# Bivariate density estimation using BV regularisation

Andreas Obereder [a], Otmar Scherzer [a], Arne Kovac [b,*]

[a] *Institute of Computer Science, University of Innsbruck, Austria*

[b] *Department of Mathematics, University of Bristol, Bristol, UK*

**Abstract**

The problem of bivariate density estimation is studied with the aim of finding the density function with the smallest number of local extreme values which is adequate with the given data. Adequacy is defined via Kuiper metrics. The concept of the taut-string algorithm which provides adequate approximations with a small number of local extrema is generalised for analysing two- and higher dimensional data, using Delaunay triangulation and diffusion filtering. Results are based on equivalence relations in one dimension between the taut string algorithm and the method of solving the discrete total variation flow equation. The generalisation and some modifications are developed and the performance for density estimation is shown.

*Key words:* Density estimation, modality, regularisation.

## 1 Introduction

### *1.1 Density estimation and local extreme values*

Local extreme values often play an important role in non-parametric statistics. Figure 1 shows observations from the *Old Faithful Geyser* in the Yellowstone National Park. Each observation consists of two measurements: the duration of the eruption and waiting time to the next eruption (both in minutes),

* Corresponding author. Tel.: +44 117 3317450; fax: +44 117 9287999.
*Email addresses:* `andreas.obereder@uibk.ac.at` (Andreas Obereder), `otmar.scherzer@uibk.ac.at` (Otmar Scherzer), `a.kovac@bristol.ac.uk` (Arne Kovac).

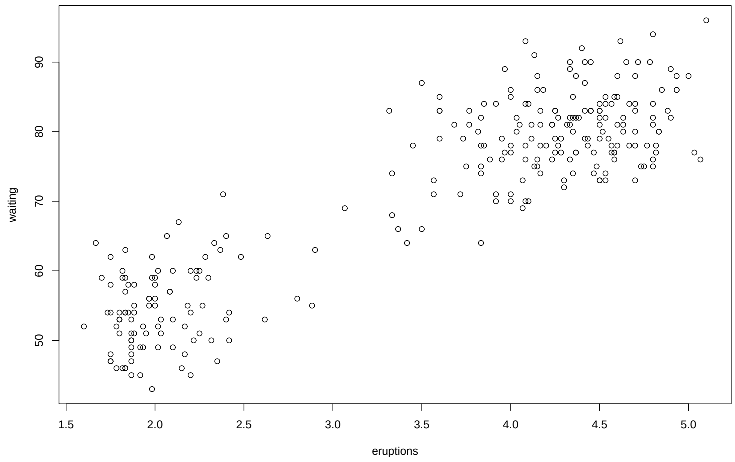


Fig. 1. 'Old Faithful' geyser data.

which are plotted against each other. The data indicate a bimodal distribution which is rare for the behaviour of a geyser and so physicists and geologists are interested in investigating possible reasons for this distribution (Azzalini and Bowman, 1990). One possible explanation for this pattern was given by Rinehart (1969) based on the temperature level of the water at the bottom of a geyser tube at the time the water at the top reaches the boiling point. This is a typical example where modality plays a crucial part in the analysis of the data. Local maxima in the density estimate reflect certain effects in the mechanism that generated the data.

We consider in this paper the problem of density estimation. Given a sample $x_1, \ldots, x_n \in \mathbb{R}^m$ the task is to specify a simple density function $f$ and hence a distribution function $F$ such that the data look like a typical sample from $F$. Simplicity is measured by the number of local extreme values, which we want to be as small as possible while still approximating the data. We concentrate on bivariate density estimations, $m = 2$, and the case $m = 1$, which serves motivation purposes. The principal concept presented in this paper can be applied to $m$-variate density estimation with $m \in \mathbb{N}$. However, the difficulty in practical realization is that commercial or open source software tools that we are using, are currently not available for $m \geq 4$.

Our approach relies on a suitable definition of adequacy (Davies, 1995). A measure of adequacy gives rise to a set of adequate functions, each of them representing a plausible model for the data in the sense that the data look like a "typical" sample from the model. The measure we employ in this paper is based on projections of the density function in the directions of its coordinates and evaluating the distance to the data with the Kuiper metric.

Having specified the set of adequacy we look for an adequate function $f$ which is as simple as possible and in particular has the *smallest* possible number of local extreme values. In particular we construct a scale of candidate functions $f^{(1)}, f^{(2)}, f^{(3)}, \ldots$ with decreasing complexity, in particular decreasing number of modes, and choose the simplest function that is still adequate as an approximation to the data.

The candidate functions are generated by generalisations of the *taut-string* algorithm (Davies and Kovac, 2001, 2004) for analysing data defined on a multi-dimensional domain. In one dimension the taut string method produces with high probability estimates with the correct modality while most other methods tend to produce superfluous local extreme values (Mammen, 1995). It is the aim of this paper to extend this concept to two and more dimensions. We propose a two step algorithm, which consists in generating an auxiliary function $y$ from the sampled data $x_i$ to which a filtering technique is applied. For the purpose of data filtering we discuss total variation regularisation, the total variation flow, and some variants (for instance based on the Fisher information), since these three techniques can be considered possible generalisations of the taut-string algorithm. Generalisations to contact problems derived from minimal surface minimisation as in Mammen and van de Geer (1997) have been studied in Hinterberger et al. (2003) and Scherzer (2005) but are not discussed further in this paper.

The outline of this paper is as follows:

In the remainder of Section 1 we give an overview of existing work on density estimation with a focus on work that is related to modality. In Section 2 we discuss some possible generalisation of the taut-string algorithm for analysing multi-dimensional data. In Section 3 an initial solution for the density estimation process is constructed. Moreover, a grid, which can be used in numerical reconstructions is provided. Section 4 discusses different diffusion filtering methods. Section 5 is concerned with the definition of adequacy for multivariate density functions. Finally in Section 6 we present results obtained with the different methods.

### *1.2 Previous work*

Existing literature on density estimation is vast. In particular the univariate situation has been studied extensively and numerous methods have been proposed. Of the different approaches the most popular one is kernel estimation. We refer to Nadaraya (1964), Watson (1964), Silverman (1986), Wand and Jones (1995) and Sain and Scott (1996) and the references given there. Other approaches are based on wavelets (Vidakovic, 1999), splines (Eilers and Marx, 1996) and mixtures of densities as analysed in the Bayesian framework by Richardson and Green (1997) and Roeder and Wasserman (1997). Multivariate density estimation has been analysed mainly in the context of kernel estimators (Scott, 1992; 2004) which have recently been adapted to the infinite case as well (Jacob and Oliveira, 1997; Ferraty and Vieu, 2006). Important applications include in particular astronomical data (Jang, 2006).

Shapes of densities have been most widely considered in the context of estimating the mode of a density. Mode estimation has been studied extensively in the case of univariate data (Parzen, 1962; Eddy, 1980; Vieu, 1996; Herrman and Ziegler, 2004), but also for multivariate data (Devroye, 1979; Abraham et al, 2003) and functional data (Gasser et al, 1998).

Work directly concerned with controlling the number of local extreme values has been done by Davies and Kovac (2001, 2004) who use the taut-string method to provide approximations for one-dimensional data with asymptotically consistent modality. A smooth version of the taut string method has recently been introduced by Kovac (2006). Further work in the context of investigating multimodality is provided by Müller and Sawitzki (1991) using their concept of excess mass. Their ideas have been extended to multidimensional distributions by Polonik (1995a, 1995b, 1999). Minotte and Scott (1993) consider mode trees where they calculate the modes of a kernel estimator for a scale of bandwidths and then plot the locations of the modes against the bandwidths. Another approach is by Hengartner and Stark (1995) who use the Kolmogoroff ball centred at the empirical distribution function to obtain nonparametric confidence bounds for shape restricted densities. Another way of controlling modality is that of mode testing. We refer to Good and Gaskins (1980), Silverman (1986), Hartigan and Hartigan (1985), Fisher, Mammen and Marron (1994) and Mammen (1995). The number of modes has also been used as a criterion for performance of a method by Park and Turlach (1992) and more recently by Davies et al (2006).

## 2 Taut strings in one and higher dimensions

One dimensional density functions approximating scattered point data can be calculated with the taut string algorithm (Davies and Kovac, 2001, 2004). The taut string method provides approximations with small number of local extreme values. It is the goal of this section to discuss generalisations of the taut string algorithm for density estimation on multi-dimensional data.

It has been shown by Mammen and van de Geer (1997) that the solution of the taut string algorithm with parameter $\lambda$ for given sampling data $\mathbf{y} = (y_1, \ldots, y_{n-1})$ sampled at the midpoints of uniformly distributed grid points $x_1 = 0, x_2 = h, x_3 = 2h, \ldots, x_n = (n-1)h$ in $\Omega = [0, 1]$ with sampling distance

$h = 1/(n-1)$ is equivalent to minimising

$$T_d(\mathbf{f}) := \frac{1}{2}\sum_{i=1}^{n-1} h\,|f_i - y_i|^2 + \lambda \sum_{i=1}^{n-2} h\frac{|f_{i+1}-f_i|}{h} .$$

where the $f_i$ are associated with a function $f$ s.t. $f(x) := f_i$ if $x \in (x_i, x_{i+1})$. The equivalence relation of the taut string algorithm and total variation flow regularisation shows that $\mathbf{f}$ is determined by $f_i = g_{i+1} - g_i$, $i = 1, \ldots, n-1$ where $\mathbf{g}$ is the minimiser of the constraint optimisation problem

$$\sum_{i=1}^{n-1}\sqrt{1 + \frac{|g_{i+1}-g_i|^2}{h^2}} \text{ subject to } \left| g_j - h\sum_{k=1}^{j} y_k \right| \le \lambda, \quad j = 1, \ldots, n\,, \tag{1}$$

and $g_1 = 0$. The $g_i$ are associated with a function $g$ which is linear in each $[x_i, x_{i+1}]$ and $g(x_i) = g_i$.

In Steidl et al. (2004) it was shown that the minimiser of $T_d$ and the solution of the *space discrete total variation flow* equation at time $t = \lambda$ which satisfies

$$\begin{aligned} \dot{f}_1 &\in sgn(f_2 - f_1), \\ \dot{f}_i &\in sgn(f_{i+1} - f_i) - sgn(f_i - f_{i-1}), \qquad i = 2, \ldots, n-2, \\ \dot{f}_{n-1} &\in -sgn(f_{n-1} - f_{n-2}), \end{aligned} \tag{2}$$

and

$$f(0) = y,$$

are identical. Here $\dot{f}_i$ denotes the time derivative of $f$ at grid point $x_i$ and $sgn$ denotes the sign-function.

These considerations reveal that for analysing sampling data with sampling points in $(0,1)$ there are at least three equivalent concepts, contact problems as formulated in Mammen and van de Geer (1997), discrete total variation regularisation (i.e. minimisation of $T_d$), and the discrete total variation flow (2). We show below that for continuous data the adequate continuous formulations of the three concepts can be generalised to multi-dimensional data. We focus on generalisations of total variation regularisation and the discrete total variation flow.

It has been shown by Grasmair (2006) (see also Pöschl and Scherzer, 2006) that for $y \in L^2(0,1)$, the minimiser of *continuous total variation minimisation*, consisting of minimisation of

$$T_c(u) := \frac{1}{2}\int_0^1 (f-y)^2 + \lambda \int_0^1 |f'|\ ,$$

is locally constant or satisfies $f = y$. If $f : \Omega \subset \mathbb{R}^2 \to \mathbb{R}$ is an image or a voxel data $f : \Omega \subset \mathbb{R}^3 \to \mathbb{R}$, then *continuous total variation minimisation* consists in minimisation of

$$\mathcal{T}_c(f) := \frac{1}{2} \int_\Omega (f - y)^2 + \lambda \, |Df| \, ,$$

where $|Df|$ is the total variation of $f$ (see e.g. Evans and Gariepy, 1992). In the image processing community $\mathcal{T}_c$ is called the R(udin)-O(sher)-F(atemi)-functional (Rudin, Osher, Fatemi, 1992).

With a similar argumentation it becomes evident that the total variation inclusion equation

$$\frac{\partial f}{\partial t} \in -\partial \, |Df| \, ,$$

where $\partial \, |Df|$ denotes the subgradient of the total variation $|Df|$ of $f$, is the continuous formulation of the discrete total variation flow equation. For more background on inclusion equations we refer to Brezis (1973). It is convenient and instructive, but not mathematically rigorous, to write

$$\partial \, |Df| = -\nabla \cdot \left( \frac{\nabla f}{|\nabla f|} \right) \, . \tag{3}$$

Total variation regularisation and total variation flow can be used if appropriate initial data $y$ has been determined from discrete sampling data. A method to construct the data $y$ which is compatible with the taut-string algorithm is presented in the next section.

## 3 Initialisation of irregularly sampled data

To calculate a density function on a one-dimensional domain given a sample $\mathbf{x}$ we first define a piecewise constant function $y$ by setting $h_i := x_{(i+1)} - x_{(i)}$ and $y(x) := 1/((n-1)h_i)$ for all $x \in [x_{(i)}, x_{(i+1)})$ and for $i = 1, \ldots, n-1$. Here $x_{(i)}$ denotes the ordered sample with $x_{(j)} \leq x_{(j+1)}$. Then the taut string algorithm can be used and equivalently be formulated as the problem of minimising $T_e$,

$$T_e(\mathbf{f}) := \frac{1}{2} \sum_{i=1}^{n-1} h_i (f_i - y_i)^2 + \lambda \sum_{i=1}^{n-2} h_i \, |\nabla_{h_i} f| \, (x_{(i)}) \, ,$$

where $|\nabla_{h_i} f| \, (x_{(i)}) = |f_{i+1} - f_i| \, / h_i$ is the absolute value of the right difference quotient of a function $f$ at $x_{(i)}$ with step size $h_i$.

We interpret $T_e$ as a quadrature rule of $\mathcal{T}_c(u)$ with sampling distances $h_i$ and partitioning sampling intervals $I_i := \left[ x_{(i)}, x_{(i+1)} \right] \, , \quad i = 1, \ldots, n-1$. The

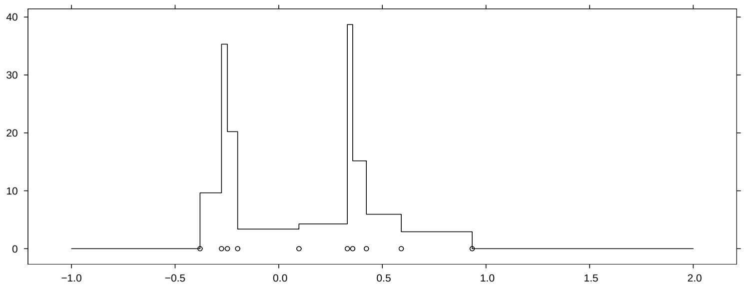


Fig. 2. Top: Randomly distributed data points with piecewise constant initial guess $y$.

method of minimising $T_e$ is called *discrete total variation minimisation with irregular samples*.

A typical example of a function $y$ is plotted in Figure 2.

We associate a grid with the nodes (i.e. data points) $x_i$ and the corresponding elements $I_i$. For multi-dimensional domains, Delaunay's triangulation can be used to determine partitioning tetrahedrons, thus generalising the concept of sampling intervals $I_i$ in space dimension one. Delaunay's triangulation and the associated Voronoi diagram are well known concepts from computational geometry and used in many applications (Aurenhammer and Klein, 2000). An excellent introduction to this topic is Edelsbrunner (2001). An example of a Delaunay triangulation is shown in Figure 3. The data used for the triangulation were 500 points randomly generated from the distribution shown in the left panel of Figure 4.

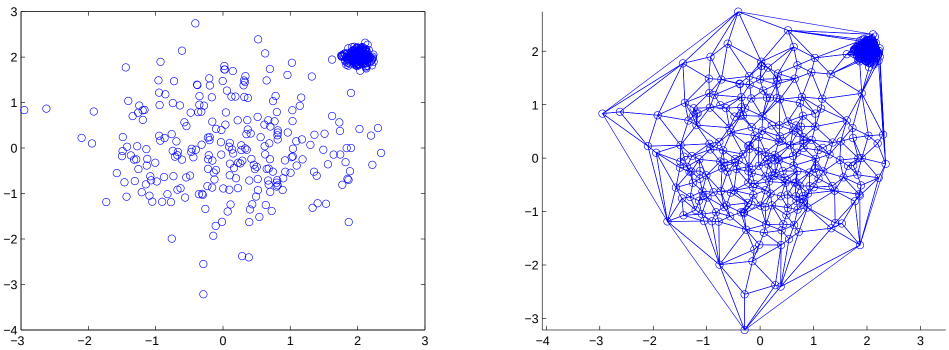

Fig. 3. Scattered data and the Delaunay triangulation

As in the the one-dimensional case we define the value of the initial solution for every grid-element $I_i$ as $1/(M \times area(I_i)), i = 1, \ldots, n-1$ where $M$ is the number of triangles. The right panel of Figure 4 shows an initial solution obtained by setting a constant value over each grid cell.

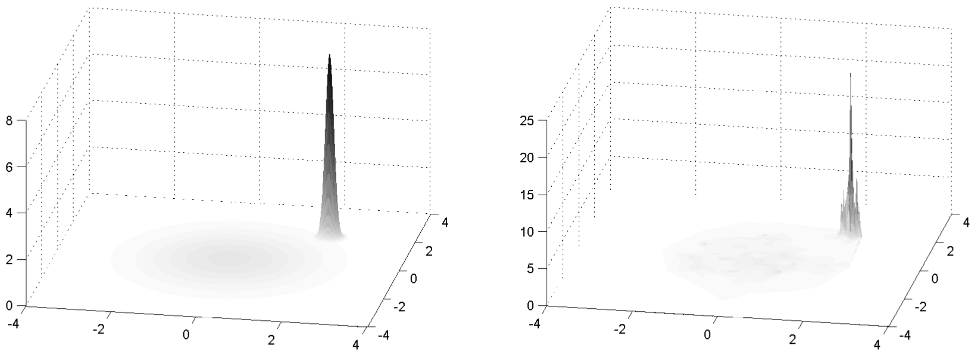


Fig. 4. Underlying distribution and Initial Guess

We mention that open source software for Delaunay triangulation for $m = 3$ is available on the internet.

## 4 Diffusion filtering after Delaunay's triangulation

For filtering data $y$ derived using Delaunay's triangulation from a discrete sample we use differential equations of the form

$$\begin{aligned} \frac{\partial f}{\partial t} &= \nabla \cdot (d(f, \nabla f)\nabla f) + e(f, \nabla f) \text{ in } \Omega\,, \\ \frac{\partial f}{\partial \mathbf{n}} &= 0 \text{ on } \partial\Omega\,, \end{aligned} \tag{4}$$

and

$$f(0) = y\,.$$

where $d(\cdot,\cdot)$ and $e(\cdot,\cdot) : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}$ are appropriate functions and $\mathbf{n}$ denotes the outer normal vector on the domain $\Omega$.

Particular examples considered in this paper are

- the *total variation flow equation* (3), where $d(f, \nabla f) = \frac{1}{|\nabla f|}$ and $e(f, \nabla f) = 0$,
- the *Fisher information flow equation* with $d(f, \nabla f) = \frac{1}{|f|}$ and $e(f, \nabla f) = \frac{f|\nabla f|^2}{2|f|^3}$, and
- the $\frac{3}{2}$ *Laplacian flow* with $d(f, \nabla f) = \frac{1}{\sqrt{|\nabla f|}}$ and $e(f, \nabla f) = 0$.

The solution of (4) at time $t$ is an approximation of the density to be estimated.

We note that the partial differential equation (4) is only formally stated and in general has to be considered as an inclusion equation.

### 4.1 The total variation flow equation

For the solution of (3) we use a standard semi-implicit scheme

$$f^{(n)} - f^{(n-1)} \in \Delta t \nabla \cdot \left( \frac{\nabla f^{(n)}}{|\nabla f^{(n-1)}|} \right) \text{ in } \Omega \text{ and } \frac{\partial f^{(n)}}{\partial \mathbf{n}} = 0 \text{ on } \partial\Omega\,. \tag{5}$$

Here we take $f^{(0)} = y$ and consider $f^{(n)}$ an approximation of $f(n\Delta t)$.

Equation (5) is related to adaptive weights smoothing (Polzehl and Spokoiny, 2000) as far as both methods use nonlinear diffusivities depending on the solution and the norm of the gradient of the solution. This brings up the idea of using general smoothing kernels of the form $d = d(f, |\nabla f|)$.

### 4.2 Fisher information minimisation

As an alternative to total variation regularisation we also consider a regularisation method (Ambrosio et al., 2005), which consists in minimisation of the functional $\mathcal{T}_{Fisher}(f)$,

$$\mathcal{T}_{Fisher}(f) = \frac{1}{2} \int_\Omega (f-y)^2 + \frac{\lambda}{2} \int_\Omega \frac{|\nabla f|^2}{|f|}\,,$$

which pronounces high peaks in $y$. The optimality condition for a minimiser $f$ is

$$\frac{f-y}{\lambda} = \nabla \cdot \left( \frac{\nabla f}{|f|} \right) + \frac{f \cdot |\nabla f|^2}{2|f|^3}\,. \tag{6}$$

Identifying $\lambda = \Delta t$, equation (6) can be interpreted as a fully implicit time step of length $\lambda$ of the following flow equation:

$$\frac{\partial f}{\partial t} = \nabla \cdot \left( \frac{\nabla f}{|f|} \right) + \frac{f \cdot |\nabla f|^2}{2|f|^3}\,. \tag{7}$$

In our numerical experiments we have implemented (7) with the semi-implicit scheme

$$f^{(n)} - f^{(n-1)} = \Delta t \left( \nabla \cdot \frac{\nabla f^{(n)}}{\sqrt{|f^{(n-1)}|^2 + \beta^2}} + \frac{f^{(n-1)} \cdot |\nabla f^{(n-1)}|^2}{\sqrt{4|f^{(n-1)}|^6 + \beta^2}} \right)\,. \tag{8}$$

Due to the diffusivity $d(f) = 1/|f|$ pronounced (high) peaks are prevented from being blurred, i.e. they remain significant. The right hand side is, as for the TV-flow, invariant to scaling of $f$.

### *4.3 The $\frac{3}{2}$ Laplacian flow equation*

In the literature the $\frac{3}{2}$ Laplacian operator is defined as

$$\partial \left( \frac{2}{3} \int_\Omega |\nabla f|^{\frac{3}{2}} \right) = -\nabla \cdot (d(f, |\nabla f|) \nabla u) \,,$$

where

$$d(f, \nabla f) = \frac{1}{\sqrt{|\nabla f|}} \,.$$

For filtering we have used the equation

$$\frac{\partial f}{\partial t} = \nabla \cdot (d(f, |\nabla f|) \nabla f) \,. \tag{9}$$

In our numerical solution we have actually approximated $d$ by

$$d_\beta(f, \nabla f) := \sqrt{1/\sqrt{|\nabla f|^2 + \beta^2}} \approx \frac{1}{\sqrt{|\nabla f|}} \,.$$

Moreover we have implemented the time steps with a semi-implicit algorithm.

This diffusion filtering approach is the steepest descent flow for the $W^{1,3/2}$-Sobolev semi norm.

Commercial and open source software for solving partial differential equations is available for three dimensional data. This is driven by the needs of engineering and industrial applications.

## 5 Approximation and Kuiper metrics

Davies and Kovac (2004) use Kuiper metrics to derive a measure of adequacy for univariate densities. This measure of adequacy combined with the taut string method provides a density estimator that has the right number of modes with high probability. Our definition of adequacy for multivariate densities is based on projecting the density in the directions of its coordinates and then analysing each projection individually with Kuiper metrics. Although this can easily be done for any dimension we concentrate on the bivariate case to keep notation as simple as possible.

Given bivariate data $x_1, \ldots, x_n$ and a bivariate density function $f(x^1, x^2)$ the problem is to decide whether the data $x_i$ look like a typical sample from $f$. We consider the marginal distribution functions $F^1(x)$ and $F^2(x)$, given by

$$F^1(x) = \int_{-\infty}^{x} \int_{-\infty}^{\infty} f(x^1, x^2) dx^2 dx^1 \text{ and } F^2(x) = \int_{-\infty}^{x} \int_{-\infty}^{\infty} f(x^1, x^2) dx^1 dx^2,$$

and define two sets of variables $u_i^1 = F^1(x_i^1)$ and $u_i^2 = F^2(x_i^2)$. If the approximation $f$ is adequate, then the new variables $u^1$ and $u^2$ should look like two samples of a uniform distribution on $[0, 1]$. Therefore we consider the empirical distribution $E^j$ of each sample

$$E^j(x) = \frac{1}{n} \sum_{i=1}^{n} 1_{\{u_i^j \le x\}}, \quad j = 1, 2,$$

and calculate the distances $d^1$ and $d^2$ to a uniform distribution in the Kuiper metric

$$d^j(u^j) = d_{Ku}(E^j, U) = (\sup_{x \in [0,1]} E^j(x) - x) - (\inf_{x \in [0,1]} E^j(x) - x),$$

where $d_{Ku}$ is the Kuiper metric and $U$ is a random variable with uniform distribution.

Let $\mathrm{qu}(n, \alpha)$ be the $\alpha$-quantile of the maximum of $d^1(U^1)$ and $d^2(U^2)$ so that

$$\mathbb{P}(\max(d^1(U^1), d^2(U^2)) \le \mathrm{qu}(n, \alpha)) = \alpha,$$

where $U_1^1, \ldots, U_n^1, U_1^2, \ldots, U_n^2$ are iid random variables with a uniform distribution on $[0, 1]$. Then a function $f$ is considered adequate with the given data if $\max(d^1(u^1), d^2(u^2)) \le \mathrm{qu}(n, \alpha)$. In this paper we always use $\alpha = 0.99$.

For small values of $n$ the quantiles $\mathrm{qu}(n, \alpha)$ can be obtained by simulation (see Table 1). For larger values of $n$ the distribution of the Kuiper distance between the uniform distribution and its empirical distribution can be approximated by a Brownian bridge and explicit expressions can be derived (Dudley, 1989) and evaluated to obtain quantiles $\mathrm{qu}(n, \alpha)$.

Given this notion of adequacy we are interested in finding an adequate density with the smallest number of local extreme values. Since an exact solution of this problem does not seem to be achievable, we consider the sequence of functions generated by the discrete diffusion filtering approaches introduced in the previous section. Then we choose the last adequate function in that sequence as an approximation to the data because it has minimal modality among all adequate functions.

| $n$ | $\alpha = 0.95$ | $\alpha = 0.99$ | $\alpha = 0.999$ |
|---|---|---|---|
| 100 | 0.175 | 0.199 | 0.228 |
| 200 | 0.125 | 0.141 | 0.160 |
| 500 | 0.081 | 0.091 | 0.105 |
| 1000 | 0.058 | 0.065 | 0.074 |
| 2000 | 0.041 | 0.046 | 0.052 |

Table 1
Quantiles for the Kuiper difference for 5 different sample sizes and 3 different values of $\alpha$.

## 6 Numerical examples

In our numerical experiments we compare the diffusion filtering methods (5), (8) and (9) with kernel estimators. To solve the flow equations we have implemented the following automatic algorithm:

(1) For given measurement data $x_i$ we calculate the Delaunay triangulation as described in Section 3 (see figure 3).
(2) Define an initial guess $y_i := 1/(M \times area(I_i))$ for every element of the Delaunay triangulation where $M$ is the number of triangles (see figure 4).
(3) We perform a piecewise linear interpolation $y$ on the Delaunay triangles, such that the integral over each triangle is one; that is the integral equals the integral of the piecewise constant initial guess on each triangle. Then we define a regular rectangular grid covering the Delaunay Triangulation and determine the values $y$ of the interpolated initial guess on the regular grid. In our numerical realisation we used a $281 \times 281$ grid.
(4) We use a Finite Difference Method to solve the flow equations on the regular grid with piecewise linear initial guess $y$ with a semi-implicit iteration as shown in equation (5) and (8), respectively.
(5) For each iteration we check whether the current function is still adequate with the data in the sense that it satisfies the Kuiper criterion from Section 5 with $\alpha = 0.99$. We stop once this is no longer the case and then choose the function from the previous iteration as an approximation to the data.

### *6.1 A simulated example*

A sample of size 500 was drawn from a mixture of two bivariate normal distributions. The underlying density function was given by

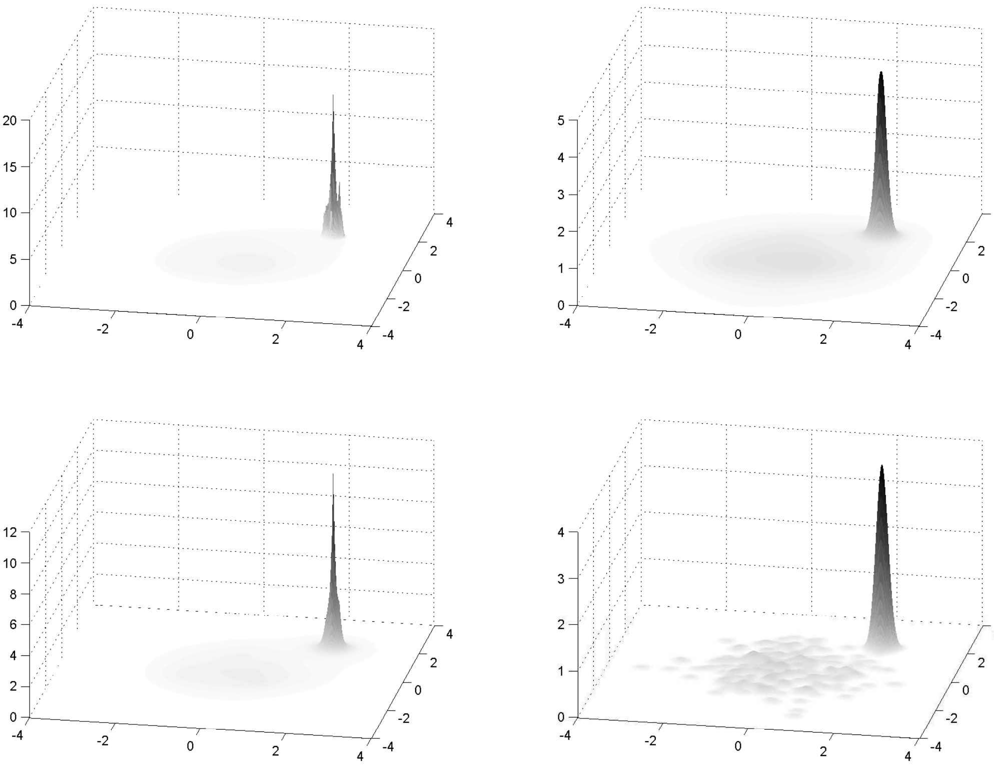


Fig. 5. Top left: TV-flow; Top right: $\frac{3}{2}$ Laplacian flow equation; Bottom left: Fisher information minimisation; Bottom right: Kernel estimator

$$f(x, y) = 0.5 \cdot \phi(x; 0, 1) \cdot \phi(y; 0, 1) + 0.5 \cdot \phi(x; 2, 0.1) \cdot \phi(y; 2, 0.1),$$

where $\phi(x; \mu, \sigma)$ denotes the density function of a normal distribution with mean $\mu$ and standard deviation $\sigma$.

The results are shown in Figure 5. The result top left was calculated using the total variation flow equation, where in the numerical realisation the time step size was set to 0.005 and where the flow was stopped by the Kuiper criterion at time $t = 0.33$, i.e. after 66 time steps.

Using equation (8) we obtain the solutions represented in the top right of the figure. Here the time step size was set to 0.05 and 49 time steps were performed. The bottom left of Figure 5 shows the result using equation (9) where the time step size was set to 0.05 and $t = 3.9$ (i.e. 78 time steps). Finally the bottom right of the figure shows the result of an kernel estimator using a Gaussian kernel and the largest bandwidth such that the Kuiper criterion was satisfied (0.947).

Comparing the results the most obvious observation is that the kernel estimator produces a much rougher approximation than the diffusion filtering methods. A close inspection reveals that the kernel estimator produces 82 local maxima whereas TV-flow needs only 6 local maxima and $\frac{3}{2}$-Laplacian and Fisher flow both produce exactly 2 local maxima.

We also note that TV-flow tends to create levels with constant values. The $\frac{3}{2}$-Laplacian flow is a good compromise between smoothing and preserving high peaks with small footpoints (base area). The best results are obtained using the Fisher information minimisation, since the solution $f$ itself is used in the diffusivity $(d(f) = 1/|f|\ )$ and the gradient of $f$ is taken into account in an energy term.

The advantage of using flow equations compared to minimising energy functionals where $\lambda$ is fixed, is that the solution for different times (i.e. different values of $\lambda$, see the results in Section 2, Steidl et al., 2004) can be calculated efficiently, i.e. no additional iteration is needed.

### *6.2 Another simulated example*

The example in this section shows that even for more complicated densities the diffusion filtering methods provide considerably simpler density estimates. We would like to stress again that the competing methods were entirely automatic and produced in each case the adequate function with the smallest number of local extreme values that could be generated with the particular method.

This time we generated a sample of size 1000 of the distribution with density function

$$\begin{aligned} f(x,y) =& 0.2 \cdot \phi(x;0,0.8) \cdot \phi(y;0,0.8) + 0.2 \cdot \phi(x;1,0.1) \cdot \phi(y;1,0.1) \\ &+ 0.2 \cdot \phi(x;-1.5,0.2) \cdot \phi(y;-1.5,0.2) \\ &+ 0.2 \cdot \phi(x;2,0.3) \cdot \phi(y;-2,0.3) + 0.2 \cdot \phi(x;-2,0.4) \cdot \phi(y;2,0.4). \end{aligned}$$

The density is shown in the top left corner of Figure 6 and the sample in the top right corner. The height of the density at the origin is only 0.016 compared with a height of 3.19 at the sharp peak. Suprisingly the mode at the origin was detected by $\frac{3}{2}$-Laplacian flow and Fisher flow and only missed out by the TV-flow. As in the example in the previous section $\frac{3}{2}$-Laplacian flow and Fisher flow produced exactly the correct number of local extreme values and placed them at the correct positions. Their estimates are shown in the middle right and bottom left panel of Figure 6. The TV flow shown in the middle left panel produced also a density with five local extreme values, however it missed out the one at the origin and placed an additional extreme value near

(1, 1). Finally, the kernel estimate, shown in the bottom right panel of Figure 6 exhibited 43 local extreme values.

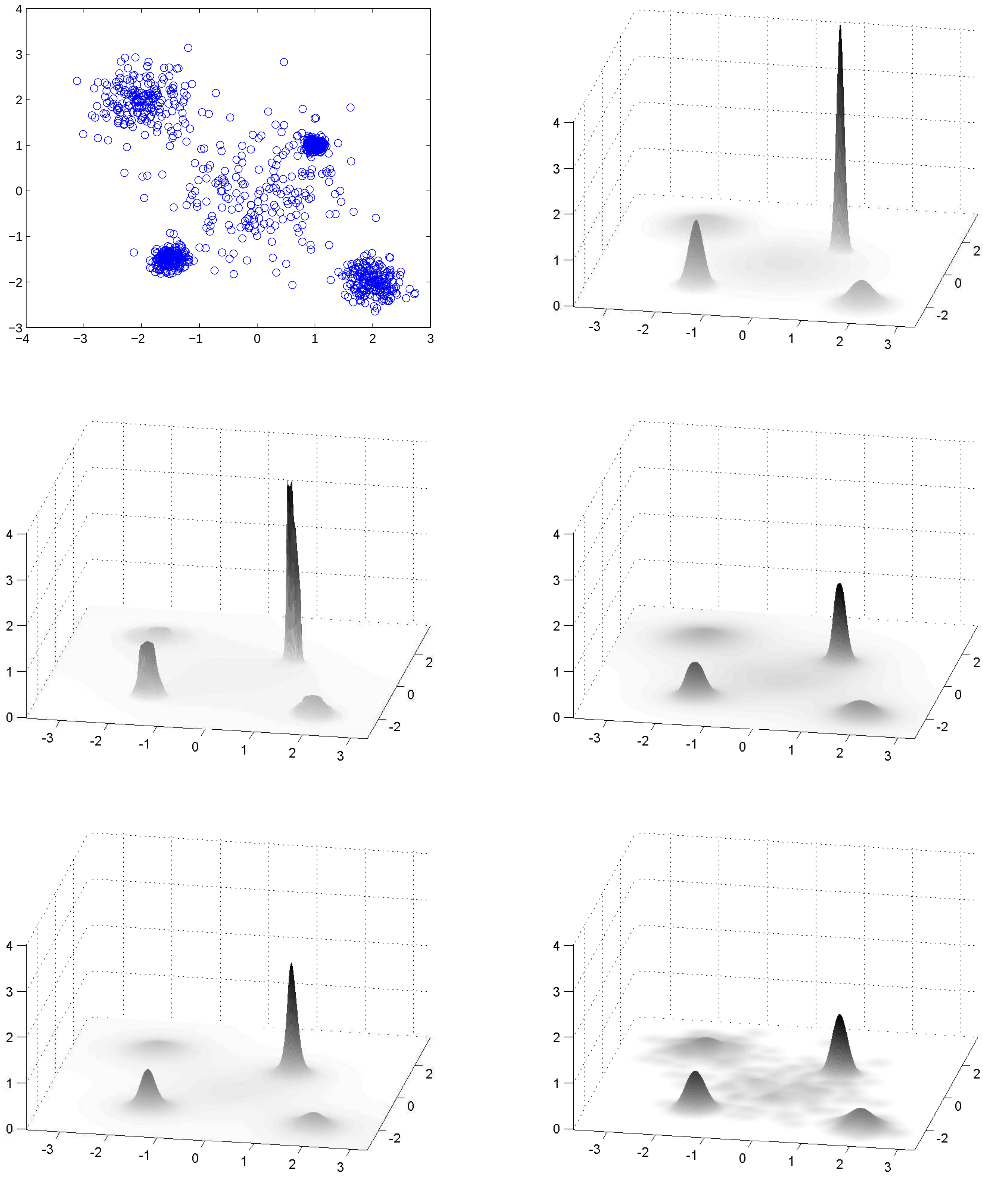

Fig. 6. Top left: Data sample; Top right: Original density function; Middle left: TV-flow; Middle right: $\frac{3}{2}$ Laplacian flow equation; Bottom left: Fisher information minimisation; Bottom right: Kernel estimator

### *6.3 The geyser data*

It is interesting to see how the diffusion filtering methods performs on the geyser data that we introduced in Section 1. Figure 7 shows perspective plots of the three methods and a contour plot for the $\frac{3}{2}$ Laplacian flow. The bimodality of the data set is clearly shown by all three methods although a close inspection of the output of the Fisher flow reveals an artificial third local extreme value to satisfy the Kuiper criterion. BV flow and Laplacian flow on the other hand both produce only the required two local extreme values.

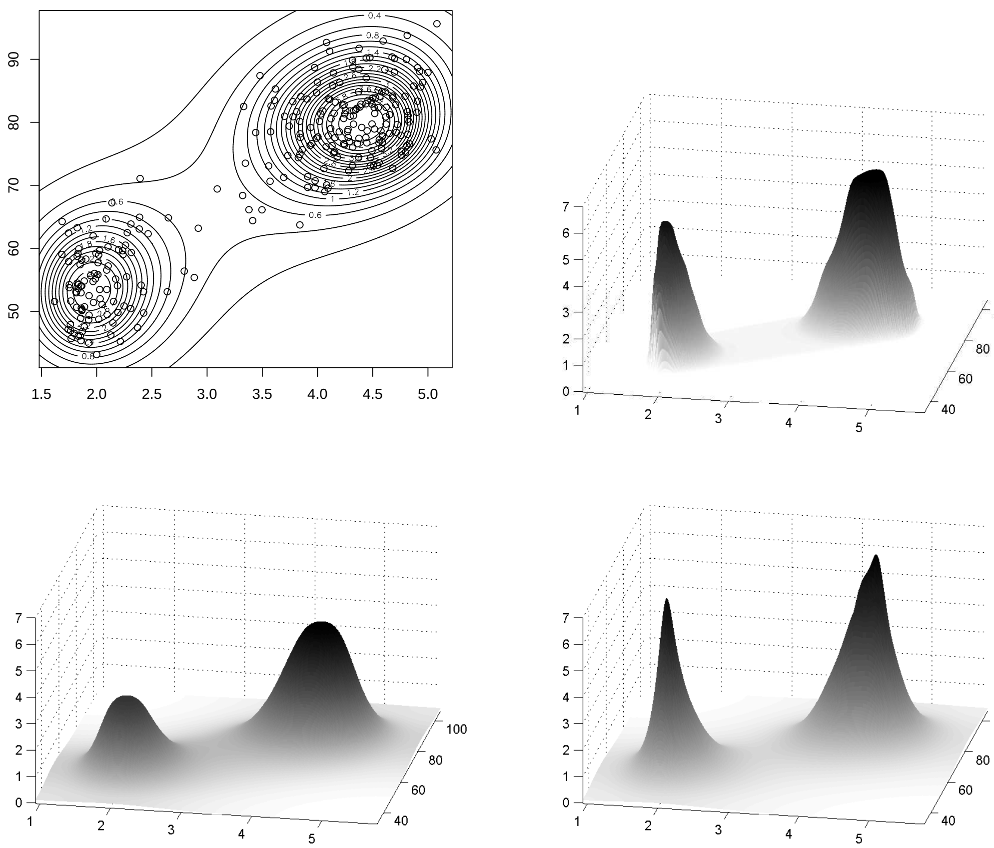


Fig. 7. Top left: Scatter plot of geyser data and contour plot of $\frac{3}{2}$ Laplacian flow; Top right: TV-flow; Bottom left: $\frac{3}{2}$ Laplacian flow; Bottom right: Fisher flow

## Conclusions

In this paper we have shown that Delaunay triangulation and diffusion filtering generalise the concept of the taut-string algorithm for analysing multidimensional data. In principle the concept applies in any space dimension, however, so far, the dimensionality is limited by the partial differential equation solver where current software can handle three space dimensions, but typically not more.

## Acknowledgement

The work of OS is supported by the Austrian Science Foundation (FWF) Projects Y-123INF, FSP 9203-N12 and FSP 9207-N12. Part of the work of AO and OS has been supported by the Austrian Ministry for Economy and Labour and by the Government of Upper Austria within the framework "Industrial Competence Centers" The work of AK was supported by the DFG within the Sonderforschungsbereich 475 at the University of Dortmund.

OS wants to thank Franz Aurenhammer and Oswin Aichholzer (Graz) for many discussions on Delaunay's triangulations. Appreciation is also due to the referees, whose comments on an earlier version of the paper have been invaluable and Oliver Johnson for proofreading the final version of the paper.

## References

Abraham, C., Biau, G., and Cadre, B. (2003). Simple estimation of the mode of a multivariate density. *Canadian Journal of Statistics*, 31:23–34.

Ambrosio, L., Gigli, N., and Savaré, G. (2005). *Gradient Flows in Metric Spaces and in the Space of Probability Measures*. Birkhäuser.

Aurenhammer, F. and Klein, R. (2000). *Handbook of Computational Geometry*, chapter 5, pages 201–290. Elsevier Science Publishing.

Bowman, A. and Azzalini, A. (1997). *Applied Smoothing Techniques for Data Analysis: the Kernel Approach with S-Plus Illustrations*. Oxford University Press, Oxford.

Brézis, H. (1973). *Opérateurs maximaux monotones et semi-groupes de contractions dans les espaces de Hilbert*. Number 5 in North-Holland Mathematics Studies. North-Holland Publishing Co, Amsterdam-London.

Davies, P. (1995). Data features. *Statistica Neerlandica*, 49:185–245.

Davies, P. L., Gather, U., and Weinert, H. (2006). Nonparametric regression as an example of model choice. Technical Report 24/06, Sonderforschungsbereich 475, Fachbereich Statistik, University of Dortmund, Germany.

Davies, P. L. and Kovac, A. (2001). Local extremes, runs, strings and multiresolution (with discussion). *Annals of Statistics*, 29:1–65.

Davies, P. L. and Kovac, A. (2004). Densities, spectral densities and modality. *Annals of Statistics*, 32(3):1093–1136.

Devroye, L. (1979). Recursive estimation of the mode of a multivariate density. *Canadian Journal of Statistics*, 7:159–167.

Dudley, R. (1989). *Real Analysis and Probability.* Wadsworth and Brooks/Cole, California.

Eddy, W. (1980). Optimum kernel estimators of the mode. *Annals of Statistics*, 8:870–882.

Edelsbrunner, H. (2001). *Geometry and Topology for Mesh Generation*. Cambridge University Press.

Eilers, P. and Marx, B. (1996). Flexible smoothing with b-splines and penalties. *Statistical Science*, 11:89–102.

Evans, L. C. and Gariepy, R. F. (1992). *Measure theory and fine properties of functions*. Studies in Advanced Mathematics. CRC Press, Boca Raton, FL.

Ferraty, F. and Vieu, P. (2006). *Nonparametric Functional Data Analysis: Theory and Practice*. Springer Series in Statistics. Springer-Verlag, New York.

Fisher, N. I., Mammen, E., and Marron, J. S. (1994). Testing for multimodality. *Computational Statistics and Data Analysis*, 18:499–512.

Gasser, T., Hall, P., and Presnell, B. (1998). Nonparametric estimation of the mode of a distribution of random curves. *Journal of the Royal Statistical Society*, 60:681–691.

Good, I. and Gaskins, R. (1980). Density estimating and bump-hunting by the penalizedlikelihood method exemplified by scattering and meteorite data. *Journal of the American Statistical Association*, 75:42–73.

Grasmair, M. (2006). The equivalence of the taut string algorithm and bv-regularization. *Journal of Mathematical Imaging and Vision*.

Hartigan, J. A. and Hartigan, P. M. (1985). The dip test of unimodality. *Annals of Statistics*, 13:70–84.

Hengartner, N. and Stark, P. (1995). Finite-sample confidence envelopes for shape-restricted densities. *Annals of Statistics*, 23:525–550.

Herrmann, E. and Ziegler, K. (2004). Rates on consistency for nonparametric estimation of the mode in absence of smoothness assumptions. *Statistics and Probability Letters*, 68:359–368.

Hinterberger et al., W. (2003). Tube methods for bv regularization. *Journal of Mathematical Imaging and Vision*, 19:219–235.

Jacob, P. and Oliveira, P. (1997). Kernel estimators of general radon-nikodym derivatives. *Statistics*, 30:25–46.

Jang, W. (2006). Nonparametric density estimation and clustering in astronomical sky surveys. *Computational Statistics and Data Analysis*, 50:760–774.

Kovac, A. (2007). Smooth functions and local extreme values. *Computational Statistics and Data Analysis*, 51:5156–5171.

Mammen, E. (1995). On qualitative smoothness of kernel density estimators. *Statistics*, 26:253–267.

Mammen, E. and van de Geer, S. (1997). Locally adaptive regression splines. *Annals of Statistics*, 25:387–413.

Minotte, M. C. and Scott, D. W. (1993). The mode tree: a tool for visualization of nonparametric density features. *Journal of Computational and Graphical Statistics*, 2:51–68.

Müller, D. W. and Sawitzki, G. (1991). Excess mass estimates and tests for multimodality. *J. Amer. Statist. Assoc.*, 86(415):738–746.

Nadaraya, E. A. (1964). On estimating regression. *Theory of Probability and its Applications*, 10:186–190.

Park, B.-U. and Turlach, B. (1992). Practical performance of several data-driven bandwidth selectors (with discussion). *Computational Statistics*, 7:251–285.

Parzen, E. (1962). On estimation of a probability density function and mode. *Annals of Mathematical Statistics*, 33:1065–1076.

Polonik, W. (1995a). Density estimation under qualative assumptions inhigher dimensions. *Journal of Mulitivariate Analysis*, 55:61–81.

Polonik, W. (1995b). Measuring mass concentrations and estimating density-contour clusters–an excess mass approach. *Annals of Statistics*, 23:855–881.

Polonik, W. (1999). Concentration and goodness of fit in higherdimensions: (aymptotically) distribution-free methods. *Annals of Statistics*, 27:1210–1229.

Polzehl, J. and Spokoiny, V. (2000). Adaptive weights smoothing with applications to image resoration. *Journal of the Royal Statistical Society, Series B*, 62:335–354.

Pöschl, C. and Scherzer, O. (2006). Characterization of minimizers of convex regularization functionals. Technical Report 42, Institute for Computer Science, University Innsbruck.

Richardson, S. and Green, P. (1997). On bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society*, 59:731–792.

Rinehart, J. S. (1969). Thermal and seismic indications of old faithful geyser's inner workings. *Journal of Geophysical Research*, 74:566.

Roeder, K. and Wasserman, L. (1997). Practical Bayesian density estimation using mixtures of normals. *J. Amer. Statist. Assoc.*, 92(439):894–902.

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

Sain, S. and Scott, D. (1996). On locally adaptive density estimation. *Journal of the American Statistical Association*, 91:1525–1534.

Scherzer, O. (2005). Taut-string algorithm and regularization programs with g-norm data fit. *Journal of Mathematical Imaging and Vision*, 23(2):135–143.

Scott, D. (1992). *Multivariate density estimation. Theory, practice and visualization.* Wiley, New York.

Scott, D. (2004). *Mulitvariate density estimation*, pages 517–538. Handbook of Statistics. Springer-Verlag.

Silverman, B. W. (1986). *Density estimation for statistics amd data analysis*. Chapman and Hall, London.

Steidl, G., Weickert, J., Brox, J., Mrazek, P., and Welk, M. (2004). On the equivalence of soft wavelet shrinkage, total variation diffusion, total variation regularization, and sides. *SIAM J. Numer. Anal.*, 42(2):686–713.

Vidakovic, B. (1999). *Statistical modeling based on wavelets*. Wiley, New York.

Vieu, P. (1996). A note on density mode estimation. *Statistics and Probability Letters*, 26:297–307.

Wand, M. P. and Jones, M. C. (1995). *Kernel smoothing*. Chapman and Hall Ltd., London.

Watson, G. S. (1964). Smooth regression analysis. *Sankhyā*, 26:101–116.