Local Polynomial Regression and Its Applications in Environmental Statistics

David Ruppert *

February 14, 1996

Abstract

Nonparametric regression estimates a conditional expectation of a response given a predictor variable without requiring parametric assumptions about this conditional expectation. There are many methods of nonparametric regression including kernel estimation, smoothing splines, regression splines, and orthogonal series. Local regression fits parametric models locally by using kernel weights. Local regression is proving to be a particularly simple and effective method of nonparametric regression.

This talk reviews recent work on local polynomial regression including estimation of derivatives, multivariate predictors, and bandwidth selection. Three applications to environmental science are discussed:

1. Estimation of the distribution of airborne mercury about an incinerator using biomonitoring data.
2. Estimation of airborne pollutants from LIDAR (LIght Detection And Ranging) data. Because of substantial heteroskedasticity, this example requires estimation of the conditional variance function as well as the conditional expectation function.
3. Estimation of gradients from elevation data. The estimated gradients are used in a model to predict soil movement during earthquakes. Data from Noshiro, Japan are used.

The first and third examples use two-dimensional spatial data. Though these problems could be analyzed by geostatistics, local polynomial regression has the advantage of modeling the heteroskedasticity in the first and second examples and being able to estimate gradients in the third.

Key words and phrases. Bandwidth Selection, Biomonitoring, Curve and surface fitting, Derivative Estimation, LIDAR, Spatial Data.

Short title: Local polynomial regression.

*David Ruppert is Professor, School of Operations Research & Industrial Engineering, Cornell University, Ithaca, New York 14853 (E-mail: davidr@orie.cornell.edu). This research was supported by NSA Grant MDA 904-95-H-1025 and NSF Grant DMS-9306196.

1 Introduction

In nonparametric regression, also called curve and surface estimation, we observe data $\{(\boldsymbol{X}_i, Y_i)\}_{i=1}^n$ and attempt to estimate $m(\boldsymbol{x}) := E(Y_i|\boldsymbol{X}_i = \boldsymbol{x})$. Here Y_i is a univariate response and $\boldsymbol{X}_i$ is a vector of covariates or predictor variables. There are many possible applications of nonparametric regression in the environmental sciences, and in this paper we will focus on three examples: biomonitoring of airborne mercury, LIDAR measurement of atmospheric pollutants, and estimation of soil gradient in geotechnical engineering.

An outline of the paper is as follows. In this Section, we will introduce the three examples. Then in Section 2 we will look at local polynomial regression, a simple yet powerful nonparametric regression methodology. In Section 3 we discuss the selection of the bandwidth or smoothing parameter by minimizing an estimator of the mean squared error. A common problem in environmental applications and elsewhere is the modeling of heteroskedasticity. This problem is discussed in Section 4 where we introduce variance function estimation, i.e., estimation of conditional variances. Finally, Section 5 returns to the three examples, and Section 6 is discussion.

The first example comes from biomonitoring of airborne mercury about a solid waste incinerator in New Jersey, USA. The data come from Opsomer, Agras, Carpi, and Rodrigues (1995). Waste incineration is a major source of environmental mercury, and the researchers wanted to know if elevated mercury concentrations could be detected near the incinerator. Pots of sphaghum moss were placed at 16 sampling locations about the solid waste incinerator and exposed to ambient conditions between July 9 and July 23, 1991. Then the moss was collected, dried, and assayed for mercury. The goals of the study include estimating the distribution of mercury about the incinerator and testing the null hypothesis that the mean mercury concentration is constant. Possible approaches to this problem are geostatistics, i.e., kriging (Cressie, 1991), smoothing splines such as thin-plate splines (Green and Silverman, 1994), regression splines such as MARS (Friedman, 1991), and local regression. We will pursue the last approach. Here $\boldsymbol{X}$ is bivariate spatial position and Y is measured mercury concentration. In Figure 1(b) the sampling locations and the location of the incinerator are shown. Also shown are the contours of mercury concentration, estimated by the local regression method that is described in this paper.

The contours indicate that mercury concentration peaks a little south of the incinerator. There are only 16 sampling locations, with replicate moss pots at 6 of these sites, for a total of 22 observations. With so few data, only gross features of mercury deposition can be resolved, but the nonparametric fit provides a nice image of these features.

The second example also involves measurement of atmospheric pollutants, but with a different measuring technique, LIDAR (LIght Detection And Ranging); see Ragnarson (1994) and Holst, Hössjer, Björklund, Ragnarson, and Edner (1994). The particular technique used was DIAL (DIfferential Absorption Lidar) where one measures reflected light from lasers at two frequencies, one off and one on the resonance frequency of the chemical

species being measured. Here $Y = \log[\mathrm{P(on)}/\mathrm{P(off)}]$, where P(on) and P(off) are the power of the reflected light on and off resonance, respectively, and the univariate X is range. In this example one is primarily interested in estimating the first derivative since $-m'(x)$ is proportional to mercury concentration at range x.

The raw data are shown in Figure 2. The steep negative slope when $550 \leq$ range ≤ 600 corresponds to an emissions plume containing mercury. One can see that the variance of Y is an increasing function of X. Local polynomial regression is particularly well suited to this problem since it easily handles both derivative estimation and estimation of nonconstant conditional variance.

The third example comes from a geotechnical engineering study of factors affecting soil movement during earthquakes. Tremors during a major quake increase the pressure of the water in the soil pores causing saturated sandy soil to liquefy by reducing friction between sand particles. At least three factors are believed to contribute to soil movements: thickness of the liquefiable layer, thickness of the top layer of nonliquefiable soil, and slope or gradient of the ground. All three factors must be smoothed and interpolated from sample data. Here we present only the results for soil gradient. As in the biomonitoring example, we have two-dimensional spatial data with $\boldsymbol{X}$ being spatial location. Here Y is elevation. As in the LIDAR example, our primary interest is not in m but a derivative, the gradient $(\partial/\partial x_1\, m, \partial/\partial x_2\, m)$.

2 Local Regression

This section is an introduction to local polynomial regression and a brief survey of some of the literature. We will assume that the pairs $\{(\boldsymbol{X}_i, Y_i)\}_{i=1}^n$ satisfy

$$Y_i = m(\boldsymbol{X}_i) + \sigma(\boldsymbol{X}_i)\epsilon_i, \tag{1}$$

where $\epsilon_1, \ldots, \epsilon_n$ are mutually independent, $E(\epsilon_i) = 0$, $\mathrm{Var}(\epsilon_i) = 1$, and m and σ are smooth functions. Because of this smoothness, $m(\boldsymbol{x})$ and $\sigma(\boldsymbol{x})$ can be estimated by fitting a polynomial to observations $(\boldsymbol{X}_i, Y_i)$ with $\boldsymbol{X}_i$ close to $\boldsymbol{x}$. To measure closeness, we use a nonnegative kernel function K defined on $\Re^d$, where d is the dimension of the $\boldsymbol{X}_i$'s, and a symmetric, positive-definite $d \times d$ bandwidth matrix $\boldsymbol{H}$. Typically, $K(\boldsymbol{u})$ is spherically contoured and decreasing in $\|\boldsymbol{u}\|$. All examples in this paper use the spherically-contoured Epanechnikov kernel

$$K(\boldsymbol{u}) = (1 - \|\boldsymbol{u}\|^2)_+.$$

This kernel is known to be optimal both for estimating m and for estimating derivatives of m of any order (Fan et al., 1995). We define $K_H(\boldsymbol{u}) = |\boldsymbol{H}|^{-1} K(\boldsymbol{H}^{-1}\boldsymbol{u})$. If K is spherically contoured then the contours of K_H are ellipses of the form $\{\boldsymbol{t} : \boldsymbol{t}^T \boldsymbol{H}^{-2} \boldsymbol{t} = k\}$, $k > 0$. Then $K_H(\boldsymbol{X}_i - \boldsymbol{x})$ is the weight given to $\boldsymbol{X}_i$ when estimating $m(\boldsymbol{x})$. For spatial data, it is often reasonable for $\boldsymbol{H}$ to be a scalar multiple of the identity matrix. In other cases, $\boldsymbol{H}$ could be a general diagonal matrix giving appropriate scaling to the different components of $\boldsymbol{x}$.

Nondiagonal $\boldsymbol{H}$ may be desirable when the $\boldsymbol{X}_i$'s are concentrated on a ridge not parallel to a coordinate axis or when m has a ridge.

For concreteness, we will assume that the dimension of $\boldsymbol{x}$ is $d = 2$. Let $\widehat{m}(\boldsymbol{x}; \boldsymbol{H}, p)$ be our estimate of m using bandwidth matrix $\boldsymbol{H}$ and local polynomials of degree p. First assume that $p = 0$, i.e., that we are using local constant approximations. Then $\widehat{m}(\boldsymbol{x}; \boldsymbol{H}, 0) = \widehat{\beta}_0$ where $\widehat{\beta}_0$ minimizes:

$$\sum_{i=1}^{n} \{Y_i - \beta_0\}^2 K_H(\boldsymbol{X}_i - \boldsymbol{x}).$$

It is easy to see that

$$\widehat{\beta}_0 = \frac{\sum_{i=1}^n Y_i \, K_H(\boldsymbol{X}_i - \boldsymbol{x})}{\sum_{i=1}^n K_H(\boldsymbol{X}_i - \boldsymbol{x})}.$$

This is the Nadaraya-Watson kernel regression estimator, which has been extensively studied but is no longer recommended because of unsuitable bias properties. First, the Nadaraya-Watson estimator may have large bias in the "boundary region" defined as the set of $\boldsymbol{x}$'s where the support of $K_H(\cdot - \boldsymbol{x})$ extends beyond the support of the $\boldsymbol{X}_i$'s; see Fan and Gijbels (1992) and Ruppert and Wand (1994). (If the X_i's are iid with density f, then their support is the support of f. In other cases we might define their support to be their convex hull.) Another problem is that the Nadaraya-Watson estimator may be badly biased at interior points due to unequal spacing of the X_i's.

An improvement over the Nadaraya-Watson estimator is offered by the local linear estimator which can "adapt" to unequal spacing and boundaries (Fan, 1992 and Ruppert and Wand, 1994). This estimator is $\widehat{m}(\boldsymbol{x}; \boldsymbol{H}, 1) = \widehat{\beta}_0$ where $(\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\beta}_2)$ minimizes

$$\sum_{i=1}^{n} \{Y_i - [\beta_0 + \beta_1(X_{i1} - x_1) + \beta_2(X_{i2} - x_2)]\}^2 K_H(\boldsymbol{X}_i - \boldsymbol{x}),$$

The local linear estimator may still be badly biased in regions of high curvature, e.g., where m has a sharp peak. A substantial reduction in this curvature-induced bias is possible if one uses the local quadratic estimator defined as $\widehat{m}(\boldsymbol{x}; \boldsymbol{H}, 2) = \widehat{\beta}_0$ where $(\widehat{\beta}_0, \widehat{\beta}_1,$ $\widehat{\beta}_2, \widehat{\beta}_{11}, \widehat{\beta}_{12}, \widehat{\beta}_{22})$ minimizes

$$\sum_{i=1}^{n} \Big\{ Y_i - [\beta_0 + \beta_1(X_{i1} - x_1) + \beta_2(X_{i2} - x_2) + \beta_{11}(X_{i1} - x_1)^2$$
$$+ \beta_{12}(X_{i1} - x_1)(X_{i2} - x_2) + \beta_{22}(X_{i2} - x_2)^2] \Big\}^2 K_H(\boldsymbol{X}_i - \boldsymbol{x}).$$

Simulation studies generally show that local polynomial models of degree $p = 3$ and higher offer little, if any, improvement over local quadratic models, unless one is estimating second or higher derivatives.

Derivatives are easily estimated by using the appropriate derivative of the local polynomial model. Thus, if $p \geq 1$, then we can estimate the gradient of m by $(\widehat{\beta}_1, \widehat{\beta}_2)$, and for $p \geq 2$ one estimates the Hessian matrix by

$$\begin{pmatrix} \widehat{\beta}_{11} & \widehat{\beta}_{12} \\ \widehat{\beta}_{12} & \widehat{\beta}_{22} \end{pmatrix}.$$

Bias considerations show that one should use local polynomials whose order is at least one greater than the order of the derivative being estimated, e.g., at least local quadratics for the gradient and at least local cubics for the Hessian.

Figure 4 illustrates local linear regression for univariate x's. The true function m shown as a dotted curve is $m(x) = 5\exp(-(x-2)^2)$. The raw data are denoted by asterisks. The local linear estimate with a global bandwidth of $h = 1$ computed on an equally-spaced grid of 31 points is shown as a dashed curve. Estimation at the grid points, $x = 2$ and $x = 5$, is illustrated. The kernels at these two points are shown as dot-and-dashed curves. The local linear fits are shown as solid lines. Notice how these lines intersect $\widehat{m}$ at $x = 2$ and $x = 5$, respectively. To produce the dashed curve, $\widehat{m}$ was computed by the same process at the other 29 grid points and these 31 points were connected piecewise linearly.

3 Bandwidth Selection

The proper choice of $\boldsymbol{H}$ is crucial. If $\boldsymbol{H}$ is too small, meaning that the support of K_H is too small, then the estimator will use too few data locally and will be highly variable and rough. Conversely, if $\boldsymbol{H}$ is too large, then the estimator will be insufficiently local and hence highly biased. The proper choice of $\boldsymbol{H}$ will depend on σ^2, the sample size, the location of the $\boldsymbol{X}_i$'s, and p. Sometimes $\boldsymbol{H}$ is chosen subjectively by trial and error, but data-based, automatic bandwidth selectors are desirable, at the very least for comparison with user-selected bandwidths. Most bandwidth selectors attempt to minimize the mean squared error (MSE) of the local regression estimator. For simplicity we will assume that $\boldsymbol{H}$ is a scalar multiple of the identity matrix, i.e., $\boldsymbol{H} = hI$, $h > 0$. Suppose that we are estimating $m^{(\boldsymbol{r})}(\boldsymbol{x}) = \partial^{r_1+r_2}/\partial x_1^{r_1}\partial x_2^{r_2}\, m(\boldsymbol{x})$ where $\boldsymbol{r} = (r_1, r_2)$ is a pair of nonnegative integers. The estimator will be denoted by $\widehat{m}^{(\boldsymbol{r})}(\boldsymbol{x}; h, p)$.

One has a choice between global bandwidths where h is independent of $\boldsymbol{x}$ and local bandwidths where h depends on $\boldsymbol{x}$. We will consider only local bandwidths since they are more adaptable to unequally spaced $\boldsymbol{x}$'s, heteroskedasticity, and regions of high curvature of m, problems often encountered in practice. Let $\mathrm{BIAS}(h; \boldsymbol{x})$ and $\mathrm{VAR}(h; \boldsymbol{x})$ be the bias and variance of $\widehat{m}^{(\boldsymbol{r})}(\boldsymbol{x}; h, p)$ conditional on $(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$. Expressions for $\mathrm{BIAS}(h; \boldsymbol{x})$ and $\mathrm{VAR}(h; \boldsymbol{x})$ and asymptotic approximations to these expressions can be found in Ruppert and Wand (1994). A local bandwidth can be selected by minimizing an estimate of $\mathrm{MSE}(h; \boldsymbol{x}) = \mathrm{BIAS}^2(h; \boldsymbol{x}) + \mathrm{VAR}(h; \boldsymbol{x})$ at each $\boldsymbol{x}$. The estimate of $\mathrm{MSE}(h; \boldsymbol{x})$ will be developed by estimating $\mathrm{BIAS}(h; \boldsymbol{x})$ and $\mathrm{VAR}(h; \boldsymbol{x})$ separately.

The simplest method of estimating $\mathrm{VAR}(h; \boldsymbol{x})$, the method used here, is to substitute an estimate $\widehat{\sigma}^2(\cdot)$ into the exact expression for $\mathrm{VAR}(h; \boldsymbol{x})$, which depends only on $\sigma^2(\cdot)$ and $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$. The estimation of σ^2 is discussed in the next section. This method is somewhat computationally expensive, but is quite feasible with current hardware. Fan and Gijbels (1995) and, following them, Ruppert (1995) use this idea. (Earlier work, e.g., Ruppert, Sheather, and Wand (1995) substitutes $\widehat{\sigma}^2(\cdot)$ and an estimate of the density of

the $\boldsymbol{X}_i$'s into an asymptotic formula. This type of bandwidth estimator was sensible with the computing equipment (386 processors) of 1992 when the research in Ruppert, Sheather, and Wand began. However, asymptotics may give only crude approximations to $\mathrm{VAR}(h;\boldsymbol{x})$, particularly in boundary regions and it now seems better to avoid their use in bandwidth selection when possible.)

Earlier methods of estimating $\mathrm{BIAS}(h;\boldsymbol{x})$ substituted estimates of appropriate higher derivatives of $\widehat{m}(\boldsymbol{x};h,p)$ into asymptotic expressions for the bias. See Ruppert, Sheather, and Wand (1995) and Fan and Gijbels (1995). An alternative introduced by Ruppert (1995) is EBBS (Empirical Bias Bandwidth Selection) where one computes $\widehat{m}(\boldsymbol{x};h,p)$ for several values of h and uses these "data" to estimate bias.

Here is a brief introduction to EBBS; for details the reader should see Ruppert (1995). Fix $\boldsymbol{x}$, $\boldsymbol{r}$, and p with $p > r_1 + r_2$. Asymptotic theory shows that as $n \to \infty$ and $h \to 0$

$$\begin{aligned} E\widehat{m}^{(\boldsymbol{r})}(\boldsymbol{x};h,p) &= bc_0 + bc_{p+1-r_1-r_2}h^{p+1-r_1-r_2} + \\ &\cdots + bc_{p+t-r_1-r_2}h^{p+t-r_1-r_2} + o(h^{p+t-r_1-r_2}). \end{aligned} \tag{2}$$

All expectations and variances in this paper, e.g., in (2), are conditional on $\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n$, though this is not always made explicit by the notation. Here $bc_0 = m^{(\boldsymbol{r})}(x)$ and $\{bc_j\}_{j>0}$ represent bias and depend on higher derivatives of m and the density of the $\boldsymbol{X}$'s and possibly the derivatives of this density. For the estimation of bias, we use (2) as a model. We will use $t = 1$ or 2 in (2) since these values have worked well in simulations.

Suppose that we wish to estimate the bias of $\widehat{m}^{(\boldsymbol{r})}(\boldsymbol{x};h,p)$ for h in an interval (a,b). Here the lower bound, a, might, for example, give a span of 30%, i.e., be the smallest bandwidth such that at least 30% of the data get positive weight, while b might give a span of 90%. (Although 30% and 90% are just examples, I use these values as defaults.) Fix $M > t$ and let $a = h_1 < h_2 < \cdots < h_M = b$. M in the range 10 to 20 and geometrically spaced h_j's are recommended. For each j in $\{2,\ldots,M-1\}$ we estimate the bias of $\widehat{m}^{(\boldsymbol{r})}(\boldsymbol{x},h_j,p)$ as follows. First, fit model (2) to the "data" $\{(h_l,\widehat{m}^{(\boldsymbol{r})}(\boldsymbol{x},h_l,p)) : l = j-1,j,j+1\}$ by least squares and let $\widehat{bc}_0(j),\widehat{bc}_{p+1-r_1-r_2}(j),\ldots,\widehat{bc}_{p+t-r_1-r_2}(j)$ be the estimated coefficients. Then the estimated bias of $\widehat{m}^{(\boldsymbol{r})}(\boldsymbol{x},h_j,p)$ is

$$\widehat{\mathrm{BIAS}}(h_j;\boldsymbol{x}) = \sum_{l=1}^{t} \widehat{bc}_{p+l-r_1-r_2}h_0^{p+l-r_1-r_2}.$$

Now let $\widehat{\mathrm{VAR}}(h_j;\boldsymbol{x})$ be the estimated variance of $\widehat{m}^{(\boldsymbol{r})}(\boldsymbol{x};h_j,p)$, which, as we have mentioned, is computed by plugging an estimate of the function $\sigma^2(\cdot)$ into an exact formula for this variance. Finally, let

$$\widehat{\mathrm{MSE}}(h_j;\boldsymbol{x}) = \widehat{\mathrm{BIAS}}^2(h_j;\boldsymbol{x}) + \widehat{\mathrm{VAR}}(h_j,\boldsymbol{x}).$$

We then let our local bandwidth, $\widehat{h}(\boldsymbol{x})$, be the *smallest* local minimum of $\widehat{\mathrm{MSE}}(h;\boldsymbol{x})$ on the grid $h = \{h_2,\ldots,h_{M-1}\}$. The smallest local minimum is used since $\widehat{\mathrm{MSE}}(h;\boldsymbol{x})$ is known to

have a global minimum at $h = \infty$ because the bias in not properly modeled by (2) for very large h. The problem is that when h is too large, all features of m are smoothed away so there is little change in $\widehat{m}$ as h varies and therefore little apparent bias.

EBBS is illustrated in Figure 5 with simulated data using $m(x) = 5\exp(-(x-2)^2)$. In (a) the raw data are shown, and in (b) the estimator $\widehat{m}$ computed using global bandwidths of 1 (solid), 1.5 (dashed), and 2 (dotted) are shown. In (c), estimation of the bias of $\widehat{m}(x; h, 1)$ is shown for $x = 2$ and $h = 1.5$. The six asterisks are at $\widehat{m}$ using $h = 0.5$, 1, 1.5, 2, 2.5, and 3. The vertical dotted line goes through the value $h = 1.5$ where we are estimating bias. The solid curve is a least squares fit of the model $\widehat{m} = bc_0 + bc_2h^2$ to the three middle points corresponding to $h = 1$, 1.5, and 2. The horizontal dotted lines intersect this curve at $h = 0$ and $h = 1.5$, so that the vertical distance between these dotted lines is the empirical estimate of bias. In (d) the same process is illustrated at $x = 5$. Notice that the estimated bias at $x = 5$ is much less than at $x = 2$; the larger bias at $x = 2$ is also quite evident in (b). The solid curves at (c) and (d) do not appear to fit the "data" shown by asterisks. However, these points have *highly correlated* random errors since they are six estimates at the same x using the same data sets. This high correlation causes random deviation of $\widehat{m}(\boldsymbol{x}; h, p)$ from $m(\boldsymbol{x})$ as h varies to appear smooth rather than scattered as would be the case for independent random deviation. In (e) the EBBS local bandwidth is shown. Notice how the bandwidth is small for x between 2 and 4 where there is substantial curvature and increases rapidly between $x = 4$ and $x = 6$ where m is very flat. In (f) we see the true m (dashed) and $\widehat{m}$ with the EBBS bandwidth (solid). The high variability near the boundary can be seen. This high boundary variance is an unavoidable feature of nonparametric regression due to the extrapolation inherent in estimation at the boundary; see Ruppert and Wand (1994).

Several refinements of this methodology are discussed in Ruppert (1995). One is to cubicly interpolate $\widehat{\mathrm{MSE}}(\cdot; \boldsymbol{x})$ onto a finer grid of h-values before minimizing. This allows a greater choice of bandwidths without increasing computational time by much. Typically $\widehat{h}(\boldsymbol{x})$ and $\widehat{m}^{(\boldsymbol{r})}(\boldsymbol{x}; \widehat{h}(\boldsymbol{x}), p)$ are computed over some regularly spaced grid of $\boldsymbol{x}$-values. A second refinement is to smooth $\widehat{h}(\boldsymbol{x})$ over this grid after $\widehat{h}(\boldsymbol{x})$ has been found by minimization, but of course before computing $\widehat{m}^{(\boldsymbol{r})}(\boldsymbol{x}; \widehat{h}(\boldsymbol{x}), p)$.

Computing $\widehat{h}(\boldsymbol{x})$ is CPU intensive since it requires M computations of $\widehat{m}$, so $\widehat{h}(\boldsymbol{x})$ is generally found on a coarse $\boldsymbol{x}$-grid. One can interpolate $\widehat{h}(\boldsymbol{x})$ onto a finer grid before computing $\widehat{m}$. This idea is particularly useful if we need fitted values, $\widehat{m}(\boldsymbol{X}_i; \widehat{h}(\boldsymbol{X}_i), p)$, at the observed $\boldsymbol{X}_i$'s and n is large. For example, fitted values are used in the next section to form residuals in order to estimate the conditional variance.

For the two-dimensional spatial data in this paper, $\widehat{h}$ is computed on a regular 5×5 grid and then interpolated onto a regular 25×25 grid. Then $\widehat{m}$ is computed on this finer grid.

4 Estimation of Conditional Variance

One estimates the conditional variance of Y given $\boldsymbol{X}$ by calculating a preliminary estimate of the mean, forming residuals, and then smoothing the squared residuals. As in parametric modeling, there is a bias due to estimating the mean, but one can correct for this. A detailed study of this method can be found in Ruppert, Wand, Holst, and Hössjer (1995), and here we only present a synposis.

Let $e_i = Y_i - \widehat{m}(\boldsymbol{X}_i; h, p)$ with h "small," e.g., giving a 10 to 30% span, so that there is little bias when estimating $m(\boldsymbol{X}_i)$. Let $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$ and $\boldsymbol{e} = (e_1, \ldots, e_n)^T$. Since $\widehat{m}(\boldsymbol{X}_i; \widehat{h}(\boldsymbol{X}_i), p)$ is linear in $\boldsymbol{Y}$ for fixed $\widehat{h}(\boldsymbol{X}_i)$, there is an $n \times n$ "smoother matrix," S_1, depending on $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$, $\widehat{h}(\boldsymbol{X}_1), \ldots, \widehat{h}(\boldsymbol{X}_n)$, and p such that

$$\boldsymbol{e} = (I - S_1)\boldsymbol{Y}.$$

S_1 will be $n \times n$. Let $\boldsymbol{e}^2$ be obtained by squaring each component of $\boldsymbol{e}$. Let $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m\}$ be a grid of $\boldsymbol{x}$-values. Suppose that we estimate $\{\sigma^2(\boldsymbol{x}_j) : j = 1, \ldots, m\}$ by smoothing $\boldsymbol{e}^2$ by local polynomial regression, which corresponds to a second smoother matrix S_2. S_2 will be $m \times n$. Thus a naive estimator of $\sigma^2(\cdot)$ is $S_2\boldsymbol{e}^2$, "naive" meaning that we have not corrected for estimation of the mean function. Let Δ be the vector of diagonal elements of $S_1 S_1^T - 2S_1$. Ruppert et al. (1995) show that the bias of $\widehat{\sigma}^2(\boldsymbol{x}_j) = (S_2\boldsymbol{e}^2)_j$ is the sum of a term from the bias of $\widehat{m}(\boldsymbol{x}_j)$ plus $\sigma^2(\boldsymbol{x}_j)\Delta_j$. Because $\widehat{m}$ uses a small bandwidth we will ignore the first term. Then the bias-corrected estimate of the variance function in Ruppert et al. (1995) is

$$\frac{S_2\boldsymbol{e}^2}{\mathbf{1} + S_2\Delta},$$

where $\mathbf{1}$ is a vector of ones and where the division is coordinate-wise.

5 Examples

In this section, the examples introduced in Section 1 are continued. All computations were performed using programs written in MATLAB by the author.

5.1 Biomonitoring of mercury

The conditional mean and variance of mercury concentration were estimated by local quadratic and local linear regression, respectively, with bandwidths selected by EBBS. Because of the small sample sizes, local ridge regression, as discussed in Cleveland and Loader (1995), was used to stabilize the estimators. Ridge regression shrinks the local pth degree polynomial estimator towards the local $(p-1)$th degree polynomial estimator and can be viewed as a way of interpolating between integer values of p. Thus, in this example, ridge regression shrunk the local quadratic estimate towards the local linear estimate in the case of the mean. For estimation of the variance, ridge regression shrunk the local linear estimator towards the local constant estimator. As explained in Ruppert (1995), EBBS bandwidth

selection is directly applicable to local ridge regression. However, the ridge coefficient, which determines the amount of shrinkage, was chosen by trial and error.

The conditional mean is plotted in Figure 1(a) and (b), and the conditional variance is in Figure 6(a) and (b). The conditional variance does not appear to be constant. The log transformation is somewhat variance stabilizing, but the estimated conditional variance of log(Hg) still appears nonconstant, though with only 22 observations one cannot say anything definitive.

The local bandwidths for estimating the mean and variance functions, respectively, are plotted in Figure 6(c) and (d). The local bandwidth for the mean is relatively constant, which seems to be common for local quadratic estimators.

One might try to analyze these data using geostatistics, but the heteroskedasticity makes standard kriging methods inappropriate, since the variogram is not defined due to the nonstationary variance. In Figure 7 we have the sample covariogram in (a) and the sample variogram in (b). In (a), for each pair, $(\boldsymbol{X}_i, Y_i)$ and $(\boldsymbol{X}_{i'}, Y_{i'})$, we plot $D_{i,i'} = \|\boldsymbol{X}_i - \boldsymbol{X}_{i'}\|$ against $(Y_i - \bar{Y})(Y_{i'} - \bar{Y})$. A curve is also fit to these points by local linear regression.

In (b) we plot $D_{ii'}$ versus $(Y_i - Y_{i'})^2$ with a local linear fit. This empirical variogram in (b) is not monotonically increasing in D as is typical of stationary data. The reason for the nonmonotonicity is that the largest distances come from pairs where both Y's have low variability. The ability of local polynomial regression to model heteroskedasticity is a distinct advantage here over standard geostatistical methods. Figure 7(c) and (d) shows the sample covariogram and variogram for the residuals. There seems to be little evidence of correlation among the residuals and again the heteroskedasticity is evident since the variogram is not monotonically increasing; in fact, it is decreasing except at very small distances.

The primary scientific question behind this study is whether the incinerator is contributing to mercury deposition. To test this we use the hypotheses

H_0: Y_i is independent of $\boldsymbol{X}_i$

H_1: $E(Y_i|\boldsymbol{X}_i)$ is not constant

We will use a test statistic familiar in linear regression

$$R^2 = 1 - \frac{\sum_{i=1}^n (Y_i - \widehat{Y}_i)^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2}.$$

The null distribution of R^2 is unknown, but this problem can be circumvented by a permutation test. R^2 was computed for the original data and for 999 pseudo data sets obtained by randomly permuting the Y_i's while keeping the $\boldsymbol{X}_i$'s fixed. For each pseudo data set, the $\widehat{Y}_i$'s were obtained by applying EBBS to that data set, i.e., *not* using the bandwidths computed from the original data. Under the null hypothesis, all possible rankings of the 1,000 R^2 values are equally likely. Ranking from largest to smallest, the original data set had rank 118 giving the p-value 0.118. Opsomer et al. (1995) tested these hypotheses by an

approximate F-test and obtained a p-value of 0.101. However, they were concerned because their p-value was based on asymptotics but there were only 22 observations. Moreover, their p-value did not account for the fact that the bandwidth was data-based. Since we compute a separate bandwidth for each data set, the extra variability due to bandwidth selection is taking fully into account by our test.

In Figure 1(c), mercury concentration (Y_i) is plotted against distance from the incinerator (D_i) with a local linear smooth superimposed. The impression is that there is a flat background concentration with a rise near the incinerator. The slight dip in the estimated concentration when D is around 10 is likely due to random variation, or perhaps it is an artifact of the way the sampling locations were located.

We also tested H_0 as given above against a more specific alternative hypothesis,

H_2: $E(Y_i|\boldsymbol{X}_i)$ is a function of distance from the incinerator

This test used the test statistic R^2 as defined above, except now the fitted values, $\widehat{Y}_i$, were found by regressing Y_i on D_i by local linear regression, both for the original data and for the pseudo data sets creating by permuting the Y_i's. This time the original data set had an R^2 value that was second largest among the 1000 data sets, giving a p-value of 0.002 when testing H_0 against H_2.

This p-value is of course far smaller than what Opsomer found when testing against H_1. This fact illustrates the general principle that one creates a more powerful test by using a more specific alternative hypothesis. Of course, one must take care that the more specific alternative was not formulated after examining the data. However, since H_2 is the alternative that the researchers were considering when designing their study, 0.002 is the more appropriate p-value. Thus, the new finding in this paper is that the effect of the incinerator on mercury deposition is not borderline significant as had been thought, but rather highly significant.

5.2 LIDAR

As has been mentioned, the raw LIDAR data are plotted in Figure 2. The mean function, m, and its first derivative, m', were estimated by local linear and local quadratic regression, respectively, and are plotted in Figure 8(a) and (c). The sharp trough in m' when the range, x, is slightly less than 600 indicates a peak in mercury concentration there.

The EBBS bandwidths for m and m' are plotted in Figure 8(b) and (d). The local bandwidths adapt to changes in curvature between range = 500 and range = 600 by becoming smaller there. Also, EBBS optimizes the bandwidths for the order of the derivative being estimated, and, as might be expected, uses a somewhat larger bandwidth for m' than for m.

5.3 Soil elevation and gradient

Now we continue the analysis of the soil gradient data illustrated in Figure 3. Elevation, which is shown in that figure both as a surface plot and as a contour plot, shows a hill in the center of the sampling region. Smaller hills in the boundary region are likely to be artifacts due to lack of data in this region. This is not an issue, since estimates of the gradient are only needed in the central region.

The partial derivatives of elevation with respect to x_1 and x_2 were estimated separately, though the separately estimated optimal bandwidths were rather similar. The EBBS bandwidths were calculated on a 5×5 grid and then interpolated onto a 25×25 grid where the derivatives were estimated. We experimented with several methods of depicting the estimated gradient. The best method seemed to be a quiver plot where the gradient at each grid point is represented by an arrow pointing in the gradient direction wtih length proportional to the gradient length.

These estimates will be used in further analyses such as modeling the relationship between the direction of the gradient and the direction of soil movement by angular regression (Rivest, 1995).

6 Discussion

We have shown that local polynomial regression is a simple but powerful method of nonparametric regression. Although not especially designed for environmental statistics or spatial data, local polynomial regression is very useful in both areas. Two important features of local regression are its ability to model nonconstant variance and the ease with which it can estimate derivatives.

The degree of smoothing is controlled by the all important bandwidth matrix. By using local bandwidth matrices, we can adapt to features typical of real data, e.g., unequal spacing of the $\boldsymbol{x}$'s. Because we have smoothed the local bandwidths, they might more properly be called "partial local"—see Hall, Marron, and Titterington (1995). Data-based bandwidth matrices can be found by several methods including Ruppert's (1995) EBBS that is illustrated in this paper. So far, EBBS has been implemented only when there is a single bandwidth parameter, e.g., if the bandwidth matrix is a scalar multiple of the identity. Another potentially useful method of bandwidth selection is "double smoothing," which so far has only been proposed for local constant smoothing of one-dimensional data; see Härdle, Hall, and Marron (1992).

We have illustrated local regression applied to two-dimensional spatial data. The extension to three dimensions is straightforward. A more interesting extension would be to spatio-temporal data. The general theory extends immediately but selection of the bandwidth matrix requires care, and clearly one needs more flexibility than a scalar multiple of the identity matrix can provide.

In model (1) we have assumed that the ϵ_i's are mutually independent. Thus, any

association betwen nearby data is due to the smoothness of m. Another possibility would be to assume some dependence between the ϵ_i's. Then we would distinguish between long range dependence due to m and short range dependence due to the errors. When estimating m, typically a larger bandwidth is needed than under independence. Some work on data-based bandwidth selection under correlation is found in Opsomer (1995), but one-dimensional x's are assumed.

Geostatistics models all association as due to a random process, except possibly for a parametrically modeled trend. Altman (1995) compares kriging, nonparametric smoothing (but not local polynomial regression), a combination of kriging and smoothing, and simple interpolation.

Local polynomial estimation is being extended beyond estimation of mean and variance functions. For example, most time series and spatial data models assume stationary correlation structure. When this is absent, analysts may divide the data into blocks with, it is hoped, nearly stationary behavior within blocks. However, the assumption that a correlation function is piecewise constant can be a crude approximation. Hyndman and Wand (1995) use local polynomial estimation to model slowly changing correlation structure in univariate time series.

An interesting account of early work on local polynomial regression can be found in Cleveland and Loader (1995). For a good introduction to kernel-based smoothing including local polynomial regression, the reader should consult Wand and Jones (1995). The program *loess* implements local polynomial regression but without automatic bandwidth selection. Instead the user specifies the span used. See Cleveland, Grosse, and Shyu (1993). Some interesting applications of local regression, including prediction of an air pollutant (ozone) and another study on predicting NO_x emissions from internal combustion engines, can be found in Cleveland and Devlin (1988).

Acknowledgement

The biomonitoring and geotechnical engineering examples came from the Environmental Statistics Program at Cornell. I thank my co-director George Casella and the past and present students in this program for making it an exciting place for interdisciplinary work and an excellent source of interesting statistical problems. The biomonitoring data was kindly supplied by Anthony Carpi and Jean-Didier Opsomer. The LIDAR data were given to me by Dr. Ulla Holst. The soil elevation data was supplied by Thomas O'Rourke, Andrew Schulman, and Samantha Williams. Figure 4 was adapted from a hand-sketched figure given to the author by Matt Wand for use in lectures. I thank Andrew Schulman, Jean-Didier Opsomer, and Steve Marron for comments on an earlier draft.

References

Altman, N. (1994). "Krige, Smooth, Both or Neither," Technical Report, Biometrics Unit, Cornell University, Ithaca, New York.

Cleveland, W.S. and Devlin, S.J. (1988), "Locally-weighted regression: an approach to regression analysis by local fitting," *Journal of the American Statistical Association*, 83, 597–610.

Cleveland, W.S., Grosse, E., and Shyu, W.M. (1993), "Local regression models," in *Statistical Models in S*, edited by J.M. Chambers and T.J. Hastie, pp. 309–376, Chapman & Hall, New York and London.

Cleveland, W.S. and Loader, C. (1995). Smoothing by local regression: principles and methods. Preprint.

Cressie, N.A.C., (1991). *Statistics for Spatial Data*, John Wiley & Sons, New York.

Fan, J. (1994). "Design-adaptive nonparametric regression," *Journal of the American Statistical Association*, 87, 998–1004.

Fan, J., Gasser, T., Gijbels, I, Brockman, M, and Engel, J. (1995). "On nonparametric estimation via local polynomial regression," Manuscript.

Fan, J., and Gijbels, I. (1992). Variable Bandwidth and local linear smoothers. *The Annals of Statistics*, 20, 2008–2036.

Fan, J., and Gijbels, I. (1995), "Data-driven bandwidth selection in local polynomial fitting: variable bandwidth and spatial adaptation," *Journal of the Royal Statistics Society, Series B*, 57, 371–394.

Friedman, J. (1991). "Multivariate adaptive regression splines (with discussion)," *The Annals of Statistics*, 19, 1–141.

Green, P.J., and Silverman, B.W. (1994). *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*, Chapman & Hall, London.

Hall, P., Marron, J.S., and Titterington, D.M. (1995). "On partial local smoothing rules for curve estimation," *Biometrika*, 82, 575–588.

Härdle, W., Hall, P., and Marron, J.S., (1992), "Regression smoothing parameters that are not far from their optimum," *Journal of the American Statistical Association*, 87, 227-233.

Holst, U. Hössjer, Ola, Björklund, C., Ragnarson, P., and Edner, H. (1994). "Locally weighted least squares kernel regression and statistical evaluation of LIDAR measurements," Manuscript.

Hyndman, R.J., and Wand, M.P. (1995). "Nonparametric autocovariance function estimation," Manuscript. (http://www.maths.monash.edu.au/~hyndman/papers.html)

Opsomer, J. (1995). "A framework for estimating an unknown function by local polynomial regression when the errors are correlated," Preprint 95-44, Statistical Laboratory, Iowa State University. (http://www.public.iastate.edu/~jopsomer/research.html)

Opsomer, J.D., Agras, J., Carpi, A., and Rodrigues, G. (1995). "An application of locally weighted regression to airborne mercury depostion around an incinerator site," *Environmetrics*, 6, 205–219.

Ragnarson, P. (1994). *Optical Techniqes for Measurement of Atmospheric Trace Gases*, Lund Reports on Atomic Physics, LRAP-152, Department of Physics, Lund Institute of Technology.

Rivest, L.-P. (1995). A decentered predictor for angular regression. Mauscript.

Ruppert, D. (1995). Empirical-bias bandwidths for local polynomial nonparametric regression and density estimation. Manuscript. (To obtain a compressed postscript file on UNIX machines anonymous ftp to ftp.orie.cornell.edu/pub/techreps/TR1137.ps.Z)

Ruppert, D. and Wand, M.P. (1994), "Multivariate locally weighted least squares regression," *The Annals of Statistics*, 22, 1346–1370.

Ruppert, D. and Wand, M.P., Holst, U., and Hössjer, O. (1995), "Local polynomial variance function estimation," manuscript. (ftp.orie.cornell.edu/pub/techreps/TR1132.ps.Z)

Ruppert, D., Sheather, S.J., and Wand, M.P. (1995), "An effective bandwidth selector for local least squares regression," *Journal of the American Statistical Association*, to appear.

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

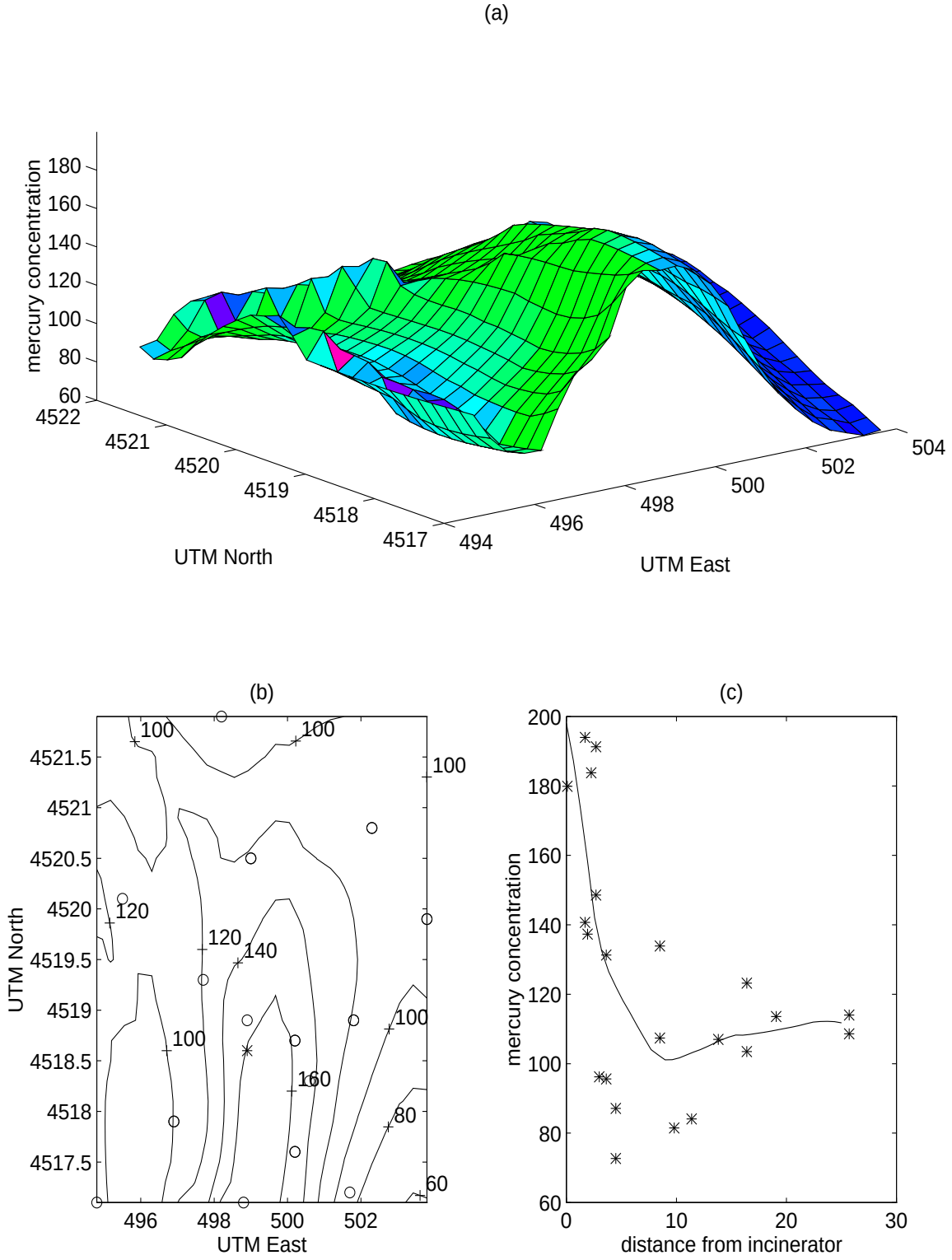

Figure 1: *Biomonitoring example. (a) Surface plot of mercury concentration estimated by local quadratic ridge regression. (b) Sampling locations (circles) and location of the incinerator (asterisk) with contours of mercury concentration as in (a). (c) Plot of mercury concentration versus distance from the incinerator with a local linear smooth added.*

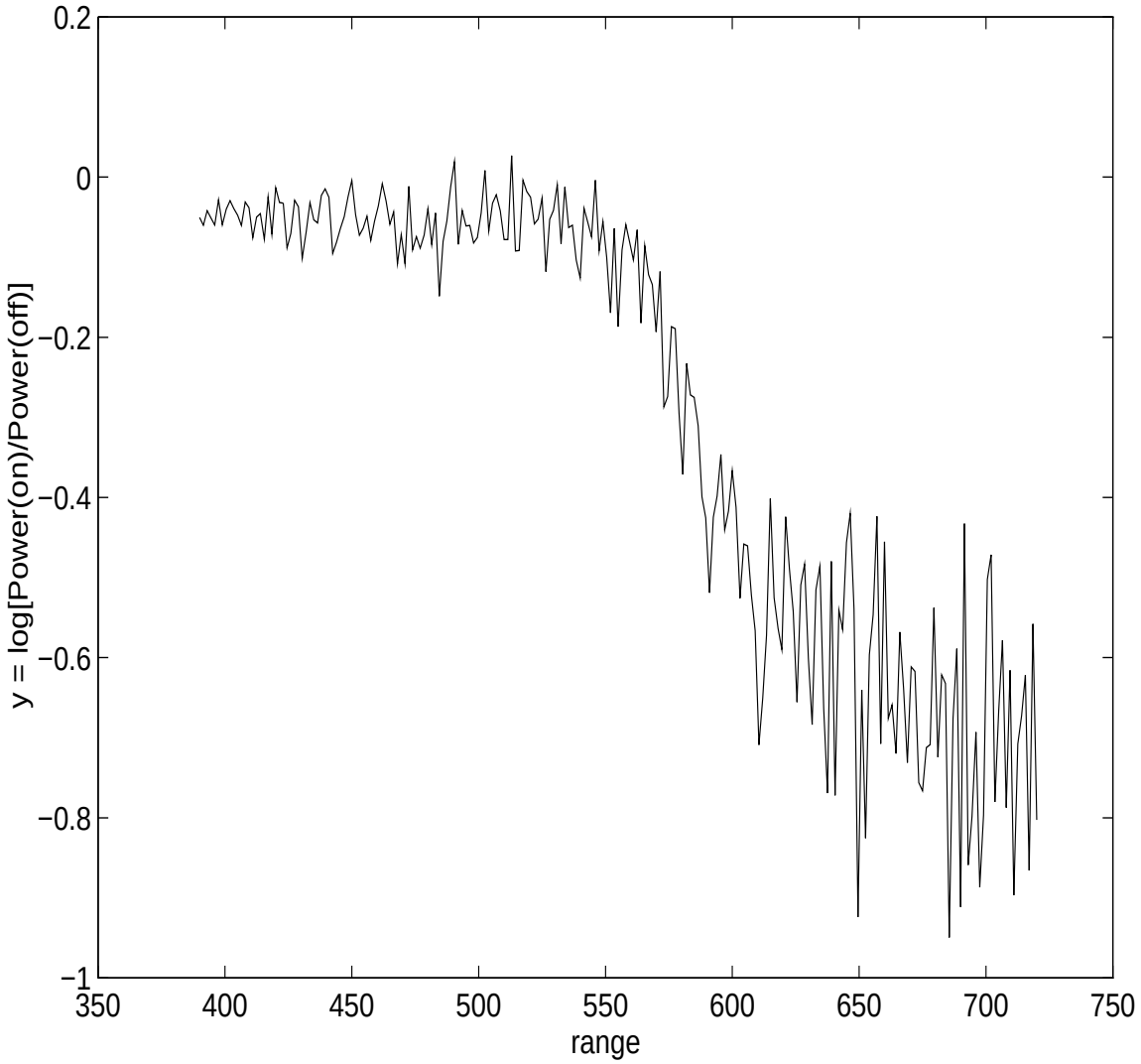

Figure 2: *LIDAR example. Raw data. Response plotted against range.*

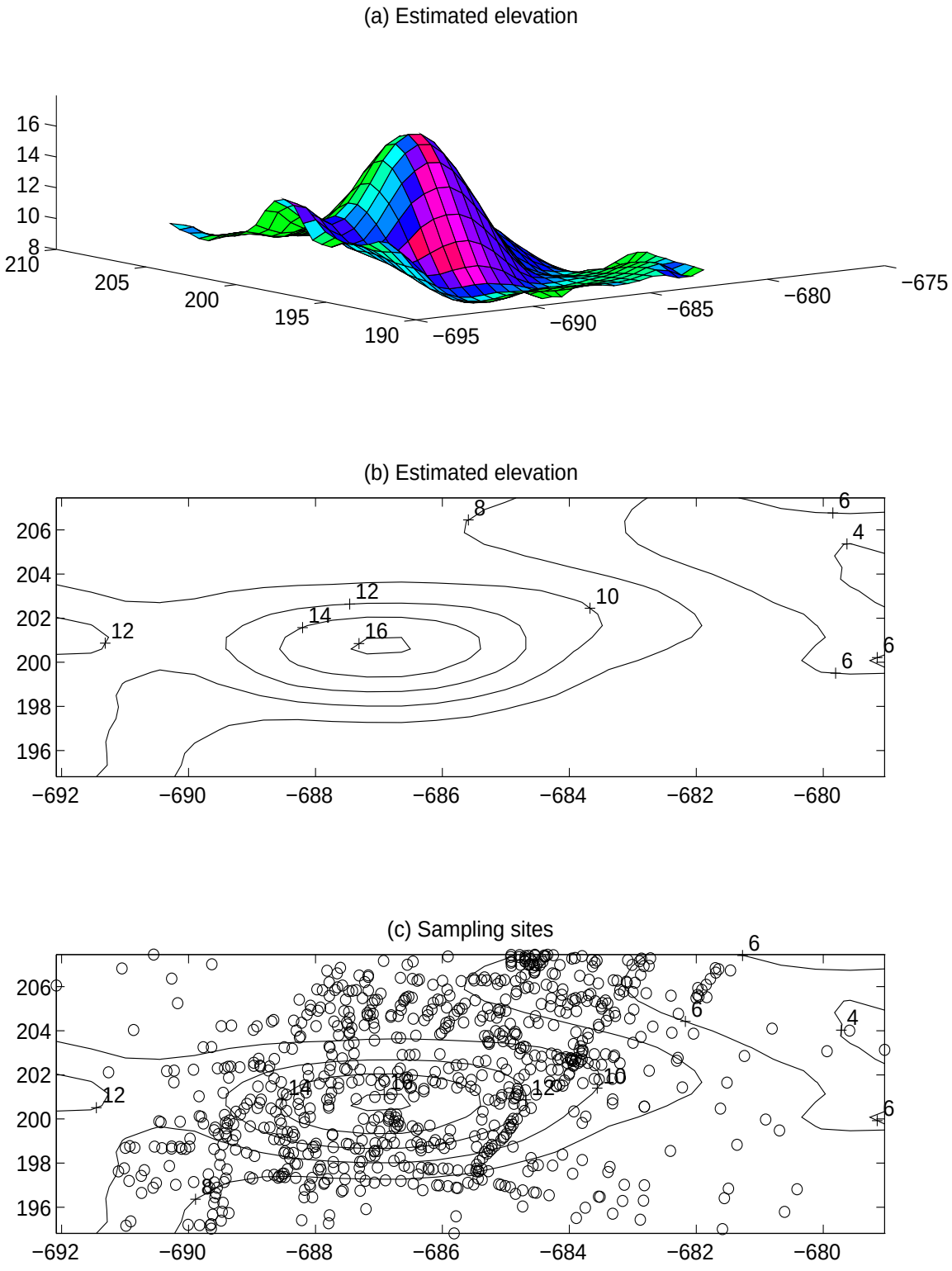

Figure 3: *Noshiro data. (a) and (b) Elevation. (c) Elevation and sampling locations.*

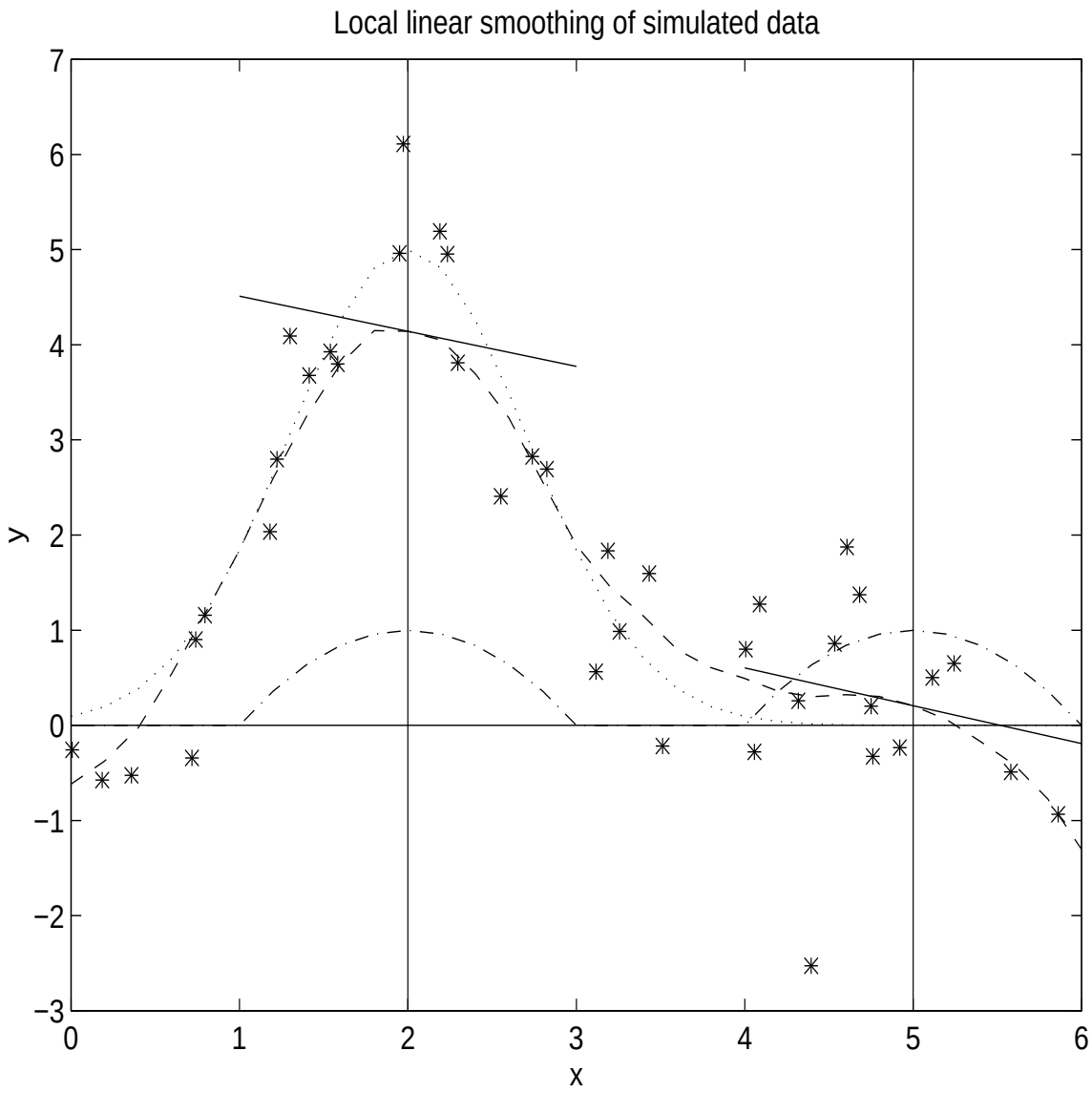

Figure 4: *Illustration of local linear regression with simulated data. See text for explanation.*

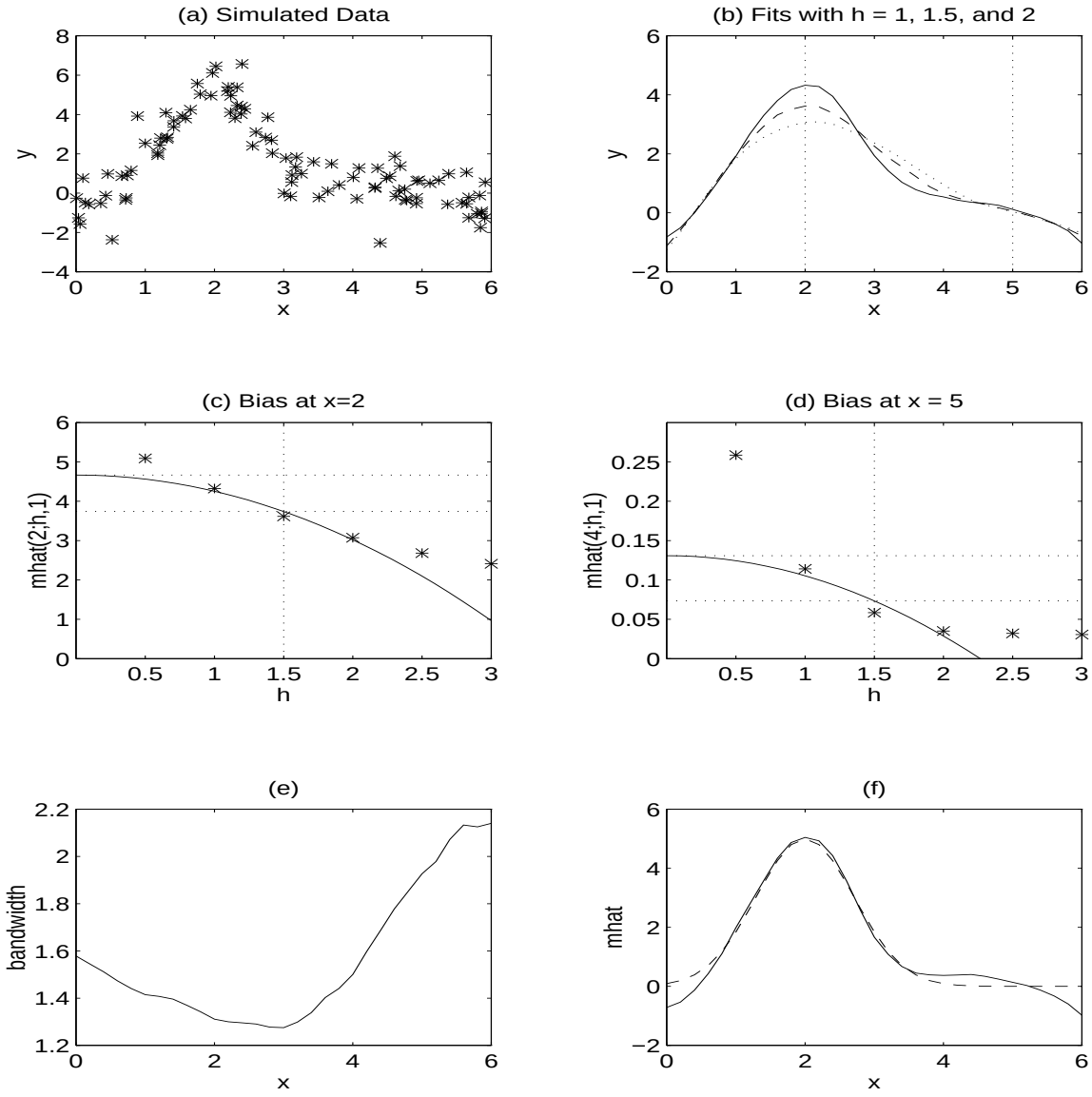

Figure 5: *Illustration of empirical bias estimation with simulated data. See text for explanation.*

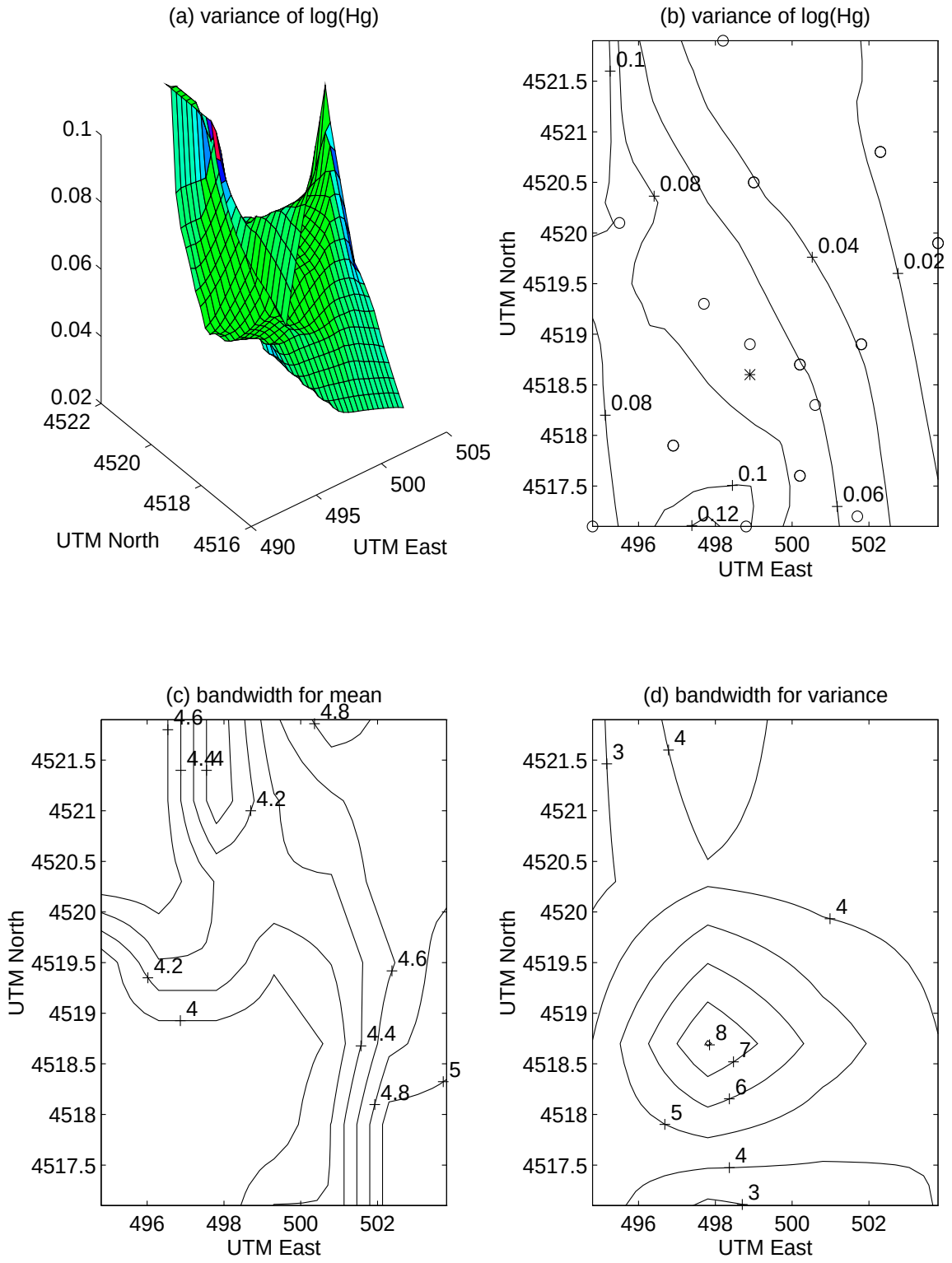

Figure 6: *Biomonitoring example. (a) Estimated variance of log(Hg) by local linear ridge regression. (b) Sampling locations (circles) and location of the incinerator (asterisk) with contours of the variance of log(Hg) estimated by local linear regression. (c) Bandwidths used in (a). (d) Bandwidths used in (b).*

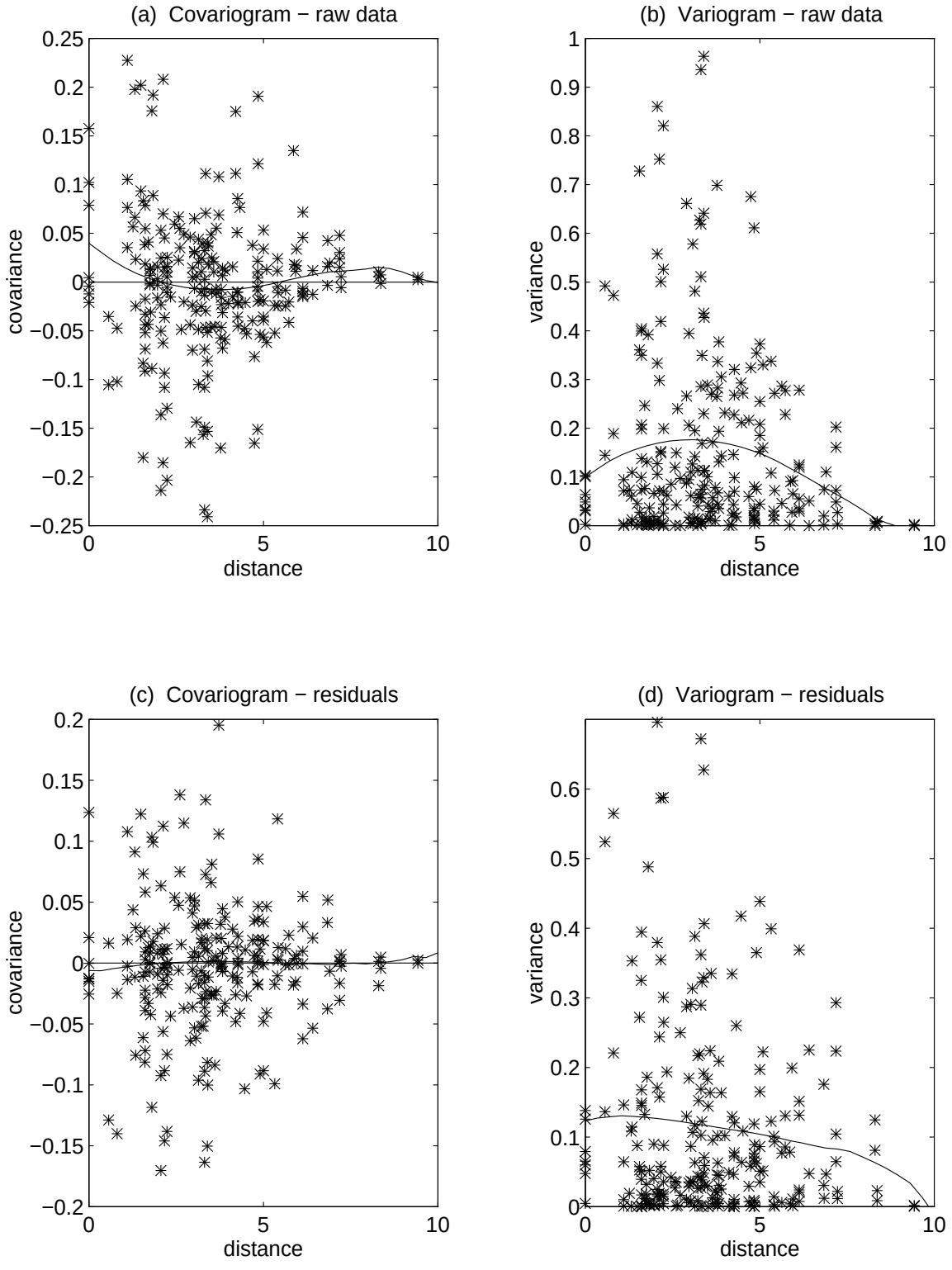

Figure 7: *Biomonitoring example. Sample covariograms and variograms of raw data and residuals.*

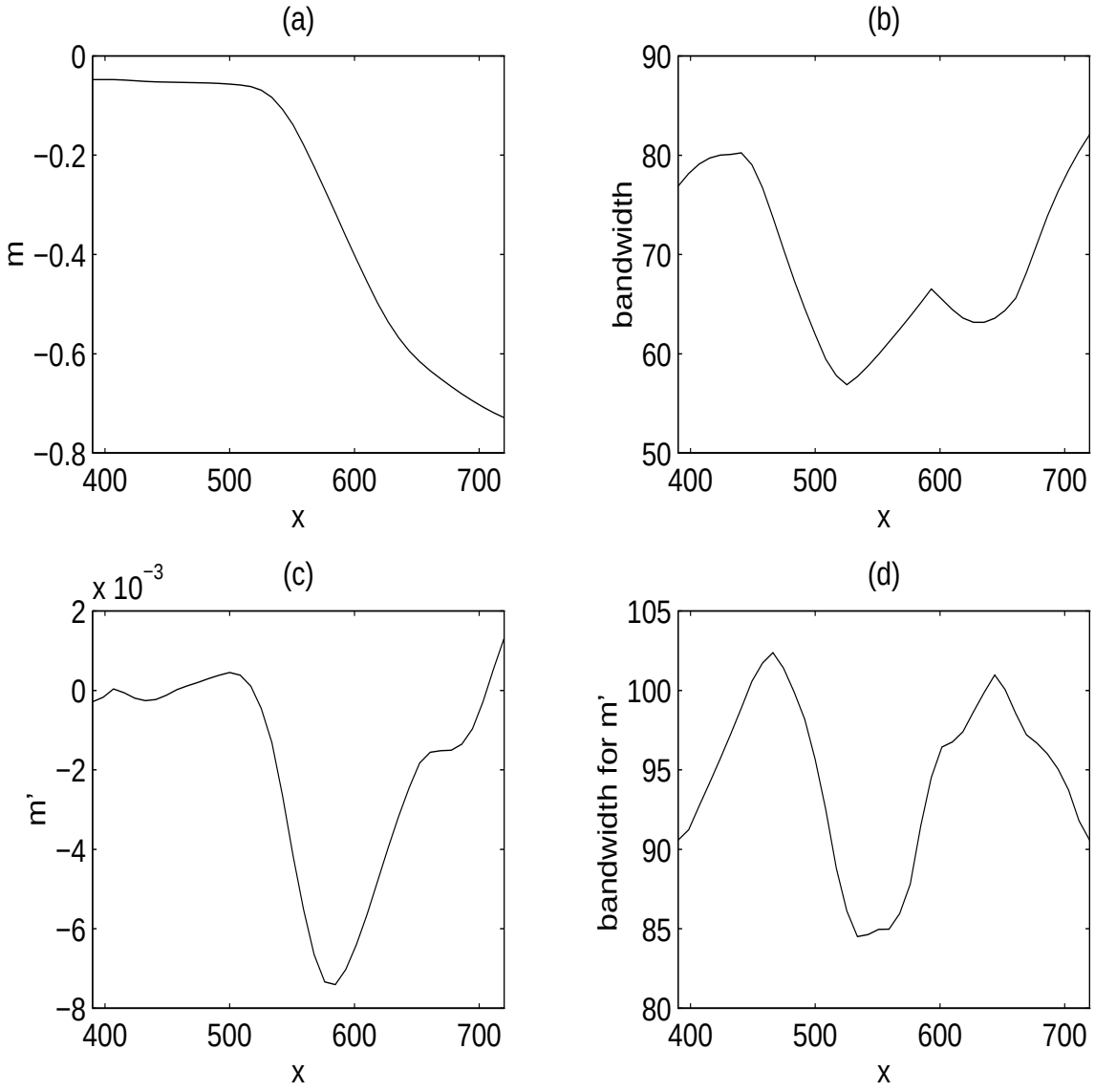

Figure 8: *LIDAR example. Raw data. (a) Local linear estimate of* m*. (b) Bandwidth for estimating* m*. (c) Local quadratic estimate of* m'*. (d) Bandwidth for estimating* m'*.*

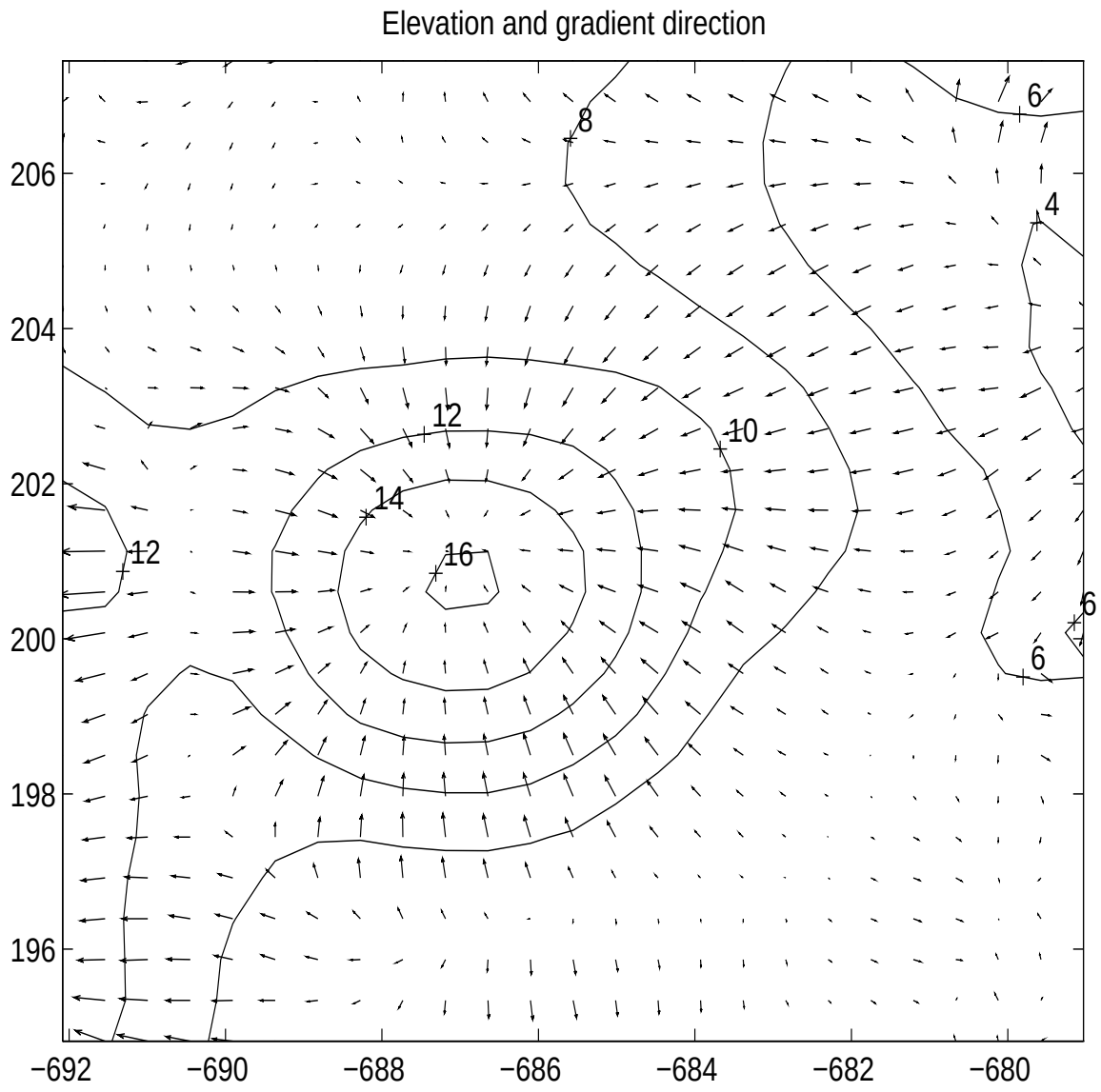

Figure 9: *Noshiro data. Elevation (contours) and gradient (arrows).*