行政院國家科學委員會專題研究計畫成果報告

估計陡坡樹狀圖

計畫編號：NSC 89-2118-M-002-007

執行期限：88 **年** 8 **月** 1 **日至** 89 **年** 9 **月** 30 **日**

主持人:：鄭明燕副教授　　國立臺灣大學數學系

一、中文摘要

我們介紹一種描述樣本點與點關係的樹狀圖。我們的方法間接利用密度函數的坡度估計值。樹狀圖中兩點相連對等於經過此兩點，沿密度函數曲面，的最陡坡曲線會合於同一聚點。這些最陡坡曲線在樣本平面上的投影稱為陡坡樹狀圖。陡坡樹狀圖有規律的結構而且由非參數密度函數估計，若其能一致估計密度函數及其導數，建造的它的類似體可對它一致估計。此外，我們建議一種樹叢圖，其中樣本點之間以與部分陡坡樹狀圖相近的線段連接。一個樹叢圖是一個規律化的最小全距圖。建構樹叢圖的密度曲面估計使用的帶寬較建構最小全距圖的密度曲面估計使用的帶寬大，所以樹叢圖具有遠較為規則的形狀。

Abstract:

We suggest new approaches to constructing tree diagrams that describe associations among points in a scatterplot. Our methods are based implicitly on gradient estimates. In our tree diagrams, two data points are 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 a regularization of a minimum spanning tree. 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.

Keywords: Density ascent line, density estimation, forest, gradient tree, minimum spanning tree, nearest neighbor methods, ridge estimation, tree diagram.

二、緣由與目的

Gradient trees capture topological features of multivariate probability densities, such as

modes and ridges. 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. We also propose a new technique for describing, and presenting information about, neighbor relationships for spatial data.

The gradient curves of a multivariate density f are the curves of steepest ascent up the surface S defined by $y = f(x)$. The representations of gradient curves, in the sample space, are 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.

The most familiar tree diagram in multivariate analysis is the minimum spanning tree, or MST (Florek *et al.*, 1951; Friedman and Rafsky, 1981, 1983), which is the graph of minimum total length connecting all sample points. The MST is an estimator of the gradient tree that arises when we take $\hat{f}$ to be the nearest neighbor density estimators, in which the estimate at each point is inversely proportional to a monotone function of the distance to the closest sample point. This is a poor estimator of f, and so it is not surprising that the MST is a poor estimator of the population gradient tree. We suggest gradient tree estimators that are asymptotically consistent for the population gradient tree, and which also improve on the MST.

We also suggest algorithms for drawing `forests', using either the full dataset or subsets that have been identified by the gradient tree. A forest provides information about relationships among neighboring data.

三、結果與討論

Let $\mathcal{X} = \{X_1, ..., X_n\}$ be a sample observed from f. Write Π for the sample space. 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 S in the direction of steepest ascent.

If the trajectory on S is represented as the locus of points $\left(x^{(1)}(s), x^{(2)}(s), y(s)\right)$, 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 $\left(x^{(1)}(s), x^{(2)}(s)\right)$, 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.

Let $D(f) = \left(f_1^2 + f_2^2\right)^{1/2}$ and put $\mu_j = f_j / D(f)$ and $\mu = (\mu_1, \mu_2)$. Then, for $x \in S$, $\mu(x)$ is the unit vector in Π representing the direction of steepest ascent up S, at the point $(x, f(x)) \in S$. The DAL

that passes through $x \in \Pi$ is represented by the infinitesimal transformation, $x \to x + \mu(x)ds$, 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 $x \in \Pi$, and a small positive number $¢$, consider the sequence of points $P \equiv \{x_j, -\infty < j < \infty\}$ defined by , for $j \geq 1$, $x_j = x_{j-1} + \mu(x_{j-1})¢$ and $x_{-j} = x_{1-j} + \mu(x_{1-j})¢$. Thus, the DAL that passes through x_0 represents the limit, as $¢ \to 0$, of the sequence P. The algorithm is convenient for numerical calculation, provided we stop before reaching places where $D(f)$ vanishes.

We suggest a regularization of the minimum spanning tree in which links between observations are penalized if they are not sufficiently close to estimated density ascent lines. It may be applied to a subset $Y = \{Y_1, ..., Y_N\}$ of the sample i as well as to the full sample. Let $\|Y_i - Y_j\|$ be the Euclidean distance in the sample space Π. Define a penalized distance measure D by

$$D(Y_i, Y_j) = \|Y_i - Y_j\|^2 + t\left[\|Y_i - Y_j\|^2 - \{(Y_i - Y_j) \cdot \hat{\mu}(Y_i)\}^2\right].$$

Given Y_i we draw a direct line segment from Y_i to Y_j if and only if Y_j minimizes $D(Y_i, Y_j)$ over all points Y_j for which $\hat{f}(Y_j) > \hat{f}(Y_i)$. The forest is the set of these directed 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 $\hat{S}$ to give more emphasis to the uphill component of motion.

We employed two different versions of $\hat{f}$. Both were nearest neighbor 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 neighbor estimator, with $\hat{f}(x)$ equal to $k/(n®r^2)$ where $r = r(x)$ was the smallest number such that the circle centered on x and with radius r contained just k points. The second density estimator was a smoothed version of the first, equal to $2k/(n®r^2)$ where r was the solution of $\sum_{i=1}^{n} \left\{1 - \left(\frac{\|X_i - x\|^2}{r}\right)\right\}^+ = k$, where K is a kernel function.

四、計劃成果自評

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.

Our forest 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

$\hat{S}$ of the surface S. Our approach to constructing a forest allows the experimenter to choose 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 $\hat{S}$.

五、參考文獻

1. 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.
2. Friedman, J.H and Rafsky, L.C. (1981). Graphics for the multivariate two-sample problem. (With discussion.) *J. Amer. Statist. Assoc.* **76**, 277--295.
3. Friedman, J.H and Rafsky, L.C. (1983). Graph-theoretic measures of multivariate association and prediction. *Ann. Statist.* **11**, 377--391.