Aust. N. Z. J. Stat. **46**(4), 2004, 569–581

COMPARISON OF FEATURE SIGNIFICANCE QUANTILE APPROXIMATIONS

M.P. SALGANIK[1], M.P. WAND[1,2*] AND N. LANGE[1,3]

Harvard University, The University of New South Wales and McLean Hospital

Summary

Curve estimates and surface estimates often contain features such as inclines, bumps or ridges which may signify an underlying structural mechanism. However, spurious features are also a common occurrence and it is important to identify those features that are statistically significant. A method has been developed recently for recognising feature significance based on the derivatives of the function estimate. It requires simultaneous confidence intervals and tests, which in turn require quantiles for the maximal deviation statistics. This paper reviews and compares various approximations to these quantiles. Applying upcrossing-probability theory to this problem yields better quantile approximations than the use of an independent blocks method.

Key words: derivative estimation; non-parametric regression; simultaneous confidence band; SiZer; upcrossing probability.

1. Introduction

In many areas of research, data arrive as noisy curves or surfaces. Features such as peaks and ridges often signify the presence of an underlying mechanism of interest. Examples occur in our own collaborative research using flow cytometry, imaging neuroscience (Lange, 1999, 2004; Wager, Coull & Lange, 2004), and spatial epidemiology where geographical 'hot spots' of particular diseases are of interest (Ganguli & Wand, 2004). Flow cytometry studies the distributions of chemical and physical characteristics in the population of cells or other biological particles. Positron emission tomography (PET) and functional magnetic resonance imaging (FMRI) of the human brain measure local responses to controlled stimuli *in vivo*: PET records changes in cerebral blood flow and volume, and FMRI records blood oxygenation when subjects perform cognitive and physiological tasks. Statistical analysis is often needed in two intersecting components of this research: (1) the estimation of correlations between external stimuli and the FMRI time series at each image volume element; and (2) the detection of brain regions activated by the external stimulus, and contrasts or interactions between stimuli. For activation detection subsequent to time series summarization, one may embed the results from the first component within an appropriate random field and then use

Received July 2003; revised November 2003; accepted January 2004.

*Author to whom correspondence should be addressed.

[1] Dept of Biostatistics, School of Public Health, Harvard University, 665 Huntington Avenue, Boston, MA 02115, USA.

[2] Dept of Statistics, School of Mathematics, The University of New South Wales, Sydney NSW 2052, Australia. e-mail: wand@maths.unsw.edu.au

[3] Dept of Psychiatry, Harvard Medical School, 25 Shattuck Street, Boston, MA 02115, USA and Laboratory for Statistical Neuroimaging, McLean Hospital, 115 Mill Street, Belmont, MA 02478, USA.

Acknowledgments. The authors thank Bhaswati Ganguli for programming assistance and Keith Worsley for helpful comments. This research was supported by NIH grants NS37483 and MH60450 to NL.

© 2004 Australian Statistical Publishing Association Inc. Published by Blackwell Publishing Asia Pty Ltd.

thresholding methods from random field and scale–space theory (Adler, 1981; Worsley *et al.*, 1992; Worsley, 1994) to identify significant spatial peaks or clusters of brain activity.

Methods aimed at identifying features that are statistically significant, rather than mere aberrations, have developed greatly in recent years. Chaudhuri & Marron (1999) and Loader (1999) describe methods for assessing the significance of features in curves (a response variable and a univariate predictor variable). Godtliebsen, Marron & Chaudhuri (2002, 2004) and Ganguli & Wand (2004) treat surfaces (a response variable and a bivariate predictor variable) in the image, density estimation and geostatistical contexts respectively.

An important component of feature significance research is the approximation of quantiles for maximum deviation statistics. Such approximations fall into two broad categories:

- analytic; based on asymptotic or inequality arguments but prone to inaccuracy;
- simulation-based; can be made quite accurate but are much more computationally expensive.

Analytic methods used in feature significance include Bonferroni-type adjustment based on independent blocks (e.g. Härdle & Marron, 1991) and upcrossing theory as reviewed by Loader (1999).

In this paper we review and compare these analytic approximations and benchmark them against simulation-based approximations. We work at a more general level than Chaudhuri & Marron (1999). Those authors concentrated on kernel-type smoothers whereas we treat general linear smoothers — a class which also includes smoothing splines (e.g. Wahba, 1990; Green & Silverman, 1994) and penalized splines (e.g. Eilers & Marx, 1996). The analytic approximations used by Chaudhuri & Marron (1999, 2000) and Härdle & Marron (1991) are less useful than the upcrossing-probability-based approximations. Analytic theory for two-dimensional simultaneous confidence regions is not yet at the stage where it can be used for derivative-based feature significance. We identify some future research directions for quantile approximation that would be helpful for multiple degrees of freedom and analysing two-dimensional feature significance. Section 2 describes the fundamental components of feature significance. Sections 3 and 4 discuss various approaches to feature significance quantile approximation and Section 5 makes comparisons. Sections 6 and 7 make some notes on multiple degrees of freedom and two-dimensional feature significance. Section 8 gives closing discussion.

2. Elements of feature significance

2.1. Scale–space viewpoint

Consider the univariate scatterplot smoothing (or non-parametric regression) setting,

$$y_i = f(x_i) + \varepsilon_i \,,$$

where the (x_i, y_i), $1 \leq i \leq n$, are the scatterplot data, ε_i are zero mean random variables and $f(x) = \mathrm{E}(y \mid x)$ is a smooth function.

A smoother of (x_i, y_i) is a function $\hat{f}$ that summarizes the local mean structure of y as a function of x. Usually, $\hat{f}(x)$ is a member of a family of smoothers parametrized by smoothing parameters. Examples include smoothing splines (e.g. Wahba, 1990; Eubank, 1999) and local polynomials (e.g. Fan & Gijbels, 1995; Wand & Jones, 1995). In traditional non-parametric regression the target is f, and $\hat{f}$ is an estimate. However, from the scale–space viewpoint the

© 2004 Australian Statistical Publishing Association Inc.

target is $\mathrm{E}(\hat{f}(x))$, with $\hat{f}(x)$ possibly computed at several 'scales' or values of the smoothing parameter. Contemporary feature significance methods (e.g. Chaudhuri & Marron, 1999) use the scale–space viewpoint.

2.2. Linear smoothers

An estimate $\hat{f}(x)$ of $f(x)$ is a linear smoother if

$$\hat{f}(x) = \boldsymbol{\ell}_x^{\top} \boldsymbol{y} \tag{1}$$

for some $n \times 1$ vector $\boldsymbol{\ell}_x$, where $\boldsymbol{y} = (y_1, \ldots, y_n)$. Linear smoothers include kernel and local polynomial smoothers, smoothing splines, penalized splines and kriging. We can write the vector of fitted values $\hat{\boldsymbol{f}} = (\hat{f}(x_1), \ldots, \hat{f}(x_n))$ as

$$\hat{\boldsymbol{f}} = \boldsymbol{S}\boldsymbol{y}\,,$$

where $\boldsymbol{S}$ is the smoother matrix, given by

$$\boldsymbol{S} = \begin{bmatrix} \boldsymbol{\ell}_{x_1}^{\top} \\ \vdots \\ \boldsymbol{\ell}_{x_n}^{\top} \end{bmatrix}.$$

The number of degrees of freedom of the fit, df_{fit}, is given by $df_{\mathrm{fit}} = \mathrm{tr}(\boldsymbol{S})$. For most linear smoothers, including those described in Sections 2.2.1 and 2.2.2, df_{fit} is a monotone function of the smoother parameter (e.g. the bandwidth h in Section 2.2.1). We prefer to work with df_{fit} because it helps in interpreting the effective number of parameters and allows for comparison across smoothers.

In feature significance, estimates of the derivatives of $\hat{f}$ are a central tool. We denote by $\widehat{f^{(r)}}(x)$, the estimate of the rth derivative of $\hat{f}$ at x. Assuming existence, a natural candidate for $\widehat{f^{(r)}}(x)$ is

$$\widehat{f^{(r)}}(x) = \frac{d^r}{dx^r}\hat{f}(x)\,. \tag{2}$$

However, as demonstrated in Section 2.2.1, rth derivative estimators are not always of this form. Either way, derivative estimators based on splines and local polynomials are also linear smoothers and we define $\boldsymbol{\ell}_x^{(r)}$ by

$$\widehat{f^{(r)}}(x) = (\boldsymbol{\ell}_x^{(r)})^{\top}\boldsymbol{y}\,.$$

If (2) is satisfied, $\boldsymbol{\ell}_x^{(r)}$ is obtained through element-wise differentiation of $\boldsymbol{\ell}_x$ with respect to x.

The standard deviation of $\widehat{f^{(r)}}(x)$ is

$$\mathrm{sd}\big(\widehat{f^{(r)}}(x)\big) = \sqrt{(\boldsymbol{\ell}_x^{(r)})^{\top} V(\boldsymbol{y}) \boldsymbol{\ell}_x^{(r)}}\,,$$

where $V(\boldsymbol{y})$ is the variance–covariance matrix of $\boldsymbol{y}$. Assuming $V(\boldsymbol{y}) = \sigma_\varepsilon^2 \boldsymbol{I}$ this simplifies to

$$\mathrm{sd}\big(\widehat{f^{(r)}}(x)\big) = \sigma_\varepsilon \|\boldsymbol{\ell}_x^{(r)}\|\,, \tag{3}$$

where $\|\boldsymbol{v}\| = \sqrt{\boldsymbol{v}^{\top}\boldsymbol{v}}$ is the length of $\boldsymbol{v}$.

© 2004 Australian Statistical Publishing Association Inc.

2.2.1. *Local polynomial smoothers*

The local polynomial estimate of $f^{(r)}(x)$ is obtained by fitting a local polynomial of degree $p \geq r$,

$$\beta_0 + \beta_1(\cdot - x) + \cdots + \beta_p(\cdot - x)^p \,,$$

to the (x_i, y_i), using weighted least squares with kernel weights $K((x_i - x)/h)$, $1 \leq i \leq n$ (e.g. Wand & Jones, 1995). The kernel function K is usually chosen to be a unimodal probability density function that is symmetric about 0, and $h > 0$ is a parameter controlling the amount of smoothing, usually referred to as the bandwidth. The estimate $\widehat{f^{(r)}}(x)$ is $r!\hat{\beta}_r$ where $\hat{\beta}_r$ is the weighted least squares estimate of β_r. Note that

$$\boldsymbol{\ell}_x^{(r)} = \boldsymbol{W}_x \boldsymbol{X}_x (\boldsymbol{X}_x^T \boldsymbol{W}_x \boldsymbol{X}_x)^{-1} \boldsymbol{e}_{r+1} \,,$$

where

$$\boldsymbol{X}_x = \begin{bmatrix} 1 & x_1 - x & \ldots & (x_1 - x)^p \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_n - x & \ldots & (x_n - x)^p \end{bmatrix}, \qquad \boldsymbol{W}_x = \operatorname{diag}\left(K\left(\frac{x_i - x}{h}\right)\right)$$

and $\boldsymbol{e}_{r+1}$ is the $(p+1) \times 1$ vector having 1 in the $(r+1)$th entry and 0s elsewhere.

2.2.2. *Penalized spline smoothers*

Penalized spline smoothers (e.g. Eilers & Marx, 1996; Ruppert & Carroll, 2000) come in a number of forms. Here we work with radial basis smoothers due to their particularly attractive extension to higher dimensions (French, Kammann & Wand, 2001). The resulting linear smoother can be shown to be

$$\hat{f} = \boldsymbol{C}(\boldsymbol{C}^\top \boldsymbol{C} + \lambda^{2m-1} \boldsymbol{D})^{-1} \boldsymbol{C}^\top \boldsymbol{y} \,,$$

where λ is a smoothing parameter,

$$\boldsymbol{C} = [1 \;\; x_i \;\; \ldots \;\; x_i^{m-1} \;\; |x_i - \kappa_1|^{2m-1} \;\; \ldots \;\; |x_i - \kappa_K|^{2m-1}]_{1 \leq i \leq n} \,,$$

$$\boldsymbol{D} = \begin{bmatrix} \boldsymbol{0}_{m \times m} & \boldsymbol{0}_{m \times K} \\ \boldsymbol{0}_{K \times m} & (\boldsymbol{\Omega}^{1/2})^\top \boldsymbol{\Omega}^{1/2} \end{bmatrix} \qquad \text{and} \qquad \boldsymbol{\Omega} = \left[|\kappa_k - \kappa_{k'}|^{2m-1}\right]_{1 \leq k, k' \leq \kappa} \,.$$

The resulting estimate of $f(x)$ is of the form

$$\hat{f}(x) = \sum_{j=0}^{m-1} \hat{\beta}_j x^j + \sum_{k=1}^{K} \widehat{u}_k |x - \kappa_k|^{2m-1} \,,$$

where m is a fixed positive integer. It corresponds to a piecewise $(2m-1)$th degree polynomial function and can be differentiated over each piece to obtain $\widehat{f^{(r)}}(x)$.

© 2004 Australian Statistical Publishing Association Inc.

2.3. Simultaneous confidence bands

Let $\mathcal{X}$ denote the set of x values of interest. Often $\mathcal{X}$ is the smallest interval containing each of the x_i. A $100(1-\alpha)\%$ simultaneous confidence band must satisfy

$$\Pr\left(L(x) \le \mathrm{E}\big(\widehat{f^{(r)}}(x)\big) \le U(x) \quad \text{for all } x \in \mathcal{X}\right) \ge 1-\alpha\,.$$

Define the maximal deviation statistic to be

$$M_r = \sup_{x\in\mathcal{X}} \left| \frac{\widehat{f^{(r)}}(x) - \mathrm{E}(\widehat{f^{(r)}}(x))}{\mathrm{sd}(\widehat{f^{(r)}}(x))} \right|,$$

and let $M_{r,q}$ denote the qth quantile of its distribution. Then standard arguments lead to

$$\begin{aligned} L(x) &= \widehat{f^{(r)}}(x) - M_{r,1-\alpha}\,\widehat{\mathrm{sd}}\big(\widehat{f^{(r)}}(x)\big)\,,\\ U(x) &= \widehat{f^{(r)}}(x) + M_{r,1-\alpha}\,\widehat{\mathrm{sd}}\big(\widehat{f^{(r)}}(x)\big) \qquad (x\in\mathcal{X})\,, \end{aligned}$$

being a $100(1-\alpha)\%$ simultaneous confidence band for $\mathrm{E}(\widehat{f^{(r)}}(x))$ over $\mathcal{X}$. Here $\widehat{\mathrm{sd}}(\widehat{f^{(r)}}(x))$ is an estimate of $\mathrm{sd}(\widehat{f^{(r)}}(x))$. Under the constant variance assumption this involves replacing σ_ε in (3) by an estimate. Determination of the quantile $M_{r,1-\alpha}$ is the main obstacle to computing a simultaneous confidence band.

2.4. SiZer

SiZer (Chaudhuri & Marron, 1999) is an acronym for significance of zero crossings of derivatives. This technique uses estimates and simultaneous confidence bands for $f^{(1)}(x)$ and $f^{(2)}(x)$ at different degrees of freedom.

Figure 1 describes the underlying mechanics of SiZer, using fossil data from Chaudhuri & Marron (1999). It shows first and second derivative estimates, with corresponding 95% simultaneous confidence bands. The response variable y has undergone the linear transformation $y \leftarrow\!\shortmid 10\,000(y-0.707)$ to make the y axes readable. For the first derivative, the regions over which the variability band is positive correspond to those where the regression function is significantly increasing. Regions where the variability band is below the zero line correspond to those where $\mathrm{E}(\widehat{f^{(r)}}(x))$, $r=1,2$, is significantly negative. If the variability band covers a portion of the zero line then nothing can be concluded about the $\mathrm{E}(\widehat{f^{(r)}}(x))$. The bar at the base of the plot is a simple graphic showing where the derivative is positive (black), negative (white), or neither positive nor negative (dark grey); light grey shading shows the data are too sparse for use. Following Chaudhuri & Marron (1999), the light grey corresponds to the 'effective sample size' at x being less than 5. In our notation, the effective sample size is $1/\|\ell_x^{(r)}\|^2$. Analogous descriptions apply to the second derivative, with 'increasing' replaced by 'convex' and 'decreasing' replaced by 'concave'.

In Figure 1(b), around Age $= 115$, the bar goes from white to dark grey to black showing that the dip there in (a) is statistically significant. Similarly, the large hump in (a) from around 95 to 110 on the x axis is significant, since the bar in (b) goes from black to grey to white. However, there is no such change in (b) immediately about 95 and 105 on the x axis, so the secondary bumps in (a) at those ages, and the associated dip around Age $= 100$, may be aberrations. However, the second derivative plot in Figure 1(c) supports the existence of these

© 2004 Australian Statistical Publishing Association Inc.

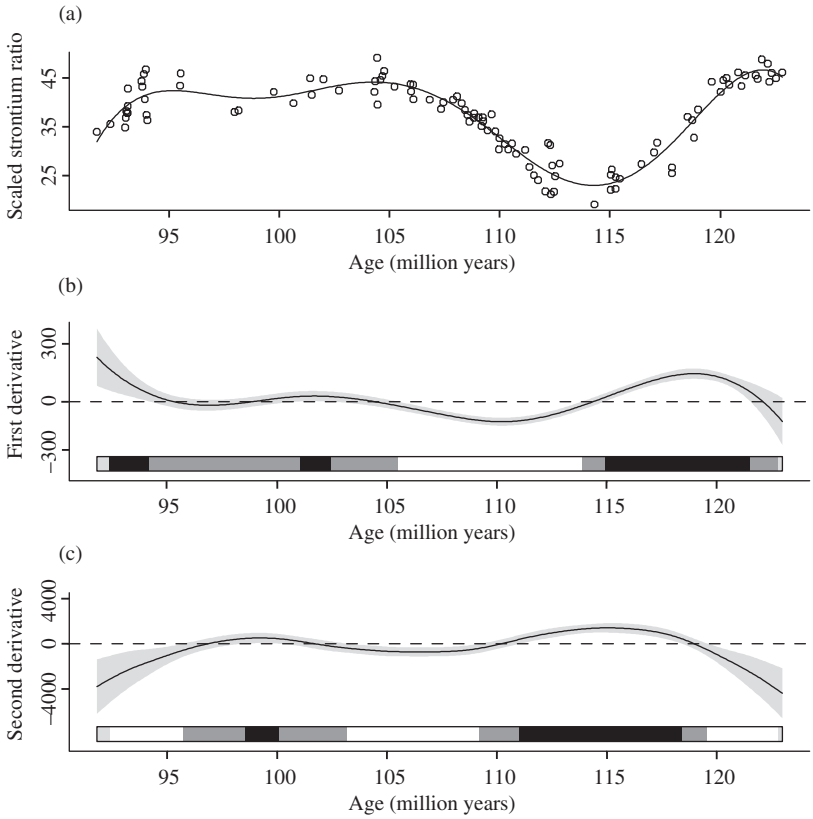

Figure 1. (a) Penalized spline smoothing of fossil data (Chaudhuri & Marron, 1999). Penalized spline estimates of (b) first and (c) second derivatives. Bar shows significance of deviation from zero as described in the text.

bumps. In (c), white corresponds to statistically significant concavity and black corresponds to statistically significant convexity, so the bar in (c) shows inflexion points at about Age = 98 to Age = 102.

A drawback of Figure 1 is that it depends on the amount of smoothing. For example, greater smoothing can remove features that are apparent in less smoothed plots. Chaudhuri & Marron (1999) propose a 'map' of significant zero crossing bars across a range of smoothing parameters, which they call a SiZer map. Figure 2 shows such a map for the fossil data. Chaudhuri & Marron (1999) also discuss adjustments for simultaneous confidence bands across degrees of freedom, but advise that simultaneous confidence bands for individual degrees of freedom in SiZer maps are often a reasonable approximation. Figure 2 uses this approximation, although further research in this direction is warranted; see Section 6.

3. Simulation-based quantile approximation

As discussed in Section 2.3, upper quantiles of

$$M_r = \sup_{x \in \mathcal{X}} \left| \frac{\widehat{f^{(r)}}(x) - \mathrm{E}(\widehat{f^{(r)}}(x))}{\mathrm{sd}(\widehat{f^{(r)}}(x))} \right|$$

© 2004 Australian Statistical Publishing Association Inc.

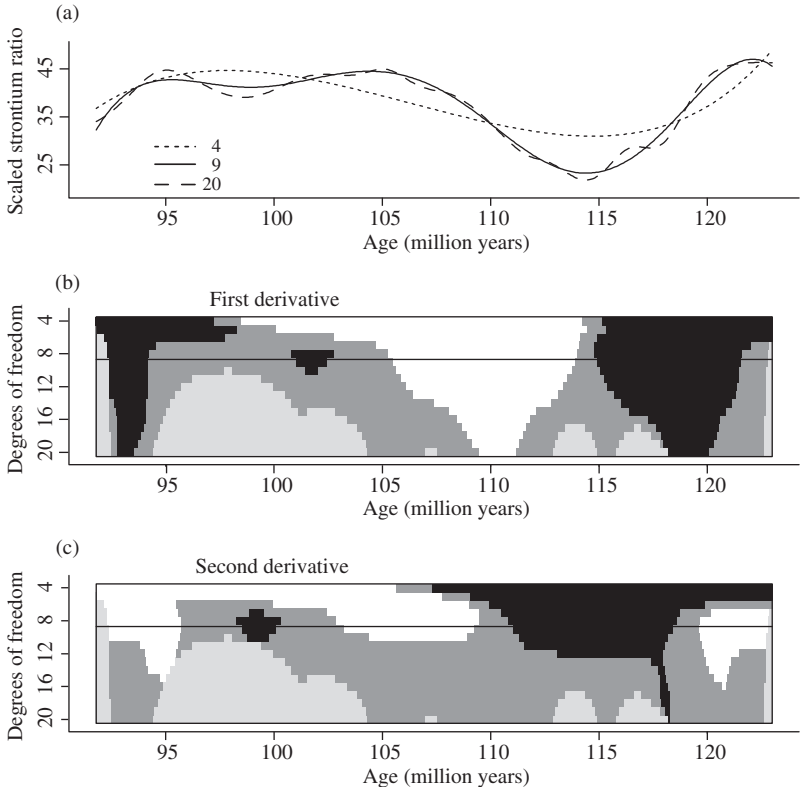

Figure 2. (a) Penalized spline smoothing of fossil data for several df_{fit} values, and maps of significance of deviations from zero for first (b) and second (c) derivatives of the smoothings. Shading is as described in the text for Figure 1. The solid line on (a) and solid horizontal lines on (b) and (c) correspond to degrees of freedom chosen by the restricted-maximum-likelihood-automatic-smoothing-parameter-selector.

are of central importance for feature significance. If normality and homoscedasticity of the errors can be assumed then it is straightforward to approximate $M_{r,1-\alpha}$ via simulation, as we now show.

For each $x \in \mathbb{R}$,

$$\left| \frac{(\boldsymbol{\ell}_x^{(r)})^\top \boldsymbol{y} - (\boldsymbol{\ell}_x^{(r)})^\top \boldsymbol{f}}{\sqrt{\sigma_\varepsilon^2 \|\boldsymbol{\ell}_x^{(r)}\|^2}} \right| = \left| \frac{(\boldsymbol{\ell}_x^{(r)})^\top (\boldsymbol{y} - \boldsymbol{f})/\sigma_\varepsilon}{\|\boldsymbol{\ell}_x^{(r)}\|} \right|.$$

Assuming ε_i independent $\mathrm{N}(0, \sigma_\varepsilon^2)$ it follows that

$$M_r = \sup_{x \in \mathcal{X}} \left| \frac{(\boldsymbol{\ell}_x^{(r)})^\top \boldsymbol{z}}{\|\boldsymbol{\ell}_x^{(r)}\|} \right| \quad \text{where } \boldsymbol{z} \stackrel{\mathrm{d}}{=} \mathrm{N}(\boldsymbol{0}, \boldsymbol{I}_n). \tag{4}$$

A sample from the distribution of M_r can be made by generating a $\mathrm{N}(\boldsymbol{0}, \boldsymbol{I}_n)$ random vector and evaluating the right-hand side of (4) (using a fine grid of x-values over $\mathcal{X}$). Suppose this process is repeated a large number of times, say $N = 10\,000$. The N simulated values are sorted from smallest to largest and the one with rank $\lceil (1-\alpha)N \rceil$ approximates $M_{r,1-\alpha}$.

© 2004 Australian Statistical Publishing Association Inc.

The low-rank nature of penalized splines affords some computational savings since

$$(\boldsymbol{\ell}_x^{(r)})^\top \boldsymbol{z} = \boldsymbol{C}_x^{(r)}(\boldsymbol{C}^\top\boldsymbol{C} + \lambda^{2m-1}\boldsymbol{D})^{-1}\boldsymbol{C}\boldsymbol{z} = \boldsymbol{C}_x^{(r)}\boldsymbol{\Delta}\,, \tag{5}$$

where
$$\boldsymbol{\Delta} \stackrel{\mathrm{d}}{=} \mathrm{N}\big(\boldsymbol{0},\ (\boldsymbol{C}^\top\boldsymbol{C} + \lambda^{2m-1}\boldsymbol{D})^{-1}\boldsymbol{C}^\top\boldsymbol{C}(\boldsymbol{C}^\top\boldsymbol{C} + \lambda^{2m-1}\boldsymbol{D})^{-1}\big) \tag{6}$$

and $\boldsymbol{C}_x^{(r)}$ is the row vector containing rth derivatives of the basis functions evaluated at x. For example, when $r = 1$ and $m = 2$,

$$\boldsymbol{C}_x^{(1)} = [0 \quad 1 \quad c_1 \quad \ldots \quad c_K]\,, \quad \text{where } c_k = 3(x - \kappa_k)|x - \kappa_k|\,.$$

Note that $\boldsymbol{\Delta}$ requires random vectors of dimension corresponding to the number of basis functions, which can be considerably less than n.

For large N, $M_{r,0.95}$ can be approximated very accurately. On five independent simulations of 10 000 draws each we obtained $M_{0,0.95} \approx 3.172, 3.198, 3.172, 3.201, 3.199$ for the fossil data. The smallest and largest of the five differ by less than 2%.

4. Analytic quantile approximation

Simulation-based approximation of $M_{r,1-\alpha}$, while accurate, is computationally expensive. Therefore, analytic approximations to $M_{r,1-\alpha}$ are also worth considering. We now describe two that are commonly used in the feature significance literature.

4.1. Independent blocks

Based on an 'independent blocks' argument and Bonferroni adjustment, originally used by Härdle & Marron (1991), Chaudhuri & Marron (1999, 2000) advocate the fast-to-compute approximation

$$M^{\mathrm{IB}}_{r,1-\alpha} = \Phi^{-1}\big(\tfrac{1}{2}\big(1 + (1-\alpha)^{1/df_{\mathrm{fit}}}\big)\big)\,.$$

This approximation is based on the number of degrees of freedom for estimation of f, and does not depend on the order of derivative being estimated.

4.2. Upcrossing theory

Loader (1999 Section 9.2) surveys asymptotics for simultaneous confidence bands based on upcrossing theory (Rice, 1945). For linear smoothers as defined by (1), upcrossing theory leads to

$$\Pr\Big(\Big|\frac{\widehat{f^{(r)}}(x) - \mathrm{E}(\widehat{f^{(r)}}(x))}{\widehat{\mathrm{sd}}(\widehat{f^{(r)}}(x))}\Big| > c \quad \text{for all } x \in \mathcal{X}\Big) \approx \frac{\kappa_r}{\pi}e^{-c^2/2} + 2\big(1 - \Phi(c)\big)\,, \tag{7}$$

where
$$\kappa_r = \int_{\mathcal{X}} \frac{\sqrt{\|\boldsymbol{\ell}_x^{(r)}\|^2\|(\boldsymbol{\ell}_x^{(r)})'\|^2 - \big((\boldsymbol{\ell}_x^{(r)})^\top(\boldsymbol{\ell}_x^{(r)})'\big)^2}}{\|\boldsymbol{\ell}_x^{(r)}\|^2}\,dx\,.$$

Such an approximation was used by Knafl, Sacks & Ylvisaker (1985) for simultaneous confidence bands for a general class of regression functions. Note that $(\boldsymbol{\ell}_x^{(r)})'$ is obtained from

© 2004 Australian Statistical Publishing Association Inc.

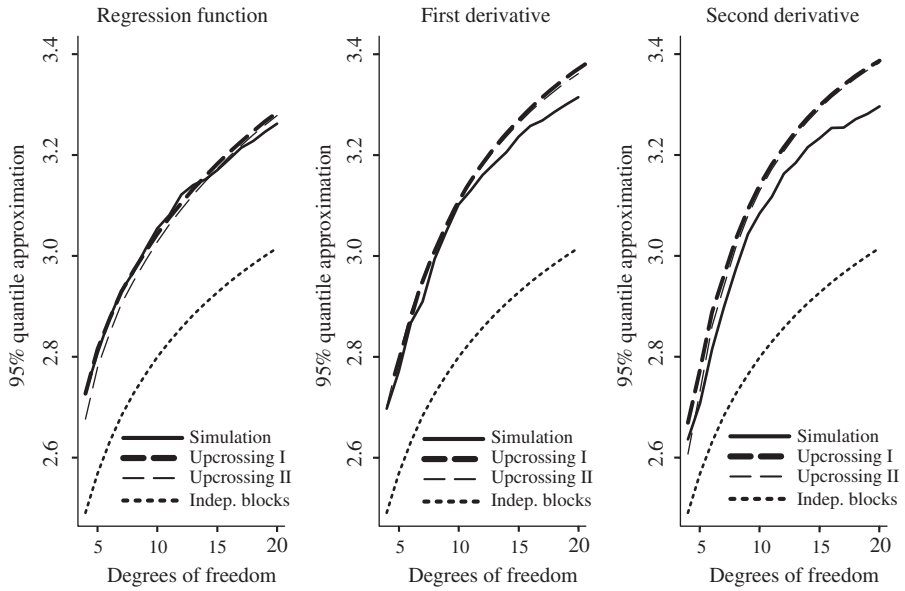

Figure 3. Comparison of $M^{\text{sim}}_{r,0.95}$, $M^{\text{UCI}}_{r,0.95}$, $M^{\text{UCII}}_{r,0.95}$ and $M^{\text{IB}}_{r,0.95}$ for the fossil data when $r = 0, 1, 2$. Penalized radial quintic splines are used throughout.

$\boldsymbol{\ell}^{(r)}$ via element-wise differentiation with respect to x. If, as in Section 2.2.2, $\widehat{f^{(r)}(x)}$ is the rth derivative of $\hat{f}(x)$ then $(\boldsymbol{\ell}^{(r)}_x)' = \boldsymbol{\ell}^{(r+1)}_x$.

The approximation to $M_{r,1-\alpha}$ implied by (7) is

$$M^{\text{UCI}}_{r,1-\alpha} = \left\{ c > 0\text{: } \frac{\kappa_r}{\pi}\, e^{-c^2/2} + 2\big(1 - \Phi(c)\big) - \alpha = 0 \right\}.$$

However, $2(1 - \Phi(c))$ is quite small for $c > 2$ and often has little effect on the result. Ignoring this term leads to the following closed-form approximation based on probability of upcrossing:

$$M^{\text{UCII}}_{r,1-\alpha} = \sqrt{2 \log_e \left(\frac{\kappa_r}{\alpha\pi} \right)}.$$

For differentiable smoothers, the vectors $\boldsymbol{\ell}^{(r)}_x$ and $(\boldsymbol{\ell}^{(r)}_x)'$ are straightforward and inexpensive to compute over a fine grid, and quadrature can be used to obtain an accurate approximation to κ_r.

5. Comparisons

5.1. Numerical comparison

We compared the values of $M^{\text{sim}}_{r,0.95}$, $M^{\text{UCI}}_{r,0.95}$, $M^{\text{UCII}}_{r,0.95}$ and $M^{\text{IB}}_{r,0.95}$ for various degrees of freedom for the fossil data, and started with penalized radial quintic splines. Figure 3 provides a visual summary of the results. It shows that each of the upcrossing approximations is very accurate. On the other hand, the independent block approximation is roughly 10% smaller than the simulation-based approximation and leads to more liberal analysis of feature significance.

© 2004 Australian Statistical Publishing Association Inc.

We also ran a simulation with sample sizes $n = 100, 1000$ and four random designs for $x_1, \ldots, x_n$. For $j = 1, 2, 3$, Design j corresponds to the Beta($\frac{1}{4}(j+5)$, $\frac{1}{4}(11-j)$) density and Design 4 corresponds to the uniform density on $(0, 1)$ Note that the quantile approximations do not depend on the y_i or σ_ε.

Figure 4 summarizes the results for $M^{\text{UCI}}_{r,0.95}/M^{\text{sim}}_{r,0.95}$. The plot for $M^{\text{UCII}}_{r,0.95}/M^{\text{sim}}_{r,0.95}$ is quite similar so is not shown. The figure shows that the ratio is always between 0.99 and 1.03. The accuracy of $M^{\text{UCI}}_{r,0.95}$ deteriorates as r increases, but slightly improves as n increases.

5.2. Heuristic comparison

For large values of c,

$$\Phi(c) \approx 1 - \frac{1}{\sqrt{2\pi}c}\exp(-\tfrac{1}{2}c^2)\,, \quad \text{and} \quad \Pr_{\text{IB}}(M_r > c) \approx df_{\text{fit}}\frac{2}{\sqrt{2\pi}c}\exp(-\tfrac{1}{2}c^2)\,.$$

The literature on the maxima of random fields (e.g. Adler, 1981, 2001) suggests that this dependence of $\Pr(M_r > c)$ on c is qualitatively incorrect because the first term of the corresponding asymptotic expansion for $\Pr(M_r > c)$ should be proportional to $\exp(-\frac{1}{2}c^2)$. In the simulation studies we found that IB-based approximation underestimates the value of $\Pr(M_r > c)$ for large threshold values of c and the resulting value of $M^{\text{IB}}_{r,1-\alpha}$ consistently underestimates the true value of $M_{r,1-\alpha}$. This discrepancy appears to be uniform across df_{fit} values, suggesting an inherent defect with $M^{\text{IB}}_{r,0.95}$.

6. Adjustment for multiple degrees of freedom

The SiZer maps in Figure 2 allow assessment of feature significance at the 0.05 level of significance for individual degrees of freedom. An adjustment could allow for simultaneous inference across the degrees of freedom. Such adjustments are discussed by Siegmund & Worsley (1995), Loader (1999), Chaudhuri & Marron (1999). We do not address adjustment for multiple degrees of freedom.

7. Two-dimensional quantile approximation

SiZer has been extended to bivariate designs by Godtliebsen, Marron & Chaudhuri (2004) in the rectangular design case, and Ganguli & Wand (2004) for general designs. The latter authors used low-rank thin-plate splines for estimating $f(\boldsymbol{x})$, where $\boldsymbol{x} \in \mathbb{R}^2$. Here, for simplicity, we describe thin-plate splines with smoothness $m = 2$ in the notation of Wahba (1990). Estimates of $f(\boldsymbol{x})$ in the form

$$\hat{f}(\boldsymbol{x}) = \hat{\beta}_0 + \hat{\boldsymbol{\beta}}_1^{\mathsf{T}}\boldsymbol{x} + \sum_{k=1}^{K}\widehat{u}_k\|\boldsymbol{x} - \boldsymbol{\kappa}_k\|^2\log(\|\boldsymbol{x} - \boldsymbol{\kappa}_k\|)\,,$$

where $\boldsymbol{\kappa}_1, \ldots, \boldsymbol{\kappa}_K \in \mathbb{R}^d$, are chosen to 'fill the space' corresponding to the $\boldsymbol{x}_i$ (e.g. Nychka & Saltzman, 1998). The gradient vector is

$$\begin{bmatrix} \widehat{f^{(1,0)}}(\boldsymbol{x}) \\ \widehat{f^{(0,1)}}(\boldsymbol{x}) \end{bmatrix} = \hat{\boldsymbol{\beta}}_1 + \sum_{k=1}^{K}\widehat{u}_k(2\log\|\boldsymbol{x} - \boldsymbol{\kappa}_k\| + 1)(\boldsymbol{x} - \boldsymbol{\kappa}_k)\,,$$

© 2004 Australian Statistical Publishing Association Inc.

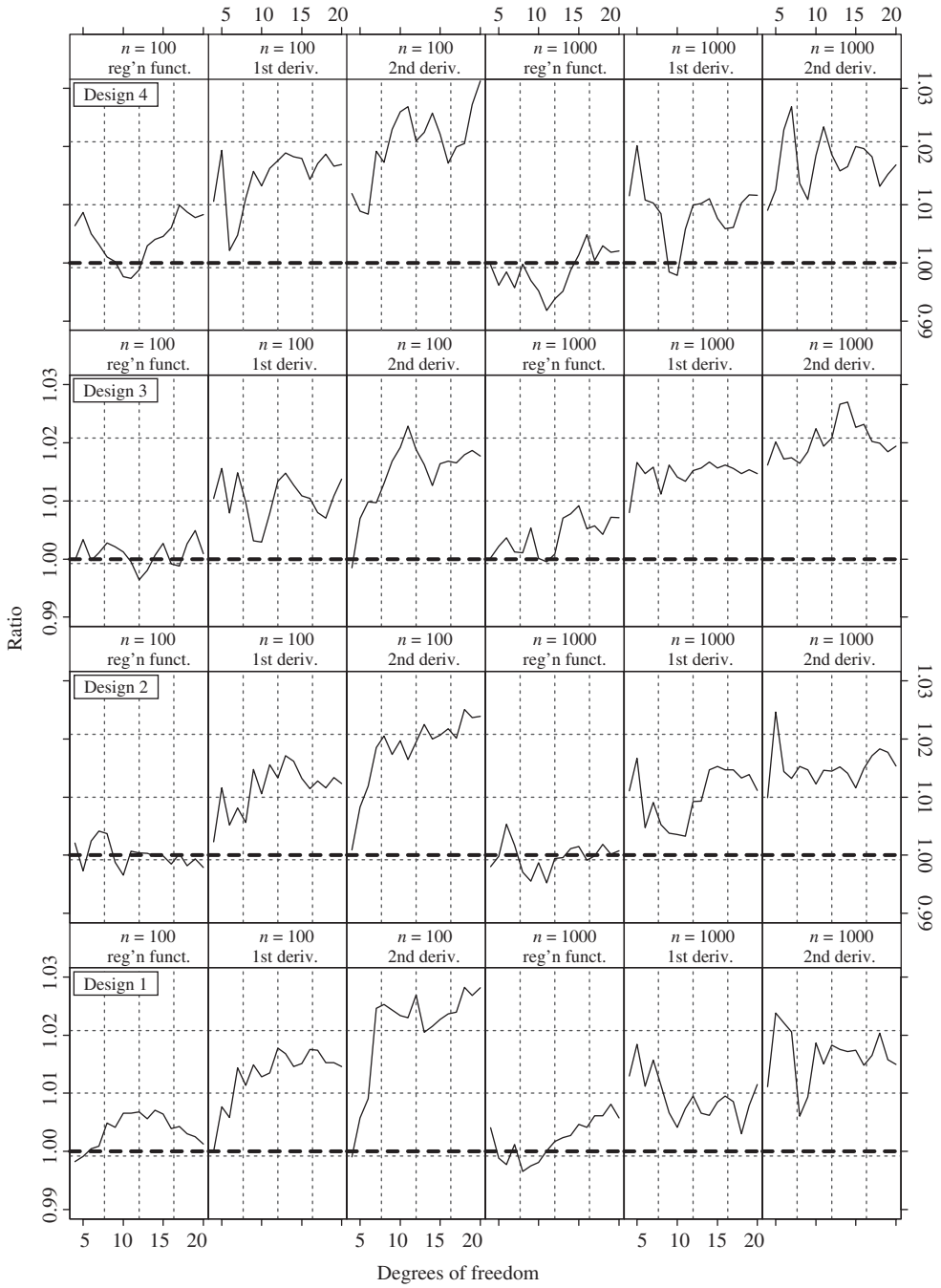

Figure 4. Summary of simulation comparison for $M^{\mathrm{UCI}}_{r,0.95}$ and $M^{\mathrm{sim}}_{r,0.95}$. Each panel shows $M^{\mathrm{UCI}}_{r,0.95}/M^{\mathrm{sim}}_{r,0.95}$ for $r = 0, 1, 2$, $n = 100, 1000$ and four random designs. The vertical range of all panels is 0.99 to 1.03.

© 2004 Australian Statistical Publishing Association Inc.

where $f^{(1,0)}$ and $f^{(0,1)}$ denote the partial derivatives of f with respect to the first and second arguments respectively. The analogue of (5) for $\widehat{f^{(1,0)}}$ is

$$\widehat{\boldsymbol{f}^{(1,0)}}(\boldsymbol{x}) - \mathrm{E}\big(\widehat{\boldsymbol{f}^{(1,0)}}(\boldsymbol{x})\big) = \boldsymbol{C}_{\boldsymbol{x}}^{(1,0)}\boldsymbol{\Delta}\,,$$

where $\boldsymbol{C}_{\boldsymbol{x}}^{(1,0)} = [0 \quad 1 \quad c_1 \quad \ldots \quad c_K]\,, \quad c_k = (2\log\|\boldsymbol{x}-\boldsymbol{\kappa}_k\|+1)(\boldsymbol{x}-\boldsymbol{\kappa}_k)\,,$

and $\boldsymbol{\Delta}$ is defined analogously to (6). Similar expressions apply to $\hat{f}^{(0,1)}$.

First derivative feature significance requires simultaneous inference concerning the gradient-size function

$$G(\boldsymbol{x}) = \sqrt{\widehat{f^{(1,0)}}(\boldsymbol{x})^2 + \widehat{f^{(0,1)}}(\boldsymbol{x})^2}\,.$$

Departure of $G(\boldsymbol{x})$ from 0 over $\mathcal{X} \subseteq \mathbb{R}^2$ is based on the statistic

$$\sup_{\boldsymbol{x}\in\mathcal{X}} \left| \begin{bmatrix} \widehat{f^{(1,0)}}(\boldsymbol{x}) - \mathrm{E}\big(\widehat{f^{(1,0)}}(\boldsymbol{x})\big) \\ \widehat{f^{(0,1)}}(\boldsymbol{x}) - \mathrm{E}\big(\widehat{f^{(0,1)}}(\boldsymbol{x})\big) \end{bmatrix}^{\mathrm{T}} V\left(\begin{bmatrix} \widehat{f^{(1,0)}}(\boldsymbol{x}) \\ \widehat{f^{(0,1)}}(\boldsymbol{x}) \end{bmatrix}\right)^{-1} \begin{bmatrix} \widehat{f^{(1,0)}}(\boldsymbol{x}) - \mathrm{E}\big(\widehat{f^{(1,0)}}(\boldsymbol{x})\big) \\ \widehat{f^{(0,1)}}(\boldsymbol{x}) - \mathrm{E}\big(\widehat{f^{(0,1)}}(\boldsymbol{x})\big) \end{bmatrix} \right|.$$

Let $M^G_{r,1-\alpha}$ denote the $1-\alpha$ quantile of this statistic. It is apparent from its expression that the simulation-based approximation of $M^G_{r,1-\alpha}$ is achievable through extension of the approach described in (3). Ganguli & Wand (2004) ran some tests on this approximation. For fixed amounts of smoothing several simulations were drawn and the resulting $M^G_{r,1-\alpha}$ values were compared. For simulation sizes of 10 000 the Monte Carlo error was found to be insignificant in each case.

An interesting open problem is the derivation of an analytic approximation to $M^G_{r,q}$.

8. Closing remarks

Feature significance methods such as SiZer maps have a promising future because many datasets (e.g. those arising in flow cytometry) contain signals at different levels of spatial resolution. However, the quantile approximation needs fine tuning for high quality testing of features. In this paper we have shown that upcrossing-probability theory yields better quantile approximations than the independent blocks method. As mentioned in Sections 6 and 7 respectively, further research is warranted for multiple degrees of freedom adjustment and gradient approximation in two dimensions.

References

ADLER, R.J. (1981). *The Geometry of Random Fields*. New York: Wiley.

ADLER, R.J. (2001). On excursion sets, tube formulas and maxima of random fields. *Ann. Appl. Probab.* **10**, 1–74.

CHAUDHURI, P. & MARRON, J.S. (1999). SiZer for exploration of structures in curves. *J. Amer. Statist. Assoc.* **94**, 807–823.

CHAUDHURI, P. & MARRON, J.S. (2000). Scale space view of curve estimation. *Ann. Statist.* **28**, 408–428.

EILERS, P.H.C. & MARX, B.D. (1996). Flexible smoothing with B-splines and penalties (with discussion). *Statist. Sci.* **11**, 89–121.

EUBANK, R.L. (1999). *Spline Smoothing and Nonparametric Regression*. New York: Marcel Dekker.

FAN, J. & GIJBELS, I. (1995). *Local Polynomial Modelling and Its Applications*. London: Chapman and Hall.

© 2004 Australian Statistical Publishing Association Inc.

FRENCH, J.L., KAMMANN, E.E. & WAND, M.P. (2001). Comment on Ke and Wang. *J. Amer. Statist. Assoc.* **96**, 1285–1288.

GANGULI, B. & WAND, M.P. (2004). Feature significance in geostatistics. *J. Comput. Graph. Statist.* (in press).

GODTLIEBSEN, F., MARRON, J.S. & CHAUDHURI, P. (2002). Significance in scale space for bivariate density estimation. *J. Comput. Graph. Statist.* **11**, 1–22.

GODTLIEBSEN, F., MARRON, J.S. & CHAUDHURI, P. (2004). Statistical significance of features in digital images. *Image and Vision Computing* **13**, 1093–1104.

GREEN, P.J. & SILVERMAN, B.W. (1994). *Nonparametric Regression and Generalized Linear Models.* London: Chapman and Hall.

HÄRDLE, W. & MARRON, J.S. (1991). Bootstrap simultaneous error bars for nonparametric regression. *Ann. Statist.* **19**, 778–796.

KNAFL, G., SACKS, J. & YLVISAKER, D. (1985). Confidence bands for regression functions. *J. Amer. Statist. Assoc.* **80**, 683–691.

LANGE, N. (1999). Statistical procedures for functional MRI. In *Medical Radiology — Diagnostic Imaging and Radiation Oncology: Functional MRI*, eds C. Moonen & P.A. Bandettini, **27**, 301–335. New York: Springer.

LANGE, N. (2004). What can modern statistics offer imaging neuroscience? *Statistical Methods in Medical Research* **12**, 447–469.

LOADER, C. (1999). *Local Regression and Likelihood.* New York: Springer-Verlag.

NYCHKA, D. & SALTZMAN, N. (1998). Design of air quality monitoring networks. In *Case Studies in Environmental Statistics*, Lecture Notes in Statistics, Vol. 132, eds D. Nychka, W. Piegorsch & L. Cox, pp . 51–76. New York: Springer-Verlag.

RICE, S.O. (1945). Mathematical analysis of random noise. *Bell System Technical J.* **24**, 46–156.

RUPPERT, D. & CARROLL, R.J. (2000). Spatially-adaptive penalties for spline fitting. *Aust. N. Z. J. Stat.* **42**, 205–224.

SIEGMUND, D.O. & WORSLEY, K.J. (1995). Testing for a signal with unknown location and scale in a stationary Gaussian random field. *Ann. Statist.* **23**, 608–639.

WAGER, C.G., COULL, B.A. & LANGE, N. (2004). Modelling spatial intensity for replicated inhomogeneous point patterns in brain imaging. *J. Roy. Statist. Soc. Ser. B* **66**, 429–446.

WAHBA, G. (1990). *Spline Models for Observational Data.* Philadelphia: SIAM.

WAND, M.P. & JONES, M.C. (1995). *Kernel Smoothing.* London: Chapman and Hall.

WORSLEY, K.J. (1994). Local maxima and the expected Euler characteristic of excursion sets of chi-squared, F, and t fields. *Adv. in Appl. Probab.* **26**, 13–42.

WORSLEY, K.J., EVANS, A.C., MARRETT, S. & NEELIN, P. (1992). Determining the number of statistically significant areas of activation in subtracted studies from PET. *Journal of Cerebral Blood Flow and Metabolism* **12**, 900–918.

© 2004 Australian Statistical Publishing Association Inc.