Statistica Sinica **9**(1999), 571-594

DOMAIN SPLITTING IN REGRESSION ANALYSIS

Hans-Georg Müller and Peng-Liang Zhao

University of California, Davis and Merck Research Laboratories

Abstract: We introduce Domain Splitting as a new tool for regression analysis. This device corresponds to splitting the domain of a regression function into m subdomains, where m is varied, and fitting a linear model on each subdomain. The residual sums of squares from these various fits are compared graphically. Domain Splitting provides a visual diagnostic, as well as a model-independent estimate of the error variance. We investigate the asymptotic behavior of Domain Splitting for the cases of an underlying linear model and that of a smooth regression function. The asymptotic findings are illustrated in simulations and examples.

Key words and phrases: Diagnostic plot, goodness-of-fit, linear model, model selection, smooth regression, variance estimation.

1. Introduction

Consider the fixed design regression model

$$y_i = g(x_i) + \varepsilon_i,\ i = 1, \ldots, n,$$

with a "smooth" regression function g on $\Re^p$, $p \geq 1$, and i.i.d. errors ε_i. If a linear model provides a reasonable fit to the data, say $g(x) = \alpha + \beta^T x$, this would usually be preferred over the alternative, a regression function which is only smooth. In the latter case, one has to resort to nonparametric regression procedures and to deal with the associated problems of bias assessment and bandwidth choice. This is particularly cumbersome in higher dimensions.

A related question is how to find a good estimate of the variance of the errors. Such variance estimates are needed for such tasks as the construction of confidence regions, model-based tests, model selection procedures, and signal-to-noise ratio determination. Some of these applications are related to estimating the "functional correlation", $\rho^2 = 1 - \mathrm{Var}\,(\varepsilon)/\{\mathrm{Var}\,(\varepsilon) + \int [g(x) - \int g(y)dy]^2 dx\}$, for the case of equidistant data. Having found a good estimator $\hat{\sigma}^2$ of $\mathrm{Var}\,(\varepsilon_i)$, and using the sample variance s_y^2 of the data $y_1, \ldots, y_n$, one may then estimate ρ^2 for regular designs by $\hat{\rho}^2 = 1 - \hat{\sigma}^2/s_y^2$. The classical variance estimate under

the linear model assumption is the Mean Square due to Error (MSE). Given least squares estimates $\hat{\alpha}$ and $\hat{\beta}$ of the parameters α, β, the MSE is defined as

$$MSE = \frac{1}{n-(p+1)} \sum_{i=1}^{n} e_i^2,$$

where $e_i = y_i - (\hat{\alpha} + \hat{\beta}^T x_i)$ are the residuals. This is a standard output of regression routines.

On the other hand, for smooth regression models the MSE is biased as a variance estimator. Alternative variance estimators for the smooth regression case, based on difference schemes, were proposed by Rice (1984). This approach was further developed by Gasser, Sroka and Jennen-Steinmetz (1986), Hall, Kay and Titterington (1990), and Müller and Stadtmüller (1987, 1993), among others. MSE's derived from local nonparametric fits have been considered by Buckley, Eagleson and Silverman (1988). Another proposal by Breiman and Meisel (1976), to fit local planes on subdomains for obtaining estimates of σ^2, is also of interest in this context and will be further discussed and contrasted with our proposal in Section 6. Other related approaches are piecewise polynomial regression trees and the SUPPORT algorithm (Chaudhuri, Huang, Loh and Yao (1994)); the emphasis in the latter is, however, on multidimensional regression modelling and less on diagnostics and error variance estimation as in this paper. We also note connections to the piecewise fitting of step functions (Yao and Au (1989)), although the choice of the locations of the jumps is not an issue here, as jump locations are chosen equidistantly by design.

The proposed Domain Splitting method provides an alternative approach to variance estimation. Compelling results on the use of the Domain Splitting concept for the purpose of variance estimation can be found in Section 5. A second goal is a visual diagnostic assessment of the model fit of a supposed parametric model such as a linear model. Visual checks for goodness-of-fit have retained their popularity, because they allow one to gather useful information about the nature of possible deviations from an assumed model, compare for instance Mansfield and Conerly (1987). A p-value obtained from a formal inference based goodness-of-fit test alone (compare, e.g., Eubank and Spiegelman (1990), among others) is often less informative.

The basic idea of Domain Splitting can be briefly summarized as follows: split the domain on which the regression data are collected into an increasing number of smaller subdomains. An assumed parametric model is fitted on each of the subdomains, including the original global domain. For each of these fits,

one obtains the MSE as defined above, but calculated separately for each subdomain. Plotting these various MSE's versus the number (size) of the corresponding subdomains then provides the visualization of Domain Splitting. We call this the Domain Splitting Plot. Biases will inflate the MSE's for large subdomains, but less so for small subdomains, if the supposed model does not fit. Bias corresponding to lack of fit will then reveal itself as a downwards trend in the Domain Splitting Plot as the number of subdomains increases.

A formal definition of Domain Splitting is given in the next section (Section 2), and two simple simulation examples are discussed. Asymptotic properties of these plots, which aid in their interpretation, are provided in Section 3. The application of Domain Splitting for variance estimation is explored in Section 5, and simulation results are reported there. Section 4 is devoted to a discussion of practical aspects of the multivariate case. We conclude with a discussion of some pertinent issues and extensions in Section 6. Details of proofs and auxiliary results are compiled in Appendix A, and the extension to the multivariate case can be found in Appendix B.

2. The Domain Splitting Principle

Assume we observe data generated by a fixed design regression model

$$y_i = g(x_i) + \varepsilon_i, \quad 1 \le i \le n. \tag{M0}$$

We consider here the one-dimensional case with $x_i \in \Re$. The domain of the model is defined as the range of the x_i. We assume that the measurement grid $x_i = x_{i,n}, 1 \le i \le n$, is generated by a regular design density in the sense of Sacks and Ylvisaker (1970), and so need not be equidistant. More precisely,

(M1) There exists a distribution function F with compact support D, three times continuously differentiable on its support, such that $x_i = x_{i,n} = F^{-1}((i-1)/(n-1))$, $1 \le i \le n$.

Without loss of generality, we take $D = [0, 1]$ and then require that the *design density* $f = F'$ satisfies

$$f(x) > 0 \text{ for } x \in int(D). \tag{M2}$$

The "smoothness" of the regression function g is measured by its differentiability:

(M3) The regression function g is twice continuously differentiable on the domain D.

For the errors $\varepsilon_i = \varepsilon_{i,n}$, we need the existence of fourth moments. Some of our results include the case of heteroscedastic errors, while for other results, homoscedasticity is needed. More specifically, we require that

(M4) The errors $\varepsilon_i = \varepsilon_{i,n}$, $1 \le i \le n$, are independent. There exist functions $\sigma^2(x), \mu_3(x)$ and $\mu_4(x)$, twice continuously differentiable on D, such that

$$E(\varepsilon_{i,n}) = 0, E(\varepsilon_{i,n}^2) = \sigma^2(x_{i,n}), E(\varepsilon_{i,n}^3) = \mu_3(x_{i,n}), E(\varepsilon_{i,n}^4) = \mu_4(x_{i,n}).$$

This condition allows us to include the heteroscedastic case. In this case, the variance target is the average variance

$$\sigma^2 = \int_D \sigma^2(x) f(x) dx,$$

f being the design density. This simplifies to the usual error variance $\sigma^2 = \mathrm{Var}\,(\varepsilon_i)$ for the homoscedastic case. For some of the following results we will invoke the more restrictive condition:

(M5) The errors $\varepsilon_i = \varepsilon_{i,n}$ are independent and identically distributed, with $E(\varepsilon_{i,n}) = 0, E(\varepsilon_{i,n}^2) = \sigma^2, E(\varepsilon_{i,n}^3) = 0$ and $E(\varepsilon_{i,n}^4) = 3\sigma^4$.

The simple linear regression model, which we use to illustrate Domain Splitting, is

(L) There exist constants α, β such that $g(x) = \alpha + \beta x$, $x \in D$.

We will omit indices n whenever feasible to simplify the notation. Consider $m = 1, 2, \ldots, [\frac{n}{3}]$, where $[x]$ is the largest integer smaller or equal to x. Define a triangular scheme of left open and right closed *domain splitting intervals* or *subdomains* $D_{jm} = (F^{-1}((j-1)/m), F^{-1}(j/m)]$, $1 < j \le m, 1 \le m \le [\frac{n}{3}]$, $D_{1m} = [F^{-1}(0), F^{-1}(m^{-1})]$, so that always $D_{11} = [0,1] = D$. Note that the important special case of an equidistant regression design is characterized by $F(x) = x$. In this case, $D_{jm} = ((j-1)/m, j/m]$, $j = 1, \ldots, m$.

In general, $x_i \in D_{jm}$ if and only if $\frac{j-1}{m} < \frac{i-1}{n-1} \le \frac{j}{m}$. Define *subdomain centers* s_{jm} for the subdomains D_{jm} by

$$s_{jm} = n_{jm}^{-1} \sum_{x_i \in D_{jm}} x_i, \quad 1 \le j \le m, \quad 1 \le m \le [\frac{n}{3}],$$

where

$$n_{jm} = \sum_{x_i \in D_{jm}} 1$$

is the number of x_i falling into D_{jm}.

Now fit the subdomain simple regression models

$$g_{jm}(x) = \alpha_{jm} + \beta_{jm}(x - s_{jm}), \quad x \in D_{jm}, \quad 1 \le j \le m, 1 \le m \le [\frac{n}{3}] \tag{2.1}$$

to the data on each subdomain, by ordinary least squares.

Setting

$$SS_{jm}(\alpha, \beta) = \sum_{x_i \in D_{jm}} \left\{ Y_i - [\alpha + \beta(x_i - s_{jm})] \right\}^2, \tag{2.2}$$

one has the *subdomain least squares estimates*

$$(\hat{\alpha}_{jm}, \hat{\beta}_{jm}) = \arg\min_{\alpha,\beta} SS_{jm}(\alpha, \beta). \tag{2.3}$$

Define the *subdomain error mean squares*

$$\hat{\sigma}^2_{jm} = MSE_{jm} = (df_{jm})^{-1} SS_{jm}(\hat{\alpha}_{jm}, \hat{\beta}_{jm}), \quad 1 \le j \le m, \tag{2.4}$$

where $df_{jm} = n_{jm} - 2$ are the error degrees of freedom on each subdomain. These are the MSE's over each subdomain when splitting the domain into m subdomains.

We define the *average subdomain error mean squares* by

$$\hat{\sigma}^2_m = MSE_m = \sum_{j=1}^{m} \left(df_{jm} \hat{\sigma}^2_{jm} \right) / (n - 2m), \tag{2.5}$$

Notice that $\sum_{j=1}^{m} df_{jm} = n - 2m$.

The results obtained by splitting the domain and fitting the models on the subdomains can be visually summarized in the Domain Splitting Plot.

(DSP) The *Domain Splitting Plot* (DSP) consists of two parts:

(a) A plot formed by the $\frac{1}{2}[\frac{n}{3}]([\frac{n}{3}] + 1)$ points (x, y) defined by their x-and y-coordinates

$$\left(m, \hat{\sigma}^2_{1m}\right), \ldots, \left(m, \hat{\sigma}^2_{mm}\right), \quad 1 \le m \le [\frac{n}{3}].$$

(b) A graph consisting of lines connecting the points $(m, \hat{\sigma}^2_m)$, $1 \le m \le [\frac{n}{3}]$. The two parts are shown simultaneously on one plot.

To demonstrate the concept of the Domain Splitting Plot, we discuss two simple examples. Both are for simulated regression data with $y_i = g(x_i) + \varepsilon_i$, $i = 1, \ldots, 100$, where the $\varepsilon_i \sim 0.5^* \mathcal{N}(0, 1)$. The design density is $f(x) = \frac{1}{2} 1_{[-1,1]}(x)$,

i.e., the locations of the observations are spaced equidistantly on the domain $D = [-1, 1]$.

Example 1. The regression function is $g_1(x) = x$. The proposed Domain Splitting Plot for these data is in Figure 1. The maximal number of possible subdomains is here $[\frac{n}{3}] = 33$. In this case we know that the best fit is obtained by fitting one global regression line, corresponding to the choice $m = 1$.

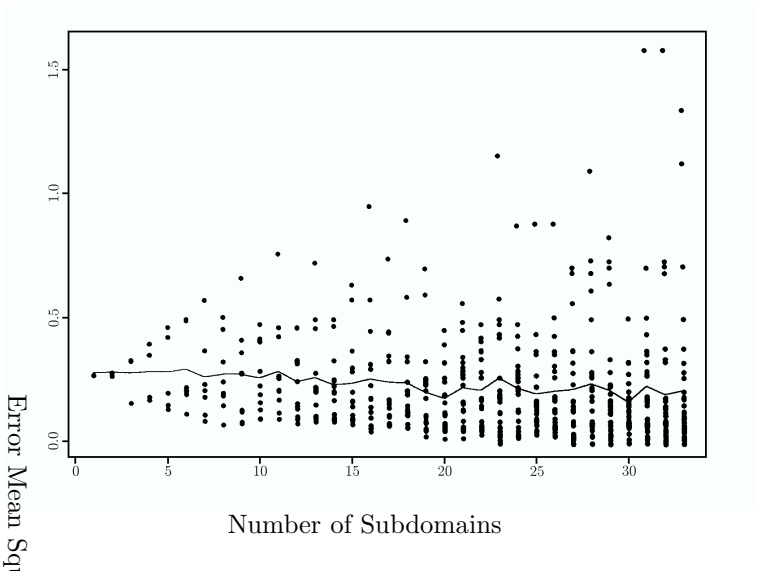

Figure 1. Domain Splitting Plot for 100 simulated data points generated from the simple linear model $y_i = x_i + 0.5^* \mathcal{N}(0, 1)$ (Example 1).

The dots in the scatterplot of Figure 1 correspond to the various observed values of $\hat{\sigma}^2_{jm} = MSE_{jm}$ (subdomain error mean squares, see (2.4)) obtained over the subdomains. They provide an indication of the degree of variability of the individual $\hat{\sigma}^2_{jm}$ values.

The average subdomain error mean squares $\hat{\sigma}^2_m$ are connected by the solid graph in Figure 1. The graph appears fairly flat, with increasing random fluctuation towards larger values for m starting from $m \approx 10$. This behavior of the graph in the Domain Splitting Plot is typical for cases where the assumed regression model provides an adequate global fit to the data. We observe that, in this case, all of the $\hat{\sigma}^2_m{}'s$ are unbiased estimators of the error variance σ^2. Consequently, the graph in the Domain Splitting Plot will not show any systematic trend. However, the variation in the $\hat{\sigma}^2_{jm}{}'s$ increases with m, as for larger m the individual $\hat{\sigma}^2_{jm}{}'s$ are based on shorter spans of data. This increased variation of

individual $\hat{\sigma}^2_{jm}{}'s$ translates into increased variation of the averaged $\hat{\sigma}^2_m{}'s$, as m increases.

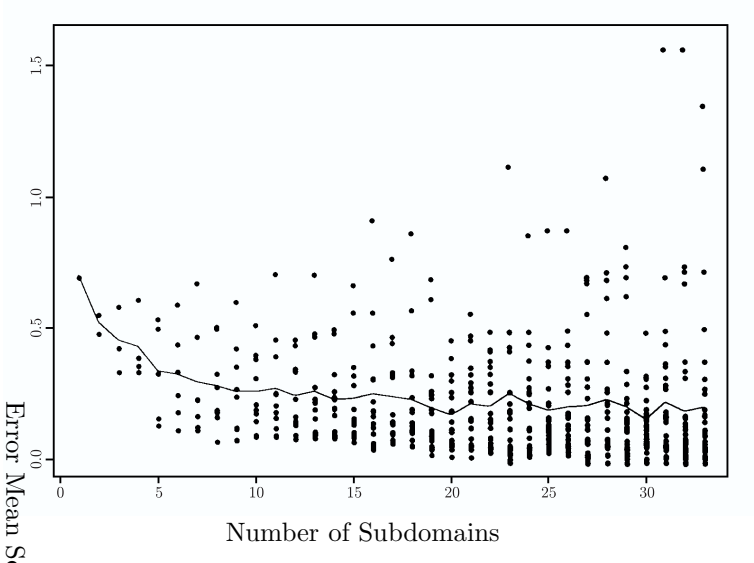

Figure 2. Domain Splitting Plot for 100 simulated data points generated from the model $y_i = \sin(2\pi x_i) + 0.5^*\mathcal{N}(0,1)$ (Example 2).

Example 2. Here the regression function is the nonlinear function $g_2(x) = \sin(2\pi x)$. Figure 2 displays the Domain Splitting Plot . The lack of fit of a simple linear regression function is revealed in the Domain Splitting Plot of Figure 2 by a strongly declining initial trend in the solid graph connecting the average subdomain error mean squares for m between 1 and 10, say. This trend reflects the bias in the $\hat{\sigma}^2_m{}'s$ when we employ them as estimates of σ^2 for small values of m. This bias arises as the residual sums of squares pick up a component measuring the distance between the true regression function and its projection on the assumed parametric function, in addition to the sum of squared errors. As m increases this bias becomes negligible and the graph flattens out towards the right, as seen in Figure 1. For values of m larger than about 17, the graph is dominated by random fluctuations.

We demonstrate next that the behavior of the Domain Splitting Plots in Examples 1 and 2 is in accordance with asymptotic predictions.

3. Asymptotic Behavior

We note that the least squares estimates $\hat{\alpha}_{jm}, \hat{\beta}_{jm}$ for the coefficients α_{jm}, β_{jm} of the subdomain regression models are given as solutions of the normal

equations

$$\hat{\alpha}_{jm} = n_{jm}^{-1} \sum_{x_i \in D_{jm}} y_i, \quad \hat{\beta}_{jm} = \sum_{x_i \in D_{jm}} y_i(x_i - s_{jm})/v_{jm},$$

where

$$n_{jm} = \sum_{x_i \in D_{jm}} 1, \quad v_{jm} = \sum_{x_i \in D_{jm}} (x_i - s_{jm})^2.$$

The asymptotic behavior of these solutions and of average subdomain error mean squares $\hat{\sigma}_m^2$ (2.5) which are displayed in the Domain Splitting Plot graph (DSP graph) as a function of m is analyzed in detail in the Appendix.

We investigate the asymptotic trend of the DSP graph by means of a bias approximation (Theorem 1), the behavior of $E(\hat{\sigma}_m^2)$ (Theorem 2) and of $\mathrm{Var}\,(\hat{\sigma}_m^2)$ (Theorem 3). The bias contained in $\hat{\sigma}_m^2$ as estimator of σ^2, where $\sigma^2 = \mathrm{Var}\,(\varepsilon_i)$ in the homoscedastic case, and $\sigma^2 = \int_D \sigma^2(x) f(x) dx$ in the heteroscedastic case, needs to be quantified. For this purpose, define

$$B(m,n) = \frac{1}{n} \sum_{j=1}^{m} \Big\{ \min_{\alpha_j, \beta_j} \sum_{x_i \in D_{jm}} [g(x_i) - (\alpha_j + \beta_j(x_i - s_{jm}))]^2 \Big\}, \tag{3.1}$$

where m is the number of splits considered and n is the number of data points. This expression can be interpreted as the sum of the lengths of the projections of g on subdomains D_{jm} onto the space orthogonal to the mean space spanned by the model, and is a measure of the deviation of g from the simple linear model on the subdomains. The proofs of the following results are in the Appendix.

Theorem 1. *Assume* (M0)-(M4). *Then, as* $m \le n/3$,
(a) $B(m,n) = O(\frac{1}{m^4} + \frac{m}{n})$ *as* $m \to \infty$.
(b) *If in addition the design is equidistant, i.e.,* $f \equiv c_0$ *for a constant* $c_0 > 0$,

$$B(m,n) = \frac{1}{720 c_o^3 m^4} \int_D g^{(2)}(x)^2 dx + o\left(\frac{1}{m^4}\right) + O\left(\frac{m}{n}\right). \tag{3.2}$$

This result relates $B(m,n)$ to the behavior of the second derivative $g^{(2)}$ of the regression function. A similar but more complex result holds for the case of a nonequidistant design. Analogous results hold for the multivariate situation.

Theorem 2. *Under* (M0)-(M4), *as* $n \to \infty$,
(a) *for* $m \to \infty$ *and for a possibly heteroscedastic model,*

$$E(\hat{\sigma}_m^2) = \frac{n}{n - 2m} B(m,n) + \int_D \sigma^2(x) f(x) dx (1 + o(1)); \tag{3.3}$$

(b) *for arbitrary m and a homoscedastic model,*

$$E(\hat{\sigma}_m^2) = \frac{n}{n-2m} B(m,n) + \sigma^2. \tag{3.4}$$

Unbiasedness of $\hat{\sigma}_m^2$ as an estimator of $\sigma^2 = \mathrm{Var}\,(\varepsilon_{i,n})$ in the homoscedastic case, or of $\sigma^2 = \int \sigma^2(x) f(x) dx$ in the heteroscedastic case, is equivalent to $B(m,n) \to 0$ as $n \to \infty$. If $\int [g^{(2)}(x)]^2 dx \neq 0$, this usually requires that $m \to \infty$, as seen from (3.2). In the homoscedastic case, the rate of convergence of the bias of $\hat{\sigma}_m^2$ is seen to be $O(m^{-4})$.

Regarding the variability of DSP graphs, we obtain under the additional homoscedasticity/normal moments assumption (M5) the following exact result.

Theorem 3. *Assume* (M0)-(M5). *Then, for any m, n,*

$$\mathrm{Var}\,(\hat{\sigma}_m^2) = \frac{2\sigma^2}{n-2m} \left[\frac{2n}{n-2m} B(m,n) + \sigma^2 \right]. \tag{3.5}$$

This indicates that $\mathrm{Var}\,(\hat{\sigma}_m^2)$ is monotonously increasing as m increases, whenever $B(m,n) = 0$ or $B(m,n) << \sigma^2$, usually the case for large $m \to [\frac{n}{3}]$ in view of Theorem 1. With Theorem 1 we find $\mathrm{Var}\,(\hat{\sigma}_m^2) \sim n/(m^4(n-2m)^2) + 1/(n-2m)$, and this is of order n^{-1}. A more complicated expression for $\mathrm{Var}\,(\hat{\sigma}_m^2)$ in the general heteroscedastic case is given in (A.23) of the Appendix.

We note that under (M0)-(M5), Theorem 2(b) and (3.5) together imply that

$$\mathrm{Var}\,(\hat{\sigma}_m^2) = \frac{2\sigma^2}{n-2m} \left[E(\hat{\sigma}_m^2) + \frac{n}{n-2m} B(m,n) \right]. \tag{3.6}$$

This implies that whenever $B(m,n)$ can be neglected,

$$\left(\frac{n}{2} - m \right) \mathrm{Var}\,(\hat{\sigma}_m^2) = \sigma^2 E(\hat{\sigma}_m^2), \tag{3.7}$$

which indicates that the $\hat{\sigma}_m^2$ should behave like quasi-Poisson random variables with overdispersion factor $2\sigma^2/(n-2m)$.

4. Multivariate Case and Data Illustrations

Domain Splitting can be extended to the multivariate case. The details of this extension are described in Appendix B.

As a multivariate example for $d = 2$, illustrating how the Domain Splitting works in a two-dimensional situation, we consider the DSP for data simulated from the model $y_i = -0.5x_{1i} + x_{2i} + 2\sin(2\pi x_{1i})\cos(\pi x_{2i}) + \varepsilon_i$, $i = 1, \ldots, 200$, in Figure 3. Here $\varepsilon_i \sim 0.5 * \mathcal{N}(0,1)$. Domain Splitting shows a sharp decline in the graph and in this example does indicate problems with the fit, while classical residual plots do not pinpoint the lack of fit in the model.

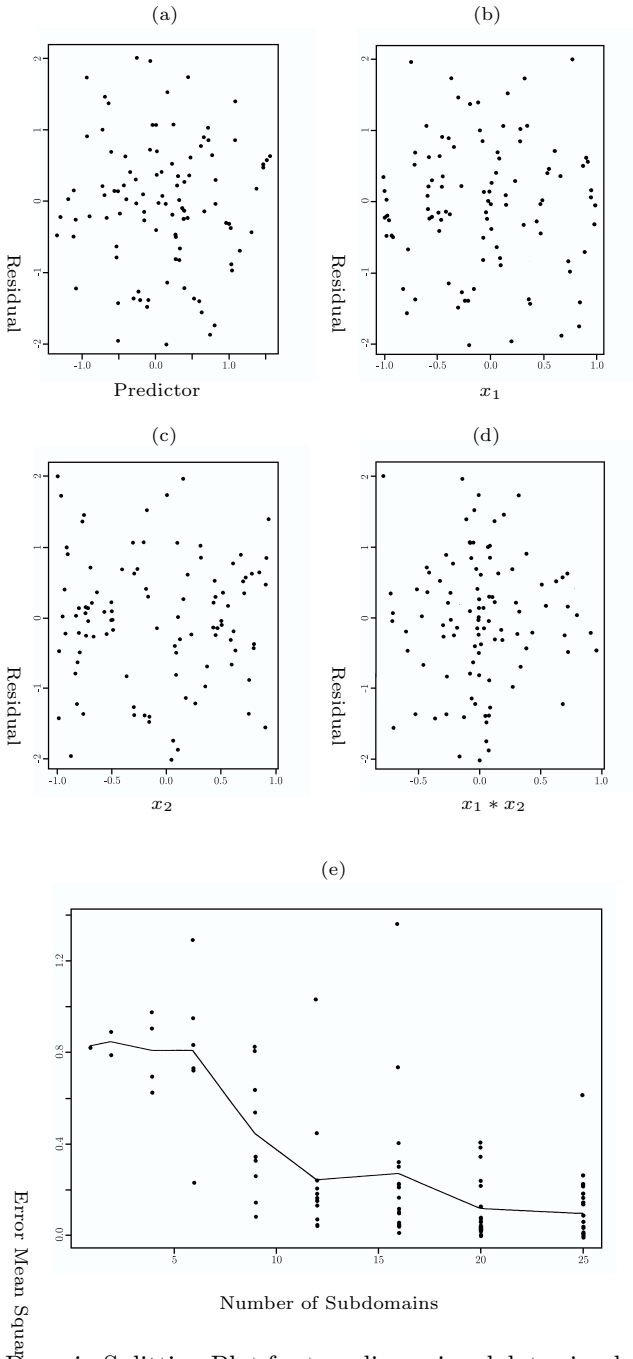

Figure 3. Domain Splitting Plot for two-dimensional data simulated from the model $y_i = -0.5x_{1i} + x_{2i} + 2\sin(2\pi x_{1i})\cos(\pi x_{2i}) + 0.5^*\mathcal{N}(0,1)$. (a) Residual plot versus predicted (b) Residual plot versus x_1(c) Residual plot versus x_2 (d) Residual plot versus x_1x_2 (e) Domain Splitting Plot.

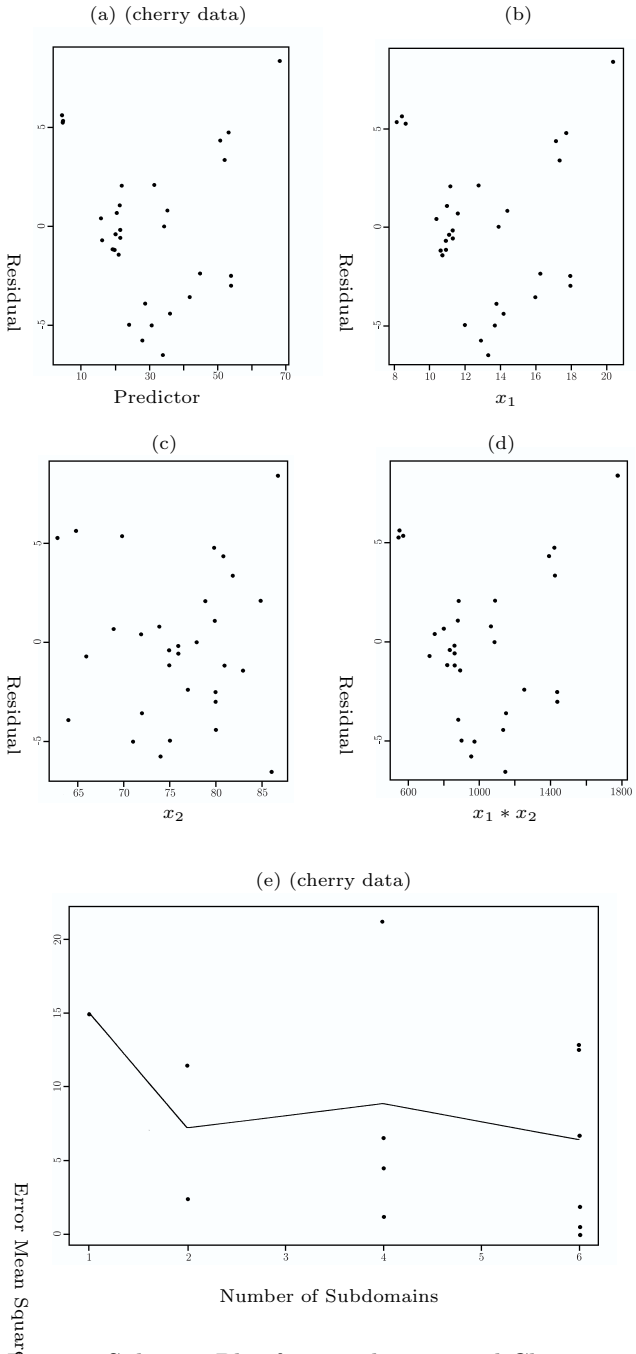

Figure 4. Domain Splitting Plot for two-dimensional Cherry Tree Data with $n = 31$. (a) Residual plot versus predicted (b) Residual plot versus x_1(c) Residual plot versus x_2 (d) Residual plot versus x_1x_2 (e) Domain Splitting Plot.

We can apply the technique to the Cherry Tree Data in Hand, Daly, Lunn, McConway and Ostrowski (1994). Here $n = 31$, and both the Domain Splitting Plot (DSP) and residual plots indicate that there are problems with the fit (Figure 4).

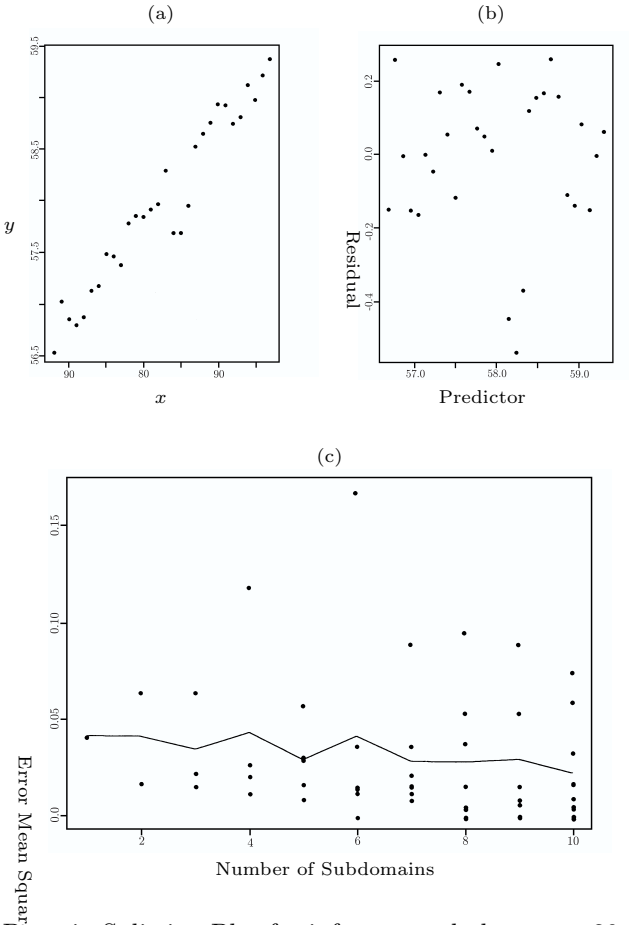

Figure 5. Domain Splitting Plot for infant growth data, $n = 30$. (a) Scatterplot: age in days on x-axis, height in cm on y-axis.(b) Residual plot versus predicted (c) Domain Splitting Plot.

Another data example of interest in this context is infant growth, where one question is whether growth is linear over a one-month period in males of approximately 2-3 months (see Heinrichs, Munson, Counts, Cutler and Baron (1994), for a discussion of an underlying scientific controversy on the nature of infant growth). A scatterplot of the data, the residual plot after fitting a simple linear regression line, and the DSP are shown in Figure 5. We note that the declining trend of the DSP indicates departure from the linear model assumption. The DSP graph shows a flat plateau only for seven or more subdomains. The large subdomain error mean squares observed for four or six

subdomains are caused by the cluster of three low lying points with predictor values around 85 days, seen in the scatterplot of Figure 5a.

5. Variance Estimation Via Domain Splitting

We note that the Domain Splitting principle provides a multitude of possible error variance estimates, namely $\hat{\sigma}_1^2, \hat{\sigma}_2^2, \ldots, \hat{\sigma}_m^2$. Once the number of splits m has been chosen, the corresponding $\hat{\sigma}_m^2$ (average subdomain mean squared error for this particular m, see (2.5)) is the natural estimator for σ^2. A simple non-automatic way to choose m is to view the Domain Splitting Plot (DSP) and to decide visually where an initial downward trend due to bias has ended, then take the smallest m where this is the case. Although this method appears to work quite well, it is desirable to provide a fully automatic implementation for the purpose of estimating the error variance. A heuristic approach which we found to work well is based on the following idea, which attempts to mimic the way a statistician might proceed.

Find the smallest m^* such that $\hat{\sigma}_m^2$ changes as little as possible in the window $m^* \le m \le m^* + m_w$, for a suitably chosen window size m_w; m^* will mark the beginning of a "flat" part of the DSP graph. Then find the smallest m, $m \le m^*$, such that the difference between $\hat{\sigma}_{m^*}^2$ and $\hat{\sigma}_m^2$ is reasonably small.

We implement this idea by choosing $m_w = 4$ and defining $R_m = \max\{\hat{\sigma}_m^2, \ldots, \hat{\sigma}_{m+4}^2\} - \min\{\hat{\sigma}_m^2, \ldots, \hat{\sigma}_{m+4}^2\}$, then finding $m^* = \arg\min_{1 \le m \le [n/3]-4} R_m$. Then we compare $\hat{\sigma}_m^2$ and $\hat{\sigma}_{m^*}^2$ by calculating $F_m = \frac{[(n-2m)\hat{\sigma}_m^2 - (n-2m^*)\hat{\sigma}_{m^*}^2]/(2m^*-2m)}{\hat{\sigma}_{m^*}^2}$ for $1 \le m \le m^* - 1$, and set

$$\hat{m} = \operatorname*{arg\,min}_{1 \le m \le m^*} \arg\min \left\{ F_m \le F_{2m^*-2m;n-2m^*;0.95} \right\}, \tag{5.1}$$

where $F_{2m^*-2m;n-2m^*;0.95}$ is the .95 quantile of the F distribution with $2m^* - 2m$ and $n - 2m^*$ degrees of freedom. If $F_m > F_{2m^*-2m;n-2m^*;0.95}$ for all $1 \le m < m^*$, choose $\hat{m} = m^*$. The resulting error variance estimate is

$$\hat{\sigma}_{\hat{m}}^2 = \frac{1}{n - 2\hat{m}} \sum_{j=1}^{\hat{m}} (n_{j\hat{m}} - 2)\hat{\sigma}_{j\hat{m}}^2. \tag{5.2}$$

(We also investigated the choice $\hat{m}' = m^*$, but found that this $\hat{m}'$ was sometimes located in a part of the DSP graph where the variance started to increase. The second step of choosing $\hat{m} \le m^*$ has a beneficial effect on the quality of the resulting variance estimates.)

Applying this procedure to the data of Example 1 as seen in Figure 1 (described in Section 2), one obtains $\hat{m} = 1$ and the variance estimate $\hat{\sigma}^2 = \hat{\sigma}_1^2 = 0.277$ (true value $\sigma^2 = 0.25$). For Example 2 (see Section 2, Figure 2),

this procedure yields $\hat{m} = 12$ (with $m^* = 14$) and an error variance estimate $\hat{\sigma}^2 = \hat{\sigma}^2_{12} = 0.247$ (true value $\sigma^2 = 0.25$). For the human infant growth data in Figure 5, one finds $\hat{m} = 3$ (with $m^* = 5$) and an error variance estimate $\hat{\sigma}^2 = \hat{\sigma}^2_3 = 0.035$. The analogous procedure for the two-dimensional case applied to the cherry tree data (see Figure 4) yields $\hat{\sigma}^2 = \hat{\sigma}^2_2 = 7.24$.

The proposed criterion (5.1) for selecting m was evaluated in a simulation study. Assumed sample sizes were $n = 100, 200$ and 500, and $N = 100$ Monte Carlo runs were made. Consider the case of a simple linear model $y = x + 0.5^*\mathcal{N}(0,1)$, with x_i equidistant in $[-1,1]$ (compare Example 1). The correct choice here is $m = 1$, as the simple linear model is the true model. Using (5.1), the distribution of the estimated $\hat{m}'s$ is listed in Table 1.

Table 1. Observed frequencies of estimates $\hat{m}$ (5.1) for various sample sizes in the case of a simple linear model, for $N = 100$ Monte Carlo runs.

Estimate $\hat{m} =$	1	2	3	4	5	≥ 6
Frequency for $n = 100$	96	2	0	1	1	0
Frequency for $n = 200$	94	1	2	1	2	0
Frequency for $n = 500$	96	3	0	1	0	0

For the situation corresponding to Example 2, where the underlying function is $y = \sin(2\pi x) + 0.5^*\mathcal{N}(0,1)$, we find the distribution of estimates for $\hat{m}$ (5.1) in Table 2. The format is the same as in Table 1, where again sample sizes $n = 100, 200, 500$ are considered and $N = 100$ Monte Carlo runs are made.

Table 2. Observed frequencies of estimates $\hat{m}$ (5.1) for various sample sizes in the case of a nonlinear sine function, for $N = 100$ Monte Carlo runs.

Estimate $\hat{m} =$	≤ 3	4	5	6	7	8	9	10	11	12	13	14	15	16	≥ 17
Frequency for $n = 100$	0	5	57	22	11	3	1	0	0	1	0	0	0	0	0
Frequency for $n = 200$	0	0	34	36	17	8	3	1	0	1	0	0	0	0	0
Frequency for $n = 500$	0	0	7	26	23	21	13	4	1	1	2	1	0	1	0

Inspecting the distribution of the values of $\hat{m}$, the need to split the domain in this case is well-recognized when using criterion (5.1). We note that the means of estimates $\hat{m}$ increase with n.

To study the practical behavior of the Domain Splitting variance estimator (5.2), we compared the following methods in a simulation study:

(a) Choosing $m = 1$, i.e., using the classical error mean square $\hat{\sigma}^2_1$;

(b) Choosing $m = [\frac{n}{3}]$, assuming nonlinearity of unspecified degree, and using $\hat{\sigma}^2_{[n/3]}$. This corresponds essentially to the prescription of a difference scheme for error variance estimation, such as the one proposed by Rice (1984), and variants thereof due to Gasser, Sroka and Jennen-Steinmetz (1986), Hall, Kay and Titterington (1990), as well as others.

(c) Choosing the proposed Domain Splitting variance estimator $\hat{\sigma}^2_{\hat{m}}$ of (5.1).

The results for the case of Example 1, are listed in Table 3, and for Example 2, in Table 4.

Table 3. Estimated error variances from simulation in the case of a simple linear model: means, standard deviations and mean squared errors (MSE) of estimated variances for choices $m = 1$, $m = \hat{m}$ and $m = [n/3]$ are compared for sample sizes $n = 100, 200, 500$ for $N = 100$ Monte Carlo runs. True variance $\sigma^2 = 0.25$.

	Mean			Standard Deviation			MSE		
	$m=1$	$m=\hat{m}$	$m=[n/3]$	$m=1$	$m=\hat{m}$	$m=[n/3]$	$m=1$	$m=\hat{m}$	$m=[n/3]$
$n=100:$	.249	.248	.255	.0346	.0349	.0553	.00120	.00122	.00309
$n=200:$	.250	.249	.256	.0247	.0249	.0450	.00060	.00062	.00206
$n=500:$	.250	.249	.250	.0155	.0155	.0265	.00024	.00024	.00070

Table 4. Same as Table 4 for the case of a nonlinear sine function. True variance $\sigma^2 = 0.25$.

	Mean			Standard Deviation			MSE		
	$m=1$	$m=\hat{m}$	$m=[n/3]$	$m=1$	$m=\hat{m}$	$m=[n/3]$	$m=1$	$m=\hat{m}$	$m=[n/3]$
$n=100:$	.673	.262	.255	.0706	.0447	.0553	.18353	.00216	.00308
$n=200:$	.680	.259	.256	.0483	.0271	.0450	.18733	.00081	.00206
$n=500:$	.669	.253	.250	.0377	.0173	.0265	.17691	.00031	.00070

We find that in the case of the linear model (Table 3) the loss for not knowing that $m = 1$, and instead using $m = \hat{m}$, is negligible in terms of MSE. However using $m = [n/3]$ in this situation, which amounts to a difference scheme variance estimator, would lead to a MSE about three times as big. This is due to a substantially larger variance of these (essentially unbiased) estimates.

In contrast, we see from Table 4 that using $m = 1$ in the case of a nonlinear model will be punished by a 100-fold increase in MSE as compared to the proposed choice $m = \hat{m}$. The choice $m = [n/3]$ again suffers from unnecessarily large variability as compared to $m = \hat{m}$, although it is orders of magnitude better than $m = 1$. The picture emerges that, although for the case of an underlying simple linear regression model the choice $m = 1$ is optimal, using $m = \hat{m}$ as protection against bias costs very little in terms of increased variance. On the other hand, using the difference scheme choice $m = [n/3]$ as an insurance against

bias is routinely associated with higher cost in terms of variance. This makes the recommendation to use $m = \hat{m}$ and $\hat{\sigma}^2_{\hat{m}}$ as variance estimate quite compelling.

6. Discussion

We have seen that the concept of domain splitting leads to superior error variance estimates for smooth nonparametric regression functions. It also provides the Domain Splitting Plot, a useful graphical tool for regression diagnostics in one or higher dimensions.

The idea of Breiman and Meisel (1976), for the estimation of the intrinsic variability of data in a nonlinear regression model by approximating the regression function with "piecewise linear patches", is a precursor of both regression trees as well as the domain splitting approach which is explored in this paper. Breiman and Meisel suggested fitting a parametric model by least squares on subsets of the domain which could potentially be split further into two smaller subsets of approximately equal size by a dividing hyperplane. The residual variance is estimated by combining the residual sums of squares of the linear models fitted on the various subdomains appropriately.

The main difference between our approach and the method of Breiman and Meisel is that, in the latter method, a sequential splitting strategy is proposed to arrive at the best segmentation and a good estimate for the residual variance. In our approach, we simultaneously consider an entire sequence of segmentations and the residual variances estimated for the various segments. We introduced the Domain Splitting Plot as a graphical device, for several reasons to find the "best" segmentation in the sequence of segmentations, for use as a model diagnostic, and ultimately for variance estimation. We also provided a variety of asymptotic results, including the mean squared error of the various variance estimates, to lend support to the graphical procedure. The resulting variance estimates were compared with alternative classes of difference based variance estimates, and were found to compare quite favorably.

We now discuss various extensions and potential applications of the Domain Splitting principle introduced in this paper. First of all, one may consider the application to piecewise continuous rather than smooth regression functions. As the number of subdomains increases, the cutpoints defining the subdomains will eventually come arbitrarily close to the locations of discontinuities of the regression function, and the Domain Splitting Plot will indicate that point by turning from a decline to a flat part. The asymptotic analysis of this case is similar to the one given above for the smooth case.

On the other hand, given a suitable number of subdomains, one might fit functions which are linear on the various subdomains of the data. The squared residuals from that fit would lead to very good variance estimates and, if the

discontinuities in the fit are not of much concern, this approach can provide a good fit to the data (the smoothing parameter choice would be implicitly based on the Domain Splitting Plot). As a variant, one could apply a simple kernel smoother in a second step with a suitably small bandwidth, smoothing over the discontinuous fit in order to obtain a smooth function or surface estimate.

A referee suggested unequal subdomains, choosing higher resolution on intervals where the regression function varies more rapidly and less on the flat parts. This seemingly attractive method may lead to better fits to the data, and it corresponds roughly to the idea of local smoothing parameter choice in nonparametric regression. While such a modification detracts somewhat from the overall simplicity of Domain Splitting Plots, versions of such a method would not be too difficult to implement. For instance, one could create sub-DSPs for given subdomains and subject the further splitting of such subdomains to the behavior of these sub-DSPs. However, this would add substantial complexity to graphical interpretation, and for splitting with different resolution a modification of a method with sequential random splitting such as CART, SUPPORT, or Breiman and Meisel's (1976) method may be more appropriate.

We note that instead of fitting simple linear regression lines or planes on the subdomains, an alternative is to fit other parametric models such as quadratic functions or surfaces, as was already suggested by Breiman and Meisel (1976). This is of interest when the goal is to find a piecewise parametric approximation to the unknown regression function.

The level of resolution in terms of the number of subdomains determined visually, or according to criterion (5.1), contains implicit information about the smoothness of the regression function as measured in terms of $\int_D g^{(2)}(x)^2 dx$. Since $B(m,n)$, is closely related to $\int_D g^{(2)}(x)^2 dx$, (3.2), the behavior of $\hat{\sigma}_m^2$ when regressed against $E(\hat{\sigma}_m^2)$ as given in (3.4) can provide such information. Together with $\hat{\sigma}_m^2$, this may lead to useful new plug-in bandwidth selection formulas for nonparametric regression. We conclude from these considerations that the potential applications of the Domain Splitting approach extend beyond variance estimation, goodness-of-fit diagnostics and piecewise linear approximation.

Acknowledgement

The authors wish to thank two referees for helpful comments. This research was supported in part by NSA Grant MDA 904-96-10026 and NSF Grant DMS-96-25984.

Appendix A. Auxiliary Results and Proofs

This section contains proofs of the results stated in Section 3. The basic assumptions (M0)-(M4) are assumed to hold, and the notation introduced in

Section 2 is used throughout. We first establish auxiliary results regarding the representation of functions on subdomains D_{jm} by orthonormal polynomials. The first of these is a well-known result and the proof is omitted.

Lemma 1. *Let $\phi_\ell, \ell \geq 0$, be a system of orthonormal polynomials on $L^2(D_{jm})$, where $D_{jm} = (F^{-1}((j-1)/m), \quad F^{-1}(j/m)]$, i.e., ϕ_ℓ is a polynomial of degree ℓ, and the ϕ_ℓ satisfy*

$$\int_{D_{jm}} \phi_\ell(x)\phi_k(x)dx = \delta_{\ell k}, \tag{A.1}$$

$\delta_{\ell k} = 1_{\{\ell=k\}}$ being the Kronecker symbol. Let

$$g(x) = \Sigma\lambda_\ell\phi_\ell(x), \quad x \in D_{jm}, \tag{A.2}$$

be the unique representation of the function g on D_{jm} in this orthonormal system. Then it holds for any polynomial p_k of degree k that

$$\int (g(x) - \sum_{\ell=0}^{k} \lambda_\ell\phi_\ell(x))^2 dx \leq \int (g(x) - p_k(x))^2 \, dx. \tag{A.3}$$

Lemma 2. *For $1 \leq j \leq m$,*

$$\min_{\alpha,\beta} \int_{D_{jm}} (g(x) - (\alpha + \beta(x - s_{jm})))^2 \, dx = \frac{g^{(2)}(s_{jm})^2}{720\, m^5 f^5(s_{jm})} + o(\frac{1}{m^5}), \text{ as } m \to \infty.$$

Proof. Let $a_{jm} = \frac{1}{2}\left(F^{-1}(j/m) - F^{-1}((j-1)/m)\right)$ and $x_{jm} = F^{-1}((j-1)/m) + a_{jm}$, so that $D_{jm} = (x_{jm} - a_{jm}, x_{jm} + a_{jm}]$. From (A.2) and (A.3) we find

$$\min_{\alpha,\beta} \int_{D_{jm}} (g(x) - (\alpha + \beta(x - s_{jm})))^2 dx = \sum_{\ell=2}^{\infty} \lambda_\ell^2, \tag{A.4}$$

where

$$\lambda_\ell = \int_{D_{jm}} g(x)\phi_\ell(x)dx. \tag{A.5}$$

By a Taylor expansion, using (M3),

$$g(x) = \sum_{i=0}^{2} \frac{g^{(i)}(x_{jm})}{i!}(x - x_{jm})^i + R(x), \quad x \in D_{jm},$$

where $R(x) = o((x - x_{jm})^2)$. With (M2),

$$\sum_{\ell=3}^{\infty} \lambda_\ell^2 \leq \int_{D_{jm}} R^2(x)dx = o\left(\frac{1}{m^5}\right), \tag{A.6}$$

and

$$\lambda_2^2 = \frac{g^{(2)}(x_{jm})^2}{4}\Big[\int_{D_{jm}}(x-x_{jm})^2\phi_2(x)dx\Big]^2 + o\Big(\Big[\int_{D_{jm}}(x-x_{jm})^2\phi_2(x)dx\Big]^2\Big). \tag{A.7}$$

We find for the relation of the normalized Legendre polynomials $\bar{\phi}_\ell$ on $[-1,1]$ with the normalized Legendre polynomials ϕ_ℓ on D_{jm}, that

$$\phi_\ell(x) = \frac{1}{\sqrt{a_{jm}}}\bar{\phi}_\ell\Big(\frac{x-x_{jm}}{a_{jm}}\Big), \quad x \in D_{jm}. \tag{A.8}$$

Further, according to Abramowitz and Stegun (1964),

$$\bar{\phi}_2(x) = \sqrt{\frac{5}{8}}(-1+3x^2), \quad x \in [-1,1]. \tag{A.9}$$

From (A.7)-(A.9), one obtains $\{\int_{D_{jm}}(x-x_{jm})^2\phi_2(x)dx\}^2 = \frac{8}{45}a_{jm}^5$, and with (M1), (M2),

$$\lambda_2^2 = \frac{1}{720}[g^{(2)}(x_{jm})]^2\Big(\frac{1}{mf(x_{jm})}\Big)^5 + o(\frac{1}{m^5}). \tag{A.10}$$

Here we observe that $F^{-1}(j/m) - F^{-1}((j-1)/m) = [mf(x_{jm})]^{-1}(1+o(1))$ implies $a_{jm} = [2mf(x_{jm})]^{-1} + o(m^{-1})$. Lemma 2 follows from the continuity of $g^{(2)}(\cdot)/f^5(\cdot)$.

Proof of Theorem 1. A useful approximation for Lipschitz continuous functions H is (see (M1), (M2)) $n^{-1}\sum_{x_i\in D_{jm}}H(x_i) = \int_{D_{jm}}H(x)f(x)dx + O(n^{-1})$. We conclude $n^{-1}\sum_{x_i\in D_{jm}}(g(x_i)-(\alpha+\beta(x_i-s_{jm})))^2 = \int_{D_{jm}}(g(x)-(\alpha+\beta(x-s_{jm})))^2 f(x)dx + O(n^{-1})$. One finds that the $O(\cdot)$-term is uniform in $1 \le j \le m$, in m, and on compact sets (α,β). Lemma 2 then implies for the equidistant case ($f \equiv c_o$), $\min_{\alpha,\beta} n^{-1}\sum_{x_i\in D_{jm}}[g(x_i)-(\alpha+\beta(x_i-s_{jm}))]^2 = c_o[g^{(2)}(s_{jm})]^2[720\,m^5c_o^5]^{-1} + o(m^{-5}) + O(n^{-1})$, as $n \le m \to \infty$. Theorem 1(b) follows by summation, and Theorem 1(a) by a Taylor expansion of g around s_{jm} within D_{jm}, which gives $\min_{\alpha,\beta}\int_{D_{jm}}[g(x)-\alpha+\beta(x-s_{jm})]^2 f(x)dx = O\{\int_{D_{jm}}[g^{(2)}(s_{jm})(x-s_{jm})^2 + o((x-s_{jm})^2)]^2 f(x)dx\} = O(m^{-5})$.

Proof of Theorem 2. It is useful to introduce some matrix notation. We fix m and j $(1 \le j \le m)$ and rewrite model (M0) restricted to subdomain D_{jm} as follows:

$$y_i^* = g(x_i^*) + \varepsilon_i^*, \quad i = 1,\dots,n_{jm},$$

where the x_i^* are an enumeration of all $x_i \in D_{jm}$, and y_i^*, ε_i^* are the corresponding concomitants. We then define n_{jm}-vectors $Y_{jm} = (y_1^*,\dots,y_{n_{jm}}^*)^T, E_{jm} =$

$(\varepsilon_1^*, \ldots, \varepsilon_{n_{jm}}^*)^T, G_{jm} = (g(x_1^*), \ldots, g(x_{n_{jm}}^*))^T$, where A^T denotes the transpose of a vector or matrix A. Let

$$X_{jm} = \begin{pmatrix} 1 & x_1^* \\ \vdots & \vdots \\ 1 & x_{n_{jm}}^* \end{pmatrix}$$

denote the $n_{jm} \times 2$ design matrix of the simple linear model on subdomain D_{jm}. Model (M0) can then be written as $Y_{jm} = G_{jm} + E_{jm}$, $1 \le j \le m$. The hat matrix corresponding to projecting on the mean space D_{jm} is the $n_{jm} \times n_{jm}$ matrix $H_{jm} = X_{jm}(X_{jm}^T X_{jm})^{-1} X_{jm}^T$, with $\hat{Y}_{jm} = H_{jm} Y_{jm}$, where $\hat{Y}_{jm}$ denotes the vector of fitted values with i-th element $\hat{y}_i^* = \hat{\alpha} + \hat{\beta}(x_i^* - s_{jm})$, $1 \le i \le n_{jm}$ (see Cook and Weisberg (1982), p.11). The projection on the orthogonal space is $A_{jm} = I_{n_{jm} \times n_{jm}} - H_{jm}$, where $I_{\ell \times \ell}$ denotes the ℓ-dimensional identity matrix.

Note that

$$A_{jm}^T = A_{jm}, \quad A_{jm}^2 = A_{jm}. \tag{A.11}$$

Defining $v_{jm} = \sum_{x_i^* \in D_{jm}} (x_i^* - s_{jm})^2$, and writing $A_{jm} = (a_{\ell k})_{1 \le \ell, k \le n_{jm}}$, we find (compare Cook and Weisberg (1982), p.12)

$$a_{\ell k} = \delta_{\ell k} - \frac{1}{n_{jm}} - \frac{(x_\ell^* - s_{jm})(x_k^* - s_{jm})}{v_{jm}}, \quad 1 \le \ell, k \le n_{jm}. \tag{A.12}$$

Note that by definition of the subdomains

$$n_{jm} = \frac{n}{m}(1 + o(1)), \tag{A.13}$$

which implies

$$v_{jm} = \sum_{x_i^* \in D_{jm}} x_i^{*2} - n_{jm} s_{jm}^2 = n\Big\{ \int_{D_{jm}} x^2 f(x) dx - m\Big[\int_{D_{jm}} x f(x) dx \Big]^2 \Big\}(1 + o(1)). \tag{A.14}$$

We obtain for $\hat{\sigma}_m^2$, (2.5),

$$\hat{\sigma}_m^2 = \frac{1}{n - 2m} \sum_{j=1}^m \{ E_{jm}^T A_{jm}^T A_{jm} E_{jm} + 2 G_{jm}^T A_{jm}^T A_{jm} E_{jm} + G_{jm}^T A_{jm}^T A_{jm} G_{jm} \}, \tag{A.15}$$

and for $B(m, n)$, (3.1),

$$B(m, n) = \frac{1}{n} \sum_{j=1}^m G_{jm}^T A_{jm}^T A_{jm} G_{jm}. \tag{A.16}$$

By (A.11)-(A.16),

$$\hat{\sigma}_m^2 = \frac{n}{n - 2m} B(m, n) + \frac{1}{n - 2m} \sum_{j=1}^m \left\{ E_{jm}^T A_{jm} E_{jm} + 2 G_{jm}^T A_{jm} E_{jm} \right\}, \tag{A.17}$$

and, using (M4),

$$E\hat{\sigma}_m^2 = \frac{n}{n-2m}B(m,n) + \frac{1}{n-2m}\sum_{j=1}^{m} E\left\{\left(E_{jm}^T A_{jm} E_{jm}\right)\right\}$$
$$= \frac{n}{n-2m}B(m,n)+\frac{1}{n-2m}\sum_{j=1}^{m}\Big(\sum_{x_i\in D_{jm}}\Big(1-\frac{1}{n_{jm}}-\frac{(x_i-s_{jm})^2}{v_{jm}}\Big)\sigma^2(x_i)\Big). \tag{A.18}$$

Using (M1), (M2), (A.13), (A.14), and Riemann sum approximations,

$$\sum_{j=1}^{m}\sum_{x_i\in D_{jm}}\frac{\sigma^2(x_i)}{n_{jm}} = m\int \sigma^2(x)f(x)dx(1+o(1)), \tag{A.19}$$

$$\sum_{j=1}^{m}\sum_{x_i\in D_{jm}}\frac{(x_i-s_{jm})^2}{v_{jm}}\sigma^2(x_i)$$
$$=\sum_{j=1}^{m}\frac{\int_{D_{jm}}\Big(x-m\int_{D_{jm}} yf(y)dy\Big)^2\sigma^2(x)f(x)dx}{\int_{D_{jm}}\Big(x-m\int_{D_{jm}} yf(y)dy\Big)^2 f(x)dx}(1+o(1)). \tag{A.20}$$

When $m\to\infty$, using the mean value theorem for integration,

$$\sum_{j=1}^{m}\sum_{x_i\in D_{jm}}\frac{(x_i-s_{jm})^2}{v_{jm}}\sigma^2(x_i) = m\int\sigma^2(x)f(x)dx(1+o(1)). \tag{A.21}$$

When the homoscedastic case applies, we find

$$\sum_{j=1}^{m}\sum_{x_i\in D_{jm}}\Big(1-\frac{1}{n_{jm}}-\frac{(x_i-s_{jm})^2}{v_{jm}}\Big)\sigma^2 = (n-2m)\sigma^2. \tag{A.22}$$

This completes the proof.

Proof of Theorem 3. Let $C = A_{jm}G_{jm}$, $C=(c_1^*,\dots,c_{n_{jm}}^*)^T$, with A_{jm} having the elements $a_{\ell k}$ as in (A.12). Denote by $Tr(M)$ the trace of a matrix M. We find

$$\text{Var}\left(E_{jm}^T A_{jm}E_{jm}\right) = \sum_{\ell=1}^{n_{jm}} a_{\ell\ell}^2\left(\mu_4(x_\ell^*)-3\sigma^2(x_\ell^*)\right)+2\sum_{\ell,k=1}^{n_{jm}} a_{\ell k}^2\sigma^2(x_\ell^*)\sigma^2(x_k^*),$$

$$\text{Var}\left(G_{jm}^T A_{jm}E_{jm}\right) = \sum_{\ell=1}^{n_{jm}} c_\ell^{*2}\sigma^2(x_\ell^*),$$

$$\text{Cov}\left(E_{jm}^T A_{jm}E_{jm}, G_{jm}^T A_{jm}E_{jm}\right) = \sum_{\ell=1}^{n_{jm}} a_{\ell\ell}c_\ell^*\mu_3(x_\ell^*),$$

which together imply

$$\begin{aligned}\mathrm{Var}\,(\hat{\sigma}_m^2) = \frac{1}{(n-2m)^2}\Big\{ \sum_{j=1}^{m} \Big[\sum_{x_\ell^* \in D_{jm}} \Big(1 - \frac{1}{n_{jm}} - \frac{(x_\ell^* - s_{jm})^2}{v_{jm}}\Big)^2 (\mu_4(x_\ell^*) - 3\sigma^2(x_\ell^*)) \\ + 2 \sum_{x_\ell^*, x_k^* \in D_{jm}} \Big(\delta_{\ell k} - \frac{1}{n_{jm}} - \frac{(x_\ell^* - s_{jm})(x_k^* - s_{jm})}{v_{jm}}\Big)^2 \sigma^2(x_\ell^*)\sigma^2(x_k^*) \\ + 4 \sum_{x_\ell^* \in D_{jm}} \Big[\sum_{x_k^* \in D_{jm}} \Big(\delta_{\ell k} - \frac{1}{n_{jm}} - \frac{(x_\ell^* - s_{jm})(x_k^* - s_{jm})}{v_{jm}}\Big) g(x_k^*)\Big]^2 \sigma^2(x_\ell^*) \\ + 4 \sum_{x_\ell^* \in D_{jm}} \Big(1 - \frac{1}{n_{jm}} - \frac{(x_\ell^* - s_{jm})^2}{v_{jm}}\Big) \Big[\sum_{x_k^* \in D_{jm}} \Big(\delta_{\ell k} - \frac{1}{n_{jm}} \\ - \frac{(x_\ell^* - s_{jm})(x_k^* - s_{jm})}{v_{jm}}\Big) g(x_k^*)\Big] \mu_3(x_\ell^*)\Big]\Big\}. \quad \text{(A.23)}\end{aligned}$$

Observing that

$$\begin{aligned}\sum_{j=1}^{m} \sum_{x_\ell^*, x_k^* \in D_{jm}} \Big(\delta_{\ell k} - \frac{1}{n_{jm}} - \frac{(x_\ell^* - s_{jm})(x_k^* - s_{jm})}{v_{jm}}\Big)^2 \sigma^4 = \sigma^4 \sum_{j=1}^{m} \sum_{x_\ell^*, x_k^* \in D_{jm}} a_{\ell k}^2 \\ = \sigma^4 \sum_{i=1}^{m} Tr(A_{jm}^2) = \sigma^4 \sum_{i=1}^{m} Tr(A_{jm}) = \sigma^4 \sum_{i=1}^{m} (n_{jm} - 2) = \sigma^4 (n - 2m),\end{aligned}$$

we find that under (M5), (A.23) simplifies to (3.6).

Appendix B. The Multivariate Case

We use the same general notation as in Section 2. Assume $d \geq 1$ is the dimension of the vector of predictor variables $\mathbf{X}$, and that the values of the predictors are in

$$\bigtimes_{\ell=1}^{d} D^{(\ell)} = D \subset \Re^d,$$

where $D^{(\ell)}$ includes the domain of the ℓth predictor variable, $\ell = 1, \ldots, d$. Consider partitioning sets $D_{\gamma_\ell, \alpha_\ell}^{(\ell)}$ of $D^{(\ell)}$, where γ_ℓ, α_ℓ are integers, $1 \leq \gamma_\ell \leq \alpha_\ell$, such that

$$\bigcup_{\gamma_\ell=1}^{\alpha_\ell} D_{\gamma_\ell, \alpha_\ell}^{(\ell)} = D^{(\ell)} \text{ and } D_{\gamma_\ell, \alpha_\ell}^{(\ell)} \cap D_{\gamma_\ell', \alpha_\ell}^{(\ell)} = \emptyset$$

for $\gamma_\ell \neq \gamma_\ell'$. Given a number $\alpha_\ell \geq 1$ of partitioning sets for $D^{(\ell)}$, the partitioning sets $D_{\gamma_\ell, \alpha_\ell}^{(\ell)}$, $1 \leq \gamma_\ell \leq \alpha_\ell$, for $D^{(\ell)}$ can be chosen to be of equal size or so as to have approximately equal numbers of observations in them (in analogy to the

univariate case). Given $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_d)$, and $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_d), 1 \leq \gamma_\ell \leq \alpha_\ell, 1 \leq \ell \leq d$, we define a partition of D by means of the partitioning sets

$$D_{\gamma\boldsymbol{\alpha}} = \underset{\ell=1}{\overset{d}{X}} D^{(\ell)}_{\gamma_\ell, \alpha_\ell}.$$

Consider the sequence of partition-inducing multiindices $\boldsymbol{\alpha}_1 = (1, \ldots, 1)$, $\boldsymbol{\alpha}_2 = (2, 1, \ldots, 1)$, $\boldsymbol{\alpha}_3 = (2, 2, 1, \ldots, 1), \ldots, \boldsymbol{\alpha}_{d+1} = (2, \ldots, 2)$, multiindices $\boldsymbol{\alpha}_{d+2} = (3, 2, \ldots, 2), \ldots,$ $\boldsymbol{\alpha}_{2d+1} = (3, \ldots, 3), \ldots$, and the associated sequence of integers $m_1 = \boldsymbol{\alpha}_1! = 1, m_2 = \boldsymbol{\alpha}_2! = 2, m_3 = \boldsymbol{\alpha}_3! = 4, \ldots, m_j = \boldsymbol{\alpha}_j!, \ldots$, where $\boldsymbol{\alpha}! = \prod \alpha_\ell$ if $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_d)$. The jth element in this sequence defines a partition of D consisting of $m_j = \boldsymbol{\alpha}_j!$ partitioning sets $D_{\gamma\boldsymbol{\alpha}_j}, 1 \leq \gamma_\ell \leq \alpha_{j\ell}$, which we re-index and denote as $D_{1m}, \ldots, D_{mm}$, for all m contained in the sequence $m_1, m_2, \ldots$, as long as $m \leq [\frac{n}{d+2}]$.

As an illustration of this multivariate partitioning scheme, consider the two dimensional case $d = 2$ with $n = 25$ and $D^{(1)} = D^{(2)} = [0, 1]$, so that $D = [0, 1]^2$. Assuming that the domain splits lead to equal size-subdomains, we find $m = 9$ and the partitioning sequence $D_{1,9} = \{D\}, \boldsymbol{\alpha}_1 = (1, 1)$, $m_1 = 1$; $D_{2,9} = \{[0, \frac{1}{2}] \times [0, 1], [\frac{1}{2}, 1] \times [0, 1]\}, \boldsymbol{\alpha}_2 = (2, 1)$, $m_2 = 2$; $D_{3,9} = \{[0, \frac{1}{2}] \times [0, \frac{1}{2}], [0, \frac{1}{2}] \times [\frac{1}{2}, 1], [\frac{1}{2}, 1] \times [0, \frac{1}{2}], [\frac{1}{2}, 1] \times [\frac{1}{2}, 1]\}, \boldsymbol{\alpha}_3 = (2, 2)$, $m_3 = 4$; $D_{4,9} = \{[0, \frac{1}{3}] \times [0, \frac{1}{2}], [0, \frac{1}{3}] \times [\frac{1}{2}, 1], [\frac{1}{3}, \frac{2}{3}] \times [0, \frac{1}{2}], [\frac{1}{3}, \frac{2}{3}] \times [\frac{1}{2}, 1],\ [\frac{2}{3}, 1] \times [0, \frac{1}{2}], [\frac{2}{3}, 1] \times [\frac{1}{2}, 1]\}, \boldsymbol{\alpha}_4 = (3, 2)$, $m_4 = 6$.

We obtain the d-dimensional version of the DSP by plotting $\hat{\sigma}^2_{jm}$ versus m, where $\hat{\sigma}^2_{jm}$ is as in (2.4), with SS_{jm} replaced by

$$SS_{jm}(\alpha, \boldsymbol{\beta}) = \sum_{\mathbf{X}_i \in D_{jm}} \left\{ y_i - \left[\alpha + \boldsymbol{\beta}^T (\boldsymbol{X}_i - \boldsymbol{s}_{jm}) \right] \right\}^2.$$

We note that other orderings of the sequence $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2 \cdots$, are also possible and could be used alternatively. A further constraint is that each of the partitioning sets D_{jm} contains at least $(d + 2)$ observations. Should one of the partitioning sets $D_{\gamma\boldsymbol{\alpha}_j}$ contain less than $(d + 2)$ observations, a possible strategy is as follows: Given the partition inducing $\boldsymbol{\alpha}_{j-1}$, we obtain the next such index $\boldsymbol{\alpha}_j$ in such a way that (i) $\alpha_{jp} = \alpha_{j-1,p} + 1$ for a p, $1 \leq p \leq d$, whereas $\alpha_{jq} = \alpha_{j-1,q}$ for $1 \leq q \leq d$, $q \neq p$; (ii) all $D_{\gamma\boldsymbol{\alpha}_j}$ have at least $(d + 2)$ observations; (iii) p is the smallest index which achieves (i) and (ii). If no such $\boldsymbol{\alpha}_j$ exists, the DSP stops at $m = \boldsymbol{\alpha}_{j-1}!$.

References

Abramowitz, M. and Stegun, I. (1964). *Handbook of Mathematical Functions.* National Bureau of Standards, Washington, D. C.

Breiman, L. and Meisel, W. S. (1976). General estimates of the intrinsic variability of data in nonlinear regression models. *J. Amer. Statist. Assoc.* **71**, 301-307.

Buckley, M. J., Eagleson, G. K. and Silverman, B. W. (1988). The estimation of residual variance in nonparametric regression. *Biometrika* **75**, 189-199.

Chaudhuri, P., Huang, M.-C., Loh, W.-Y. and Yao, R. (1994). Piecewise-polynomial regression trees. *Statistica Sinica* **4**, 143-167.

Cook, D. and Weisberg, S. (1982). *Residuals and Inference in Regression*. Chapman and Hall, London.

Eubank, R. L. and Spiegelman, C. H. (1990). Testing the goodness of fit of a linear model via nonparametric regression techniques. *J. Amer. Statist. Assoc.* **85**, 387-392.

Gasser, T., Sroka, L. and Jennen-Steinmetz, C. (1986). Residual variance and residual pattern in nonlinear regression. *Biometrika* **73**, 625-633.

Hall, P., Kay, J. W. and Titterington, D. M. (1990). Asymptotically optimal difference-based estimation of variance in nonparametric regression. *Biometrika* **77**, 521-528.

Hand, D. J., Daly, F., Lunn, A. D., McConway, K. J. and Ostrowski, E. (1994). *Small Data Sets.* Chapman and Hall, London.

Heinrichs, C., Munson, P. J., Counts, D. R., Cutler, G. B. and Baron, J. (1994). Patterns of human growth. *Science* **268**, 442-445.

Mansfield, E. R. and Conerly, M. D. (1987). Diagnostic value of residual and partial residual plots. *American Statistician* **33**, 108-115.

Müller, H. G. and Stadtmüller, U. (1987). Estimation of heteroscedasticity in regression analysis. *Ann. Statist.* **15**, 182-201.

Müller, H. G. and Stadtmüller, U. (1993). On variance function estimation with quadratic forms. *J. Statist. Plann. Inference* **35**, 213-231.

Rice, J. (1984). Bandwidth choice for nonparametric kernel regression. *Ann. Statist.* **12**, 1215-1230.

Sacks, J. and Ylvisaker, D. (1970). Designs for regression problems with correlated errors III. *Ann. Math. Statist.* **41**, 2057-2074.

Yao, Y. C. and Au, S. T. (1989). Least-squares estimation of a step function. *Sankhyā* **51**, 370-381.

Division of Statistics, University of California, Davis, Davis, CA 95616, U.S.A.

E-mail: mueller@wald.ucdavis.edu

Department of Biostatistics, Merck Research Laboratories, P.O. Box 2000, RY 70-38, Rahway, NJ 07065-0900, U.S.A.

E-mail: peng-liang_zhao@merck.com

(Received August 1997; accepted July 1998)