

Nonparametric Estimation of Residual Moments and Covariance

BY DAFYDD EVANS AND ANTONIA J. JONES

School of Computer Science, Cardiff University, 5 The Parade, Cardiff CF24 3AA, UK

Abstract: The aim of nonparametric regression is to model the behaviour of a response vector Y in terms of an explanatory vector X , based only on a finite set of empirical observations. This is usually performed under the additive hypothesis $Y = f(X) + R$, where $f(X) = \mathcal{E}(Y|X)$ is the true regression function and R is the true residual variable. Subject to a Lipschitz condition on f , we propose new estimators for the moments (scalar response) and covariance (vector response) of the residual distribution, derive their asymptotic properties and discuss their application in practical data analysis.

Keywords: nonparametric regression, exploratory data analysis, difference-based methods, nearest neighbours

1. Introduction

In the analysis of complex systems, it is often the case that we have no prior information regarding the system under investigation, and the analysis must proceed based only on empirical observations of the system behaviour. Nonparametric regression methods attempt to model the behaviour of an observable random vector $Y \in \mathbb{R}^n$ in terms of another observable random vector $X \in \mathbb{R}^m$, based only on a finite set of observations.

The relationship between the response vector Y and the explanatory vector X is often assumed to satisfy the additive hypothesis:

$$Y = f(X) + R \tag{1.1}$$

where $f(X) = \mathcal{E}(Y|X)$ is the true regression function and R is the true residual vector, which we call *noise*.

For a scalar response ($n = 1$) we present new nonparametric estimators for the moments $\mathcal{E}(R^q)$ of the noise distribution ($q = 2, 3, 4, \dots$), and for a vector response ($n > 1$) we describe a new estimator for the covariance matrix $\mathcal{E}(RR^T)$. Our estimators are computed from a finite sample of independent and identically distributed observations of the joint variable (X, Y) by exploiting the nearest neighbour structure of the explanatory observations. Using techniques first described in Bentley (1975), our estimates can be computed with asymptotic time complexity of order $\mathcal{O}(N \log N)$ as the number of points $N \rightarrow \infty$.

Our main assumptions are that the true regression function $\mathcal{E}(Y|X)$ is a smooth function of the explanatory vector X , and that the zero mean noise vector R is independent of the explanatory vector (homoscedastic). The noise vector represents all

variation in the response vector which is not accounted for by any smooth transformation of the explanatory vector, and includes what might be called deterministic variation due to non-random factors that influence the response vector but which are not included in the explanatory vector, as well as stochastic variation due to purely random factors such as measurement error.

Our estimators belong to the class of difference-based methods. These date back to von Neumann (1941) who, for the case $m = 1$ and under the assumption of homoscedastic noise, proposed that the residual variance can be estimated by the average of squared successive differences of the observations,

$$\widehat{\mathcal{E}(R^2)} = \frac{1}{2(N-1)} \sum_{i=2}^N (Y_i - Y_{i-1})^2 \quad (1.2)$$

This estimator was investigated by Rice (1984) and subsequently extended by Gasser *et al.* (1986) and Hall *et al.* (1990) by using interpolation to reduce the bias of the so-called ‘pseudo-residual’ $Y_i - Y_{i-1}$. For the case $m = 1$, Levine (2006) and Levine & Brown (2007) consider heteroscedastic noise, and propose difference-based estimators for the residual variance function $\mathcal{E}(R^2|X)$ using weighted local averages of the squared pseudo-residuals. A related paper by Wang *et al.* (2008) investigates the effect that the mean function $f(X) = \mathcal{E}(Y|X)$ has on estimators of the variance function. For the case $m \geq 1$, Munk *et al.* (2005) show that the bias of simple difference-based estimators such as (1.2) have asymptotic order $\mathcal{O}(N^{-1/m})$ as $N \rightarrow \infty$ and are therefore not \sqrt{N} -consistent for $m \geq 4$. They also describe estimators that allow better control of the bias and remain consistent when $m \geq 4$. The case $m \geq 1$ is also discussed in Evans & Jones (2002) and Bock *et al.* (2007), where the assumption that the explanatory vectors lie on a rectangular grid is relaxed by considering their nearest neighbour structure. For the case $m \geq 1$ and under the assumption of homoscedastic noise, Durrant (2001) presents difference-based estimators for the higher moments $\mathcal{E}(R^q)$ of symmetric noise distributions. To the best of our knowledge, this is the only previous attempt at estimating the higher moments of residual distributions.

(a) Notation and Conditions

Let $(X_1, Y_1), \dots, (X_N, Y_N)$ be a sample of independent and identically distributed observations of the joint variable (X, Y) . To establish asymptotic bounds on the bias and variance of our estimators as the number of points $N \rightarrow \infty$, we impose some fairly strong conditions on the distribution of the explanatory observations X_1, \dots, X_N . To this end, let $\Phi : \mathbb{R}^m \rightarrow [0, 1]$ be a probability distribution whose density $\phi(x) = \Phi'(x)$ exists at every point $x \in \mathbb{R}^m$.

Definition 1. A probability distribution Φ satisfies a smooth density condition if its density has bounded partial derivatives at every point $x \in \mathbb{R}^m$.

For $x \in \mathbb{R}^m$ let $B_x(s)$ denote the ball of radius s centred at x , and let $\omega_x(s)$ denote its probability measure:

$$\omega_x(s) = \int_{B_x(s)} \phi(\xi) d\xi \quad (1.3)$$

Definition 2. A probability distribution Φ satisfies a positive density condition if there exist constants $a > 1$ and $\delta > 0$ such that

$$\frac{s^m}{a} \leq \omega_x(s) \leq as^m \quad \text{for all } 0 \leq s \leq \delta \quad (1.4)$$

The positive density condition (Gruber 2004) ensures that the probability measure of an arbitrary ball can be bounded in terms of its radius. If a distribution satisfies a smooth density condition, Evans *et al.* (2002) show that the distribution also satisfies a positive density condition provided the support S_ϕ of the density is a compact convex body in \mathbb{R}^m , where

$$S_\phi = \{x \in \mathbb{R}^m : \phi(x) > 0\} \quad (1.5)$$

That a distribution satisfies the positive density condition essentially means that its density is well behaved in small neighbourhoods. If the distribution also satisfies a smooth density condition, then its density is approximately uniform in small neighbourhoods. The left hand inequality of (1.4) ensures that the density is zero outside some bounded region. The positive density condition also ensures that the probability measure of small neighbourhoods scale with their Lebesgue measure, and thus excludes the case where the support of the density is a fractal set.

We state the following conditions:

- C.1 The true regression function $f(X) = \mathcal{E}(Y|X)$ is Lipschitz continuous. Let b denote the associated Lipschitz constant, so that $\|f(x) - f(x')\| \leq b\|x - x'\|$ for all $x, x' \in \mathbb{R}^m$.
- C.2 The explanatory observations X_1, \dots, X_N are independent and identically distributed random vectors in \mathbb{R}^m , and their common distribution Φ satisfies smooth and positive density conditions. Furthermore, the support S_ϕ of the associated density is a compact subset of \mathbb{R}^m such that $\|x - x'\| \leq 1$ for all $x, x' \in S_\phi$.
- C.3 The (implicit) noise observations R_1, \dots, R_N are independent and identically distributed random vectors, and to estimate $\mathcal{E}(R^q)$ their common distribution Ψ must have bounded moments up to order $2q \in \mathbb{N}$.
- C.4 For every observation pair (X_i, Y_i) , the explanatory vector X_i and the (implicit) noise vector R_i are independent (homoscedastic).

We call Φ the sampling distribution, and Ψ the noise distribution or the true residual distribution. The condition that $\|x - x'\| \leq 1$ for all $x, x' \in S_\phi$ is imposed to ensure that $\|x - x'\|^t \leq \|x - x'\|$ for all $t \geq 1$. In practice, this condition can be enforced by an appropriate scaling of the explanatory observations X_1, \dots, X_N , which might lead to a modification of the Lipschitz constant b defined in condition C.1.

(b) Optimal estimators

A number of algorithms for nonparametric regression exist in the literature, most of which require an estimate of the noise variance $\mathcal{E}(R^2)$ to control the amount of

smoothing applied to the data. Too little smoothing means that noise is incorporated into the estimate, while too much smoothing leads to certain characteristics of the underlying regression function being lost. An estimate of $\mathcal{E}(R^2)$ can help to identify suitable bandwidths for kernel regression (Silverman 1986), second-derivative penalties for spline smoothing (Eubank 1999), stopping criteria for neural network training (Jones 2004) and thresholds for wavelet smoothing (Donoho & Johnstone 1995).

To evaluate an estimator $\hat{f}(X)$ of the true regression function $f(X)$, we consider the associated error function,

$$e(X) = f(X) - \hat{f}(X) \quad (1.6)$$

In practice, $e(X)$ is unknown and instead we must consider the *empirical* error function,

$$\hat{e}(X, Y) = Y - \hat{f}(X) \quad (1.7)$$

Since the explanatory vector X and the noise vector R are assumed to be independent, it follows by (1.1) that the empirical mean squared error satisfies

$$\mathcal{E}(\hat{e}^2|X) = \mathcal{E}(R^2) + e(X)^2 \quad (1.8)$$

Hence if $\mathcal{E}(\hat{e}^2|X)$ is close to the noise variance $\mathcal{E}(R^2)$, the error function must be close to zero at X and we might therefore define an optimal estimator to be one for which $\mathcal{E}(\hat{e}^2|X) = \mathcal{E}(R^2)$ for all $X \in S_\phi$. In general, the q th moment of the empirical error function satisfies

$$\mathcal{E}(\hat{e}^q|X) = \mathcal{E}(R^q) + \sum_{t=1}^q \binom{q}{t} \mathcal{E}(R^{q-t})e(X)^t \quad (1.9)$$

and we might instead define an optimal estimator to be one for which $\mathcal{E}(\hat{e}^q|X) = \mathcal{E}(R^q)$ for all $q = 1, 2, \dots$ and $X \in S_\phi$. Thus we are motivated to estimate the moments of the noise distribution.

2. Products of Differences

Let $\mathcal{X} = (X_1, \dots, X_N)$ and $\mathcal{Y} = (Y_1, \dots, Y_N)$ denote the marginal samples. Under the additive hypothesis (1.1), the noise variance $\mathcal{E}(R^2)$ can be estimated by the mean squared difference between pairs of response vectors Y and Y' for which the corresponding explanatory vectors X and X' are nearest neighbours in the marginal sample \mathcal{X} . By (1.1), since R and R' are independent and identically distributed zero-mean random variables it follows that

$$\frac{1}{2}\mathcal{E}(Y - Y')^2 = \mathcal{E}(R^2) + \frac{1}{2}\mathcal{E}(f(X) - f(X'))^2 \quad (2.1)$$

where the second term on the right hand side represents the bias of $\frac{1}{2}\mathcal{E}(Y - Y')^2$ as an estimator for $\mathcal{E}(R^2)$. Rather than using squared differences, we propose an alternative estimator for $\mathcal{E}(R^2)$ based on *products* of differences $(Y - Y')(Y - Y'')$, where Y' and Y'' are the response vectors for which the corresponding explanatory vectors X' and X'' are respectively the first and second nearest neighbours of X .

By (1.1), since R , R' and R'' are independent and identically distributed zero-mean random variables we have that

$$\mathcal{E}((Y - Y')(Y - Y'')) = \mathcal{E}(R^2) + \mathcal{E}((f(X) - f(X'))(f(X) - f(X''))) \quad (2.2)$$

where the second term on the right hand side represents the bias of $\mathcal{E}((Y - Y')(Y - Y''))$ as an estimator for $\mathcal{E}(R^2)$. By the Lipschitz condition on f ,

$$\mathcal{E}((Y - Y')(Y - Y'')) = \mathcal{E}(R^2) + \mathcal{O}(\mathcal{E}(\|X - X''\|)) \quad \text{as } \|X - X''\| \rightarrow 0 \quad (2.3)$$

Estimators for the higher order moments $\mathcal{E}(R^q)$ can be similarly constructed using q -fold products of differences over the first q nearest neighbours of X . Let $i(k)$ denote the index of the k th nearest neighbour of X_i among the points X_1, \dots, X_N . We propose the following sample mean estimator for $\mathcal{E}(R^q)$:

$$\Gamma_N(\mathcal{X}, \mathcal{Y}) = \frac{1}{N} \sum_{i=1}^N \gamma_i(\mathcal{X}, \mathcal{Y}) \quad \text{where} \quad \gamma_i(\mathcal{X}, \mathcal{Y}) = \prod_{k=1}^q (Y_i - Y_{i(k)}) \quad (2.4)$$

To compute the q th moment, we need not restrict ourselves to the first q nearest neighbours of X_i , since any set of q points in the neighbourhood of X can be used to compute a q -fold product of differences and hence an estimate for $\mathcal{E}(R^q)$. Although using more distant neighbours is likely to increase the bias of the estimate, these products might be useful to reduce statistical error when the number of observations is small. We can also form products of differences over a set of neighbours that are not necessarily distinct. In this case, some differences are repeated in the product, and the method returns an estimate for a known polynomial function of the moments up to order q . The precise form of the polynomial depends on how many differences are repeated, and how many times.

(a) Covariance

For the case where the response is a vector $Y \in \mathbb{R}^n$, since the noise vector has mean zero it follows by (1.1) that

$$\mathcal{E}(YY^T) = \mathcal{E}(RR^T) + \mathcal{E}(f(X)f(X)^T) \quad (2.5)$$

To estimate $\mathcal{E}(RR^T)$, let (X', Y') and (X'', Y'') denote the sample points for which X' and X'' are respectively the first and second nearest neighbours of the point X in the marginal sample \mathcal{X} . As above, by the Lipschitz condition on f the absolute differences $\|f(X) - f(X')\|$ and $\|f(X) - f(X'')\|$ are small provided the corresponding distances $\|X - X'\|$ and $\|X - X''\|$ are also small. Thus, because the noise components R , R' and R'' are independent and identically distributed zero mean random vectors, we have

$$\mathcal{E}((Y - Y')(Y - Y'')^T) = \mathcal{E}(RR^T) + \mathcal{O}(\mathcal{E}(\|X - X''\|)) \quad \text{as } \|X - X''\| \rightarrow 0 \quad (2.6)$$

and we propose the following sample mean estimator for $\mathcal{E}(RR^T)$:

$$\Gamma_N^{\text{cov}}(\mathcal{X}, \mathcal{Y}) = \frac{1}{N} \sum_{i=1}^N \gamma_i^{\text{cov}}(\mathcal{X}, \mathcal{Y}) \quad \text{where} \quad \gamma_i^{\text{cov}}(\mathcal{X}, \mathcal{Y}) = (Y_i - Y'_i)(Y - Y''_i)^T \in \mathbb{R}^{n \times n} \quad (2.7)$$

3. Asymptotic Results

We now proceed to establish an asymptotic upper bound on the bias of our moment estimator Γ_N , and verify its probabilistic convergence as the number of observations $N \rightarrow \infty$. To this end, we first consider the mean distance from a point to its q th nearest neighbour in the marginal sample $\mathcal{X} = (X_1, \dots, X_N)$, which we denote by

$$\Delta_N(\mathcal{X}) = \frac{1}{N} \sum_{i=1}^N \delta_i(\mathcal{X}) \quad \text{where} \quad \delta_i(\mathcal{X}) = \|X_i - X_{i(q)}\| \quad (3.1)$$

An almost identical argument to one found in (Evans *et al.* 2002) leads to the following result, which we state without proof.

Lemma 3.1. *Subject to condition C.2,*

$$\mathcal{E}(\Delta_N) \leq \frac{a\Gamma(q+1/m)}{\Gamma(q)} N^{-1/m} \left(1 + \mathcal{O}\left(\frac{1}{N}\right)\right) \quad \text{as} \quad N \rightarrow \infty \quad (3.2)$$

where a is the constant implied by the positive density condition on the sampling distribution, and Γ denotes the Euler Gamma function.

The following theorem shows that Γ_N is an asymptotically unbiased estimator for $\mathcal{E}(R^q)$ as the number of points $N \rightarrow \infty$.

Theorem 3.2. *Subject to conditions C.1 – C.4,*

$$|\mathcal{E}(\Gamma_N - \mathcal{E}(R^q))| \leq cN^{-1/m} \left(1 + \mathcal{O}\left(\frac{1}{N}\right)\right) \quad \text{as} \quad N \rightarrow \infty \quad (3.3)$$

where

$$c = c(m, q, a, b) = \frac{a\Gamma(q+1/m)}{\Gamma(q)} \sum_{t=1}^q b^t \binom{q}{t} |\mathcal{E}(R^{q-t})| \quad (3.4)$$

where a is the constant implied by the positive density condition on the sampling distribution, b is the constant implied by the Lipschitz condition on the regression function and Γ denotes the Euler Gamma function.

Proof. Let $A = \{1, \dots, q\}$ and let $A(t)$ denote the set of all subsets $B \subset A$ containing exactly t elements. Applying the identity

$$\prod_{k \in A} (a_k + b_k) = \sum_{t=0}^q \sum_{B \in A(t)} \left(\prod_{k \in A \setminus B} a_k \right) \left(\prod_{k \in B} b_k \right) \quad (3.5)$$

we obtain

$$\gamma_i(\mathcal{X}, \mathcal{Y}) = \sum_{t=0}^q \sum_{B \in A(t)} \left(\prod_{k \in A \setminus B} (R_i - R_{i(k)}) \right) \left(\prod_{k \in B} (f(X_i) - f(X_{i(k)})) \right) \quad (3.6)$$

Hence the conditional expectation of $\gamma_i(\mathcal{X}, \mathcal{Y})$ given a fixed realisation of the explanatory sample \mathcal{X} satisfies

$$\mathcal{E}(\gamma_i | \mathcal{X}) = \sum_{t=0}^q \sum_{B \in A(t)} \mathcal{E} \left(\prod_{k \in A \setminus B} (R_i - R_{i(k)}) \middle| \mathcal{X} \right) \prod_{k \in B} (f(X_i) - f(X_{i(k)})) \quad (3.7)$$

Expanding the product $\prod_{k \in A \setminus B} (R - R_{i(k)})$, since $A \setminus B$ contains $q - t$ distinct integers, the first term of the expansion is R_i^{q-t} while the remaining terms are products in which at least one of the $R_{i(k)}$ occurs exactly once. Since the noise variables R_j are independent and identically distributed with $\mathcal{E}(R_j) = 0$, and because they are also independent of the sample \mathcal{X} , the expected value of the first term is equal to $\mathcal{E}(R_i^{q-t})$ and the expected value of each remaining term is zero. Thus it follows that

$$\mathcal{E}(\gamma_i | \mathcal{X}) = \sum_{t=0}^q \mathcal{E}(R_i^{q-t}) \sum_{B \in A(t)} \prod_{k \in B} (f(X_i) - f(X_{i(k)})) \quad (3.8)$$

$$= \mathcal{E}(R^q) + \sum_{t=1}^q \mathcal{E}(R^{q-t}) \sum_{B \in A(t)} \prod_{k \in B} (f(X_i) - f(X_{i(k)})) \quad (3.9)$$

Hence the bias of Γ_N is given by

$$\mathcal{E}(\Gamma_N - \mathcal{E}(R^q)) = \sum_{t=1}^q \mathcal{E}(R^{q-t}) \sum_{B \in A(t)} \mathcal{E}(\Lambda_{B,N}) \quad (3.10)$$

where

$$\Lambda_{B,N} = \frac{1}{N} \sum_{i=1}^M \lambda_{B,i} \quad \text{with} \quad \lambda_{B,i} = \prod_{k \in B} (f(X_i) - f(X_{i(k)})) \quad (3.11)$$

For $t \geq 1$,

$$|\lambda_{B,i}| \leq b^t \prod_{k \in B} \|X_i - X_{i(k)}\| \leq b^t \|X_i - X_{i(q)}\|^t \leq b^t \|X_i - X_{i(q)}\| = b^t \delta_i \quad (3.12)$$

where the first inequality follows by the Lipschitz condition on $f(x)$, and the last by the assumption that $\|X_i - X_{i(q)}\| \leq 1$. Hence for $B \in A(t)$ with $t \geq 1$ it follows that

$$|\Lambda_{B,N}| \leq \frac{1}{N} \sum_{i=1}^N |\lambda_{B,i}| \leq b^t \frac{1}{N} \sum_{i=1}^N \delta_i = b^t \Delta_N \quad (3.13)$$

Thus, because $A(t)$ contains $\binom{q}{t}$ elements it follows by (3.10) and (3.13) that

$$|\mathcal{E}(\Gamma_N - \mathcal{E}(R^q))| \leq \mathcal{E}(\Delta_N) \sum_{t=1}^q b^t \binom{q}{t} |\mathcal{E}(R^{q-t})| \quad (3.14)$$

Hence by Lemma 3.1 we conclude that

$$|\mathcal{E}(\Gamma_N - \mathcal{E}(R^q))| \leq c(m, q, a, b) N^{-1/m} \left(1 + \mathcal{O}\left(\frac{1}{N}\right) \right) \quad \text{as } N \rightarrow \infty \quad (3.15)$$

where

$$c(m, q, a, b) = \frac{a\Gamma(q+1/m)}{\Gamma(q)} \sum_{t=1}^q b^t \binom{q}{t} |\mathcal{E}(R^{q-t})| \quad (3.16)$$

as required. \square

In another paper (Evans 2008) we prove the following result:

Theorem 3.3. *Subject to Conditions C.1 – C.4,*

$$\text{var}(\Gamma_N) \leq \left(\frac{1 + qm2^m}{N} \right) \text{var}(\gamma_i) \quad (3.17)$$

If the first $2q$ moments of the noise distribution are absolutely bounded, $\text{var}(\gamma_i)$ is also bounded. Chebyshev's inequality then yields the following corollary of Theorems 3.2 and 3.3, which shows that Γ_N is a weakly consistent estimator for $\mathcal{E}(R^q)$ as the number of observations $N \rightarrow \infty$.

Corollary 3.4. *Subject to Conditions C.1 – C.4 and the additional requirement that $|\mathcal{E}(R^t)| < \infty$ for all $t = 1, \dots, 2q$,*

$$\Gamma_N = \mathcal{E}(R^q) + \mathcal{O}(N^{-1/m}) + \mathcal{O}_P(N^{-1/2}) \quad \text{as } N \rightarrow \infty \quad (3.18)$$

The first and second error terms in (3.18) correspond to the bias and variance of Γ_N respectively. Theorem 3.2 shows that the bias of Γ_N is dominated by the expected mean nearest neighbour distance $\mathcal{E}(\Delta_N)$, which by Lemma 3.1 is of asymptotic order $\mathcal{O}(N^{-1/m})$ as $N \rightarrow \infty$. Hence the rate at which the bias of Γ_N converges to zero decreases as the dimension of the explanatory vectors increases. In contrast, the rate at which the second error terms in (3.18) converges (in probability) to zero does not depend on the dimension, but only on the fact that $\text{var}(\gamma_i)$ is bounded. An improved rate of convergence is obtained if there exists some $\alpha > 0$ such that $\text{var}(\gamma_i) = \mathcal{O}(N^{-\alpha})$ as $N \rightarrow \infty$, in which case the variance of Γ_N will be of asymptotic order $\mathcal{O}_P(N^{-1/2-\alpha/2})$ as $N \rightarrow \infty$. However, subject to conditions C.1 – C.4 it is easy to show that

$$\text{var}(\gamma_i) = \mathcal{E}(R^{2q}) + q\mathcal{E}(R^2) - \mathcal{E}(R^q)^2 + \mathcal{O}(\delta_i) \quad \text{as } N \rightarrow \infty \quad (3.19)$$

so we cannot improve on the rate at which the variance of Γ_N converges to zero. Instead of using (2.4), it is interesting to speculate whether γ_i can be defined to ensure that both $\mathcal{E}(\gamma_i) \rightarrow \mathcal{E}(R^q)$ and $\text{var}(\gamma_i) \rightarrow 0$ as $N \rightarrow \infty$. In particular, if $\text{var}(\gamma_i)$ can be bounded by some constant multiple of the expected nearest neighbour distance $\mathcal{E}(\delta_i)$, it follows by Lemma 3.1 that the variance of Γ_N will be of asymptotic order $\mathcal{O}_P(N^{-1/2-1/2m})$ as $N \rightarrow \infty$, which would ensure the \sqrt{N} -consistency of Γ_N for all $m \geq 1$ (Munk *et al.*, 2005).

4. Practical Considerations

In practical nonparametric data analysis, the implied constants in the asymptotic expression (3.18) are unknown, and thus we are not able to establish analytic confidence intervals on an estimate computed by Γ_N .

Let (\tilde{x}, \tilde{y}) be a realisation of the random sample $(\mathcal{X}, \mathcal{Y})$, with $\tilde{x} = (x_1, \dots, x_N)$ and $\tilde{y} = (y_1, \dots, y_N)$. For the estimate $\Gamma_N(\tilde{x}, \tilde{y})$, we refer to the errors incurred due to the bias and variance of Γ_M as the *systematic error* and the *statistical error* respectively.

The systematic error associated with $\Gamma_N(\tilde{x}, \tilde{y})$ is due to the variability of the regression function $f(x) = \mathcal{E}(Y|X = x)$ over the q -neighbourhood ball of each

sample point $x_i \in \tilde{x}$ (defined to be the ball centred at x_i and having the q th nearest neighbour $x_{i(q)}$ of x_i on its boundary). By the Lipschitz condition on f , the extent to which $f(x)$ varies over the q -neighbourhood ball of x_i can be bounded in terms of its radius $\delta_i(\tilde{x})$. Hence for a particular realisation (\tilde{x}, \tilde{y}) of the random sample $(\mathcal{X}, \mathcal{Y})$, to reduce the systematic error of our estimate it makes sense to consider only those $\gamma_i(\tilde{x}, \tilde{y})$ for which the associated value of $\delta_i(\tilde{x})$ is small.

Thus we reorder the sequence $(\delta_1(\tilde{x}), \gamma_1(\tilde{x}, \tilde{y})), \dots, (\delta_N(\tilde{x}), \gamma_N(\tilde{x}, \tilde{y}))$ to be increasing with respect to $\delta_i(\tilde{x})$, and define a sequence of estimates

$$\Gamma_M(\tilde{x}, \tilde{y}) = \frac{1}{M} \sum_{i=1}^M \gamma_i(\tilde{x}, \tilde{y}) \quad (1 \leq M \leq N) \quad (4.1)$$

The question now becomes whether or not M can be chosen so that the total error is as small as possible. In the spirit of exploratory statistics, we plot the sequence Γ_M against M and visually assess the behaviour of Γ_M as $M \rightarrow N$. If Γ_M is seen to approach a stable value as M increases, we conclude that there are sufficient data to ensure that the total error is small, and proceed to estimate the moment $\mathcal{E}(R^q)$ by inspection.

To reduce systematic error, it may be that the $(q+1)$ -neighbourhood ball of a point x_i is significantly smaller than the q -neighbourhood ball of another point x_j , in which case we might expect that the q -fold product $\gamma'_i = \prod_{k=2}^{q+1} (y_i - y_{i(k)})$ will incur less systematic error than $\gamma_j = \prod_{k=1}^q (y_j - y_{j(k)})$. However, replacing γ_j by γ'_i does not necessarily lead to a better estimate, because the empirical noise realisations corresponding to γ_i and γ'_i are estimated by almost identical products of noise differences. Consequently, the estimator suffers a loss of information regarding the noise distribution, and statistical error increases accordingly. Systematic error might also be reduced by attempting to estimate the limit of Γ_N as $\Delta_N \rightarrow 0$ (Evans & Jones 2002; Tong & Wang 2005). Thus for $k = 1, \dots, p$ we might define

$$\gamma_{i,k} = \prod_{\ell=k}^{k+q-1} (Y_i - Y_{i(\ell)}) \quad \text{and} \quad \delta_{i,k} = \|X_i - X_{i(k+q-1)}\| \quad (4.2)$$

along with the associated sample means $\Gamma_N(k)$ and $\Delta_N(k)$. The limit of the $\Gamma_N(k)$ as $\Delta_N(k) \rightarrow 0$ can then be estimated using simple linear regression on the pairs $(\Delta_N(k), \Gamma_N(k))$, which yields the estimator

$$\tilde{\Gamma}_N = \arg_a \min_{a,b} \left(\sum_{k=1}^p (\Gamma_N(k) - (a + b\Delta_N(k)))^2 \right) \quad \text{where} \quad a, b \in \mathbb{R} \quad (4.3)$$

The success of this approach depends on the nearest neighbour distances $\delta_{i,k}$ being sufficiently small to ensure that the systematic error scales approximately linearly with $\Delta_N(k)$. Empirical results indicate that $\tilde{\Gamma}_N$ does not offer any significant advantages over Γ_N .

(a) Covariance

Turning our attention to the covariance estimator Γ_N^{cov} , for a particular sample realisation (\tilde{x}, \tilde{y}) we reorder the sequence $(\delta_1(\tilde{x}), \gamma_1^{\text{cov}}(\tilde{x}, \tilde{y})), \dots, (\delta_N(\tilde{x}), \gamma_N^{\text{cov}}(\tilde{x}, \tilde{y}))$ to be increasing with respect to $\delta_i(\tilde{x})$, and define a sequence of estimates

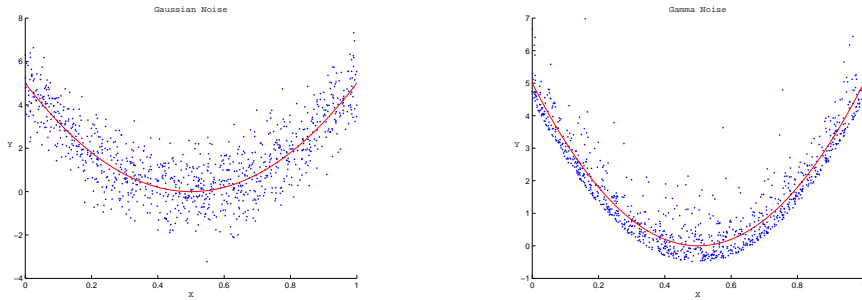


Figure 1. A sample realisation (\tilde{x}, \tilde{y}) for the Gaussian noise experiment (left) and the Gamma noise experiment (right), with the regression function $f(x) = 20(x - 1/2)^2$ shown in red.

$$\Gamma_M^{\text{cov}}(\tilde{x}, \tilde{y}) = \frac{1}{M} \sum_{i=1}^M \gamma_i^{\text{cov}}(\tilde{x}, \tilde{y}) \in \mathbb{R}^{n \times n} \quad (1 \leq M \leq N) \quad (4.4)$$

For each pair of coordinates $1 \leq \alpha, \beta \leq n$, we plot the corresponding component sequence $\Gamma_M^{\text{cov}}(\alpha, \beta)$ against M , then visually assess its behaviour as $M \rightarrow N$. If the curve appears to approach a stable value as $M \rightarrow N$, we proceed to estimate the component by inspection.

5. Experimental Results

(a) Moment estimation

To investigate our moment estimators, we generate samples (\tilde{x}, \tilde{y}) of $N = 1000$ observation pairs $(x_i, y_i) \in \mathbb{R} \times \mathbb{R}$ where

- The regression function is defined to be $f(x) = 20(x - 1/2)^2$.
- The explanatory samples x_i are selected independently according to the uniform distribution over the unit interval $[0, 1]$.
- The residual samples r_i are selected independently according to
 - (1) the standard Gaussian distribution $N(0, 1)$, and
 - (2) the Gamma distribution, normalised to have mean zero and variance 0.25.
- The response samples y_i are set to equal $f(x_i) + r_i$.

Figure 1 shows a single sample realisation for the Gaussian noise experiment (left) and the Gamma noise experiment (right), while Table 1 shows the estimated moments $\Gamma_N^{(q)}$ and the empirical moments $\langle r_i^q \rangle$ for the sample realisations shown in Figure 1.

In Figure 2, we plot the estimate sequences $\Gamma_M^{(q)}$ for the sample realisation of the Gaussian noise experiment illustrated in Figure 1, with the empirical moments $\langle r_i^q \rangle$ shown in green ($q = 1, 2, 3, 4$). It is evident from the figures that the final

q	Gaussian Noise		Gamma Noise	
	$\langle r_i^q \rangle$	$\Gamma_N^{(q)}$	$\langle r_i^q \rangle$	$\Gamma_N^{(q)}$
1	0.0486	0.0097	0.0153	0.0078
2	1.0050	1.0105	0.2888	0.2914
3	0.1295	-0.0743	0.3899	0.4069
4	2.8960	2.5537	1.1424	1.2307

Table 1. Empirical moments $\langle r_i^q \rangle$ and estimated moments $\Gamma_N^{(q)}$ for one sample realisation of both the Gaussian noise and Gamma noise experiments.

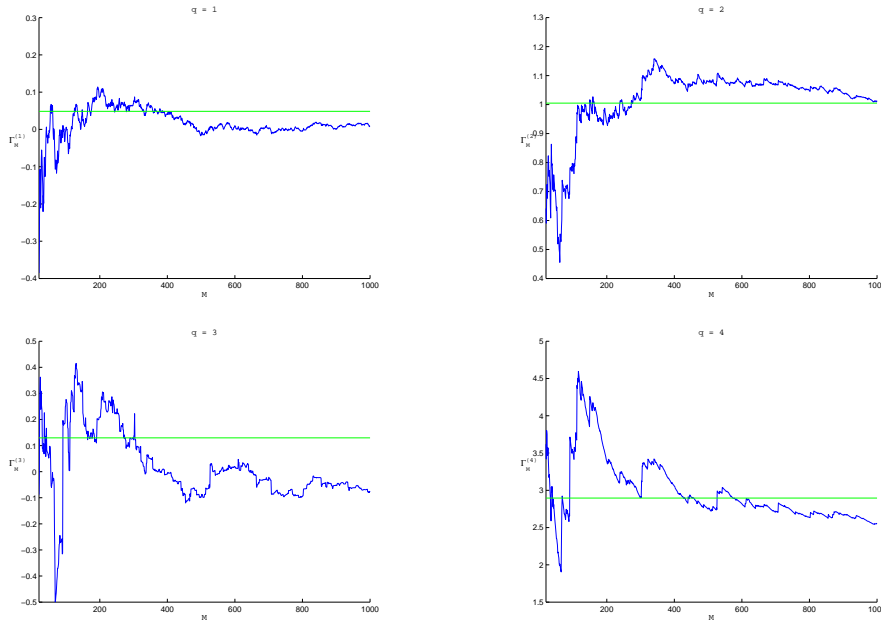


Figure 2. Visual representation of the estimate sequences $\Gamma_M^{(q)}$ for a single run of the Gaussian noise experiment, with the associated empirical moments $\langle r_i^q \rangle$ shown in green ($q = 1, 2, 3, 4$).

values $\Gamma_N^{(q)}$ are likely to provide reasonable estimates for the corresponding moments of the noise distribution. Similarly, in Figure 3 we plot the estimate sequences $\Gamma_M^{(q)}$ for the sample realisation of the Gamma noise experiment illustrated in Figure 1, with the empirical moments $\langle r_i^q \rangle$ shown in green ($q = 1, 2, 3, 4$). In this case the plots are more ‘jagged’ than those for the Gaussian noise experiment shown in Figure 2. This is due to the heavy tail of the Gamma distribution, which produces outliers that can be seen in Figure 1. These ‘jagged’ features become increasingly apparent as q increases, illustrating the well-known fact that sample estimates of higher order moments are negatively affected by outliers. In spite of this, it appears that the curves stabilise as $M \rightarrow N$, so that the final values $\Gamma_N^{(q)}$ can be adopted as reasonable estimates for the corresponding moments of the noise distribution.

To investigate the variance of our moment estimators, we perform 100 repetitions of both the Gaussian noise and Gamma noise experiments. In Figure 4 we

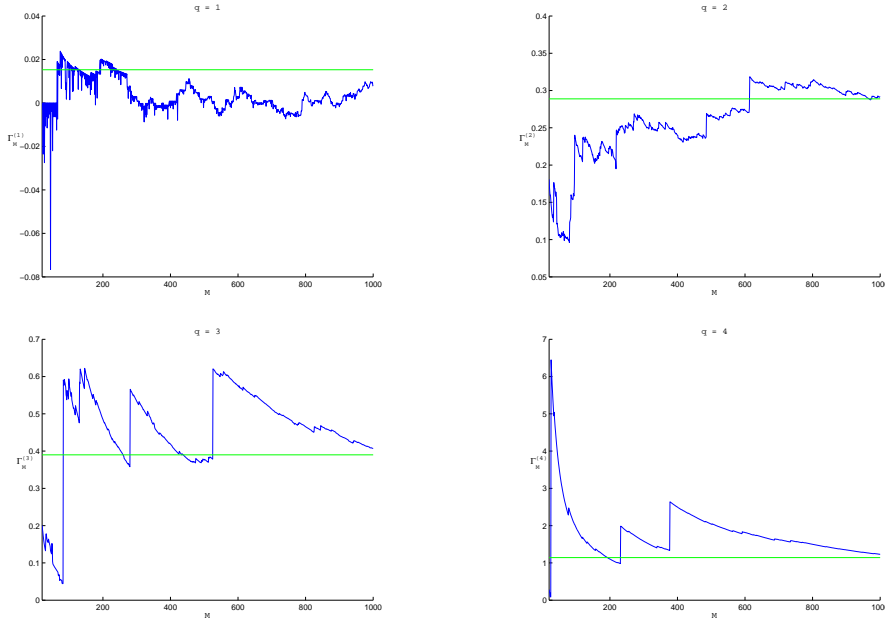


Figure 3. Plot of the estimate sequences $\Gamma_M^{(q)}$ for a single run of the Gamma noise experiment, with the associated empirical moments $\langle r_i^q \rangle$ shown in green ($q = 1, 2, 3, 4$).

plot the mean absolute estimation error of the sequences $\Gamma_M^{(q)}$ for the Gaussian noise experiment, along with error bars representing one standard error of the mean over these repetitions. Figure 5 shows similar plots for the Gamma noise experiment.

(b) Covariance estimation

To investigate our covariance estimators, we generate samples (\tilde{x}, \tilde{y}) of $N = 1000$ observation pairs $(x_i, y_i) \in \mathbb{R}^2 \times \mathbb{R}^2$ where

- The regression function is defined to be $f(x_1, x_2) = 20(x_1 - 1/2)^2 + 20(x_2 - 1/2)^2$.
- The explanatory samples x_i are selected independently according to the uniform distribution over the unit square $[0, 1]^2$.
- The residual samples r_i are selected independently according to the multivariate Gaussian distribution $N(0, \Sigma)$ where

$$\Sigma = \begin{pmatrix} 2 & 1 \\ 1 & 3 \end{pmatrix}$$

- The response samples y_i are set to equal $f(x_i) + r_i$.

For a single run of this experiment, Figure 6 shows the estimate sequences $\Gamma_M^{\text{cov}}(\alpha, \beta)$, along with the associated empirical covariances $\langle r_{i\alpha} r_{i\beta} \rangle$ shown in green ($1 \leq \alpha, \beta \leq 2$). It is evident from the plots that the final values $\Gamma_N^{\text{cov}}(\alpha, \beta)$ are likely to provide reasonable estimates for the corresponding residual covariances.

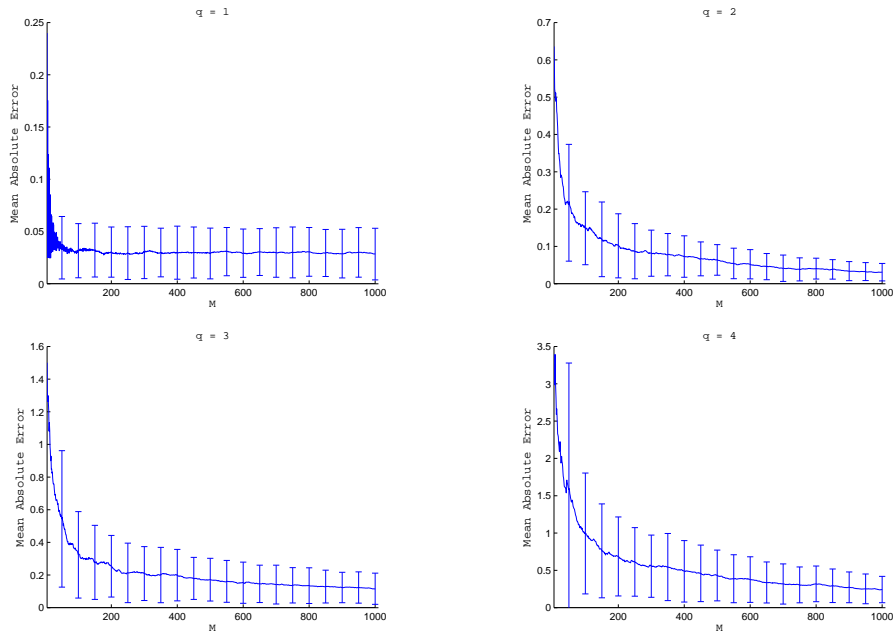


Figure 4. Mean absolute estimation error of the sequences $\Gamma_M^{(q)}$ over 100 repetitions of the Gaussian noise experiment. The error bars represent one standard error of the mean over these repetitions.

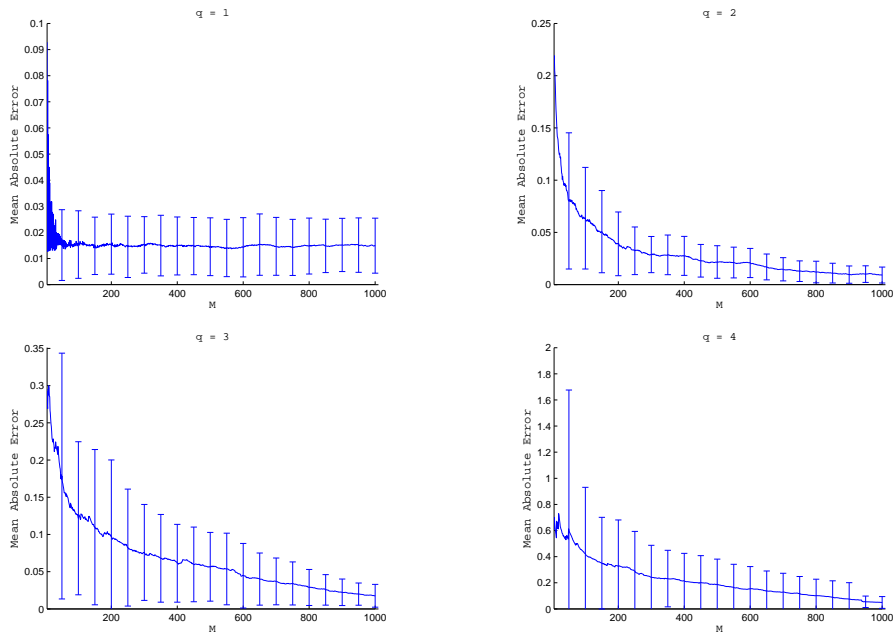


Figure 5. Mean absolute estimation error of the sequences $\Gamma_M^{(q)}$ over 100 repetitions of the Gamma noise experiment. The error bars represent one standard error of the mean over these repetitions.

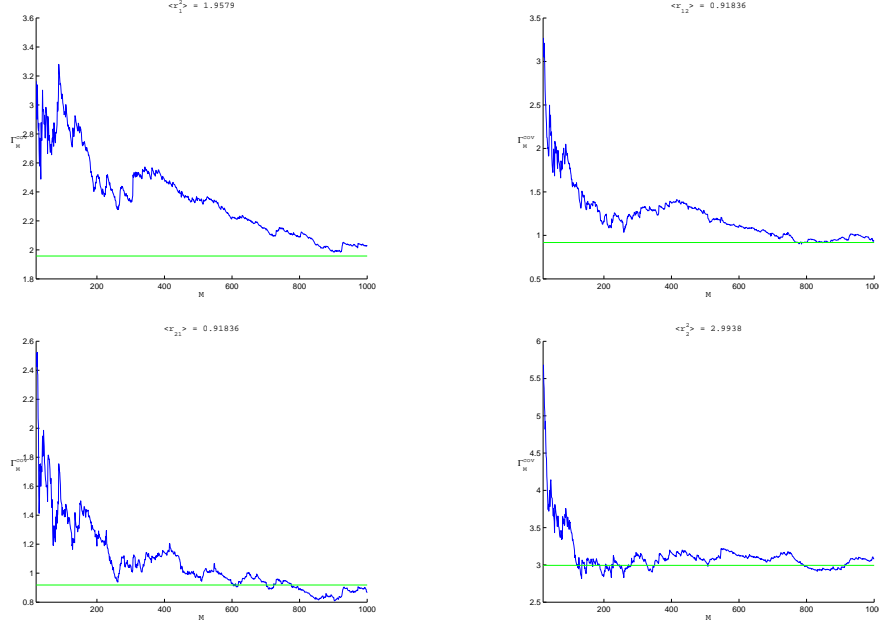


Figure 6. Visual representation of the estimate sequences $\Gamma_M^{\text{cov}}(\alpha, \beta)$ for a single run of the covariance estimation experiment, with the associated empirical covariances $\langle r_{i\alpha} r_{i\beta} \rangle$ shown in green ($1 \leq \alpha, \beta \leq 2$).

The empirical residual covariance matrix and the estimated residual covariance matrix are shown in (5.1). Because covariance matrices are symmetric by definition, we should replace $\Gamma_N^{\text{cov}}(1, 2)$ and $\Gamma_N^{\text{cov}}(2, 1)$ by their mean value, which in this case would lead to an improved estimator for $\langle r_{i1} r_{i2} \rangle$.

$$\langle r_i r_i^T \rangle = \begin{pmatrix} 1.9579 & 0.9184 \\ 0.9184 & 2.9938 \end{pmatrix} \quad \Gamma_N^{\text{cov}} = \begin{pmatrix} 2.0257 & 0.9393 \\ 0.8654 & 3.0833 \end{pmatrix} \quad (5.1)$$

(c) Smoothing windows

For scalar data $(x_i, y_i) \in \mathbb{R} \times \mathbb{R}$, one of the simplest nonparametric modelling techniques is to apply a *smoothing window* to the data, where the regression function value $f(x_i)$ at x_i is estimated simply by the average of the associated point y_i and a certain number of its nearest neighbours,

$$\hat{f}_W(x_i) = \frac{1}{2W+1} \sum_{j=i-W}^{i+W} y_j \quad (5.2)$$

where the *window size* W controls the amount of smoothing applied to the data. Consider the empirical error function associated with a particular value of W :

$$\hat{e}_W = \langle y_i - \hat{f}_W(x_i) \rangle \quad (5.3)$$

In view of (1.9), we define the optimal window size to be that for which the moments of the empirical error function \hat{e}_W best match the estimated moments $\Gamma_N^{(q)}$ of the

	Gaussian noise	Gamma noise
$\hat{W}_{\text{opt}}^{(2)}$	0.0434	0.0118
$\hat{W}_{\text{opt}}^{(3)}$	0.0433	0.0105
$\hat{W}_{\text{opt}}^{(4)}$	0.0403	0.0224

Table 2. The average mean squared error of the models $\hat{f}_{W_{\text{opt}}^{(q)}}$ over 100 repetitions of the Gaussian noise and Gamma noise experiments.

noise distribution. In fact, for $q = 2, 3$ and 4 we propose the following notions of what constitutes an optimal window size:

$$\hat{W}_{\text{opt}}^{(q)} = \arg \min_{1 \leq W \leq W_{\text{max}}} \sum_{k=2}^q | \langle \hat{e}_W^k \rangle - \Gamma_N^{(k)} | \quad (5.4)$$

In Table 2 we show the average mean squared estimation error of the models $\hat{f}_{W_{\text{opt}}^{(q)}}(x_i)$, computed over 100 repetitions of both the Gaussian noise experiment and the Gamma noise experiment. For Gaussian noise, the table shows that improved models can be obtained using $\hat{W}_{\text{opt}}^{(4)}$ instead of $\hat{W}_{\text{opt}}^{(2)}$. The results also show that $\hat{W}_{\text{opt}}^{(3)}$ offers no advantage over $\hat{W}_{\text{opt}}^{(2)}$, which is reasonable in view of the fact that the Gaussian distribution is symmetric and therefore has third moment equal to zero. In contrast, the table shows that $\hat{W}_{\text{opt}}^{(3)}$ does offer an advantage over $\hat{W}_{\text{opt}}^{(2)}$ for the Gamma noise distribution. This is also reasonable since the Gamma distribution is non-symmetric, and therefore has non-zero third moment. While the advantage over $\hat{W}_{\text{opt}}^{(2)}$ of $\hat{W}_{\text{opt}}^{(4)}$ for Gaussian noise and $\hat{W}_{\text{opt}}^{(3)}$ for Gamma noise appears to be very slight, it must be remembered that smoothing windows are rather crude nonparametric regression methods, and it may be that more significant advantages can be achieved with more sophisticated modelling techniques.

6. Conclusion

We have proposed new estimators of residual moments and covariance for nonparametric data analysis. Standardised moments (skewness, kurtosis) and correlation can be estimated by normalising with respect to the appropriate standard deviation estimates. The estimators are computationally efficient, with running time of order $\mathcal{O}(N \log N)$ as the number of data points $N \rightarrow \infty$. We have shown that our moment estimators are asymptotically unbiased and weakly consistent as $N \rightarrow \infty$.

In practical applications, the sample mean sequence Γ_M is plotted to see whether it approaches a stable value as M increases. If this occurs, we conclude that there are sufficient data to ensure that the estimation error is small, and proceed to estimate the moment or covariance component by inspection. If Γ_M does not appear to approach a stable value as $M \rightarrow N$, we must conclude that we have insufficient data to allow accurate noise estimates using the methods described in this paper.

Dafydd Evans would like to thank the Royal Society for supporting his research through its University Research Fellowship scheme.

References

- Bentley, J. L. 1975 Multidimensional binary search trees used for associative search. *Comm. ACM* **18**, 309–517.
- Bock M., Bowman A. W. & Ismail B. 2007 Estimation and inference for error variance in bivariate nonparametric regression. *Stat. Comput.* **17**, 39–47.
- Brown L. D. & Levine M. 2007 Variance estimation in nonparametric regression via the difference sequence method. *Ann. Stat.* **35**, 2219–2232.
- Donoho, D. L. & Johnstone, I. M. 1995 Adapting to unknown smoothness via wavelet shrinkage. *J. Am. Stat. Assoc.* **90**(432), 1200–1224.
- Durrant, P. J. 2001 *winGamma: a non-linear data analysis and modelling tool with applications to flood prediction*, PhD Thesis, Cardiff University.
- Eubank, R. L. 1999 *Nonparametric Regression and Spline Smoothing*. Marcel Dekker Ltd.
- Evans, D. 2008 A Law of Large Numbers for Nearest Neighbour Statistics. Preprint, Cardiff University.
- Evans, D. & Jones, A. J. 2002 A proof of the Gamma test. *Proc. R. Soc. A* **458**(2027), 2759–2799.
- Evans, D., Jones, A. J. & Schmidt W. M. 2002 Asymptotic moments of near neighbour distance distributions. *Proc. R. Soc. A* **458**(2028), 2839–2849.
- Gasser, T., Sroka, L. & Jennen-Steinmetz, C. 1986 Residual variance and residual pattern in nonlinear regression. *Biometrika* **73**(3), 625–633.
- Gruber, P. M. 2004 Optimum quantization and its applications. *Adv. Math.* **186**, 456–497.
- Hall, P., Kay, J. W. & Titterington, D.M. 1990 Asymptotically optimal difference-based estimation of variance in nonparametric regression. *Biometrika*, **77**, 521–528.
- Jones, A. J. 2004 New tools in non-linear modelling and prediction. *Comput. Man. Sci.* **1**, 109–149.
- Levine M. 2006 Bandwidth selection for a class of difference-based variance estimators in the nonparametric regression: A possible approach. *Comput. Stat. Data An.* **50**, 3405–3431.
- Munk, A., Bissantz, N., Wagner, T. & Freitag, G. 2005 On difference-based variance estimation in nonparametric regression when the covariate is high dimensional. *J. Roy. Stat. Soc. B* **67**, 19–41.
- Rice, J. 1984 Bandwidth choice for nonparametric kernel regression. *Ann. Stat.* **12**, 1215–1230.
- Silverman, B. W. 1986 *Density Estimation for Statistics and Data Analysis*. Chapman and Hall.
- von Neumann, J. 1941 Distribution of the ratio of the mean squared successive difference to the variance. *Ann. Math. Stat.* **12**, 367–395.
- Tong, T. & Wang, Y. 2005 Estimating residual variance in nonparametric regression using least squares. *Biometrika* **92**(4), 821–830.
- Wang L., Brown L. D., Cai T. T. & Levine, M. 2008 Effect of mean on variance function estimation in nonparametric regression. *Ann. Stat.* **36**, 646–664.