

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

6 June 2007

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. (2002). The theoretical framework is complemented by novel visualisation 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.

Keywords: Curvature estimators, density derivatives, feature significance testing, gradient estimators, modal regions.

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 visualising these features. Hannig and Marron (2006) produced distribution theoretic 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 visualisation 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.

Kernel density estimation has been a popular technique for analysing one and two-dimensional data; see Bowman and Azzalini (1997), Scott (1992), Simonoff (1996), Wand and Jones (1995) for examples. 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 all dimensions, thus reducing the two-dimensional problem to one dimension. For the single diagonal bandwidth one 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 visualisation tools. Finally the Appendix contains proofs of the theoretical results in Section 2.

2 Kernel density derivative estimation

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

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

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

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

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

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

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

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

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

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

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

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

Similarly the kernel curvature estimator

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

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

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

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 and begin with the asymptotic distribution of $\widehat{\nabla} f$ of (2).

Result 1. *The approximate distribution of the kernel gradient estimator $\widehat{\nabla} f(\mathbf{x}; \mathbf{H})$ as $n \rightarrow \infty$ is*

$$\widehat{\nabla} f(\mathbf{x}; \mathbf{H}) \overset{\text{approx.}}{\sim} \mathcal{N}\left(\nabla f(\mathbf{x}), \Sigma^{(1)}(\mathbf{x})\right)$$

where

$$\Sigma^{(1)}(\mathbf{x}) = n^{-1} |\mathbf{H}|^{-1/2} \mathbf{H}^{-1/2} \mathbf{R}(\nabla K) \mathbf{H}^{-1/2} f(\mathbf{x}), \quad (4)$$

and for any square integrable vector-valued function $\mathbf{g}(\cdot)$,

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

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

$$\Sigma^{(1)}(\mathbf{x}) = \frac{1}{2}(2\pi^{1/2})^{-d}n^{-1}(h_1 \dots h_d)^{-1} \text{diag}(h_1^{-2}, \dots, h_d^{-2})f(\mathbf{x}).$$

The asymptotic means and variances in Result 1 follow from Lemmas 1 to 3. Although not proved here, the asymptotic normality of the estimator follows since the kernel gradient estimator is a sample mean and an appropriate central limit theorem therefore yields the desired result.

The special case of the diagonal bandwidth matrix is included as it explicitly shows the dependence of the variance on the density and the bandwidths 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 than on matrices is easier, and for this reason it is natural to consider matrix norms based on MISE. Since the p -norms ($p = 1, 2, \dots$) 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

$$\begin{aligned} \text{TAMISE}^{(1)}(\mathbf{H}) &\equiv \text{TAMISE} \widehat{\nabla} f(\cdot; \mathbf{H}) \\ &= n^{-1}|\mathbf{H}|^{-1/2} \text{tr}(\mathbf{H}^{-1}\mathbf{R}(\nabla K)) + \frac{1}{4}\mu_2(K)^2(\text{vech}^T \mathbf{H})\Psi_6(\text{vech} \mathbf{H}). \end{aligned} \quad (5)$$

The first summand is derived from the variance, and the second from the squared bias. We have used the following notation: tr denotes the trace of a matrix, $\mu_2(K)\mathbf{I} = \int \mathbf{x}\mathbf{x}^T K(\mathbf{x})$ with \mathbf{I} the $d \times d$ identity matrix; vech is the vector half operator which transforms a symmetric $d \times d$ matrix 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 symmetric matrix this operation results in

$$\text{vech} \begin{bmatrix} a & b & c \\ b & d & e \\ c & e & f \end{bmatrix} = [a \ b \ c \ d \ e \ f]^T, \quad (6)$$

the case for general d follows the same pattern. The term Ψ_6 involves higher order derivatives of f , and its subscript, here 6, indicates the order of derivatives used. It is a symmetric matrix with d^* rows. We defer its definition until the proof of Theorem 1. With this Ψ_6 matrix, the second term of (5) is a quadratic form in $\text{vech} \mathbf{H}$, which is easy to differentiate with respect to $\text{vech} \mathbf{H}$.

Theorem 1. *Let $\mathbf{H}_T^{(1)}$ be the bandwidth matrix which minimises 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 1 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 (6). We present our results for the curvature estimator in this latter form and begin with its distribution.

Result 2. *The approximate distribution of the kernel curvature estimator $\text{vech} \widehat{\nabla^{(2)}} f(\mathbf{x}; \mathbf{H})$ as $n \rightarrow \infty$ is*

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

where

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

For the normal kernel K the variance reduces to

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

where \mathbf{D}_d is the duplication matrix of order d , see Magnus and Neudecker (1999); and $\mathbf{A} \otimes \mathbf{B}$ denotes the tensor product of the matrices \mathbf{A} and \mathbf{B} .

The kernel curvature estimator is a sample mean and so by applying an appropriate central limit theorem the asymptotic normality follows.

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

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

where $\boldsymbol{\Psi}_8$ is the square matrix of size d^* which contains derivatives of order 8, and \mathbf{D}_d as in Result 2.

Theorem 2. *Let $\mathbf{H}_T^{(2)}$ be the bandwidth matrix which minimises 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 2 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. Although the distributional results of these 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 $\mathbf{x} \in \mathbb{R}^d$, a null hypothesis for gradient testing is

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

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

Consider a bandwidth matrix \mathbf{H} suitable for kernel density gradient estimators. From Result 1 the approximate asymptotic null distribution after normalisation is

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

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

$$W^{(1)}(\mathbf{x}) = \|\{\boldsymbol{\Sigma}^{(1)}(\mathbf{x})\}^{-1/2} \widehat{\nabla} f(\mathbf{x}; \mathbf{H})\|^2, \quad (8)$$

which requires an estimate for $\boldsymbol{\Sigma}^{(1)}$. A natural choice is to replace the true density f at \mathbf{x} in (4) by its kernel density estimator $\hat{f}(\mathbf{x}; \mathbf{H})$ and to use

$$\widehat{\boldsymbol{\Sigma}}^{(1)}(\mathbf{x}) = n^{-1} |\mathbf{H}|^{-1/2} \mathbf{H}^{-1/2} \mathbf{R}(\nabla K) \mathbf{H}^{-1/2} \hat{f}(\mathbf{x}; \mathbf{H}) \quad (9)$$

as an estimator for $\boldsymbol{\Sigma}^{(1)}$. This choice, which we make, has the added advantage that the theoretical properties of the estimator follow directly from Result 1. In contrast, for Godtliebsen et al. (2002)'s estimator

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

no such distributional results are available.

Since the scalar term of $\widehat{\boldsymbol{\Sigma}}^{(1)}(\mathbf{x})$ in (9), namely $n^{-1} |\mathbf{H}|^{-1/2} \hat{f}(\mathbf{x}; \mathbf{H})$, 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

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

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

Our aim is to test simultaneously at all points \mathbf{x} whether $\widehat{W}^{(1)}(\mathbf{x})$ shows a significant deviation from zero. Such tests are highly correlated since for nearby points \mathbf{x}_1 and \mathbf{x}_2 ,

$\widehat{W}^{(1)}(\mathbf{x}_1)$ and $\widehat{W}^{(1)}(\mathbf{x}_2)$ 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 $P_{(1)}, \dots, P_{(m)}$, corresponding to the null hypotheses $H_{0,(1)}, \dots, H_{0,(m)}$. If $P_{(j)} \leq \alpha/(m-j+1)$, then we reject all null hypotheses $H_{0,(1)}, \dots, H_{0,(j)}$. We find the largest such $j = j_{\max}$ and

$$\text{reject } H_{0,(1)}, \dots, H_{0,(j_{\max})} \text{ where } j_{\max} = \operatorname{argmax}_{1 \leq j \leq m} P_{(j)} \leq \alpha/(m-j+1). \quad (10)$$

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

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

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

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 hull hypothesis is

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

Consider a bandwidth matrix \mathbf{H} suitable for kernel density curvature estimators. From Result 2, the approximate asymptotic null distribution of $\text{vech } \widehat{\nabla^{(2)} f(\mathbf{x}; \mathbf{H})}$ after normalising is

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

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

$$W^{(2)}(\mathbf{x}) = \|\{\boldsymbol{\Sigma}^{(2)}(\mathbf{x})\}^{-1/2} \text{vech } \widehat{\nabla^{(2)} f(\mathbf{x}; \mathbf{H})}\|^2. \quad (11)$$

This choice of test statistic is markedly different from those of Godtliebsen et al. (2002). These authors used as their test statistic $\max\{|\lambda_1(\mathbf{x})|, \dots, |\lambda_d(\mathbf{x})|\}$ where $\lambda_j(\mathbf{x})$ is the j^{th} eigenvalue of the normalised $\widehat{\nabla^{(2)} f(\mathbf{x}; \mathbf{H})}$. They wished to characterise 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 characterise 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 Godtlielsen et al. (2002)'s test statistic, since $W^{(2)}(\mathbf{x})$ of (11) has an approximate closed form null distribution.

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

$$\widehat{W}^{(2)}(\mathbf{x}) = \|\{\widehat{\Sigma}^{(2)}(\mathbf{x})\}^{-1/2} \text{vech} \widehat{\nabla^{(2)}} f(\mathbf{x}; \mathbf{H})\|^2 \stackrel{\text{approx.}}{\sim} \chi_{d^*}^2$$

where $\widehat{\Sigma}^{(2)}(\mathbf{x}) = n^{-1} |\mathbf{H}|^{-1/2} \mathbf{R}(\text{vech}(\mathbf{H}^{-1/2} \nabla^{(2)} K \mathbf{H}^{-1/2})) \hat{f}(\mathbf{x}; \mathbf{H})$. This estimator of $\Sigma^{(2)}(\mathbf{x})$ is derived from Result 2 analogously to the variance estimator in (9), and is an alternative to the estimator of Godtlielsen et al. (2002). The latter estimator relies on individually estimating the elements of $\Sigma^{(2)}(\mathbf{x})$ 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 $\Sigma^{(2)}(\mathbf{x})$ 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 $\text{vech} \widehat{\nabla^{(2)}} f(\mathbf{x}; \mathbf{H})$. Godtlielsen et al. (2002)'s approach also relies on using the most restricted parameterisation $\mathbf{H} = h^2 \mathbf{I}$ whereas our method allows more general bandwidth matrices.

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

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 test '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 the 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. One 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 the smoothing parameters could 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), one needs 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 focussed 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 parametrisation 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 parametrisation of a plug-in selector is an obvious candidate. To the authors’ knowledge no such method is 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.

From our Theorems 1 and 2 we find that the bandwidth matrices which minimise TAMISE are of order $n^{-2/(d+6)}$ for the gradient and of order $n^{-2/(d+8)}$ for the curvature estimator. We adjust the bandwidth matrices by taking into account the marginal sample variances S_i and the sample size n . Starting with a kernel density bandwidth matrix $\mathbf{H} = \text{diag}(h_1^2, \dots, h_d^2)$, we obtain the gradient bandwidth matrix

$$\mathbf{H}^{(1)} = \text{diag}(h_{g,1}^2, \dots, h_{g,d}^2) \quad \text{with} \quad h_{g,i}^2 = h_i^2 \frac{(nS_i)^{-2/(d+6)}}{(nS_i)^{-2/(d+4)}}, \quad (12)$$

and similarly the curvature bandwidth matrix

$$\mathbf{H}^{(2)} = \text{diag}(h_{c,1}^2, \dots, h_{c,d}^2) \quad \text{with} \quad h_{c,i}^2 = h_i^2 \frac{(nS_i)^{-2/(d+8)}}{(nS_i)^{-2/(d+4)}}. \quad (13)$$

These rules represent a simplistic way of choosing diagonal bandwidth matrices. Other more rigorous choices could be explored.

4 Real data examples

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

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 of 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 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.

Figure 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 Figure 1. This red cluster represents the points inside a hyperrectangular gate obtained from a flow cytometry scientist for these data. We compare these gated cells 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 parameterisation of the plug-in selector $\mathbf{H} = \text{diag}(11.3^2, 6.26^2, 10.0^2)$ of Duong and Hazelton (2003). We apply the transformations given in (12) and (13) to \mathbf{H} and display the significant gradient regions in green and the significant curvature regions in blue in Figure 2. Here we have used the gradient bandwidth matrix $\mathbf{H}^{(1)} = \text{diag}(19.48^2, 10.45^2, 17.13^2)$ on the left and the curvature bandwidth matrix $\mathbf{H}^{(2)} = \text{diag}(27.53^2, 12.48^2, 24.13^2)$ on the right. Overlaid in red are the gated points from Figure 1. On the left, the red gated points intersect with one significant gradient region, and the right, they contain one significant curvature region and intersects 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 in darker green 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

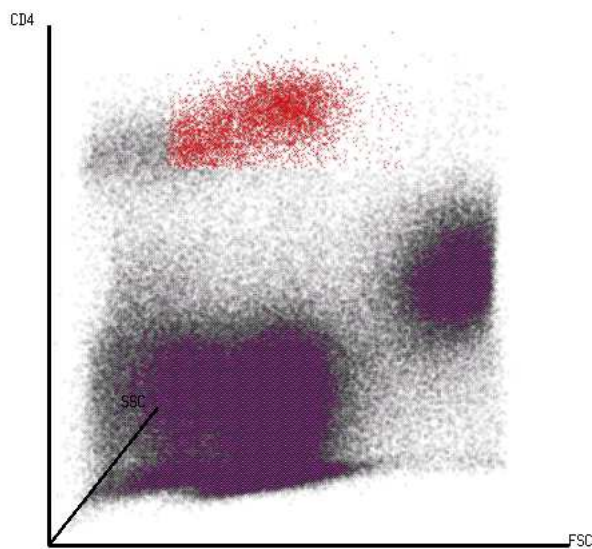


Figure 1: Scatterplot of three-dimensional flow cytometric data set with subjectively gated lymphocytes (red)

not change much. In the next part we will therefore focus on curvature based tests only.

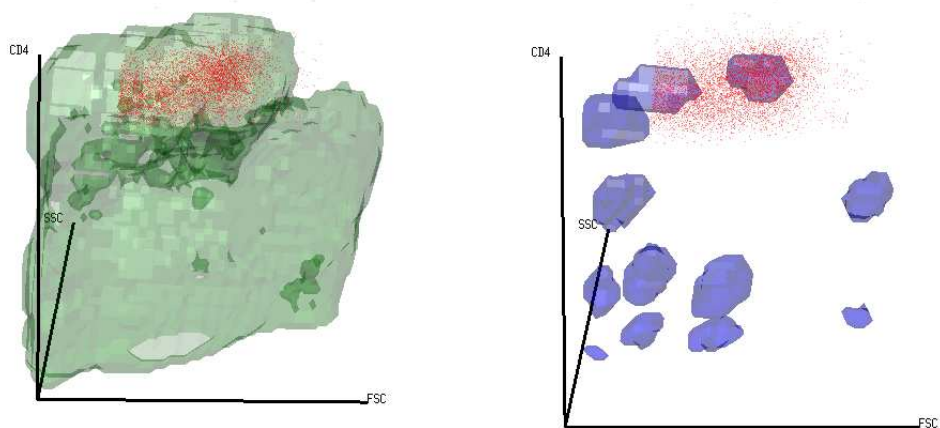


Figure 2: Significant gradient (green) and curvature regions (blue) with subjectively gated lymphocytes (red) and adjusted plug-in bandwidth matrices $\mathbf{H}^{(1)}$ for the gradient and $\mathbf{H}^{(2)}$ for the curvature tests.

In their two-dimensional scale space approach Godtlielsen 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 $\mathbf{H} = \text{diag}(11.3^2, 6.26^2, 10.0^2)$, namely 0.8, 1.1, 1.5 and 2. The resulting significant curvature regions are shown in blue in Figure 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 Figure 2 – in the bottom left figure – significant feature regions appear again in the region of the red gated points, and as the bandwidths increase further, the number of significant regions decreases whilst their individual volumes increase. The sequence of bandwidths plots confirms the adequacy of our adjusted plug-in bandwidth matrix $\mathbf{H}^{(2)}$ when a single bandwidth matrix is required.

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 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 Figure 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 Figure 4 suggests that there are a number of modal and anti-modal regions which are marked by the darker green 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 $\mathbf{H} = \text{diag}(0.222^2, 0.216^2, 0.196^2)$ 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 in blue

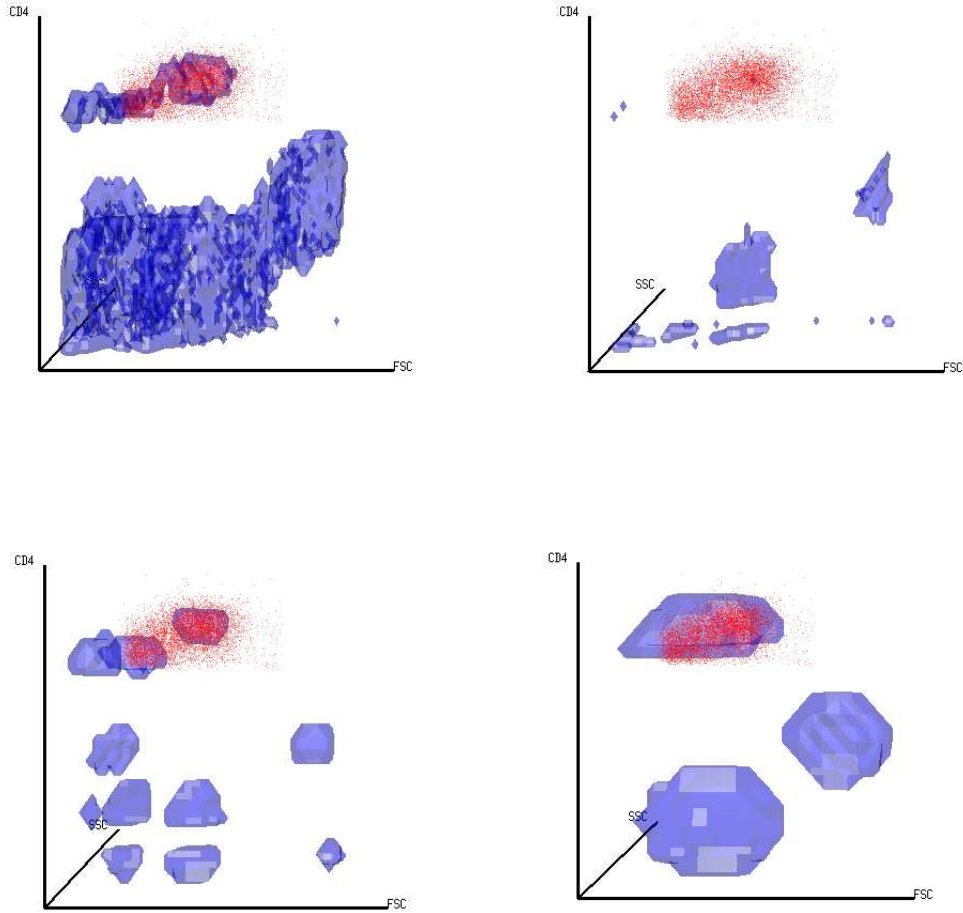


Figure 3: Significant curvature regions for different bandwidth matrices: Scale space for flow cytometry data – significant curvature regions (blue). Upper left – $\mathbf{H}^{0.8}$ (top left), $\mathbf{H}^{1.1}$ (top right), $\mathbf{H}^{1.5}$ (bottom left) and \mathbf{H}^2 (bottom right).

on the green significant gradient regions. The results are shown in Figure 5. For the bandwidths used in Figure 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.

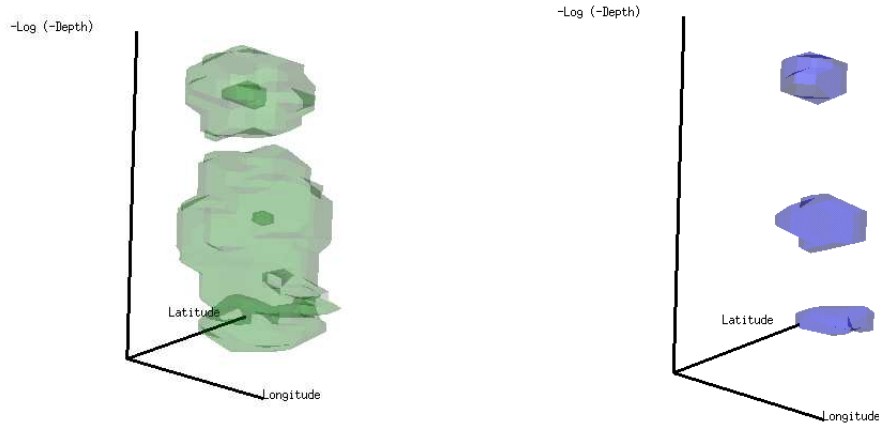


Figure 4: Significant gradient (green) and curvature (blue) regions for Mt St Helens earthquake data. Left with bandwidths $\mathbf{H}^{(1)}$, right with $\mathbf{H}^{(2)}$.

5 Software

Accompanying this paper is an R library, named `feature` and 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 bandwidth and visualise 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 and Murdoch, 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 visualisation 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

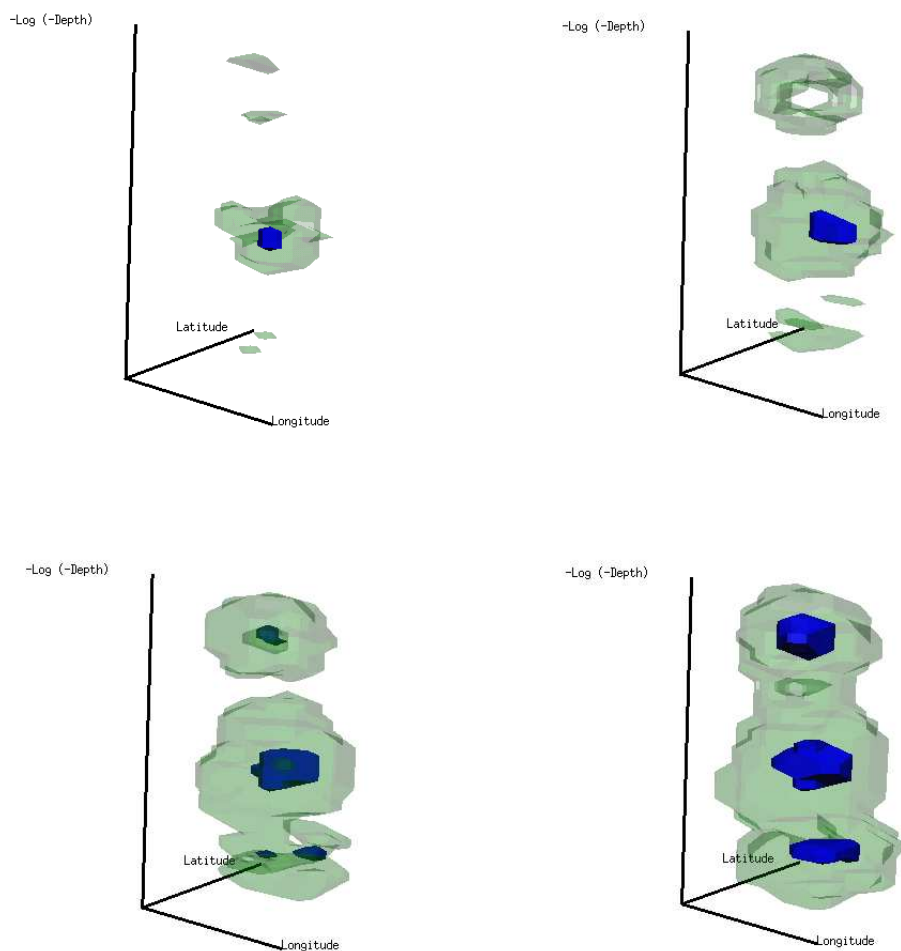


Figure 5: Significant gradient (in green) and curvature regions (in blue) for different bandwidth matrices: $\mathbf{H}^{1.1}$ (top left), \mathbf{H} (top right), $\mathbf{H}^{0.9}$ (bottom left) and $\mathbf{H}^{0.8}$ (bottom right).

based tests of significant gradient regions can enhance the understanding of significant structure in data.

7 Acknowledgements

The authors acknowledge the financial support of an Australian Research Council (Project DP0556518).

A Proofs

A.1 Kernel gradient estimation

Assume the following conditions hold.

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

The proof of Theorem 1 requires the definition of the term

$$\Psi_6 = \int_{\mathbb{R}^d} \mathbf{D}_d^T (\nabla \otimes \nabla^{(2)})f(\mathbf{x})(\nabla^T \otimes \nabla^{(2)})f(\mathbf{x})\mathbf{D}_d d\mathbf{x} \quad (14)$$

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

$$\text{vec} \begin{bmatrix} a & b & c \\ b & d & e \\ c & e & f \end{bmatrix} = [a \ b \ c \ b \ d \ e \ c \ e \ f]^T.$$

For details see Magnus and Neudecker (1999). The subscript of Ψ_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 a matrix of third order differentials.

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

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

If we take the trace of this and integrate then

$$\text{TAMISE}^{(1)}(\mathbf{H}) = n^{-1}|\mathbf{H}|^{-1/2} \text{tr}(\mathbf{H}^{-1}\mathbf{R}(\nabla K)) + \frac{1}{4}\mu_2(K)^2(\text{vech}^T \mathbf{H})\Psi_6(\text{vech} \mathbf{H}).$$

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

$$\begin{aligned}\nabla_{\mathbf{H}} \text{TAMISE}^{(1)}(\mathbf{H}) &\equiv \frac{\partial \text{TAMISE}^{(1)}(\mathbf{H})}{\partial \text{vech } \mathbf{H}} \\ &= -\frac{1}{2}n^{-1}|\mathbf{H}|^{-1/2} \left[\text{tr}(\mathbf{H}^{-1}\mathbf{R}(\nabla K))\mathbf{D}_d^T(\text{vec } \mathbf{H}^{-1}) \right. \\ &\quad \left. - 2\mathbf{D}_d^T(\mathbf{H}^{-1} \otimes \mathbf{H}^{-1}) \text{vec } \mathbf{R}(\nabla K) \right] + \frac{1}{2}\mu_2(K)^2\Psi_6(\text{vech } \mathbf{H})\end{aligned}$$

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

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

and

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

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 $\nabla_{\mathbf{H}} \text{TAMISE}^{(1)}(\mathbf{H}) = \mathbf{0}$ 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 . \square

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

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

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

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

So then

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

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

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

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

Proof. The variance is

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

We have

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

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

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

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

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

Substitute this and $\mathbf{H} = \text{diag}(h_1^2, \dots, h_d^2)$ into $\Sigma^{(1)}(\mathbf{x}) = n^{-1} |\mathbf{H}|^{-1/2} \mathbf{H}^{-1/2} \mathbf{R}(\nabla K) \mathbf{H}^{-1/2} f(\mathbf{x})$. \square

A.2 Kernel curvature estimation

Assume the following conditions hold.

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

(B2) All entries of $\nabla^{(2)} \otimes \nabla^{(2)} f(\mathbf{x})$ are bounded, continuous and square integrable.

(B3) The kernel K is a symmetric probability density function and $\int_{\mathbb{R}^d} K(\mathbf{x}) \mathbf{x} \mathbf{x}^T d\mathbf{x} = \mu_2(K) \mathbf{I}$ with $\mu_2(K) < \infty$. This is the same as (A3).

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

Let $\mathbf{D}_d^+ = (\mathbf{D}_d^T \mathbf{D}_d)^{-1} \mathbf{D}_d^T$ be the unique Moore-Penrose inverse of \mathbf{D}_d , then $\text{vech } \mathbf{H} = \mathbf{D}_d^+ \text{vec } \mathbf{H}$, and $(\mathbf{D}_d^+)^T \mathbf{D}_d^+ = \mathbf{D}_d (\mathbf{D}_d^T \mathbf{D}_d)^{-2} \mathbf{D}_d^T$. See Magnus and Neudecker (1999, p. 49).

From TAMISE for curvature

$$\Psi_8 = \int_{\mathbb{R}^d} \mathbf{D}_d^T (\nabla^{(2)} \otimes \nabla^{(2)}) f(\mathbf{x}) \mathbf{D}_d (\mathbf{D}_d^T \mathbf{D}_d)^{-2} \mathbf{D}_d^T (\nabla^{(2)} \otimes \nabla^{(2)}) f(\mathbf{x}) \mathbf{D}_d d\mathbf{x}. \quad (15)$$

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

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

If we take the trace of this and integrate then

$$\begin{aligned} \text{TAMISE}^{(2)}(\mathbf{H}) &= n^{-1} |\mathbf{H}|^{-1/2} \text{tr}((\mathbf{H}^{-1} \otimes \mathbf{H}^{-1}) \mathbf{R}(\text{vec } \nabla^{(2)} K) \mathbf{D}_d (\mathbf{D}_d^T \mathbf{D}_d)^{-2} \mathbf{D}_d^T) \\ &\quad + \frac{1}{4} \mu_2(K)^2 (\text{vec}^T \mathbf{H}) \Psi_8(\text{vech } \mathbf{H}). \end{aligned}$$

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

$$\begin{aligned} \nabla_{\mathbf{H}} \text{TAMISE}^{(2)}(\mathbf{H}) &= -\frac{1}{2} n^{-1} |\mathbf{H}|^{-1/2} \left\{ \text{tr} [(\mathbf{H}^{-1} \otimes \mathbf{H}^{-1}) \mathbf{D}_d (\mathbf{D}_d^T \mathbf{D}_d)^{-2} \mathbf{D}_d^T \mathbf{R}(\text{vec } \nabla^{(2)} K)] \mathbf{D}_d^T (\text{vec } \mathbf{H}^{-1}) \right. \\ &\quad \left. - 2 \mathbf{D}_d^T (\mathbf{H}^{-1} \otimes \mathbf{H}^{-1}) \mathbf{D}_d (\mathbf{D}_d^T \mathbf{D}_d)^{-2} \mathbf{D}_d^T \mathbf{R}(\text{vec } \nabla^{(2)} K) (\text{vec } \mathbf{H}^{-1}) \right\} \\ &\quad + \frac{1}{2} \mu_2(K)^2 \Psi_8(\text{vech } \mathbf{H}) \end{aligned}$$

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

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

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

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

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

Proof. The expected value is

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

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

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

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

Proof. We have

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

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

Lemma 6. *Assume that the conditions (B1) – (B2) hold. Further assume that K is the normal kernel then $\boldsymbol{\Sigma}^{(2)}(\mathbf{x}) = \frac{1}{4} (4\pi)^{-d/2} |\mathbf{H}|^{-1/2} [2(\mathbf{D}_d^T \mathbf{D}_d)^{-1} \mathbf{D}_d^T (\mathbf{H}^{-1} \otimes \mathbf{H}^{-1}) \mathbf{D}_d (\mathbf{D}_d^T \mathbf{D}_d)^{-1} + (\text{vech} \mathbf{H}^{-1})(\text{vech}^T \mathbf{H}^{-1})] f(\mathbf{x})$.*

Proof. An alternative expression for the variance of the curvature estimator is

$$\begin{aligned}
\text{Var} \text{vech} \widehat{\nabla^{(2)}} f(\mathbf{x}; \mathbf{H}) &= |\mathbf{H}|^{-1/2} \mathbf{D}_d^+ (\mathbf{H}^{-1/2} \otimes \mathbf{H}^{-1/2}) \mathbf{R}(\text{vec} \nabla^{(2)} K) (\mathbf{H}^{-1/2} \otimes \mathbf{H}^{-1/2}) (\mathbf{D}_d^+)^T f(\mathbf{x}) \\
&\quad + O(|\mathbf{H}|^{-1/2} \|\text{vech} \mathbf{H}^{-1}\|).
\end{aligned}$$

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

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

where \mathbf{N}_d is a $d^2 \times d^2$ symmetric 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

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

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

References

- Adler, D. and Murdoch, D. (2006). *rgl: 3D visualization device system (OpenGL)*. R package version 0.67-2.
- Bowman, A. W. and Azzalini, A. (1997). *Applied Smoothing Techniques for Data Analysis*. Oxford University Press, Oxford.
- Chaudhuri, P. and Marron, J. S. (1999). SiZer for exploration of structures in curves. *Journal of the American Statistical Association*, 94:807–823.
- Duong, T. and Hazelton, M. L. (2003). Plug-in bandwidth matrices for bivariate kernel density estimation. *Journal of Nonparametric Statistics*, 15:17–30.
- Duong, T. and Hazelton, M. L. (2005). Cross-validation bandwidth matrices for multivariate kernel density estimation. *Scandinavian Journal of Statistics*, 32:485–506.
- Feng, D. and Tierney, L. (2005). *misc3d: Miscellaneous 3D Plots*. R package version 0.3-1.
- Givan, A. L. (2001). *Flow Cytometry: First Principles*. Wiley-Liss, New York, 2nd edition.
- 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.
- Hannig, J. and Marron, J. S. (2006). Advanced distribution theory for SiZer. *Journal of the American Statistical Association*, 101:484–499.

- Härdle, W., Marron, J. S., and 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. and 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., and Moodie, Z. (2005). *rflowcyt: Statistical tools and data structures for analytic flow cytometry*. R package version 1.0.1.
- Ruppert, D. (1997). Empirical-bias bandwidths for local polynomial nonparametric regression and density estimation. *Journal of the American Statistical Association*, 92.
- 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*, 46:561–84.
- 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. and Jones, M. C. (1994). Multivariate plug-in bandwidth selection. *Computational Statistics*, 9:97–116.
- Wand, M. P. and Jones, M. C. (1995). *Kernel Smoothing*. Chapman and Hall Ltd., London.
- Waterloo Maple Inc. (2004). *Maple*. Waterloo, Canada.
- Wu, T.-J. (1997). Root n bandwidth selectors for kernel estimation of density derivatives. *Journal of the American Statistical Association*, 92:536–547.