Biometrika (2020), **107**, 2, *pp.* 293–310
Printed in Great Britain
doi: 10.1093/biomet/asz071
Advance Access publication 29 January 2020

Adaptive nonparametric regression with the K-nearest neighbour fused lasso

BY OSCAR HERNAN MADRID PADILLA

Department of Statistics, University of California, 520 Portola Plaza, Los Angeles, California 90095, U.S.A.
oscar.madrid@stat.ucla.edu

JAMES SHARPNACK

Department of Statistics, University of California, One Shields Avenue, Davis, California 95616, U.S.A.
jsharpna@ucdavis.edu

YANZHEN CHEN

Department of Information Systems, Business Statistics and Operations Management, Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong
imyanzhen@ust.hk

AND DANIELA M. WITTEN

Department of Statistics, University of Washington, Seattle, Washington 98195, U.S.A.
dwitten@uw.edu

SUMMARY

The fused lasso, also known as total-variation denoising, is a locally adaptive function estimator over a regular grid of design points. In this article, we extend the fused lasso to settings in which the points do not occur on a regular grid, leading to a method for nonparametric regression. This approach, which we call the K-nearest-neighbours fused lasso, involves computing the K-nearest-neighbours graph of the design points and then performing the fused lasso over this graph. We show that this procedure has a number of theoretical advantages over competing methods: specifically, it inherits local adaptivity from its connection to the fused lasso, and it inherits manifold adaptivity from its connection to the K-nearest-neighbours approach. In a simulation study and an application to flu data, we show that excellent results are obtained. For completeness, we also study an estimator that makes use of an ϵ-graph rather than a K-nearest-neighbours graph and contrast it with the K-nearest-neighbours fused lasso.

Some key words: Fused lasso; Local adaptivity; Manifold adaptivity; Nonparametric regression; Total variation.

1. INTRODUCTION

This article considers the nonparametric regression setting in which we have n observations, $(x_1, y_1), \ldots, (x_n, y_n)$, of the pair of random variables $(X, Y) \in \mathcal{X} \times \mathbb{R}$, where $\mathcal{X}$ is a metric space

with metric $d_{\mathcal{X}}$. We assume that the model

$$y_i = f_0(x_i) + \varepsilon_i \quad (i = 1, \ldots, n) \tag{1}$$

holds, where $f_0 : \mathcal{X} \to \mathbb{R}$ is an unknown function that we wish to estimate. This problem arises in many settings, including demographic applications (Petersen et al., 2016a; Sadhanala & Tibshirani, 2018), environmental data analysis (Hengl et al., 2007), image processing (Rudin et al., 1992) and causal inference (Wager & Athey, 2018).

A substantial body of work has dealt with estimating the function f_0 in (1) at the observations $X = x_1, \ldots, x_n$, i.e., denoising, as well as at other values of the random variable X, i.e., prediction. This includes the seminal papers by Duchon (1977), Breiman et al. (1984) and Friedman (1991), as well as more recent work by Petersen et al. (2016a,b) and Sadhanala & Tibshirani (2018). A number of previous papers have focused in particular on manifold adaptivity, i.e., adapting to the dimensionality of the data; these include work on local polynomial regression by Bickel & Li (2007) and Cheng & Wu (2013), K-nearest-neighbours regression by Kpotufe (2011), Gaussian processes by Yang & Tokdar (2015) and Yang & Dunson (2016), and tree-based estimators such as those in Kpotufe (2009) and Kpotufe & Dasgupta (2012). We refer the reader to Györfi et al. (2006) for a detailed survey of other classical nonparametric regression methods. The vast majority of these methods perform well in function classes with variation controlled uniformly throughout the domain, such as Lipschitz and L_2 Sobolev classes. Donoho & Johnstone (1998) and Härdle et al. (2012) generalized this setting by considering functions of bounded variation and Besov classes. In this article, we focus on piecewise-Lipschitz and bounded-variation functions, as these classes can have functions with nonsmooth regions as well as smooth regions (Wang et al., 2016).

Recently, interest has focused on so-called trend filtering (Kim et al., 2009), which seeks to estimate $f_0(\cdot)$ under the assumption that its discrete derivatives are sparse, in a setting in which one has access to an unweighted graph that quantifies the pairwise relationships between the n observations. In particular, the fused lasso, also known as zeroth-order trend filtering or total variation denoising (Rudin et al., 1992; Mammen & van de Geer, 1997; Tibshirani et al., 2005; Wang et al., 2016), solves the optimization problem

$$\underset{\theta \in \mathbb{R}^n}{\text{minimize}} \left\{ \frac{1}{2} \sum_{i=1}^{n} (y_i - \theta_i)^2 + \lambda \sum_{(i,j) \in E} |\theta_i - \theta_j| \right\}, \tag{2}$$

where λ is a nonnegative tuning parameter and $(i, j) \in E$ if and only if there is an edge between the ith and jth observations in the underlying graph. Then $\hat{f}(x_i) = \hat{\theta}_i$. Computational aspects of the fused lasso have been studied extensively in the case of chain graphs (Davies & Kovac, 2001; Johnson, 2013; Barbero & Sra, 2017) and for general graphs (Chambolle & Darbon, 2009; Hoefling, 2010; Chambolle & Pock, 2011; Tibshirani & Taylor, 2011; Landrieu & Obozinski, 2016). Furthermore, the fused lasso is known to have excellent theoretical properties. In one dimension, Mammen & van de Geer (1997) and Tibshirani (2014) showed that the fused lasso attains nearly minimax rates in mean squared error for estimating functions of bounded variation. More recently, also in one dimension, Guntuboyina et al. (2018) and Lin et al. (2017) independently proved that the fused lasso is nearly minimax under the assumption that f_0 is piecewise constant. In grid graphs, Hutter & Rigollet (2016) and Sadhanala et al. (2016, 2017) proved minimax results for the fused lasso when estimating signals of interest in applications of image denoising. In more general graph structures, Padilla et al. (2018) showed that the fused lasso is consistent for denoising problems, provided that the underlying signal has total variation along the graph which when divided by n goes to zero. Other graph models that have been studied in

the literature include tree graphs in Ortelli & van de Geer (2018) and Padilla et al. (2018), and star and Erdős-Rényi graphs in Hutter & Rigollet (2016).

In this paper, we extend the utility of the fused lasso approach by combining it with the K-nearest-neighbours, or K-NN, procedure. The K-NN has been well-studied from theoretical (Stone, 1977; Alamgir et al., 2014; Chaudhuri & Dasgupta, 2014; Von Luxburg et al., 2014), methodological (Dasgupta, 2012; Dasgupta & Kpotufe, 2014; Kontorovich et al., 2016; Singh & Póczos, 2016) and algorithmic (Friedman et al., 1977; Zhang et al., 2012; Dasgupta & Sinha, 2013) perspectives. One key feature of K-NN methods is that they automatically have a finer resolution in regions with a higher density of design points; this is particularly consequential when the underlying density is highly nonuniform. We study the extreme case in which the data are supported over multiple manifolds of mixed intrinsic dimension. An estimator that adapts to this setting is said to achieve manifold adaptivity.

We exploit recent theoretical developments in the fused lasso and the K-NN procedure to derive a single approach that inherits the advantages of both methods. In greater detail, we extend the fused lasso to the general nonparametric setting of (1) by performing a two-step procedure.

Step 1. We construct a K-NN graph by placing an edge between each observation and the K observations to which it is closest in terms of the metric $d_{\mathcal{X}}$.

Step 2. We apply the fused lasso to this K-NN graph.

The resulting K-NN fused lasso estimator appeared in the context of image processing in Elmoataz et al. (2008) and Ferradans et al. (2014), and more recently in an application of graph trend filtering in Wang et al. (2016). The present article is the first to study its theoretical properties. We also consider a variant obtained by replacing the K-NN graph in Step 1 with an ϵ-nearest-neighbour, ϵ-NN, graph, which contains an edge between x_i and x_j only if $d_{\mathcal{X}}(x_i, x_j) < \epsilon$.

The main contributions of this paper are the following.

(i) Local adaptivity. We show that provided f_0 has bounded variation and satisfies an additional condition that generalizes piecewise-Lipschitz continuity, then the mean squared errors of both the K-NN fused lasso estimator and the ϵ-NN fused lasso estimator scale like $n^{-1/d}$, ignoring logarithmic factors; here, $d > 1$ is the dimension of $\mathcal{X}$. In fact, this matches the minimax rate for estimating a two-dimensional Lipschitz function (Györfi et al., 2006), but over a much wider function class.

(ii) Manifold adaptivity. Suppose that the covariates are independent and identically distributed samples from a mixture model $\sum_{l=1}^{\ell} \pi_l^* p_l$, where $p_1, \ldots, p_\ell$ are unknown bounded densities and the weights $\pi_l^* \in [0, 1]$ satisfy $\sum_{l=1}^{\ell} \pi_l^* = 1$. Suppose further that for $l = 1, \ldots, \ell$, the support $\mathcal{X}_l$ of p_l is homeomorphic to $[0, 1]^{d_l} = [0, 1] \times [0, 1] \times \cdots \times [0, 1]$, where $d_l > 1$ is the intrinsic dimension of $\mathcal{X}_l$. We show that under mild conditions, if the restriction of f_0 to $\mathcal{X}_l$ is a function of bounded variation, then the K-NN fused lasso estimator attains the rate $\sum_{l=1}^{\ell} \pi_l^* (\pi_l^* n)^{-1/d_l}$. For intuition about this rate, observe that $\pi_l^* n$ is the expected number of samples from the lth component, and hence $(\pi_l^* n)^{-1/d_l}$ is the expected rate for the lth component. Therefore, our rate is the weighted average of the expected rates for the different components.

2. Methodology

2.1. *The K-NN and ϵ-NN fused lasso estimators*

Both the K-NN and the ϵ-NN fused lasso approaches are simple two-step procedures. The first step involves constructing a graph on the n observations. The K-NN graph, $G_K = (V, E_K)$, has

vertex set $V = \{1, \ldots, n\}$, and its edge set E_K contains the pair (i, j) if and only if x_i is among the K nearest neighbors of x_j, with respect to the metric $d_{\mathcal{X}}$, and vice versa. By contrast, for the ϵ-graph $G_\epsilon = (V, E_\epsilon)$, the pair (i, j) is in E_ϵ if and only if $d_{\mathcal{X}}(x_i, x_j) < \epsilon$.

After constructing the graph, the fused lasso is applied to $y = (y_1, \ldots, y_n)^{\mathrm{T}}$ over the graph G (either G_K or G_ϵ). We can rewrite the fused lasso optimization problem (2) as

$$\hat{\theta} = \underset{\theta \in \mathbb{R}^n}{\arg\min} \left\{ \frac{1}{2} \sum_{i=1}^{n} (y_i - \theta_i)^2 + \lambda \, \| \nabla_G \theta \|_1 \right\}, \tag{3}$$

where $\lambda > 0$ is a tuning parameter and ∇_G is an oriented incidence matrix of G; each row of ∇_G corresponds to an edge in G. For instance, if the kth edge in G connects the ith and jth observations, then

$$(\nabla_G)_{k,l} = \begin{cases} 1, & l = i, \\ -1, & l = j, \\ 0, & \text{otherwise}, \end{cases}$$

and so $(\nabla_G \theta)_k = \theta_i - \theta_j$. This definition of ∇_G implicitly assumes an ordering of the nodes and edges, which may be chosen arbitrarily without loss of generality. In this paper we mostly focus on the setting where $G = G_K$ is the K-NN graph. We also include an analysis of the ϵ-graph, which results from taking $G = G_\epsilon$, as a point of contrast.

Given the estimator $\hat{\theta}$ defined in (3), we predict the response at a new observation $x \in \mathcal{X} \backslash \{x_1, \ldots, x_n\}$ according to

$$\hat{f}(x) = \frac{1}{\sum_{j=1}^{n} k(x_j, x)} \sum_{i=1}^{n} \hat{\theta}_i \, k(x_i, x). \tag{4}$$

In the case of K-NN fused lasso, we take $k(x_i, x) = \mathbb{1}_{\{x_i \in \mathcal{N}_K(x)\}}$, where $\mathcal{N}_K(x)$ is the set of K nearest neighbours of x in the training data. For the ϵ-NN fused lasso, we take $k(x_i, x) = \mathbb{1}_{\{d_{\mathcal{X}}(x_i, x) < \epsilon\}}$. Given a set A, $\mathbb{1}_A(x)$ is the indicator function that equals 1 if $x \in A$ and 0 otherwise. For the ϵ-NN fused lasso estimator, the prediction rule in (4) may not be well-defined if all the training points are farther than ϵ from x. When that is the case, we set $\hat{f}(x)$ to equal the fitted value of the nearest training point.

We construct the K-NN and ϵ-NN graphs using standard Matlab functions such as `knnsearch` and `bsxfun`; this results in a computational complexity of $O(n^2)$. We solve the fused lasso with the parametric max-flow algorithm of Chambolle & Darbon (2009). The procedure is in practice much faster than its worst-case complexity of $O(m\, n^2)$, where m is the number of edges in the graph (Boykov & Kolmogorov, 2004; Chambolle & Darbon, 2009).

In ϵ-NN and K-NN, the values of ϵ and K directly affect the sparsity of the graphs and hence the computational performance of the fused lasso estimators. Corollary 3.23 in Miller et al. (1997) provides an upper bound on the maximum degree of arbitrary K-NN graphs in $\mathbb{R}^d$.

2.2. *Example*

To illustrate the main advantages of the K-NN fused lasso, we construct a simple example. The ability to adapt to the local smoothness of the regression function will be referred to as local adaptivity, and the ability to adapt to the density of the design points will be referred to as manifold adaptivity. The performance gains of the K-NN fused lasso are most pronounced

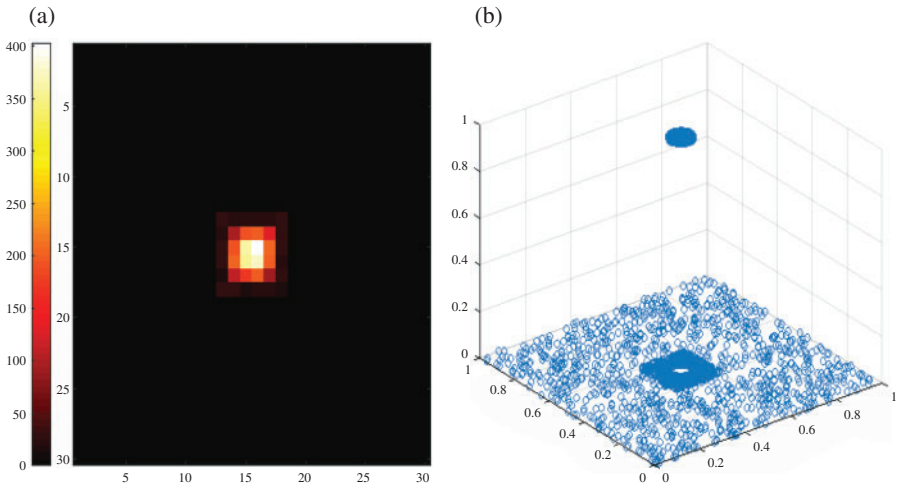

Fig. 1. (a) Heatmap of $n = 5000$ draws from (5). (b) $n = 5000$ samples generated as in (1), with independent and identically distributed $\varepsilon_i \sim N(0, 0.5)$, X having probability density function as in (5), and f_0 as given in (6); the vertical axis corresponds to $f_0(x_i)$ and the other two axes display the two covariates.

when these two effects happen in concert, i.e., when the regression function is less smooth where design points are denser. These properties are manifested in the following example.

We generate $X \in \mathbb{R}^2$ according to the probability density function

$$p(x) = \frac{1}{5}\mathbb{1}_{\{[0,1]^2\setminus[0.4,0.6]^2\}}(x) + \frac{16}{25}\mathbb{1}_{\{[0.45,0.55]^2\}}(x) + \frac{4}{25}\mathbb{1}_{\{[0.4,0.6]^2\setminus[0.45,0.55]^2\}}(x). \quad (5)$$

Thus, p concentrates 64% of its mass in the small interval $[0.45, 0.55]^2$ and 80% of its mass in $[0.4, 0.6]^2$. Figure 1(a) displays a heatmap of $n = 5000$ observations drawn from (5).

We define $f_0 : \mathbb{R}^2 \to \mathbb{R}$ in (1) to be the piecewise-constant function

$$f_0(x) = \mathbb{1}_{\left\{\|x-\frac{1}{2}(1,1)^{\mathrm{T}}\|_2^2 \leqslant \frac{2}{1000}\right\}}(x). \quad (6)$$

We then generate $\{(x_i, y_i)\}_{i=1}^n$ with $n = 5000$ from (1); the regression function is displayed in Fig. 1(b). This simulation study has the following characteristics: the function f_0 in (6) is not Lipschitz, but does have low total variation; and the probability density function p is nonuniform with higher density in the region where f_0 is less smooth.

We compared the following methods in this example:

(i) K-NN fused lasso, with the number of neighbours set to $K = 5$ and the tuning parameter λ chosen to minimize the average mean squared error over 100 Monte Carlo replicates;
(ii) classification and regression trees, CART (Breiman et al., 1984), with the complexity parameter chosen to minimize the average mean squared error over 100 Monte Carlo replicates;
(iii) K-NN regression (see, e.g., Stone, 1977), with the number of neighbours K set to minimize the average mean squared error over 100 Monte Carlo replicates.

The estimated regression functions resulting from these three approaches are displayed in Fig. 2. We see that the K-NN fused lasso can adapt to low-density and high-density regions of the distribution of covariates, as well as to the local structure of the regression function. By contrast,

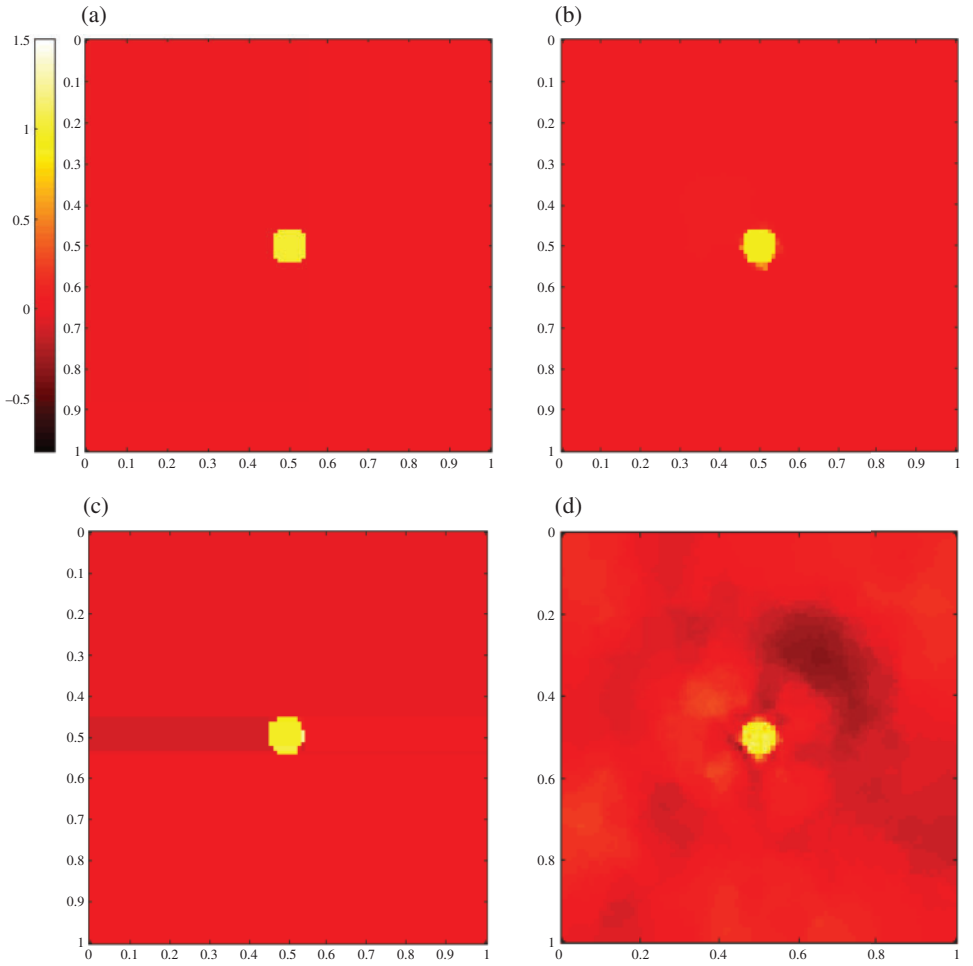

Fig. 2. (a) The function f_0 in (6), evaluated on an evenly spaced grid of size 100×100 in $[0,1]^2$; (b) the estimate of f_0 obtained via the K-NN fused lasso; (c) the estimate of f_0 obtained via CART; (d) the estimate of f_0 obtained via K-NN regression.

the method of Breiman et al. (1984) displays some artifacts due to the binary splits that make up the decision tree, and K-NN regression undersmooths in large areas of the domain.

In practice, we anticipate that the K-NN fused lasso will outperform its competitors when the data are highly concentrated around a low-dimensional manifold, and the regression function is nonsmooth in that region, as in the above example. In our theoretical analysis, we will consider the special case in which the data lie precisely on a low-dimensional manifold or a mixture of low-dimensional manifolds.

3. LOCAL ADAPTIVITY OF THE K-NN AND ϵ-NN FUSED LASSO APPROACHES

3.1. *Assumptions*

We assume that in (1) the elements of $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^{\mathrm{T}}$ are independent and identically distributed zero-mean sub-Gaussian random variables:

$$E(\varepsilon_i) = 0, \quad \mathrm{pr}(|\varepsilon_i| > t) \leqslant C \exp\{-t^2/(2\sigma^2)\} \quad (i = 1, \ldots, n) \quad \text{for all } t > 0, \tag{7}$$

for some positive constants σ and C. Furthermore, we assume that ε is independent of X.

In addition, for a set $A \subset \mathcal{A}$ with $(\mathcal{A}, d_{\mathcal{A}})$ a metric space, we write $B_\epsilon(A) = \{a : \text{there exists } a' \in A \text{ with } d_{\mathcal{A}}(a, a') \leqslant \epsilon\}$. Let ∂A denote the boundary of the set A. The mean squared error of $\hat{\theta}$ is defined as $\|\hat{\theta} - \theta^*\|_n^2 = n^{-1} \sum_{i=1}^n (\hat{\theta}_i - \theta_i^*)^2$. The Euclidean norm of a vector $x \in \mathbb{R}^d$ is denoted by $\|x\|_2 = (x_1^2 + \cdots + x_d^2)^{1/2}$. For $s \in \mathbb{N}$, write $\mathbb{1}_s = (1, \ldots, 1)^{\mathrm{T}} \in \mathbb{R}^s$. In the covariate space $\mathcal{X}$, we consider the Borel σ-algebra $\mathcal{B}(\mathcal{X})$ induced by the metric $d_{\mathcal{X}}$. Let μ be a measure on $\mathcal{B}(\mathcal{X})$. We complement the model in (1) by assuming that the covariates independently satisfy $x_i \sim p(x)$. Thus, p is the probability density function of the distribution of the x_i with respect to the measure space $\{\mathcal{X}, \mathcal{B}(\mathcal{X}), \mu\}$. Note that $\mathcal{X}$ can be a manifold of dimension d in a space of much higher dimension.

We begin by stating assumptions on the distribution of the covariates $p(\cdot)$ and on the metric space $(\mathcal{X}, d_{\mathcal{X}})$. In the theoretical results in Györfi et al. (2006, § 3), it is assumed that p is the probability density function of the uniform distribution on $[0, 1]^d$. In this section we will require only that p be bounded above and below. This condition appeared in the framework for studying K-NN graphs in Von Luxburg et al. (2014) and in the work on density quantization by Alamgir et al. (2014).

Assumption 1. The density p satisfies $0 < p_{\min} < p(x) < p_{\max}$ for all $x \in \mathcal{X}$, where $p_{\min}, p_{\max} \in \mathbb{R}$.

Although we do not require that $\mathcal{X}$ be a Euclidean space, we do require that balls in $\mathcal{X}$ have volume, with respect to μ, that behaves similarly to the Lebesgue measure of balls in $\mathbb{R}^d$. This is expressed in the next assumption, which appeared as part of the definition of a valid region in Von Luxburg et al. (2014, Definition 2).

Assumption 2. The base measure μ in $\mathcal{X}$ satisfies

$$c_{1,d} r^d \leqslant \mu\{B_r(x)\} \leqslant c_{2,d} r^d \quad \text{for all } x \in \mathcal{X},$$

for all $0 < r < r_0$, where r_0, $c_{1,d}$ and $c_{2,d}$ are positive constants and $d \in \mathbb{N} \backslash \{0, 1\}$ is the intrinsic dimension of $\mathcal{X}$.

Next, we make an assumption about the topology of the space $\mathcal{X}$. We require that the space have no holes and be topologically equivalent to $[0, 1]^d$, in the sense that there exists a continuous bijection between $\mathcal{X}$ and $[0, 1]^d$.

Assumption 3. There exists a homeomorphism $h : \mathcal{X} \to [0, 1]^d$, i.e., a continuous bijection with a continuous inverse, such that

$$L_{\min} d_{\mathcal{X}}(x, x') \leqslant \|h(x) - h(x')\|_2 \leqslant L_{\max} d_{\mathcal{X}}(x, x') \quad \text{for all } x, x' \in \mathcal{X},$$

for some positive constants $L_{\min}$ and $L_{\max}$, where $d \in \mathbb{N} \backslash \{0, 1\}$ is the intrinsic dimension of $\mathcal{X}$.

Assumptions 2 and 3 immediately hold if we take $\mathcal{X} = [0, 1]^d$, with $d_{\mathcal{X}}$ the Euclidean distance, h the identity mapping in $[0, 1]^d$, and μ the Lebesgue measure in $[0, 1]^d$. A metric space $(\mathcal{X}, d_{\mathcal{X}})$ that satisfies Assumption 3 is a special case of a differential manifold; the intuition is that the space $\mathcal{X}$ is a chart of the atlas for this differential manifold.

In Assumptions 2 and 3 we assume $d > 1$, since local adaptivity in nonparametric regression is well understood in one dimension. For example, see Tibshirani (2014), Wang et al. (2016), Guntuboyina et al. (2018) and references therein.

We now state conditions on the regression function f_0 defined in (1). The first assumption simply requires bounded variation of the composition of the regression function with the homeomorphism h from Assumption 3.

Assumption 4. The function $g_0 = f_0 \circ h^{-1}$ has bounded variation, i.e., $g_0 \in \text{BV}\{(0,1)^d\}$, and is also bounded. Here $(0,1)^d$ is the interior of $[0,1]^d$, and $\text{BV}\{(0,1)^d\}$ is the class of functions in $(0,1)^d$ of bounded variation. We refer the reader to the Supplementary Material for the explicit construction of the $\text{BV}\{(0,1)^d\}$ class. The function h was defined in Assumption 3.

If $\mathcal{X} = [0,1]^d$ and $h(\cdot)$ is the identity function in $[0,1]^d$, then Assumption 4 simply says that f_0 has bounded variation. However, to allow for more general scenarios, the condition is stated in terms of the function g_0 which has domain in the unit box, whereas the domain of f_0 is the more general set $\mathcal{X}$.

We now recall the definition of a piecewise-Lipschitz function, which induces a much larger class than the set of Lipschitz functions, as it allows for discontinuities.

DEFINITION 1. *Let* $\Omega_\epsilon := [0,1]^d \backslash B_\epsilon(\partial[0,1]^d)$. *We say that a bounded function* $g : [0,1]^d \to \mathbb{R}$ *is piecewise Lipschitz if there exists a set* $\mathcal{S} \subset (0,1)^d$ *that has the following properties.*

(i) *The set* $\mathcal{S}$ *has Lebesgue measure zero.*

(ii) *For some constants* $C_\mathcal{S}, \epsilon_0 > 0$, *we have that* $\mu[h^{-1}\{B_\epsilon(\mathcal{S}) \cup ([0,1]^d \backslash \Omega_\epsilon)\}] \leqslant C_\mathcal{S}\,\epsilon$ *for all* $0 < \epsilon < \epsilon_0$.

(iii) *There exists a positive constant* L_0 *such that if* z *and* z' *belong to the same connected component of* $\Omega_\epsilon \backslash B_\epsilon(\mathcal{S})$, *then* $|g(z) - g(z')| \leqslant L_0 \|z - z'\|_2$.

Roughly speaking, Definition 1 says that g is piecewise Lipschitz if there exists a small set $\mathcal{S}$ that partitions $[0,1]^d$ in such a way that g is Lipschitz within each connected component of the partition. Theorem 2.2.1 in Ziemer (2012) implies that if g is piecewise Lipschitz, then g has bounded variation on any open set within a connected component.

Theorem 1 will require Assumption 5, which is a milder condition on g_0 than piecewise Lipschitz continuity. We now define some notation that is needed in order to introduce Assumption 5.

For $\epsilon > 0$ small enough, we denote by $\mathcal{P}_\epsilon$ a rectangular partition of $(0,1)^d$ induced by $\{0, \epsilon, 2\epsilon, \ldots, \epsilon(\lfloor 1/\epsilon \rfloor - 1), 1\}$, so that all the elements of $\mathcal{P}_\epsilon$ have volume of order ϵ^d. Define $\Omega_{2\epsilon} = [0,1]^d \backslash B_{2\epsilon}(\partial[0,1]^d)$. Then, for a set $\mathcal{S} \subset (0,1)^d$, define

$$\mathcal{P}_{\epsilon,\mathcal{S}} := \{A \cap \Omega_{2\epsilon} \backslash B_{2\epsilon}(\mathcal{S}) : A \in \mathcal{P}_\epsilon,\ A \cap \Omega_{2\epsilon} \backslash B_{2\epsilon}(\mathcal{S}) \neq \emptyset\};$$

this is the partition induced in $\Omega_{2\epsilon} \backslash B_{2\epsilon}(\mathcal{S})$ by the grid $\mathcal{P}_\epsilon$.

For a function g with domain $[0,1]^d$, define

$$S_1(g, \mathcal{P}_{\epsilon,\mathcal{S}}) = \sum_{A \in \mathcal{P}_{\epsilon,\mathcal{S}}} \sup_{z_A \in A} \frac{1}{\epsilon} \int_{B_\epsilon(z_A)} \big|g(z_A) - g(z)\big|\, \mathrm{d}z. \tag{8}$$

If g is piecewise Lipschitz, then $S_1(g, \mathcal{P}_{\epsilon,\mathcal{S}})$ is bounded; see the Supplementary Material.

Next, define

$$S_2(g, \mathcal{P}_{\epsilon,\mathcal{S}}) := \sum_{A \in \mathcal{P}_{\epsilon,\mathcal{S}}} \sup_{z_A \in A} T(g, z_A)\epsilon^d \tag{9}$$

with

$$T(g, z_A) = \sup_{z \in B_\epsilon(z_A)} \sum_{l=1}^{d} \left| \int_{\|z'\|_2 \leqslant \epsilon} \frac{\partial \psi(z'/\epsilon)}{\partial z_l} \left\{ \frac{g(z_A - z') - g(z - z')}{\|z - z_A\|_2 \, \epsilon^d} \right\} \mathrm{d}z' \right|, \tag{10}$$

where ψ is a test function; see the Supplementary Material. Thus (9) is the summation, over evenly sized rectangles of volume ϵ that intersect $\Omega_{2\epsilon} \backslash B_{2\epsilon}(\mathcal{S})$, of the supremum values of the function in (10). The latter, for a function g, can be thought as the average Lipschitz constant near z_A, see the expression within curly braces in (10), weighted by the derivative of a test function. The scaling factor ϵ^d in (10) arises because the integral is taken over a set of measure proportional to ϵ^d.

As with $S_1(g, \mathcal{P}_{\epsilon,\mathcal{S}})$, one can verify that if g is a piecewise-Lipschitz function, then $S_2(g, \mathcal{P}_{\epsilon,\mathcal{S}})$ is bounded.

We now make use of (8) and (9) to state our next condition on $g_0 = f_0 \circ h^{-1}$. This next condition is milder than assuming that g_0 is piecewise Lipschitz; see Definition 1.

Assumption 5. Let $\Omega_\epsilon := [0,1]^d \backslash B_\epsilon(\partial[0,1]^d)$. There exists a set $\mathcal{S} \subset (0,1)^d$ that has the following properties.

(i) The set $\mathcal{S}$ has Lebesgue measure zero.

(ii) For some constants $C_\mathcal{S}, \epsilon_0 > 0$, we have that $\mu(h^{-1}[B_\epsilon(\mathcal{S}) \cup \{(0,1)^d \backslash \Omega_\epsilon\}]) \leqslant C_\mathcal{S}\, \epsilon$ for all $0 < \epsilon < \epsilon_0$.

(iii) The summations $S_1(g_0, \mathcal{P}_{\epsilon,\mathcal{S}})$ and $S_2(g_0, \mathcal{P}_{\epsilon,\mathcal{S}})$ are bounded:

$$\sup_{0<\epsilon<\epsilon_0} \max\{S_1(g_0, \mathcal{P}_{\epsilon,\mathcal{S}}),\, S_2(g_0, \mathcal{P}_{\epsilon,\mathcal{S}})\} < \infty.$$

We refer the reader to the Supplementary Material for a discussion on Assumptions 4 and 5. In particular, we present an example illustrating that the class of piecewise-Lipschitz functions is, in general, different from the class of functions for which Assumptions 4 and 5 hold. However, both classes contain the class of Lipschitz functions, which is obtained by taking $\mathcal{S} = \emptyset$ in Definition 1.

3.2. *Results*

Letting $\theta_i^* = f_0(x_i)$, we express the mean squared errors of the K-NN fused lasso and the ϵ-NN fused lasso in terms of the total variation of θ^* with respect to the K-NN and ϵ-NN graphs.

THEOREM 1. *Let $K \asymp \log^{1+2r} n$ for some $r > 0$. Then under Assumptions 1–3, with an appropriate choice of the tuning parameter λ, the K-NN fused lasso estimator $\hat{\theta}$ satisfies*

$$\|\hat{\theta} - \theta^*\|_n^2 = O_{\mathrm{pr}}\left(\frac{\log^{1+2r} n}{n} + \frac{\log^{1.5+r} n}{n} \|\nabla_{G_K} \theta^*\|_1 \right).$$

This upper bound also holds for the ϵ-NN fused lasso estimator with $\epsilon \asymp \log^{(1+2r)/d} n / n^{1/d}$ if we replace $\|\nabla_{G_K}\theta^*\|_1$ by $\|\nabla_{G_\epsilon}\theta^*\|_1$ and make an appropriate choice of λ.

Clearly, the upper bound in Theorem 1 is a function of $\|\nabla_{G_K}\theta^*\|_1$ or $\|\nabla_{G_\epsilon}\theta^*\|_1$ for the K-NN or ϵ-NN graph, respectively. For the grid graph considered in Sadhanala et al. (2016), $\|\nabla_G \theta^*\|_1 \asymp n^{1-1/d}$, leading to the rate $n^{-1/d}$. However, for a general graph, there is no a priori reason to expect

that $\|\nabla_G \theta^*\|_1 \asymp n^{1-1/d}$. Our next result shows that $\|\nabla_G \theta^*\|_1 \asymp n^{1-1/d}$ for $G \in \{G_K, G_\epsilon\}$, under the assumptions discussed in § 3.1.

THEOREM 2. *Under Assumptions 1–5 or under Assumptions 1–3 and piecewise Lipschitz continuity of $f_0 \circ h^{-1}$, if $K \asymp \log^{1+2r} n$ for some $r > 0$, then for an appropriate choice of the tuning parameter λ, the K-NN fused lasso estimator defined in (3) satisfies*

$$\|\hat{\theta} - \theta^*\|_n^2 = O_{\mathrm{pr}}\left(\frac{\log^\alpha n}{n^{1/d}}\right) \tag{11}$$

with $\alpha = 3r + 5/2 + (2r+1)/d$. Moreover, under Assumptions 1–3 and piecewise Lipschitz continuity of $f_0 \circ h^{-1}$, $\hat{f}$ defined in (4) with the K-NN fused lasso estimator satisfies

$$E_{X\sim p}\left\{\left|f_0(X) - \hat{f}(X)\right|^2\right\} = O_{\mathrm{pr}}\left(\frac{\log^\alpha n}{n^{1/d}}\right). \tag{12}$$

Furthermore, under the same assumptions, (11) and (12) hold for the ϵ-NN fused lasso estimator with $\epsilon \asymp \log^{(1+2r)/d} n / n^{1/d}$.

Theorem 2 indicates that under Assumptions 1–5 or under Assumptions 1–3 and piecewise Lipschitz continuity of $f_0 \circ h^{-1}$, both the K-NN fused lasso and the ϵ-NN fused lasso estimators attain a convergence rate of $n^{-1/d}$, ignoring logarithmic terms. Importantly, Theorem 3.2 of Györfi et al. (2006) shows that in the two-dimensional setting, this rate is actually minimax for estimation of Lipschitz-continuous functions when the design points are uniformly drawn from $[0,1]^2$. Thus, when $d = 2$, both the K-NN fused lasso and the ϵ-NN fused lasso are minimax for estimating functions in the class implied by Assumptions 1–5, and also in the class of piecewise-Lipschitz functions implied by Assumptions 1–3 and Definition 1. In higher dimensions ($d > 2$), by the lower bound in Castro et al. (2005, Proposition 2), we can conclude that both estimators attain nearly minimax rates for estimating piecewise-Lipschitz functions, whereas it is unknown whether the same is true under Assumptions 1–5. A different method, similar in spirit to the method of Breiman et al. (1984), was introduced in Castro et al. (2005, Appendix E). Castro et al. (2005) showed that this approach is also nearly minimax for estimating elements in the class of piecewise-Lipschitz functions, although is unclear whether a computationally feasible implementation of their algorithm is available.

We see from Theorem 2 that both of the fused lasso estimators are locally adaptive, in the sense that they can adapt to the form of the function f_0. Specifically, these estimators do not require knowledge of the set $\mathcal{S}$ in Assumption 5 or Definition 1. This is similar in spirit to the one-dimensional fused lasso, which does not require knowledge of the breakpoints when estimating a piecewise-Lipschitz function.

There is, however, an important difference in the applicability of Theorem 2 to the K-NN fused lasso and to the ϵ-NN fused lasso. To attain the rate in Theorem 2, the ϵ-NN fused lasso requires knowledge of the dimension d, since this quantity appears in the rate of decay of ϵ; but in practice the value of d may not be clear. For instance, suppose that $\mathcal{X} = [0,1]^2 \times \{0\}$; this is a subset of $[0,1]^3$, but it is homeomorphic to $[0,1]^2$, so $d = 2$. If d is unknown, then it can be challenging to choose ϵ for the ϵ-NN fused lasso. By contrast, the choice of K in the K-NN fused lasso involves only the sample size n. Consequently, local adaptivity of the K-NN fused lasso may be much easier to achieve in practice.

4. Manifold adaptivity of the K-NN fused lasso

In this section, we allow the observations $\{(x_i, y_i)\}_{i=1}^n$ to be drawn from a mixture distribution in which each mixture component satisfies the assumptions in § 3. Under these assumptions, we show that the K-NN fused lasso estimator can still achieve a desirable rate.

We assume

$$\begin{aligned} y_i &= \theta_i^* + \varepsilon_i \quad (i = 1, \ldots, n), \\ \theta_i^* &= f_{0,z_i}(x_i), \\ x_i &\sim p_{z_i}(x), \\ \mathrm{pr}(z_i = l) &\sim \pi_l^* \quad (l = 1, \ldots, \ell), \end{aligned} \tag{13}$$

where ε satisfies (7), $\pi_l^* \in [0, 1]$ with $\sum_{l=1}^{\ell} \pi_l^* = 1$, p_l is a density with support $\mathcal{X}_l \subset \mathcal{X}$, $f_{0,l} : \mathcal{X}_l \to \mathbb{R}$, and $\{\mathcal{X}_l\}_{l=1,\ldots,\ell}$ is a collection of subsets of $\mathcal{X}$. For simplicity, we will assume that $\mathcal{X} \subset \mathbb{R}^d$ for some $d > 1$ and that $d_{\mathcal{X}}$ is the Euclidean distance. In (13), the observed data are $\{(x_i, y_i)\}_{i=1}^n$. The remaining ingredients in (13) are either latent or unknown.

We further assume that each set $\mathcal{X}_l$ is homeomorphic to a Euclidean box of dimension depending on l, as follows.

Assumption 6. For $l = 1, \ldots, \ell$, the set $\mathcal{X}_l$ satisfies Assumptions 1–3 with metric given by $d_{\mathcal{X}}$, dimension $d_l \in \mathbb{N}\backslash\{0, 1\}$, and μ equal to some measure μ_l. In addition, the following hold.

(i) There exists a positive constant $\tilde{c}_l$ such that the set $\partial X_l = \bigcup_{l' \neq l} \mathcal{X}_{l'} \cap \mathcal{X}_l$ satisfies

$$\mu_l\{B_\epsilon(\partial X_l) \textstyle\bigcap X_l\} \leqslant \tilde{c}_l \epsilon \tag{14}$$

for any small enough $\epsilon > 0$.

(ii) There exists a positive constant r_l such that for any $x \in \mathcal{X}_l$, either

$$\inf_{x'' \in \partial X_l} d_{\mathcal{X}}(x, x'') < d_{\mathcal{X}}(x, x') \quad \text{for all } x' \in \mathcal{X}\backslash\mathcal{X}_l \tag{15}$$

or $B_\epsilon(x) \subset \mathcal{X}_l$ for all $\epsilon < r_l$.

The constraints implied by Assumption 6 are very natural. Inequality (14) states that the intersections of the manifolds $\mathcal{X}_1, \ldots, \mathcal{X}_\ell$ are small. To put this into perspective, if the extrinsic space ($\mathcal{X}$) were $[0, 1]^d$ with Lebesgue measure, then balls of radius of ϵ would have measure ϵ^d, which is less than ϵ for all $d > 1$, and the set $B_\epsilon(\partial[0, 1]^d) \cap [0, 1]^d$ would have measure that scales like ϵ, which is the same scaling as in (14). Furthermore, (15) holds if $\mathcal{X}_1, \ldots, \mathcal{X}_\ell$ are compact and convex subsets of $\mathbb{R}^d$ whose interiors are disjoint.

We are now ready to extend Theorem 2 to the framework described in this section.

Theorem 3. *Suppose the data are generated as in* (13) *and that Assumption* 6 *holds. Suppose also that the functions $f_{0,1}, \ldots, f_{0,\ell}$ either satisfy Assumptions* 4 *and* 5 *or are piecewise Lipschitz in the domain $\mathcal{X}_l$. Then for an appropriate choice of the tuning parameter λ, the K-NN fused lasso estimator defined in* (3) *satisfies*

$$\|\hat{\theta} - \theta^*\|_n^2 = O_{\mathrm{pr}}\left\{\mathrm{poly}(\log n) \sum_{l=1}^{\ell} \frac{\pi_l^*}{(\pi_l^* n)^{1/d_l}}\right\},$$

provided that $n \min\{\pi_l^* : l \in [\ell]\} \geqslant c_0 n^{r_0}$ *and* $K \asymp \log^{1+2r} n$ *for some constants* $c_0, r_0, r > 0$, *where* poly$(\cdot)$ *is a polynomial function. Here, the* π_l^* *are allowed to change with* n.

When $d_l = d$ for all $l \in [\ell]$ in Theorem 3, we obtain, ignoring logarithmic factors, the rate $n^{-1/d}$, which is minimax when the functions $f_{0,l}$ are piecewise Lipschitz. The rate is also minimax when $d = 2$ and the functions $f_{0,l}$ satisfy Assumptions 4 and 5. In addition, our rates can be compared with those in the existing literature on manifold adaptivity. Specifically, when $d = 2$, the rate $n^{-1/2}$ is attained by local polynomial regression (Bickel & Li, 2007) and Gaussian process regression (Yang & Dunson, 2016) for the class of differentiable functions with bounded partial derivatives, and by K-NN regression for Lipschitz functions (Kpotufe, 2011). In higher dimensions, the methods of Bickel & Li (2007), Yang & Dunson (2016) and Kpotufe (2011) attain better rates than $n^{-1/d}$ on smaller classes of functions that do not allow for discontinuities.

Finally, we refer the reader to the Supplementary Material for an example suggesting that the ϵ-NN fused lasso estimator may not be manifold adaptive.

5. EXPERIMENTS

5.1. *Simulated data*

Throughout this section, we take $d_{\mathcal{X}}$ to be Euclidean distance. We compare the following approaches:

(i) the ϵ-NN fused lasso, with ϵ held fixed and λ treated as a tuning parameter;
(ii) the K-NN fused lasso, with K held fixed and λ treated as a tuning parameter;
(iii) CART (Breiman et al., 1984), implemented in the `R` (R Development Core Team, 2020) package `rpart`, with the complexity parameter treated as a tuning parameter;
(iv) multivariate adaptive regression splines, MARS (Friedman, 1991), implemented in the `R` package `earth`, with the penalty parameter treated as a tuning parameter;
(v) random forests, RF (Breiman, 2001), implemented in the `R` package `randomForest`, with the number of trees fixed at 800 and with the minimum size of each terminal node treated as a tuning parameter;
(vi) K-NN regression (e.g., Stone, 1977), implemented in Matlab using the function `knnsearch`, with K treated as a tuning parameter.

We evaluate each method's performance in terms of the mean squared error, as defined in § 3.1. Specifically, we apply each method to 150 Monte Carlo datasets with a range of tuning parameter values. For each method, we then identify the tuning parameter value that leads to the smallest average mean squared error over the 150 datasets. We refer to this smallest average mean squared error as the optimized mean squared error in what follows.

In our first two scenarios we consider $d = 2$ covariates and let the sample size n vary.

Scenario 1. The function $f_0 : [0,1]^2 \to \mathbb{R}$ is piecewise constant,

$$f_0(x) = \mathbb{1}_{\left\{t \in \mathbb{R}^2 : \|t - \frac{3}{4}(1,1)^{\mathrm{T}}\|_2 < \|t - \frac{1}{2}(1,1)^{\mathrm{T}}\|_2\right\}}(x).$$

The covariates are drawn from a uniform distribution on $[0,1]^2$. The data are generated as in (1) with $N(0,1)$ errors.

Scenario 2. The function $f_0 : [0,1]^2 \to \mathbb{R}$ is as in (6), with generative density for X as in (5). The data are generated as in (1) with $N(0,1)$ errors.

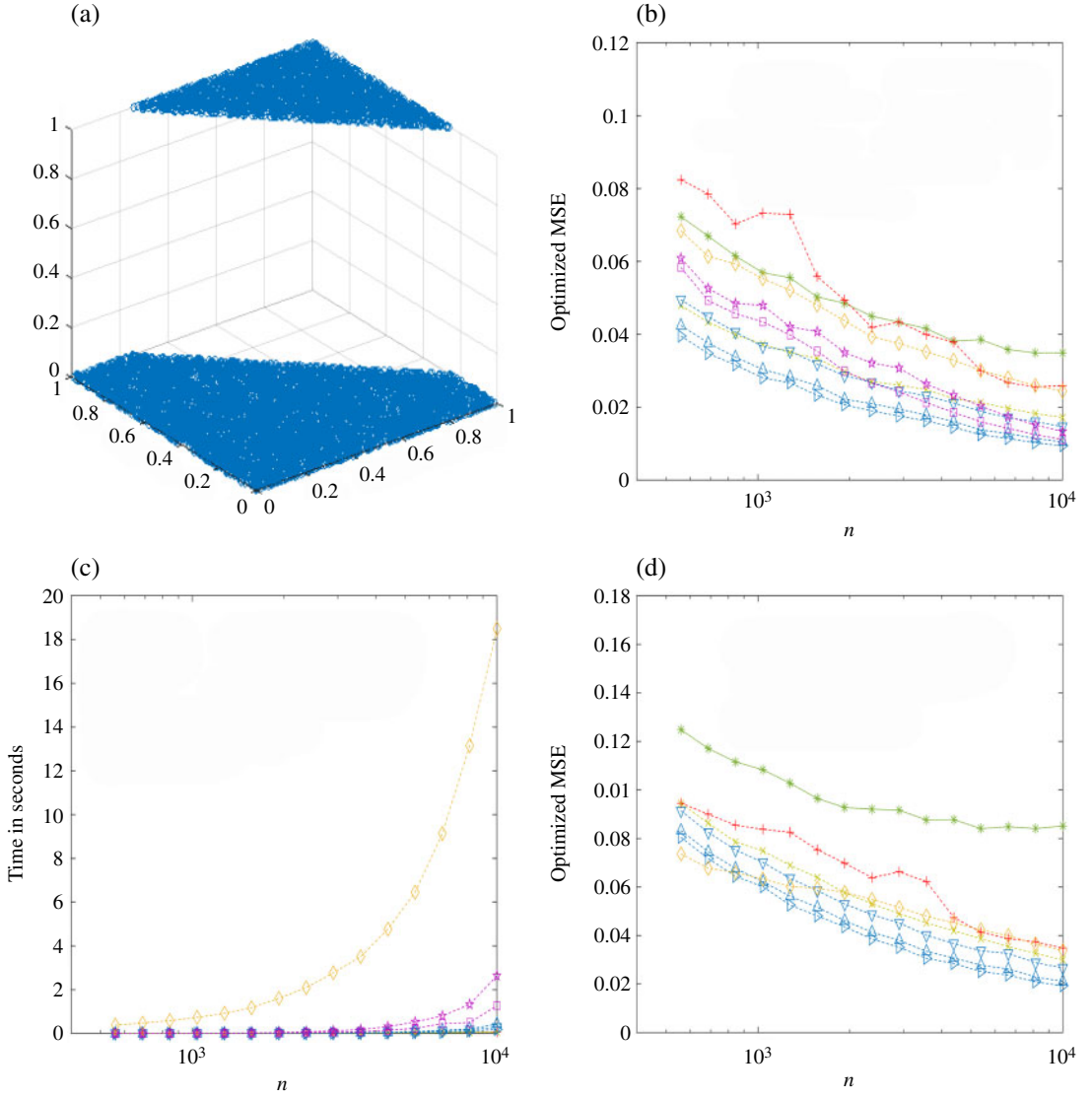

Fig. 3. (a) Scatterplot of data generated under Scenario 1; the vertical axis displays $f_0(x_i)$ and the other two axes display the two covariates. (b) Optimized mean squared error, MSE, averaged over 150 Monte Carlo simulations, of competing methods under Scenario 1; here $\epsilon_1 = (3/4)(\log n/n)^{1/2}$ and $\epsilon_2 = (\log n/n)^{1/2}$. (c) Computational time, in seconds for Scenario 1, averaged over 150 Monte Carlo simulations. (d) Optimized mean squared error, averaged over 150 Monte Carlo simulations, of competing methods under Scenario 2. The methods under comparison are MARS (green solid line and asterisks), CART (red dashed line and plus signs), K-NN (olive dashed line and crosses), 3-NN fused lasso (blue dashed line and downward-pointing triangles), 4-NN fused lasso (blue dashed line and upward-pointing triangles), 5-NN fused lasso (blue dashed line and rightward-pointing triangles), ϵ_1-NN fused lasso (purple dashed line and squares), ϵ_2-NN fused lasso (purple dashed line and stars), and RF (gold dashed line and diamonds).

Data generated under Scenario 1 are displayed in Fig. 3(a). Data generated under Scenario 2 are displayed in Fig. 1(b).

Figure 3(b) and (d) display the optimized mean squared error as a function of the sample size for Scenarios 1 and 2, respectively. The K-NN fused lasso gives the best results in both scenarios. The ϵ-NN fused lasso performs a little worse than K-NN fused lasso in Scenario 1, and very poorly in Scenario 2; the results are not shown.

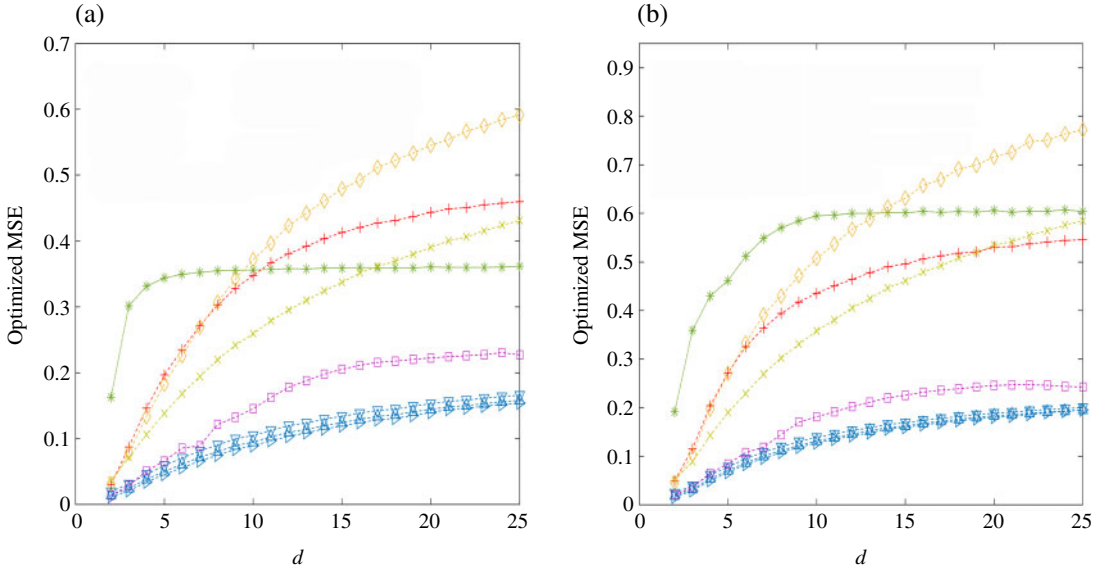

Fig. 4. Optimized mean squared error (MSE), averaged over 150 Monte Carlo simulations, for (a) Scenario 3 and (b) Scenario 4. In both scenarios, ϵ_1 is chosen to be the largest value such that the total number of edges in the graph G_{ϵ_1} is at most 50 000. The methods under comparison are MARS (green solid line and asterisks), CART (red dashed line and plus signs), K-NN (olive dashed line and crosses), 3-NN fused lasso (blue dashed line and downward-pointing triangles), 4-NN fused lasso (blue dashed line and upward-pointing triangles), 5-NN fused lasso (blue dashed line and rightward-pointing triangles), ϵ_1-NN fused lasso (purple dashed line and squares), and RF (gold dashed line and diamonds).

Timing results for all approaches under Scenario 1 are given in Fig. 3(c). For all methods, the times reported are averaged over a range of tuning parameter values. For instance, for the K-NN fused lasso, we fix K and compute the time for different choices of λ; we then report the average of those times.

For the next two scenarios, we consider $n = 8000$ and values of d in $\{2, \ldots, 25\}$.

Scenario 3. The function $f_0 : [0,1]^d \to \mathbb{R}$ is defined as

$$f_0(x) = \begin{cases} 1, & \|x - \frac{1}{4}\mathbb{1}_d\|_2 < \|x - \frac{3}{4}\mathbb{1}_d\|_2, \\ -1, & \text{otherwise}, \end{cases}$$

and the density p is uniform in $[0,1]^d$. The data are generated as in (1) with independent $\varepsilon_i \sim N(0, 0.3)$.

Scenario 4. The function $f_0 : [0,1]^d \to \mathbb{R}$ is defined as

$$f_0(x) = \begin{cases} 2, & \|x - q_1\|_2 < \min\{\|x - q_2\|_2, \|x - q_3\|_2, \|x - q_4\|_2\}, \\ 1, & \|x - q_2\|_2 < \min\{\|x - q_1\|_2, \|x - q_3\|_2, \|x - q_4\|_2\}, \\ 0, & \|x - q_3\|_2 < \min\{\|x - q_1\|_2, \|x - q_2\|_2, \|x - q_4\|_2\}, \\ -1, & \text{otherwise}, \end{cases}$$

where $q_1 = (\frac{1}{4}\mathbb{1}^{\mathrm{T}}_{\lfloor d/2 \rfloor}, \frac{1}{2}\mathbb{1}^{\mathrm{T}}_{d-\lfloor d/2 \rfloor})^{\mathrm{T}}$, $q_2 = (\frac{1}{2}\mathbb{1}^{\mathrm{T}}_{\lfloor d/2 \rfloor}, \frac{1}{4}\mathbb{1}^{\mathrm{T}}_{d-\lfloor d/2 \rfloor})^{\mathrm{T}}$, $q_3 = (\frac{3}{4}\mathbb{1}^{\mathrm{T}}_{\lfloor d/2 \rfloor}, \frac{1}{2}\mathbb{1}^{\mathrm{T}}_{d-\lfloor d/2 \rfloor})^{\mathrm{T}}$ and $q_4 = (\frac{1}{2}\mathbb{1}^{\mathrm{T}}_{\lfloor d/2 \rfloor}, \frac{3}{4}\mathbb{1}^{\mathrm{T}}_{d-\lfloor d/2 \rfloor})^{\mathrm{T}}$. Once again, the generative density for X is uniform in $[0,1]^d$. The data are generated as in (1) with independent $\varepsilon_i \sim N(0, 0.3)$.

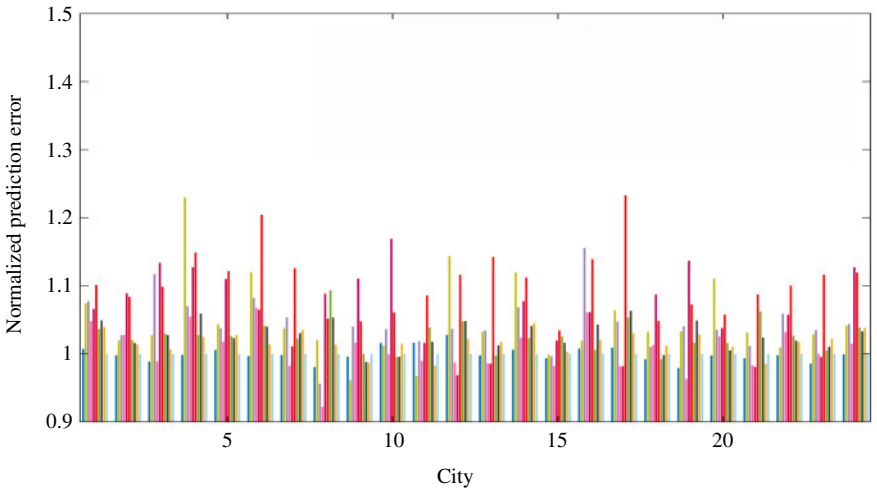

Fig. 5. Results for the flu data; the normalized prediction error was obtained by dividing each method's test set prediction error by the test set prediction error of the 7-NN fused lasso. The methods under comparison are the 5-NN fused lasso (dark blue), neural networks (olive green), ϵ_1-NN fused lasso (purple), ϵ_2-NN fused lasso (light pink), ϵ_3-NN fused lasso (dark pink), CART (red), MARS (green), TPS (dark grey), RF (gold) and 7-NN fused lasso (light blue).

The optimized mean squared error for each approach is displayed in Fig. 4. When d is small, most methods perform well; however, as d increases, the performance of the competing methods quickly deteriorates, whereas the K-NN fused lasso continues to perform well.

5.2. *Flu data*

The data consist of flu activity and atmospheric conditions between 1 January 2003 and 31 December 2009 in different cities across the U.S. state of Texas. Our data-use agreement does not permit dissemination of the flu activity data, which come from medical records. The atmospheric conditions, which include temperature and air quality, can be obtained directly from `http://wonder.cdc.gov/`. Using the number of flu-related doctor's office visits as the dependent variable, we fit a separate nonparametric regression model to each of 24 cities; each day is treated as a separate observation, so that the number of samples is $n = 2556$ in each city. Five independent variables are included in the regression: maximum and average observed concentrations of particulate matter, maximum and minimum temperatures, and day of the year. All variables are scaled to lie in $[0, 1]$. We performed 50 75%/25% splits of the data into a training set and a test set. All models were fitted on the training data, using five-fold cross-validation to select tuning parameter values. Then prediction performance was evaluated on the test set.

We apply the K-NN fused lasso with $K \in \{5, 7\}$ and the ϵ-NN fused lasso with $\epsilon = j/n^{1/d}$ for $j \in \{1, 2, 3\}$, which is motivated by Theorem 2, and with larger choices of ϵ, leading to worse performance. We also fit neural networks (Hagan et al., 1996; implemented in Matlab using the functions `newfit` and `train`), thin plate splines (TPS, Duchon, 1977; implemented using the `R` package `fields`), and MARS, CART and RF as described in § 5.1.

The average test set prediction error across the 50 test sets is displayed in Fig. 5. It can be seen that the K-NN fused lasso and the ϵ-NN fused lasso have the best performances. In particular, the K-NN fused lasso performs best in 13 out the 24 cities, and second best in 6 cities. In 8 of the 24 cities, the ϵ-NN fused lasso performs best.

We contend that the K-NN fused lasso achieves superior performance because it adapts to heterogeneity in the density of design points p, i.e., manifold adaptivity, and adapts to heterogeneity in the smoothness of the regression function f_0, i.e., local adaptivity. In our theoretical

results, we have substantiated this contention through prediction error rate bounds for a large class of regression functions of heterogeneous smoothness and a large class of underlying measures with heterogeneous intrinsic dimensionality. Our experiments demonstrate that these theoretical advantages translate into practical performance gains.

ACKNOWLEDGEMENT

Sharpnack was partially supported by the U.S. National Science Foundation. Witten was partially supported by the U.S. National Institutes of Health, a National Science Foundation CAREER Award, and a Simons Investigator Award in Mathematical Modeling of Living Systems.

SUPPLEMENTARY MATERIAL

Supplementary material available at *Biometrika* online contains proofs of all the theoretical results.

REFERENCES

ALAMGIR, M., LUGOSI, G. & LUXBURG, U. (2014). Density-preserving quantization with application to graph downsampling. *Proc. Mach. Learn. Res.* **35**, 543–59. Proceedings of the 27th Annual Conference on Learning Theory.

BARBERO, Á. & SRA, S. (2017). Modular proximal optimization for multidimensional total-variation regularization. *arXiv:* 1411.0589v3.

BICKEL, P. J. & LI, B. (2007). Local polynomial regression on unknown manifolds. In *Complex Datasets and Inverse Problems*. Beachwood, Ohio: Institute of Mathematical Statistics, pp. 177–86.

BOYKOV, Y. & KOLMOGOROV, V. (2004). An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *IEEE Trans. Pat. Anal. Mach. Intel.* **26**, 1124–37.

BREIMAN, L. (2001). Random forests. *Mach. Learn.* **45**, 5–32.

BREIMAN, L., FRIEDMAN, J., STONE, C. J. & OLSHEN, R. A. (1984). *Classification and Regression Trees*. Boca Raton, Florida: CRC Press.

CASTRO, R. M., WILLETT, R. & NOWAK, R. (2005). Faster rates in regression via active learning. In *Proc. 18th Int. Conf. Neural Information Processing Systems.* pp. 179–86.

CHAMBOLLE, A. & DARBON, J. (2009). On total variation minimization and surface evolution using parametric maximum flows. *Int. J. Comp. Vis.* **84**, 288–307.

CHAMBOLLE, A. & POCK, T. (2011). A first-order primal-dual algorithm for convex problems with applications to imaging. *J. Math. Imag. Vis.* **40**, 120–45.

CHAUDHURI, K. & DASGUPTA, S. (2014). Rates of convergence for nearest neighbor classification. In *Proc. 27th Int. Conf. Neural Information Processing Systems (NIPS'14)*, vol. 2. Cambridge, Massachusetts: MIT Press, pp. 3437–45.

CHENG, M.-Y. & WU, H.-T. (2013). Local linear regression on manifolds and its geometric interpretation. *J. Am. Statist. Assoc.* **108**, 1421–34.

DASGUPTA, S. (2012). Consistency of nearest neighbor classification under selective sampling. *Proc. Mach. Learn. Res.* **23**, 18.1–15. Proceedings of the 25th Annual Conference on Learning Theory.

DASGUPTA, S. & KPOTUFE, S. (2014). Optimal rates for k-NN density and mode estimation. In *Advances in Neural Information Processing Systems 27 (NIPS'14)*. San Diego, California: Neural Information Processing Systems Foundation, pp. 2555–63.

DASGUPTA, S. & SINHA, K. (2013). Randomized partition trees for exact nearest neighbor search. *JMLR Workshop Conf. Proc.* **30**, 317–37.

DAVIES, P. L. & KOVAC, A. (2001). Local extremes, runs, strings and multiresolution. *Ann. Statist.* **29**, 1–65.

DONOHO, D. L. & JOHNSTONE, I. M. (1998). Minimax estimation via wavelet shrinkage. *Ann. Statist.* **26**, 879–921.

DUCHON, J. (1977). Splines minimizing rotation-invariant semi-norms in Sobolev spaces. In *Constructive Theory of Functions of Several Variables*. Berlin: Springer, pp. 85–100.

ELMOATAZ, A., LEZORAY, O. & BOUGLEUX, S. (2008). Nonlocal discrete regularization on weighted graphs: A framework for image and manifold processing. *IEEE Trans. Image Proces.* **17**, 1047–60.

FERRADANS, S., PAPADAKIS, N., PEYRÉ, G. & AUJOL, J.-F. (2014). Regularized discrete optimal transport. *SIAM J. Imag. Sci.* **7**, 1853–82.

FRIEDMAN, J. H. (1991). Multivariate adaptive regression splines. *Ann. Statist.* **19**, 1–67.

Friedman, J. H., Bentley, J. L. & Finkel, R. A. (1977). An algorithm for finding best matches in logarithmic expected time. *ACM Trans. Math. Software* **3**, 209–26.

Guntuboyina, A., Lieu, D., Chatterjee, S. & Sen, B. (2018). Adaptive risk bounds in univariate total variation and trend filtering. *arXiv:* 1702.05113v2.

Györfi, L., Kohler, M., Krzyzak, A. & Walk, H. (2006). *A Distribution-Free Theory of Nonparametric Regression*. New York: Springer.

Hagan, M. T., Demuth, H. B., Beale, M. H. & De Jesús, O. (1996). *Neural Network Design*. Boston: PWS Publishing Co.

Härdle, W., Kerkyacharian, G., Picard, D. & Tsybakov, A. (2012). *Wavelets, Approximation, and Statistical Applications*, vol. 129 of *Lecture Notes in Statistics*. New York: Springer.

Hengl, T., Heuvelink, G. B. & Rossiter, D. G. (2007). About regression-kriging: From equations to case studies. *Comp. Geosci.* **33**, 1301–15.

Hoefling, H. (2010). A path algorithm for the fused lasso signal approximator. *J. Comp. Graph. Statist.* **19**, 984–1006.

Hutter, J.-C. & Rigollet, P. (2016). Optimal rates for total variation denoising. *Proc. Mach. Learn. Res.* **29**, 1115–46. 29th Annual Conference on Learning Theory.

Johnson, N. (2013). A dynamic programming algorithm for the fused lasso and l_0-segmentation. *J. Comp. Graph. Statist.* **22**, 246–60.

Kim, S.-J., Koh, K., Boyd, S. & Gorinevsky, D. (2009). ℓ_1 trend filtering. *SIAM Rev.* **51**, 339–60.

Kontorovich, A., Sabato, S. & Urner, R. (2016). Active nearest-neighbor learning in metric spaces. In *Proc. 30th Int. Conf. Neural Information Processing Systems (NIPS'16)*. New York: Curran Associates, pp. 856–64.

Kpotufe, S. (2009). Escaping the curse of dimensionality with a tree-based regressor. *arXiv:* 0902.3453.

Kpotufe, S. (2011). k-NN regression adapts to local intrinsic dimension. In *Proc. 24th Int. Conf. Neural Information Processing Systems (NIPS'11)*. New York: Curran Associates, pp. 729–37.

Kpotufe, S. & Dasgupta, S. (2012). A tree-based regressor that adapts to intrinsic dimension. *J. Comp. Syst. Sci.* **78**, 1496–515.

Landrieu, L. & Obozinski, G. (2016). Cut pursuit: Fast algorithms to learn piecewise constant functions on general weighted graphs. In Proc. 19th International Conference on Artificial Intelligence and Statistics, vol. 51, 1384–93. Cadiz, Spain: PMLR.

Lin, K., Sharpnack, J. L., Rinaldo, A. & Tibshirani, R. J. (2017). A sharp error analysis for the fused lasso, with application to approximate changepoint screening. In *Proc. 31st Int. Conf. Neural Information Processing Systems (NIPS'17)*. New York: Curran Associates, pp. 6887–96.

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

Miller, G. L., Teng, S.-H., Thurston, W. & Vavasis, S. A. (1997). Separators for sphere-packings and nearest neighbor graphs. *J. Assoc. Comp. Mach* **44**, 1–29.

Ortelli, F. & van de Geer, S. (2018). On the total variation regularized estimator over a class of tree graphs. *Electron. J. Statist.* **12**, 4517–70.

Padilla, O. H. M., Scott, J. G., Sharpnack, J. & Tibshirani, R. J. (2018). The DFS fused lasso: Linear-time denoising over general graphs. *J. Comp. Graph. Statist.* **18**, 1–36.

Petersen, A., Simon, N. & Witten, D. (2016a). Convex regression with interpretable sharp partitions. *J. Comp. Graph. Statist.* **17**, 3240–70.

Petersen, A., Witten, D. & Simon, N. (2016b). Fused lasso additive model. *J. Comp. Graph. Statist.* **25**, 1005–25.

R Development Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. http://www.R-project.org.

Rudin, L., Osher, S. & Faterni, E. (1992). Nonlinear total variation based noise removal algorithms. *Physica* D **60**, 259–68.

Sadhanala, V. & Tibshirani, R. J. (2018). Additive models with trend filtering. *arXiv:* 1702.05037v4.

Sadhanala, V., Wang, Y.-X., Sharpnack, J. L. & Tibshirani, R. J. (2017). Higher-order total variation classes on grids: Minimax theory and trend filtering methods. In *Proc. 31st Int. Conf. Neural Information Processing Systems (NIPS'17)*. New York: Curran Associates, pp. 5796–806.

Sadhanala, V., Wang, Y.-X. & Tibshirani, R. J. (2016). Total variation classes beyond 1d: Minimax rates, and the limitations of linear smoothers. In *Proc. 30th Int. Conf. Neural Information Processing Systems (NIPS'16)*. New York: Curran Associates, pp. 3513–21.

Singh, S. & Póczos, B. (2016). Analysis of k-nearest neighbor distances with application to entropy estimation. *arXiv:* 1603.08578v2.

Stone, C. J. (1977). Consistent nonparametric regression. *Ann. Statist.* **5**, 595–620.

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

Tibshirani, R. J. (2014). Adaptive piecewise polynomial estimation via trend filtering. *Ann. Statist.* **42**, 285–323.

Tibshirani, R. J. & Taylor, J. (2011). The solution path of the generalized lasso. *Ann. Statist.* **39**, 1335–71.

Von Luxburg, U., Radl, A. & Hein, M. (2014). Hitting and commute times in large graphs are often misleading. *J. Mach. Learn. Res.* **15**, 1751–98.

WAGER, S. & ATHEY, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *J. Am. Statist. Assoc.* **113**, 1228–42.
WANG, Y.-X., SHARPNACK, J., SMOLA, A. & TIBSHIRANI, R. J. (2016). Trend filtering on graphs. *J. Mach. Learn. Res.* **17**, 1–41.
YANG, Y. & DUNSON, D. B. (2016). Bayesian manifold regression. *Ann. Statist.* **44**, 876–905.
YANG, Y. & TOKDAR, S. T. (2015). Minimax-optimal nonparametric regression in high dimensions. *Ann. Statist.* **43**, 652–74.
ZHANG, C., LI, F. & JESTES, J. (2012). Efficient parallel kNN joins for large data in MapReduce. In *Proc. 15th Int. Conf. Extending Database Technology*. New York: Assocation of Computing Machinery, pp. 38–49.
ZIEMER, W. P. (2012). *Weakly Differentiable Functions: Sobolev Spaces and Functions of Bounded Variation*. New York: Springer.

[*Received on* 30 *July* 2018. *Editorial decision on* 4 *July* 2019]