMM were less consistent, ranging from a few hundred to a few thousand.

For the geometric network case, we performed experiments by using four examples: the San Diego and Baghdad data sets from Figs 2 and 6, in addition to similar data sets in Davis, California. One of these is a fused density estimator with domain as the road network in downtown Davis, and the other is on the *entire city* of Davis—our largest example in this paper—which has 19000 variables and 25000 constraints in the dual formulation (5). We choose $\lambda$ in a range that progresses from overfitting to underfitting the data. By overfit, we mean that we choose $\lambda$ as small as possible to make the FDE problem still feasible. By underfit, we mean that the fused density estimator is a constant function. We record a dash when a solver does not run to successful completion. All experiments were run on a computer with 8 Gbytes of memory, an Intel processor with four cores at 2.50 GHz and a 64-bit Linux operating system.

From these experiments we see that the augmented Lagrangian method outperforms the interior point on the geometric network examples. The lack of regularity in the matrices $D_1$ and $D_2$, and the large-scale matrix factorizations that are associated with the Newton step, limits this method in comparison with the ADMM. On smaller, well-structured problems, like in the univariate examples, interior point methods are often faster. On these well-structured problems, however, the gain in performance is negligible (of the order of a tenth of a second). In contrast, the speed and versatility of OSQP especially in the context of large, irregular network structure, leads us to recommend the ADMM as the method to solve the fused density estimator problem (5). This supports the suggestion of using the ADMM for trend filtering in Wang *et al.* (2016) and extends their recommendation beyond the $2 \times 2$ grid graph.

## 4. Statistical rates

In this section we prove a squared Hellinger rate of convergence for FDE when the true log-density is of bounded variation. The Hellinger distance $h$ between functions $f$ and $f_0$ on $G$ is defined as

$$h^2(f, f_0) = \frac{1}{2} \int_G (\surd f - \surd f_0)^2 \, \mathrm{d}x,$$

where $\mathrm{d}x$ is the base measure over the edges in the geometric network $G$; in the univariate setting, this is just the Lebesgue measure. The factor of $\frac{1}{2}$ is a convention that ensures that the Hellinger distance is bounded above by 1. The Hellinger distance is a natural choice for quantifying rates of convergence for density estimators because it is tractable for product measures and provides bounds for rates in other metrics (Le Cam, 1973; 2012; Gibbs and Su, 2002). The squared Hellinger risk of an estimator $\tilde{f}$ for $f_0$ is $\mathbb{E}[h^2(\tilde{f}, f_0)]$. The minimax squared Hellinger risk over a set of densities $\mathcal{H}$, for a sample size $n$, is

$$\min_{\tilde{f}} \max_{f \in \mathcal{H}} \mathbb{E}_f[h^2(\tilde{f}, f)].$$

The minimum is over all estimators $\tilde{f}$ which are measurable maps from the sample space of $x_1, \ldots, x_n$ to $\mathcal{H}$.

We find that FDE achieves a rate of convergence in squared Hellinger risk which matches the minimax rate over all univariate densities in $\mathcal{F}$-densities of log-bounded variation, when the underlying geometric network is simply a compact interval. In this sense, univariate FDE has the best possible squared Hellinger rate of convergence over this function class. The rate that we attain is $n^{-2/3}$, and the equivalence of rates is asymptotic. On an arbitrary connected geometric network, minimax rates for density estimation can depend on the network, but our results demonstrate that FDE on a geometric network has squared Hellinger rate at most the univariate minimax rate.

We begin by establishing the minimax rate over the class $\mathcal{F}$, which gives a lower bound on the squared Hellinger rate for FDE. To establish the lower bound, it is sufficient to examine the minimax rate of convergence over a set of densities contained in $\mathcal{F}$. Fixing a constant $C$ and compact interval $I$, we consider the set of functions $g: I \to \mathbb{R}$

$$\mathrm{BV}(C) := \{g : \mathrm{TV}(g) \leqslant C, \|g\|_\infty < C\}.$$

Recall that, for a given radius $\epsilon$, the packing entropy of a set $S$ with respect to a metric $d: S \times S \to \mathbb{R}_+$ is the logarithm of its packing number, the size of the largest collection of points in $S$ which are at least $\epsilon$ separated with respect to the metric $d$. Because $\mathrm{BV}(C)$ is bounded below, the packing entropy of $\mathrm{BV}(C)$ and $\widetilde{\mathrm{BV}}(C) := \{\exp(g) : g \in \mathrm{BV}(C)\}$ are of the same order. From example 6.4 in Yang and Barron (1999), we have that $\mathrm{BV}(C)$ has $L_2$ packing entropy of order $1/\epsilon$. Applying theorem 5 from Yang and Barron (1999) gives the minimax squared Hellinger rate over densities $\{f / \int f : f \in \widetilde{\mathrm{BV}}(C)\}$ as $n^{-2/3}$. In theorem 2, we show that FDE attains the rate of $n^{-2/3}$ over the larger class $\mathcal{F}$. Therefore, the minimax squared Hellinger rate over $\mathcal{F}$ must also equal $n^{-2/3}$, so we have proved the following theorem. For sequences $a_n$ and $b_n$, we write $a_n \asymp b_n$ if $a_n = O(b_n)$ and $b_n = O(a_n)$.

*Theorem 3.* The minimax squared Hellinger rate over $\mathcal{F}$, the set of densities $f$ with $\log(f)$ of bounded variation, is $n^{-2/3}$, i.e.

$$\min_{\tilde{f}} \max_{f \in \mathcal{F}} \mathbb{E}_f[h^2(\tilde{f}, f)] \asymp n^{-2/3}.$$

To prove the FDE rate of convergence for univariate density estimation, we extend techniques that were developed for the theory of $M$-estimators (van de Geer, 2000) and locally adaptive regression splines in Gaussian models (Mammen and van de Geer, 1997). A detailed proof of our main result can be found in the on-line appendix A. This rate bound for FDE is based on novel empirical process bounds for log-densities of bounded variation, and these are used in conjunction with peeling arguments to provide a uniform bound on the Hellinger error. The empirical process bounds in appendix A.3 rely on new Bernstein difference metric covering number bounds for functions of bounded variation, which can be found in appendix B. We extend the FDE rates for the univariate setting to arbitrary geometric networks in Section 4.2; this requires embedding the geometric network onto the real line. This embedding is constructed from the depth-first search algorithm, which was a technique that was used in Padilla *et al.* (2017) for regression over graphs, and is described in appendix A.

The subsections in this section follow this outline. In Section 4.1, we provide a proof sketch of the squared Hellinger rate of convergence for univariate FDE. In the on-line appendix A.3, we detail the lemmas that are used to prove the main result. In Section 4.2, we extend these rate results from the univariate setting to arbitrary geometric networks.

### *4.1. Upper bounds for rate of univariate fused density estimation*

In this subsection we prove a squared Hellinger rate of $n^{-2/3}$ for univariate FDE. Let the geometric network $G$ be a closed interval $[a,b]$ (a single edge connecting nodes $a$ and $b$). Recall the definition of $\mathcal{F}$ as the set of densities $f$ with $\log(f)$ of bounded variation. Let $f_0 \in \mathcal{F}$ be a fixed density on $G$, so that the total variation $\mathrm{TV}\{\log(f_0)\}$ is constant as $n$ increases.

*Theorem 4.* Let $\hat{f}_n$ be the fused density estimator of an independent and identically distributed sample of $n$ points drawn from a univariate density $f_0$. There is an $f_0$-dependent sequence $\lambda_n$ such that $\lambda_n = O_P(n^{-2/3})$, the FDE is well defined, and

$$\mathbb{E}_{f_0}[h^2(\hat{f}_n, f_0)] = O(n^{-2/3}).$$

Combined with the lower bound in theorem 4, this gives that univariate FDE attains the minimax rate over densities in $\mathcal{F}$.

*Proof.* (For a detailed proof see the on-line appendix A.) To control the Hellinger error for FDE, we rely on the fact that the fused density estimator is the minimizer of problem (1). We derive an inequality involving the squared Hellinger distance, an empirical process and fusion penalty terms. This inequality (and in general inequalities serving this purpose; see van de Geer (2000)) is referred to as a basic inequality. To reduce the notation, we introduce the shorthand $\hat{h} = h(\hat{f}_n, f_0)$, $I(f) = \mathrm{TV}\{\log(f)\}$, $\hat{I} = I(\hat{f}_n)$, $I_0 = I(f_0)$ and

$$p_f = \frac{1}{2}\log\left(\frac{f+f_0}{2f_0}\right).$$

We arrive at the following basic inequality by manipulating the optimality condition, $-P_n\{\log(\hat{f}_n)\} + \lambda_n \hat{I} \leqslant -P_n\{\log(f_0)\} + \lambda_n I_0$. In fact, from the definition of FDE we have the stronger condition $-P_n\{\log(\hat{f}_n)\} + \lambda_n \hat{I} \leqslant -P_n\{\log(f)\} + \lambda_n I(f)$ for all $f \in \mathcal{F}$, but the weaker condition will suffice.

*Lemma 3* (basic inequality).

$$\hat{h}^2 \leqslant 16(P_n - P)(p_{\hat{f}_n}) + 4\lambda_n(I_0 - \hat{I}).$$

Squared Hellinger rates now follow from controlling the right-hand side. We do so by considering two cases. When $\hat{h}$ is small, we show that

$$(P_n - P)(p_{\hat{f}_n}) = O_P\{n^{-2/3}(1 + I_0 + \hat{I})\}.$$

From the basic inequality, this gives

$$\begin{aligned}\hat{h}^2 &= O_P\{16n^{-2/3}(1 + I_0 + \hat{I}) + 4\lambda_n(I_0 - \hat{I})\}\\ &= O_P\{4(4n^{-2/3} - \lambda_n)\hat{I} + 4(4n^{-2/3} + \lambda_n)I_0 + 16n^{-2/3}\}.\end{aligned} \tag{6}$$

Excluding details, when $\lambda_n$ is chosen to dominate $4n^{-2/3}$, the first term in equation (6) is negative, so we conclude that $\hat{h}^2 = O_P(\max\{n^{-2/3}, \lambda_n\})$.

The condition 'when $\hat{h}$ is small', and the corresponding control on $(P_n - P)(p_{\hat{f}_n})$, can be formalized in the following theorem.

*Theorem 5.*

$$\sup_{h(f,f_0) \leqslant n^{-1/3}\{1+I(f)+I_0\}} \frac{n^{2/3}|(P_n - P)(p_f)|}{1 + I(f) + I_0} = O_P(1),$$

where the supremum is taken over all $f \in \mathcal{F}$.

When $\hat{h}$ is large, in contrast, we show that $(P_n - P)(p_{\hat{f}_n}) = O_P\{n^{-1/2}\hat{h}^{1/2}(1 + \hat{I} + I_0)^{1/2}\}$. From the basic inequality, this gives

$$\surd n\hat{h}^2 = O_P\{16\hat{h}^{1/2}(1 + I_0 + \hat{I})^{1/2} + 4\surd n\lambda_n(I_0 - \hat{I})\},$$

whence we conclude that $\hat{h}^2 = O_P(\max\{n^{-2/3}, \lambda_n\})$. The next theorem is an analogue to theorem 5 when $\hat{h}$ is large.

*Theorem 6.*

$$\sup_{h > n^{-1/3}\{1+I(f)+I_0\}} \frac{n^{1/2}|(P_n - P)(p_f)|}{h^{1/2}(f, f_0)\{1 + I(f) + I_0\}^{1/2}} = O_P(1),$$

where the supremum is taken over all $f \in \mathcal{F}$.

To summarize our conclusions so far: the squared Hellinger rate is $\max\{n^{2/3}, \lambda_n\}$ when $\lambda_n$ balances the competing terms in equation (6). By choosing

$$\lambda_n = \max\left\{\sup_{h(f,f_0) \leqslant n^{-1/3}\{1+I(f)+I_0\}} \frac{4|(P_n - P)(p_f)|}{1 + I(f) + I_0}, n^{-2/3}\right\},$$

we have a minimal $\lambda_n$ which dominates in equation (6). Furthermore, this choice of $\lambda_n$ satisfies $\lambda_n = O_P(n^{-2/3})$ by theorem 5. We have established a squared Hellinger rate of $n^{-2/3}$ for both the cases of $\hat{h}$ considered. Lastly, this choice of $\lambda_n$ satisfies the condition on $\lambda$ in theorem 2, so the fused density estimator is well defined.

Theorems 5 and 6 are essential components of the proof that was outlined above. Both of these results are new and of independent interest. Their derivation requires the following lemma.

*Lemma 4.* Let $M \in \mathbb{R}$ and $\mathcal{P}_M = \{p_f : 1 + I(f) + I_0 \leqslant M\}$. There is a constant $C$ and choice of $c_1$ such that for all $C_1 \geqslant c_1$ and $\delta \geqslant (M/2)n^{1/3}$

$$\mathbb{P}\left(\sup_{p_f \in \mathcal{P}_M, h(f,f_0) \leqslant \delta} |\surd n(P_n - P)(p_f)| \geqslant 2C_1\surd M\delta^{1/2}\right) \leqslant C\exp\left(-\frac{C_1 M\delta^{-1}}{4C^2}\right).$$

Lemma 4 can be used to prove theorems 5 and 6 by applying the peeling device twice: once each for the parameters $M$ and $\delta$.

The proof of lemma 4 requires three basic ingredients: control of the bracketing entropy of $\mathcal{P}_M$, a uniform bound on $\mathcal{P}_M$ and a relationship dictating how $M$ scales with control of the Hellinger distance. These ingredients have the same motivation as in Mammen and van de Geer (1997), where the authors used a total variation penalty to construct adaptive estimators in the context of regression. Mammen and van de Geer (1997) assumed sub-Gaussian errors and proved bounds on metric entropy for functions of bounded variation. The sub-Gaussian assumption provides local error bounds, and the metric entropy condition bounds the number of sets on which we must control that error. Though similarly motivated, our context is more complicated. To control the $M$ in $\mathcal{P}_M$ with the Hellinger distance, we consider coverings in the Bernstein difference metric instead of the $L_2(P)$ metric. Using the Bernstein difference enables us to achieve the results in lemma 4, but its use requires control of generalized bracketing entropy—bracketing with the Bernstein difference—instead of the usual bracketing entropy with the $L_2(P)$ metric. In addition, the uniform bound that we require is now on the Bernstein difference over $\mathcal{P}_M$.

In the on-line appendix B, we show that the bracketing entropy of $\mathcal{P}_M$, with bracketing radius $\delta$, is of order $M/\delta$. This bracketing entropy result implies generalized bracketing entropy bounds and can be proved similarly to results in monotonic shape-constrained estimation (Van Der Vaart and Wellner, 1996). To achieve the finite sample bounds that are necessary to achieve these rates, Bernstein's inequality is used to provide concentration inequalities that are critical to bounding the basic inequality. With this combination of local error bounds and bracketing rates, we can apply results in the spirit of generic chaining (Talagrand, 2006) to obtain lemma 4.

Lastly, we translate the probabilistic results into bounds on Hellinger risk. In general, one cannot prove expected risk rates from convergence in probability because the tails may not decay sufficiently quickly to give a finite expectation. But, out of the proofs of theorems 5 and 6, we can derive exponential tail bounds for $h^2(\hat{f}_n, f_0)$. This enables us to translate our probabilistic rates into rates on the Hellinger risk; doing so requires some care to apply the rates in theorems 5 and 6 simultaneously. These details are provided in the expanded proof in the on-line appendix A.

### 4.2. Guarantees for connected geometric networks

In the previous sections we proved an $n^{-2/3}$ rate of convergence for univariate fused density estimators. The following theorem extends this result to arbitrary geometric networks.

*Theorem 7.* Let $\hat{f}_n$ be the fused density estimator of an independent and identically distributed sample over a connected geometric network with true density $f_0$. Then there is a choice of $\lambda_n$, dependent on $f_0$, such that $\lambda_n = O_P(n^{-2/3})$ and

$$\mathbb{E}_{f_0}[h^2(\hat{f}_n, f_0)] = O(n^{-2/3}).$$

We prove this theorem by using an embedding lemma, lemma A.11 in the on-line appendix, which states that for any fixed geometric network $G$ there is a measure preserving embedding $\gamma$ of $G$ into $\mathbb{R}$ that preserves densities and Hellinger distances. Furthermore, for any function $g$ on $G$, the (univariate) total variation of the embedded function $g \circ \gamma^{-1}$ never exceeds twice that of the graph-valued total variation. With this lemma in hand, theorem 7 is proved by strategically bounding terms in our analysis by their univariate counterparts. A detailed proof can be found in the on-line supplementary material.

Theorem 7 provides an upper bound on the minimax Hellinger rate for densities of log-bounded variation on geometric networks. Unlike the univariate case, however, we do not have a lower bound as in theorem 4, so we cannot conclude that FDE attains the minimax rate for arbitrary geometric networks. This mirrors similar results found in the Gaussian regression setting; see for example Padilla *et al.* (2017). The minimax squared Hellinger rate for density estimation on geometric networks is at least as small as the univariate rate, and it is reasonable to suspect that the minimax rate on some graphs may be strictly better than the univariate rate. Though the univariate case may seem simple, and hence we might expect it to be easier, the sparse connectivity of the underlying graph in univariate estimation negatively affects its minimax rate of convergence. In some sense this is intuitive. For example, adding cycles to a graph increases the total variation compared with the same graph with the cycles removed. The increased total variation can be seen as applying more shrinkage in the context of estimation, which makes total variation balls smaller and the problem easier. Similarly, tree graphs (graphs without cycles) have more connectivity than the univariate chain graph and larger total variation for a function that is defined on it. This intuition is consistent with the formal results from Hütter and Rigollet (2016) in the regression setting. Although there may be networks for which the FDE and minimax squared Hellinger rates may decrease more quickly than $n^{-2/3}$, we leave that study to future work.

## Acknowledgements

RB was supported in part by US Office of Naval Research grant N00014-17-2372. JS was supported in part by National Science Foundation grant DMS-1712996. We thank Ryan Tibshirani, Roger Wets and Matthias Köppe for helpful conversations.

## References

Arnold, T. B. and Tibshirani, R. J. (2016) Efficient implementations of the generalized lasso dual path algorithm. *J. Computnl Graph. Statist.*, **25**, 1–27.

Beck, A. and Teboulle, M. (2009) Fast gradient-based algorithms for constrained total variation image denoising and deblurring problems. *IEEE Trans. Im. Process.*, **18**, 2419–2434.

Bertsekas, D. P. (2014) *Constrained Optimization and Lagrange Multiplier Methods*. Cambridge: Academic Press.

Birgé, L. and Massart, P. (1997) From model selection to adaptive estimation. In *Festschrift for Lucien le Cam* (ed. D. Pollard), pp. 55–87. Berlin: Springer.

Birgé, L. and Rozenholc, Y. (2006) How many bins should be put in a regular histogram? *Eur. Ser. Appl. Indust. Math. Probab. Statist.*, **10**, 24–45.

Boyd, S., Parikh, N., Chu, E., Pelato, B. and Eckstein, J. (2011) Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundns Trends Mach. Learn.*, **3**, 1–122.

Boyd, S. and Vandenberghe, L. (2004) *Convex Optimization*. New York: Cambridge University Press.

Caron, S. (2018) QPSolvers: wrappers for quadratic programming in Python. (Available from `https://github.com/stephane-caron/qpsolvers`.)

Denby, L. and Mallows, C. (2009) Variations on the histogram. *J. Computnl Graph. Statist.*, **18**, 21–31.

Devroye, L. and Györfi, L. (1985) *Nonparametric Density Estimation: the L1 View*. New York: Wiley.

Donoho, D. L., Johnstone, I. M., Kerkyacharian, G. and Picard, D. (1996) Density estimation by wavelet thresholding. *Ann. Statist.*, **24**, 508–539.

Fithian, W., Sun, D. and Taylor, J. (2014) Optimal inference after model selection. *Preprint arXiv:1410.2597*. University of California at Berkeley, Berkeley.

Freedman, D. and Diaconis, P. (1981) On the histogram as a density estimator: L 2 theory. *Zeits. Wahrsch. Ver. Geb.*, **57**, 453–476.

van de Geer, S. A. (2000) *Empirical Processes in M-estimation*, vol. 6. New York: Cambridge University Press.

Gibbs, A. L. and Su, F. E. (2002) On choosing and bounding probability metrics. *Int. Statist. Rev.*, **70**, 419–435.

Hall, P. and Wand, M. P. (1988) Minimizing L1 distance in nonparametric density estimation. *J. Multiv. Anal.*, **26**, 59–88.

Hütter, J.-C. and Rigollet, P. (2016) Optimal rates for total variation denoising. In *Proc. Conf. Learning Theory* (eds V. Feldman, A. Rakhlin and O. Shamir), pp. 1115–1146.

Kim, S.-J., Koh, K., Boyd, S. and Gorinevsky, D. (2009) $\ell_1$ trend filtering. *SIAM Rev.*, **51**, 339–360.

Koenker, R. and Mizera, I. (2007) Density estimation by total variation regularization. In *Advances in Statistical*

*Modeling and Inference: Essays in Honor of Kjell A Doksum* (ed. V. Nair), pp. 613–633. Singapore: World Scientific Publishing.

Koo, J.-Y. and Kim, W.-C. (1996) Wavelet density estimation by approximation of log-densities. *Statist. Probab. Lett.*, **26**, 271–278.

Koo, J.-Y. and Kooperberg, C. (2000) Logspline density estimation for binned data. *Statist. Probab. Lett.*, **46**, 133–147.

LaFree, G. and Dugan, L. (2007) Introducing the global terrorism database. *Terrorsm Polit. Violnce*, **19**, 181–204.

LeCam, L. (1973) Convergence of estimates under dimensionality restrictions. *Ann. Statist.*, **1**, 38–53.

Le Cam, L. (2012) *Asymptotic Methods in Statistical Decision Theory*. New York: Springer Science and Business Media.

Le Cam, L. and Yang, G. L. (2012) *Asymptotics in Statistics: Some Basic Concepts*. New York: Springer Science and Business Media.

Li, H., Munk, A., Sieling, H. and Walther, G. (2016) The essential histogram. *Preprint arXiv:1612.07216*. Georg-August-Universität Göttingen, Göttingen.

Mammen, E. and van de Geer, S. (1997) Locally adaptive regression splines. *Ann. Statist.*, **25**, 387–413.

Meir, A. and Drton, M. (2017) Tractable post-selection maximum likelihood inference for the lasso. *Preprint arXiv:1705.09417*. University of Washington, Seattle.

Nesterov, Y. and Nemirovskii, A. (1994) *Interior-point Polynomial Algorithms in Convex Programming*, vol. 13. Philadelphia: Society for Industrial and Applied Mathematics.

Padilla, O. H. M. and Scott, J. G. (2015) Nonparametric density estimation by histogram trend filtering. *Preprint arXiv:1509.04348*. University of Texas at Austin, Austin.

Padilla, O. H., Sharpnack, J., Scott, J. G. and Tibshirani, R. J. (2017) The DFS fused lasso: linear-time denoising over general graphs. *J. Mach. Learn. Res.*, **18**, no. 176.

Rockafellar, R. T. (2015) *Convex Analysis*. Princeton: Princeton University Press.

Rockafellar, R. T. and Wets, R. J.-B. (2009) *Variational Analysis*. New York: Springer Science and Business Media.

Royset, J. O. and Wets, R. J. B. (2013) Nonparametric density estimation via exponential epi-eplines: fusion of soft and hard information. *Technical Report*.

Rudin, L. I., Osher, S. and Fatemi, E. (1992) Nonlinear total variation based noise removal algorithms. *Physica* D, **60**, 259–268.

Sardy, S. and Tseng, P. (2010) Density estimation by total variation penalized likelihood driven by the sparsity 1 information criterion. *Scand. J. Statist.*, **37**, 321–337.

Scott, D. W. (1979) On optimal and data-based histograms. *Biometrika*, **66**, 605–610.

Shuman, D. I., Narang, S. K., Frossard, P., Ortega, A. and Vandergheynst, P. (2013) The emerging field of signal processing on graphs: extending high-dimensional data analysis to networks and other irregular domains. *IEEE Signl Process. Mag.*, **30**, no. 3, 83–98.

Silverman, B. W. (1982) On the estimation of a probability density function by the maximum penalized likelihood method. *Ann. Statist.*, **10**, 795–810.

Stellato, B., Banjac, G., Goulart, P., Bemporad, A. and Boyd, S. (2017) OSQP: an operator splitting solver for quadratic programs. *Preprint arXiv:1711.08013*. Massachusetts Institute of Technology, Cambridge.

Sturges, H. A. (1926) The choice of a class interval. *J. Am. Statist. Ass.*, **21**, 65–66.

Talagrand, M. (2006) *The Generic Chaining: Upper and Lower Bounds of Stochastic Processes*. New York: Springer Science and Business Media.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J. and Knight, K. (2005) Sparsity and smoothness via the fused lasso. *J. R. Statist. Soc.* B, **67**, 91–108.

Tibshirani, R. J., Taylor, J. *et al.* (2012) Degrees of freedom in lasso problems. *Ann. Statist.*, **40**, 1198–1232.

Van Der Vaart, A. W. and Wellner, J. A. (1996) *Weak Convergence and Empirical Processes*. New York: Springer.

Wahba, G. (1990) *Spline Models for Observational Data*. Philadelphia: Society for Industrial and Applied Mathematics.

Wang, Y.-X., Sharpnack, J., Smola, A. J. and Tibshirani, R. J. (2016) Trend filtering on graphs. *J. Mach. Learn. Res.*, **17**, 1–41.

Willett, R. M. and Nowak, R. D. (2007) Multiscale Poisson intensity and density estimation. *IEEE Trans. Inform. Theory*, **53**, 3171–3187.

Wright, S. J. (1997) *Primal-dual Interior-point Methods*. New York: Society for Industrial and Applied Mathematics.

Yang, Y. and Barron, A. (1999) Information-theoretic determination of minimax rates of convergence. *Ann. Statist.*, **27**, 1564–1599.

*Supporting information*

Additional ‘supporting information’ may be found in the on-line version of this article:

‘Supplementary material: Fused density estimation: theory and methods’.