Available online at www.sciencedirect.com
ScienceDirect
Computational Statistics and Data Analysis 52 (2008) 4225–4242
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

Feature significance for multivariate kernel density estimation

Tarn Duong, Arianna Cowling, Inge Koch*, M.P. Wand

School of Mathematics & Statistics, University of New South Wales, Sydney, Australia

Received 7 June 2007; received in revised form 25 February 2008; accepted 25 February 2008
Available online 8 March 2008

Abstract

Multivariate kernel density estimation provides information about structure in data. Feature significance is a technique for deciding whether features – such as local extrema – are statistically significant. This paper proposes a framework for feature significance in d-dimensional data which combines kernel density derivative estimators and hypothesis tests for modal regions. For the gradient and curvature estimators distributional properties are given, and pointwise test statistics are derived. The hypothesis tests extend the two-dimensional feature significance ideas of Godtliebsen et al. [Godtliebsen, F., Marron, J.S., Chaudhuri, P., 2002. Significance in scale space for bivariate density estimation. Journal of Computational and Graphical Statistics 11, 1–21]. The theoretical framework is complemented by novel visualization for three-dimensional data. Applications to real data sets show that tests based on the kernel curvature estimators perform well in identifying modal regions. These results can be enhanced by corresponding tests with kernel gradient estimators.

1. Introduction

In the analysis of three or higher-dimensional data significant features are of prime interest rather than estimation of the whole data density. For one- and two-dimensional data Chaudhuri and Marron (1999) and Godtliebsen et al. (2002) regarded as significant features local extrema, valleys, ridges, saddle points and steep gradients, and they developed techniques for determining and visualizing these features. Chaudhuri and Marron (2000) and Hannig and Marron (2006) produced asymptotic distributional results for the one-dimensional case. As the number of dimensions increases, local maxima become the single most important features, and identification of these maxima is the goal.

In this paper we present a treatment of feature significance for multivariate data and kernel density estimation which extends the one- and two-dimensional results of Chaudhuri and Marron (1999) and Godtliebsen et al. (2002) in two important ways. We consider multivariate data in arbitrary dimensions, and we describe novel visualization of significant features in three-dimensional data. The key components are kernel density derivative estimators and their tuning parameters. We derive theoretical properties of these estimators and provide a framework for hypothesis testing.

Multivariate feature significance is not just a technically more challenging problem than uni- or bivariate feature significance. Rather, it extends as well as differs from the previous cases in a number of ways. In dimensions greater

* Corresponding author.
E-mail address: inge@unsw.edu.au (I. Koch).

0167-9473/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2008.02.035

than two only modes are well-defined features, while one- and two-dimensional saddle points or ridges do not extend in an obvious way, and valleys are often of less interest as the dimension increases. This difference in the set of relevant features impacts on the framework and hypothesis tests we propose. For multivariate kernel density estimation general bandwidth matrices induce rotation in the kernels with no analogue in one dimension. Such bandwidth matrices are better suited to deal with correlated data or with variables whose ranges differ substantially. Further, the computational complexity of multivariate or high-dimensional data requires different estimation approaches in particular when combined with sparsity of data in high dimensions.

Kernel density estimation has been a popular technique for analysing one and two-dimensional data; see Bowman and Azzalini (1997), Scott (1992), Simonoff (1996) and Wand and Jones (1995) for examples. Scott (1992)'s book applies to multivariate density estimation, and Stone (1980) focuses on bandwidth selection for general multivariate data based on diagonal bandwidth matrices. More general bandwidth choices for multivariate density estimation are described in Duong and Hazelton (2005). Density estimates provide useful information about features in the data. Since the quantitative information about features is contained in the first and second derivative of the true density f, natural estimators of the derivatives of f are the derivatives of the kernel density estimator $\hat{f}$.

In the context of kernel density or kernel density derivative estimation bandwidth selection affects the performance of an estimator. In feature significance we focus on a range of bandwidths, rather than a 'best' bandwidth selection. For one-dimensional data this approach leads to the 'Sizer' plots of Chaudhuri and Marron (1999). For bivariate data Godtliebsen et al. (2002) employ diagonal bandwidth matrices with the same bandwidth for both dimensions, thus reducing the two-dimensional problem to one dimension. For the single diagonal bandwidth we can then proceed as in the one-dimensional 'Sizer' case. We regard the bandwidth selection from a practical point only and briefly suggest – in Section 3.4 – possible ways of selecting bandwidths in general d-dimensional applications.

Section 2 establishes our framework and describes the kernel derivative estimators. Theoretical results of our estimators are presented in this section. Simultaneous hypothesis tests based on the kernel derivative estimators are described in Section 3. Section 4 shows how our results work in practice and demonstrates our visualization tools. Finally the Appendix contains proofs of the theoretical results in Section 2.

2. Kernel density derivative estimation

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

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

where K is a kernel function which is a symmetrical probability density function, $\mathbf{H}$ is a positive-definite, symmetrical bandwidth matrix, and $K_{\mathbf{H}}$ is the scaled kernel function which is defined by

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

with $|\mathbf{H}|$ the determinant of the matrix $\mathbf{H}$.

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

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

In this paper, the first and second derivatives are of interest. We write $\widehat{\nabla f}$ instead of $\widehat{\nabla^{(1)} f}$ for the kernel gradient estimator

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

where ∇ is the column vector of the d partial first derivatives and

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

By convention we take derivatives after scaling with the bandwidth matrix $\mathbf{H}$.

Similarly the kernel curvature estimator

$$\widehat{\nabla^{(2)} f}(\boldsymbol{x}; \mathbf{H}) = n^{-1} \sum_{i=1}^{n} \nabla^{(2)} K_{\mathbf{H}}(\boldsymbol{x} - \boldsymbol{X}_i), \tag{3}$$

where $\nabla^{(2)}$ denotes the matrix of all second-order partial derivatives, and

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

We now turn to properties of these two estimators, which we present separately for the gradient and curvature estimator.

2.1. Kernel gradient estimation

We now focus on gradient estimation. For the asymptotic distribution of $\widehat{\nabla f}$ of (2) we require statements (A1)–(A4) below:

(A1) All entries of $\mathbf{H} \to 0$ and $n^{-1}|\mathbf{H}|^{-1/2}\mathbf{H}^{-1} \to 0$, as $n \to \infty$.
(A2) All entries of $\nabla \otimes \nabla^{(2)} f(\boldsymbol{x})$ are bounded, continuous and square integrable.
(A3) The kernel K is a symmetrical probability density function, ∇K and $(\nabla_{\mathbf{H}} \otimes \nabla) K$ are bounded, and $\mu_2(K)\mathbf{I} \equiv \int_{\mathbb{R}^d} K(\boldsymbol{x}) \boldsymbol{x}\boldsymbol{x}^{\mathrm{T}} \mathrm{d}\boldsymbol{x}$ satisfies $\mu_2(K) < \infty$ where $\mathbf{I}$ is the $d \times d$ identity matrix.
(A4) All entries of $\mathbf{R}(\nabla K) < \infty$ where $\mathbf{R}(\boldsymbol{g}) = \int_{\mathbb{R}^d} \boldsymbol{g}(\boldsymbol{x}) \boldsymbol{g}(\boldsymbol{x})^{\mathrm{T}} \mathrm{d}\boldsymbol{x}$.

In statements (A1)–(A4) we introduced the notation $\mathbf{A} \otimes \mathbf{B}$ for the Kronecker or tensor product of matrices $\mathbf{A}$ and $\mathbf{B}$; and hence $\nabla \otimes \nabla^{(2)}$ refers to the partial derivatives of third order. $\nabla_{\mathbf{H}} \equiv \partial/(\partial \operatorname{vech} \mathbf{H})$ and $\operatorname{vech} \mathbf{H}$ is the vector half operator which transforms a symmetric $d \times d$ matrix, here $\mathbf{H}$, into a vector of length $d^* = (d+1)d/2$ by stacking the elements of the upper triangular half of the matrix. For a 3×3 symmetrical matrix this operation results in

$$\operatorname{vech} \begin{bmatrix} a & b & c \\ b & d & e \\ c & e & f \end{bmatrix} = \begin{bmatrix} a & b & c & d & e & f \end{bmatrix}^{\mathrm{T}}, \tag{4}$$

the case for general d follows the same pattern.

Theorem 1. *Assume that* (A1)–(A4) *hold. The asymptotic distribution of the kernel gradient estimator* $\widehat{\nabla f}(\boldsymbol{x}; \mathbf{H})$ *as* $n \to \infty$ *is*

$$n^{1/2}|\mathbf{H}|^{1/4}\mathbf{H}^{1/2} \left[\widehat{\nabla f}(\boldsymbol{x}; \mathbf{H}) - \nabla f(\boldsymbol{x})\right] \to \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Sigma}^{(1)}(\boldsymbol{x})\right),$$

where the convergence is in distribution, $\boldsymbol{\Sigma}^{(1)}(\boldsymbol{x}) = \mathbf{R}(\nabla K) f(\boldsymbol{x})$*, and for any square integrable vector-valued function* $\boldsymbol{g}$*,*

$$\mathbf{R}(\boldsymbol{g}) = \int_{\mathbb{R}^d} \boldsymbol{g}(\boldsymbol{x}) \boldsymbol{g}(\boldsymbol{x})^{\mathrm{T}} \mathrm{d}\boldsymbol{x}.$$

For the special case $\mathbf{H} = \operatorname{diag}(h_1^2, \ldots, h_d^2)$ *and the normal kernel* K *the variance reduces to*

$$\boldsymbol{\Sigma}^{(1)}(\boldsymbol{x}) = \frac{1}{2}(2\pi^{1/2})^{-d} f(\boldsymbol{x}).$$

The proof of Theorem 1 is given in the Appendix. The special case of the diagonal bandwidth matrix is included in the statement of the theorem, as it explicitly shows the dependence of the variance on the density and the bandwidth parameters.

A common performance measure for kernel estimators is the Mean Integrated Squared Error (MISE) and its asymptotic approximation AMISE. For a vector-valued estimator, such as the gradient estimator, the MISE is a matrix. A performance comparison based on scalar quantities rather than on matrices is easier, and for this reason it is natural to consider matrix norms based on MISE. Since the p-norms ($p = 1, 2, \ldots$) are equivalent, we propose to use the

trace- or one-norm which is simple to calculate. The Trace of Asymptotic Mean Integrated Squared Error (TAMISE) of the kernel gradient estimator is

$$\begin{aligned}\mathrm{TAMISE}^{(1)}(\mathbf{H}) &\equiv \mathrm{TAMISE}\widehat{\nabla f}(\cdot;\mathbf{H})\\ &= n^{-1}|\mathbf{H}|^{-1/2}\mathrm{tr}(\mathbf{H}^{-1}\mathbf{R}(\nabla K)) + \frac{1}{4}\mu_2(K)^2(\mathrm{vech}^{\mathrm{T}}\,\mathbf{H})\boldsymbol{\Psi}_6(\mathrm{vech}\,\mathbf{H}).\end{aligned} \tag{5}$$

The first summand is derived from the variance, and the second from the squared bias. We use tr to denote the trace of a matrix.

The term $\boldsymbol{\Psi}_6$ involves higher order derivatives of f, and its subscript, here 6, indicates the order of derivatives used. It is a symmetrical matrix with d^* rows. We defer its definition until the proof of Theorem 2. With this $\boldsymbol{\Psi}_6$ matrix, the second term of (5) is a quadratic form in vech $\mathbf{H}$, which is easy to differentiate with respect to vech $\mathbf{H}$.

Theorem 2. *Assume that* (A1)–(A4) *hold, and that f has derivatives up to 6th order. Let $\mathbf{H}_T^{(1)}$ be the bandwidth matrix which minimizes the TAMISE of* (5) *for the kernel gradient estimator. Then $\mathbf{H}_T^{(1)} = O(n^{-2/(d+6)})\mathbf{J}$ where $\mathbf{J}$ is the $d \times d$ matrix of ones.*

The proof of Theorem 2 is given in the Appendix.

The result shows the dependence of the asymptotic order on the dimension. As the dimension increases, the order decreases. Furthermore, the diagonal elements of the general matrix selector $\mathbf{H}_T^{(1)}$ are the same order as $(h_0^{(1)})^2$, where $\mathbf{H}_0^{(1)} = (h_0^{(1)})^2\mathbf{I}$ is the MISE-optimal bandwidth for the first derivative (eg Stone (1980)).

We indicate practical choices of the bandwidth matrix in Section 3.4, as they apply to both derivative estimators.

2.2. Kernel curvature estimation

In (3) of Section 2 we introduced the kernel curvature estimator in matrix form. It is more convenient to express this estimator in vector form using the vech operator defined in (4). We present our results for the curvature estimator in this latter form. For the distributional results of the kernel curvature estimator we require statements (B1)–(B4).

(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 $\nabla^{(2)} \otimes \nabla^{(2)} f(\boldsymbol{x})$ are bounded, continuous and square integrable.
(B3) The kernel K is a symmetrical probability density function, $\nabla^{(2)}K$ and $(\nabla_{\mathbf{H}} \otimes \nabla^{(2)})K$ are bounded, and $\int_{\mathbb{R}^d} K(\boldsymbol{x})\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}\,\mathrm{d}\boldsymbol{x} = \mu_2(K)\mathbf{I}$ with $\mu_2(K) < \infty$.
(B4) All entries of $\mathbf{R}(\mathrm{vech}\,\nabla^{(2)}K) < \infty$.

Theorem 3. *Assume that* (B1)–(B4) *hold. The asymptotic distribution of the kernel curvature estimator* $\mathrm{vech}\,\widehat{\nabla^{(2)} f}(\boldsymbol{x};\mathbf{H})$ *as $n \to \infty$ is*

$$n^{1/2}|\mathbf{H}|^{1/4}\mathrm{vech}\left[\mathbf{H}^{1/2}\left(\widehat{\nabla^{(2)} f}(\boldsymbol{x};\mathbf{H}) - \nabla^{(2)} f(\boldsymbol{x})\right)\mathbf{H}^{1/2}\right] \to \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Sigma}^{(2)}(\boldsymbol{x})\right),$$

where the convergence is in distribution, and $\boldsymbol{\Sigma}^{(2)}(\boldsymbol{x}) = \mathbf{R}\left(\mathrm{vech}\,\nabla^{(2)}K\right) f(\boldsymbol{x})$.

For the normal kernel K the variance

$$\boldsymbol{\Sigma}_{\mathbf{H}}^{(2)}(\boldsymbol{x}) = n^{-1}|\mathbf{H}|^{-1/2}\mathbf{R}\left(\mathrm{vech}\,(\mathbf{H}^{-1/2}\nabla^{(2)}K\mathbf{H}^{-1/2})\right) f(\boldsymbol{x}) \tag{6}$$

of $\widehat{\nabla^{(2)} f}(\boldsymbol{x};\mathbf{H})$ *reduces to*

$$\begin{aligned}\boldsymbol{\Sigma}_{\mathbf{H}}^{(2)}(\boldsymbol{x}) = \frac{1}{4}(4\pi)^{-d/2}|\mathbf{H}|^{-1/2}[&2(\mathbf{D}_d^{\mathrm{T}}\mathbf{D}_d)^{-1}\mathbf{D}_d^{\mathrm{T}}(\mathbf{H}^{-1}\otimes\mathbf{H}^{-1})\mathbf{D}_d(\mathbf{D}_d^{\mathrm{T}}\mathbf{D}_d)^{-1}\\ &+ (\mathrm{vech}\,\mathbf{H}^{-1})(\mathrm{vech}^{\mathrm{T}}\,\mathbf{H}^{-1})]f(\boldsymbol{x}),\end{aligned}$$

where $\mathbf{D}_d$ is the duplication matrix of order d, see Magnus and Neudecker (1999).

The proof of Theorem 3 is similar to that of Theorem 1.

We next turn to the performance of the kernel curvature estimator. As for the gradient estimator, we consider the TAMISE for vech $\widehat{\nabla^{(2)} f}(\boldsymbol{x}; \mathbf{H})$.

$$\begin{aligned}\mathrm{TAMISE}^{(2)}(\mathbf{H}) &\equiv \mathrm{TAMISE}(\mathrm{vech}\, \widehat{\nabla^{(2)} f}(\cdot; \mathbf{H})) \\ &= n^{-1}|\mathbf{H}|^{-1/2}\mathrm{tr}(\mathbf{D}_d(\mathbf{D}_d^{\mathrm{T}}\mathbf{D}_d)^{-2}\mathbf{D}_d^{\mathrm{T}}(\mathbf{H}^{-1} \otimes \mathbf{H}^{-1})\mathbf{R}(\mathrm{vec}\, \nabla^{(2)} K)) \\ &\quad + \frac{1}{4}\mu_2(K)^2(\mathrm{vech}^{\mathrm{T}}\, \mathbf{H})\boldsymbol{\Psi}_8(\mathrm{vech}\, \mathbf{H}), \end{aligned} \tag{7}$$

where $\boldsymbol{\Psi}_8$ is the square matrix of size d^* which contains derivatives of order 8, and $\mathbf{D}_d$ is as in Theorem 3.

Theorem 4. *Assume that* (B1)–(B4) *hold, and that f has derivatives up to* 8*th order. Let $\mathbf{H}_T^{(2)}$ be the bandwidth matrix which minimizes the TAMISE of* (7) *for the kernel curvature estimator. Then $\mathbf{H}_T^{(2)} = O(n^{-2/(d+8)})\mathbf{J}$ where $\mathbf{J}$ is the $d \times d$ matrix of ones.*

The proof of Theorem 4 is given in the Appendix, by drawing on Lemmas 4 and 5.

A comparison of our TAMISE optimal bandwidth matrices with the results obtained in Stone (1980) shows that our results extend the optimal bandwidth of order $n^{-1/(2r+d+4)}$ for his single scalar $h_0^{(r)}$ to the full bandwidth matrix.

3. Hypothesis tests for derivative estimators

The basic methodology of feature significance requires testing for regions in which the derivatives are significantly different from zero. We propose separate tests for the gradient and curvature derivatives. Our tests differ in a number of ways from those of Chaudhuri and Marron (2000) and Godtliebsen et al. (2002). We discuss these differences in Sections 3.1 and 3.2 and summarize them at the end of Section 3.2. Although the distributional results of our two types of tests are similar, the regions in which we expect to reject the null hypothesis are very different, or even complementary. We explore these *null rejection regions* in Section 3.3 and indicate consequences for applications.

3.1. Tests for the kernel density gradient estimator

For $\boldsymbol{x} \in \mathbb{R}^d$, a null hypothesis for gradient testing is

$$H_0 : \|\nabla f(\boldsymbol{x})\| = 0,$$

where $\|\cdot\|$ is the Euclidean norm.

Consider a bandwidth matrix $\mathbf{H}$ suitable for kernel density gradient estimators. For large n put

$$\boldsymbol{\Sigma}_{\mathbf{H}}^{(1)}(\boldsymbol{x}) = n^{-1}|\mathbf{H}|^{-1/2}\mathbf{H}^{-1/2}\mathbf{R}(\nabla K)\mathbf{H}^{-1/2} f(\boldsymbol{x}). \tag{8}$$

Using this notation and Theorem 1, the asymptotic null distribution after normalization is

$$\{\boldsymbol{\Sigma}_{\mathbf{H}}^{(1)}(\boldsymbol{x})\}^{-1/2}\widehat{\nabla f}(\boldsymbol{x}; \mathbf{H}) \sim \mathcal{N}(\mathbf{0}, \mathbf{I}).$$

Since $\|\nabla f(\boldsymbol{x})\|$ is the quantity of interest in the hypothesis test, an appropriate test statistic is

$$W^{(1)}(\boldsymbol{x}) = \|\{\boldsymbol{\Sigma}_{\mathbf{H}}^{(1)}(\boldsymbol{x})\}^{-1/2}\widehat{\nabla f}(\boldsymbol{x}; \mathbf{H})\|^2, \tag{9}$$

which requires an estimate for $\boldsymbol{\Sigma}_{\mathbf{H}}^{(1)}(\boldsymbol{x})$. A natural choice is to replace the true density f at $\boldsymbol{x}$ in (8) by its kernel density estimator $\hat{f}(\boldsymbol{x}; \mathbf{H})$ and to use

$$\widehat{\boldsymbol{\Sigma}}_{\mathbf{H}}^{(1)}(\boldsymbol{x}) = n^{-1}|\mathbf{H}|^{-1/2}\mathbf{H}^{-1/2}\mathbf{R}(\nabla K)\mathbf{H}^{-1/2}\hat{f}(\boldsymbol{x}; \mathbf{H}) \tag{10}$$

as an estimator for $\boldsymbol{\Sigma}_{\mathbf{H}}^{(1)}(\boldsymbol{x})$. This choice has the advantage that the theoretical properties of the estimator follow directly from Theorem 1. In contrast, for Godtliebsen et al. (2002)'s estimator

$$n^{-1}(n-1)^{-1}\sum_{i=1}^{n}\left[\nabla K_{\mathbf{H}}(\boldsymbol{x} - \boldsymbol{X}_i) - \widehat{\nabla f}(\boldsymbol{x}; \mathbf{H})\right]\left[\nabla K_{\mathbf{H}}(\boldsymbol{x} - \boldsymbol{X}_i) - \widehat{\nabla f}(\boldsymbol{x}; \mathbf{H})\right]^{\mathrm{T}}$$

no such distributional results are available.

Since the scalar term of $\widehat{\mathbf{\Sigma}}_{\mathbf{H}}^{(1)}(\boldsymbol{x})$ in (10), namely $n^{-1}|\mathbf{H}|^{-1/2}\hat{f}(\boldsymbol{x};\mathbf{H})$, is positive and its matrix component, $\mathbf{H}^{-1/2}\mathbf{R}(\nabla K)\mathbf{H}^{-1/2}$, is positive definite and invertible, we put

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

It follows that $\widehat{W}^{(1)}(\boldsymbol{x})$ is approximately asymptotically chi-squared with d degrees of freedom, so $\widehat{W}^{(1)}(\boldsymbol{x}) \overset{\text{approx.}}{\sim} \chi^2_d$.

Our aim is to test simultaneously at all points $\boldsymbol{x}$ whether $\widehat{W}^{(1)}(\boldsymbol{x})$ shows a significant deviation from zero. Such tests are highly correlated since for nearby points $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, $\widehat{W}^{(1)}(\boldsymbol{x}_1)$ and $\widehat{W}^{(1)}(\boldsymbol{x}_2)$ are highly correlated. The approach of Godtliebsen 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, as outlined in Shaffer (1995). We follow Hochberg (1988) which is based on Simes (1986). We choose Hochberg's procedure 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. Assume that the p-values for each of the m individual tests are 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 $H_{0,(1)}, \ldots, H_{0,(j)}$. We find the largest such $j = j_{\max}$ and

$$\text{reject } H_{0,(1)}, \ldots, H_{0,(j_{\max})} \text{ where } j_{\max} = \operatorname*{argmax}_{1\leq j\leq m} P_{(j)} \leq \alpha/(m-j+1). \tag{11}$$

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

In our case, the p-value at $\boldsymbol{x}$ is $\mathbb{P}(U > \widehat{W}^{(1)}(\boldsymbol{x}))$ where $U \sim \chi^2_d$. Like Godtliebsen et al. (2002), we restrict significance testing to regions with a sufficient number of data points, namely an 'effective sample size' $\text{ESS}(\boldsymbol{x}) \geq n_0$ where

$$\text{ESS}(\boldsymbol{x}) = \sum_{i=1}^{n} K_{\mathbf{H}}(\boldsymbol{x}-\boldsymbol{X}_i) \Big/ K_{\mathbf{H}}(\mathbf{0}).$$

We choose $n_0 = 5$ as a minimum effective sample size. This is analogous to choosing 5 as the minimum individual expected cell counts in a χ^2 test of independence.

3.2. Tests for the kernel density curvature estimator

For curvature testing, our null hypothesis is

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

Consider a bandwidth matrix $\mathbf{H}$ suitable for kernel density curvature estimators and $\mathbf{\Sigma}_{\mathbf{H}}^{(2)}(\boldsymbol{x})$ as in (6). From Theorem 3, the asymptotic null distribution of $\text{vech}\,\widehat{\nabla^{(2)} f}(\boldsymbol{x};\mathbf{H})$ after normalizing is

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

where $\mathbf{I}_{d^*}$ is the $d^* \times d^*$ identity matrix with $d^* = (d+1)d/2$. In analogy with the gradient test statistic $W^{(1)}(\boldsymbol{x})$ the curvature test statistic is

$$W^{(2)}(\boldsymbol{x}) = \|\{\mathbf{\Sigma}_{\mathbf{H}}^{(2)}(\boldsymbol{x})\}^{-1/2}\text{vech}\,\widehat{\nabla^{(2)} f}(\boldsymbol{x};\mathbf{H})\|^2. \tag{12}$$

This choice of test statistic is markedly different from those of Godtliebsen et al. (2002). These authors used as their test statistic $\max\{|\lambda_1(\boldsymbol{x})|, \ldots, |\lambda_d(\boldsymbol{x})|\}$ where $\lambda_j(\boldsymbol{x})$ is the jth eigenvalue of the normalized $\widehat{\nabla^{(2)} f}(\boldsymbol{x};\mathbf{H})$. They wished to characterize the different types of significant curvature such as significant positive, significant negative or non-significant eigenvalues. This inclusive approach is able to detect, in addition to modes, other features such as valleys, ridges and saddle points in bivariate data. In higher dimensions, however, the identification and interpretation of equivalents of the latter mathematical features is not clear. As pointed out in the introduction, for three or more

dimensions, significant features refer to significant modes (which *do* have a clear interpretation in any dimension). Thus we only need to establish the existence of significant curvature without having to characterize it by its eigenspace structure. A side effect of our approach is that we circumvent the need to simulate critical points of the null distribution, which is required in Godtliebsen et al. (2002)'s test statistic, since $W^{(2)}(\boldsymbol{x})$ of (12) has an approximate closed form null distribution.

An estimate of $W^{(2)}(\boldsymbol{x})$ is

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

where

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

This estimator of $\boldsymbol{\Sigma}_{\mathbf{H}}^{(2)}(\boldsymbol{x})$ is derived from Theorem 3 analogously to the variance estimator in (10), and is an alternative to the estimator of Godtliebsen et al. (2002). The latter estimator relies on individually estimating the elements of $\boldsymbol{\Sigma}_{\mathbf{H}}^{(2)}(\boldsymbol{x})$ by the individual sample variances and covariances. However, individually estimating the elements of a matrix quantity is not always optimal and cannot guarantee positive definiteness of the matrix of individual estimates. On the other hand, our method estimates $\boldsymbol{\Sigma}_{\mathbf{H}}^{(2)}(\boldsymbol{x})$ in a matrix-wise procedure, and positive definiteness is guaranteed. Another advantage of our approach is that it is computationally more efficient since it no longer requires the computationally intensive step of calculating the (many) sample variances and covariances of the elements of $\operatorname{vech}\widehat{\nabla^{(2)} f}(\boldsymbol{x};\mathbf{H})$. Our method allows general bandwidth matrices, whereas Godtliebsen et al. (2002)'s approach relies on using the most restricted parameterization $\mathbf{H} = h^2\mathbf{I}$.

In analogy with the gradient test, we use a Hochberg procedure to test simultaneously for significance of $\widehat{W}^{(2)}(\boldsymbol{x})$. For curvature tests the p-value at $\boldsymbol{x}$ is $\mathbb{P}(U > \widehat{W}^{(2)}(\boldsymbol{x}))$ where $U \sim \chi^2_{d^*}$, so the main difference to the gradient test is the increase in the degrees of freedom – from d to $d^* = (d+1)d/2$ – of the chi-squared distribution.

A comparison with the tests used in Godtliebsen et al. (2002) shows a number of differences:

1. Our testing framework focuses on modes, applies to arbitrary data dimensions and allows general bandwidth selectors.
2. The theoretical properties of our variance estimator (10) follow from Theorem 1. Similar theoretical results are not readily available for their estimator.
3. Our variance estimation of kernel density derivative estimators is matrix-based rather than element-wise based, so preserves positive definiteness. As our variance estimate requires only kernel density estimates rather than variances and covariances of kernel derivative estimates, it is computationally less intensive and therefore more adaptable to higher dimensions.
4. The tests for nearby points are highly correlated. We apply the multiple testing procedures of Hochberg (1988) directly to the tests, while Godtliebsen et al. (2002) reduce the dependent tests first to independent ones and then use Bonferroni-type tests.
5. Our test statistic for the density curvature estimator is based on a Wald-type statistic which differs greatly from the test statistic based on eigenvalues. Since our test statistic has a closed form, critical points of the null distribution do not need to be simulated – the latter is required in Godtliebsen et al. (2002). Our approach therefore leads to a computational advantage especially for higher dimensions.

3.3. The gradient and curvature rejection regions

A comparison of the gradient- and curvature-based tests for modal regions shows that they are complementary in the following sense. For the curvature-based test the rejection region should include the modal regions, since significant deviations from the null hypothesis are expected to occur at or near the true modes. So the rejection regions of curvature-based tests 'cover' the modal regions. In contrast, for gradient-based tests the rejection region excludes the modal and anti-modal regions, but can contain all other regions provided sufficient data are available. These observations suggest that curvature-based tests are the prime object of interest. The gradient-based tests can further enhance the results of the curvature-based tests, but may not be as informative by themselves.

Next we consider the effect of sample size. As the sample size grows, the null hypothesis is rejected more frequently. This is a consequence of the framework of classical testing and the asymmetry between acceptance and

rejection region. For curvature-based testing the rejection region will grow with sample size and therefore contain modal and near modal regions. This is a desirable outcome. For gradient-based testing the rejection regions will also grow with sample size, but the crucial difference is that these rejection regions could 'eat' into the modal regions and eventually absorb the modal regions. We could try to lower the significance level of the gradient test and thus find a compromise between sample size and rejection region. We explore the effect of lowering the significance level in our applications.

3.4. Practical bandwidth choices

The calculation of either of the two test statistics requires a bandwidth matrix. For the d-dimensional setting a range of smoothing parameter methods is required to cover the needs of users and data. For a quick glance at data the same single bandwidth for all dimensions will give some understanding, but for a deeper analysis of correlated data or variables with different ranges, a general bandwidth matrix may do more justice to the data and is therefore more appropriate. This is especially the case if the computational effort is not the prime concern. General bandwidth matrices will also lead to an understanding of good theoretical and practical bandwidth choices. To appeal to many applications our method handles smoothing parameters which range from user supplied values to an objectively chosen and data driven bandwidth matrix.

While we do not want to exclude the former (our software allows the user to vary the bandwidth for each variable), we need to look at other choices. Full plug-in selectors like Wand and Jones (1994) or Duong and Hazelton (2003) could be developed for derivative estimators, but this is beyond the scope of this paper. Published research on the selection of smoothing parameters for kernel density derivative estimators has focused on univariate estimators – see for example Singh (1987), Härdle et al. (1990) and Wu (1997). For bivariate data Godtliebsen et al. (2002) – and for general multivariate data Stone (1980) – focus on diagonal bandwidth matrices with the same bandwidth for all dimensions. Such univariate-like parameterization represents a useful bandwidth choice as it gives the user the whole range of one-dimensional bandwidths as in 'Sizer'.

If this single-parameter choice is too restrictive, one could progress to a diagonal bandwidth matrix with different bandwidths for each variable. The diagonal parameterization of a plug-in selector is an obvious candidate. To the authors' knowledge no such method is currently available for density derivative estimation. We therefore adjust Duong and Hazelton (2003)'s diagonal selectors to derivatives as an 'ad hoc' way of obtaining a diagonal bandwidth matrix.

We take into account the marginal sample variances S_i and the sample size n, and starting with a kernel density bandwidth matrix $\mathbf{H} = \operatorname{diag}(h_1^2, \ldots, h_d^2)$, we obtain the gradient bandwidth matrix

$$\mathbf{H}^{(1)} = \operatorname{diag}(h_{g,1}^2, \ldots, h_{g,d}^2) \quad \text{with } h_{g,i}^2 = h_i^2 \frac{(nS_i)^{-2/(d+6)}}{(nS_i)^{-2/(d+4)}}, \tag{13}$$

and similarly the curvature bandwidth matrix

$$\mathbf{H}^{(2)} = \operatorname{diag}(h_{c,1}^2, \ldots, h_{c,d}^2) \quad \text{with } h_{c,i}^2 = h_i^2 \frac{(nS_i)^{-2/(d+8)}}{(nS_i)^{-2/(d+4)}}. \tag{14}$$

These rules represent a simplistic way of choosing diagonal bandwidth matrices. They form a compromise between single bandwidth choice and full bandwidth matrix if *one* bandwidth matrix is desired rather than a sequence of bandwidth matrices as advocated in Figs. 3 and 5. The bandwidth matrices (13) and (14) are optimal in the sense that they have the same asymptotic order as the TAMISE optimal selectors which are given in Theorems 2 and 4, namely $n^{-2/(d+6)}$ for the gradient and $n^{-2/(d+8)}$ for the curvature estimator. Further they are analogous to the normal reference or 'rule-of-thumb' selectors and form a starting point for more elaborate choices which could be explored.

4. Real data examples

We apply the gradient and curvature tests to two examples: one of moderate sample size, the Mt St Helens data, and a very large dataset from flow cytometry.

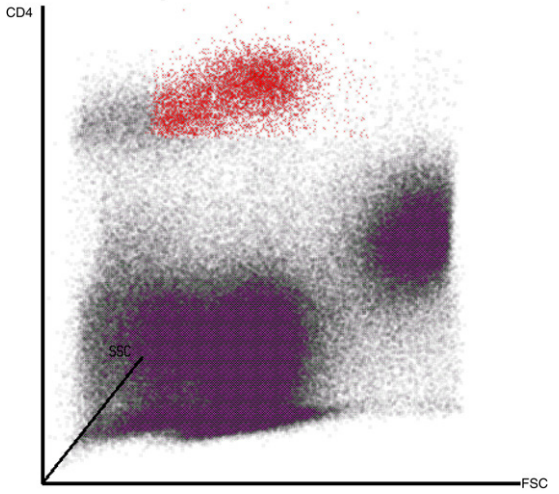

Fig. 1. Scatterplot of three-dimensional flow cytometric data set with subjectively gated lymphocytes (red). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

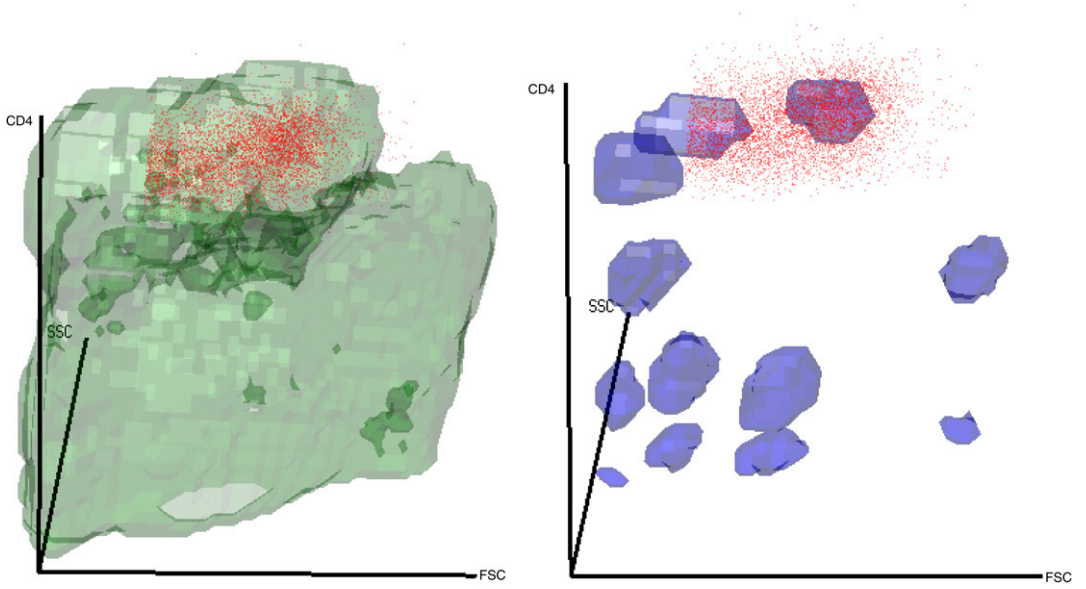

Fig. 2. Significant gradient (left) and curvature regions (right) with subjectively gated lymphocytes (red cluster) and adjusted plug-in bandwidth matrices $\mathbf{H}^{(1)}$ for the gradient and $\mathbf{H}^{(2)}$ for the curvature tests. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

4.1. Flow cytometry data

Flow cytometry is an emerging technique for identifying structure and properties of large cell populations from measurements of fluorescent intensity emitted by the cells. Typically 3–5 different measurements are made but more recent technologies allow up to 15 different measurements. The data are used in biotechnology to separate or *gate* different cell types, for example, to further the understanding of the differences in the cells of HIV− and HIV+ individuals. For details on flow cytometry see Givan (2001).

We apply the results of the previous sections to the flow cytometry data `unst.DRT` which is available from the `rflowcyt` package (Rossini et al., 2005) in the statistical programming language `R` (R Development Core Team, 2005). It contains measurements of forward scatter intensity (FSC), side scatter intensity (SSC) and the level of CD4

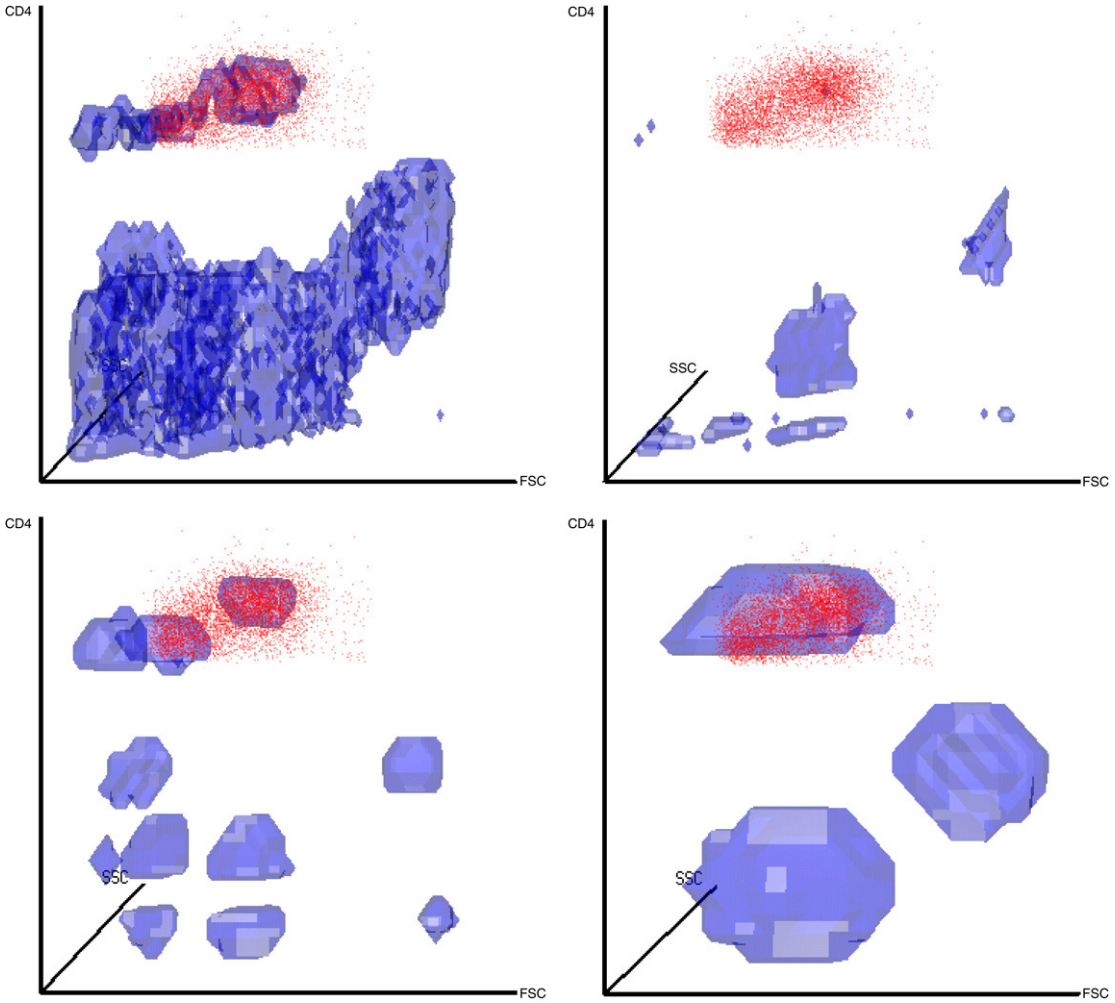

Fig. 3. Significant curvature regions for different bandwidth matrices: Scale space for flow cytometry data — significant curvature regions. Upper left – $\mathbf{H}^{0.8}$ (top left), $\mathbf{H}^{1.1}$ (top right), $\mathbf{H}^{1.5}$ (bottom left) and $\mathbf{H}^{2}$ (bottom right).

antigen (CD4) for 194 629 cells in a mixed cell population containing lymphocytes, monocytes and granulocytes from an HIV– individual. The aim of our analysis is two-fold: we want to extract significant structure in the data, but we also want to provide a more objective method for gating. Gating refers to the isolating or separating of cells of a particular type, for example, lymphocytes in a mixed cell population which are a 'particular type of white blood cell involved in many of an organism's immune response', see Givan (2001, p. 249). Current gating methods are subjective and based on the user's experience.

Fig. 1 shows a three-dimensional scatter plot of these data after debris cells have been removed. A total of 175 098 cells are used in this scatter plot. Of special interest are lymphocytes with higher levels of CD4 antigen expression (which are marked in red in Fig. 1). This cluster represents the points inside a hyperrectangular gate obtained from a flow cytometry scientist for these data. We compare these gated cells or points with the significant feature regions we obtain for this part of the data.

We calculate significant regions for the $\alpha = 0.05$ level of significance – the default throughout this paper. Our estimates are based on the diagonal parameterization of the plug-in selector $\mathbf{H} = \mathrm{diag}(11.3^2, 6.26^2, 10.0^2)$ of Duong and Hazelton (2003). We apply the transformations given in (13) and (14) to $\mathbf{H}$ and display the significant gradient regions (left panel) and the significant curvature regions (right panel) in Fig. 2. Here we have used the gradient bandwidth matrix $\mathbf{H}^{(1)} = \mathrm{diag}(19.48^2, 10.45^2, 17.13^2)$ on the left and the curvature bandwidth matrix $\mathbf{H}^{(2)} = \mathrm{diag}(27.53^2, 12.48^2, 24.13^2)$ on the right. Overlaid are the gated points from Fig. 1. On the left, the gated

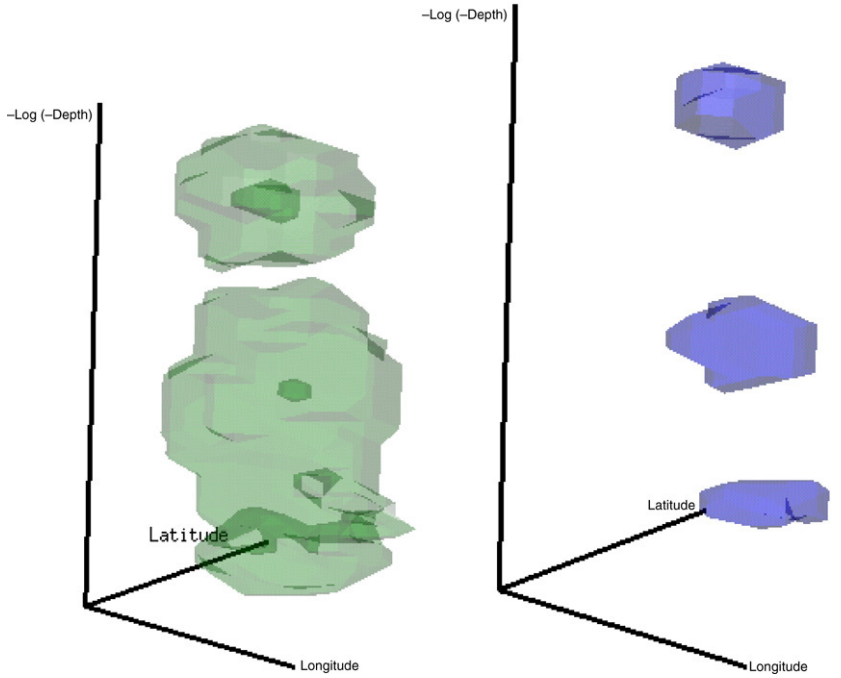

Fig. 4. Significant gradient (left) and curvature (right) regions for Mt St Helens earthquake data. Left with bandwidths $\mathbf{H}^{(1)}$, right with $\mathbf{H}^{(2)}$.

points intersect with one significant gradient region, and the right, they contain one significant curvature region and intersect with another. The plot on the right shows significant regions and suggests that they could lead to a more automatic and objective way of defining gates. The gradient plot on the left indicates a clear anti-mode – shown darker here and separating the higher levels of CD4 from the lower ones. As expected from the large sample size, the rejection region of the gradient test results has become too large. We tried much smaller values for the significance level, but even for very small values, the rejection regions did not change much. In the next part we will therefore focus on curvature-based tests only.

In their two-dimensional scale space approach Godtliebsen et al. (2002) use a range of bandwidths. These show how different significant features emerge or merge and thus give a better understanding of the underlying structure in the data. We mimic their approach by showing a sequence of plots for four different bandwidth matrices. In each case we use different powers of the density plug-in bandwidth matrix $\mathbf{H} = \mathrm{diag}(11.3^2, 6.26^2, 10.0^2)$, namely 0.8, 1.1, 1.5 and 2. The resulting significant curvature regions are shown in Fig. 3. The bandwidth matrix for the top left figure is very small. As a result a large number of modal regions appear which seem to merge into two big regions. The top right figure is also undersmoothed and results in some spurious modal regions. When the bandwidth matrix is close to that of Fig. 2 – in the bottom left figure – significant feature regions appear again in the region of the gated points, and as the bandwidths increase further, the number of significant regions decreases whilst their individual volumes increase. The sequence of bandwidth plots confirms the adequacy of our adjusted plug-in bandwidth matrix $\mathbf{H}^{(2)}$ when a single bandwidth matrix is required.

The sequence of figures, especially those arising from significant curvature regions, allow determination of cell subpopulations in mixed cell populations which is of particular interest to biologists. Such an interpretation suggests a link between significant features in multivariate density estimation and unsupervised classification of data into different classes.

4.2. Mt St Helens data

The Mt St Helens earthquake data are taken from Scott (1992). The data consist of measurements of the epicentres of 510 earthquakes which took place beneath the Mt St Helens volcano before its 1982 eruption. We take the first three variables longitude (degrees), latitude (degrees) and depth (km), from the full set of 5 variables. Following Scott (1992, Color Plate 8), the depth variable z is transformed to $-\log(-z)$, where negative depths indicate distances

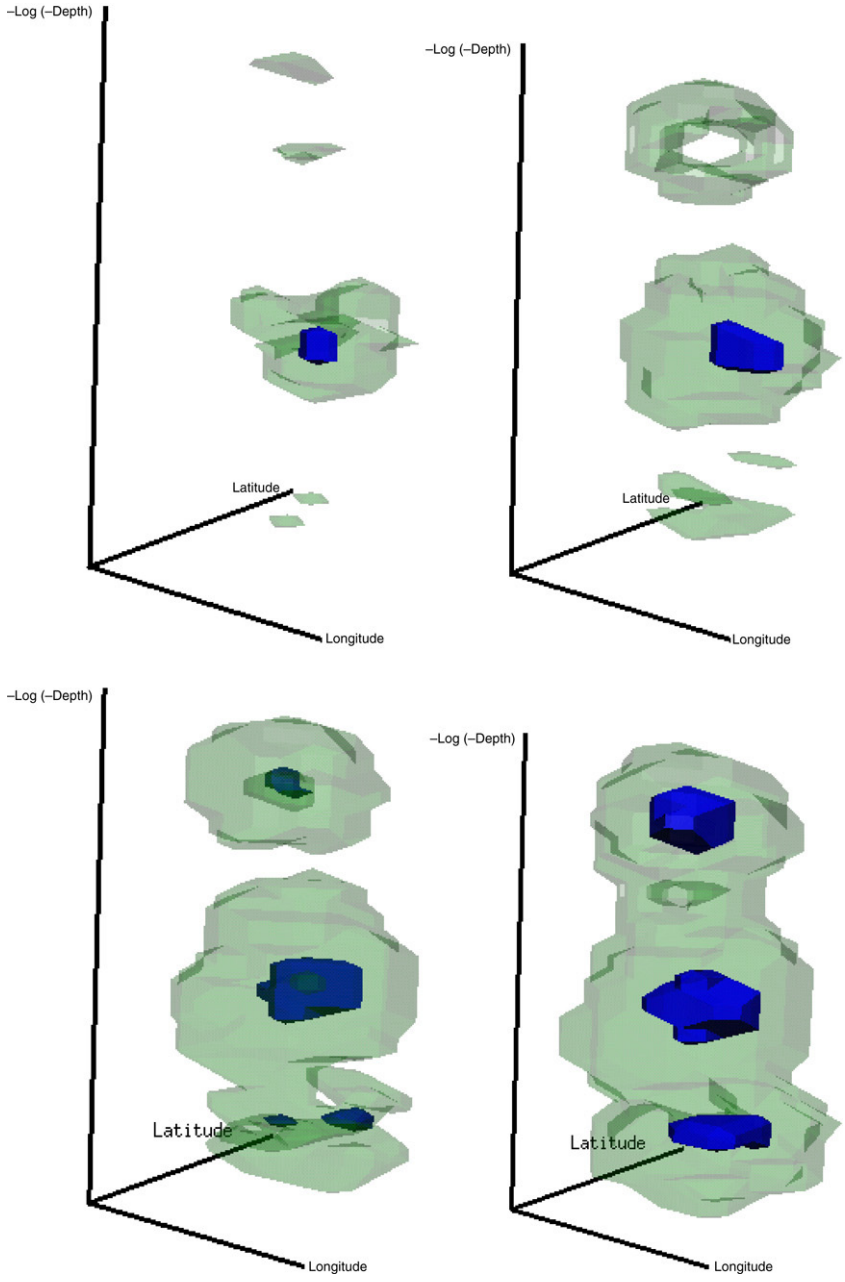

Fig. 5. Significant gradient and curvature regions (darker/blue) for different bandwidth matrices: $\mathbf{H}^{1.1}$ (top left), $\mathbf{H}$ (top right), $\mathbf{H}^{0.9}$ (bottom left) and $\mathbf{H}^{0.8}$ (bottom right).

beneath the Earth's surface. We scale the data, separately for each variable. This sort of scaling is appropriate when the variables are measured in units that are not comparable. As in the previous data, we use the diagonal plug-in selector of Duong and Hazelton (2003) and transform to the gradient and curvature bandwidth selectors. These result in the bandwidth matrix $\mathbf{H}^{(1)} = \text{diag}(0.271^2, 0.263^2, 0.239^2)$ for the gradient, and $\mathbf{H}^{(2)} = \text{diag}(0.308^2, 0.298^2, 0.271^2)$ for the curvature estimator. The significant gradient and curvature regions are shown in Fig. 4 for the scaled data, on the left for $\mathbf{H}^{(1)}$ and on the right for $\mathbf{H}^{(2)}$.

The figures show three major modal regions, corresponding to three different depths of the significant curvature regions. These agree with the three modal regions shown in Scott (1992, Color Plate 8). Unfortunately an exact

quantitative comparison of these modal regions is not possible, since Scott's figure is not marked with axes, nor does he reveal the bandwidths used. Nonetheless, our results qualitatively agree with Scott's, in the sense that we agree on the number of modal regions (three) and the relative sizes of the modal regions (the middle depth modal region is the largest). The left subplot in Fig. 4 suggests that there are a number of modal and anti-modal regions which are indicated by the darker colour. Together with the modal regions of the right subplot, these regions provide a better understanding of the structure of the data.

As for the flow cytometry data, we show a progression of bandwidth matrices through the Mt St Helens data, this time for gradient-based tests, too. We use the density plug-in bandwidth matrix $\mathbf{H} = \operatorname{diag}(0.222^2, 0.216^2, 0.196^2)$ as basis, and consider bandwidth matrices $\mathbf{H}^r$ for the different powers $r = 1.1, 1, 0.9$ and 0.8 of this density plug-in bandwidth matrix $\mathbf{H}$. For each subfigure the significant curvature regions are superimposed as darker (blue) regions on the significant gradient regions. The results are shown in Fig. 5. For the bandwidths used in Fig. 4, $\mathbf{H}^{(1)}$ is 'in between' $\mathbf{H}^{0.9}$ and $\mathbf{H}^{0.8}$, and $\mathbf{H}^{(2)}$ is approximately equal to $\mathbf{H}^{0.8}$. These results suggest that a single well-chosen bandwidth matrix suffices in establishing modal regions in three-dimensional data. Our figures also show that a more comprehensive picture emerges when the significant regions of the two tests are combined.

5. Software

Accompanying this paper is an `R` library, named `feature`, which is available on CRAN at http://cran.r-project.org. It is able to produce feature significance plots for one- to three-dimensional data, though we have only demonstrated here its functionality for three dimensions. The data displays are interactive i.e. the user is able to choose bandwidth(s), and the significant regions are computed in real time. Interactivity is possible since we use a fast implementation of the kernel density derivative estimators with diagonal bandwidth matrices for the required test statistics. This enables the user to explore a range of bandwidths and visualize the significant modal regions in one to three dimensions. The three-dimensional data displays are based on the RGL-type interfaces of the `rgl` (Adler et al., 2006) and `misc3d` (Feng and Tierney, 2005) `R` packages. The plug-in bandwidths of Duong and Hazelton (2003) are implemented in the `R` packages `ks`.

6. Conclusion

Feature significance for modal regions has been extended to d-dimensional data. The theoretical framework combines ideas from kernel density derivative estimation with test statistics based on the first and second derivative estimators and is complemented by novel visualization of interesting regions in three-dimensional data. Feature significance for one- and two-dimensional data refers to local extrema, ridges, valleys and steep gradients. For three- and higher-dimensional data, significant features are essentially captured by significant modal regions or local maxima.

Our examples of real data demonstrate the value of curvature-based tests in finding significant modal regions which can be displayed in an easily interpretable way. Gradient-based tests of significant gradient regions can enhance the understanding of significant structure in data.

Acknowledgement

The authors acknowledge the financial support of the Australian Research Council Project DP055651.

Appendix. Proofs

A.1. Kernel gradient estimation

Lemmas 1–3 are required for the proof of Theorem 1, and the first two are required for the proof of Theorem 2.

Lemma 1. *Under the conditions* (A1)–(A4)*, the expected value of the kernel gradient estimator* $\widehat{\nabla f}$ *is*

$$\mathbb{E}\widehat{\nabla f}(\boldsymbol{x}; \mathbf{H}) = \nabla f(\boldsymbol{x}) + \frac{1}{2}\mu_2(K)(\nabla^{\mathrm{T}} \otimes \nabla^{(2)}) f(\boldsymbol{x})(\operatorname{vec} \mathbf{H}) + O(\|\operatorname{vech} \mathbf{H}^2\|).$$

Proof. The expected value can be expressed as a convolution:

$$\mathbb{E}\widehat{\nabla f}(\boldsymbol{x};\mathbf{H}) = \mathbb{E}\nabla K_{\mathbf{H}}(\boldsymbol{x}-\boldsymbol{X}) = (\nabla K_{\mathbf{H}} * f)(\boldsymbol{x}) = (K_{\mathbf{H}} * \nabla f)(\boldsymbol{x}).$$

So then

$$\begin{aligned}\mathbb{E}\widehat{\nabla f}(\boldsymbol{x};\mathbf{H}) &= \int_{\mathbb{R}^d} K_{\mathbf{H}}(\boldsymbol{x}-\boldsymbol{y})\nabla f(\boldsymbol{y})\,\mathrm{d}\boldsymbol{y}\\ &= \int_{\mathbb{R}^d} K(\boldsymbol{w})\nabla f(\boldsymbol{x}+\mathbf{H}^{1/2}\boldsymbol{w})\,\mathrm{d}\boldsymbol{w}\\ &= \int_{\mathbb{R}^d} K(\boldsymbol{w})\left[\nabla f(\boldsymbol{x}) + \nabla^{(2)} f(\boldsymbol{x})\mathbf{H}^{1/2}\boldsymbol{w}\right.\\ &\quad \left. + \frac{1}{2}(\nabla^{\mathrm{T}}\otimes\nabla^{(2)}) f(\boldsymbol{x})\mathrm{vec}\,(\boldsymbol{w}\boldsymbol{w}^{\mathrm{T}}\mathbf{H})\right]\mathrm{d}\boldsymbol{w} + O(\|\mathrm{vech}\,\mathbf{H}^2\|)\\ &= \nabla f(\boldsymbol{x}) + \frac{1}{2}\mu_2(K)(\nabla^{\mathrm{T}}\otimes\nabla^{(2)}) f(\boldsymbol{x})(\mathrm{vec}\,\mathbf{H}) + O(\|\mathrm{vech}\,\mathbf{H}^2\|)\end{aligned}$$

as condition (A3) gives $\int_{\mathbb{R}^d} K(\boldsymbol{w})\boldsymbol{w}\boldsymbol{w}^{\mathrm{T}}\,\mathrm{d}\boldsymbol{w} = \mu_2(K)\mathbf{I}$. □

Lemma 2. *Under the conditions* (A1)–(A4)*, the variance of the kernel gradient estimator* $\widehat{\nabla f}$ *is*

$$\mathrm{Var}\widehat{\nabla f}(\boldsymbol{x};\mathbf{H}) = n^{-1}|\mathbf{H}|^{-1/2}\mathbf{H}^{-1/2}\mathbf{R}(\nabla K)\mathbf{H}^{-1/2} f(\boldsymbol{x}) + O(n^{-1}|\mathbf{H}|^{-1/2}).$$

Proof. The variance $\mathbf{v} \equiv \mathrm{Var}\widehat{\nabla f}(\boldsymbol{x};\mathbf{H})$ becomes

$$\begin{aligned}\mathbf{v} &= n^{-1}\mathrm{Var}\nabla K_{\mathbf{H}}(\boldsymbol{x}-\boldsymbol{X})\\ &= n^{-1}\{\mathbb{E}[\nabla K_{\mathbf{H}}(\boldsymbol{x}-\boldsymbol{X})][\nabla^{\mathrm{T}} K_{\mathbf{H}}(\boldsymbol{x}-\boldsymbol{X})] - \mathbb{E}[\nabla K_{\mathbf{H}}(\boldsymbol{x}-\boldsymbol{X})]\mathbb{E}[\nabla^{\mathrm{T}} K_{\mathbf{H}}(\boldsymbol{x}-\boldsymbol{X})]\}.\end{aligned}$$

We have

$$\begin{aligned}&\mathbb{E}[\nabla K_{\mathbf{H}}(\boldsymbol{x}-\boldsymbol{X})][\nabla^{\mathrm{T}} K_{\mathbf{H}}(\boldsymbol{x}-\boldsymbol{X})]\\ &= \int_{\mathbb{R}^d} |\mathbf{H}|^{-1}\mathbf{H}^{-1/2}[\nabla K(\mathbf{H}^{-1/2}(\boldsymbol{x}-\boldsymbol{y}))][\nabla^{\mathrm{T}} K(\mathbf{H}^{-1/2}(\boldsymbol{x}-\boldsymbol{y}))]\mathbf{H}^{-1/2} f(\boldsymbol{y})\,\mathrm{d}\boldsymbol{y}\\ &= \int_{\mathbb{R}^d} |\mathbf{H}|^{-1/2}\mathbf{H}^{-1/2}[\nabla K(\boldsymbol{w})][\nabla^{\mathrm{T}} K(\boldsymbol{w})]\mathbf{H}^{-1/2} f(\boldsymbol{x}+\mathbf{H}^{1/2}\boldsymbol{w})\,\mathrm{d}\boldsymbol{w}\\ &= |\mathbf{H}|^{-1/2}\mathbf{H}^{-1/2}\mathbf{R}(\nabla K)\mathbf{H}^{-1/2} f(\boldsymbol{x}) + O(|\mathbf{H}|^{-1/2})\end{aligned}$$

which dominates $\mathbb{E}[\nabla K_{\mathbf{H}}(\boldsymbol{x}-\boldsymbol{X})]\mathbb{E}[\nabla^{\mathrm{T}} K_{\mathbf{H}}(\boldsymbol{x}-\boldsymbol{X})]$. □

Lemma 3. *Assume that the conditions* (A1)–(A2) *hold. Further assume that* K *is the normal kernel and the bandwidth matrix is parameterized* $\mathbf{H} = \mathrm{diag}(h_1^2,\ldots,h_d^2)$*, then* $\boldsymbol{\Sigma}_{\mathbf{H}}^{(1)}(\boldsymbol{x}) = \frac{1}{2}(4\pi)^{-d/2}n^{-1}(h_1\ldots h_d)^{-1}\mathrm{diag}(h_1^{-2},\ldots,h_d^{-2}) f(\boldsymbol{x})$.

Proof. For the multivariate normal kernel $K = \phi_{\mathbf{I}}$, the gradient is $\nabla K(\boldsymbol{x}) = -\phi_{\mathbf{I}}(\boldsymbol{x})\boldsymbol{x}$ then

$$\mathbf{R}(\nabla K) = \int_{\mathbb{R}^d} [\nabla K(\boldsymbol{x})][\nabla^{\mathrm{T}} K(\boldsymbol{x})]\,\mathrm{d}\boldsymbol{x} = \int_{\mathbb{R}^d} \boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}\phi_{\mathbf{I}}(\boldsymbol{x})^2\,\mathrm{d}\boldsymbol{x} = \frac{1}{2}(4\pi)^{-d/2}\mathbf{I}.$$

We substitute this and $\mathbf{H} = \mathrm{diag}(h_1^2,\ldots,h_d^2)$ into

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

Proof of Theorem 1. The asymptotic means and variances follow from the three lemmas above. In their Theorem 3.1 Chaudhuri and Marron (2000) establish the asymptotic normality for kernel density derivative estimators for univariate data with a scalar bandwidth.

Since tr$\mathbf{H}$ is a scalar that is asymptotically at least as large as all elements of $\mathbf{H}$, we apply their theorem componentwise on $\boldsymbol{x}$ with tr$\mathbf{H}$. This establishes the asymptotic normality of $\widehat{\nabla^{(r)}} f(\boldsymbol{x};\mathbf{H})$. □

The proof of Theorem 2 requires the definition of the term

$$\mathbf{\Psi}_6 = \int_{\mathbb{R}^d} \mathbf{D}_d^{\mathrm{T}}(\nabla \otimes \nabla^{(2)}) f(\boldsymbol{x})(\nabla^{\mathrm{T}} \otimes \nabla^{(2)}) f(\boldsymbol{x}) \mathbf{D}_d \, \mathrm{d}\boldsymbol{x}, \tag{15}$$

where $\mathbf{D}_d$ is the duplication matrix of order d so that $\mathbf{D}_d \operatorname{vech} \mathbf{H} = \operatorname{vec} \mathbf{H}$, and the vec notation is illustrated for a 3×3 matrix:

$$\operatorname{vec} \begin{bmatrix} a & b & c \\ b & d & e \\ c & e & f \end{bmatrix} = \begin{bmatrix} a & b & c & b & d & e & c & e & f \end{bmatrix}^{\mathrm{T}}.$$

For details see Magnus and Neudecker (1999). The subscript of $\mathbf{\Psi}_6$ indicates the order of derivatives involved. Expressions like $\nabla \otimes \nabla^{(2)}$ involve 'multiplication' of differentials in the sense that

$$\frac{\partial}{\partial x_i} \frac{\partial}{\partial x_j} = \frac{\partial^2}{\partial x_i \partial x_j}$$

so $\nabla \otimes \nabla^{(2)}$ is an operator of third order differentials.

Proof of Theorem 2. The MSE (Mean Squared Error) of $\widehat{\nabla f}(\boldsymbol{x}; \mathbf{H})$, using Lemmas 1 and 2, is

$$\begin{aligned} \operatorname{MSE} \widehat{\nabla f}(\boldsymbol{x}; \mathbf{H}) &= \operatorname{Var} \widehat{\nabla f}(\boldsymbol{x}; \mathbf{H}) + [\mathbb{E} \widehat{\nabla f}(\boldsymbol{x}; \mathbf{H}) - \nabla f(\boldsymbol{x})][\mathbb{E} \widehat{\nabla f}(\boldsymbol{x}; \mathbf{H}) - \nabla f(\boldsymbol{x})]^{\mathrm{T}} \\ &= n^{-1} |\mathbf{H}|^{-1/2} \mathbf{H}^{-1/2} \mathbf{R}(\nabla K) \mathbf{H}^{-1/2} f(\boldsymbol{x}) \\ &\quad + \frac{1}{4} \mu_2(K)^2 (\nabla^{\mathrm{T}} \otimes \nabla^{(2)}) f(\boldsymbol{x}) (\operatorname{vec} \mathbf{H})(\operatorname{vec}^{\mathrm{T}} \mathbf{H})(\nabla \otimes \nabla^{(2)}) f(\boldsymbol{x}) \\ &\quad + O(n^{-1} |\mathbf{H}|^{-1/2} + \| \operatorname{vech} \mathbf{H}^3 \|). \end{aligned}$$

If we take the trace of this and integrate then

$$\operatorname{TAMISE}^{(1)}(\mathbf{H}) = n^{-1} |\mathbf{H}|^{-1/2} \operatorname{tr}(\mathbf{H}^{-1} \mathbf{R}(\nabla K)) + \frac{1}{4} \mu_2(K)^2 (\operatorname{vech}^{\mathrm{T}} \mathbf{H}) \mathbf{\Psi}_6 (\operatorname{vech} \mathbf{H}).$$

Differentiating this with respect to $\operatorname{vech} \mathbf{H}$

$$\begin{aligned} \nabla_{\mathbf{H}} \operatorname{TAMISE}^{(1)}(\mathbf{H}) &\equiv \frac{\partial \operatorname{TAMISE}^{(1)}(\mathbf{H})}{\partial \operatorname{vech} \mathbf{H}} \\ &= -\frac{1}{2} n^{-1} |\mathbf{H}|^{-1/2} \Big[\operatorname{tr}(\mathbf{H}^{-1} \mathbf{R}(\nabla K)) \mathbf{D}_d^{\mathrm{T}} (\operatorname{vec} \mathbf{H}^{-1}) \\ &\quad + 2 \mathbf{D}_d^{\mathrm{T}} (\mathbf{H}^{-1} \otimes \mathbf{H}^{-1}) \operatorname{vec} \mathbf{R}(\nabla K) \Big] + \frac{1}{2} \mu_2(K)^2 \mathbf{\Psi}_6 (\operatorname{vech} \mathbf{H}) \end{aligned}$$

since we have for a $d \times d$ matrix $\mathbf{A}$,

$$\nabla_{\mathbf{H}} \operatorname{tr}(\mathbf{H}^{-1} \mathbf{A}) = -\mathbf{D}_d^{\mathrm{T}} (\mathbf{H}^{-1} \otimes \mathbf{H}^{-1}) (\operatorname{vec} \mathbf{A}) = -\mathbf{D}_d^{\mathrm{T}} \operatorname{vec} (\mathbf{H}^{-1} \mathbf{A} \mathbf{H}^{-1})$$

and

$$\nabla_{\mathbf{H}} |\mathbf{H}|^{-1/2} = -\frac{1}{2} |\mathbf{H}|^{-1/2} \mathbf{D}_d^{\mathrm{T}} \operatorname{vec} \mathbf{H}^{-1}.$$

Let $\mathbf{H} = O(h^2)\mathbf{J}$ where $\mathbf{J}$ is the $d \times d$ matrix of ones. Then the first term is $O(n^{-1} h^{-d-4})$ and the second term is $O(h^2)$. Let $\mathbf{H}_T^{(1)}$ be a solution of $\nabla_{\mathbf{H}} \operatorname{TAMISE}^{(1)}(\mathbf{H}) = \mathbf{0}$ then $\mathbf{H}_T^{(1)} = O(n^{-2/(d+6)})\mathbf{J}$. This is a larger bandwidth than the AMISE-optimal $O(n^{-2/(d+4)})\mathbf{J}$ bandwidth for estimating f.

A.2. Kernel curvature estimation

Lemmas 4–6 are required in the proof of Theorem 3, while only Lemmas 4 and 5 are necessary for the proof of Theorem 4.

For the proof of Lemma 4 and of Theorem 4, we require Moore–Penrose inverses. Let $\mathbf{D}_d^+ = (\mathbf{D}_d^{\mathrm{T}}\mathbf{D}_d)^{-1}\mathbf{D}_d^{\mathrm{T}}$ be the unique Moore–Penrose inverse of $\mathbf{D}_d$, then $\operatorname{vech}\mathbf{H} = \mathbf{D}_d^+\operatorname{vec}\mathbf{H}$, and $(\mathbf{D}_d^+)^{\mathrm{T}}\mathbf{D}_d^+ = \mathbf{D}_d(\mathbf{D}_d^{\mathrm{T}}\mathbf{D}_d)^{-2}\mathbf{D}_d^{\mathrm{T}}$. See Magnus and Neudecker (1999, p. 49).

Lemma 4. *Under the conditions* (B1)–(B4)*, the expected value of the kernel curvature estimator* $\operatorname{vech}\widehat{\nabla^{(2)}f}$ *is*

$$\mathbb{E}\operatorname{vech}\widehat{\nabla^{(2)}f}(\boldsymbol{x};\mathbf{H}) = \operatorname{vech}\nabla^{(2)}f(\boldsymbol{x}) + \frac{1}{2}\mu_2(K)\operatorname{vech}(\nabla^{(2)}\mathbf{H}\nabla^{(2)})f(\boldsymbol{x}) + O(\|\operatorname{vech}\mathbf{H}^2\|).$$

Proof. The expected value is

$$\begin{aligned}
\mathbb{E}\operatorname{vec}\widehat{\nabla^{(2)}f}(\boldsymbol{x};\mathbf{H}) &= \int_{\mathbb{R}^d} K_{\mathbf{H}}(\boldsymbol{x}-\boldsymbol{y})\operatorname{vec}\nabla^{(2)}f(\boldsymbol{y})\,\mathrm{d}\boldsymbol{y}\\
&= \int_{\mathbb{R}^d} K(\boldsymbol{w})\operatorname{vec}(\nabla^{(2)}f(\boldsymbol{x}+\mathbf{H}^{1/2}\boldsymbol{w}))\,\mathrm{d}\boldsymbol{w}\\
&= \int_{\mathbb{R}^d} K(\boldsymbol{w})\operatorname{vec}\nabla^{(2)}f(\boldsymbol{x}) + (\nabla\otimes\nabla^{(2)})f(\boldsymbol{x})\mathbf{H}^{1/2}\boldsymbol{w}\\
&\quad + \frac{1}{2}(\nabla^{(2)}\otimes\nabla^{(2)})f(\boldsymbol{x})\operatorname{vec}(\boldsymbol{w}\boldsymbol{w}^{\mathrm{T}}\mathbf{H}) + O(\|\operatorname{vec}\mathbf{H}^2\|)\\
&= \operatorname{vec}\nabla^{(2)}f(\boldsymbol{x}) + \frac{1}{2}\mu_2(K)(\nabla^{(2)}\otimes\nabla^{(2)})f(\boldsymbol{x})(\operatorname{vec}\mathbf{H}) + O(\|\operatorname{vech}\mathbf{H}^2\|)\\
&= \operatorname{vec}\nabla^{(2)}f(\boldsymbol{x}) + \frac{1}{2}\mu_2(K)\operatorname{vec}(\nabla^{(2)}\mathbf{H}\nabla^{(2)})f(\boldsymbol{x}) + O(\|\operatorname{vech}\mathbf{H}^2\|)
\end{aligned}$$

where $\operatorname{vec}(\mathbf{ABC}) = (\mathbf{C}^{\mathrm{T}}\otimes\mathbf{A})\operatorname{vec}\mathbf{B}$. Then we premultiply by $\mathbf{D}_d^+$ to convert from vec to vech. □

Lemma 5. *Under the conditions* (B1)–(B4)*, the variance of the kernel curvature estimator* $\operatorname{vech}\widehat{\nabla^{(2)}f}$ *is*

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

Proof. We have

$$\begin{aligned}
&\mathbb{E}[\operatorname{vech}\nabla^{(2)}K_{\mathbf{H}}(\boldsymbol{x}-\boldsymbol{X})][\operatorname{vech}^{\mathrm{T}}\nabla^{(2)}K_{\mathbf{H}}(\boldsymbol{x}-\boldsymbol{X})]\\
&= \int_{\mathbb{R}^d}|\mathbf{H}|^{-1}[\operatorname{vech}(\mathbf{H}^{-1/2}\nabla^{(2)}K(\mathbf{H}^{-1/2}(\boldsymbol{x}-\boldsymbol{y}))\mathbf{H}^{-1/2})]\\
&\quad\times[\operatorname{vech}^{\mathrm{T}}(\mathbf{H}^{-1/2}\nabla^{(2)}K(\mathbf{H}^{-1/2}(\boldsymbol{x}-\boldsymbol{y}))\mathbf{H}^{-1/2})]f(\boldsymbol{y})\,\mathrm{d}\boldsymbol{y}\\
&= \int_{\mathbb{R}^d}|\mathbf{H}|^{-1/2}[\operatorname{vech}(\mathbf{H}^{-1/2}\nabla^{(2)}K(\boldsymbol{w})\mathbf{H}^{-1/2})][\operatorname{vech}^{\mathrm{T}}(\mathbf{H}^{-1/2}\nabla^{(2)}K(\boldsymbol{w})\mathbf{H}^{-1/2})]f(\boldsymbol{x}+\mathbf{H}^{1/2}\boldsymbol{w})\,\mathrm{d}\boldsymbol{w}\\
&= |\mathbf{H}|^{-1/2}\mathbf{R}(\operatorname{vech}(\mathbf{H}^{-1/2}\nabla^{(2)}K\mathbf{H}^{-1/2}))f(\boldsymbol{x}) + O(|\mathbf{H}|^{-1/2}\|\operatorname{vech}\mathbf{H}^{-1}\|).
\end{aligned}$$

This expression dominates $\mathbb{E}[\operatorname{vech}\nabla^{(2)}K_{\mathbf{H}}(\boldsymbol{x}-\boldsymbol{X})]\mathbb{E}[\operatorname{vech}^{\mathrm{T}}\nabla^{(2)}K_{\mathbf{H}}(\boldsymbol{x}-\boldsymbol{X})]$. □

Similar to $\boldsymbol{\Psi}_6$ for the gradient result in Theorem 2, for the TAMISE for curvature we define

$$\boldsymbol{\Psi}_8 = \int_{\mathbb{R}^d}\mathbf{D}_d^{\mathrm{T}}(\nabla^{(2)}\otimes\nabla^{(2)})f(\boldsymbol{x})\mathbf{D}_d(\mathbf{D}_d^{\mathrm{T}}\mathbf{D}_d)^{-2}\mathbf{D}_d^{\mathrm{T}}(\nabla^{(2)}\otimes\nabla^{(2)})f(\boldsymbol{x})\mathbf{D}_d\,\mathrm{d}\boldsymbol{x}. \tag{16}$$

Lemma 6. *Assume that the conditions* (B1)–(B2) *hold. Further assume that* K *is the normal kernel then*

$$\boldsymbol{\Sigma}^{(2)}(\boldsymbol{x}) = \frac{1}{4}(4\pi)^{-d/2}|\mathbf{H}|^{-1/2}[2(\mathbf{D}_d^{\mathrm{T}}\mathbf{D}_d)^{-1}\mathbf{D}_d^{\mathrm{T}}(\mathbf{H}^{-1}\otimes\mathbf{H}^{-1})\mathbf{D}_d(\mathbf{D}_d^{\mathrm{T}}\mathbf{D}_d)^{-1} + (\operatorname{vech}\mathbf{H}^{-1})(\operatorname{vech}^{\mathrm{T}}\mathbf{H}^{-1})]f(\boldsymbol{x}).$$

Proof. An alternative expression for the variance $\mathbf{v}$ of the curvature estimator is

$$\mathbf{v} = |\mathbf{H}|^{-1/2}\mathbf{D}_d^+(\mathbf{H}^{-1/2}\otimes\mathbf{H}^{-1/2})\mathbf{R}(\operatorname{vec}\nabla^{(2)}K)(\mathbf{H}^{-1/2}\otimes\mathbf{H}^{-1/2})(\mathbf{D}_d^+)^{\mathrm{T}}f(\boldsymbol{x}) + O(|\mathbf{H}|^{-1/2}\|\operatorname{vech}\mathbf{H}^{-1}\|).$$

For the multivariate normal kernel $K = \phi_{\mathbf{I}}$, the curvature is $\nabla^{(2)} K(\boldsymbol{x}) = \phi_{\mathbf{I}}(\boldsymbol{x})[\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}} - \mathbf{I}]$. From Schott (1996, Theorem 9.19)

$$\begin{aligned}\mathbf{R}(\operatorname{vec}\nabla^{(2)}\phi_{\mathbf{I}}) &= \int_{\mathbb{R}^d} \phi_{\mathbf{I}}(\boldsymbol{x})^2 \Big[\operatorname{vec}(\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}})\operatorname{vec}^{\mathrm{T}}(\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}) - (\operatorname{vec}\mathbf{I})\operatorname{vec}^{\mathrm{T}}(\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}) \\ &\quad - \operatorname{vec}(\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}})(\operatorname{vec}\mathbf{I}) + (\operatorname{vec}\mathbf{I})(\operatorname{vec}^{\mathrm{T}}\mathbf{I})\Big]\,\mathrm{d}\boldsymbol{x} \\ &= \int_{\mathbb{R}^d} \phi_{\mathbf{I}}(\boldsymbol{x})^2 (\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}} \otimes \boldsymbol{x}\boldsymbol{x}^{\mathrm{T}})\mathrm{d}\boldsymbol{x} \\ &= \frac{1}{4}(4\pi)^{-d/2}[2\mathbf{N}_d + (\operatorname{vec}\mathbf{I}_d)(\operatorname{vec}^{\mathrm{T}}\mathbf{I}_d)],\end{aligned}$$

where $\mathbf{N}_d$ is a $d^2 \times d^2$ symmetrical matrix such that $\mathbf{N}_d(\mathbf{A}\otimes\mathbf{A}) = (\mathbf{A}\otimes\mathbf{A})\mathbf{N}_d$ for a $d \times d$ matrix $\mathbf{A}$. Then

$$\begin{aligned}&(\mathbf{H}^{-1/2}\otimes\mathbf{H}^{-1/2})\mathbf{R}(\operatorname{vec}\nabla^{(2)}\phi_{\mathbf{I}})(\mathbf{H}^{-1/2}\otimes\mathbf{H}^{-1/2}) \\ &\quad = \frac{1}{4}(4\pi)^{-d/2}[2(\mathbf{H}^{-1}\otimes\mathbf{H}^{-1})\mathbf{N}_d + (\operatorname{vec}\mathbf{H}^{-1})(\operatorname{vec}^{\mathrm{T}}\mathbf{H}^{-1})],\end{aligned}$$

and we apply $\mathbf{N}_d(\mathbf{D}_d^+)^{\mathrm{T}} = (\mathbf{D}_d^+)^{\mathrm{T}}$ from Magnus and Neudecker (1999, p. 50) and $\mathbf{D}_d^+ = (\mathbf{D}_d^{\mathrm{T}}\mathbf{D}_d)^{-1}\mathbf{D}_d^{\mathrm{T}}$. □

Proof of Theorem 4. The MSE vech $\widehat{\nabla^{(2)} f}(\boldsymbol{x};\mathbf{H})$, using Lemmas 4 and 5, is

$$\begin{aligned}&\operatorname{MSEvech}\widehat{\nabla^{(2)} f}(\boldsymbol{x};\mathbf{H}) \\ &\quad = n^{-1}|\mathbf{H}|^{-1/2}\mathbf{D}_d^+(\mathbf{H}^{-1/2}\otimes\mathbf{H}^{-1/2})\mathbf{R}(\operatorname{vec}\nabla^{(2)}K)(\mathbf{H}^{-1/2}\otimes\mathbf{H}^{-1/2})(\mathbf{D}_d^+)^{\mathrm{T}} f(\boldsymbol{x}) \\ &\qquad + \frac{1}{4}\mu_2(K)^2\mathbf{D}_d^+(\nabla^{(2)}\otimes\nabla^{(2)})f(\boldsymbol{x})(\operatorname{vec}\mathbf{H})(\operatorname{vec}^{\mathrm{T}}\mathbf{H})(\nabla^{(2)}\otimes\nabla^{(2)})f(\boldsymbol{x})(\mathbf{D}_d^+)^{\mathrm{T}} \\ &\qquad + O(n^{-1}|\mathbf{H}|^{-1/2}\|\operatorname{vech}\mathbf{H}^{-1}\| + \|\operatorname{vech}\mathbf{H}^3\|).\end{aligned}$$

If we take the trace of this and integrate then

$$\begin{aligned}\operatorname{TAMISE}^{(2)}(\mathbf{H}) &= n^{-1}|\mathbf{H}|^{-1/2}\operatorname{tr}((\mathbf{H}^{-1}\otimes\mathbf{H}^{-1})\mathbf{R}(\operatorname{vec}\nabla^{(2)}K)\mathbf{D}_d(\mathbf{D}_d^{\mathrm{T}}\mathbf{D}_d)^{-2}\mathbf{D}_d^{\mathrm{T}}) \\ &\quad + \frac{1}{4}\mu_2(K)^2(\operatorname{vech}^{\mathrm{T}}\mathbf{H})\boldsymbol{\Psi}_8(\operatorname{vech}\mathbf{H}).\end{aligned}$$

Differentiating this with respect to $\operatorname{vech}\mathbf{H}$, we obtain

$$\begin{aligned}&\nabla_{\mathbf{H}}\operatorname{TAMISE}^{(2)}(\mathbf{H}) \\ &\quad = -\frac{1}{2}n^{-1}|\mathbf{H}|^{-1/2}\Big\{\operatorname{tr}[(\mathbf{H}^{-1}\otimes\mathbf{H}^{-1})\mathbf{D}_d(\mathbf{D}_d^{\mathrm{T}}\mathbf{D}_d)^{-2}\mathbf{D}_d^{\mathrm{T}}\mathbf{R}(\operatorname{vec}\nabla^{(2)}K)]\mathbf{D}_d^{\mathrm{T}}(\operatorname{vec}\mathbf{H}^{-1}) \\ &\qquad - 2\mathbf{D}_d^{\mathrm{T}}(\mathbf{H}^{-1}\otimes\mathbf{H}^{-1})\mathbf{D}_d(\mathbf{D}_d^{\mathrm{T}}\mathbf{D}_d)^{-2}\mathbf{D}_d^{\mathrm{T}}\mathbf{R}(\operatorname{vec}\nabla^{(2)}K)(\operatorname{vec}\mathbf{H}^{-1})\Big\} + \frac{1}{2}\mu_2(K)^2\boldsymbol{\Psi}_8(\operatorname{vech}\mathbf{H})\end{aligned}$$

since we have for a $d^2 \times d^2$ matrix $\mathbf{A}$,

$$\begin{aligned}\nabla_{\mathbf{H}}\operatorname{tr}((\mathbf{H}^{-1}\otimes\mathbf{H}^{-1})\mathbf{A}) &= -\mathbf{D}_d^{\mathrm{T}}(\mathbf{H}^{-1}\otimes\mathbf{H}^{-1})(\mathbf{I}_{d^2}\otimes\operatorname{vec}\mathbf{H}^{-1} + \operatorname{vec}\mathbf{H}^{-1}\otimes\mathbf{I}_{d^2})\operatorname{vec}\mathbf{A} \\ &= -2\mathbf{D}_d^{\mathrm{T}}(\mathbf{H}^{-1}\otimes\mathbf{H}^{-1})\mathbf{A}\operatorname{vec}\mathbf{H}^{-1}.\end{aligned}$$

Let $\mathbf{H}_T^{(2)} = O(h^2)\mathbf{J}$ be a solution to $\nabla_{\mathbf{H}}\operatorname{TAMISE}^{(2)}(\mathbf{H}) = \mathbf{0}$ then $\mathbf{H}_T^{(2)} = O(n^{-2/(d+8)})\mathbf{J}$. □

References

Adler, D., Murdoch, D., 2006. rgl: 3D visualization device system (OpenGL). R package version 0.67–2.
Bowman, A.W., Azzalini, A., 1997. Applied Smoothing Techniques for Data Analysis. Oxford University Press, Oxford.
Chaudhuri, P., Marron, J.S., 1999. SiZer for exploration of structures in curves. Journal of the American Statistical Association 94, 807–823.
Chaudhuri, P., Marron, J.S., 2000. Scale space view of curve estimation. The Annals of Statistics 28, 408–428.
Duong, T., Hazelton, M.L., 2003. Plug-in bandwidth matrices for bivariate kernel density estimation. Journal of Nonparametric Statistics 15, 17–30.
Duong, T., Hazelton, M.L., 2005. Cross-validation bandwidth matrices for multivariate kernel density estimation. Scandinavian Journal of Statistics 32, 485–506.

Feng, D., Tierney, L., 2005. misc3d: Miscellaneous 3D Plots. R package version 0.3–1.
Givan, A.L., 2001. Flow Cytometry: First Principles, 2nd ed. Wiley-Liss, New York.
Godtliebsen, F., Marron, J.S., Chaudhuri, P., 2002. Significance in scale space for bivariate density estimation. Journal of Computational and Graphical Statistics 11, 1–21.
Hannig, J., Marron, J.S., 2006. Advanced distribution theory for SiZer. Journal of the American Statistical Association 101, 484–499.
Härdle, W., Marron, J.S., Wand, M.P., 1990. Bandwidth choice for density derivatives. Journal of the Royal Statistical Society. Series B 52, 223–232.
Hochberg, Y., 1988. A sharper Bonferroni procedure for multiple tests of significance. Biometrika 75, 800–802.
Magnus, J.R., Neudecker, H., 1999. Matrix Differential Calculus with Applications in Statistics and Econometrics: Revised edition. John Wiley & Sons, Chichester.
R Development Core Team, 2005. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.
Rossini, A., Wan, J., Moodie, Z., 2005. rflowcyt: Statistical tools and data structures for analytic flow cytometry. R package version 1.0.1.
Schott, J.R., 1996. Matrix Analysis for Statistics. John Wiley & Sons Inc., New York.
Scott, D.W., 1992. Multivariate Density Estimation: Theory, Practice, and Visualization. John Wiley & Sons Inc., New York.
Shaffer, J.P., 1995. Multiple hypothesis testing. Annual Review of Psychology 46, 561–584.
Simes, R.J., 1986. An improved Bonferroni procedure for multiple tests of significance. Biometrika 73, 751–754.
Simonoff, J.S., 1996. Smoothing Methods in Statistics. Springer-Verlag, New York.
Singh, R.S., 1987. MISE of kernel estimates of a density and its derivatives. Statistics & Probability Letters 5, 153–159.
Stone, C.J., 1980. Optimal convergence rates for nonparametric estimators. The Annals of Statistics 8, 1348–1360.
Wand, M.P., Jones, M.C., 1994. Multivariate plug-in bandwidth selection. Computational Statistics 9, 97–116.
Wand, M.P., Jones, M.C., 1995. Kernel Smoothing. Chapman and Hall Ltd., London.
Wu, T.-J., 1997. Root n bandwidth selectors for kernel estimation of density derivatives. Journal of the American Statistical Association 92, 536–547.