A Festschrift for Herman Rubin
Institute of Mathematical Statistics
Lecture Notes – Monograph Series
Vol. 45 (2004) 237–249

Estimating gradient trees

Ming-Yen Cheng[1], **Peter Hall**[2] **and John A. Hartigan**[3]

National Taiwan University, Australian National University and Yale University

Abstract: With applications to cluster analysis in mind, we suggest new approaches to constructing tree diagrams that describe associations among points in a scatterplot. Our most basic tree diagram results in two data points being associated with one another if and only if their respective curves of steepest ascent up the density or intensity surface lead toward the same mode. The representation, in the sample space, of the set of steepest ascent curves corresponding to the data, is called the gradient tree. It has a regular, octopus-like structure, and is consistently estimated by its analogue computed from a nonparametric estimator which gives consistent estimation of both the density surface and its derivatives. We also suggest 'forests', in which data are linked by line segments which represent good approximations to portions of the population gradient tree. A forest is closely related to a minimum spanning tree, or MST, defined as the graph of minimum total length connecting all sample points. However, forests use a larger bandwidth for constructing the density-surface estimate than is implicit in the MST, with the result that they are substantially more orderly and are more readily interpreted. The effective bandwidth for the MST is so small that even the corresponding density-surface estimate, let alone its derivatives, is inconsistent. As a result, relationships that are suggested by the MST can change considerably if relatively small quantities of data are added or removed. Our trees and forests do not suffer from this problem. They are related to the concept of gradient traces, introduced by Wegman, Carr and Luo (1993) and Wegman and Carr (1993) for purposes quite different from our own.

1. Introduction

Gradient trees capture topological features of multivariate probability densities, such as modes and ridges. In this paper we suggest methods for estimating gradient trees based on a sample of n observations from the density. Each estimator is in the form of a tree with $n-1$ linear links, connecting the observations. The methods will be evaluated in terms of their accuracy in estimating the population gradient tree, and their performance for real data sets. We also propose a new technique for describing, and presenting information about, neighbour relationships for spatial data.

To define a gradient tree, note that the gradient curves of a multivariate density f are the curves of steepest ascent up the surface $\mathcal{S}$ defined by $y = f(x)$. The representations of gradient curves, in the sample space, will be called *density ascent lines*, or DALs. The tree-like structure that they form is the gradient tree. This theoretical quantity may be estimated by replacing f by a nonparametric density estimator, $\hat{f}$ say, and then following the prescription for computing DALs and the gradient tree.

[1] Department of Mathematics, National Taiwan University, Taipei 106, Taiwan. e-mail: `cheng@math.ntu.edu.tw`

[2] Centre for Mathematics and its Applications, Australian National University, Canberra, ACT 0200, Australia. e-mail: `halpstat@maths.anu.edu.au`

[3] Department of Statistics, Yale University, Box 2179, Yale Station, New Haven, CT 06520, USA. e-mail: `john.hartigan@yale.edu`

Keywords and phrases: density ascent line, density estimation, forest, gradient trace, minimum spanning tree, nearest neighbour methods, ridge estimation, tree diagram.

AMS 2000 subject classifications: primary 62H30; secondary 62H20.

A gradient tree may be viewed as a modification the concept of a 'gradient trace', introduced by Wegman, Carr and Luo (1993) and Wegman and Carr (1993). The goal of these authors was to use gradient traces to compute 'k-skeletons', which are k-dimensional analogues of the mode and represent nonlinear regression-like summary statistics. Our purpose is quite different. We view gradient trees as a tool for cluster analysis, and argue that in this context the concept has advantages over more familiar methodology such as minimum spanning trees, or MSTs, introduced by Florek *et al.* (1951); see also Friedman and Rafsky (1981, 1983).

An MST is the graph of minimum total length connecting all sample points. It is an estimator of the gradient tree that arises when we take $\hat{f}$ to be the most basic of nearest neighbour density estimators, in which the estimate at each point is inversely proportional to a monotone function of the distance to the closest sample point. However, this is a poor estimator of the population density, let alone its gradient, and so it is not surprising that the MST is a poor estimator of the corresponding population gradient tree. We suggest gradient tree estimators that are asymptotically consistent for the corresponding population gradient tree, and which also improve on the MST for small sample sizes.

We also suggest algorithms for drawing 'forests', using either the full dataset or subsets that have been identified by the gradient tree. Like the MST, a forest provides information about relationships among neighbouring data, but like our gradient tree it has the advantage that it is based on relatively accurate, and statistically consistent, information about gradients. In contrast with the MST, a forest is based on directed line segments, with the direction corresponding to movement up an estimate $\widehat{\mathcal{S}}$ of the surface $\mathcal{S}$. Our approach to constructing a forest allows the experimenter to choose, when describing relationships between points, how much emphasis will be given to a relatively conventional Euclidean measure of closeness of the points, and how much will be given to a measure of closeness related to movement up $\widehat{\mathcal{S}}$.

Although we work mainly in the bivariate case, our methods are certainly not limited to two dimensions. One way of treating high-dimensional data is of course to form bivariate scatterplots by projection, and apply our methods to the individual plots. Tools for manipulating two-dimensional projections of three- or higher-dimensional data include Asimov's (1985) grand tour, Tierney's (1990) Lisp-Stat, or Swayne, Cook and Buja's (1991) XGobi; see also Cook, Buja, Cabrera and Hurley's (1995) grand-tour projection-pursuit.

Moreover, density ascent lines and gradient trees have analogues when the sample space is of arbitrarily high dimension, rather than simply bivariate. (Analogues of forests may be constructed too, but the formula for a certain penalty term that is needed to define a forest is more complex in higher dimensions.) Hence, rather than compute these quantities for bivariate scatterplots, their multivariate forms (represented as lines in space, rather than lines in the plane) could be calculated and then viewed through their bivariate projections, or through rotations of trivariate projections.

Density-based approaches to assessing relationship have also been considered by Hartigan (1975), who took clusters to be maximal connected sets (that enjoyed at least a certain level of likelihood) of points of density exceeding a certain level. See also the discussion of tree diagrams by Hartigan (1982). Alternative approaches include methods based on measures of distance that satisfy the triangle inequality (e.g. Jardine and Sibson, 1971; Hubert, 1974) and techniques founded on parametric mixtures (e.g. Everitt, 1974; Kuiper and Fisher, 1975). Wishart (1969) was an early user of near neighbour methods to construct clusters.

Pruzansky, Tversky and Carroll (1982) compared spatial and tree representations of data.

2. Gradient trees and ridges

We begin by defining a 'true' density ascent line, when the density f of the bivariate distribution of a general data point X is assumed known. Then we discuss computation of this line, and calculation of its sample version.

Let $\mathcal{S}$ be the surface defined by the equation $y = f(x)$, and assume that both the first derivatives of f are continuous everywhere. Suppose too that the set of positive density is connected, and contains at most a finite number of stationary points. A *density ascent line* (DAL) for f, starting at a point x in the plane Π that denotes the sample space, is defined to be the projection, into Π, of the trajectory formed by climbing $\mathcal{S}$ in the direction of steepest ascent. Henceforth we shall call the 'projection' of a three-dimensional structure into Π, the 'representation' of that structure in Π, and reserve the term 'projection' for other purposes.

If the trajectory on $\mathcal{S}$ is represented as the locus of points $(x^{(1)}(s), x^{(2)}(s), y(s))$, where $s \in (0, s_0)$ is a convenient parameter such as distance along the trajectory from one of its ends, then the corresponding DAL will be the curve formed by the locus of points $(x^{(1)}(s), x^{(2)}(s))$, for $s \in (0, s_0)$, in Π. If f_1, f_2 denote the derivatives of f in the two coordinate directions then the curve of steepest ascent is in the direction (f_1, f_2), and is well defined except at stationary points of the density. The gradient tree is the collection of closures of DALs.

Next we give more detail about a DAL, and then an explicit method for computing one. Let $D(f) = (f_1^2 + f_2^2)^{1/2}$ denote the length of $\nabla f = (f_1, f_2)$, and put $\omega_j = f_j/D(f)$ and $\omega = (\omega_1, \omega_2)$. Then, for $x \in \mathcal{S}$, $\omega(x)$ is the unit vector in Π representing the direction of steepest ascent up $\mathcal{S}$, at the point $(x, f(x)) \in \mathcal{S}$. The DAL that passes through $x \in \Pi$ may be thought of as having been obtained, starting at a point on the line, by stepping along the line in the direction indicated by ω. Formally, the DAL that passes through $x \in \Pi$ may be represented by the infinitesimal transformation, $x \mapsto x + \omega(x)\, ds$, where ds is an element of displacement along the DAL, denoting the length of one of the aforementioned steps.

This suggests the following algorithm for computation. Given $x_0 \in \Pi$, and a small positive number δ, consider the sequence of points $\mathcal{P} \equiv \{x_j : -\infty < j < \infty\}$ defined by $x_j = x_{j-1} + \omega(x_{j-1})\,\delta$ and $x_{-j} = x_{1-j} + \omega(x_{1-j})\,\delta$, for $j \geq 1$. Thus, the DAL that passes through x_0 represents the limit, as $\delta \to 0$, of the sequence $\mathcal{P}$. The algorithm is convenient for numerical calculation, provided we stop before reaching places where $D(f)$ vanishes.

In empirical work, where we compute estimators of DALs, we of course replace f, f_1, f_2 in the algorithm by their estimators $\hat{f}, \hat{f}_1, \hat{f}_2$. We used the algorithm described above, with a suitably small value of δ, to calculate the empirical DALs shown in Section 4. Alternatively, one could recognise that DALs are integral lines of the gradient field of a smooth density function, implying that in principle they could be computed using an ordinary differential equation solver.

There is no commonly accepted definition of a *ridge* (or antiridge) of a surface such as $\mathcal{S}$, and in fact four different approaches, framed in terms of indices of 'ridgeness', were suggested by Hall, Qian and Titterington (1992). The following is related to the second definition there, and is chosen partly for ease of computation in the present context; its representation in Π is easily calculated from the functional $D(f)$. Moreover, the representation is itself a DAL, and it admits an elementary (and computable) generalisation to high-dimensional settings.

Given a point P on $\mathcal{S}$, let $\Pi' = \Pi'(P)$ denote the plane that contains P and is parallel to Π, and let $\mathcal{C}$ be the curve formed by the intersection of Π' with $\mathcal{S}$. If the steepest ascent curve up $\mathcal{S}$, starting from P, is perpendicular to $\mathcal{C}$ at P, then we say that P is a point on a ridge (or an antiridge) of $\mathcal{S}$. The ridge or antiridge itself is a locus of such points, and is the curve of steepest ascent on $\mathcal{S}$ that passes through P. (Therefore, its representation in Π is a DAL.) The point P is on a ridge, rather than an antiridge, if the curvature at P of the curve formed by the intersection of $\mathcal{S}$ with a plane perpendicular to Π, and containing P, is negative; and on an antiridge if the curvature is positive.

A ridge can bifurcate at a point which represents a location on $\mathcal{S}$ where three or more ridges join. The trajectories of steepest ascent that climb up the surface between two ridges meeting at a bifurcation point B, necessarily join one another at B. From there they have a common path, along an ascending ridge that leads away from B; and they continue together until they terminate at a local maximum, perhaps passing through other bifurcation points on the way.

The representation, in the plane Π, of a ridge and a bifurcation point will be called a ridge line (RL) and a branchpoint, respectively. The DALs corresponding to the representations (in Π) of ridges have different paths until they meet their first branchpoint, after which they are the same until they terminate at a mode. An RL is essentially what Wegman and Carr (1993) call 1-skeleton, the main difference being in the definition of a ridge.

Therefore, the DALs that comprise a gradient tree do have a tree-like structure, in the following way. Individual points in the sample space, representing leaves of the tree, are at first linked to branchpoints through distinct DAL paths. Beyond the first branchpoint the consolidated bundle of DAL paths, representing a branch of the tree, may be joined at subsequent branchpoints by other branches, until they finally reach a mode.

In theory, more complex structures are also possible, for example when two branches lead away from a branchpoint and come together again at a mode or at another branchpoint. However, it is rare in practice for such features to occur in DALs computed from data via nonparametric density estimators, and so we shall not consider them further here.

Two points $x_1, x_2 \in \Pi$ that are linked to the same mode by a DAL, may be said to lie in the same cluster. Thus, DALs divide the plane into clusters. Ridge lines divide the sample space in a DEFANGED different manner, in a sense orthogonal to the division into clusters. They give neither a subclassification nor a higher-level classification, but provide information of a different type, as follows.

If the ridge that produced an RL were almost horizontal, and lay between two local maxima of $\mathcal{S}$, occurring at points $x_{\max,1}$ and $x_{\max,2}$, say, in Π, then the points along that RL would have no clear allocation to the clusters corresponding to $x_{\max,1}$ and $x_{\max,2}$. Therefore, the RL would represent a watershed in the division of the sample space into clusters. On the other hand, a point that lay on either side of, and sufficiently close to, the RL would be more definitively allocated to just one of the clusters represented by $x_{\max,1}$ and $x_{\max,2}$.

More generally, we might fairly say that points that lie on one side or other of an RL are less ambiguously associated with their corresponding mode, at least if they are sufficiently close to the RL, than are points that lie directly on the RL. Indeed, if two points $x_1, x_2 \in \Pi$ lie on opposite sides of, and sufficiently close to, an RL, then all points x_3 that lie between x_1 and x_3 can be said to be more ambiguously associated with their corresponding modes than either x_1 or x_2.

In addition to their role in defining such a gradation of the sample space, the fact that RLs of density or intensity estimators represent the 'backbone' and 'ribs' of the structure of those quantities means that they provide valuable quantitative information about structure. Indeed, they are sometimes used to approximate the locations of physical structures associated with scatterplots, for example positions of the subterranean fault lines that give rise to earthquake epicentres (see Jones and Stewart, 1997).

Relative to ridge lines, antiridge lines have more connection with clustering in the usual sense, since they represent boundaries between regions where points are assigned to different clusters. However, they are typically computed from relatively little data, and so their locations may not be known as precisely as those of ridge lines.

Next we describe a method for locating, and computing, an RL, given the density f. A locus of points on $\mathcal{S}$, all of which have the same height above Π, is called a level set of $\mathcal{S}$. Its representation in Π is a contour of $\mathcal{S}$. An RL may be reached from another point in Π by moving around a contour. The orientation of the contour passing through x is the direction of the unit vector $\omega_{\mathrm{perp}}(x)$, say, defined as being orthogonal to $\omega(x)$ and determined up to a change of sign. Therefore, the contour is defined by the infinitesimal transformation $x \mapsto x \pm \omega_{\mathrm{perp}}(x)\, ds$, where ds is an infinitesimal unit of length around the contour. The point at which this contour cuts an RL is a local minimum of $D(f)$; a local maximum corresponds to cutting the representation in Π of an antiridge.

Hence, to find a point x on an RL we move around the contour, computing $D(f)$ as we go, until we find a local minimum of $D(f)$. Then, moving along the RL is equivalent to moving up the DAL starting from x, or down the DAL leading to x; we have already described how this may be done. It is helpful to note that turning points of $D(f)$ are solutions of the equation

$$f_{12}\left(f_1^2 - f_2^2\right) = f_1 f_2 \left(f_{11} - f_{22}\right),$$

where $f_{ij}(x) = \partial^2 f(x_1, x_2)/\partial x_i\, \partial x_j$. Of course, descending the DAL that defines a ridge is equivalent to traversing the line defined by $x \mapsto x - \omega(x)\, ds$, where now ds is an infinitesimal unit of length along the DAL.

More generally, if the sample space Π is p-dimensional, where $p \geq 2$; and if we define $D = (\sum_i f_i^2)^{1/2}$, where f_i equals the derivative of f in the direction of the ith coordinate direction, for $1 \leq i \leq p$; then a ridge line or antiridge line is a locus in Π of turning points of $D(f)$. It may be calculated by generalising the method suggested above. DEFANGED A practicable, computational algorithm for an RL may be obtained as before, replacing the infinitesimal ds by a small positive number δ. The empirical version, in which density f is replaced by the density estimator $\hat{f}$, also follows as before; we used this method to compute the RLs shown in Section 4. Tests for significance of empirical modes may be based on work of Silverman (1981), Hartigan and Hartigan (1985), Müller and Sawitzki (1991) or Cheng and Hall (1999), for example.

3. Forests based on distance and density

While the minimum spanning tree is not consistent for the population gradient tree, it provides some information about relationships among neighbouring data values. In this section we suggest a regularisation of the minimum spanning tree in which links between observations are penalised if they are not sufficiently close to estimated density ascent lines. It may be applied to a subset $\mathcal{Y} = \{Y_1, \ldots, Y_N\}$

of the sample $\mathcal{X} = \{X_1, \dots, X_n\}$, for example to those data that are linked to the same mode in the gradient tree, as well as to the full sample.

Let $\|Y_i - Y_j\|$ denote Euclidean distance in the sample space Π, and let $d(Y_i, Y_j)$ be some other measure of distance between Y_i and Y_j. It is not necessary that $d(\cdot,\cdot)$ be a metric; appropriate definitions of d are powers of Euclidean distance in Π, i.e. $d(Y_i, Y_j) \equiv \|Y_i - Y_j\|^s$, and powers of Euclidean distance on $\widehat{\mathcal{S}}$, i.e.

$$d(Y_i, Y_j) \equiv \left[\|Y_i - Y_j\|^2 + \{\hat f(Y_i) - \hat f(Y_j)\}^2\right]^{s/2},$$

where $s > 0$. In our numerical work in Section 4 we shall use the first of these definitions, with $s = 2$.

Now add a penalty to $d(Y_i, Y_j)$, proportional to the squared length of the projection of $Y_i - Y_j$ orthogonal to $\widehat{\omega}(Y_i)$. (Here, $\widehat{\omega}(x)$ denotes the empirical form of $\omega(x)$, computed with $\hat f$ replacing f.) Equivalently, the penalty is proportional to the area of the triangle that has one side equal to the length of the line joining Y_i and Y_j, and another equal to the length of the representation in Π of a straight-line approximation, of the same length as the previous side, to the gradient curve. The area in question equals half the value of $\|Y_i - Y_j\|^2 - \{(Y_i - Y_j) \cdot \widehat{\omega}(Y_i)\}^2$, if the vertex of the triangle is at Y_i. We apply these penalties in proportion to a tuning parameter $t \ge 0$, obtaining symmetrically and asymmetrically penalised versions, respectively, of $d(Y_i, Y_j)$:

$$\begin{aligned} D(Y_i, Y_j) &= d(Y_i, Y_j) + t\left[\|Y_i - Y_j\|^2 - \{(Y_i - Y_j) \cdot \widehat{\omega}(Y_i)\}^2\right] \quad \text{or} \qquad (3.1)\\ D(Y_i, Y_j) &= d(Y_i, Y_j) + t\left[\|Y_i - Y_j\|^2 - \{(Y_i - Y_j) \cdot \widehat{\omega}(Y_i)\}^2\right]\\ &\quad + t\left[\|Y_i - Y_j\|^2 - \{(Y_i - Y_j) \cdot \widehat{\omega}(Y_j)\}^2\right]. \qquad (3.2) \end{aligned}$$

Using a large value of t amounts to placing more emphasis on point pairs whose interconnecting line segment lies close to a gradient curve.

We are now in a position to construct the forest corresponding to the dataset $\mathcal{Y}$ and the penalised distance measure D. Given Y_i, we draw a directed line segment from Y_i to Y_j if and only if Y_j minimises $D(Y_i, Y_j)$ over all points Y_j for which $\hat f(Y_j) > \hat f(Y_i)$. The forest is the set of these directed segments. If $\mathcal{Y}$ is a cluster, and if we adjoin to $\mathcal{Y}$ the unique mode associated with that structure, then with probability 1 there is exactly one point Y_i (the mode) in $\mathcal{Y}$ for which the directed line segment does not exist. As we climb higher up the surface the directed line segments tend to coalesce, producing a tree structure sprouting from the mode (although it was constructed from the opposite direction).

If we define $D(\cdot,\cdot)$ as at (3.1) then taking $t = 0$ produces a forest that is similar in both definition and appearance to the minimum spanning tree, although based on directed line segments. Choosing a relatively large value of t imposes greater penalty for not walking as nearly as possible along the DAL that starts at Y_i, when passing from Y_i to Y_j. The extent to which line segments cross over in the forest may be reduced by increasing t, thereby forcing the direction of movement on $\widehat{\mathcal{S}}$ to give more emphasis to the uphill component of motion. The advantage of (3.2) over (3.1) is that in the former the tree treats the notions of 'uphill' and 'downhill' symmetrically, but in practice, forests defined by (3.1) and (3.2) are virtually identical.

4. Numerical examples

Rees (1993) determined the 'proper motions' of 515 stars in the region of the globular cluster M5. Using the proper motions and radial velocity dispersions he estimated the probability that each star belonged to the cluster. The analysis below is

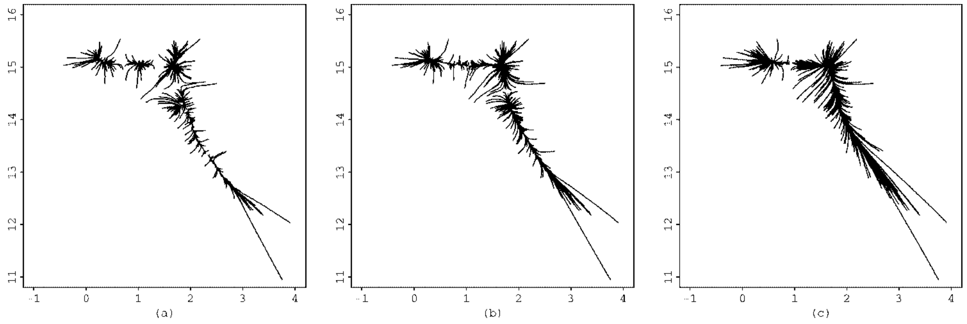

Figure 1: *Steepest Ascent Trees.* Panels (a), (b) and (c) depict DALs for the smoothed nearest neighbour estimator corresponding to $k = 25, 50, 100$, respectively.

based on the Herzprung-Russell diagram, a plot of magnitude versus temperature, for the 463 stars that were determined by Rees to have probability of at least 0.99 of belonging to the cluster.

We employed two different versions of $\hat{f}$. Both were nearest neighbour methods, which we chose for reasons that were both pragmatic (the adaptivity of NN methods means that they have less tendency than other density estimation techniques to suffer from spurious islands of mass) and didactic (NN methods are commonly used in classification problems). The first version of $\hat{f}$ was a standard k'th nearest neighbour estimator, with $\hat{f}(x)$ equal to $k/(n\pi r^2)$ where $r = r(x)$ was the smallest number such that the circle centred on x and with radius r contained just k points. The second density estimator was a smoothed version of the first, equal to $2k/(n\pi r^2)$ where r was the solution of

$$\sum_{i=1}^{n} \left\{ 1 - \left(\frac{\|X_i - x\|}{r} \right)^2 \right\}^+ = k.$$

See Section 5 for discussion of this technique. Since our graphs remain unchanged if we multiply $\hat{f}$ by a constant factor then it is not necessary to normalise, and so the factor $k/n\pi$ may be dropped.

Figure 1 depicts the gradient tree, or collection of DALs, for $k = 25, 50, 100$. In constructing figures 1 and 2 we used only the second, smoothed nearest neighbour estimator $\hat{f}$. Note that as k increases the number of empirical modes decreases; the number is $7, 4, 2$ for $k = 25, 50, 100$ respectively. The gradient trees indicate which points are most closely associated with the respective modes. The orientations and spacings of the tentacles of these 'octopus diagrams' provide information about the steepness of $\hat{f}$ in different places.

Figure 2 shows the RLs for the same values of k. Ridge lines are depicted by solid lines, and antiridge lines by dashed lines. The main RL, in the lower right of the figure, is clearly depicted; it is in a sense the backbone of the surface defined by the density estimator. Other RLs represent relatively minor 'creases' in the surface, and play more the role of 'ribs'.

The gradient trees provide only minimal information about interpoint relationships. Detail of that type is more readily supplied by forests, depicted in figures 3 and 4 for the two respective density estimators. We used the distance function defined at (3.1), with $d(Y_i, Y_j) = \|Y_i - Y_j\|^2$. The six panels in each figure represent different pairs of values of the smoothing parameter $k = 25, 50, 100$ and gradient

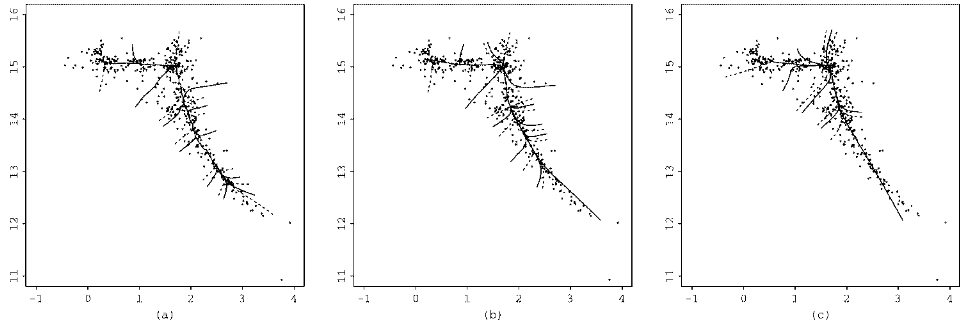

Figure 2: *Ridge Projections*. Panels (a), (b) and (c) show the ridge lines (solid) and antiridge lines (dashed) corresponding to the respective DALs in figure 1. To illustrate relationships to the data, a scatterplot of the data is included in each panel.

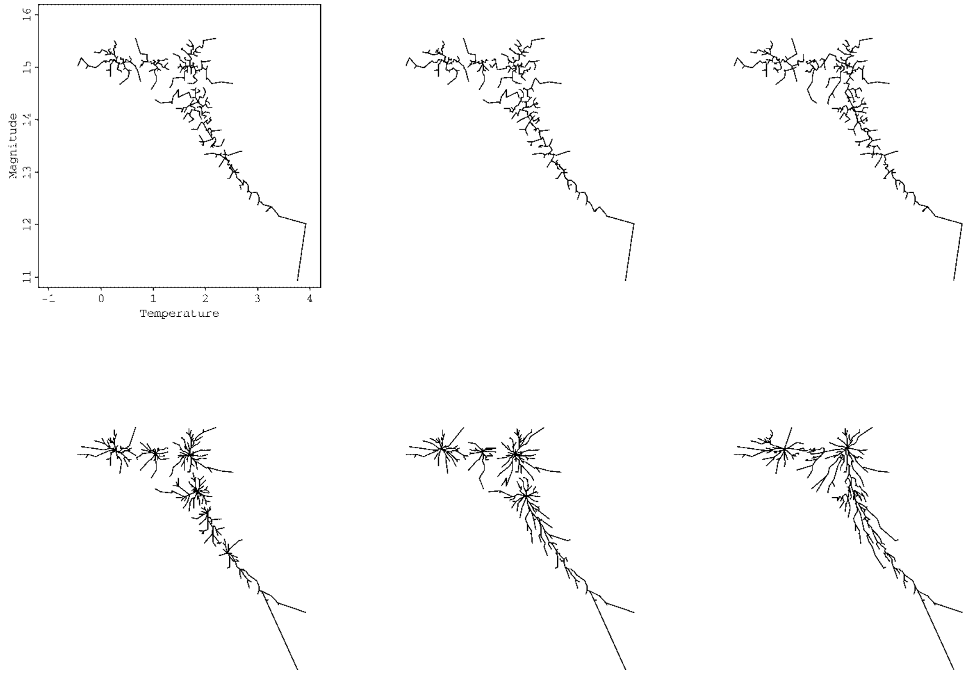

Figure 3: *Forests*. Forests drawn using the unsmoothed nearest neighbour estimator, with $t = 0$ (top row) and $t = 10$ (bottom row), and $k = 25$, 50, 100 (columns 1–3).

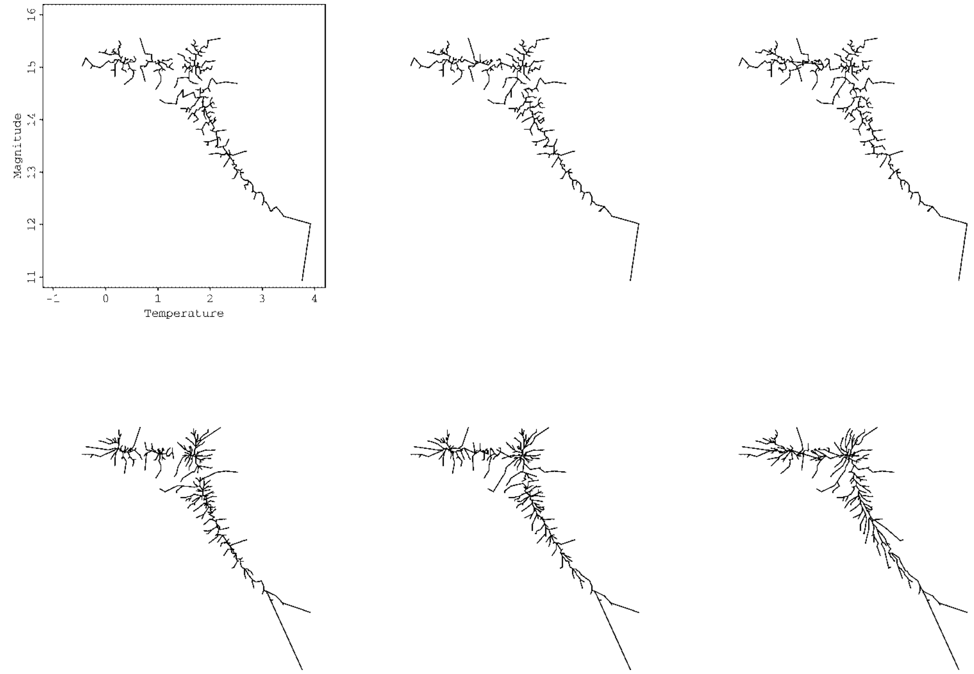

Figure 4: *Forests.* Forests drawn using the smoothed nearest neighbour estimator, with panels ordered as in figure 3.

weight $t = 0, 10$. Taking $t = 0$ produces directed line segments based almost entirely on distances between points, except that the direction of the segment is always that of increasing estimated density. The resulting forest is comparable to the minimum spanning tree, and its links have almost random orientation. On the other hand, using $t = 10$ gives heavy weight to segments that lie close to the representation in Π of the estimated gradient curve, and (for both density estimators) produces a more orderly presentation of the links.

Overall, the data show strong evidence of a northwest to southeast ridge, and at least three modes. Smoothing the density estimator produces some regularisation of forests, but choice of k has much greater effect on our graphs than estimator type.

In order to further illustrate performance of the gradient tree approach, these methods, along with two conventional graphical tools (contour plots and perspective mesh plots), were applied to two simulated data sets. In these examples, which are discussed below, smoothed nearest neighbour estimators were employed whenever estimation of the density and its gradients were required.

In the first example, 500 random variates were generated from the bimodal Normal mixture,

$$0.7\,N\left(\begin{pmatrix}0\\0\end{pmatrix},\begin{pmatrix}1&0\\0&1\end{pmatrix}\right)+0.3\,N\left(\begin{pmatrix}2\\2\end{pmatrix},\begin{pmatrix}0.26&-0.13\\-0.13&0.65\end{pmatrix}\right). \qquad (4.1)$$

The smoothing parameter was $k = 45$, and gradient weight was $t = 10$. The data, contour plots, and perspective mesh plots based on the density estimator, are shown in panels (a) and (b) of figure 5, which provide evidence of bimodality. However, the density ascent lines, ridge lines and forests, depicted in panels (c), (d) and (e)

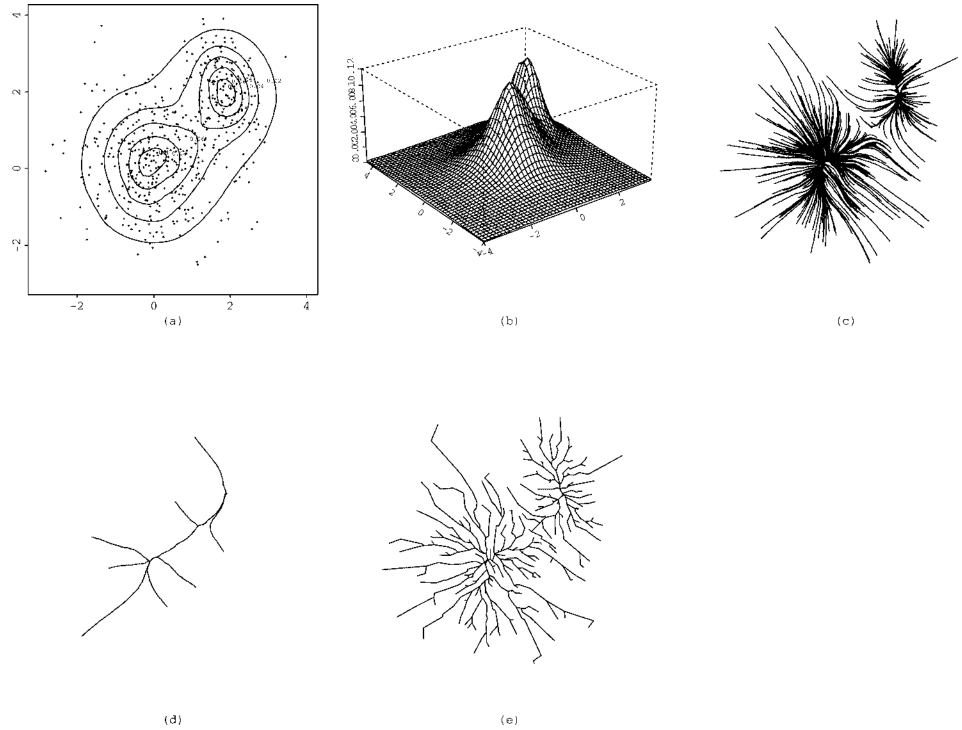

Figure 5: *Bimodal data example.* A scatterplot of 500 random numbers simulated from model (4.1) is shown in panel (a). Panels (a), (b), (c), (d) and (e) depict respectively contour plots, a perspective mesh plot, density ascent lines, ridge lines, and forests based on the smoothed nearest neighbour estimator with $k = 45$ and $t = 10$.

of figure 5, show more clearly than panels (a) and (b) structure of the surface, and in particular the locations of the two modes and the steepest ascent directions up the surface.

Each of the graphical methods illustrated in panels (c) and (d) divides the 500 data points into two subgroups, in which each point is connected to the centre of the subgroup to which it belongs. The directions of the density ascent curves, and hence information about the way in which the surface increases as one moves in different directions, are conveyed much better by these two graphics than by those in panels (a) and (b). Most importantly, panels (c) and (d) allow the reader to extract point-to-point relationships from the data to a significant extent; such information cannot be so readily obtained from the contour plot (panel (a)) or the perspective mesh plot (panel (b)).

The second example is of data simulated from a model, described below, which has more complex structure than that described at (4.1). Let U, V, W, Z be independent random variables, with U and V having the $N(0, 0.06^2)$ distribution, W being uniformly distributed on the interval $(-1, 1)$, and Z having density $g(z) = 0.2z+0.5$ for $|z| \leq 1$. Put

$$X = \text{sgn}(W)\,(0.6 - Z)\,I(-1 \leq Z \leq 0.6) + U, \qquad Y = Z + V, \tag{4.2}$$

where $I(\cdot)$ denotes the indicator function. The surface defined by the joint density of (X, Y) has two ridges, represented by the lines $x = \pm(0.6 - y)$ for $-1 \leq y \leq 0.6$,

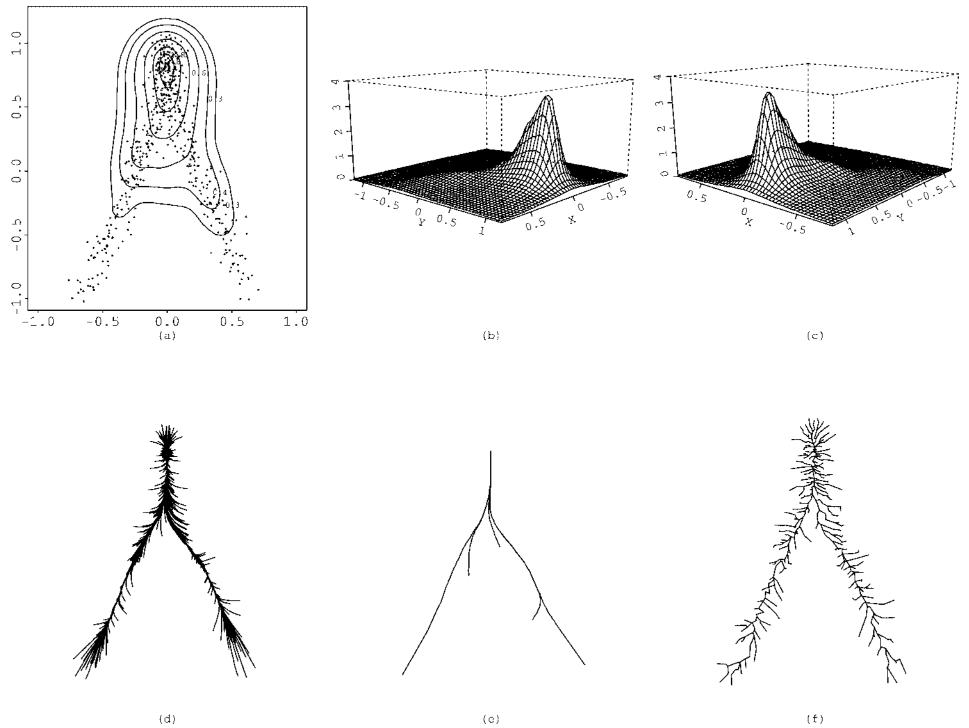

Figure 6: *Ridge data example.* A scatterplot of 500 observations simulated from model (4.2), and the corresponding contour plots, are shown in panel (a). Perspective mesh plots from different angles, showing the two ridge branches, are given in panels (b) and (c). Panels (d), (e) and (f) depict respectively density ascent lines, ridge lines, and forests. Graphics here used the smoothed nearest neighbour estimator with $k = 55$ and $t = 10$.

which merge at $(0, 0.6)$ and then continue together along the line $x = 0$ until the point $(0, 1)$ is reached. The height of the surface increases steadily as one travels along any of these ridges in a direction that has a northbound component.

We generated 500 observations from model at (4.2). The smoothing parameter was taken to be $k = 55$, and the gradient weight was $t = 10$. Panel (a) of figure 6 incorporates a scatterplot of the dataset. The contour plots and perspective mesh plots, given in panels (a)–(c) of figure 6, provide only a vague impression of the bi-ridge nature of the data. In contrast, the density ascent lines, ridge and antiridge lines and forests, shown in panels (d)–(f) of figure 6, provide substantially less ambiguous information about the ridges and, more generally, about the nature of the scatterplot.

The tree and forest structures in different datasets, for example those in our last two examples, are readily compared. In particular the very different characters of the 'octopus plots' (tree structures made up of density ascent lines) in panel (c) of figure 5, and panel (d) of figure 6, are immediately apparent. The first shows two approximately symmetric clusters about single centres, with little evidence of ridges, while the second demonstrates marked asymmetry and 'ridginess'. Likewise, the forests in panel (d) of figure 5, and panel (f) of figure 6, show very different hierarchical structures. The first demonstrates a relatively low level of relationship among different points in the cluster, with many of the branches of the forest joining the cluster relatively close to the respective mode, and so being related to other branches (and hence other points in the cluster) largely through that mode. On the other hand, panel (f) of figure 6 shows a strong degree of hierarchy, with each branch of the forest joining its respective 'ridge branch' after travelling only a short distance, and being linked to other branches though the ridge.

5. Density estimators and theory

The two-dimensional nearest-neighbour density estimators used in Section 4 may be described as follows. Given a kernel K, put $\hat{f}(x) = \hat{f}(x|R) = R/nh_x^2$ where h_x is given by

$$\sum_{i=1}^{n} K\left(\frac{X_i - x}{h_x}\right) = R.$$

If K is the uniform kernel, equal to $1/\pi$ within a region $\mathcal{R}$ and 0 elsewhere, then this prescription requires h_x to be such that R data values are contained within the region $x \oplus h_x\mathcal{R}$, which of course is the standard near-neighbour construction. A disadvantage of the uniform kernel, however, is that the resulting estimator is very rough. The second approach discussed in Section 4 uses a bivariate form of the Epanechnikov kernel. Alternatively we could use bivariate biweight or triweight kernels.

We employed the same value of R for all x, so that the bandwidth h_x was relatively small in regions of high data density. Assuming that $R = R(n) \to \infty$ and $R/n \to 0$ as $n \to \infty$ it may be shown that $h_x \sim \{R/n\kappa_1 f(x)\}^{1/2}$ as $n \to \infty$, where $\kappa_j = \int K(v)^j \, dv$. In particular, the effective bandwidth is of size $(R/n)^{1/2}$. Assuming that K is symmetric and f has two bounded derivatives, the bias and variance of $\hat{f}$ are of sizes R/n and $(n/R^3)^{1/2}$, respectively. Therefore, optimal mean-square performance of the estimator $\hat{f}$ is obtained with R of size $n^{5/7}$, in which case mean squared error equals $O(n^{-4/7})$, just as it would be for a traditional second-order kernel estimator. Variance is asymptotic to $(nf^5\kappa_1^3/R^3)^{1/2}\kappa_2$.

Note particularly that, using bandwidths of these sizes, our gradient estimators are consistent for the true gradients. That is not true for the implicit gradient estimators employed in a minimum spanning tree, which are in effect based on a bandwidth that is of size $n^{-1/2}$. This means that the error-about-the-mean term in the estimator of f, let alone for estimators of the derivatives of f, does not converge to zero, which accounts for the haphazard, complex structure of minimum spanning tree diagrams.

References

[1] Asimov, D. (1985). The grand tour: a tool for viewing multidimensional data. *SIAM J. Sci. Statist. Comput.* **6**, 128–143. MR773286

[2] Cheng, M.-Y. and Hall, P. (1998). Calibrating the excess mass and dip tests of modality. *J. Roy. Statist. Soc. Ser. B* **60**, 579–590 MR1625938

[3] Cook, D., Buja, A., Cabrera, J. and Hurley, C. (1995). Grand tour and projection pursuit. *J. Computat. Graph. Statist.* **4**, 155–172.

[4] Everitt, B. S. (1974). *Cluster Analysis.* Halstead, London. MR455213

[5] Florek, K., Lukaszwicz, J., Perkal, J., Steinhaus, H., and Zubrzycki, S. (1951). Sur la liaison et la division des points d'un ensemble finit. *Colloq. Math.* **2**, 282–285. MR48832

[6] Friedman, J. H. and Rafsky, L. C. (1981). Graphics for the multivariate two-sample problem. (With discussion.) *J. Amer. Statist. Assoc.* **76**, 277–295. MR624331

[7] Friedman, J. H. and Rafsky, L. C. (1983). Graph-theoretic measures of multivariate association and prediction. *Ann. Statist.* **11**, 377–391. MR696054

[8] Hall, P., Qian, W. and Titterington, D. M. (1992). Ridge finding from noisy data. *J. Computat. Graph. Statist.* **1**, 197–211. MR1270818

[9] Hartigan, J. A. (1975). *Clustering Algorithms*. Wiley, New York. MR405726

[10] Hartigan, J. A. (1982). Classification. In: *Encyclopedia of Statistical Sciences* **2**, Eds. S. Kotz and N. L. Johnson, pp. 1–10. Wiley, New York. MR670950

[11] Hartigan, J. A. and Hartigan, P. M. (1985). The DIP test of unimodality. *Ann. Statist.* **13**, 70–84. MR773153

[12] Hubert, L. J. (1974). Approximate evaluation techniques for the single-link and complete-link hierarchical clustering procedures. *J. Amer. Statist. Assoc.* **69**, 698–704. MR373170

[13] Jardine, N. and Sibson, R. (1971). *Mathematical Taxonomy*. Wiley, New York. MR441395

[14] Jones, R. H. and Stewart, R. C. (1997). A method for determining significant structures in a cloud of earthquakes. *J. Geophysical Res.* **102**, 8245–8254.

[15] Kuiper, F. K. and Fisher, L. (1975). A Monte Carlo comparison of six clustering procedures. *Biometrics* **31**, 777–784.

[16] Müller, D. W. and Sawitzki, G. (1991). Excess mass estimates and tests for multimodality. *J. Amer. Statist. Assoc.* **86**, 738–746. MR1147099

[17] Pruzansky, S., Tversky, A. and Carroll, J. D. (1982). Spatial versus tree representations of proximity data. *Psychometrika* **47**, 3–24.

[18] Rees Jr., R. F. (1993). New proper motions in the globular cluster M5. *Astron. J.* **106**, 1524–1532.

[19] Silverman, B. W. (1981). Using kernel density estimates to investigate multimodality. *J. Roy. Statist. Soc. Ser. B* **43**, 97–99. MR610384

[20] Swayne, D. F., Cook, D. and Buja, A. (1991). XGobi: Interactive dynamic graphics in the X window system with a link to S. In *ASA Proceedings of the Section on Statistical Graphics*, pp. 1–8.

[21] Tierney, L. (1990). *LISP-STAT, An Object-Oriented Environment for Statistics and Dynamic Graphics*. Wiley, New York.

[22] Wegman, E. J. and CARR, D. B. (1993). Statistical graphics and visualization. In: *Handbook of Statistics* **9***: Computational Statistics*, Ed. C. R. Rao, pp. 857–958. North Holland, Amsterdam.

[23] Wegman, E. J., Carr, D. B. and LUO, Q. (1993). Visualizing multivariate data. In: *Multivariate Analysis: Future Directions*, Ed. C. R. Rao, pp. 423–466. North Holland, Amsterdam. MR1246351

[24] Wishart, D. (1969). A generalization of nearest neighbour which reduces chaining effects. In: *Numerical Taxonomy*, Ed. A. J. Cole, pp. 282–311. Academic, London.