Consequences of Spurious Modes in Density Estimates for Defining Clusters in Multiple Dimensions

David W. Scott[1] and William F. Szewczyk[2]

October 20, 1998

SUMMARY: The use of nonparametric density estimators to define of clusters is growing as a serious analytical tool. Clusters may be defined by high density regions, or by the location of sample modes. Minnotte and Scott (1993) showed how a tree directly analogous to a dendrogram can be produced by recording the location of sample modes as the smoothing parameter varies. A number of general principles of density estimation are relevant. First, smoothing may obscure features in the data/density. Second, the number of sample modes decreases (monotonically) as the smoothing parameter increases. Third, dimension reduction can lead to superposition of features and loss of distinct clusters. In this note, we provide counterexamples to each notion.

KEY WORDS: Kernel function, mode tree, projection, dendrogram.

[Note: Shorten title? Start with "Examples of Spurious..."?]

[Note: Start with mode tree of Iris data??? They are rounded so...]

[1]David W. Scott is Professor, Department of Statistics (MS–138), Rice University, 6100 Main Street, Houston, TX 77005-1892. This research was supported in part by the National Science Foundation and the Department of Defense.

[2]William F. Szewczyk is Senior Mathematician with the Department of Defense.

1. INTRODUCTION

Clustering remains one of the great challenges in statistical practice, requiring great skill and artfulness on the part of the practitioner. Hierarchical methods based on various distance metrics and agglomeration rules have been widely applied with much success but well-documented limitations. The results of such an analysis may be summarized graphically by a tree figure such as the dendrogram, in which the vertical scale indicates the value for which clusters are formed.

Density estimators have also been used to define clusters through the notion of high-density regions. Since no single fixed-bandwidth kernel estimator can capture all modes, this approach has not been widely adopted. However, Silverman (1981) proposed a bootstrap test for the number of modes that introduced the notion of a "critical bandwidth." When the smoothing parameter takes on the value of a critical bandwidth, there are k modes and one saddle point in the estimate. Any decrease in the bandwidth results in an additional sample mode. For univariate data, Silverman proved that a Gaussian kernel was sufficient for the monotonicity property. Minnotte and Scott demonstrated that a number of other kernels did not have the monotonicity property. xxx proved that the Gaussian kernel was necessary (again, in the univariate case).

Minnotte and Scott (1993) first pointed out that the modes of the entire family of kernel density estimates can be captured in the mode tree. The mode tree assumes the appearance of a dendrogram, but the vertical dimension is the value of the smoothing parameter and the mode tree contains the number (and location in low dimensions) of the sample modes at that bandwidth. As the smoothing parameter increases, modes collapse into bigger modes in a monotone fashion. Minnotte demonstrated a local version of Silverman's bootstrap test that was more successful in discovering the correct number of modes when different bandwidths are required. The mode tree was used to record the results of the many local simulations.

Of course, associating clusters with modes is more restrictive that assuming a mixture model. In the latter case, clusters may manifest in the density as modes, bumps, or neither. However, in multiple dimensions, the shape of the mixing density has a large impact on the model fit. We find the mode–cluster duality appealing. The curse of dimensionality suggests both finding modes in high dimensions as well as the problem of spurious modes are both likely. Bumps in higher dimensions are even more susceptible to noise. Thus we focus on modes for the moment.

In the following, we demonstrate some features of real density estimation that must be taken into account when using multivariate density estimation to determine clusters. These lessons have application to mixtures, as a kernel estimate is in fact a (large) mixture density.

2. MONOTONICITY OF MULTIVARIATE MODES UNDER PROJECTION

The monotonicity of the number of modes in dimensions of two or more seems reasonable. Monotonicity in multiple dimensions takes on two distinct forms. One corresponds to our discussion in the previous section, namely, increasing values of the smoothing parameter give fewer modes. The second notion is that the projection of a multimodal density should have fewer modes. However, the image processing community has been aware of counterexamples of the second type found in smoothing images by convolution operations. Here, we give an explicit demonstration of this "large sample" problem. Unexpectedly, we follow with a "sample sample" example that is completely different from the first type.

The problem arises in images when objects have narrow features. We can mimic that with a density function that is like a barbell, but unimodal. We begin by defining a unimodal mixture, with two (weighted) standard normal densities at the ends ($\mu_1 = 0$ and $\mu_2 = 10$, connected by a mixture of 101 equally spaced normals between the means, but with $\sigma = 1/10$. The weights on the 101 normals were selected so that the "ridge" function was monotone decreasing for $x > 0$, and hence, unimodal. In Figure 1, the 103 normal components are displayed together with their sum (the "ridge trace"). Figure 1 actually displays the conditional density, $f(x|y = 0)$. The full bivariate density, $f(x, y)$, was obtained by making each mixture (univariate) density a bivariate normal, with the same variance along each coordinate. A perspective plot of the bivariate (unimodal) density is shown in Figure 2.

Since the projection of a mixture density is a mixture of the projection of the individual densities, finding the projection of the density in Figure 2 onto the x-axis is particularly simple. The result is shown in Figure 3, which displays a clearly bimodal univariate projection. Intuitively, the ridge connecting the two barbells is relatively narrow; hence, it does not contain much probability mass. Making the ridge even narrower would further enhance the bimodality of the projection.

This example illustrates how high-dimensional features such as bumps (not containing a mode) may manifest themselves as simple modes in a lower dimension. Thus a mode in a projected view may or may not correspond to a mode in the full dimension. Furthermore, a projected view may contain more (not always less) modes than in the full dimension.

3. MONOTONICITY OF MULTIVARIATE MODES VERSUS BANDWIDTH

Following the univariate result that the normal kernel is necessary and sufficient for monotonicity in the number of sample modes, it clearly follows that any multivariate kernel estimator satisfying this property must have kernel that has normal slices when sliced in any direction. Thus a bivariate normal kernel is required in two dimensions with $\Sigma = I_2$. (Imagine the bivariate data lying close to a straight line.)

We experimented with kernel estimates of small data sets with points on a regular grid. Such configurations are important in practice because pre-binning algorithm are commonly employed to speed computation of multivariate densities (Scott, 1992). Much to our surprise, we found that three data points on the vertices of an equilateral triangle could give rise to four (4!) sample modes. Such a configuration might arise with hexagonal binning (Scott, 1988; Carr, 19xx). Taking 3 points on the unit circle (shifted to the left for plotting purposes) and a bandwidth of $h = 0.7355$ gives four modes of essentially equal height; see Figure 4. The three data points are highlighted by the triangle for clarity.

The fourth (false) mode is located at the mean of the three data points. For h sufficiently large, the density estimate is in fact unimodal at this center point. As the bandwidth h decreases, we expect the sequence $1 \to 3$ for the number of modes. However, as may be seen in Figure 5, the sequence $1 \to 4 \to 3$ is in fact observed as h decreases.

Figure 5 clearly demonstrates that we are not observing an event with zero probability. In fact, $h = 0.739098$ is the smallest bandwidth displaying exact one mode, while $h = 0.709967$ is the largest bandwidth displaying exactly three modes. In between, the density has four modes. While clustering algorithms can be designed to ignore such "phantom" or "anticipatory" modes, numerical algorithms can be directed to such modes if care is not exercised.

Thus no necessary and sufficient conditions can be given for monotonicity of modes in the multivariate case. In recent presentations, Marchette (1997) has indicated that filtered kernel estimates (unequal smoothing parameters and/or weights) need not be monotone even in the univariate setting. Again, such surfaces are very nearly flat (and more so than Figure 5 may suggest, since extra contours were added to the equally spaced contours levels to emphasize the modal structure).

The three points need not lie exactly on the vertices of an equilateral triangle to produce four modes. However, no other regular polygon in two dimensions produces phantom modes. For example, Figure 6 displays the sequence of modes (such as would be obtained for a mode tree) for $n = 5$ points located at the vertices of a regular pentagon (shown). Notice that the five modes emerge continuously from the center as the bandwidth decreases, eventually merging with the original data points. This pattern is observed in all situations we examined. We have not looked closely at higher dimensions.

4. SMOOTHING CAN INDUCE FALSE MODES

In general, smoothing obscures or downplays features (negative bias at peaks; positive bias in valleys). However, it is easy to construct a situation where this is not the case. In fact, we have already provided such as example in Figure 2. Drawing a sample of size $n = 5{,}000$ from this mixture density and constructing a bivariate averaged shifted histogram density estimator (shown in Figure 7), we see that the estimate is at least bimodal. This example is actually more in line with the observation of the image processing community discussed in Section 2. Any convolution of the mixture density rapidly brings down the ridge, leaving the two barbells more prominently featured.

5. THEORETICAL DERIVATION OF THE EXTRA MODE

We have analyzed the normal mixture model in 2–4 dimensions to discover when the extra mode will appear for data points on a regular simplex.

In two dimensions, we consider three data points on the vertices of an equilateral triangle (different orientation than previous figures). Without loss of generality, we let $h = 1$ and vary the radius of a circle about the origin that intersects the vertices (one of which is located at the point $(0, r)$). It is sufficient to examine the y-axis for the locations of modes as the radius, r, varies. Clearly, for large r, we will have 3 modes located a distance r away from the origin. For r near to zero, we will have one mode at the origin. What happens in the transition is depicted in Figure 8. Let the mixture density be denoted by $f_M(x, y)$. Then we examine the set of points where $\nabla_{xy} f_M(x, y) = (0, 0)$. By symmetry, $\partial f_M(x, y)/\partial x = 0$ everywhere on the y-axis, for all radii. The combination of values (r, y) where $\partial f_M(x, y)/\partial y = 0$ is shown by the solid line in Figure 8. Notice that the origin ($y = 0$ and any $r \geq 0$) is always a critical point. (As we shall demonstrate, for $r < \sqrt{2}$, the origin is a mode; otherwise, it is a saddle point.) There is a second solid line where the gradient vanishes. This line approaches the asymptote $y = r$, as we would expect for the locations of the three modes when the components are widely separated. However, when $r < 2$, then the location of the modes is closer to the origin than the location of the three kernels.

To determine which critical points are in fact modes, we must determine if the Hessian, $\nabla^2_{xy} f_M(x, y)$, is negative definite at $(x, y) = (0, r)$ and at $(x, y) = (0, y(r))$, where $y(r)$ represents the gradient solution away from the origin. Fortunately, $\partial^2 f_M(x, y)/\partial x \partial y = 0$ everywhere along the y-axis, so that the eigenvalues of the Hessian matrix are just the values of $\partial^2 f_M(x, y)/\partial x^2$ and $\partial^2 f_M(x, y)/\partial y^2$. The boundary $\partial^2 f_M(x, y)/\partial x^2 = 0$ is shown by the dashed line in Figure 8, while the boundary $\partial^2 f_M(x, y)/\partial y^2 = 0$ is shown by the dot-dash-dot line. The region where both are negative is shaded. The most interesting region around the point $(r, y) = (\sqrt{2}, 0)$ is blown up in Figure 9. From this figure, we see that the origin is in fact a mode for all $r < \sqrt{2}$, while the modes away from the origin abruptly end at the value $y = 0.5103$, which occurs when the three points are at a distance $r = 1.353$ from the origin (determined by using FindRoot in Mathematica). Thus there are four modes with three data points when $1.353 < r < \sqrt{2}$. Clearly, the extra mode is not a zero-probability event, and can occur occasionally in a number of situations.

In Figure 6, we saw that the five modes smoothly approach the origin. In fact, the same is true for all regular polygons with more than three vertices. Consider the case of 4 vertices, with one vertex again located at the point $(0, y)$. In Figure 10, we plot the same information contained in Figure 8. Observe how the gradient and Hessian contours intersect only at the point $(r, y) = (\sqrt{2}, 0)$. All calculations were done exactly in Mathematica.

What about higher dimensions? We have examined a number of special cases, and found that the bivariate case generalizes. In particular, a placing four trivariate normal kernels at the vertices of a tetrahedron centered at the origin and at a distance r from the origin will display five (5) modes when $1.5537 < r < sqrt3$. If one vertex is located at the point $(x, y, z) = (0, 0, r)$, then it is sufficient to look at all critical points located on the z-axis. By symmetry, the partials with respect to x and y vanish, and the Hessian matrix is again diagonal. Furthermore, the entries $\nabla_{xx} f_M(x, y, z)$ and $\nabla_{yy} f_M(x, y, z)$ are identical along the z-axis. Therefore, we may construct Figure 11, which exactly parallels the bivariate case shown in Figure 9. (Eight points on a cube, not shown, did not manifest an extra mode.)

We went one dimension further, and examined the simplex with five vertices in $\Re^4$. With one vertex located at the point $(x, y, z, t) = (0, 0, 0, r)$, we find a sixth mode whenever $1.6887 < r < 2$; see Figure 12. If anything, the likelihood of an extra mode is increasing with dimension.

6. DISCUSSION

Kernel estimation and projection are both smoothing operations. In general, such operations tend to either reduce the prominence of real features or mask them through superposition. However, we have shown examples that are not so rare in practice where smoothing in fact has the opposite effect. Algorithms for clustering and projection can be designed to work around these facts, but this is not enough.

If, indeed, we accept the mode-cluster duality then tests must be developed to help identify true modes in the data. Much work has been done in the univariate case (e.g. Hartigan and Hartigan (19xx), Silverman (1981), and Minnotte and Scott (1993)), but little has been done in the multivariate case. We believe that careful application of the results in this paper (and other multivariate phenomena less well understood) can help.

REFERENCES

Carr, D.B. (19xx). Histogram glyph paper.

Hartigan, J.A. (19??). High density cluster paper.

Hartigan, J.A. and Hartigan, ?? (19xx). Dip test.

Marchette, D.? (1997), "???," *Computing Science*, in press.

Minnotte, M.?., and Scott, D.W. (1993). "The Mode Tree: A Tool for Visualization ...," *J. Computational and Graphical Statistics*, xx:xxx-yyy.

Polonik et al (19??). Excess mass paper.

Rudemo, M.?. (19??), Reference talks about necessity of Normal kernel???

Scott, D.W. (1988). "A Note on Choice of Bivariate Histogram Bin Shape," *J. Official Statistics*, 4:47–51.

Scott, D.W. (1992). *Multivariate Density Estimation: Theory, Practice, and Visualization*. John Wiley & Sons, New York.

Silverman, B.W. (1981). "Using Kernel Density Estimates to Investigate Multimodality," *J. Royal Statistical Society, B*, 43:97–99.

[Hartigan paper first to push density – cluster link....also kNN way]

[Look over Everitt clustering book again for other examples.]

[Find IEEE paper on the necessity of normal kernel in 1-D.]

[Mixture pdf's. Mention Titterington, too?]

[Bopok — image processing examples a la Marron's reference.]

[Marchette — other mode tree ideas]

[Discussion: some relevance to pre-binning....]

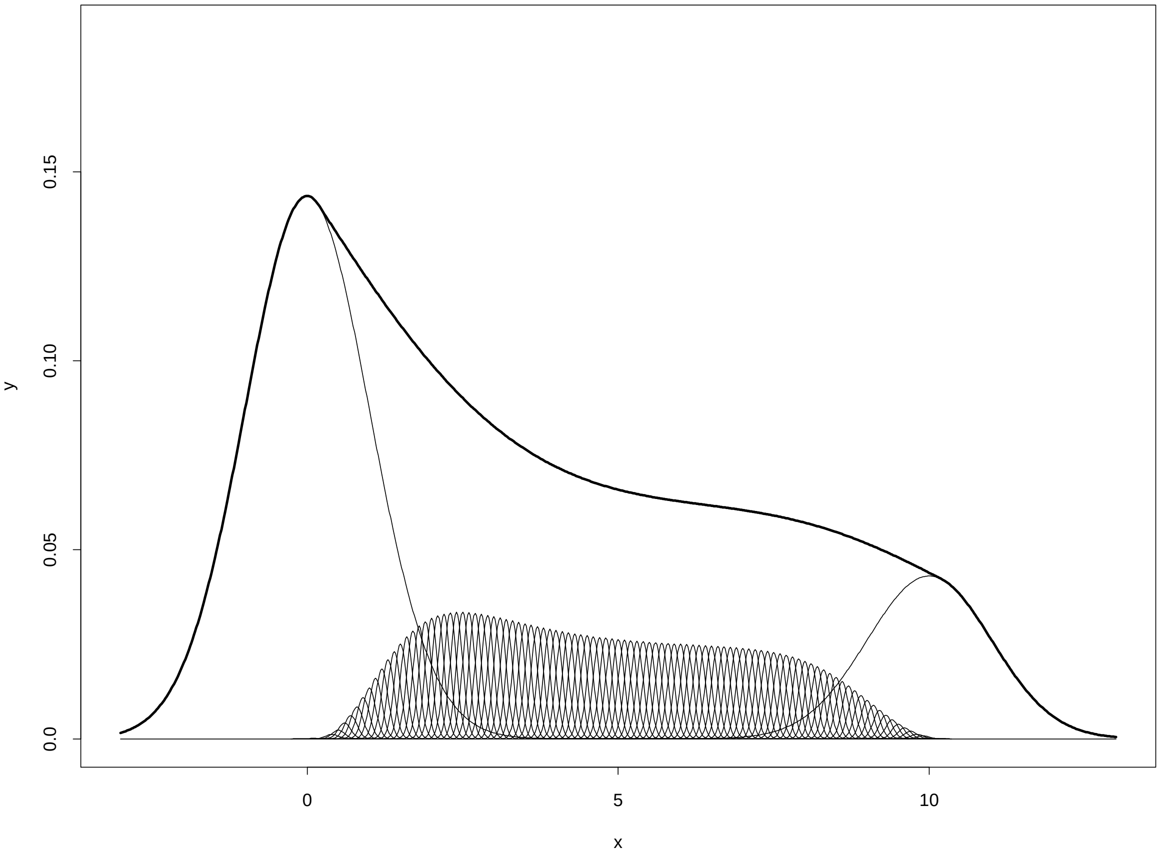

Figure 1: Construction of 1-D ridge of bivariate density.

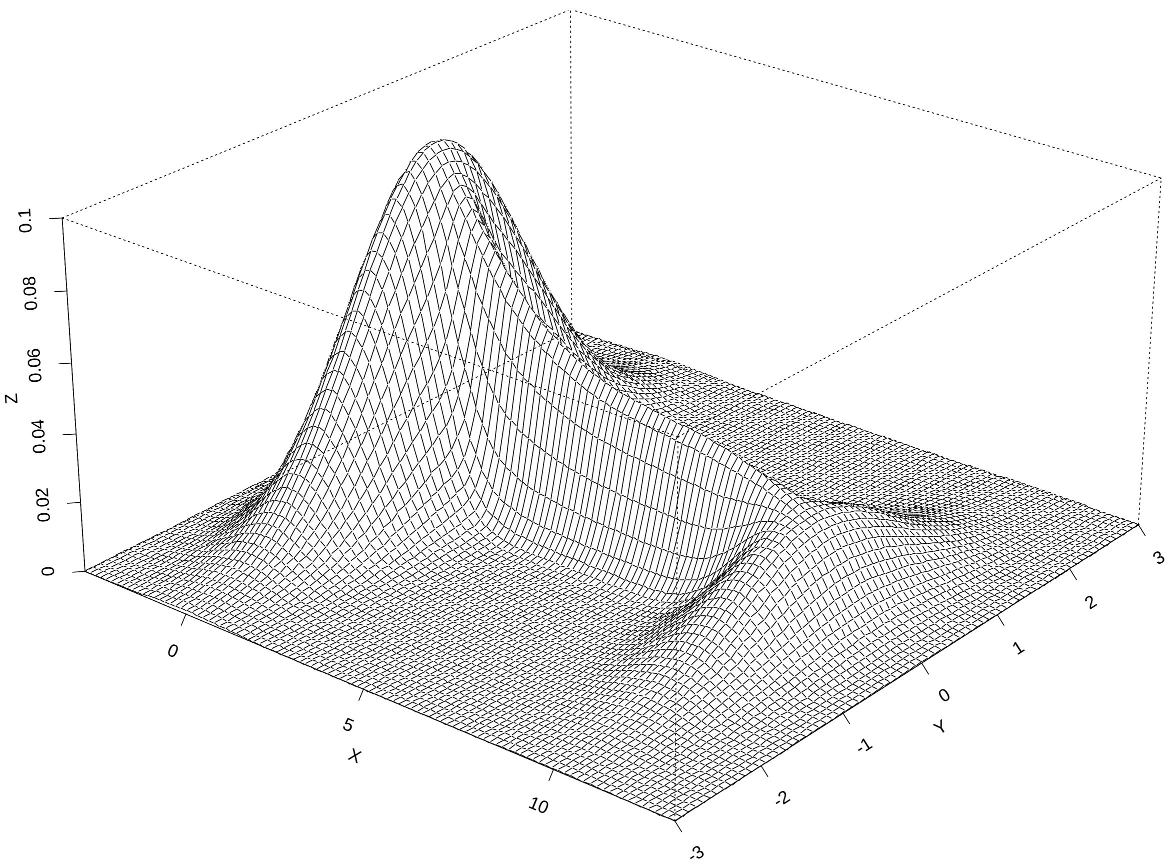

Figure 2: Resulting bivariate mixture density, which is clearly unimodal.

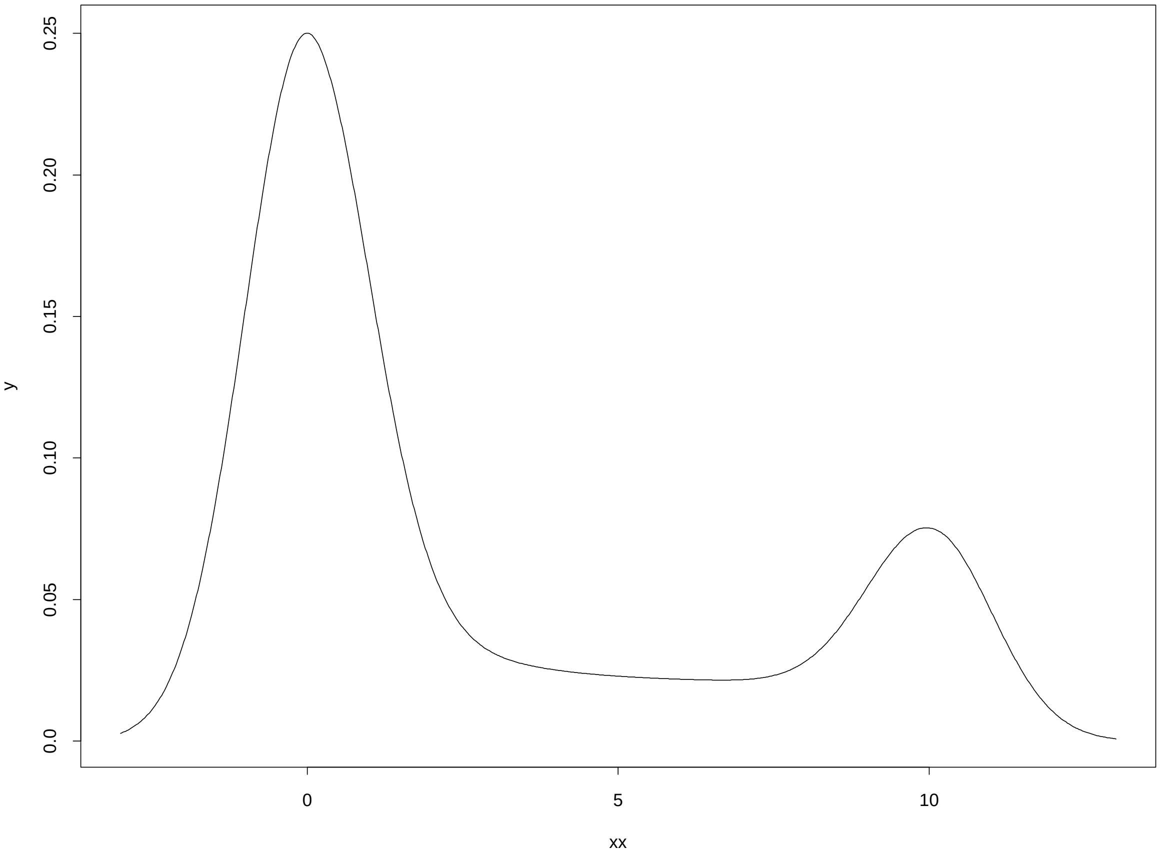

Figure 3: Marginal density along x-axis is bimodal.

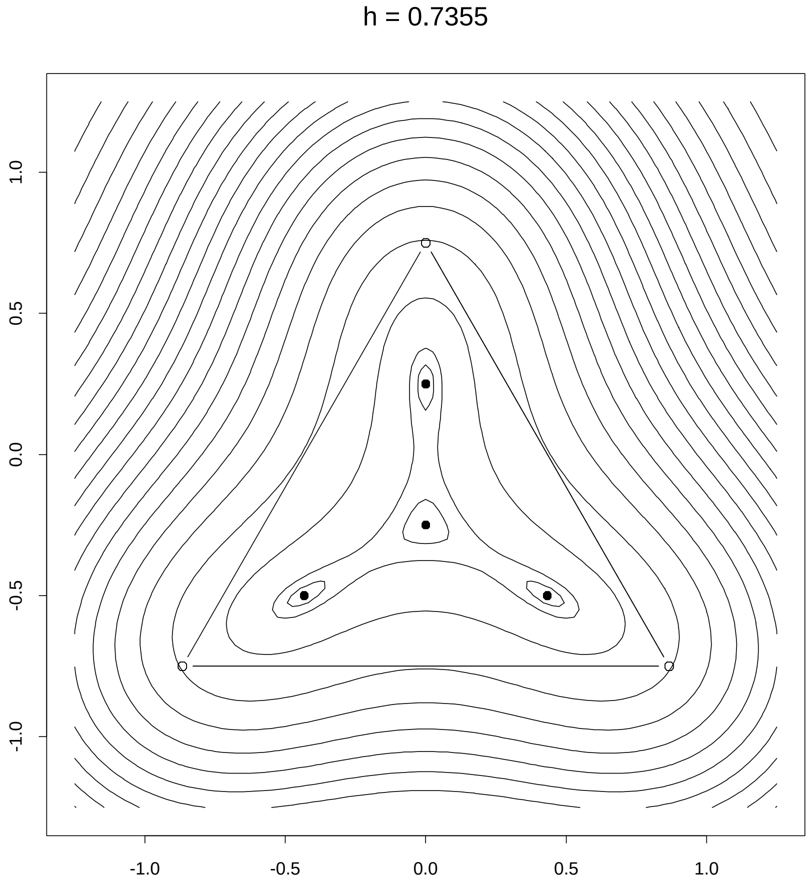

Figure 4: Contour of mixture of 3 bivariate normals giving rise to 4 modes.

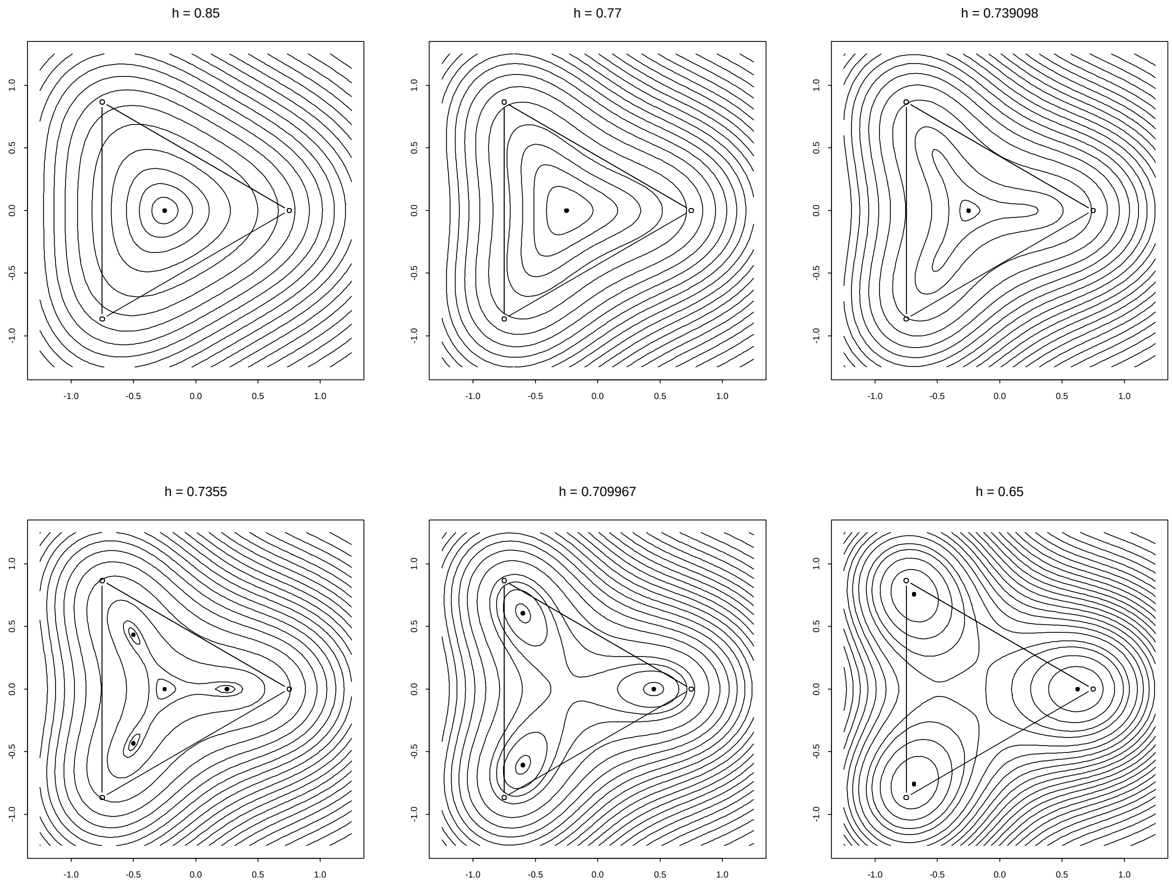

Figure 5: Sequence of bandwidths around the 4-mode setting.

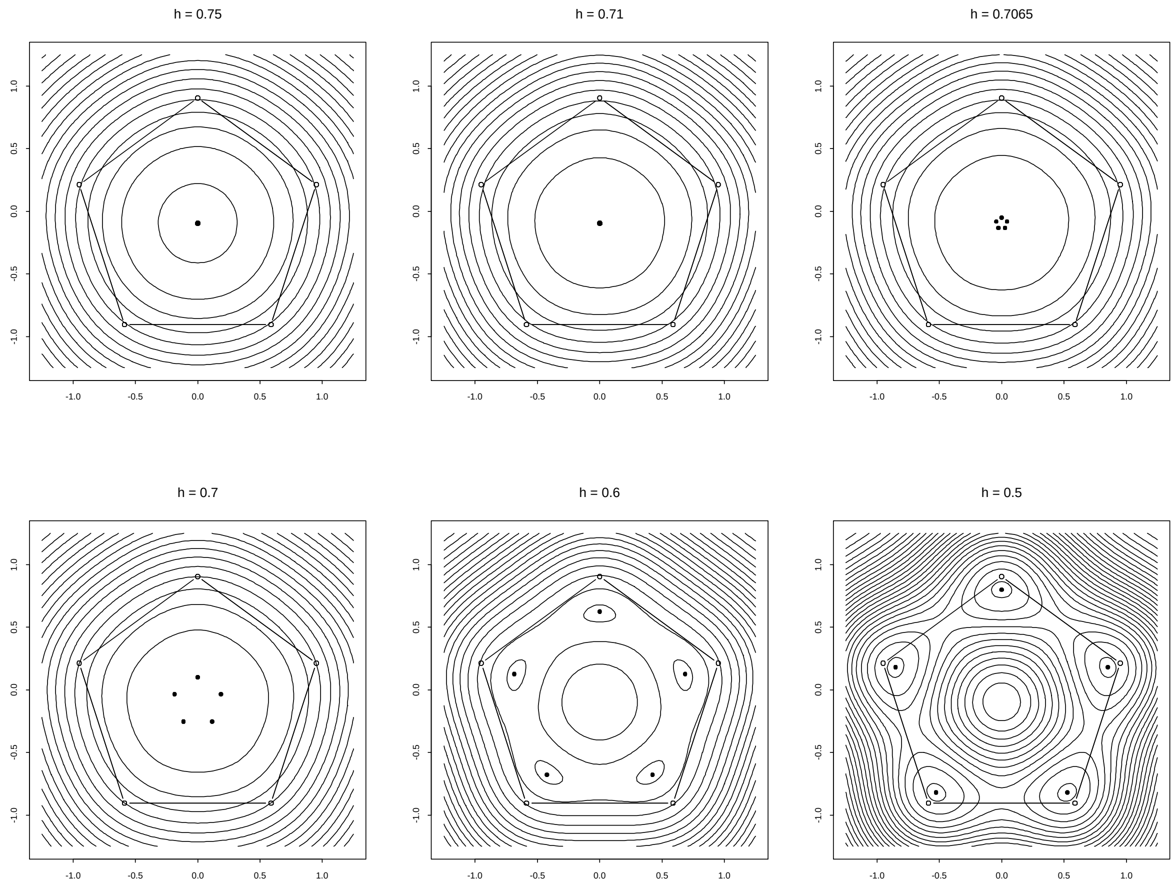

Figure 6: Sequence of bandwidths with mixture of 5 bivariate normals showing monotone and smooth change in the location of modes, with no extra sixth mode.

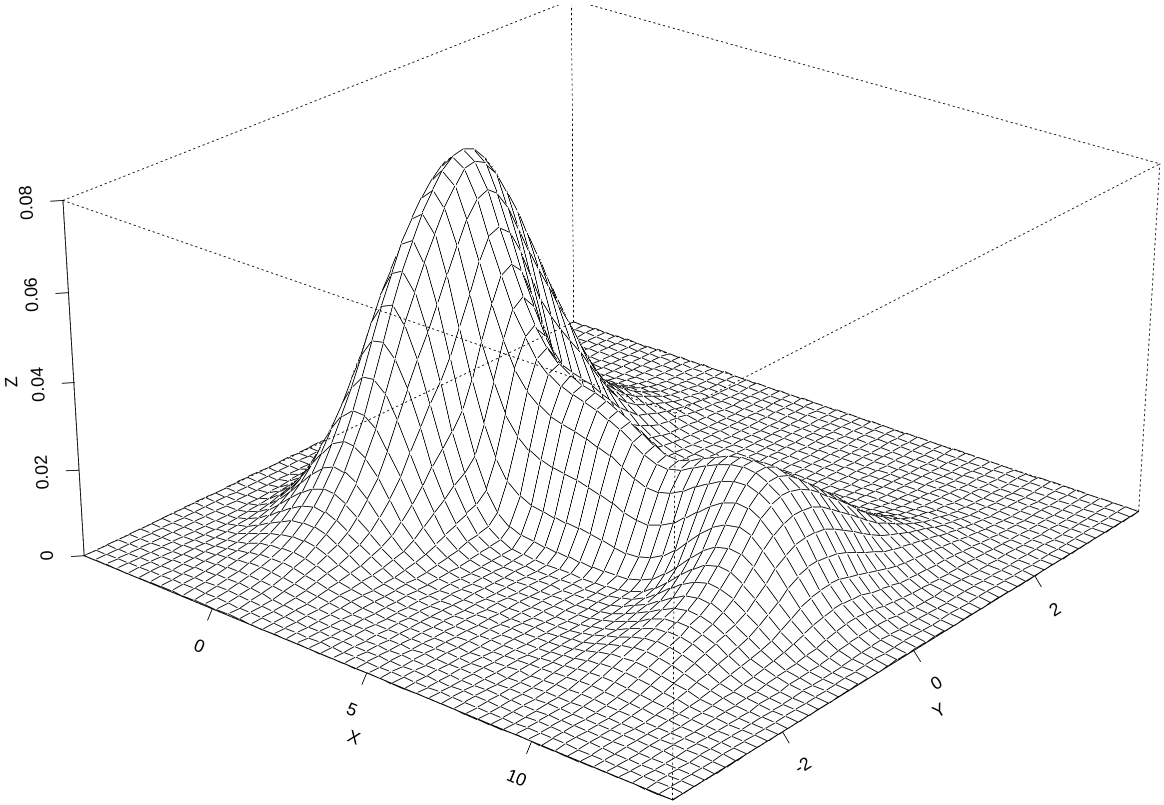

Figure 7: Bivariate ASH estimate of a sample of 5,000 points from the bivariate mixture shown in Figure 2. Notice the smoothing results in a bimodal bivariate density estimate
.

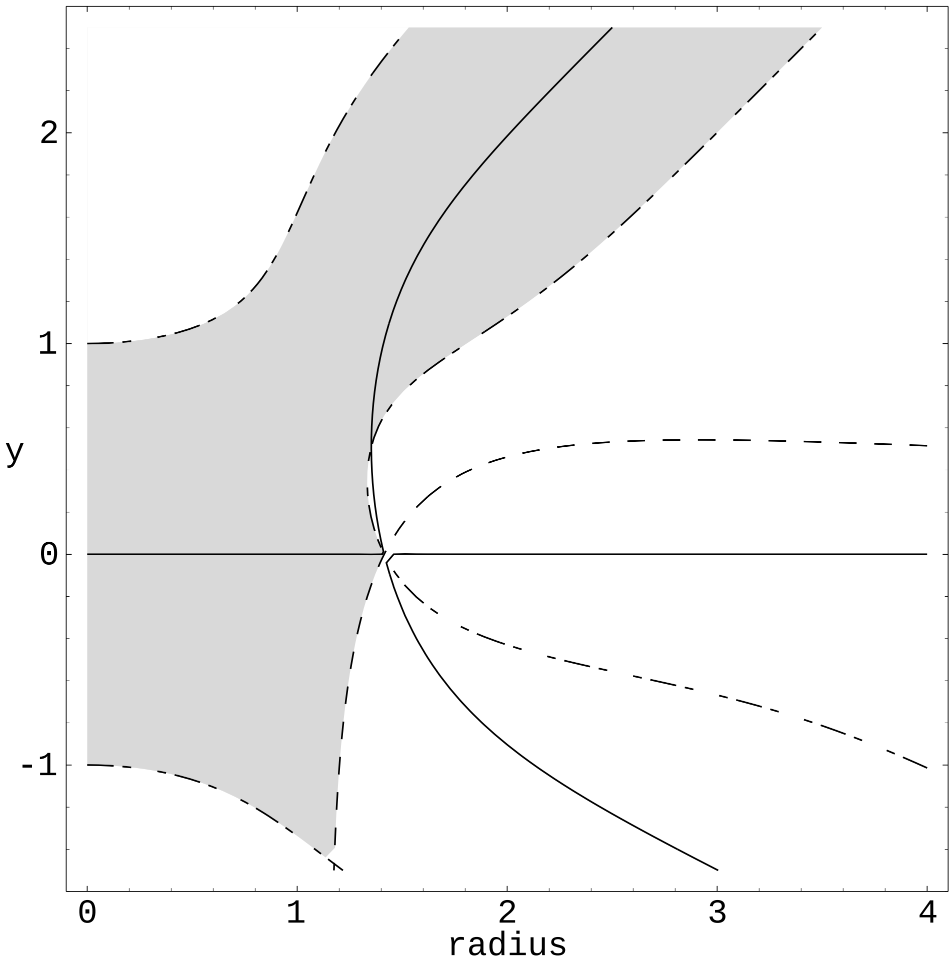

Figure 8: Depiction of region with extra mode.

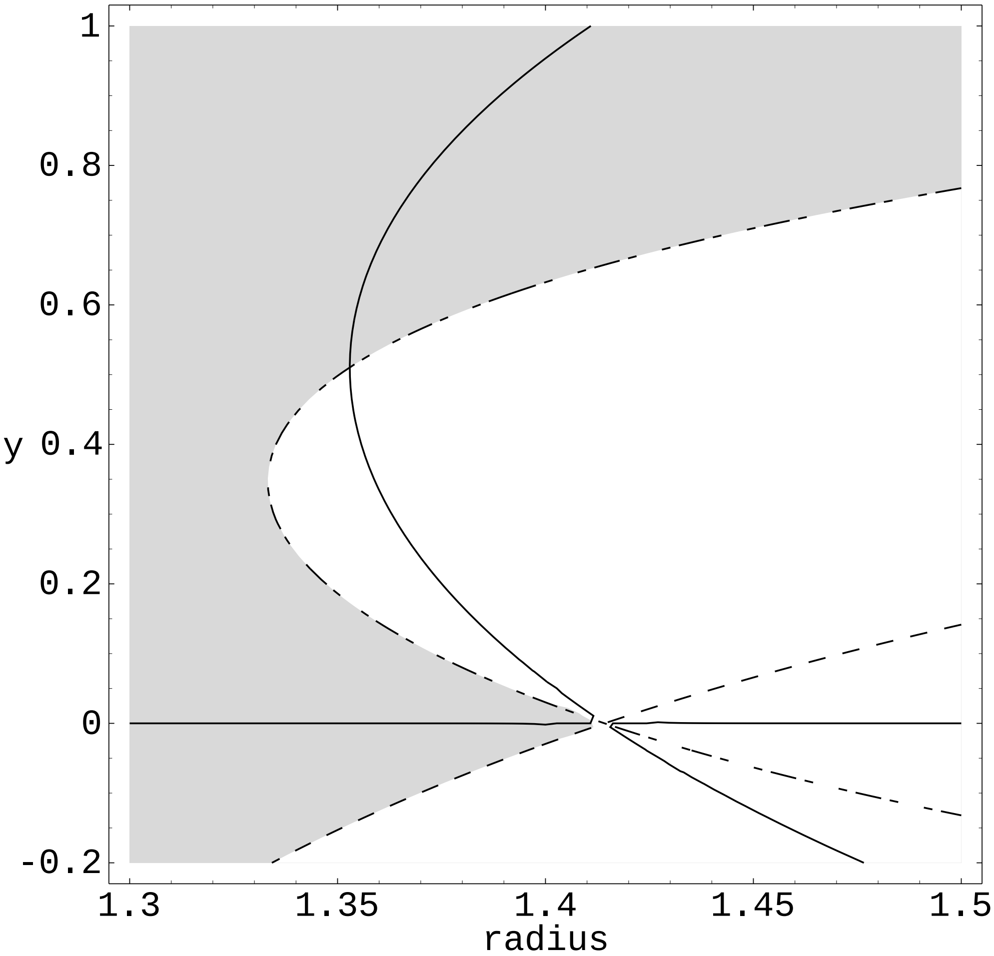

Figure 9: Blowup of region with extra mode.

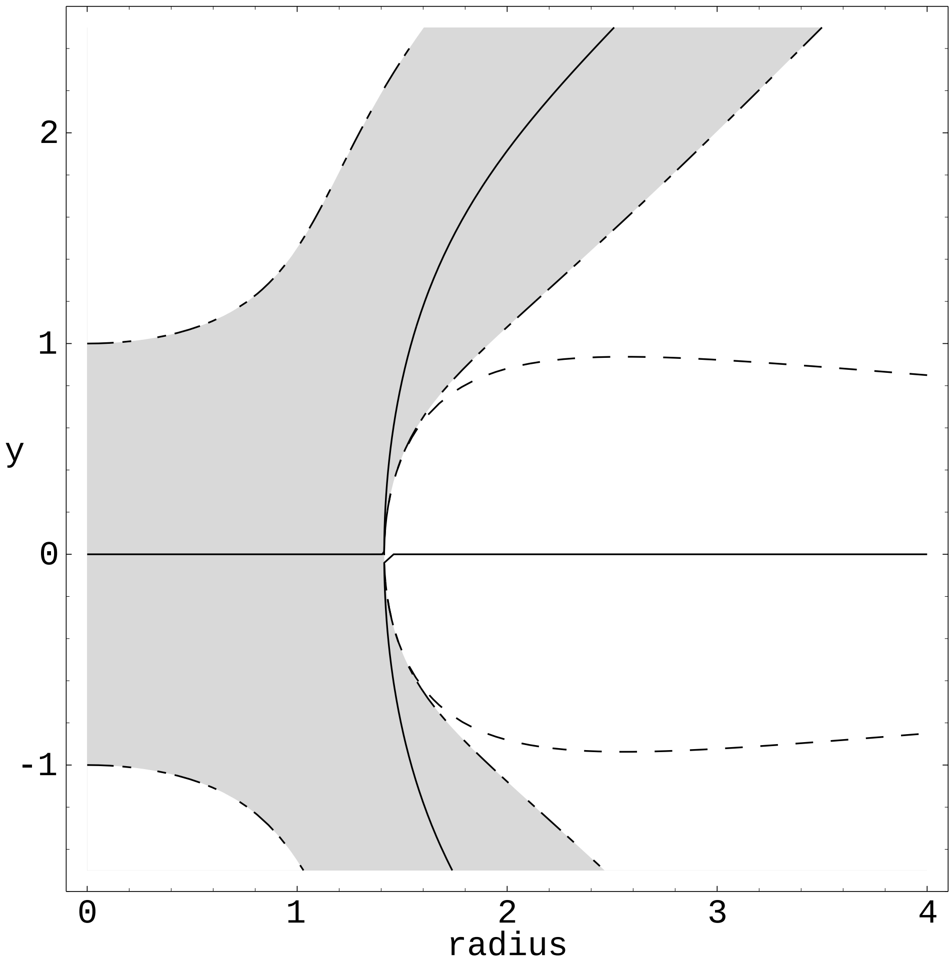

Figure 10: Vertices on a square do not have an extra mode.

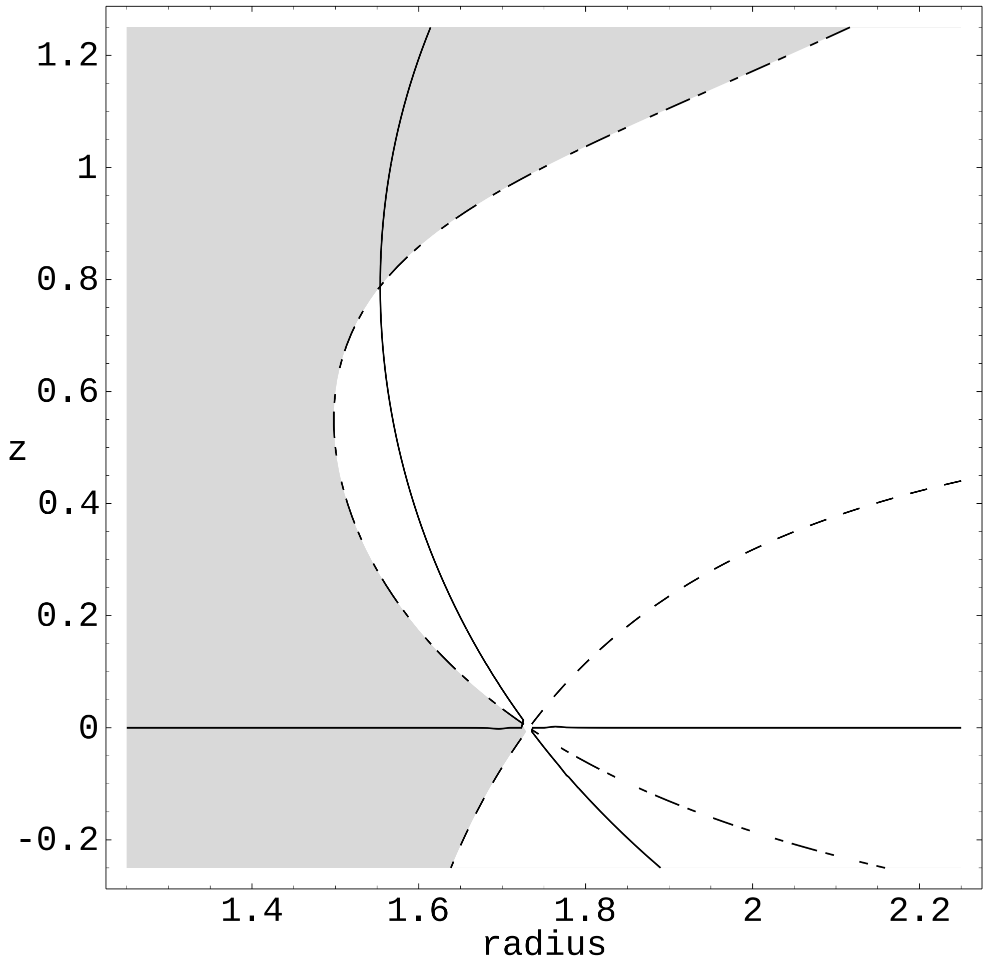

Figure 11: Vertices on a tetrahedron have an extra mode.

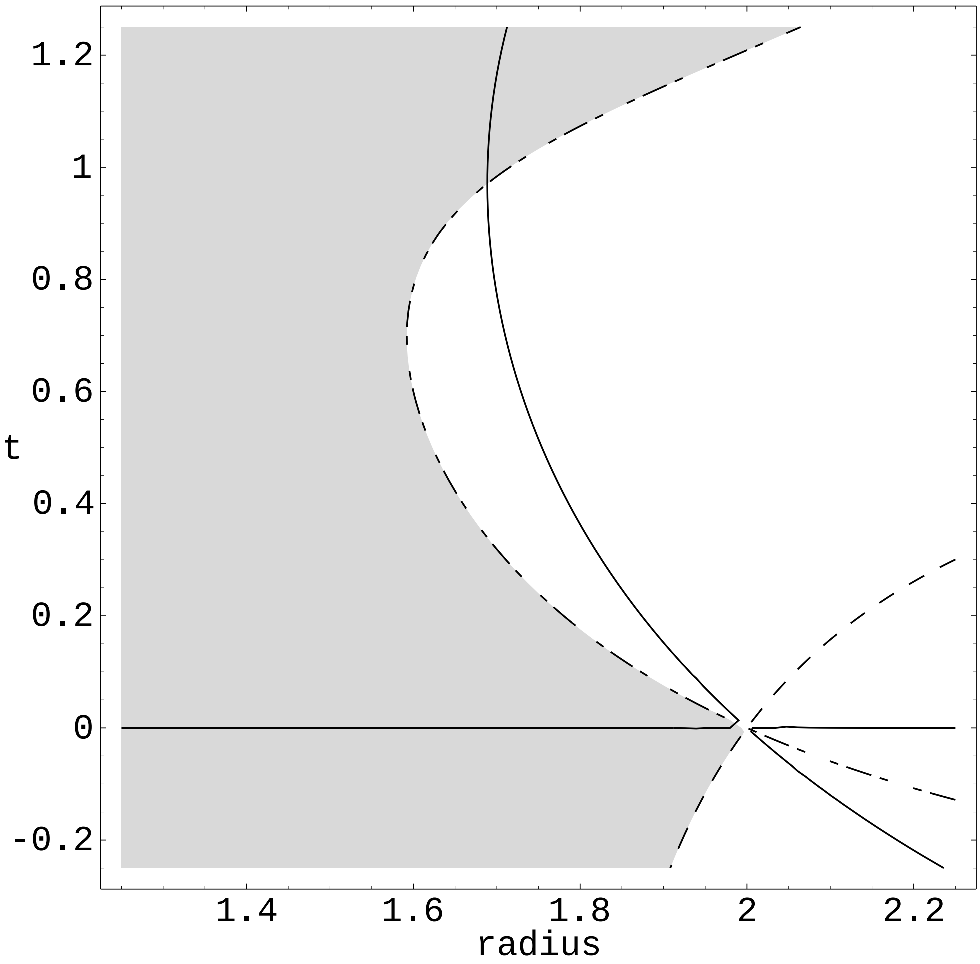

Figure 12: Vertices on 4-D simplex have an extra mode.