The Annals of Statistics
1998, Vol. 26, No. 6, 2218–2245

DISCONTINUOUS REGRESSION SURFACES FITTING

BY PEIHUA QIU

University of Minnesota

We suggest a three-stage procedure to recover discontinuous regression surfaces when noisy data are present. In the first stage, jump candidate points are detected using a jump detection criterion. A local principal component line is then fitted through these points in a neighborhood of a design point. This line provides a first-order approximation to the true jump location curve in that neighborhood. In the third stage, observations on the same side of the line as the given point are combined using a weighted average procedure to fit the surface at that point. If there are no jump candidate points in the neighborhood, then all observations in that neighborhood are used in the surface fitting. If, however, the center of the neighborhood is on a jump location curve, only those observations on one side of the line are used. Thus blurring is automatically avoided around the jump locations. This methodology requires $O(N(k^*)^2)$ computation, where N is the sample size and k^* is the window width. Its assumptions on the model are flexible. Some numerical results are presented to evaluate the surface fit and to discuss the selection of the window widths.

1. Introduction. This paper provides a methodology to fit discontinuous regression surfaces (DRSs) in the presence of noisy data. Geologists, for example, often need to estimate mine surfaces from mineral samples. Because of the earth's movement, the mine surface often splits into several segments. It is important to know the split locations of the mine surface so precautionary actions can be taken to reduce accidents. In image processing, the image intensity function has step discontinuities (called step edges) at the outlines of the objects [Gonzalez and Woods (1992)]. Since much of the information in an image is conveyed by the edges and our eye–brain system has evolved to extract edges by preprocessing that begins right at the retina [Bracewell (1995), Chapter 5], edge detection and edge-preserving image reconstruction are important topics in image analysis. From a statistical viewpoint, each of these problems can be regarded as an application of DRSs fitting. We find similar problems in many other application fields, such as meteorology and oceanography.

Figure 1b displays a noisy data set. The original regression surface is shown in Figure 1a. It has jumps along a circle. If we ignore the jump structure, then "blurring" is present in the surface fitting around the jump

Received May 1997; revised July 1998.

AMS 1991 *subject classifications*. Primary 62G05; secondary 62G20.

Key words and phrases. Discontinuous regression surfaces, image processing, jump detection criterion, jump location curves, least squares coefficients, principal component line, threshold value.

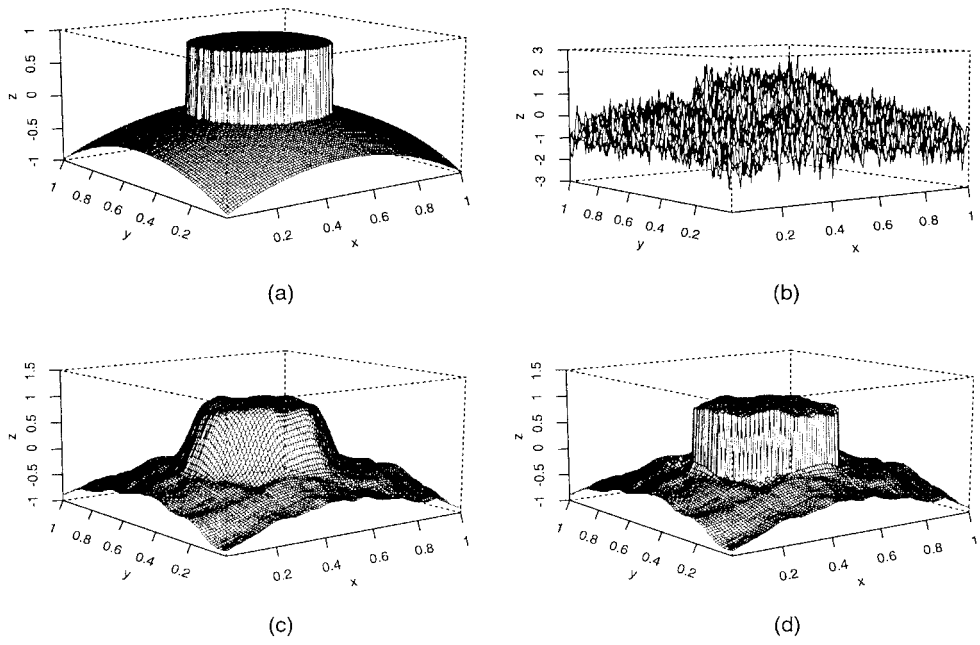

FIG. 1. (a) *The true regression surface*; (b) *the noisy data*; (c) *the conventional kernel surface fit*; (d) *the fitted surface by our discontinuity-preserving procedure*.

positions (Figure 1c). Our purpose is to reduce such blurring or to make the surface fitting be discontinuity preserving (Figure 1d). Additional discussion of these plots is found in Section 6.

Image reconstruction is essentially the same problem as DRS fitting except that the image intensity function usually takes discrete values while the response variable in the regression set-up is generally continuous. In most situations, this is not a big difference, since the value of the intensity function most often has 256 levels (corresponding to an eight-bit record in computer memory). However, two-color images or images with more than two colors also play an important role in image processing, especially in statistical image analysis. For further discussion about types of images, see Switzer (1986) or Section 7.4 of Cressie (1991).

When the image has a limited number of colors, discriminant analysis is a natural choice to classify each pixel as the most probable color type. An obvious drawback of this method is that it ignores the relationship among neighboring pixels. Several modifications exist in the literature to partially overcome this difficulty: such as pre- or post-smoothing [Switzer (1983), Switzer, Kowalick and Lyon (1982)] or the use of contextual information [Owen (1984)]. Among the possible modifications that have been proposed, Markov random field (MRF) methods have become an active research area in recent years. The true image is assumed to be a MRF, or equivalently, has a

Gibbs distribution. Geman and Geman (1984) suggest reconstructing the image by maximizing a posteriori (MAP) with a restoration algorithm which is based on stochastic relaxation and annealing. Discontinuities are preserved by using a line process. Besag (1986) suggests maximizing the marginal posterior distribution at each pixel by using the iterated conditional modes (ICM) algorithm. See Besag, Green, Higdon and Mengersen (1995) for a discussion of recent developments in this area, especially for image reconstruction with Markov chain Monte Carlo (McMC).

Another kind of image reconstruction method uses a regularization framework approach [Titterington (1985)]. Under this framework, the reconstructed image is a solution of a minimization problem. The minimization criterion usually consists of two terms. The first term measures the distance between candidate image and data. The second term measures the smoothness of the candidate image or the interaction among neighboring pixels. A smoothing parameter controls their trade-off. Several authors [Sinha and Schunck (1992); Li (1995); Yi and Chelberg (1995)] suggest choosing this parameter locally to accommodate discontinuities. If a pixel is on or close to a jump location curve, as judged by the estimates of derivatives around that pixel, then the smoothing parameter is chosen to be zero or very small. This kind of regularization method is closely related to the maximum-entropy methods [Titterington (1985)] and the Bayesian methods [Li (1995)]. For image reconstruction in emission tomography, see Green (1990) and O'Sullivan (1995) and references cited there.

A natural way to recover the DRS is to initially detect the jump location curves (JLCs) and then to fit the regression surface separately in regions separated by the JLCs. In order to recover the surface in connected regions, the estimated JLCs should also have compact forms. To achieve this, some assumptions on the JLCs are needed. When the number of JLCs is known and the population candidate JLCs is also known, the best candidates by some criterion could be chosen from the population as estimates of the JLCs by solving a minimization problem [Müller and Song (1994); O'Sullivan and Qian (1994)]. In the case that there is only one JLC, Qiu (1997) suggested a so-called *rotational difference kernel estimator* of the JLC by using a rotation transformation. Korostelev and Tsybakov (1993) investigated several kinds of design and jump boundaries. They suggested approximating the JLCs by piecewise polynomials and then estimating the coefficients by maximum likelihood estimation. Rudemo and Stryhn (1994) studied two types of two-region image models with a univariate boundary representation and suggested a nonparametric histogram-like contour estimator. When the JLCs are completely known, Shiau, Wahba and Johnson (1986) fit the DRS by partial spline method.

In this paper, we introduce another approach to fit the DRS that reduces computational requirements while providing greater generality by weakening some of the assumptions traditionally found in the literature. The methodology consists of three steps. First, we detect a set of possible jump positions (we call them jump candidate points) with a jump detection criterion. These

jump candidate points may not form curves. This kind of flexibility allows us not to impose restrictive conditions on the JLCs. Second, we fit a local principal component (PC) line through the jump candidate points in a neighborhood of a design point. This PC line can be regarded as a first-order approximation to the true JLC. The third stage uses only observations on the same side of the line as the given design point and combines them using a weighted average procedure to estimate the DRS. As a consequence, blurring is avoided in the neighborhood of JLCs. In the continuous regions, however, very few or no jump candidate points are detected in the neighborhood of each design point. Consequently, almost all observations in the neighborhood are actually used in the surface fit, making the fit the conventional local smoothing estimate.

The MRF methods and other related Bayesian methods generally require the following assumptions: (1) a prior distribution is assumed on the real image; (2) the observations have a known conditional distribution (usually Normal distribution) conditional on the real image; and (3) a correlation structure (such as MRF) is assumed on the real image. Some non-Bayesian methods assume various structure on the JLCs. For example, O'Sullivan and Qian (1994) assumed the JLCs to be "smooth, simple and closed" curves. Müller and Song (1994) assumed that the population of the JLCs was known. Qiu (1997) assumed that there was only one JLC and the JLC satisfied the Lipschitz (1) condition. The JLCs were assumed to have a univariate representation in Rudemo and Stryhn (1994). Shiau, Wahba and Johnson (1986) assumed the JLCs to be completely known. All of these non-Bayesian methods require that the number of JLCs should be known beforehand.

As more structure is assumed on the jump locations, the fitting of the jump surface becomes easier. But some real applications are excluded at the same time. Part of the purpose of this paper is to develop a method to fit the jump surface under mild conditions on the jump locations. We do not assume any prior distribution on the real image and on the observations. The assumptions we impose on the JLCs are also more general.

From a computational point of view, most Bayesian methods are designed for images with only several colors. One reason behind this is the process of maximizing a posteriori which is a main part of the Bayesian image reconstruction. If the number of image colors is large (for most images, this number is 256) or the color is continuous (this latter case is what interests us in this paper), then the maximization process at each pixel requires extensive computation. There are two maximization processes involved in the method suggested by Müller and Song (1994). One is to maximize the difference of the directional limits with respect to the direction at each point on the JLC, and the second is to search for the optimal JLC candidate from the candidate population. Each of these two processes requires extensive computation. In the minimax estimation of the jump boundary [Korostelev and Tsybakov (1993)], it is difficult numerically to compute the coefficients of the piecewise polynomials by using the maximum likelihood estimation. The algorithm suggested in this paper requires $O(N(k^*)^2)$ computation, where N is the

sample size and k^* is the width of the neighborhood. This property makes it appropriate to handle large data sets.

One-dimensional discontinuous regression curve fitting is discussed by many authors. McDonald and Owen (1986) suggested the "split linear smoother" algorithm. Hall and Titterington (1992) proposed a method to detect the jumps by establishing some relations among three local linear smoothers. Müller (1992), Qiu (1994), Qiu, Asano and Li (1991), Wu and Chu (1993a, b), among many others, discussed various kernel-type methods. Eubank and Speckman (1993) treated the one-dimensional discontinuous regression model as a semiparametric model. Loader (1996) suggested a jump detector based on local polynomial kernel estimates. Qiu and Yandell (1998) suggested detecting jumps based on local least squares fits. Wang (1995) detected jumps with wavelet transformations.

This paper is organized as follows. In next section, we discuss detection of the jump candidate points. In Section 3, the local PC line is introduced. The DRS fit is discussed in Section 4. Properties of the surface fit are discussed in Section 5. In Section 6, we present some numerical results concerning the goodness-of-fit and the selection of window widths. The final section (Section 7) contains some concluding remarks.

2. Detection of the possible jumps. Most edge detection methods can be used as the first step in our three-stage DRS fitting procedure. The proposal presented below is only one possibility. We use it here because we know some of its theoretical properties and have the required software. For an overview on edge detection techniques in image processing, please read Bhandarkar, Zhang and Potter (1994), Gonzalez and Woods (1992), Qiu and Bhandarkar (1996), Rosenfeld and Kak (1982), Torre and Poggio (1986) and the references cited there.

Suppose that the regression model concerned is

$$z_{ij} = f(x_i, y_j) + \epsilon_{ij}, \qquad i, j = 1, 2, \ldots, n, \tag{2.1}$$

where $\{(x_i, y_j) = (i/n, j/n), i, j = 1, 2, \ldots, n\}$ are equally spaced design points in $[0, 1] \times [0, 1]$, $\{\epsilon_{ij}\}$ are i.i.d. errors with mean 0 and unknown variance σ^2, $f(x, y)$ is an unknown nonparametric regression function, and $N = n^2$ is the sample size. We further assume that there exists a partition $\{\Lambda_i, i = 1, 2, \ldots, s\}$ of the design space $[0, 1] \times [0, 1]$ such that: (1) each Λ_i is a connected region in the design space; (2) $\bigcup_{i=1}^{s} \Lambda_i = [0, 1] \times [0, 1]$; (3) $f(x, y)$ is continuous in $\Lambda_i \setminus \partial\Lambda_i$, for $i = 1, 2, \ldots, s$, where $\partial\Lambda_i$ denotes the boundary point set of Λ_i and (4) there exist at most finite points $\{(x_j^*, y_j^*), j = 1, 2, \ldots, s^*\}$ in $[\bigcup_{i=1}^{s} \partial\Lambda_i] \cap [(0, 1) \times (0, 1)]$ such that for each point (x_j^*, y_j^*), $j = 1, 2, \ldots, s^*$, there are $\Lambda_{j1}^*, \Lambda_{j2}^* \in \{\Lambda_i, i = 1, 2, \ldots, s\}$ satisfying (1) $(x_j^*, y_j^*) \in \partial\Lambda_{j1}^* \cap \partial\Lambda_{j2}^*$ and (2) $\lim_{(x, y) \to (x_j^*, y_j^*), (x, y) \in \Lambda_{j1}^*} f(x, y) = \lim_{(x, y) \to (x_j^*, y_j^*), (x, y) \in \Lambda_{j2}^*} f(x, y)$. We call $D := [\bigcup_{i=1}^{s} \partial\Lambda_i] \cap [(0, 1) \times (0, 1)]$ *the jump location curves* of $f(x, y)$.

A point (x, y) is defined as *a singular point* of the JLCs if it is on a JLC and satisfies any one of the following three conditions.

1. There exists some constant $\zeta_0 > 0$ such that for any $0 < \zeta < \zeta_0$ the neighborhood of (x, y) with diameter ζ is divided into more than two connected regions by the JLCs.
2. There do not exist two orthogonal lines crossing at (x, y) such that two opposite quadrants formed by these two lines belong to two different regions separated by a JLC in a neighborhood of (x, y).
3. (x, y) is one of $\{(x_j^*, y_j^*), j = 1, 2, \ldots, s^*\}$.

The above condition (1) actually defines cross points of the JLCs (Figure 2a). Condition (2) corresponds to acute angles of the JLCs (Figure 2b). The jump magnitude is 0 at (x_j^*, y_j^*), for $j = 1, 2, \ldots, s^*$, if (x_j^*, y_j^*) is not a cross point of the JLCs. When a JLC has a unique tangent line at (x, y) (Figure 2c) or a JLC has an obtuse angle at (x, y) (Figure 2d), (x, y) is a nonsingular point as long as it is not one of $\{(x_j^*, y_j^*), j = 1, 2, \ldots, s^*\}$.

A natural way to detect a jump at (x_i, y_j) is to consider two small regions on two different sides of (x_i, y_j) along some direction. The absolute difference of the weighted averages of the observations in these two regions could be maximized with respect to the direction. This maximized value could be used as a jump detection criterion [Müller and Song (1994); Qiu (1997)]. However, this maximization procedure requires extensive computation. Instead, we suggest using the following least squares (LS) coefficients to approximate the gradient direction of the regression surface and then to detect jumps.

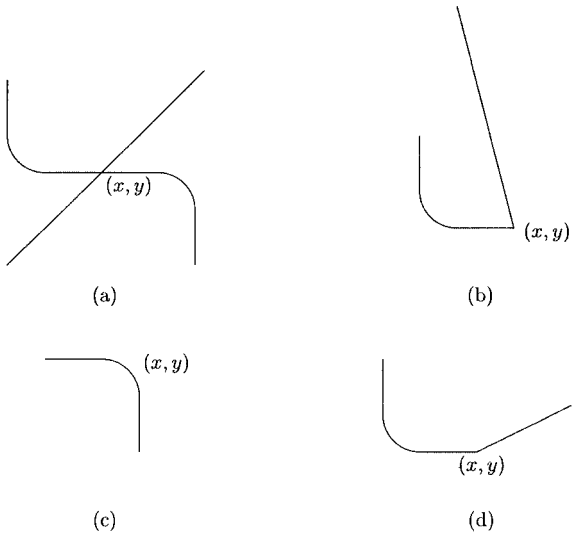

FIG. 2. (a) *Point* (x, y) *is a cross point of several JLCs*; (b) *a JLC has an acute angle at* (x, y); (c) *a JLC has a unique tangent line at* (x, y); (d) *a JLC has an obtuse angle at* (x, y). *The point* (x, y) *is a singular point in cases* (a) *and* (b). *It is a nonsingular point in cases* (c) *and* (d), *as long as it is not one of* $\{(x_j^*, y_j^*), j = 1, 2, \ldots, s^*\}$.

At any design point (x_i, y_j), $l+1 \le i, j \le n-l$, we consider its neighborhood $N(x_i, y_j) := \{(x_{i+s}, y_{j+t}),\ s,t = -l, -l+1, \ldots, 0, \ldots, l-1, l\}$ with window width $k = 2l+1 \ll n$, where l is a nonnegative integer. A least squares plane is fitted in this neighborhood,

$$\hat{z}_{ij}(x,y) = \hat{\beta}_0^{(i,j)} + \hat{\beta}_1^{(i,j)}(x - x_i) + \hat{\beta}_2^{(i,j)}(y - y_j), \qquad (x,y) \in N(x_i, y_j).$$

Qiu and Yandell (1997) proved the following results.

THEOREM 2.1. *In model* (2.1), *suppose that* $f(x,y)$ *has continuous first-order partial derivatives over* $(0,1)\times(0,1)$ *except on the JLCs at which the first-order right and left partial derivatives have limits. The window width* k *satisfies the conditions that* $\lim_{n\to\infty} k = \infty$ *and* $\lim_{n\to\infty} k/n = 0$. *Let* (x,y) *be a point in* $(0,1)\times(0,1)$. *If* (x,y) *is not on any JLC, then when* n *is large enough,*

$$\hat{\beta}_1^{(i,j)} = f_x'(x,y) + O\left(\frac{n\sqrt{\log\log k}}{k^2}\right) \quad a.s.$$

and

$$\hat{\beta}_2^{(i,j)} = f_y'(x,y) + O\left(\frac{n\sqrt{\log\log k}}{k^2}\right) \quad a.s.,$$

where $\hat{\beta}_1^{(i,j)}$ *and* $\hat{\beta}_2^{(i,j)}$ *are the LS coefficients in* $N(x_i, y_j)$ *and* (x_i, y_j) *is the design point that is closest to* (x,y). *If* (x,y) *is on a JLC and it is not a singular point, then*

$$\hat{\beta}_1^{(i,j)} = f_x'(x_i, y_j) + h_1^{(i,j)}C(x,y) + \gamma_1 C_x(x,y) + O\left(\frac{n\sqrt{\log\log k}}{k^2}\right) \quad a.s.$$

and

$$\hat{\beta}_2^{(i,j)} = f_y'(x_i, y_j) + h_2^{(i,j)}C(x,y) + \gamma_2 C_y(x,y) + O\left(\frac{n\sqrt{\log\log k}}{k^2}\right) \quad a.s.,$$

where (x_i, y_j) *is the design point that is closest to* (x,y) *among the design points on the same side of the JLC as* (x,y); $C(x,y)$ *is the jump magnitude of* $f(x,y)$ *at* (x,y); $C_x(x,y)$ *and* $C_y(x,y)$ *are the jump magnitudes of the first order* x *and* y *partial derivatives of* $f(x,y)$ *at* (x,y), *respectively*; $h_1^{(i,j)}$ *and* $h_2^{(i,j)}$ *are two constants satisfying*

$$3n(k+1)/(4k^2) \le \sqrt{\left(h_1^{(i,j)}\right)^2 + \left(h_2^{(i,j)}\right)^2} \le 3\sqrt{2}\,n(k+1)/(2k^2);$$

γ_1 *and* γ_2 *are two constants between* -1 *and* 1.

In Theorem 2.1, the term $O(n\sqrt{\log\log k}\,/k^2)$ is due to the noise. We can see that the slopes $\hat{\beta}_1^{(i,j)}$ and $\hat{\beta}_2^{(i,j)}$ carry both the continuous and the jump information. When (x,y) is on a JLC and it is not a singular point, the leading terms in slopes are $h_1^{(i,j)}C(x,y)$ and $h_2^{(i,j)}C(x,y)$ which are of order n/k and tend to infinity when n increases. Consequently, $\vec{v}_{ij} := (\hat{\beta}_1^{(i,j)}, \hat{\beta}_2^{(i,j)})$ can be regarded as the approximate direction of jump at (x,y). We should

point out that some quantities used in this paper such as k, $\hat{\beta}_1^{(i,j)}$, $\hat{\beta}_2^{(i,j)}$, $h_1^{(i,j)}$ and $h_2^{(i,j)}$ depend on n. For simplicity, we did not make this explicit in notation. Their meaning should be clear from the context.

We then consider two neighboring design points (x_{N1}, y_{N1}) and (x_{N2}, y_{N2}) along the direction of $\vec{v}_{ij}$ with neighborhoods nonoverlapping but adjacent to $N(x_i, y_j)$ on either side (cf. Figure 3). If (x, y) is a nonsingular point of a JLC and (x_i, y_j) is the design point closest to (x, y), then as a direct conclusion of Theorem 2.1, $N(x_{N1}, y_{N1})$ and $N(x_{N2}, y_{N2})$ are on two different sides of that JLC when n is large enough.

An obvious criterion to detect a jump at (x_i, y_j) is then to use the difference of the averages of the observations in $N(x_{N1}, y_{N1})$ and $N(x_{N2}, y_{N2})$. That is, we can use

$$\delta_{ij}^* := |\bar{z}_{N1} - \bar{z}_{N2}|,$$

where $\bar{z}_{N1}$ and $\bar{z}_{N2}$ are the sample means in $N(x_{N1}, y_{N1})$ and $N(x_{N2}, y_{N2})$, respectively.

EXAMPLE 2.1. Let $f(x, y) = I_{\{x > 0.5\}}$ where $I_{\{\cdot\}}$ is the usual indicator function. Then $f(x, y)$ has a unique JLC: $\{(x, y)\colon x = 0.5,\ 0 \le y \le 1\}$ with constant jump magnitude 1. Figure 4a shows a cross section of this function at $y = 0.5$ (a three-dimensional plot of $f(x, y)$ could be found in Figure 9a in Section 6). Consider a sample of size 100^2 in design space $[0, 1] \times [0, 1]$ and let k be 9. If the noise in the data is ignored, then δ_{ij}^* in the cross section of $y = 0.5$ are shown in Figure 4b by the dotted curve.

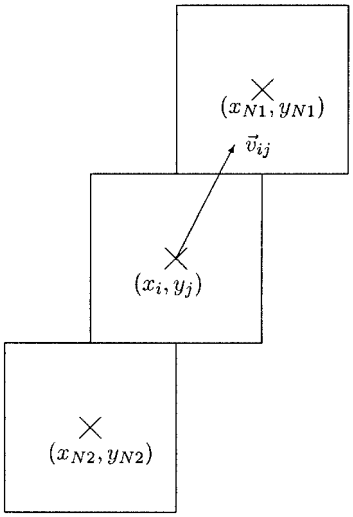

FIG. 3. *For design point (x_i, y_j), its two neighboring design points (x_{N1}, y_{N1}) and (x_{N2}, y_{N2}) are defined as the ones whose neighborhoods are adjacent to $N(x_i, y_j)$ on either side along the direction of $\vec{v}_{ij}$.*

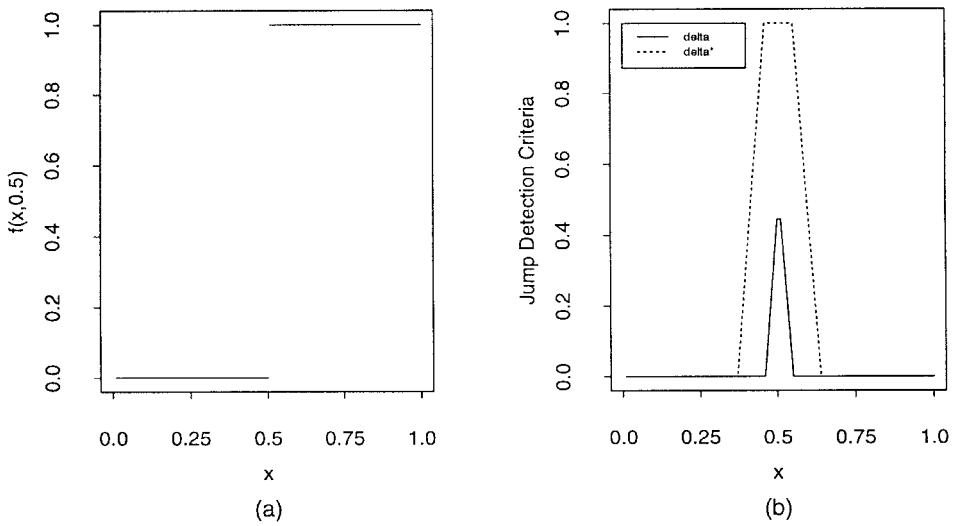

FIG. 4. (a) *A cross section of* $f(x, y) = I_{\{x > 0.5\}}$ *at* $y = 0.5$. (b) *The solid and dotted curves represent* δ_{ij} *and* δ_{ij}^*, *respectively, in the cross section of* $y = 0.5$.

By using δ_{ij}^*, design points in the band $\{(x, y)\colon 0.37 < x < 0.63,\ 0 \le y \le 1\}$ could possibly be detected in the case of Example 2.1, since the jump information makes their δ_{ij}^* values quite large. Consequently, the detected jumps are relatively thick. We suggest using the following criterion:

$$\delta_{ij} := \min\Big\{|\bar{z}_{ij} - \bar{z}_{N1}|, |\bar{z}_{ij} - \bar{z}_{N2}|\Big\}, \tag{2.2}$$

where $\bar{z}_{ij}$ is the sample mean in $N(x_i, y_j)$.

In the case of Example 2.1, δ_{ij} in the cross section of $y = 0.5$ are shown in Figure 4b by the solid curve. We can see that when (x_i, y_j) is far away from the JLC such that there are no jumps in $N(x_i, y_j)$, $N(x_{N1}, y_{N1})$ and $N(x_{N2}, y_{N2})$, then δ_{ij} is small. If (x_i, y_j) is close to the JLC, then δ_{ij} is relatively large. By using criterion δ_{ij}, the detected jumps are relatively thin. Design points outside the band $\{(x, y)\colon 0.46 < x < 0.54,\ 0 \le y \le 1\}$ are not likely to be detected. We also notice that δ_{ij} is less than δ_{ij}^* around the JLC. Therefore it could lead to a relatively conservative jump detection by using δ_{ij}.

For a constant $b > 0$, if no jumps exist in $N(x_i, y_j) \cup N(x_{N1}, y_{N1}) \cup N(x_{N2}, y_{N2})$, then

$$\begin{aligned} P\big(\delta_{ij} > b\big) &\le P\Big(|\bar{z}_{ij} - \bar{z}_{N1}| > b\Big) \\ &= E\Big\{P\Big(|\bar{z}_{ij} - \bar{z}_{N1}| > b \;\Big|\; \hat{\beta}_1^{(i,j)}, \hat{\beta}_2^{(i,j)}\Big)\Big\}. \end{aligned}$$

It is not hard to check that both $\bar{z}_{ij}$ and $\bar{z}_{N1}$ are approximately uncorrelated with $\hat{\beta}_1^{(i,j)}$ and $\hat{\beta}_2^{(i,j)}$ and $\bar{z}_{ij} - \bar{z}_{N1}$ is approximately normally distributed with mean 0 and variance $\mathrm{Var}(\bar{z}_{ij} - \bar{z}_{N1}) = 2\sigma^2/k^2$. Therefore a natural threshold value for δ_{ij} is

$$b = \frac{\sqrt{2}\,\hat{\sigma}}{k} Z_{\alpha_n/2}, \tag{2.3}$$

where $\hat{\sigma}$ is a consistent estimate of σ and $Z_{\alpha_n/2}$ is the $1 - \alpha_n/2$ quantile of the standard normal distribution. Relationship between $Z_{\alpha_n/2}$ and α_n can be described by

$$\begin{aligned}\left[\sqrt{2/\pi}\, Z_{\alpha_n/2}/\left(1 + Z_{\alpha_n/2}^2\right)\right]\exp\left(-Z_{\alpha_n/2}^2/2\right) \\ \leq \alpha_n \leq \left[\sqrt{2/\pi}\,/Z_{\alpha_n/2}\right]\exp\left(-Z_{\alpha_n/2}^2/2\right).\end{aligned}$$

Design points in $\hat{D}_n := \{(x_i, y_j)\colon \delta_{ij} > b$ are then flagged as jump candidate points.

THEOREM 2.2. *Under the conditions in Theorem* 2.1, *if*:

(i) $E|\epsilon_{11}|^p < \infty$, $p > 4$;
(ii) $\lim_{n\to\infty} Z_{\alpha_n/2}/\log n = \infty$;
(iii) $\lim_{n\to\infty}(Z_{\alpha_n/2}/n^{2/p}) = 0$;
(iv) $0 < \lim_{n\to\infty}(k/n^{2/p}) < \infty$,

then $d_H(D \cap \mathscr{D}_\varepsilon, \hat{D}_n \cap \mathscr{D}_\varepsilon) = O(n^{-1+2/p})$, *a.s.*, *where* $\mathscr{D}_\varepsilon := [\varepsilon, 1-\varepsilon] \times [\varepsilon, 1-\varepsilon] \backslash \{(x, y)\colon \|(x, y) - (x_*, y_*)\| < \varepsilon$, (x_*, y_*) *is some singular point of the JLCs*$\}$, ε *is any constant in* $(0, 0.5)$, $\|\cdot\|$ *is the Euclidean norm and* $d_H(\mathscr{A}, \mathscr{B})$ *denotes the Hausdorff distance between two point sets* $\mathscr{A}$ *and* $\mathscr{B}$ *which is defined by* $d_H(\mathscr{A}, \mathscr{B}) := \max(\sup_{\xi\in\mathscr{A}} \inf_{\eta\in\mathscr{B}} \|\xi - \eta\|, \sup_{\xi\in\mathscr{B}} \inf_{\eta\in\mathscr{A}} \|\xi - \eta\|)$.

The proof of Theorem 2.2 is given in Appendix A.

The domain $\mathscr{D}_\varepsilon$ defined in Theorem 2.2 excludes the border area of the design space and the neighborhoods of the singular points. In $\mathscr{D}_\varepsilon$, $\hat{D}_n$ converges almost surely to D in Hausdorff distance by Theorem 2.2. Most local smoothing methods share the "boundary problem." When the regression function is continuous, several proposals have been suggested in the literature to handle this problem [e.g., Gasser and Müller (1979); Rice (1984)]. When the regression function is discontinuous, further research is needed to detect jumps in the border area. In our simulation study presented in Section 6, we use a "padding" method to overcome this difficulty. In the case of jump detection around singular points, future research is also needed.

In Qiu and Bhandarkar (1996) and Qiu and Yandell (1997), there is some related discussion about edge detection based on LS coefficients. Two postprocessing procedures suggested by Qiu and Yandell (1997) can be applied to the edge detection procedure (2.2) and (2.3) to delete two kinds of deceptive jump candidates. In Qiu and Yandell (1997), we compared the performance of our

edge detectors with two popular edge detectors in image processing literature: Sobel [Rosenfeld and Kak (1982), Section 10.2], and Laplacian of Gaussian [Marr and Hildreth (1980)].

3. First-order approximation to the JLCs. At the completion of the first stage, the possible jump positions are known. In regions without jump candidate points, the regression surface could be fitted as usual (by using kernel method, spline method or others). In regions with the jump candidate points, however, special treatment is necessary.

First of all, we introduce some notation. Consider a design point (x_i, y_j) and its neighborhood $N^*(x_i, y_j) := \{(x_{i+s}, y_{j+t}): s, t = -l^*, \ldots, 0, \ldots, l^*\}$ with window width $k^* = 2l^* + 1$. The values of k^* and l^* could be different from k and l introduced in the previous section. The jump candidate points in that neighborhood are denoted by $\{(w_r, v_r),\ r = 1, 2, \ldots, m\}$. We use (W, V) to denote a vector variable taking values over $\{(w_r, v_r),\ r = 1, 2, \ldots, m\}$; $\overline{w}, \overline{v}, \sigma_{ww}$ and σ_{vv} to denote the sample means and variances of w's and v's; σ_{wv} to denote their covariance.

Most jump candidate points are in the bands centered at the JLCs with width $\sqrt{2}\,k/n$ when n is large enough, since these points have high probability for large jump detection criterion. The closer to the JLCs, the denser the jump candidate points tend to be because more jump information is involved. We then search for a line $a(W - \overline{w}) + b(V - \overline{v}) = 0$ with $a^2 + b^2 = 1$ such that

$$\text{(3.1)}\quad \operatorname{Var}\big(a(W - \overline{w}) + b(V - \overline{v})\big) = \min_{\tilde{a}^2 + \tilde{b}^2 = 1} \operatorname{Var}\big(\tilde{a}(W - \overline{w}) + \tilde{b}(V - \overline{v})\big),$$

where $\operatorname{Var}(\cdot)$ denotes the sample variance. An extreme case is that $\operatorname{Var}(a(W - \overline{w}) + b(V - \overline{v})) = 0$. In that case, all jump candidate points are on the line $a(W - \overline{w}) + b(V - \overline{v}) = 0$.

The vector (a, b) determined by (3.1) is actually the second principal component (PC), the normalized eigenvector of the smaller eigenvalue of the covariance matrix of (W, V). The line $a(W - \overline{w}) + b(V - \overline{v}) = 0$ is the usual PC line which indicates the direction that the jump candidate points have the biggest dispersion. After some calculations, that line turns out to be

$$\text{(3.2)}\qquad \sigma_{wv}(W - \overline{w}) + (\lambda_1 - \sigma_{ww})(V - \overline{v}) = 0,$$

where

$$\lambda_1 := \left(\sigma_{ww} + \sigma_{vv} - \sqrt{(\sigma_{ww} - \sigma_{vv})^2 + 4\sigma_{wv}^2}\right)\Big/2$$

and λ_1 is the smaller eigenvalue of the covariance matrix of (W, V).

It is not hard to check that $\lambda_1 - \sigma_{ww} \le 0$. Hence the PC line has an increasing trend when w's and v's are positively correlated ($\sigma_{wv} > 0$) and vice versa. This is intuitively reasonable.

THEOREM 3.1. *Besides the conditions stated in Theorems* 2.1 *and* 2.2, *we suppose that* $\lim_{n\to\infty} k^*/n = 0$ *and* $\lim_{n\to\infty} k/k^* = 0$. *If* (x_i, y_j) *is the closest design point to a given nonsingular point* P *on a JLC at which the JLC has a*

unique tangent line, then the principal component line in $N^*(x_i, y_j)$ *converges to the tangent line at P both pointwise and in direction with probability* 1. *The convergence rate could be* $O(n^{-1+2/p})$.

The above result establishes the first order approximation to the JLCs of the local PC lines. Second or higher order approximation might also be possible. However, this kind of generalization is not trivial. It might relate to the curve–surface description in computer-aided design [Farin (1993)]. We leave this issue to future research. The proof of Theorem 3.1 is given in Appendix B.

4. Fit the jump regression surfaces. Observations on the same side of the PC line as a selected design point are then combined as a weighted average to fit the regression surface. The subscripts of these design points are determined by the following set:

$$S(x_i, y_j) := \begin{cases} \{(s,t): \sigma_{wv}(x_{i+s} - \overline{w}) + (\lambda_1 - \sigma_{ww})(y_{j+t} - \overline{v}) \geq 0, \\ \qquad\qquad s,t = -l^*, \ldots, 0, \ldots, l^*\}, \\ \quad \text{if } \sigma_{wv}(x_i - \overline{w}) + (\lambda_1 - \sigma_{ww})(y_j - \overline{v}) \geq 0; \\ \{(s,t): \sigma_{wv}(x_{i+s} - \overline{w}) + (\lambda_1 - \sigma_{ww})(y_{j+t} - \overline{v}) < 0, \\ \qquad\qquad s,t = -l^*, \ldots, 0, \ldots, l^*\}, \\ \quad \text{otherwise.} \end{cases} \tag{4.1}$$

Different smoothing methods could be used here to average the observations. For simplicity of presentation, we apply the kernel smoothing method. Let $K(x, y)$ be a bivariate kernel function defined on $[-L/2, L/2] \times [-L/2, L/2]$ with $L > 0$ a constant. Then the surface fit at (x_i, y_j) is defined by

$$\hat{f}(x_i, y_j) := \frac{\sum_{s,t} K(Ls/k^*, Lt/k^*) Z_{i+s, j+t} I_{\{(s,t) \in S(x_i, y_j)\}}}{\sum_{s,t} K(Ls/k^*, Lt/k^*) I_{\{(s,t\} \in S(x_i, y_j)\}}}. \tag{4.2}$$

When there are no jump candidate points in $N^*(x_i, y_j)$, then σ_{wv} and $\lambda_1 - \sigma_{ww}$ are both zero. By the above procedure, all observations are actually used in the surface fitting. If the number of jump candidate points is very small such that no JLS is possible in the neighborhood (e.g., in the case that several jump candidate points exist in a corner of the neighborhood), then few observations are dropped from the surface fit. Although our procedure works well in these cases, we still suggest making a judgment to insure that a JLC is possible in the neighborhood (e.g., by the fact that the number of jump candidate points is more than, say, $k^*/2$) before finding the PC line. The computation involved in each of such judgments is much less than the amount involved in finding the PC line. This could save considerable computation time since most design points are not jump candidate points. We also want to mention that the first two stages (namely, detecting the jump

candidate points and defining the local PC lines) of the three-stage procedure could be updated from one design point to the next because only a few points change. That will further reduce computation time.

5. Consistency of the fitted surface.

THEOREM 5.1. *Besides the conditions in Theorems* 2.1, 2.2 *and* 3.1, *suppose that* $f(x, y)$ *has continuous second-order partial derivatives in the continuous regions and the kernel function* $K(x, y)$ *satisfies the conditions that*:

(i) $K(x, y) > 0$ *for* $(x, y) \in [-L/2, L/2] \times [-L/2, L/2]$;
(ii) $\int_{-L/2}^{L/2} \int_{-L/2}^{L/2} K(x, y)\, dx\, dy = 1$;
(iii) $\int_{-L/2}^{L/2} K(x, y) x\, dx\, dy = \int_{-L/2}^{L/2} K(x, y) y\, dx\, dy = 0$.

Then the surface fit (4.2) *is uniformly* L^2 *consistent in*

$$\mathscr{D}_\varepsilon^* := \mathscr{D}_\varepsilon \setminus \{(x, y): \|(x, y) - (x_*, y_*)\| < \varepsilon, \; (x_*, y_*) \text{ is some point on the JLCs}\},$$

for any constant $0 < \varepsilon < 0.5$. *The convergence rate is* $O(n^{-4/3})$ *if* $k^* = O(n^{2/3})$. *The surface fit is* L^2 *consistent with rate* $O(n^{-1})$ *in* $\mathscr{D}_\varepsilon \setminus \mathscr{D}_\varepsilon^*$ *if* $k^* = O(n^{1/2})$.

Proof of Theorem 5.1 is given in Appendix C.

The above result basically says that the surface fit (4.2) behaves as well as the conventional kernel estimator in the continuous regions [cf. Stone (1982) for the optimal convergence rates for nonparametric regression]. In the neighborhood of the JLCs, however, it sacrifices some convergence rate because we only use part of the observations in the surface fit. This result is compatible with that of Theorem 5.1 in Müller and Song (1994).

6. Simulation study. We present our simulation study in three parts. In Section 6.1, accuracy of the fitted surface of the three-stage surface-fitting algorithm is evaluated by continuing our discussion about the example of Figure 1. Its performance around a singular point of the JLCs or when the jump magnitude is small is investigated in Section 6.2. The algorithm is applied to a global topographical elevation dataset in Section 6.3.

6.1. *Numerical performance of the fitted surface.* We study the numerical performance of the three-stage algorithm by revisiting the example of Figure 1 first. The data in Figure 1b is from model (2.1) for equally spaced design points $\{(x_i, y_j) = (i/n, j/n),\ i, j = 1, 2, \ldots, n\}$, with errors from $N(0, \sigma^2)$, $\sigma = 0.5$ and $n = 100$. The regression function $f(x, y) = -2(x - 0.5)^2 - 2(y - 0.5)^2 + I_{\{(x-0.5)^2+(y-0.5)^2 \le 0.25^2\}}$. It has jumps at the jump location curve $(x - 0.5)^2 + (y - 0.5)^2 = 0.25^2$ with constant jump magnitude 1.

The product of two Epanechnikov kernels is used in our simulation as the bivariate kernel function needed in the local average procedure of the algorithm. That is, $K(x, y) = \frac{9}{16}(1 - x^2)(1 - y^2) I_{\{(x, y) \in [-1, 1] \times [-1, 1]\}}$. The significance level α_n in (2.3) is set to be 0.001. Mean squared error (MSE) is used to

evaluate the surface fit. Figure 5 shows the MSE values (all MSE values used in this section are averages of 100 replications) for several k and k^*. As we expected, for a fixed k value, MSE decreases and then increases as k^* increases. The window width k^* works as a tuning parameter to balance "underfit" and "overfit." The minimal MSE with respect to k^* depends on k. Thus we do need to balance the values of k and k^*. If k is very small, few jumps are detected. In that case, the accuracy of our surface fit is almost the same as that of the conventional kernel surface fit. If k is too large, many false jumps are detected. That could deteriorate the surface fitting since many observations are not used in the fitted surface. The overall minimal MSE (with respect to both k and k^*) is about 0.008 which corresponds to $k = 9$ and $k^* = 15$. Figure 1d shows one realization of the surface fit in this case. Figure 1c shows the conventional kernel surface fit with window width 15.

Figure 5 also shows that when k^* is too small or too large, MSE is not sensitive to the value of k. This implies that the surface fit would not change much with or without using our discontinuous surface fit procedure. This is also shown in Figure 6. The solid curve in Figure 6 represents MSE values of the surface fits from our procedure with $k = 9$. The dotted curve denotes the MSE values of the conventional kernel surface fits. The optimal MSE of the

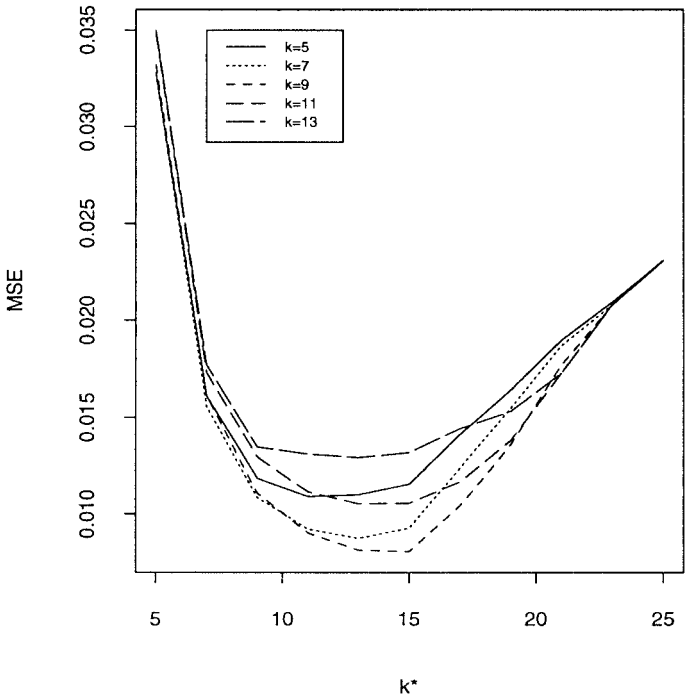

FIG. 5. *MSE values with several k and k*.*

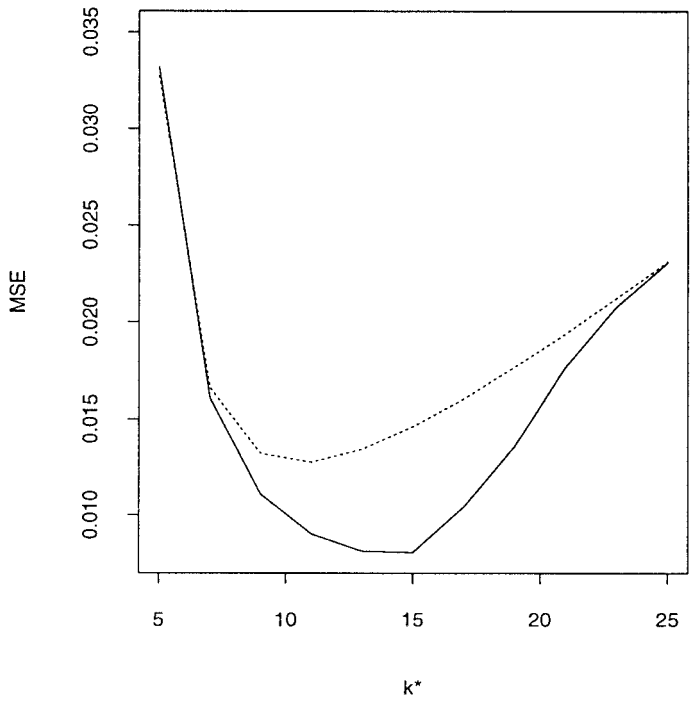

FIG. 6. *The solid line represents the MSE values of the surface fits of the discontinuity-preserving algorithm with* $k = 9$. *The dotted line denotes the MSE values of the conventional kernel surface fits.*

conventional kernel fit is 0.0128. The optimal MSE of our procedure explains 62.5% of that value. Since the MSE values are calculated in the entire design space, our procedure does improve the accuracy of the fitted surface much.

In the algorithm, we use the so-called "padding" method to define neighborhoods of the design points in border area. That is, we expand the design space in such a way that an "observation" in the expanded area takes the same value as the observation in the original design space which is closest to it. The expanded design space should be large enough to define neighborhoods of all design points in the border area of the original design space.

The 2.5 and 97.5 percentiles of 100 replications of the surface fit in the cross section of $y = 0.5$ are plotted in Figure 7a with the lower and upper dashed curves, respectively. These two percentiles are defined by the third smallest and the third largest fits of the 100 replications. The solid curve is the cross section of the regression surface. The dotted curve is the average of the 100 fits. We can see that the averaged fit and the percentiles all have sharp breaks nearby the true jump locations. But these breaks are about 0.01 (less than $1/10$ of k^*/n) away from the true jumps. This amount of difference could be explained by the curvature of the jump location curve. If we

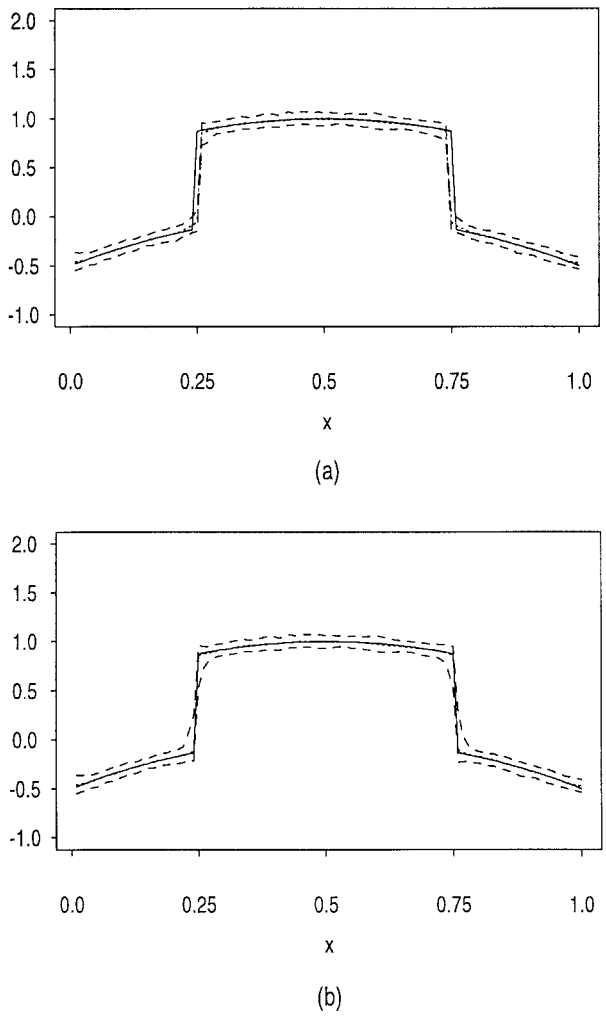

FIG. 7. *The lower and upper dashed curves denote the* 2.5 *and* 97.5 *percentiles of* 100 *replications of the surface fit in the cross section of* $y = 0.5$. *The solid curve is the cross section of the true regression surface. The dotted curve is the averaged fit*. (a) *The jump location curve is a circle* $(x - 0.5)^2 + (y - 0.5)^2 = 0.25^2$; (b) *the jump location curves are two straight lines* $x = 0.25$ *and* $x = 0.75$.

replace the circle by two straight lines $x = 0.25$ and $x = 0.75$ (the regression function becomes $f(x, y) = -2(x - 0.5)^2 - 2(y - 0.5)^2 + I_{\{0.25 \le x \le 0.75\}}$), then the corresponding percentiles are shown in Figure 7b. We can see that the difference disappears. This phenomenon may imply that the first order approximation to the JLCs can be improved by the second or higher order approximations.

We then perform the simulations with several σ and n values. The optimal MSE value and the corresponding pair (k, k^*) in each combination of σ and n are presented in Table 1. From this table, we can see that (1) k^* is bigger than k; (2) both k and k^* increase with increases in sample size and increases in σ; (3) MSE decreases with increases in sample size and reductions in σ. The first finding suggests the condition $\lim_{n\to\infty} k/k^* = 0$ used in Theorem 3.1 is reasonable in applications. The third finding might reflect the consistency of the fitted surface.

6.2. *Performance around a singular point of the JLCs.* The regression function used in the above example has three preferable properties. First, there does not exist any singular point on the JLC. Second, the limit of the gradient direction of the surface at the JLC is consistent with the jump direction. Third, the JLC is far away from the border region of the design space. These properties should be helpful for the surface fitting. Next we consider regression function $f(x, y) = x + I_{\{y > 0.375 + 0.25x \text{ or } x > 0.5\}}$ which does not have any of these three properties.

Figure 8a shows the original surface. It has jumps at JLC: $\{(x, y)$: $y = 0.375 + 0.25x$ and $0 \le x \le 0.5\} \cup \{(x, y)$: $x = 0.5$ and $0 \le y \le 0.5\}$ with constant jump magnitude 1. Point $(0.5, 0.5)$ is a singular point of the JLC. The gradient direction of the surface in the continuous regions is $(1, 0)$, which is different from the jump direction at the left part of the JLC. As before, observations are generated from model (2.1) for equally spaced design points in design space $[0, 1] \times [0, 1]$, with errors from $N(0, \sigma^2)$, $\sigma = 0.5$ and $n = 100$. The data points are presented in Figure 8b. Figure 8c shows the gradient directions $\{\vec{v}_{ij}\}$ of the fitted local LS planes. It is apparent that $\{\vec{v}_{ij}\}$ still reveal the JLC well even when the gradient direction of the surface disagrees with the jump direction. The jump detection criterion δ_{ij} is shown in Figure

TABLE 1
The minimal MSE values (in parentheses) and the corresponding pairs of (k, k^*) for several combinations of n and σ

	n			
σ	**71**	**100**	**141**	**200**
0.25	5, 9 (0.0060)	7, 11 (0.0043)	7, 11 (0.0029)	7, 13 (0.0021)
0.5	7, 11 (0.0112)	9, 15 (0.0081)	11, 17 (0.0055)	13, 19 (0.0042)
0.75	9, 13 (0.0181)	9, 15 (0.0128)	13, 21 (0.0079)	17, 25 (0.0063)
1	9, 13 (0.0242)	13, 17 (0.0182)	13, 21 (0.0111)	19, 29 (0.0083)

8d by an image plot. The darker the color, the larger the value. This figure shows that it is not unreasonable to use the "padding" method to handle the "boundary problem" in this case. The averaged surface fit of our three-stage procedure based on 100 replications is presented in Figure 8e. The procedure works well except at the places around the singular point. The 2.5 and 97.5 percentiles of the 100 replications of the surface fit in the cross section of $x = 0.25$ are plotted in Figure 8f by the lower and upper dashed curves, respectively.

The jump detection algorithm discussed in Section 2 often fails to detect jumps with magnitudes less than the threshold value b defined in (2.3). Thus it is of some interest to evaluate the performance of the three-stage surface fitting procedure when the jump magnitude is small. Consider the regression function $f(x, y) = CI_{\{x > 0.5\}}$ which has a unique JLC: $\{(x, y)\colon x = 0.5, 0 \le y \le 1\}$ with constant jump magnitude C. Figure 9a shows the regression function when $C = 1$. Let C change from 0 to 2 with step 0.1. For each C value, observations are generated in the same way as in the previous examples with $\sigma = 0.5$ and $n = 100$. The MSE values of the three-stage procedure and the conventional kernel smoothing method are presented in Figure 9b by the solid and the dotted curves, respectively. From the plot, it is clear that the two methods do not differ very much when C is small. In this case the jump detection procedure only identifies a few jumps, making the fitted surface close to the conventional kernel estimate [cf. (4.2) and the discussion in the last paragraph of Section 4]. We plot the detected jumps when $C = 0.6$ in Figure 9c. More jumps are detected when C gets larger. This can be seen from Figure 9d in which the detected jumps when $C = 1$ are presented. In such a case our procedure outperforms the conventional method using the MSE criterion.

6.3. *Application to global topographical elevation data*. Figure 10a shows the global topographical elevation data contaminated by i.i.d. noise with $N(0, 1500^2)$ distribution. The darker the color, the higher the elevation. The resolution of the data is 1 degree $\times$ 1 degree in longitude and latitude. Hence the sample size is $181 \times 360 = 65{,}160$. The elevation measurement is relative to the sea surface level, ranging from $-8{,}964.0$ meters to 6,096.0 meters in this data set. As mentioned in Section 1, it is not easy to apply the Bayesian methods to this data set since the response is a continuous variable. Because the number of JLCs cannot be counted accurately, some non-Bayesian methods such as those suggested by Korostelev and Tsybakov (1993), Müller and Song (1994), O'Sullivan and Qian (1994), and Qiu (1997) are not appropriate for this data set either.

In image processing literature, local smoothing operators are often used for two purposes. One is to remove the noise in an image [Bow (1992), Section 9.3]. The second is to change the image from one resolution to another. Commonly used smoothing operators include the local averaging smoothers and the local median smoother [Gonzalez and Woods (1992), Section 4.3]. The proposal suggested in this paper provides an alternative way to do local

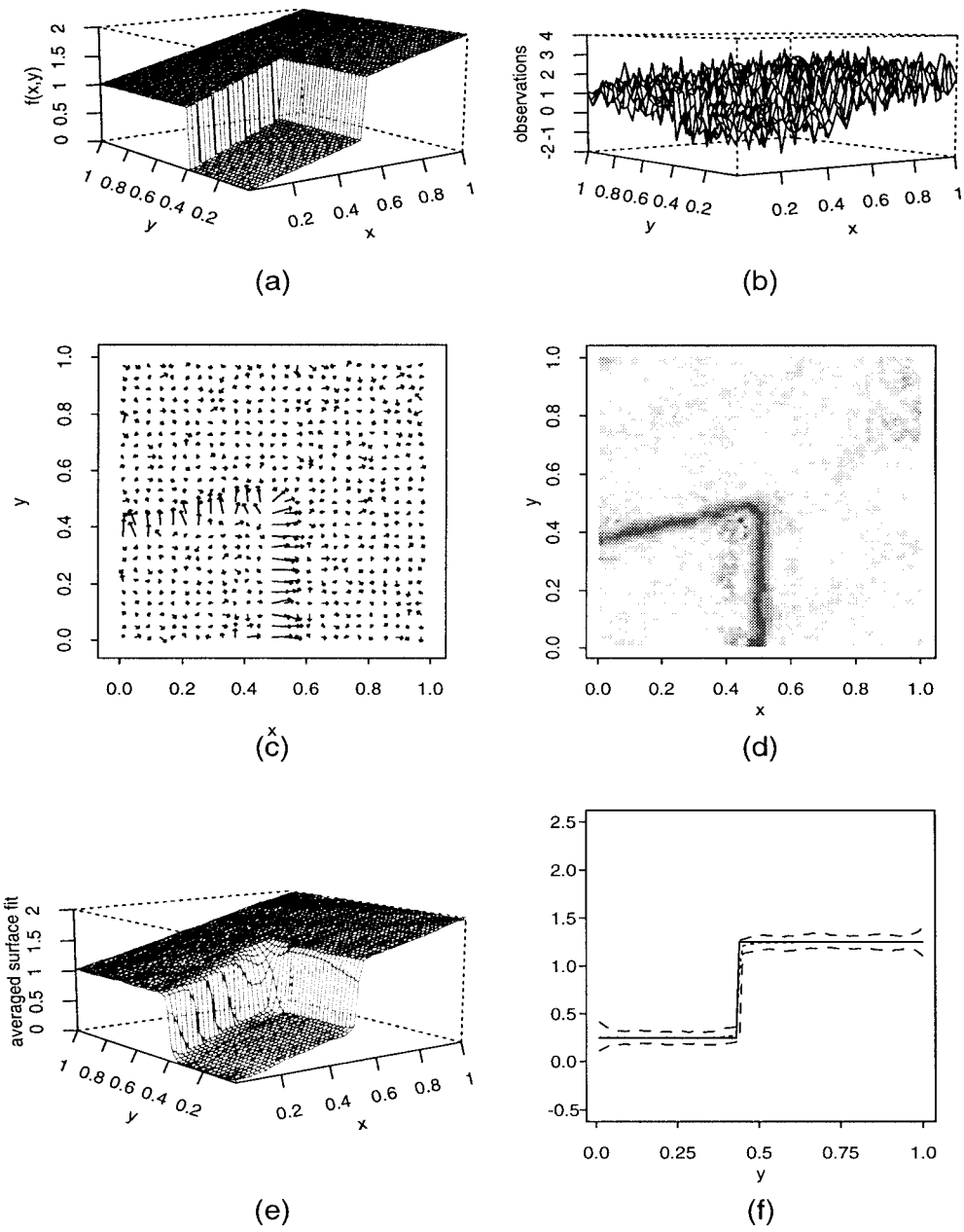

FIG. 8. (a) *The original regression surface*; (b) *the observations*; (c) *the gradient directions* $\{\vec{v}_{ij}\}$ *of the fitted local LS planes*; (d) *the jump detection criterion* $\{\delta_{ij}\}$; (e) *the averaged surface fit based on* 100 *replications*; (f) *the* 2.5 *and* 97.5 *percentiles of the* 100 *replications of the surface fit in the cross section of* $x = 0.25$. *In plot* (f), *the dashed curves represent the two percentiles. The solid curve is the cross section of the true regression surface. The dotted curve is the averaged fit.*

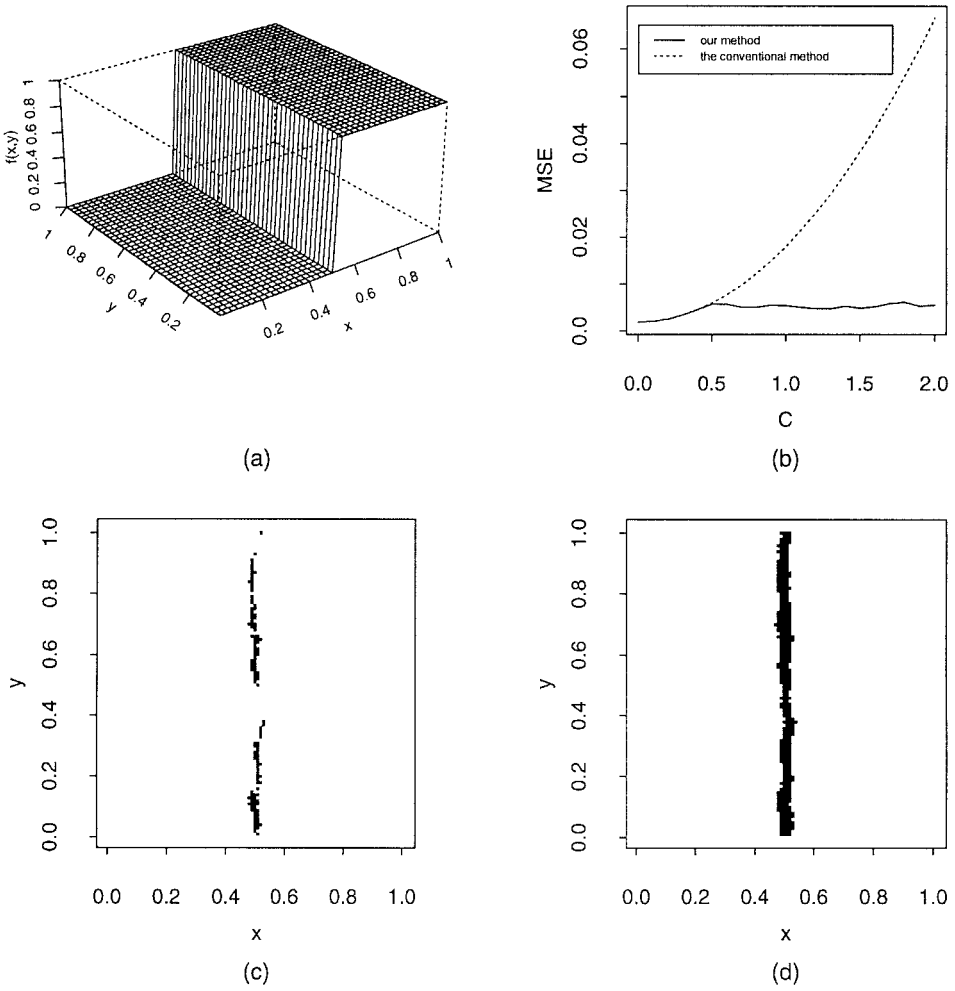

FIG. 9. (a) *The regression surface with jump magnitude* $C = 1$. (b) *The MSE values of the three-stage surface fitting algorithm and the conventional kernel smoothing method when* C *changes from* 0 *to* 2. (c) *The detected jumps when* $C = 0.6$. (d) *The detected jumps when* $C = 1$.

smoothing but with the jumps preserved, which might be preferable in such applications.

Theoretically speaking, the topographical elevation changes gradually and abrupt jumps can hardly be found. However given the low resolution of the data, the jumps are quite obvious, particularly between land and sea. With the jump detection algorithm introduced in Section 2, the detected jump locations are presented in Figure 10e. Its modified version by the modification procedures in Qiu and Yandell (1997) is shown in Figure 10f. In the algo-

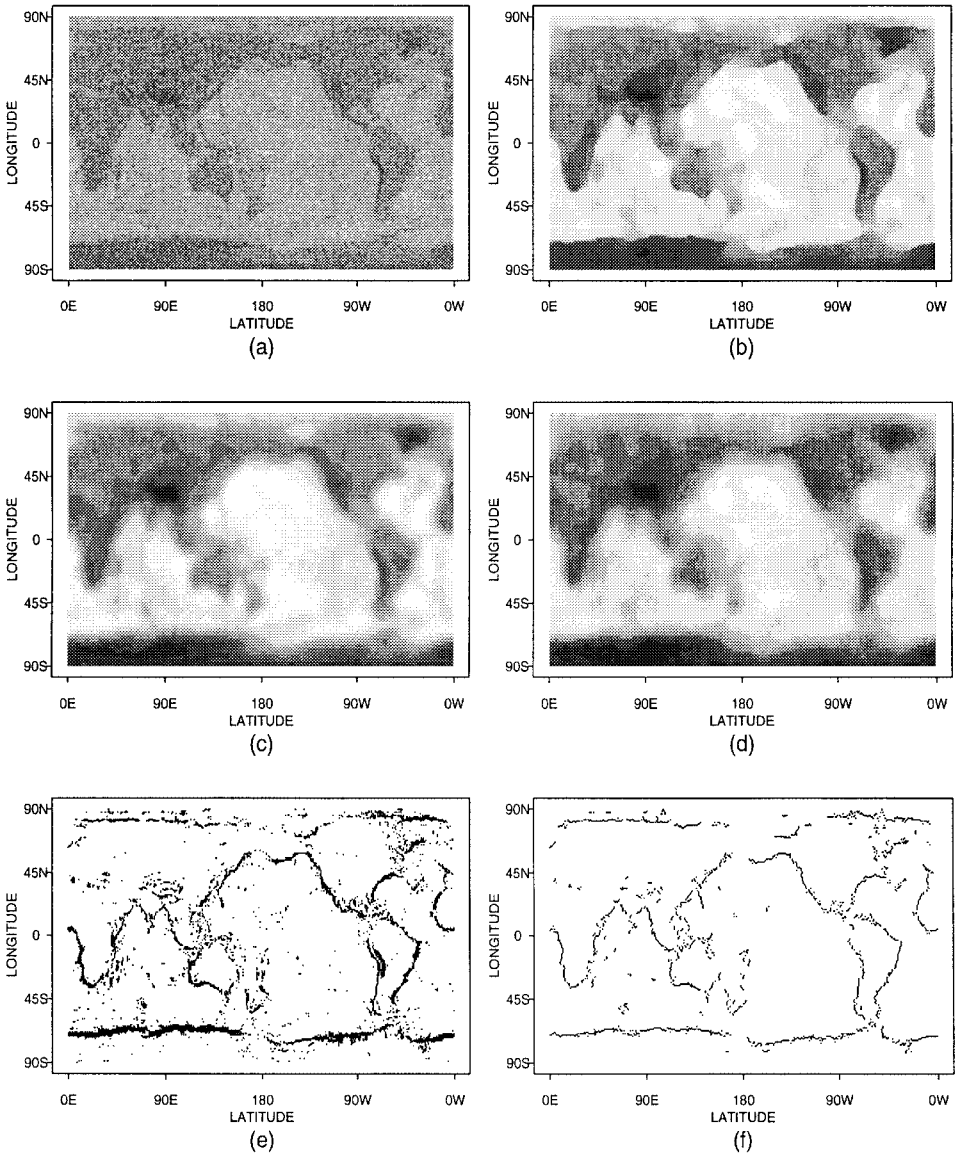

FIG. 10. (a) *A global topographical elevation data contaminated by i.i.d. noise with* $N(0, 1500^2)$ *distribution.* (b) *The fitted surface by the three-stage jump-preserving algorithm.* (c) *The fitted surface by the conventional kernel smoothing method.* (d) *The fitted surface by the local median smoothing method.* (e) *The detected jump locations by the jump detection procedure discussed in Section* 2. (f) *The modified version of the image in plot* (e) *by the modification procedures in Qiu and Yandell* (1997).

rithm, α_n is set to be 0.001 as in the previous examples. The window width k is chosen to be 5 for good visual impression.

We then apply the jump-preserving surface fitting procedure to this data set. The fitted surface with $k^* = 15$ is presented in Figure 10b. As a comparison, we plot the fitted surfaces by the conventional kernel smoothing method and the local median smoothing method in Figures 10c and d, respectively. The same window sizes are used in these three methods. We can see that our procedure performs quite well in preserving the jumps. But some angles of the JLCs seem to be smoothed away. This phenomenon might be explained by the following two reasons. One is that the angles are close to some singular points. Another is the nature of the *first*-order approximation to the JLCs of the local PC lines. The MSE values of these three fitted surfaces are 5.979×10^5, 6.195×10^5 and 8.684×10^5, respectively.

For this data set, there is no "boundary problem" since the data are from a globe. Longitude 0 degree west is the same location as that of longitude 0 degree east although they are at two ends in the plot. In the Arctic Circle and the Antarctic Circle, neighborhood relationship can also be carefully defined.

7. Concluding remarks. We have presented a three-stage procedure to fit discontinuous regression surface in the presence of noisy data. The main features of the method are its simple computation requirements and the weak assumptions imposed on the model. Simulation studies show that it works well in applications. However, some quantities used in the procedure, such as the two window widths, are still not well defined in finite sample cases. Additional simulation studies are needed to provide more guidelines on selection of these parameters for practical use. Alternative ways exist to detect jump points in the first stage of our procedure and how to average the observations in the third stage. Future research is needed to evaluate these alternatives. The principal component lines provide first-order approximations to the jump location curves. As mentioned earlier, second or higher order approximations might improve the procedure. All these issues should be future research topics.

APPENDIX

A. Proof of Theorem 2.2.

LEMMA A.1. *Let ν be a positive number and $\{\beta_n\}$ be a sequence of positive constants diverging to ∞ as $n \to \infty$. In model* (2.1), *if* (i) $E|\epsilon_{11}|^p < \infty$, $p > 2$; (ii) $n^{4\nu}/(k^2\beta_n) = O(1)$ *and* (iii) $n^{2\nu+2/p}/(k^2\beta_n \log n) = o(1)$, *then*

$$A_{n,\varepsilon} := \max_{(i/n, j/n) \in I_\varepsilon} \left| \frac{1}{k^2} \sum_{s=-l}^{l} \sum_{t=-l}^{l} \epsilon_{i+s, j+t} \right| = o\left(n^{-2\nu}\beta_n \log n\right) \quad a.s.,$$

where $0 < \varepsilon < 0.5$ is a constant, $I_\varepsilon = \{(i/n, j/n)\colon \varepsilon \le i/n, j/n \le 1 - \varepsilon\}$, $l = (k-1)/2$, and k is a positive integer satisfying $\lim_{n\to\infty} k = \infty$.

Lemma A.1 is a special case of Theorem 2.1 in Qiu (1997). Its one-dimensional version was given by Cheng and Lin (1981).

We now start to prove Theorem 2.2. Let (x_i, y_j) be a design point in $\mathscr{D}_\varepsilon$. If (x_i, y_j) is more than $3\sqrt{2}\,k/(2n)$ away from any JLC, then $N(x_i, y_j)$, $N(x_{N1}, y_{N1})$ and $N(x_{N2}, y_{N2})$ are all located in a same continuous region, say Λ_1. From (2.2),

$$\delta_{ij} \le |\bar{z}_{ij} - \bar{z}_{N1}| \le |\bar{f}_{ij} - \bar{f}_{N1}| + 2A_{n,\varepsilon},$$

where

$$\bar{f}_{ij} = (1/k^2) \sum_{s,t=-l}^{l} f(x_i + s/n, y_j + t/n)$$

and

$$\bar{f}_{N1} = (1/k^2) \sum_{s,t=-l}^{l} f(x_{N1} + s/n, y_{N1} + t/n).$$

By Lemma A.1, if $0 < \lim_{n\to\infty}(k/n^{2/p}) < \infty$, then there exists a sample subspace Ω such that $P(\Omega) = 1$ and for each $\omega \in \Omega$,

$$A_{n,\varepsilon} = O(n^{-2/p} \log n). \tag{A.1}$$

On the other hand, $\vec{v}_{ij}$ and consequently (x_{N1}, y_{N1}) are uniquely determined for each $\omega \in \Omega$. By the condition that $f(x, y)$ has continuous first order partial derivatives in Λ_1,

$$\begin{aligned} |\bar{f}_{ij} - \bar{f}_{N1}| &\le \sqrt{\left[f'_x(\eta_x, \eta_y)\right]^2 + \left[f'_y(\eta_x, \eta_y)\right]^2} \\ &\quad \times \sqrt{\left(\eta_x^{(1)} - \eta_x^{(2)}\right)^2 + \left(\eta_y^{(1)} - \eta_y^{(2)}\right)^2} \\ &\le B\frac{2\sqrt{2}\,k}{n}, \end{aligned} \tag{A.2}$$

where $(\eta_x^{(1)}, \eta_y^{(1)})$ and $(\eta_x^{(2)}, \eta_y^{(2)})$ are two points in $[x_i - k/(2n), x_i + k/(2n)] \times [y_j - k/(2n), y_j + k/(2n)]$ and $[x_{N1} - k/(2n), x_{N1} + k/(2n)] \times [y_{N1} - k/(2n), y_{N1} + k/(2n)]$, respectively; (η_x, η_y) is a point between them; and

$$B := max_{(x,y)\in \cup_{i=1}^{s}(\Lambda_i \setminus \partial\Lambda_i)} \sqrt{\left[f'_x(x, y)\right]^2 + \left[f'_y(x, y)\right]^2} < \infty.$$

By (A.1), (A.2) and the condition that $\lim_{n\to\infty} Z_{\alpha_n/2}/\log n = \infty$, when n is large enough

$$\delta_{ij} < b = O\left(\frac{Z_{\alpha_n/2}}{k}\right) \quad \text{a.s.},$$

and this is uniformly true for all (x_i, y_j) in $\mathscr{D}_\varepsilon$ and more than $3\sqrt{2}\,k/(2n)$ away from any JLC. Therefore

$$\sup_{x\in\hat{D}_n\cap\mathscr{D}_\varepsilon} \inf_{y\in D\cap\mathscr{D}_\varepsilon} \|x - y\| = O(k/n) \quad \text{a.s.} \tag{A.3}$$

Next, we assume that (x, y) is a nonsingular point on a JLC and (x_i, y_j) is the design point closest to (x, y) among the design points on the same side of the JLC as (x, y). Without loss of generality, we further assume that $(x, y) \in$

$\partial\Lambda_1 \cap \partial\Lambda_2$ with a positive jump magnitude $C(x, y)$ from Λ_1 to Λ_2. Then

$$\begin{aligned}\bar{f}_{ij} &= \frac{1}{k^2}\sum_{(s,t)\in I_1} f(x_{i+s}, y_{j+t}) + \frac{1}{k^2}\sum_{(s,t)\in I_2} f(x_{i+s}, y_{j+t}) \\ &= \frac{1}{k^2}\sum_{(s,t)\in I_1} f_-(x, y) + \frac{1}{k^2}\sum_{(s,t)\in I_2} f_+(x, y) + O(k/n) \\ &= f_-(x, y) + \frac{\#I_2}{k^2}C(x, y) + O(k/n),\end{aligned}$$

where $I_a := \{(s, t): (x_{i+s}, y_{j+t}) \in N(x_i, y_j) \cap \Lambda_a\}$, $\#I_a$ denotes the number of design points in I_a, $a = 1, 2$, $f_-(x, y) := \lim_{(x', y')\to(x, y),(x', y')\in\Lambda_1} f(x', y')$, and $f_+(x, y) := \lim_{(x', y')\to(x, y),(x', y')\in\Lambda_2} f(x', y')$. In the above expression, we have used the property of $f(x, y)$ that $|f(x_{i+s}, y_{j+t}) - f_-(x, y)| \le B(\sqrt{2}\,k/n)$ when $(s, t) \in I_1$; and $|f(x_{i+s}, y_{j+t}) - f_+(x, y)| \le B(\sqrt{2}\,k/n)$ when $(s, t) \in I_2$. Similarly,

$$\bar{f}_{N1} = f_-(x, y) + O(k/n) \quad \text{a.s.} \quad \text{and} \quad \bar{f}_{N2} = f_+(x, y) + O(k/n) \quad \text{a.s.}$$

So

$$|\bar{z}_{ij} - \bar{z}_{N1}| \le \frac{\#I_2}{k^2}C(x, y) + O(k/n) + 2A_{n,\varepsilon} \quad \text{a.s.}$$

and

$$|\bar{z}_{ij} - \bar{z}_{N2}| \le \frac{\#I_1}{k^2}C(x, y) + O(k/n) + 2A_{n,\varepsilon} \quad \text{a.s.}$$

Consequently,

$$\delta_{i,j} \le \frac{\min\{\#I_1, \#I_2\}}{k^2}C(x, y) + O(k/n) + 2A_{n,\varepsilon} \quad \text{a.s.}$$

Since (x, y) is a nonsingular point of the JLCs, it is not difficult to check that $1/4 \le \min\{\#I_1, \#I_2\}/k^2 \le 1/2$. Therefore, (x_i, y_j) could be detected when n is large enough. This conclusion is uniformly true for $(x, y) \in D \cap \mathscr{D}_\varepsilon$ since $\min_{(x,y)\in D\cap\mathscr{D}_\varepsilon} C(x, y) > 0$. Therefore,

$$\sup_{x\in D\cap\mathscr{D}_\varepsilon}\ \inf_{y\in\hat{D}_n\cap\mathscr{D}_\varepsilon} \|x - y\| = O(1/n) \quad \text{a.s.} \tag{A.4}$$

The theorem is proved after combining (A.3) and (A.4). □

B. Proof of Theorem 3.1. Suppose that the point P has coordinate (x, y) and design point (x_i, y_j) is closest to P. Then

$$|x - x_i| < \frac{1}{2n}, \qquad |y - y_j| < \frac{1}{2n}.$$

Let Q_1 denote the set of the design points in $N^*(x_i, y_j)$ which are also in the band of the JLC with width $3\sqrt{2}\,k/(2n)$. Then Q_2 denotes the set of the

design points in $N^*(x_i, y_j)$ and outside this band. From the proof of Theorem 2.2, we know that almost surely none of the design points in Q_2 could be detected as jump candidates. Let $\underline{\Omega}$ be all possible samples in the sample space such that there are no design points in Q_2 detected as jump candidates. Then $P(\underline{\Omega}) = 1$ and for any sample $\omega \in \underline{\Omega}$, the detected jump points are all in Q_1. Next we concentrate on the samples in $\underline{\Omega}$. The slope of the tangent line of the JLC at P is assumed to be b. Without loss of generality, we assume that $b \neq \infty$. Then detected jump candidate (w_r, v_r) in $N^*(x_i, y_j)$ has the following relationship by the Taylor expansion:

$$v_r - y_j = b(w_r - x_i) + o(k^*/n). \tag{B.1}$$

From (B.1), we have

$$\begin{aligned} \sigma_{wv} &= b\sigma_{ww} + o(k^*/n), \\ \sigma_{vv} &= b^2\sigma_{ww} + o(k^*/n). \end{aligned} \tag{B.2}$$

The slope of the PC line defined by (3.2) is

$$\begin{aligned} \frac{\sigma_{wv}}{\sigma_{ww} - \lambda_1} &= \frac{2\sigma_{wv}}{\sigma_{ww} - \sigma_{vv} + \sqrt{(\sigma_{ww} - \sigma_{vv})^2 + 4\sigma_{wv}^2}} \\ &= \frac{2b\sigma_{ww} + o(k^*/n)}{\sigma_{ww} - b^2\sigma_{ww} + \sqrt{(\sigma_{ww} - b^2\sigma_{ww})^2 + 4b^2\sigma_{ww}^2} + o(k^*/n)} \\ &= b + o(k^*/n). \end{aligned}$$

In the second equation of the above expression, (B.2) is used. Hence the slope of the PC line converges to the slope of the tangent line with rate $o(k^*/n)$. It is obvious that $(\overline{w}, \overline{v})$ converges to P with the same rate. From (3.2), $(\overline{w}, \overline{v})$ is on the PC line. Hence the PC line converges to the tangent line at P with rate $o(k^*/n)$. Since k^* is any positive integer satisfying $\lim_{n\to\infty} k/k^* = 0$, this rate can be $O(n^{-1+2/p})$.

C. Proof of Theorem 5.1. If $(x_i, y_j) \in \mathscr{D}_\varepsilon^*$, then after some routine algebra manipulation [cf. Section 2.5 of Wand and Jones (1995)], we have

$$\begin{aligned} \operatorname{Var}\left(\hat{f}(x_i, y_j)\right) &\approx \frac{L^2\sigma^2}{(k^*)^2} \int_{-L/2}^{L/2}\int_{-L/2}^{L/2} K^2(u,v)\, du\, dv; \\ \text{Bias} &:= E\left(\hat{f}(x_i, y_j)\right) - f(x_i, y_j) \\ &\approx \frac{(k^*)^2}{n^2L^2}\left[f''_{xx}(x_i, y_j)\int_{-L/2}^{L/2}\int_{-L/2}^{L/2} K(u,v)u^2\, du\, dv \right. \\ &\qquad + f''_{xy}(x_i, y_j)\int_{-L/2}^{L/2}\int_{-L/2}^{L/2} K(u,v)uv\, du\, dv \\ &\qquad \left. + f''_{yy}(x_i, y_j)\int_{-L/2}^{L/2}\int_{-L/2}^{L/2} K(u,v)v^2\, du\, dv \right] \\ &\quad + O(1/k^*). \end{aligned}$$

Hence if $k^* = O(n^{2/3})$, then $\mathrm{MSE} := \mathrm{Var}(\hat{f}(x_i, y_j)) + \mathrm{Bias}^2 = O(n^{-4/3})$. Since $f(x, y)$ has continuous second-order partial derivatives in $\mathscr{D}_\varepsilon^*$, the above convergence rate is uniformly true in $\mathscr{D}_\varepsilon^*$.

Now if $(x_i, y_j) \in \mathscr{D}_\varepsilon \setminus \mathscr{D}_\varepsilon^*$, then we have

$$\text{(C.1)}\quad \mathrm{Bias} \approx \frac{k^*}{nL} \frac{f'_x(x_i, y_j) \iint_{\mathscr{T}} K(u, v)\, u\, du\, dv + f'_y(x_i, y_j) \iint_{\mathscr{T}} K(u, v)\, u\, du\, dv}{\iint_{\mathscr{T}} K(u, v)\, du\, dv},$$

$$\text{(C.2)}\quad \mathrm{Var}\left(\hat{f}(x_i, y_j)\right) \approx \frac{L^2 \sigma^2}{(k^*)^2} \frac{\iint_{\mathscr{T}} K^2(u, v)\, du\, dv}{\left[\iint_{\mathscr{T}} K(u, v)\, du\, dv\right]^2}.$$

The integration area $\mathscr{T}$ is a counterpart in $[-L/2, L/2] \times [-L/2, L/2]$ [the area of definition of $K(x, y)$] of $S(x_i, y_j)$ defined in (4.1). If (x_i, y_j) is exactly on a JLC, then $f'_x(x_i, y_j)$ and $f'_y(x_i, y_j)$ in (C.1) and (C.2) could be replaced by $f'_x(x_{i+s}, y_{j+t})$ and $f'_y(x_{i+s}, y_{j+t})$, respectively, where (s, t) is some point in $S(x_i, y_j)$. When $k^* = O(n^{1/2})$, $\mathrm{MSE} = O(n^{-1})$. □

Acknowledgments. The author is grateful to the Editor, Associate Editor and two anonymous referees, whose stimulating comments and suggestions greatly improved the article, and to Jeff Douglas, Gilbert Nestel and Brian Yandell for helpful discussions.

REFERENCES

BESAG, J. (1986). On the statistical analysis of dirty pictures (with discussion). *J. Roy. Statist. Soc. Ser. B* **48** 259–302.

BESAG, J., GREEN, P., HIGDON, D. and MENGERSEN, K. (1995). Bayesian computation and stochastic systems (with discussions). *Statist. Sci.* **10** 3–66.

BHANDARKAR, S. M., ZHANG, Y. and POTTER, W. D. (1994). An edge detection technique using genetic algorithm-based optimization. *Pattern Recognition* **27**, 1159–1180.

BOW, S. T. (1992). *Pattern Recognition and Image Preprocessing*. Marcel Dekker, New York.

BRACEWELL, R. N. (1995). *Two-Dimensional Imaging*. Prentice-Hall, Englewood Cliffs, NJ.

CHENG, K. F. and LIN, P. E. (1981). Nonparametric estimation of a regression function. *Z. Wahrsch. Verw. Gebiete* **57** 223–233.

CRESSIE, N. (1991). *Statistics for Spatial Data*. Wiley, New York.

EUBANK, R. L. and SPECKMAN, P. L. (1994). Nonparametric estimation of functions with jump discontinuities. In *Change-Point Problems* (E. Carlstein, H. G. Müller and D. Siegmund, eds.) 130–144. IMS, Hayward, CA.

FARIN, G. (1993). *Curves and Surfaces for Computer Aided Geometric Design*, 3rd ed. Academic Press, New York.

GASSER, T. and MÜLLER, H. G. (1979). Kernel estimation of regression functions. *Lecture Notes in Math.* **757** 23–68. Springer, Berlin.

GEMAN, S. and GEMAN, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Trans. Pattern Analysis and Machine Intelligence* **6** 721–741.

GONZALEZ, R. C. and WOODS, R. E. (1992). *Digital Image Processing*. Addison-Wesley, Reading, MA.

GREEN, P. J. (1990). Bayesian reconstruction from emission tomography data using a modified EM algorithm. *IEEE Trans. Medical Imaging* **9** 84–93.

HALL, P. and TITTERINGTON, M. (1992). Edge-preserving and peak-preserving smoothing. *Technometrics* **34** 429–440.

KOROSTELEV, A. P. and TSYBAKOV, A. B. (1993). *Minimax Theory of Image Reconstruction*. *Lecture Notes in Statist*. **82**. Springer, New York.

LI, S. Z. (1995). On discontinuity-adaptive smoothness priors in computer vision. *IEEE Trans. Pattern Analysis and Machine Intelligence* **17** 576–586.

LOADER, C. R. (1996). Change point estimation using nonparametric regression. *Ann. Statist*. **24** 1667–1678.

MARR, D. and HILDRETH, E. (1980). Theory of edge detection. *Proc. Roy. Soc. London Ser. A* **207** 187–217.

MCDONALD, J. A. and OWEN, A. B. (1986). Smoothing with split linear fits. *Technometrics* **28** 195–208.

MÜLLER, H. G. (1992). Change-points in nonparametric regression analysis. *Ann. Statist*. **20** 737–761.

MÜLLER, H. G. and SONG, K. S. (1994). Maximin estimation of multidimensional boundaries *J. Multivariate Anal*. **50** 265–281.

O'SULLIVAN, F. (1995). A study of least squares and maximum likelihood for image reconstruction in positron emission tomography. *Ann. Statist*. **23** 1267–1300.

O'SULLIVAN, F. and QIAN, M. (1994). A regularized contrast statistic for object boundary estimation—implementation and statistical evaluation. *IEEE Trans. Pattern Analysis and Machine Intelligence* **16** 561–570.

OWEN, A. (1984). A neighborhood-based classifier for Landsat data. *Canad. J. Statist*. **12** 191–200.

QIU, P. (1994). Estimation of the number of jumps of the jump regression functions. *Comm. Statist. Theory Methods* **23** 2141–2155.

QIU, P. (1997). Nonparametric estimation of jump surface. *Sankhyā Ser. A* **59** 268–294.

QIU, P., ASANO, CHI. and LI, X. (1991). Estimation of jump regression functions. *Bull. Inform. Cybernet*. **24** 197–212.

QIU, P. and BHANDARKAR, S. M. (1996). An edge detection technique using local smoothing and statistical hypothesis testing. *Pattern Recognition Lett*. **17** 849–872.

QIU P. and YANDELL, B. (1997). Jump detection in regression surfaces. *J. Comput. Graph. Statist*. **6** 332–354.

QIU, P. and YANDELL, B. (1998). A local polynomial jump detection algorithm in nonparametric regression. *Technometrics* **40** 141–152.

RICE, J. (1984). Boundary modification for kernel regression. *Comm. Statist. Theory Methods* **13** 893–900.

ROSENFELD, A. and KAK, A. C. (1982). *Digital Picture Processing* **1**, **2**. Academic Press, New York.

RUDEMO, M. and STRYHN, H. (1994). Approximating the distribution of maximum likelihood contour estimators in two-region images. *Scand. J. Statist*. **21** 41–55.

SHIAU, J. H., WAHBA, G. and JOHNSON, D. R. (1986). Partial spline models for the inclusion of tropopause and frontal boundary information in otherwise smooth two- and three-dimensional objective analysis. *J. Atmospheric and Oceanic Technology* **3** 714–725.

SINHA, S. S. and SCHUNK, B. G. (1992). A two-stage algorithm for discontinuity-preserving surface reconstruction. *IEEE Trans. Pattern Analysis and Machine Intelligence* **14** 36–55.

STONE, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *Ann. Statist*. **10** 1040–1053.

SWITZER, P. (1983). Some spatial statistics for the interpretation of satellite data. *Bull. Internat. Statist. Inst*. **50** 962–972.

SWITZER, P. (1986). Statistical image processing. In *Quantitative Analysis of Mineral and Energy Resources* (C. F. Chung, A. G. Fabbri and R. Sinding-Larsen, eds.) 271–282. Reidel, Dordrecht.

SWITZER, P., KOWALIK, W. S. and LYON, R. J. P. (1982). A prior method for smoothing discriminant analysis classification maps. *J. Internat. Assoc. Math. Geology* **14** 433–444.

TITTERINGTON, D. M. (1985). Common structure of smoothing techniques in statistics. *Internat. Statist. Rev*. **53** 141–170.

TORRE, V. and POGGIO, T. A. (1986). On edge detection. *IEEE Trans. Pattern Analysis and Machine Intelligence* **8** 147–163.

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

WANG, Y. (1995). Jump and sharp cusp detection by wavelets. *Biometrika* **82** 385–397.

WU, J. S. and CHU, C. K. (1993a). Kernel type estimators of jump points and values of a regression function. *Ann. Statist.* **21** 1545–1566.

WU. J. S. and CHU, C. K. (1993b). Nonparametric function estimation and bandwidth selection for discontinuous regression functions. *Statist. Sinica* **3** 557–576.

YI, J. H. and CHELBERG, D. M. (1995). Discontinuity-preserving and viewpoint invariant reconstruction of visible surfaces using a first order regularization. *IEEE Trans. Pattern Analysis and Machine Intelligence* **17** 624–629.

SCHOOL OF STATISTICS
UNIVERSITY OF MINNESOTA
270 VINCENT HALL
206 CHURCH STREET S.E.
MINNEAPOLIS, MINNESOTA 55455
E-MAIL: qiu@stat.umn.edu