Feature Significance for Multivariate Data and Kernel Density Estimation

Arianna Cowling, Tarn Duong, Inge Koch & M. P. Wand
Statistics, School of Mathematics,
University of New South Wales,
Sydney NSW 2052, Australia
email for Inge Koch: inge@unsw.edu.au

Abstract

Multivariate kernel density estimation is a useful technique for extracting valuable information such as the number and location of modes form multivariate data. Many methods (theoretically founded as well as implemented by efficient computer software) exist for one and two dimensional data. For three and higher dimensional data the lack of practical implementation of kernel density estimates and their visualisation has restricted their applicability.

We extend and combine two current techniques to higher dimensional data: multivariate density estimation based on the optimal (full) bandwidth matrix, and Godtliebsen, Marron & Chaudhuri's feature significance (or significance in scale space) which determines statistically significant features such as local extrema. We propose methods for visualising such data, its kernel density estimate and significant features. We apply these methods to data from flow cytometry, a fast growing technique for measuring characteristics of cells.

1 Introduction

Kernel density estimation has become an invaluable tool for extracting information and structure – including visual informations – in data. For one and two-dimensional data good methods are available which have theoretical foundations as well as software implementations, see Bowman and Azzalini (1997), Scott (1992), Simonoff (1996), or Wand and Jones (1995) for examples.

In this paper we want to use kernel density estimation as a means for determining features, in particular modes, in three and higher dimensional data. Non-parametric density estimation provides us with density estimates which we combine with feature significance, a more formal framework for making inferences by evaluating whether these features have statistical significance. It turns out that these features correspond to regions of the sample space where the gradient and/or the curvature are significantly different from zero.

Feature significance techniques have been devised for one-dimensional data by Chaudhuri and Marron (1999) and for two-dimensional data by Godtliebsen *et al* (2002). These papers include techniques for visualising significant gradient and curvature regions which greatly assist in their interpretation. In this paper, we extend the framework established by these authors to multivariate data in two ways. First we provide a more general theoretical treatment of feature significance for multivariate kernel density estimators. Second we provide novel visualisations of three- and higher-dimensional data.

The data which have motivated this research are from flow cytometry, a technique for measuring characteristics of cells suspended in a stream of fluid. In a single flow cytometry experiment, several thousand cells are analysed per second, and various physical and chemical variables, such as light scattering properties and cytoplasmic granularity, are measured and recorded. A beam of light, usually laser light, is directed onto a point through which the stream of fluid containing the cells passes. As each cell passes through the laser beam, some of the light is scattered, and fluorescent chemicals in the cells are excited into emitting light. Detectors pick up the intensity of this scattered

and fluorescent light. For more details on flow cytometry, see Givan (2001).

The data set used here is known as `unst.DRT` and is taken from the `rflowcyt` library (see Rossini *et al* (2005)) in the statistical programming language `R` see R Development Core Team (2005) . It contains measurements of forward scatter intensity (FSC), side scatter intensity (SSC), level of CD3 antigen (CD3) and level of CD4 antigen (CD4) for 194629 HIV– individuals.

The long term goals (see p.225–226 Shapiro (2003)) in the analysis of flow cytometric data are

1. Cell counting;
2. Characterisation of pure cell populations;
3. Identifying and isolating cells in mixed populations; and
4. Characterization of cell subpopulations.

Our more immediate aim is the development of a method for finding and visualising modes in such data. These methods will assist in defining gates in order to isolate cells which are usually of a particular type or satisfy particular properties and will thus provide partial answers to 3 above. For example, we often wish to isolate lymphocytes from a mixed cell population as lymphocytes are of particular interest since they are a "particular type of white blood cell that is involved in many of an organism's immune response" see Givan (2001). So by identifying the feature(s) which corresponds to lymphocytes, we are able form a gate to isolate a lymphocytes sample. By linking gates to multivariate feature significance we introduce more objectivity and more statistical rigour. This is an improvement to the current situation where gate boundaries are subjectively chosen.

Evidence of the connection between significant features and lymphocyte gates is shown in Figure 1. The figure contains a random subsample of 1000 points of the first 3 dimensions i.e. (FSC, SSC, CD3) to avoid over-cluttering in the scatterplot: the red points are inside the gate, whereas the black points are outside.

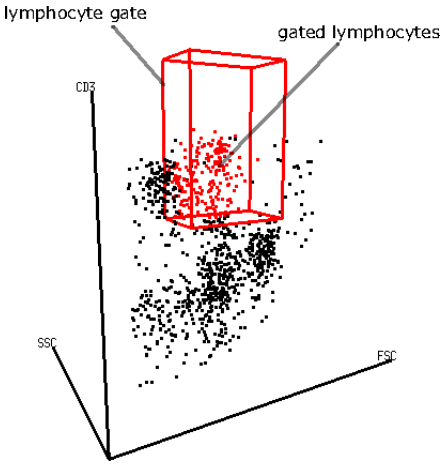

Figure 1: Data and lymphocyte gate

2 Kernel density estimation for 3+ dimensions

In this paper we restrict attention to regions with significant curvature, since these regions characterise the modes of the distribution, and are therefore the most important ones in terms of the aims above. Results similar to those for curvature also hold for regions with significant gradients, and generally these latter calculations are a little simpler. A more detailed exposition of our results, including the analogous gradient results and proofs of the results quoted here can be found in Duong *et al* (2006).

2.1 The kernel curvature estimator

Let the d-variate random sample be $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$, drawn from a common density f. The kernel density estimator is

$$\hat{f}(\boldsymbol{x}; \mathbf{H}) = n^{-1} \sum_{i=1}^{n} K_{\mathbf{H}}(\boldsymbol{x} - \boldsymbol{X}_i) \tag{1}$$

where $K(\boldsymbol{x})$ is the kernel function, $\mathbf{H}$ is the (positive-definite, symmetric) bandwidth matrix, and $K_{\mathbf{H}}$ is the scaled kernel function $K_{\mathbf{H}}(\boldsymbol{x}) = |\mathbf{H}|^{-1/2} K(\mathbf{H}^{-1/2}\boldsymbol{x})$.

The kernel density derivative estimator of $\nabla^{(r)} f$, the rth derivative of f, is

$$\widehat{\nabla^{(r)} f}(\boldsymbol{x}; \mathbf{H}) = n^{-1} \sum_{i=1}^{n} \nabla^{(r)} K_{\mathbf{H}}(\boldsymbol{x} - \boldsymbol{X}_i). \tag{2}$$

The curvature operator leading to the second derivative estimator, $\nabla^{(2)} f$, is

$$\nabla^{(2)} = \begin{bmatrix} \frac{\partial^2}{\partial x_1^2} & \cdots & \frac{\partial^2}{\partial x_1 \partial x_d} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2}{\partial x_1 \partial x_d} & \cdots & \frac{\partial^2}{\partial x_d^2} \end{bmatrix} \tag{3}$$

For $\boldsymbol{s} = (s_1, \ldots, s_d)$, we write

$$f^{(\boldsymbol{s})} = \frac{\partial^{s_1 + \cdots + s_d} f}{\partial x_1^{s_1} \ldots \partial x_d^{s_d}}.$$

So the (j,k)th element of $\nabla^{(2)} f$

$$[\nabla^{(2)} f]_{jk} = \frac{\partial^2 f}{\partial x_j \partial x_k} = \frac{\partial^2 f}{\partial x_k \partial x_j} = f^{(\boldsymbol{e}_j + \boldsymbol{e}_k)}.$$

Here we assume that f is sufficiently regular so that the order of differentiation can be interchanged.

The kernel curvature estimator $\widehat{\nabla^{(2)} f}$ is

$$\widehat{\nabla^{(2)} f}(\boldsymbol{x}; \mathbf{H}) = n^{-1} \sum_{i=1}^{n} \nabla^{(2)} K_{\mathbf{H}}(\boldsymbol{x} - \boldsymbol{X}_i)$$

where

$$\nabla^{(2)} K_{\mathbf{H}}(\boldsymbol{x}) = |\mathbf{H}|^{-1/2} \mathbf{H}^{-1/2} \nabla^{(2)} K(\mathbf{H}^{-1/2}\boldsymbol{x}) \mathbf{H}^{-1/2}.$$

In vectorised form, this is

$$\operatorname{vech} \widehat{\nabla^{(2)} f}(\boldsymbol{x}; \mathbf{H}) = n^{-1} \sum_{i=1}^{n} \operatorname{vech} \nabla^{(2)} K_{\mathbf{H}}(\boldsymbol{x} - \boldsymbol{X}_i) \tag{4}$$

where vech is the vector half operator which takes the elements of the lower triangular half of a symmetric matrix and stacks them into a vector e.g.

$$\operatorname{vech} \begin{bmatrix} a & b & c \\ b & d & e \\ c & e & f \end{bmatrix} = \begin{bmatrix} a \\ b \\ c \\ d \\ e \\ f \end{bmatrix}.$$

2.2 Bandwidth selection

In this section we present a method for estimating the full bandwidth matrix $\mathbf{H}$. In this we follow Duong (2004). Most of the material in this section applies to arbitrary dimension d, however the computatinal complexity increases rapidly as the dimension increases. For the acutal computations and the visualisation part, we are primarily interested in the case $d = 3$.

We measure the performance of $\widehat{f}$ with the mean integrated squared error (or MISE) criterion,

$$\text{MISE}\, \widehat{f}(\,\cdot\,;\mathbf{H}) = \mathbb{E} \int_{\mathbb{R}^d} [\widehat{f}(\boldsymbol{x};\mathbf{H}) - f(\boldsymbol{x})]^2\, d\boldsymbol{x}. \tag{5}$$

As MISE does not have a closed form, unless f is a normal mixture and K is a normal kernel we use AMISE, the asymptotic mean integrated squared error instead. For details on conditions and a proof of the AMISE result in the following proposition, see p94ff of Wand and Jones (1995).

Proposition 1 *Under suitable regularity conditions on the Hessian matrix $\nabla^{(2)} f$ of f, the bandwidth matrix $\mathbf{H}$ and moments of the kernel K,*

$$\begin{aligned} \mathit{AMISE}\, \widehat{f}(\,\cdot\,;\mathbf{H}) &= n^{-1}|\mathbf{H}|^{-1/2} R(K) + \frac{1}{4}\mu_2(K)^2 \int tr^2(\mathbf{H}\nabla^{(2)} f(\boldsymbol{x})\, df(\boldsymbol{x}) \\ &= n^{-1}|\mathbf{H}|^{-1/2} R(K) + \frac{1}{4}\mu_2(K)^2 (\text{vech}^T\, \mathbf{H}) \boldsymbol{\Psi}_4 (\text{vech}\, \mathbf{H}) \end{aligned} \tag{6}$$

where $R(K) = \int K(\boldsymbol{x})^2\, d\boldsymbol{x} < \infty$, and $\mu_2(K)$, where $\mu_2(K) < \infty$, Here $\boldsymbol{\Psi}_4$ is the $\frac{1}{2}d(d+1) \times \frac{1}{2}d(d+1)$ matrix given by

$$\boldsymbol{\Psi}_4 = \int \text{vech}[2\nabla^{(2)} f(\boldsymbol{x}) - \text{diag}\nabla^{(2)} f(\boldsymbol{x})]\, \text{vech}^T [2\nabla^{(2)} f(\boldsymbol{x}) - \text{diag}\nabla^{(2)} f(\boldsymbol{x})]\, d\boldsymbol{x}. \tag{7}$$

For details on the precise regularity conditions, and a proof of the result, see p 94ff of Wand and Jones (1995).

To determine the bandwidth matrix $\mathbf{H}$ which minimises AMISE, a knowledge of $\boldsymbol{\Psi}_4$ is required. As this matrix is unknown, we follow Duong (2004) and replace the unknown $\boldsymbol{\Psi}_4$ by a *plug-in* criterion and then find the minimiser of

$$\text{PI}(\mathbf{H}) = n^{-1} R(K) |\mathbf{H}|^{-1/2} + \frac{1}{4}\mu_2(K)^2 (\text{vech})^T \mathbf{H}) \widehat{\boldsymbol{\Psi}}_4 (\text{vech}\, \mathbf{H}). \tag{8}$$

Note that PI($\mathbf{H}$) is just AMISE with $\boldsymbol{\Psi}_4$ replaced by an estimator $\widehat{\boldsymbol{\Psi}}_4$.

The advantage in replacing AMISE by PI is that the elements of the matrix $\widehat{\boldsymbol{\Psi}}_4$, being estimates of elements of $\boldsymbol{\Psi}_4$, can be calculated from the data. The elements in $\boldsymbol{\Psi}_4$ are of the form

$$\psi_{\mathbf{r}} = \int_{\mathbb{R}^d} f^{(\mathbf{r})}(\mathbf{x}) f(\mathbf{x})\, d\mathbf{x} = \mathbb{E} f^{(\mathbf{r})}(\mathbf{X})$$

and so a natural estimator of $\psi_{\mathbf{r}}$ is the sample mean of the $\widehat{f}^{(\mathbf{r})}(\mathbf{X}_i)$, known as the leave-in-diagonals estimator. That is,

$$\widehat{\psi}_{\mathbf{r}}(\mathbf{G}) = n^{-1} \sum_{i=1}^{n} \widehat{f}^{(\mathbf{r})}(\mathbf{X}_i;\mathbf{G}) = n^{-2} \sum_{i=1}^{n} \sum_{j=1}^{n} K_{\mathbf{G}}^{(\mathbf{r})}(\mathbf{X}_i - \mathbf{X}_j). \tag{9}$$

for a bandwidth matrix $\mathbf{G}$.

For computational ease, take $\mathbf{G} = g^2\mathbf{I}$ and $|\mathbf{r}| = j$ and obtain

Proposition 2 *Let $\mathbf{G} = g^2\mathbf{I}$, then*

$$\begin{aligned} \mathit{AMSE}\, \widehat{\psi}_{\boldsymbol{r}}(g) &= \widehat{\psi}_{\boldsymbol{r}}(g) = 2n^{-2} g^{-d-2j} \psi_{\boldsymbol{0}} R(K^{(\boldsymbol{r})}) \\ &\quad + \left[n^{-1} g^{-d-j} K^{(\boldsymbol{r})}(\boldsymbol{0}) + \frac{1}{2} g^2 \mu_2(K) \sum_{i=1}^{d} \psi_{\boldsymbol{r}+2\boldsymbol{e}_i} \right]^2. \end{aligned} \tag{10}$$

Furthermore, approximating AMSE by

$$SAMSE_j(g) = \sum_{\boldsymbol{r}:|\boldsymbol{r}|=j} AMSE\, \widehat{\psi}_{\boldsymbol{r}}(g)$$

yields the minimiser

$$g_{j,SAMSE} = \left[\frac{A_1(4j+4d)}{n(-(j+d-2)A_2 + \sqrt{(j+d-2)^2 A_2^2 + (8j+8d)A_1 A_3})} \right]^{1/(j+d+2)} . \qquad (11)$$

where

$$\begin{aligned} A_0 &= \psi_{\boldsymbol{0}} \sum_{\boldsymbol{r}:|\boldsymbol{r}|=j} R(K^{(\boldsymbol{r})}) \\ A_1 &= \sum_{\boldsymbol{r}:|\boldsymbol{r}|=j} K^{(\boldsymbol{r})}(\boldsymbol{0})^2 \\ A_2 &= \mu_2(K) \sum_{\boldsymbol{r}:|\boldsymbol{r}|=j} K^{(\boldsymbol{r})}(\boldsymbol{0}) \left(\sum_{i=1}^{d} \psi_{\boldsymbol{r}+2\boldsymbol{e}_i} \right) \\ A_3 &= \mu_2(K)^2 \sum_{\boldsymbol{r}:|\boldsymbol{r}|=j} \left(\sum_{i=1}^{d} \psi_{\boldsymbol{r}+2\boldsymbol{e}_i} \right)^2 \end{aligned}$$

More details of this proposition can be found in Cowling (2005).

As in Duong (2004) we now reverse the whole process:

1. For $j = 4$ use the *plug-in* estimate $g_{4,\mathrm{SAMSE}}$ to produce the kernel estimate $\widehat{\boldsymbol{\Psi}}_4$.
2. Use $\widehat{\boldsymbol{\Psi}}_4$ to calculate the PI($\mathbf{H}$). in eq (8)
3. Numerically minimise PI($\mathbf{H}$) to obtain required plug-in bandwidth $\widehat{\mathbf{H}}_{\mathrm{PI,SAMSE}}$.

Remark. If the dispersion of the data varies significantly between the different coordinate directions, a pilot bandwidths of the form $\mathbf{G} = g^2\mathbf{I}$ may not be appropriate directly. Instead it is advisable to scale or sphere the data prior to bandwidth matrix for density estimation. this reason we should transform the data before any pilot bandwidth selection is carried out. There are two common transformations which can be used.

2.3 Properties of the second derivative estimatior

We now return to the expression of the kernel curvature estimator in vectorised form as in eq (4)

$$\operatorname{vech} \widehat{\nabla^{(2)} f}(\boldsymbol{x}; \mathbf{H}) = n^{-1} \sum_{i=1}^{n} \operatorname{vech} \nabla^{(2)} K_{\mathbf{H}}(\boldsymbol{x} - \boldsymbol{X}_i)$$

and exhibit its distributional properties. These are crucial for the development of the significance testing in the next section.

Theorem 3 *Assume the following conditions hold.*

(B1) All entries of $\mathbf{H} \to 0$ *and* $n^{-1}|\mathbf{H}|^{-1/2}\mathbf{H}^{-2} \to 0$, *as* $n \to \infty$.

(B2) All entries of $\nabla^{(4)} f(\boldsymbol{x})$ *are bounded, continuous and square integrable.*

(B3) The kernel K *is a symmetric probability density function and*

$$\int_{\mathbb{R}^d} K(\boldsymbol{x})\boldsymbol{x}\boldsymbol{x}^T \, d\boldsymbol{x} = R(K)\mathbf{I} \text{ with } R(K) < \infty.$$

(B4) All entries of $\mathbf{R}(\operatorname{vech} \nabla^{(2)} K) < \infty$.

Then the approximate asymptotic distribution of the kernel curvature estimator $\widehat{\nabla^{(2)} f}$ *is*

$$\operatorname{vech} \widehat{\nabla^{(2)} f}(\boldsymbol{x}; \mathbf{H}) \overset{\text{approx}}{\sim} N\left(\operatorname{vech} \nabla^{(2)} f(\boldsymbol{x}), \boldsymbol{\Sigma}^{(2)}(\boldsymbol{x})\right)$$

where $\boldsymbol{\Sigma}^{(2)}(\boldsymbol{x}) = n^{-1}|\mathbf{H}|^{-1/2}\mathbf{R}(\operatorname{vech}(\mathbf{H}^{-1/2}\nabla^{(2)} K \mathbf{H}^{-1/2}))f(\boldsymbol{x})$.

For a proof of this result see Duong *et al* (2006).

A version of this result for normal kernels and diagonal bandwidth matrices is difficult to state concisely for general dimension d since $\mathbf{R}(\operatorname{vech}(\mathbf{H}^{-1/2}\nabla^{(2)} K \mathbf{H}^{-1/2}))$ does not have simple concise algebraic form. Here we give the special case for $d = 3$.

Corollary 4 *Assume the conditions (B1) – (B2) from Theorem 3 hold. Further assume that* K *is the normal kernel and the bandwidth matrix is parameterised* $\mathbf{H} = \operatorname{diag}(h_1^2, \ldots, h_d^2)$*, then*

$$\operatorname{vech} \widehat{\nabla^{(2)} f}(\boldsymbol{x}; h_1, \ldots, h_d) \overset{\text{approx}}{\sim} N\left(\operatorname{vech} \nabla^{(2)} f(\boldsymbol{x}), \boldsymbol{\Sigma}^{(2)}(\boldsymbol{x})\right)$$

and for $d = 3$,

$$\boldsymbol{\Sigma}^{(2)}(\boldsymbol{x}) = (32\pi^{3/2})^{-1} n^{-1} (h_1 h_2 h_3)^{-1} \times \begin{bmatrix} 3h_1^{-4} & 0 & 0 & h_1^{-2}h_2^{-2} & 0 & h_1^{-2}h_3^{-2} \\ 0 & h_1^{-2}h_2^{-2} & 0 & 0 & 0 & 0 \\ 0 & 0 & h_1^{-2}h_3^{-2} & 0 & 0 & 0 \\ h_1^{-2}h_2^{-2} & 0 & 0 & 3h_2^{-4} & 0 & h_2^{-2}h_3^{-2} \\ 0 & 0 & 0 & 0 & h_2^{-2}h_3^{-2} & 0 \\ h_1^{-2}h_3^{-2} & 0 & 0 & h_2^{-2}h_3^{-2} & 0 & 3h_3^{-4} \end{bmatrix} f(\boldsymbol{x});$$

3 Testing for significant modal regions

For the two-dimensionsional feature significane testing Godtliebsen *et al* (2002) used the test statistic $\max\{|\lambda_1(\boldsymbol{x})|, \ldots, |\lambda_d(\boldsymbol{x})|\}$ where $\lambda_j(\boldsymbol{x})$ is the jth eigenvalue of the normalised $\widehat{\nabla^{(2)} f}(\boldsymbol{x}; \mathbf{H})$, which, in addition to modes, enables the detection of valleys, ridges, saddle points etc.

For three and higher dimensional data, it is not clear what these features correspond to, or whether they are even important. For this reason, we choose a different approach.

Our aim is to test simultaneously at all test points $\boldsymbol{x}$ whether $\widehat{\nabla^{(2)} f}(\boldsymbol{x}; \mathbf{H})$ indicates a significant difference from zero. These tests are highly correlated since for nearby points $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, $\widehat{\nabla^{(2)} f}(\boldsymbol{x}_1; \mathbf{H})$ and $\widehat{\nabla^{(2)} f}(\boldsymbol{x}_2; \mathbf{H})$ are highly correlated. The approach of Godtliebse *et al* (2002) is to reduce this series of dependent tests to an equivalent series of independent ones, and then to use a classical Bonferroni-type simultaneous test. Our approach is different in that we use a multiple testing procedure which is suitable for a series of dependent tests. There are many such testing procedures. The one we use is due to Hochberg (1988), as it is easy to implement and to interpret. This procedure, a modification of the classical Bonferroni one, is described as follows.

Let the nominal level of significance be α. Let the p-values for each of the m individual tests be ordered in ascending order $P_{(1)}, \ldots, P_{(m)}$, corresponding to the null hypotheses $H_{0,(1)}, \ldots, H_{0,(m)}$. If $P_{(j)} \leq \alpha/(m - j + 1)$ then we reject all null hypotheses $H_{0,(1)}, \ldots, H_{0,(j)}$. Usually we find the maximum where this occurs i.e.

$$\text{If } j_{\max} = \operatorname*{argmax}_{1 \leq j \leq m} P_{(j)} \leq \alpha/(m - j + 1) \text{ then reject } H_{0,(1)}, \ldots, H_{0,(j_{\max})}. \tag{12}$$

See Hochberg (1988) for a proof that the overall level of significance is α.

Our null hypothesis is

$$H_0 : \|\operatorname{vech} \nabla^{(2)} f(\boldsymbol{x})\| = 0.$$

From Result 3, the approximate asymptotic null distribution of vech $\widehat{\nabla^{(2)} f}(\boldsymbol{x}; \mathbf{H})$ after normalising is

$$\boldsymbol{\Sigma}^{(2)}(\boldsymbol{x})^{-1/2} \operatorname{vech} \widehat{\nabla^{(2)} f}(\boldsymbol{x}; \mathbf{H}) \overset{\text{approx}}{\sim} N(\mathbf{0}, \mathbf{I}_{d^*})$$

where $d^* = \frac{1}{2}d(d+1)$ is the length of a vech'ed $d \times d$ matrix. An appropriate curvature test statistic is therefore

$$W^{(2)}(\boldsymbol{x}) = \|\boldsymbol{\Sigma}^{(2)}(\boldsymbol{x})^{-1/2} \operatorname{vech} \widehat{\nabla^{(2)} f}(\boldsymbol{x}; \mathbf{H})\|^2 \overset{\text{approx}}{\sim} \chi^2_{d^*}. \tag{13}$$

Like Godtliebsen *et al* (2002), we restrict significance testing to points where there is sufficient number of data points.

An advantage of our approach is that we circumvent the need to simulate critical points of the null distribution, as is the case for Godtliebsen *et al*'s test statistic, since $W^{(2)}(\boldsymbol{x})$ has an approximate closed form null distribution, given by the estimate

$$\widehat{W^{(2)}}(\boldsymbol{x}) = \|\widehat{\boldsymbol{\Sigma}^{(2)}}(\boldsymbol{x})^{-1/2} \operatorname{vech} \widehat{\nabla^{(2)} f}(\boldsymbol{x}; \mathbf{H})\|^2$$

where

$$\widehat{\boldsymbol{\Sigma}^{(2)}}(\boldsymbol{x}) = n^{-1}|\mathbf{H}|^{-1/2}\mathbf{R}(\operatorname{vech} \mathbf{H}^{-1/2}\nabla^{(2)} K \mathbf{H}^{-1/2})\hat{f}(\boldsymbol{x}; \mathbf{H}),$$

and our p-value at $\boldsymbol{x}$ is $\mathbb{P}(X^2 > \widehat{W^{(2)}}(\boldsymbol{x}))$ where $X^2 \sim \chi^2_{d^*}$.

Thus, points $\boldsymbol{x}$ for which $\widehat{\nabla^{(2)} f}(\boldsymbol{x}; \mathbf{H}) < 0$ belong to a significant modal region.

This estimator of $\boldsymbol{\Sigma}^{(2)}(\boldsymbol{x})$ is from Theorem 3, and is an alternative to the estimator from Godtliebsen *et al* (2002). The latter estimator relies on individually estimating the elements of $\boldsymbol{\Sigma}^{(2)}(\boldsymbol{x})$ by the individual sample variances and covariances. However, individually estimating the elements of a matrix quantity is not always optimal e.g. it cannot guarantee positive definiteness of the matrix of individual estimates. On the other hand, our method, which estimates the $\boldsymbol{\Sigma}^{(2)}(\boldsymbol{x})$ matrix in a matrix-wise procedure, *can* guarantee this positive definiteness.

Another advantage of our approach is that it is more computationally efficient since it no longer requires the computationally intensive step of calculating the (many) sample variances and covariances of the elements of vech $\widehat{\nabla^{(2)} f}(\boldsymbol{x}; \mathbf{H})$. Godtliebsen *et al*'s approach furthermore relies on using the most restricted parametrisation $\mathbf{H} = h^2\mathbf{I}$ whereas our method is for general bandwidth matrices.

4 Results

The method described in the previous sections is applied to the flow cytometric data set partially displayed in Figure 1. As the number of data records is very large (126,675 records), we first bin the data using linear binning as described in Cowling (2005). Binning greatly increases computational efficiency of the subsequent density estimation. The computations were carried out in Matlab and `R`.

We demonstrate our method for a series of bandwidths. When considering three dimensional data, and calculating density estimates for such data, it is not possible to display all information in a single picture. Here we use isosurfaces, that is, surfaces for which the density estimate has a fixed value. This corresponds to and extends contour plots of fixed contour levels in two dimensions. A sequence of such isolevels will show the change of the modal regions with the bandwidths. Of course it is also possible to overlay or embed isosurfaces of different levels *eg* by the use of different colours, or show them in continuous time.

For Figure 2 the data was first scaled and binned. A normal kernel with a diagonal bandwidth matrix was used. For simplicity the same bandwidth was used in all three directions. The 6 subplots in Figure 2 show the results of varying the bandwidths from $h = 0.0145$ to $h = 0.1661$. As the bandwidth increases – going from left to right and then along the bottom row – the modal regions become larger and new modes appear. The location of the main three modes seem to remain the same, but the extra modal regions which apppeared at bandwidth $h = 0.0429$ have disappeared again for bandwidth $h = 0.1661$. These plots illustrate the effectiveness of feature significance for three dimensional data.

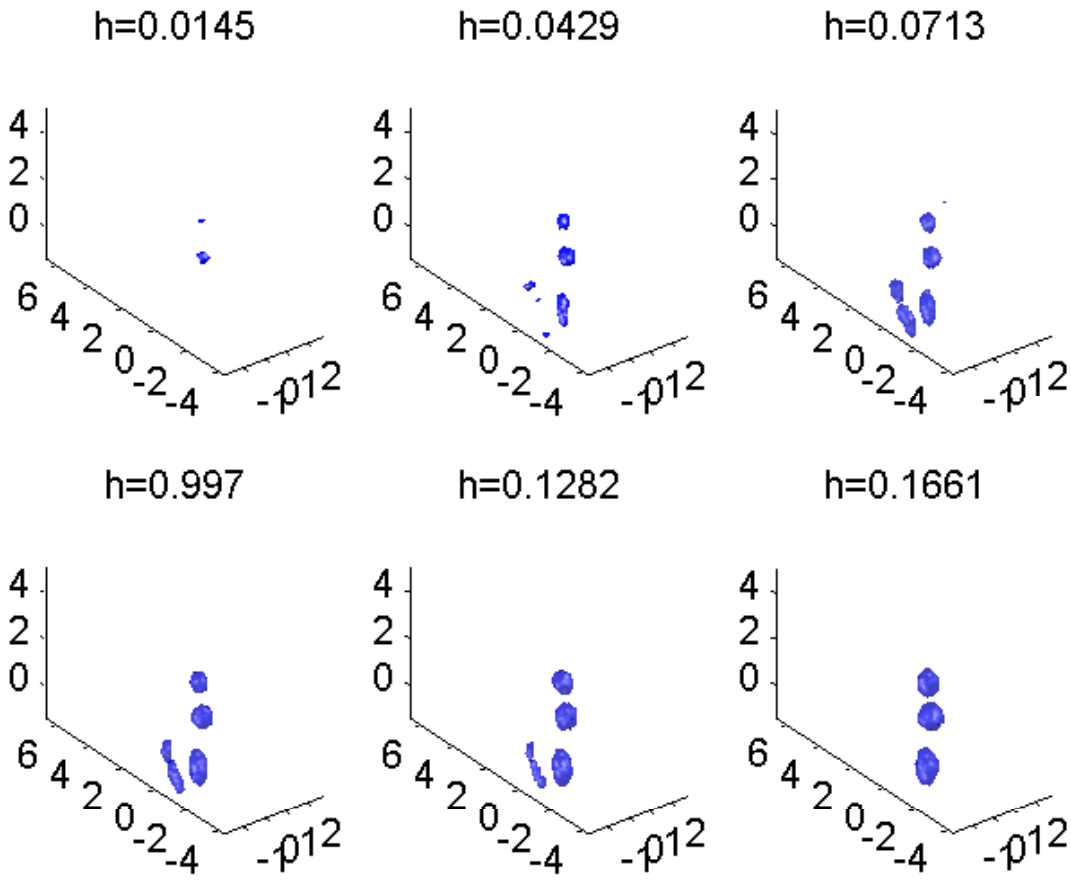

Figure 2: Modal regions for increasing bandwidths

References

[1] Bowman, A. W. and Azzalini, A. (1997). *Applied Smoothing Techniques for Data Analysis*, Oxford University Press, Oxford.

[2] Chaudhuri, P. and Marron, J. S. (1999) SiZer for exploration of structures in curves, *Journal of the American Statistical Association* **94**, 807–823.

[3] Cowling, A. (2005). *Feature Significance for Flow Cytometry Data*, Honours Thesis, University of New South Wales.

[4] Duong, T. (2004). *Bandwidth matrices for mulitivariate kernel density estimation*, PhD Thesis.

[5] Duong, T., Cowling, A., Koch, I. and Wand, M.P. (2006). Feature Significance for Multivariate Kernel Density Estimation. *Preprint, UNSW.*

[6] Givan, A. L. (2001). *Flow Cytometry: First Principles*, 2nd edn, Wiley-Liss, New York.

[7] Godtliebsen, F., Marron, J. S. and Chaudhuri, P. (2002). Significance in scale space for bivariate density estimation, *Journal of Computational and Graphical Statistics* **11**, 1–21.

[8] Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance, *Biometrika* **75**, 800–802.

[9] R Development Core Team (2005). *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria.

[10] Rossini, A., Wan, J. and Moodie, Z. (2005). *rflowcyt: Statistical tools and data structures for analytic flow cytometry.* R package version 1.0.1.

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

[12] Shapiro, H. M. (2003) *Practical Flow Cytometry: Fourth Edition.*

[13] Simonoff, J. S. (1996). *Smoothing Methods in Statistics*, Springer-Verlag, New York.

[14] Wand, M. P. and Jones, M. C. (1995). *Kernel Smoothing*, Chapman and Hall Ltd., London.