Smoothings, Ridges, and Bumps

Edward J. Wegman
Center for Computational Statistics
George Mason University

and

Qiang Luo
Real Networks, Inc.

1. Introduction

In this paper, we seek to formulate a geometrically based alternative to nonlinear, nonparametric regression. Hastie and Stuetzle (1989) introduced the ideas of principal curves and surfaces. These were 1- and 2-dimensional non-linear regression-type estimators, which were consistent with respect to the mean, i.e. involved conditional expectations. We have developed a similar notion, but with two additional features. First, we propose to construct mode-like estimators rather than mean-like estimators. Second, the geometrically-based procedures we invoke are easily generalizable to higher dimensions.

Consider the intuitive notion of a ridge on a two-dimensional surface. The ridge is, in a sense to be made more precise momentarily, a maximal one-dimensional feature on the two-dimensional surface. One simple fact to observe is that if one considers a point on the ridge and calculates the directional derivatives at that point, the directions of the derivatives with greatest magnitude will be orthogonal to the ridge. The support of the ridge is defined as the *skeleton.* For us, the utility of the skeleton is as a geometrically-derived summary statistic. In a classical bivariate normal density function with $\rho \neq 0$, the ridge will lie exactly over the major axis of the elliptical contours. Thus the skeleton will coincide with the major axis. However, because of its geometric inspiration there is no particular need for the skeleton to be linear and, in fact, there is no particular need for the skeleton to be a proper non-intersecting Euclidean manifold. Thus it is a highly flexible non-linear, nonparametric generalization of the classic regression curve.

2. Formal Definitions of Ridges and Skeletons

The ridge and skeleton notions can be generalized to arbitrary dimensions. Consider a vector $\boldsymbol{x}$ in d-dimensional space with probability density function $f(\boldsymbol{x})$. Let $\mathcal{R}$ be a compact k-dimensional, $0 < k < d$, smooth manifold on the density surface with supporting manifold $\mathcal{S}$. The *likelihood of* $\mathcal{R}$ *on* $\mathcal{S}$ is defined as the k-dimensional integral of $f(\boldsymbol{x})$ on $\mathcal{S}$ which is the $(k+1)$-dimensional hypervolume under $\mathcal{R}$. Although this is not the likelihood in the traditional statistical sense of the word, it corresponds to the appropriate marginal probability. $\mathcal{S}$ and $\mathcal{R}$ are said to be *consistent with respect to the* *mode* if each point $\boldsymbol{x}$ on $\mathcal{S}$ is the conditional mode of the density conditioned on the $(d-k)$-normal plane of $\mathcal{S}$ at $\boldsymbol{x}$. The *k-ridge of a d-dimensional density* is the k-dimensional smooth consistent manifold $\mathcal{R}$ that maximizes the likelihood defined as the hypervolume under the manifold. The *k-skeleton* is the support of the k-ridge.

Two simple observations are of interest. First, the 0-skeleton coincides with the mode. The 0-ridge is just the value of the density at the mode. Second, the ridges and skeletons are nested. That is, the k-ridge is contained in the $(k+1)$-ridge and, similarly, the k-skeleton is contained in the $(k+1)$-skeleton. This hierarchical arrangement is informative in determining the higher-dimensional structure of the data. As mentioned earlier, Hastie and Stuetzle (1989) introduced the ideas of principal curves and surfaces. These were 1- and 2-dimensional non-linear regression-type estimators, which were consistent with respect to the mean, i.e. involved conditional expectations. The 1- and 2-skeletons are their analogs except that they are consistent with respect to the mode.

3. Algorithms

The k-skeleton is computed by a techniques we call *gradient traces*. The density gradient is calculated beginning at each observation. The trace at a particular observation, say $\boldsymbol{x}_j = \boldsymbol{x}_j^0$, is computed by taking a step length proportional to the magnitude of the gradient in the direction of the gradient vector. Call the endpoint of our first step $\boldsymbol{x}_j^1$. We follow a steepest ascent-type algorithm. The gradient is recomputed at $\boldsymbol{x}_j^1$ and a second step is taken with the step length again proportional to the gradient at $\boldsymbol{x}_j^1$ in the direction of the gradient vector to compute $\boldsymbol{x}_j^2$. This process is repeated for general i to obtain $\boldsymbol{x}_j^i$. As $i \to \infty$, $\boldsymbol{x}_j^i$ approaches the local mode, but, in fact, through the nested sequence of k-ridges.

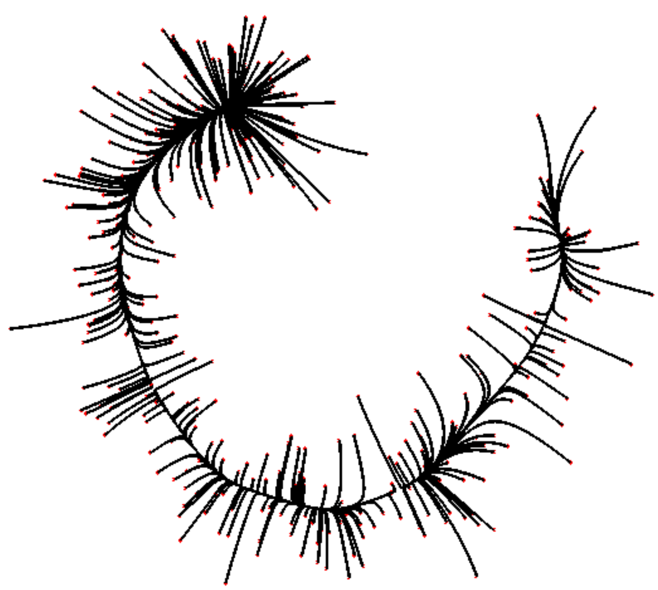

Figure 1b: Gradient trace for a highly non-linear "C-shaped" curve.

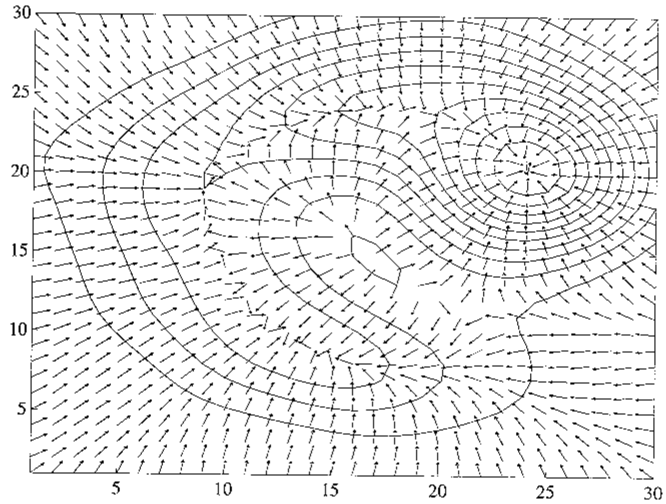

Figure 1b: The C-density viewed as a potential energy surface.

This view of the density is essentially a view of the density as a potential surface with trajectories, basins of attractors and separatrix. The separatrix essentially corresponds to the k-ridge. The nested character is a most interesting feature. Essentially the vector direction of the gradient will point to the direction of the highest dimensional ridge, say k-ridge. Once the trajectory is traced to the k-ridge, the trajectory will sharply turn in essentially an orthogonal direction and the gradient vector will point in the direction of the next highest dimensional ridge, in general the $(k-1)$-ridge, but not necessarily. This process repeats until the 0-ridge is reached. The number of turns in the trajectory is a diagnostic indicator of the dimensionality of the highest dimensional summary statistic. It may turn out, for example, that the highest dimensional ridge for three-dimensional data is one-dimensional.

Moreover by *shaving the trajectories* from the observations to the modes, i.e. by sequentially eliminating the links from $\boldsymbol{x}_j^{i-1}$ to $\boldsymbol{x}_j^{i}$ starting at $i=1,$ one obtains a nested sequence of k-ridge estimators. For at least 0-, 1-, and 2-dimensional ridges these can be visualized by standard computer graphics techniques. In addition, contours of three-dimensional ridges can also be visualized. The advantage of taking step sizes proportional to the magnitude of the gradient is that the gradient trace tends to take large steps initially, which aids in the trajectory shaving process, but much smaller steps near the local mode, which allows sorting out of fine structure near the modes.

Figure 1a illustrates the gradient trace for a two-dimensional nonlinear ridge, while Figure 1c illustrates the shaved version of that same curve, i.e. the 1-skeleton estimator. Figure 2 illustrates an example with both a 0- and a 1-dimensional ridge component. Figure 3a illustrates the gradient trace of a one-dimensional ridge in three dimensions while Figure 3b illustrates the corresponding one-dimensional skeleton after shaving..

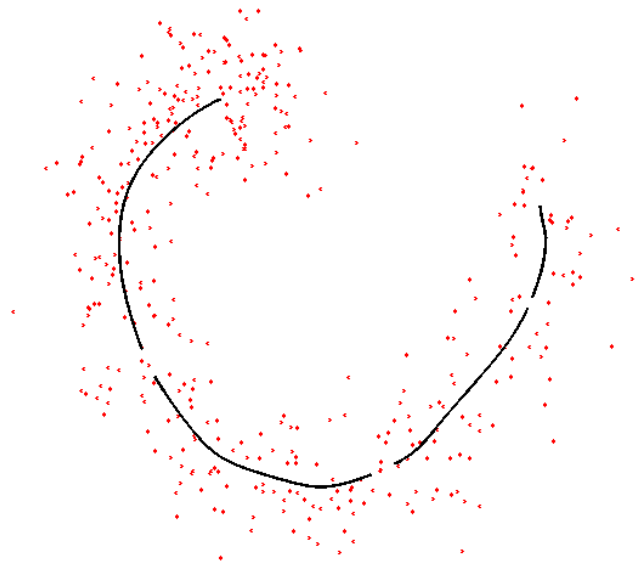

Figure 1c: Estimated 1-skeleton based on shaving the gradient traces in Figure 1a. The gaps correspond to saddle points on the density surface. Thus in this example there are 4 local modes.

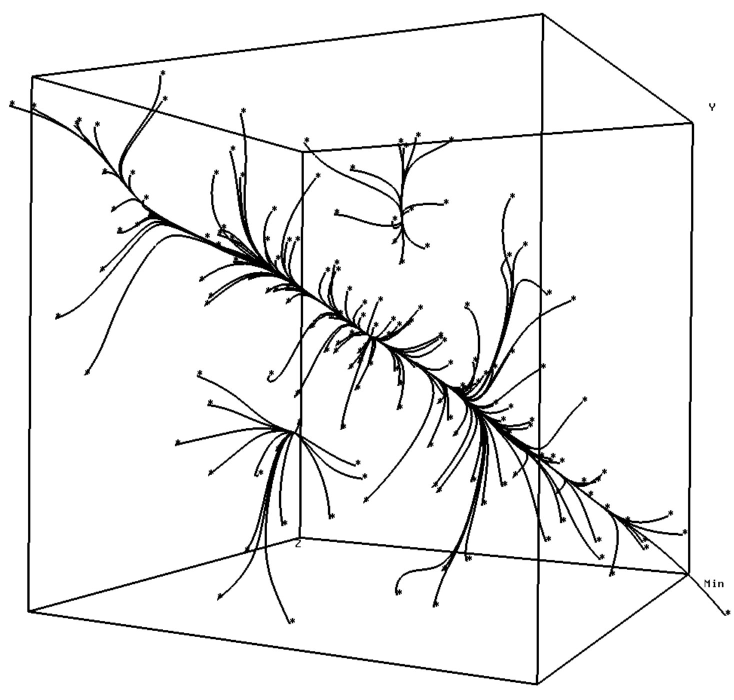

Figure 2: An example in 3 dimensions in which there are both 0-skeletons and a 1- skeleton.

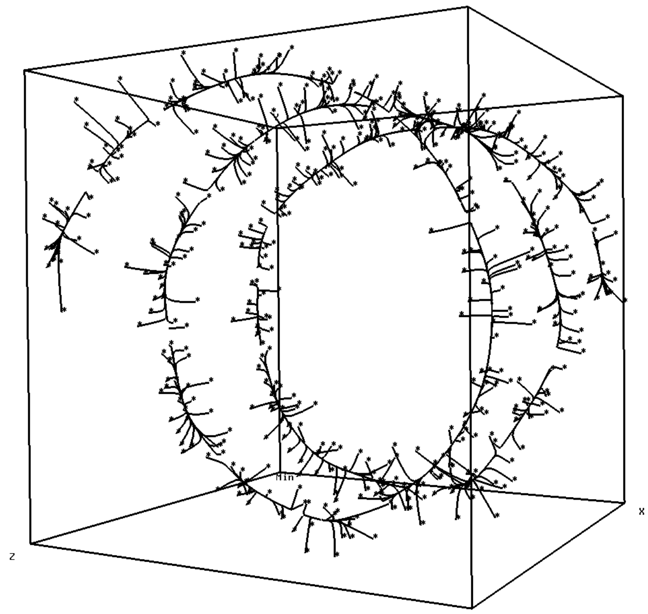

Figure 3a: A one dimensional gradient trace in three dimensions.
The basic curve is a helix.

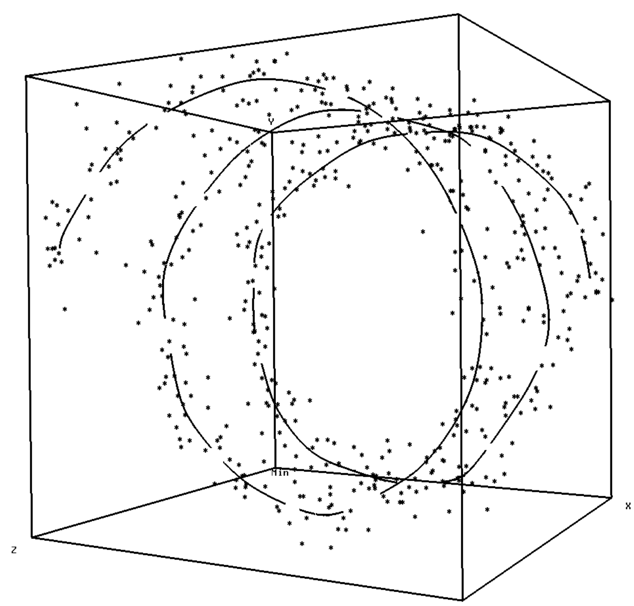

Figure 3b: 1-skeleton estimated by shaving gradient traces in Figure 3a. As in Figure 1c, the gaps correspond to saddle points on the three-dimensional density.

One interesting observation is that with a univariate normal density, the points of inflection occur at $\mu \pm \sigma$. If one computes and plots the magnitude of the gradient, the mode of the density will become a minimum and the points of inflection will become local modes (i.e. 0-ridges). In the same way, if one computes the magnitude of the gradient of a two-dimensional density, the 1-ridge becomes a local minimum and the flanking "lines-of-inflection" become 1-ridges. In general the k-ridges of the function defined as the magnitude of the gradient of the density can be interpreted as confidence bands for the k-ridge of the density. Thus this method not only gives a completely nonparametric, geometrical generalization of regression, but also yields a method for geometrically generating approximate confidence bands for the generalized, nonparametric nonlinear regression estimator.

4. Discussion

In related work, Carr et al. (1986) show an anaglyph stereo example with short segments orthogonal to steepest ascent curves located at uniform points on the domain of the surface. This work was aimed at developing algorithms for contours. Carr (1991) discusses an alternate method for calculating skeletal structures based on gray level thinning of hexagon and truncated octahedron binned 2-D and 3-D data.

Cheng, Hall, and Hartigan (2001) introduce the concept of *gradient trees* as a tool for cluster analysis as an improved version of minimal spanning trees. There gradient trees are essentially the same notion as our gradient traces. They demonstrate a connection with the estimation of ridges on the surface of two-dimensional densities and show that, under appropriate bandwidth constraints, estimated gradient trees are consistent for the true underlying gradient trees.

Acknowledgements:

This work was completed under the sponsorship of the Office of Naval Research under contract DAAD19-99-1-0314 administered by the Army Research Office, by the Air Force Office of Scientific Research under contract F49620-01-1-0274 and contract DAAD19-01-1-0464, the latter also administered by the Army Research Office and finally by the Defense Advanced Research Projects Agency through cooperative agreement 8105-48267 with the Johns Hopkins University. This paper was presented as an invited paper at JSM 2002.

References:

Carr, D. B. (1991) "Looking at large data sets using binned data plots," ***Computing and Graphics in Statistics***, Eds. A. Buja and P. Tukey, Springer-Verlag, New York. (IMA Vol. 36, Proceeding of the IMA summer program 1989.), 7 - 39.

Carr, D. B., W. L. Nicholson, R. J. Littlefield, and D. L. Hall. (1986) "Interactive color display methods for multivariate data." ***Statistical Image Processing and Graphics***, Eds. E. J. Wegman and D. J. DePriest, pp. 215-250, Marcel Decker, New York.

Cheng, M-Y, Hall, P. and Hartigan, J. (2001) "Estimating gradient trees," unpublished manuscript.

Hastie, T. and Stuetzle, W. (1989) "Principal curves," ***Journal of the American Statistical Association***, 84(406), 502-516.

Luo, Q. and Wegman, E. J. (1991) ***Mason Ridge***, copyright (c) 1990, 1991, a UNIX package for Silicon Graphics workstations for two- and three-dimensional density rendering using stereoscopic displays, transparency and lighting models and for multidimensional ridge and skeleton estimation.

Wegman, E. J., Carr, D. B. and Luo, Q. (1993) "Visualizing multivariate data," in ***Multivariate Analysis: Future Directions***, (Rao, C. R., ed.), Amsterdam: North Holland, 423-466.

Wegman E. J. and Carr, D. B. (1993) "Statistical graphics and visualization," in ***Handbook of Statistics 9: Computational Statistics***, (Rao, C. R., ed.), Amsterdam: North Holland, 857-958.