Local Smoothing Image Segmentation For Spotted Microarray Images

Peihua Qiu[1] and Jingran Sun[2]

[1]School of Statistics, University of Minnesota, 313 Ford Hall,
224 Church St. SE, Minneapolis, MN 55455

[2]US Clinical Development Biostatistics, Amgen Inc., One Amgen
Center Drive, Mailstop 24-2-C, Thousand Oaks, CA 91320

Abstract

Gene microarray data are used in a large variety of applications, including pharmaceutical and clinical research. By comparing gene expression in normal and abnormal cells, microarrays can be used for identifying genes involved in particular diseases, and then these genes can be targeted by therapeutic drugs. Most gene expression data are produced from spotted microarray images. A spotted microarray image consists of thousands of spots, with individual DNA sequences first printed at each spot and then equal amounts of probes (e.g., cDNA samples) from treatment and control cells mixed and hybridized with the printed DNA sequences. To obtain gene expression data, the image needs to be segmented first to separate foregrounds from backgrounds for individual spots, and then averages of foreground pixels are used for computing the gene expression data. So image segmentation of microarray images is related directly to the *reliability* of gene expression data. Several image segmentation procedures have been suggested in the literature, and included in software packages handling gene microarray data. In this paper, a new image segmentation methodology is proposed based on local smoothing. Theoretical arguments and numerical studies show that it has good statistical properties and would perform well in applications.

Key Words: Background; Boundary curves; Consistency; Edge detection; Foreground; Image processing; Image segmentation; Jump location curves; Gene expression data; Gray levels; Local polynomial kernel smoothing; Nonparametric regression; Spots.

1 Introduction

This paper discusses image segmentation for analyzing gene microarray images. Each human cell can be viewed as a functional unit containing 20,000-50,000 genes, whose expression levels determine

the transcriptional state of the cell. By studying the gene expression levels of different individuals in a population, we can have a better understanding of the biological differences among them. Microarrays are part of a new class of biotechnologies which allow the monitoring of expression levels for thousands of genes simultaneously (Lashkari *et al.* 1997, Chu *et al.* 1998, Ferea *et al.* 1999).

In a typical microarray experiment, thousands of individual DNA sequences are first printed in a high density array on a glass microscope slide using a robotic arrayer. Then equal amounts of probes (e.g., cDNA samples) from treatment and control cells are labeled with different colors (e.g., the red-fluorescent dye Cy5 and the green-fluorescent dye Cy3), and mixed with the arrayed DNA sequences. In each spot, the probes matching the gene of the printed DNA sequences will attach to the printed DNA sequences. After this competitive hybridization, the slides are imaged by an imaging device which makes fluorescence measurements for each dye. Then a pair of fluorescence images are obtained, one for each dye. The average fluorescence intensity of each spot can therefore measure the expression level of one specific gene in either the treatment or control cells. See Yang *et al.* (2002) for a more detailed introduction.

A regular fluorescence image includes thousands of spots arranged in rows and columns. During the *image addressing* stage, the entire image is divided into many grid cells, also in rows and columns, each grid cell includes one spot in the middle, and then coordinates are assigned to the borders and/or centers of the grid cells. *Image segmentation* is a process to classify pixels in each grid cell as either foreground or background pixels. All foreground pixels correspond to the spot of interest in that grid cell. Usually the pair of fluorescence images, made for the two different dyes, are combined in some specific way pixel by pixel and then image segmentation is performed on the combined image. After image segmentation, the average fluorescence intensity, or other more robust measures, for all foreground pixels of a given spot can be calculated, which is denoted as R and G, respectively, for the images with red-fluorescent dye and green-fluorescent dye. The *gene expression data* are then generated from them. From this brief introduction, we can see that *image segmentation of gene microarray images is crucial to the reliability of gene expression data and all subsequent statistical analysis.*

There are several image segmentation procedures in the literature for analyzing gene microarray images. Eisen (1999) provided a fixed circle segmentation procedure in the software *ScanAlyze*, which fits circles with a constant diameter to all spots in an image, and user-intervention to adjust

manually the height and width of a single spot is also allowed. The software *GenePix* (1999) provided an adaptive circle segmentation procedure, which fitted a circle to a spot with its diameter estimated separately for each spot. The seeded region growing (SRG) procedure suggested by Adams and Bischof (1994) works more flexibly as follows. For each grid cell of a microarray image, we first choose a set of seeds as foreground pixels and another set as background pixels, which is convenient to do because the center and border of the grid cell are specified in the image addressing stage. Then, at each step, those pixels whose statuses are undecided but at least one of their neighboring pixels has been allocated are considered for allocation. One such pixel whose intensity level is closest to the average intensity level of one of its neighboring regions (either foreground or background region) is allocated to that region. This process continues until all pixels are allocated. Another commonly used segmentation procedure was suggested by Chen *et al.* (1997), which segments the foreground from the background of a grid cell by thresholding the histogram of all intensities whose pixel locations are within a target mask. The threshold parameter was selected based on the Mann-Whitney statistic, and a pixel whose intensity was larger than the threshold was classified as a foreground pixel, and as a background pixel otherwise. More recent image segmentation procedures include clustering algorithms (e.g., Bozinov and Rahnenführer 2002, Glasbey and Ghazal 2003), segmentation based on Gaussian density estimation (Steinfath *et al.* 2001), segmentation using mathematical morphology (Angulo and Serra 2003), and so forth.

Limitation of the circle segmentation procedures with fixed or adaptive diameters is obvious: the actual sizes and shapes of different spots might be quite different, but these methods do not have the flexibility to accommodate such differences. The SRG procedure is more flexible and it can adapt to different shapes and sizes of the spots. But it is quite sensitive to noise because little smoothing is involved, and results are not invariant under selection of seeds. To overcome this limitation, in applications, people often pre-smooth spotted images before applying segmentation procedures, and a popular pre-smoothing operator is based on the 2-D Gaussian density function. While noise is removed by such a pre-smoother to a certain degree, some small jumps in the underlying image intensity function are also blurred, which would worsen the segmentation results of the SRG and some other existing procedures (cf., related discussion in Section 6.7 of Qiu (2005)). The segmentation procedure by thresholding the intensity histogram is easy to implement after the threshold parameter is determined, but this procedure is a global method in the sense that the threshold parameter is selected globally by all pixels in the selected target mask of the

grid cell considered, and thus it cannot easily accommodate local features of the boundary curve. Furthermore, it is sensitive to noise because little smoothing is involved. The clustering algorithm by Bozinov and Rahnenführer (2002) is based on both K-means clustering and partitioning around medoids. It takes into account some local features of the image intensities, but its segmented foregrounds and backgrounds often do not form connected regions when the true ones are actually connected regions. It is also quite sensitive to noise. The procedure by Glasbey and Ghazal (2003) thresholds the intensity histogram for image segmentation, as did by Chen *et al.* (1997); but its threshold value is chosen based on the assumption that both the foreground and background intensities have Gaussian distributions. Its results should be more efficient than the ones of Chen *et al.*'s procedure when the Gaussian distribution assumption is valid; but such an assumption may not hold in some applications. Steinfath *et al.*'s (2001) procedure is based on the assumption that the image intensity function has the parametric form of a circular 2-D Gaussian density, which might be restrictive. The superimum/infimum operators involved in the morphological procedures are sensitive to individual image intensity values. Therefore, they can provide reliable results only when the noise level is low and the observed data have no outliers.

In this paper, we suggest a new image segmentation procedure for analyzing spotted microarray images. In our procedure, each grid cell of a microarray image is regarded as a surface of the image intensity function, and the boundary curve separating the foreground region from the background region in the grid cell is estimated based on local polynomial kernel smoothing. This procedure overcomes most limitations of the existing procedures mentioned above. First, it is a local smoothing procedure. So, it should remove noise and accommodate local features of the image well. Second, it is a nonparametric procedure. We do not impose any parametric form on the true boundary curves, on the error distributions, and on the true image intensity function. Therefore, it should have the flexibility to accommodate different shapes and sizes of the true boundary curves, different error distributions, and different patterns of the true image intensity function.

Jump detection in regression surfaces is related to the image segmentation problem discussed in this paper. But the two problems are different in the sense that boundary curves in the latter problem are simple, closed, continuous curves (cf., Section 2.1 for a detailed description), and good image segmentation procedures for handling spotted microarray images should make use of these special features. For discussions about jump detection in regression surfaces, please read Hall and Rau (2000, 2002), Hall *et al.* (2001), Qiu (1997, 2002), Qiu and Bhandarkar(1996), Qiu

and Yandell (1997), and the references cited there. The image segmentation problem discussed here is also related to edge detection in image processing (cf., Qiu 2005, Chapter 6). However, the detected edges by most existing edge detectors are point sets which are scattered in the whole design space. These edge detectors can not be used directly for segmentation of spotted microarray images, because the foreground pixels and background pixels are not well defined even after the edges are detected by them.

The remaining part of the article is organized as follows. The proposed image segmentation procedure is introduced in detail in Section 2. Some of its statistical properties are discussed in Section 3. Simulation results are presented in Section 4, which show that the new procedure outperforms its peers in various cases. Then, the current procedure and several existing ones are applied to a real microarray image in Section 5. Some remarks conclude the article in Section 6. Some technical details are given in appendices.

2 Image Segmentation By Local Smoothing

This section is organized in five parts. In Section 2.1, we give a mathematical formulation of the image segmentation problem for analyzing spotted microarray images. Then our image segmentation procedure is introduced in Section 2.2. A simplified version of this procedure is discussed in Section 2.3. Selection of the bandwidth used in the proposed procedure is discussed in Section 2.4. Finally, a pseudo-code of the proposed procedure is provided in Section 2.5.

2.1 Formulation of the image segmentation problem

For a spotted microarray image, borders of its grid cells, each of which contains a spot in the middle, can be roughly specified by the arrayer in the image addressing stage. See Bergemann *et al.* (2004) for such an image addressing method, which is also used in the numerical example in Section 5. Therefore, image segmentation can be performed separately for individual grid cells, which has been a common practice in analyzing microarray images. For this reason, our discussion below is for handling a single grid cell only.

For a given grid cell, suppose that the origin of the coordinate system is at its center, pixel locations are $\{(x_i, y_j), i = 1, 2, \ldots, n_x, j = 1, 2, \ldots, n_y\}$, and observed image intensities are $\{Z_{ij},$

$i = 1, 2, \ldots, n_x$,$j = 1, 2, \ldots, n_y\}$, where n_x and n_y denote the numbers of columns and rows, respectively, of the pixels. Without loss of generality, we assume that $n_x = n_y = n$ in this paper. In most cases, the boundary curve Γ separating the foreground from the background in the grid cell can be reasonably assumed to be a continuous closed curve. Then, Γ can be expressed by

$$\begin{cases} x(\theta) &= r(\theta)\cos(\theta) \\ y(\theta) &= r(\theta)\sin(\theta), \end{cases} \tag{1}$$

where $r(\theta) > 0$ denotes the Euclidean distance from the point $(x(\theta), y(\theta))$ on Γ to the origin, and $\theta \in [0, 2\pi)$ is the angle formed by the line segment from the origin to the point $(x(\theta), y(\theta))$ and the positive direction of the x-axis. For convenience of presentation, we use $N = n^2$, N_1, and N_2 to denote the total number of pixels in the grid cell, the number of pixels in foreground, and the number of pixels in background, respectively.

Because the boundary curve Γ is assumed to be continuous and closed, $x(\theta)$ and $y(\theta)$ are both assumed to be continuous functions of θ on $[0, 2\pi)$, and $(x(0), y(0)) = \lim_{\theta \to 2\pi - 0}(x(\theta), y(\theta))$. In the framework of (1), estimation of the boundary curve Γ is equivalent to estimating $r(\theta)$ for each $\theta \in [0, 2\pi)$.

In equation (1), we do not put any restriction on the shape of the boundary curve Γ, besides it is required to be a continuous closed curve. However, this formulation does not cover the quite common phenomenon of "donut" spots, each of which has a hole around its center, and the region inside the hole is a separate part of background, besides the part of background outside the spot. To describe a "donut" spot properly, we assume that the foreground has two boundary curves Γ_1 and Γ_2, both of which are continuous closed curves that can be expressed by equation (1), with radius functions $r_1(\theta)$ and $r_2(\theta)$, respectively, satisfying $r_1(\theta) < r_2(\theta)$, for all $\theta \in [0, 2\pi)$.

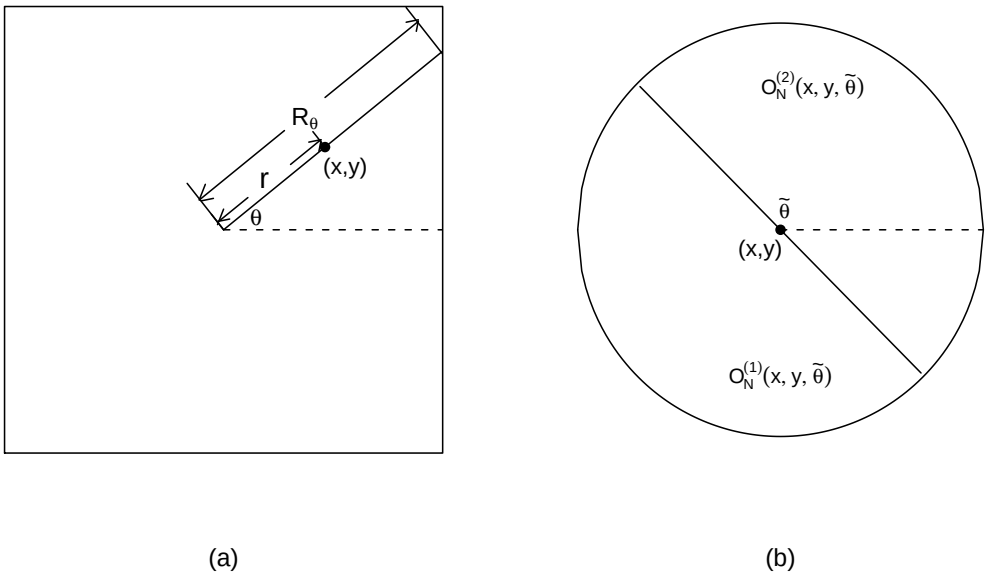
2.2 Image segmentation by searching for gradient directions

We first discuss the simpler case that the foreground has a single boundary curve Γ. To estimate Γ, let us consider a half-line starting from the origin and forming an angle θ with the positive x-axis. Any point (x, y) on this half line and within the grid cell considered has the expression $(x, y) = (r\cos(\theta), r\sin(\theta))$, where $0 \le r \le R_\theta$ and R_θ denotes the Euclidean length of the half-line segment within the grid cell, as demonstrated by Figure 1(a). For the point (x, y), let us consider its circular neighborhood $O_N(x, y) = \{(u, v) : \sqrt{(u-x)^2 + (v-y)^2} \le h_N\}$, where $h_N > 0$

is a bandwidth parameter. This neighborhood is then divided into two halves $O_N^{(1)}(x, y, \tilde{\theta})$ and $O_N^{(2)}(x, y, \tilde{\theta})$, by a line passing through (x, y) and forming an angle $\tilde{\theta}$ with the positive x-axis, where $\tilde{\theta} \in [0, \pi)$. See Figure 1(b) for a demonstration. Then, in $O_N^{(1)}(x, y, \tilde{\theta})$ and $O_N^{(2)}(x, y, \tilde{\theta})$, we fit two one-sided local constant plans by the following local constant kernel smoothing procedure:

$$\min_{a^{(\ell)} \in R} \sum_{(x_i, y_j) \in O_N^{(\ell)}(x, y, \tilde{\theta})} \left(Z_{ij} - a^{(\ell)} \right)^2 K \left(\frac{x_i - x}{h_N}, \frac{y_j - y}{h_N} \right), \text{ for } \ell = 1, 2, \tag{2}$$

where K is a radially symmetric, bivariate density kernel function with support $\{(x, y) : x^2 + y^2 \leq 1\}$. The solutions to $a^{(\ell)}$ of (2) are denoted by $\widehat{a}^{(\ell)}(x, y, \tilde{\theta})$, for $\ell = 1$ and 2, which are the one-sided local constant kernel estimators, also known as the one-sided Nadaraya-Watson kernel estimators, of the image intensity function f at (x, y). Therefore $\widehat{a}^{(\ell)}(x, y, \tilde{\theta})$, for $\ell = 1$ and 2, are weighted averages of the observations in $O_N^{(1)}(x, y, \tilde{\theta})$ and $O_N^{(2)}(x, y, \tilde{\theta})$, respectively, with the weights determined by the kernel function K.



(a) (b)

Figure 1: (a) For a given grid cell, we consider a half-line starting from the origin and forming an angle θ with the positive x-axis. Any point (x, y) on this half-line has the expression $(x, y) = (r\cos(\theta), r\sin(\theta))$, where $0 \leq r \leq R_\theta$. (b) The circular neighborhood $O_N(x, y)$ is divided into two parts $O_N^{(1)}(x, y, \tilde{\theta})$ and $O_N^{(2)}(x, y, \tilde{\theta})$ by a line passing through (x, y) and forming an angle $\tilde{\theta}$ with the positive x-axis.

Then, for any $(x, y) = (r\cos(\theta), r\sin(\theta))$ in the grid cell, we define

$$M_N(r, \theta) = \max_{\tilde{\theta} \in [0, \pi)} \left| \widehat{a}^{(1)}(x, y, \tilde{\theta}) - \widehat{a}^{(2)}(x, y, \tilde{\theta}) \right|. \tag{3}$$

Intuitively, if (x, y) is on the true boundary curve Γ and Γ has a unique tangent line at (x, y), then the maximizer of (3) should be a good estimator of the tangent direction of Γ at (x, y), and consequently $M_N(r, \theta)$ should be relatively large. On the other hand, if (x, y) is a certain distance away from Γ, then $M_N(r, \theta)$ is usually small because the two one-sided estimators are close to each other in such cases. Our estimator of Γ is then defined by

$$\widehat{\Gamma} = \{(\widehat{r}(\theta)\cos(\theta), \widehat{r}(\theta)\sin(\theta)), \text{ for } \theta \in [0, 2\pi)\}, \tag{4}$$

where $\widehat{r}(\theta)$ is the maximizer of $\max_{r\in[0,R_\theta]} M_N(r, \theta)$. Namely,

$$\widehat{r}(\theta) = \arg \max_{r\in[0,R_\theta]} M_N(r, \theta). \tag{5}$$

In (5), we assume that the maximizer $\widehat{r}(\theta)$ is unique for each $\theta \in [0, 2\pi)$. In the case of multiple maximizers for a given θ, which is an event with zero probability under some regularity conditions, $\widehat{r}(\theta)$ is defined by their simple average.

It should be pointed out that the notorious "boundary problem" of the local smoothing procedures, i.e., the bias of a local curve or surface estimator in boundary regions is usually larger than that in interior regions, is not a big issue for image segmentation of spotted microarray images, because the foreground region in each grid cell can be reasonably assumed to be in the middle (cf., Figure 4 in Section 5 for a real-life example). Thus, it is reasonable to search only a subset $[tR_\theta, (1-t)R_\theta]$ of $[0, R_\theta]$ for $\widehat{r}(\theta)$ in (5), where $0 < t < 0.5$ is a small number. We also want to mention that, when constructing the criterion $M_N(r, \theta)$, it is possible to divide the circular neighborhood $O_N(x, y)$ into two parts by a quadratic curve instead of a straight line (cf., Figure 1(b)), to reflect the fact that true boundary curves are usually closed curves and they can be approximated better by quadratic curves. However, variability of the approximated quadratic curves is usually large, especially when the sample size is small, which is often the case in the current problem. A careful investigation of this possible improvement is left to our future research.

In the estimation procedure (3)-(5), local constant kernel smoothing has been used for obtaining the two one-sided kernel estimators $\widehat{a}^{(1)}(x, y, \tilde{\theta})$ and $\widehat{a}^{(2)}(x, y, \tilde{\theta})$ (cf., equations (2) and (3)). A natural "improvement" is to use the following local linear kernel smoothing procedure:

$$\min_{a^{(\ell)},b^{(\ell)},c^{(\ell)}\in R} \sum_{(x_i,y_j)\in O_N^{(\ell)}(x,y,\tilde{\theta})} \left\{Z_{ij} - \left[a^{(\ell)} + b^{(\ell)}(x_i - x) + c^{(\ell)}(y_j - y)\right]\right\}^2 K\left(\frac{x_i - x}{h_N}, \frac{y_j - y}{h_N}\right), \tag{6}$$

for $\ell = 1, 2$. In the literature, it has been demonstrated that conventional local linear kernel estimators have less bias in the boundary regions of the design space, compared to conventional

local constant kernel estimators (e.g., Hastie and Loader 1993; Fan and Gijbels 1996). In the boundary regions, the bias of the latter estimators is mainly caused by the slope of the true regression function in such regions. Since one-sided surface estimators are used in our boundary curve detection criterion $M_N(r,\theta)$, the boundary curve Γ in the current image segmentation problem is similar in nature to the border of design space in the conventional 2-dimensional nonparametric regression problem, because the related estimators in both cases are constructed from observations located on a single side of a given point. So we can expect that the local linear kernel smoothing procedure (6) would perform better than the local constant kernel smoothing procedure (2), in excluding the impact of surface slope on boundary curve estimation. Similar statements have been made by Gijbels *et al.* (1999) in 1-D cases. However, if $M_N(r,\theta)$ is constructed from the two one-sided local linear kernel estimators obtained from (6), it would suffer the following two limitations: (i) it is relatively noisy compared to the one defined at (3) based on the one-sided local constant kernel estimators, which is especially true in the current problem due to its relatively small sample size, and (ii) for a given $\theta \in [0, 2\pi)$, it usually has three local maxima around Γ, one is close to Γ and the other two are on either side of Γ (cf., e.g., Qiu 2005, Figure 3.2(b)), and consequently the local maximum close to Γ is often difficult to select for estimating Γ, due to the relatively large variability of the one-sided local linear kernel estimators, as mentioned above. Based on our numerical experience, the benefit of using local linear kernel smoothing is not enough to offset its disadvantages mentioned above for most microarray images; see Section 4 for a numerical example. For that reason, the local constant kernel smoothing procedure (2) is used in this paper in defining the boundary detection criterion $M_N(r,\theta)$.

Now, let us discuss estimation of the boundary curves Γ_1 and Γ_2 of a "donut" spot. Let

$$\tilde{r}_1(\theta) = \arg \max_{r\in[0,R_\theta]} M_N(r,\theta), \;\; \tilde{r}_2(\theta) = \arg \max_{r\in[0,R_\theta]\setminus(\tilde{r}_1(\theta)-h_N,\tilde{r}_1(\theta)+h_N)} M_N(r,\theta). \tag{7}$$

Then, we define

$$\widehat{r}_1(\theta) = \tilde{r}_{(1)}(\theta), \;\; \widehat{r}_2(\theta) = \tilde{r}_{(2)}(\theta), \tag{8}$$

where $\tilde{r}_{(1)}(\theta) \le \tilde{r}_{(2)}(\theta)$ are the two order statistics of $\tilde{r}_1(\theta)$ and $\tilde{r}_2(\theta)$. After $\widehat{r}_1(\theta)$ and $\widehat{r}_2(\theta)$ are defined, estimators of the two boundary curves Γ_1 and Γ_2 can be defined similarly to $\widehat{\Gamma}$ in equation (4).

In applications, if it is clear based on our visual impression whether there are "donut" spots in a microarray image, then we can simply choose between the two procedures (4)–(5) and (7)–

(8). If, however, it is difficult to make such a judgment based on our visual impression alone, we suggest using the following data-driven decision rules. For a given grid cell, we compute two sets of boundary curve(s) using the procedures (4)–(5) and (7)–(8), respectively, and let (I_f, I_b) and (I_f^*, I_b^*) be pairs of averaged foreground image intensity and averaged background image intensity in the two setups. Then, we conclude that the grid cell does not contain a "donut" spot if

$$I_f - I_b > I_f^* - I_b^*, \tag{9}$$

and the opposite decision is made otherwise. It should be pointed out that more robust summary statistics, such as the trimmed means and medians, can be used in (9) in places of I_f, I_b, I_f^* and I_b^*. Also, this judgment step does not add much extra computation, because the estimator $\widehat{\Gamma}$ is readily available after the estimators $\widehat{\Gamma}_1$ and $\widehat{\Gamma}_2$ are computed.

2.3 Image segmentation by gradient estimation

In (3), $M_N(r,\theta)$ is obtained by searching all possible directions $\tilde{\theta}$ in $[0,\pi)$, at any given point $(x,y) = (r\cos(\theta), r\sin(\theta))$ in the grid cell considered. In applications, (3) needs to be replaced by its discrete version

$$M_N(r,\theta) = \max_{1\le j\le \tilde{m}} \left| \widehat{a}^{(1)}(x,y,\tilde{\theta}_j) - \widehat{a}^{(2)}(x,y,\tilde{\theta}_j) \right|,$$

where $\{\tilde{\theta}_j, j = 1,2,\ldots,\tilde{m}\}$ is a sequence of equally spaced directions in $[0,\pi)$. From our experience, the above searching algorithm is applicable in most cases, because the size of each grid cell of a typical microarray image is around 50×50, and the sequence $\{\tilde{\theta}_j, j = 1,2,\ldots,\tilde{m}\}$ should be dense enough to cover all interesting directions in $[0,\pi)$ if we choose, say, $\tilde{m} = 40$. That is, the results would hardly change if $\tilde{m}$ is chosen larger. However, sometimes we still prefer a faster boundary detection procedure, because a microarray image usually contains thousands of grid cells. This can be accomplished by replacing the searching algorithm (3) with the gradient estimation procedure introduced below.

In the neighborhood $O_N(x,y)$ of a given point (x,y) in the grid cell, let us fit a local plane by the following conventional local linear kernel smoothing procedure:

$$\min_{a,b,c\in R} \sum_{(x_i,y_j)\in O_N(x,y)} \{Z_{ij} - [a + b(x_i - x) + c(y_j - y)]\}^2 K\left(\frac{x_i - x}{h_N}, \frac{y_j - y}{h_N}\right). \tag{10}$$

The gradient direction of the fitted plane is denoted by $G(x,y) = (\widehat{b}(x,y), \widehat{c}(x,y))$, where $\hat{b}(x,y)$ and $\hat{c}(x,y)$ are solutions to b and c, respectively, of (10). Then the image intensity function f

increases fastest around this direction at (x, y). If the point (x, y) is on a boundary curve, then $G(x, y)$ should also indicate the orientation of the boundary curve at (x, y) well (cf., Figures 2(b), 2(f), and 2(j) in Section 4). More specifically, let $\widehat{\theta}_T(x, y) \in [0, \pi)$ be orthogonal to $G(x, y)$. Then $\widehat{\theta}_T(x, y)$ should be a good estimator of the tangent direction of the boundary curve at (x, y) in such a case. So the boundary detection criterion can be defined by

$$M_N^*(r, \theta) = \left| \widehat{a}^{(1)}(x, y, \widehat{\theta}_T(x, y)) - \widehat{a}^{(2)}(x, y, \widehat{\theta}_T(x, y)) \right|, \tag{11}$$

where $\widehat{a}^{(1)}$ and $\widehat{a}^{(2)}$ are solutions of (2). In the case that the grid cell has a spot with a single boundary curve Γ, the estimator of Γ can be defined by

$$\widehat{\Gamma}^* = \{(\widehat{r}^*(\theta)\cos(\theta), \widehat{r}^*(\theta)\sin(\theta)), \text{ for } \theta \in [0, 2\pi)\}, \tag{12}$$

where $\widehat{r}^*(\theta)$ is the maximizer of $\max_{r \in [0, R_\theta]} M_N^*(r, \theta)$. In the case that the grid cell has a "donut" spot, then estimators of the two boundary curves Γ_1 and Γ_2 can be defined similarly to those in equations (7) and (8), which are denoted as $\widehat{\Gamma}_1^*$ and $\widehat{\Gamma}_2^*$ below. By using criterion (11), no direction search is involved, and therefore estimators $\widehat{\Gamma}^*$, $\widehat{\Gamma}_1^*$, and $\widehat{\Gamma}_2^*$ are much easier to compute compared to estimators $\widehat{\Gamma}$, $\widehat{\Gamma}_1$, and $\widehat{\Gamma}_2$ defined by criterion (3). On the other hand, estimators based on criterion (11) may lose some efficiency, mainly due to the randomness of the estimated image gradient, although the loss is small based on our numerical studies (cf., Figure 2 and Tables 1 and 3 in Section 4).

2.4 Bandwidth selection

To compute $\widehat{\Gamma}$, $\widehat{\Gamma}^*$, or estimators of the two boundary curves of a "donut" spot, the bandwidth parameter h_N should be selected properly. Toward this end, we suggest the following cross-validation procedure. For simplicity, our discussion is for computing $\widehat{\Gamma}$ only. Bandwidth selection for computing other boundary curve estimators can be discussed similarly.

Suppose that $\widehat{\Gamma}$ divides the grid cell in question, denoted as Ω, into two parts $\widehat{\Omega}^{(1)}$ and $\widehat{\Omega}^{(2)}$, which are good estimators of the foreground region $\Omega^{(1)}$ and the background region $\Omega^{(2)}$ of the grid cell, respectively. After Γ is estimated by $\widehat{\Gamma}$, the image intensity function f can be estimated separately in the two regions $\widehat{\Omega}^{(1)}$ and $\widehat{\Omega}^{(2)}$ by the conventional local linear kernel smoothing procedure (i.e., the procedure (10) after $O_N(x, y)$ is replaced by either $\widehat{\Omega}^{(1)}$ or $\widehat{\Omega}^{(2)}$ depending on which region the point (x, y) is in). The resulting estimator of f is called the *jump-preserving*

estimator of f in this paper. If $\widehat{\Gamma}$ estimates Γ well, then the jump-preserving estimator of f should also estimate f well, and vice versa. Therefore, h_N can be chosen such that the corresponding jump-preserving estimator estimates f well. However, the jump-preserving estimator also needs a bandwidth, which could be different from the bandwidth h_N used in boundary curve estimation. That bandwidth is denoted by $\tilde{h}_N$.

Let $\widehat{f}_{-i,-j}$ be the jump-preserving estimator of f without using the (i,j)-th observation. The cross-validation (CV) score is then defined by

$$\mathrm{CV}(h_N, \tilde{h}_N) = \frac{\frac{1}{\widehat{N}_1 - 1} \sum_{(x_i,y_j)\in\widehat{\Omega}^{(1)}} \left(Z_{ij} - \widehat{f}_{-i,-j}(x_i,y_j)\right)^2}{\widehat{\sigma}_1^2} + \frac{\frac{1}{\widehat{N}_2 - 1} \sum_{(x_i,y_j)\in\widehat{\Omega}^{(2)}} \left(Z_{ij} - \widehat{f}_{-i,-j}(x_i,y_j)\right)^2}{\widehat{\sigma}_2^2}, \tag{13}$$

where $\widehat{N}_1$ and $\widehat{N}_2$ denote the numbers of pixels in $\widehat{\Omega}^{(1)}$ and $\widehat{\Omega}^{(2)}$, respectively, and $\widehat{\sigma}_1^2$ and $\widehat{\sigma}_2^2$ are some consistent estimators of the noise variances in the two regions. Let $\mathrm{CV}(h_N) = \min_{\tilde{h}_N} \mathrm{CV}(h_N, \tilde{h}_N)$. Then, the optimal bandwidth for boundary estimation can be estimated by the minimizer of $\mathrm{CV}(h_N)$.

The CV score in (13) is a sum of two terms, corresponding to two regions separated by the estimated boundary curve $\widehat{\Gamma}$, and each term is a ratio of the regional CV score to an estimated regional noise variance. A major consideration in defining such a CV score is that noise variances in foreground and background regions of a typical grid cell of a microarray image are quite different, and the two regional ratios in (13) are for accommodating such a difference. Because the center and border of each grid cell in a microarray image are determined in the image addressing stage, it is not difficult to obtain two useful estimators $\widehat{\sigma}_1^2$ and $\widehat{\sigma}_2^2$. One example to define such estimators is given in Section 5 when we analyze a microarray image. Theoretically, it has been proved that the selected bandwidth by (13) is asymptotically equivalent to the optimal bandwidth defined as the minimizer of the distance measure $d_{\Omega^{(1)},\widehat{\Omega}^{(1)}}(h_N)$ defined in Section 5. To save some space, this result is omitted here, and is included in Appendix C of Qiu and Sun (2006). We also want to mention that bootstrap-type bandwidth selection has been considered in 1-D cases recently by Gijbels and Goderniaux (2004). We did not consider such methods here because computation involved would be expensive in 2-D cases.

2.5 A pseudo-code

We provide a pseudo-code of the proposed segmentation procedure based on gradient estimation, as follows.

1. At pixel (x, y) of a given grid cell, compute the estimated gradient $G(x, y)$ using the local linear kernel smoothing procedure (10) with bandwidth h_N.
2. Construct the criterion $M_N^*(r, \theta)$ using formula (11), where the two one-sided estimators $\widehat{a}^{(1)}$ and $\widehat{a}^{(2)}$ are computed using the local constant kernel smoothing procedure (2).
3. Compute two maxima $\tilde{r}_1(\theta)$ and $\tilde{r}_2(\theta)$ of $M_N^*(r, \theta)$, as in formula (7). Then, define $\widehat{r}(\theta) = \tilde{r}_1(\theta)$ when the spot is assumed to have a single boundary curve; define $\widehat{r}_1(\theta)$ and $\widehat{r}_2(\theta)$ from $\tilde{r}_1(\theta)$ and $\tilde{r}_2(\theta)$ using formula (8) when the spot is assumed to have two boundary curves.
4. Decide whether the spot has two boundary curves using criterion (9).
5. Determine bandwidth h_N using the CV procedure (13), and obtain estimator(s) of the boundary curve(s) using the CV bandwidth.

The proposed procedures can be easily computed because of the "parallel" structure of the segmentation problem discussed here and of the local smoothing nature of the procedures. To handle one typical grid cell requires about 15 seconds CPU time on our 1.2GHz Pentium III PC running a Linux operating system.

3 Some Statistical Properties

In this section, we give some statistical properties of the estimated boundary curve by the image segmentation procedure (11)–(12), which is based on gradient estimation. By similar arguments, it can be shown that, under some regularity conditions, other boundary curve estimators (e.g., the ones defined by equations (4)–(5) and equations (7)–(8)) share the same properties given in this section. First, we have the following result about the estimated gradient direction $G(x, y) = (\widehat{b}(x, y), \widehat{c}(x, y))$ obtained in procedure (10).

Theorem 3.1 *Assume that the image intensity function f has continuous second-order derivatives in the foreground and background of the grid cell Ω considered, and it has one-sided directional*

second-order derivatives at each point of the boundary curve Γ *in the normal direction of* Γ. *It is further assumed that* Γ *is a closed curve and its radius function* $r(\theta)$ *(cf. expression (1)) has continuous second-order derivatives on* $[0, 2\pi)$, *the absolute jump magnitudes of* f *along* Γ *have a positive lower bound, the kernel function* K *is a radially symmetric, Lipschitz continuous density function on its support, the bandwidth* h_N *satisfies the conditions that* $h_N = o(1)$ *and* $\frac{\sqrt{\log(n)}}{nh_N^2} = o(1)$, *where* $N = n^2$, *and the random errors involved in observed image intensities are i.i.d. with mean zero and finite variances* σ_1^2 *and* σ_2^2, *respectively, in the foreground and background regions. Then*

(i)

$$\|\widehat{b} - f'_x\|_{\Omega_{h_N}} = o(h_N) + O\left(\frac{\sqrt{\log(n)}}{nh_N^2}\right), \text{ almost surely (a.s.)}, \tag{14}$$

$$\|\widehat{c} - f'_y\|_{\Omega_{h_N}} = o(h_N) + O\left(\frac{\sqrt{\log(n)}}{nh_N^2}\right), \text{ a.s.}, \tag{15}$$

where $\Omega_{h_N} = \{(x, y) : (x, y) \in \Omega, \sqrt{(x - x')^2 + (y - y')^2} \geq h_N$, *and* (x', y') *is any point on* Γ *or the border of* $\Omega\}$, *and* $\|f\|_{\Omega_{h_N}} = \max_{(x,y)\in\Omega_{h_N}} |f(x, y)|$;

(ii) for a point $(x, y) \in \Omega$, *if* (x_τ, y_τ) *is the closest point on* Γ *to* (x, y) *and their Euclidean distance is* τh_N *where* $0 \leq \tau < 1$ *is a constant, then*

$$\widehat{b}(x, y) = \frac{C_\tau}{K_{02}h_N} \int\int_{Q^{(2)}} uK(u, v)\, dudv + o(1/h_N), \text{ a.s.}, \tag{16}$$

$$\widehat{c}(x, y) = \frac{C_\tau}{K_{20}h_N} \int\int_{Q^{(2)}} vK(u, v)\, dudv + o(1/h_N), \text{ a.s.}, \tag{17}$$

where $C_\tau > 0$ *is the jump size of* f *at* (x_τ, y_τ), $Q_N^{(1)}(x, y)$ *and* $Q_N^{(2)}(x, y)$ *are two different parts of* $O_N(x, y)$ *separated by* Γ *with a positive jump at* (x_τ, y_τ) *from* $Q_N^{(1)}(x, y)$ *to* $Q_N^{(2)}(x, y)$, $Q^{(1)}$ *and* $Q^{(2)}$ *are the two corresponding parts of the support of* K, $K_{02} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} v^2K(u, v)dudv$ *and* $K_{20} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} u^2K(u, v)dudv$.

For people who are not familiar with the terminology of almost sure convergence, the statement that a sequence of random variables X_n converges to a random variable X almost surely implies that the event that X_n does not converge to X has zero probability. By Theorem 3.1, it can be concluded that when a point (x, y) is Euclidean distance at least h_N away from the boundary curve Γ or the border of the grid cell Ω, then the estimated gradient direction $(\widehat{b}(x, y), \widehat{c}(x, y))$ converges to the true gradient direction of f at (x, y) almost surely and uniformly with respect to (x, y). If the point (x, y) is on the boundary curve Γ and the tangent direction of Γ at (x, y) is $(\cos(\theta_T(x, y)), \sin(\theta_T(x, y)))$ with $\theta_T(x, y) \in [0, \pi)$, then the estimated gradient direction has the property that

$$\begin{aligned} \widehat{b}(x, y) &= -\frac{C_\tau C^*}{K_{02}h_N} \sin(\theta_T(x, y)) + o(1/h_N), \text{ a.s.}, \\ \widehat{c}(x, y) &= \frac{C_\tau C^*}{K_{20}h_N} \cos(\theta_T(x, y)) + o(1/h_N), \text{ a.s.}, \end{aligned}$$

where C^* is a constant. Therefore $(\widehat{b}(x,y), \widehat{c}(x,y))$ is approximately perpendicular to the tangent direction of Γ at (x,y) in such a case. Based on these properties, we have the following result about the detected boundary curve $\widehat{\Gamma}^*$.

Theorem 3.2 *Under the conditions stated in Theorem 3.1, we have*

$$\|\widehat{r}^* - r\|_{[0,2\pi)} = O(h_N), \tag{18}$$

where $r(\theta)$ and $\widehat{r}^*(\theta)$ for $\theta \in [0, 2\pi)$ are the radius functions of Γ and $\widehat{\Gamma}^*$, respectively.

Theorem 3.2 says that $\widehat{\Gamma}^*$ converges to Γ almost surely and uniformly with respect to $\theta \in [0, 2\pi)$. From (14)–(18), when $h_N \sim n^{-1/3}\sqrt{\log(n)}$, this convergence can reach the rate of $O\left(n^{-1/3}\sqrt{\log(n)}\right)$. Remember that the sample size in the grid cell Ω is $N = n^2$. So the rate is $O\left(N^{-1/6}\sqrt{\log(N)}\right)$.

4 A Simulation Study

In this section, we present some simulation results regarding the numerical performance of the proposed boundary curve estimators, discussed in the previous sections. For simplicity, our simulation is for detecting the boundary curve(s) of a single grid cell of a spotted microarray image, which is appropriate for reasons explained in Section 2.1.

Let us assume that the design space of the grid cell is $[-1/2, 1/2] \times [-1/2, 1/2]$, i.e., the origin of the coordinate system is at the center of the grid cell. We first discuss the case that the foreground has a single boundary curve Γ. In such a case, we assume that the underlying true image intensity function is $f(x,y) = 4[1 - .5(x^2+y^2)/R^2]$ if $\frac{x^2}{a^2} + \frac{y^2}{b^2} \le 1$; and $f(x,y) = 1 - .5x^2 - .5y^2$ otherwise, where $R = \max(a,b)$, and a and b are two positive parameters. So, f has an elliptical boundary curve Γ centered at the origin $(0,0)$. When $a = b = .12$, for instance, Γ is a circle with radius .12, and the jump magnitude at Γ is a constant 1.0072. Note that the performance of the proposed procedure would not change much if the form of f changes in its continuity regions, mainly due to the fact that the proposed procedure is based on nonparametric local smoothing. The above quadratic function is chosen because it can well approximate true image intensity functions of some real microarray images, including the ones used in Section 5, based on our preliminary study. It is further assumed that random errors involved in observed image intensities are i.i.d. and normally distributed with mean zero, variances σ_1^2 in the foreground and σ_2^2 in the background. When $a = b = .12$, $n = 50$, $\sigma_1 = 1$, and $\sigma_2 = .5$, one realization of the observed image intensities is presented in Figure 2(a) with whiter pixels denoting larger intensity levels.

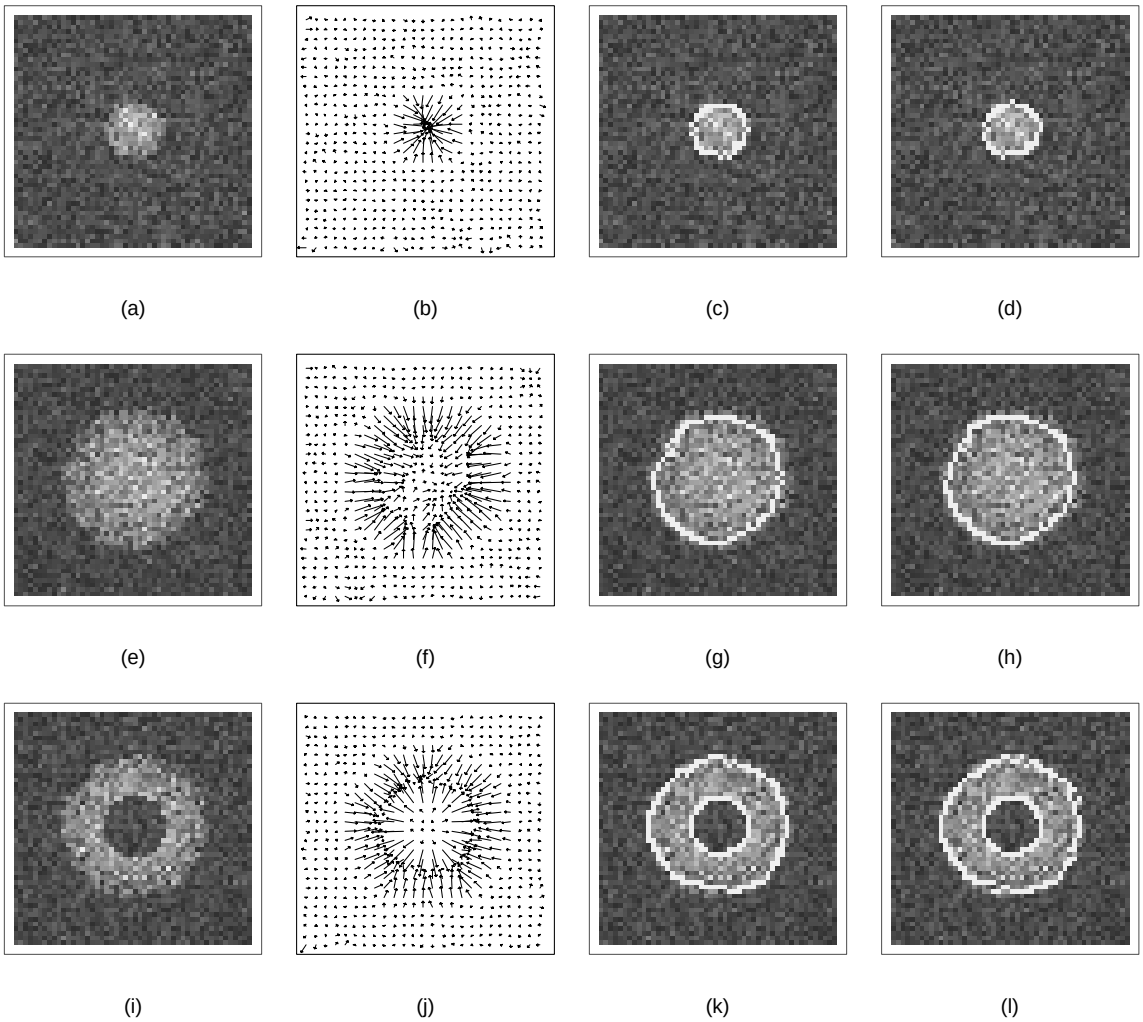


Figure 2: (a) An observed grid cell with a single boundary curve. (b) Estimated image gradient $G(x, y)$ from the data in plot (a). (c) Estimated boundary curve (white pixels) using criterion $M_N(r, \theta)$ from the data in plot (a). (d) Estimated boundary curve using criterion $M_N^*(r, \theta)$ from the data in plot (a). (e)–(h) Corresponding results with a larger boundary curve. (i) An observed "donut"-shaped spot. (j) Estimated image gradient $G(x, y)$ from the data in plot (i). (k) Estimated boundary curves using $M_N(r, \theta)$ from the data in plot (i). (l) Estimated boundary curves using $M_N^*(r, \theta)$ from the data in plot (i).

The boundary detection procedure (3)–(5) and its simplified version (11)–(12) are then applied to the data shown in Figure 2(a). In both procedures, the kernel function is chosen to be the modified, bivariate, Gaussian density function:

$$K(x,y) = \begin{cases} \frac{1}{2\pi - 3\pi e^{-.5}} \left[\exp\left(-\frac{x^2+y^2}{2} \right) - e^{-.5} \right], & \text{when } x^2+y^2 \le 1 \\ 0, & \text{otherwise,} \end{cases}$$

which has support $\{(x,y) : x^2+y^2 \le 1\}$ and is a continuous function in R^2. The bandwidths of the two procedures are both chosen to be $h_N = .07$, by the cross-validation procedure (13). For procedure (11)–(12), which is based on gradient estimation, the estimated image gradients are shown in Figure 2(b). It can be seen that (i) in continuity regions of the image intensity function, magnitudes of the estimated image gradients are small, (ii) they are relatively large around the true boundary curve Γ, and (iii) they indicate the orientation of Γ well. For both procedures, the boundary curve is detected at 40 equally spaced θ values in $[0, 2\pi)$. The detected boundary curve by procedure (3)–(5) is presented in Figure 2(c) by white pixels. The corresponding results of procedure (11)-(12) are shown in Figure 2(d). By comparing the results of the two procedures, it can be seen that the two sets of results are comparable, and the results of procedure (11)–(12) are slightly worse at several θ values. This example shows that the simplified procedure (11)–(12) does not lose much efficiency for boundary curve estimation, compared to procedure (3)–(5). Corresponding results in the case with a larger boundary curve specified by $a = b = .3$ are shown in plots (e)–(h).

Now, we discuss the case that the foreground of a grid cell has two boundary curves, or, the grid cell includes a "donut" spot. The underlying true image intensity function is assumed to be $f(x,y) = 4[1 - .5(x^2+y^2)/R^2]$ if $\frac{x^2}{a_1^2} + \frac{y^2}{b_1^2} > 1$ and $\frac{x^2}{a_2^2} + \frac{y^2}{b_2^2} \le 1$; and $f(x,y) = 1 - .5x^2 - .5y^2$ otherwise, where $R = \max(a_2, b_2)$, a_1, b_1, a_2 and b_2 are positive parameters, $a_1 < a_2$, and $b_1 < b_2$. When $a_1 = b_1 = .12$, $a_2 = b_2 = .3$, $n = 50$, $\sigma_1 = 1$, and $\sigma_2 = .5$, one realization of the observed image intensities is presented in Figure 2(i). We then use the criteria $M_N(r,\theta)$ and its simplified version $M_N^*(r,\theta)$ defined at (3) and (11), respectively, for estimating the two boundary curves, in the way as described by formulas (7) and (8). The bandwidth h_N is chosen to be .11 in both cases, by CV. The estimated image gradients and the estimated boundary curves are presented in Figures 2(j)–(l). It can be seen that the "donut" spot is segmented reasonably well.

In Section 2, we have explained in words why local constant kernel smoothing, instead of local linear kernel smoothing, is used in constructing our boundary detection criterion $M_N(r,\theta)$ (cf.,

equation (3)). Next, let us use the following numerical example to further demonstrate the benefits of using local constant kernel smoothing. Suppose that the foreground has one boundary curve as in Figure 2(a), $a = b = .12$, $\sigma_1 = 1$, $\sigma_2 = .5$, and 100 replicated simulations are performed. The boundary detection criterion $M_N(r, \theta)$ is constructed based on either local linear or local constant kernel smoothing. When $\theta = 0$, the 100 values of $M_N(r, 0)$ from the 100 replications are presented in Figure 3(a) for the local linear kernel smoothing case, and in Figure 3(b) for the local constant kernel smoothing case. In each plot, the upper, central and lower curves denote the 95^{th} percentile, the average and the 5^{th} percentile of the 100 replicated values of $M_N(r, 0)$, as functions of r. It can be seen that the criterion $M_N(r, \theta)$ based on local linear kernel smoothing is indeed not as sensitive to the boundary curve as the one based on local constant kernel smoothing in such a case.

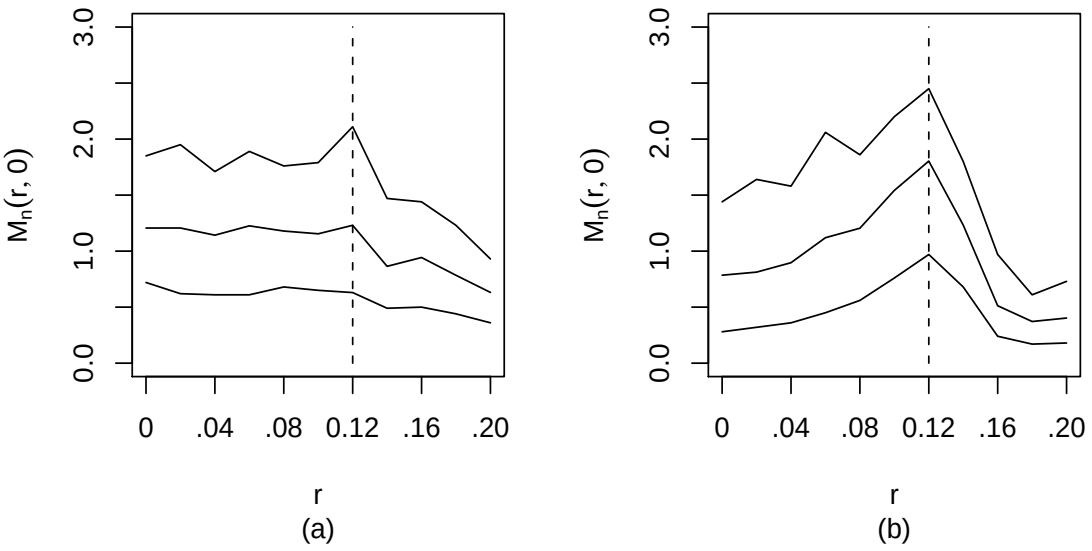


Figure 3: (a) Upper, central and lower curves denote the 95^{th} percentile, the average and the 5^{th} percentile of 100 replicated values of $M_N(r, 0)$, when it is constructed using the local linear kernel smoothing. (b) Corresponding results when $M_N(r, 0)$ is constructed using the local constant kernel smoothing.

Next, we consider the following seven cases, in each of which the true image intensity function is defined as in Figure 2. In the first four cases, the foreground is assumed to have a single boundary curve Γ as in Figure 2(a), and the parameters of Γ are (i) $a = b = .12$, (ii) $a = b = .3$, (iii) $a = .12$ and $b = .08$, and (iv) $a = .3$ and $b = .2$, respectively. The boundary curve is a circle in each of the first two cases, and the one in case (ii) is larger than the one in case (i), which is designed for investigating possible effects of the background size on numerical performance of the proposed procedures. Similarly, Γ is an ellipse in cases (iii) and (iv), and the one in case (iv) is larger than

the one in case (iii). In cases (v) and (vi), the grid cell is assumed to have a "donut" spot, as in Figure 2(i), and the "donut" is circular with parameters ($a_1 = b_1 = .12, a_2 = b_2 = .3$) in case (v), and elliptical with parameters ($a_1 = .12, b_1 = .08, a_2 = .3, b_2 = .2$) in case (vi). In all six cases described above, random errors are assumed to be Normally distributed with $\sigma_1 = 1$ and $\sigma_2 = 0.5$. To investigate possible effects of the random error distribution on performance of the proposed procedure, in case (vii), we assume that the true image intensity function is exactly the same as that in case (i), but the random errors are generated from the $\chi^2(2)$ distribution first, which is skewed to right, and then they are normalized to have mean 0 and variances 1 in the foreground and .5 in the background.

We apply procedure (3)–(5) and its simplified version (11)–(12) to cases (i)–(iv) and (vii), in which the spot has a single boundary curve, and procedure (7)–(8) and its simplified version to cases (v) and (vi), in which the grid cell includes a "donut" spot. In each of the seven cases, 100 replicated simulations are performed. In each simulation, the approximated bandwidths $\widehat{h}_N$ and $\widehat{h}^*_N$ based on criteria $M_N(r,\theta)$ and $M^*_N(r,\theta)$, respectively, are obtained by the CV procedure (13). To evaluate the performance of the CV bandwidths, we also compute the bandwidth minimizing

$$d_{\Omega^{(1)},\widehat{\Omega}^{(1)}}(h_N) = \frac{\left|\left(\Omega^{(1)} \backslash \widehat{\Omega}^{(1)}\right) \bigcup \left(\widehat{\Omega}^{(1)} \backslash \Omega^{(1)}\right)\right|}{\left|\Omega^{(1)}\right|},$$

where $|A|$ denotes the number of pixels in pointset A, $\Omega^{(1)}$ is the true foreground, and $\widehat{\Omega}^{(1)}$ is its estimator. This bandwidth is regarded as the true optimal bandwidth for boundary curve estimation, and is denoted by $h_{N,opt}$. The simulation results are summarized in Table 1. From the table, we can see that: (1) both $\widehat{h}_N$ and $\widehat{h}^*_N$ are close to the optimal bandwidth $h_{N,opt}$, (2) $d_{\Omega^{(1)},\widehat{\Omega}^{(1)}}(\widehat{h}_N)$ and $d_{\Omega^{(1)},\widehat{\Omega}^{(1)}}(\widehat{h}^*_N)$ are close to each other, and (3) they are both decent, compared to $d_{\Omega^{(1)},\widehat{\Omega}^{(1)}}(h_{N,opt})$.

Next, we compare the proposed image segmentation procedure based on M^*_N (denoted as "Kernel") with the following three existing image segmentation procedures (cf., Section 1 for a brief introduction about them): (i) SRG, (ii) image segmentation by thresholding the histogram of image intensities (denoted as "Hist"), and (iii) adaptive circle image segmentation (denoted as "Circle"). The seven cases are considered here, as in Table 1, and the true values of the averaged foreground image intensity (AFII) are 2.79, 2.94, 3.04, 3.21, 2.75, 3.05, and 2.79, respectively. In the proposed procedure, the modified bivariate Gaussian density function is used as the kernel function as before, the decision rule (9) is used for determining whether the grid cell has a "donut" spot,

Table 1: In each entry, the first line presents the averaged value of $h_{N,opt}$, $\widehat{h}_N$, or $\widehat{h}^*_N$, and its standard error (in parenthesis), obtained from 100 replications. The second line presents the averaged value of $d_{\Omega^{(1)},\widehat{\Omega}^{(1)}}(h_{N,opt})$, $d_{\Omega^{(1)},\widehat{\Omega}^{(1)}}(\widehat{h}_N)$, or $d_{\Omega^{(1)},\widehat{\Omega}^{(1)}}(\widehat{h}^*_N)$, and its standard error (in parenthesis).

Case	M_N with $h_{N,opt}$	M_N with $\widehat{h}_N$	M^*_N with $\widehat{h}^*_N$
(i)	.0756 (.0015)	.0734 (.0016)	.0722 (.0015)
	.214 (.006)	.222 (.007)	.222 (.007)
(ii)	0.1142 (.0013)	.1052 (.0014)	.1092 (.0013)
	.073 (.003)	.073 (.003)	.073 (.003)
(iii)	.0762 (.0014)	.0820 (.0020)	.0800 (.0020)
	.153 (.005)	.173 (.006)	.175 (.006)
(iv)	.1108 (.0015)	.0986 (.0015)	.0990 (.0017)
	.060 (.002)	.066 (.002)	.067 (.002)
(v)	.1070 (.0015)	.1102 (.0019)	.1090 (.0018)
	.142 (.003)	.145 (.003)	.146 (.003)
(vi)	.0954 (.0015)	.0946 (.0017)	.0908 (.0019)
	.152 (.002)	.157 (.002)	.159 (.002)
(vii)	.0760 (.0014)	.0730 (.0016)	.0724 (.0015)
	.214 (.007)	.218 (.007)	.218 (.006)

and the bandwidth is selected by the CV procedure (13). The noise variance σ_1^2 is estimated in the region within a circle of radius .08 centered at the origin, and the noise variance σ_2^2 is estimated in the region outside a circle of radius .35 centered at the origin. Both variance estimators are defined by the residual mean squares of the local linear kernel estimators of the image intensity surface in the two regions. In the SRG procedure, pixels located at the border of the grid cell are used as background seeds and pixels in a square of size $.1 \times .1$ centered at the origin are used as foreground seeds. In the Hist procedure, its circular mask is centered at the origin with radius .2 in cases (i), (iii), and (vii) when the grid cell has a single boundary curve and the foreground region is relatively small, and radius .4 in all the remaining cases, and its threshold value is determined by the Mann-Whitney statistic with significance level .05%, as used by Chen *et al.* (1997). In the adaptive circle image segmentation, the radius of the circle is searched by the approach used in the software package *Dapple* (2000), which generates a Laplacian image first from the original image, using a standard four-neighbor Laplacian mask (cf., Qiu 2005, Section 6.2), and then chooses the radius as the maximizer of function $\psi(r)$, defined as the average of all pixels in the Laplacian image whose Euclidean distances from the center are r. For each procedure, besides $d_{\Omega^{(1)},\widehat{\Omega}^{(1)}}$, we also compute

$$d^{(+)}_{\Omega^{(1)},\widehat{\Omega}^{(1)}} = \frac{|\widehat{\Omega}^{(1)} \backslash \Omega^{(1)}|}{\Omega^{(1)}}, \text{ and } d^{(-)}_{\Omega^{(1)},\widehat{\Omega}^{(1)}} = \frac{|\Omega^{(1)} \backslash \widehat{\Omega}^{(1)}|}{\Omega^{(1)}}$$

Table 2: This table presents the averaged values of $d_{\Omega^{(1)},\widehat{\Omega}^{(1)}}$, $d^{(+)}_{\Omega^{(1)},\widehat{\Omega}^{(1)}}$, $d^{(-)}_{\Omega^{(1)},\widehat{\Omega}^{(1)}}$, and $\widehat{\text{AFII}}$ from 100 replications. The numbers in parentheses are their standard errors.

Case	Method	$d_{\Omega^{(1)},\widehat{\Omega}^{(1)}}$	$d^{(+)}_{\Omega^{(1)},\widehat{\Omega}^{(1)}}$	$d^{(-)}_{\Omega^{(1)},\widehat{\Omega}^{(1)}}$	$\widehat{\text{AFII}}$
(i)	Kernel	.222 (.006)	.015 (.002)	.207 (.007)	3.01 (.01)
	SRG	.292 (.004)	.010 (.001)	.282 (.004)	3.21 (.01)
	Hist	.492 (.011)	.214 (.016)	.278 (.006)	2.90 (.03)
	Circle	.044 (.015)	0 (0)	.044 (.015)	2.85 (.02)
(ii)	Kernel	.073 (.003)	.010 (.001)	.063 (.003)	2.99 (.01)
	SRG	.226 (.002)	.005 (.001)	.221 (.002)	3.15 (.01)
	Hist	.251 (.005)	.111 (.008)	.140 (.006)	3.04 (.02)
	Circle	.004 (.004)	0 (0)	.004 (.004)	2.95 (.01)
(iii)	Kernel	.175 (.006)	.030 (.002)	.145 (.007)	3.16 (.01)
	SRG	.234 (.005)	.011 (.001)	.223 (.005)	3.35 (.02)
	Hist	.620 (.023)	.356 (.027)	.264 (.006)	2.91 (.04)
	Circle	.446 (.065)	.309 (.069)	.137 (.015)	2.88 (.05)
(iv)	Kernel	.067 (.002)	.032 (.001)	.035 (.002)	3.18 (.01)
	SRG	.169 (.002)	.007 (.001)	.162 (.002)	3.34 (.01)
	Hist	.343 (.014)	.239 (.018)	.104 (.005)	3.06 (.03)
	Circle	.310 (.010)	.155 (.017)	.155 (.011)	3.08 (.03)
(v)	Kernel	.146 (.003)	.078 (.001)	.069 (.003)	2.68 (.01)
	SRG	.699 (.008)	.224 (.001)	.475 (.008)	2.57 (.01)
	Hist	.347 (.008)	.180 (.013)	.167 (.007)	2.82 (.03)
	Circle	1.158 (.001)	.158 (.001)	1 (0)	1.00 (.01)
(vi)	Kernel	.159 (.002)	.120 (.001)	.039 (.002)	2.87 (.01)
	SRG	.576 (.006)	.242 (.001)	.334 (.006)	2.72 (.01)
	Hist	.469 (.021)	.344 (.025)	.125 (.005)	2.86 (.03)
	Circle	1.002 (.024)	.227 (.010)	.775 (.032)	1.65 (.07)
(vii)	Kernel	.218 (.006)	.018 (.002)	.200 (.007)	3.00 (.01)
	SRG	.284 (.004)	.022 (.002)	.262 (.004)	3.08 (.01)
	Hist	.484 (.012)	.146 (.015)	.338 (.017)	3.07 (.04)
	Circle	.081 (.021)	0 (0)	.081 (.021)	2.89 (.03)

to measure the amounts of false foreground pixels and false background pixels, respectively. With the estimated foreground, the estimated AFII, denoted as $\widehat{\text{AFII}}$, is also computed. The averaged values of $d_{\Omega^{(1)},\widehat{\Omega}^{(1)}}$, $d^{(+)}_{\Omega^{(1)},\widehat{\Omega}^{(1)}}$, $d^{(-)}_{\Omega^{(1)},\widehat{\Omega}^{(1)}}$, and $\widehat{\text{AFII}}$ from 100 replications are presented in Table 2, along with their standard errors; the decision rule (9) used by the proposed procedure makes 100% correct decisions in all cases.

From Table 2, it seems that the adaptive circle procedure performs the best among the four procedures in cases (i), (ii), and (vii), which is not a surprise because the grid cell has a single boundary curve and the true boundary curve is a circle in all these cases. In cases (iii)–(vi)

Table 3: In each entry, the first line presents the selected bandwidth and its standard error (in parenthesis) based on 100 replications, and the second line presents the averaged value of $d_{\Omega^{(1)},\widehat{\Omega}^{(1)}}(\widehat{h}_N)$ or $d_{\Omega^{(1)},\widehat{\Omega}^{(1)}}(\widehat{h}^*_N)$ and its standard error (in parenthesis).

CV Score	M_N with $\widehat{h}_N$	M^*_N with $\widehat{h}^*_N$
with variance	.0734 (.0016)	.0722 (.0015)
standardization	.222 (.007)	.222 (.007)
without variance	.0790 (.0020)	.0790 (.0020)
standardization	.237 (.007)	.234 (.007)

when the foreground has a single elliptical boundary curve, or, when the spot is "donut" shaped, it performs poorly because a single circle can not match the true boundary curve(s) well in such cases. Please note that, in case (v) when the spot is circularly "donut"-shaped, the estimated boundary curve by the adaptive circle procedure is completely inside the "hole" of the "donut", making its $d^{(-)}_{\Omega^{(1)},\widehat{\Omega}^{(1)}}$ value to be 1 and its $d_{\Omega^{(1)},\widehat{\Omega}^{(1)}}$ value larger than 1. Similar phenomenon can be seen in case (vi) when the spot is elliptically "donut"-shaped, although the estimated boundary curve is only partially inside the "hole" of the "donut" this time. Comparing the proposed procedure with the SRG procedure, the former outperforms the latter in all cases, with regard to $d_{\Omega^{(1)},\widehat{\Omega}^{(1)}}$, $d^{(-)}_{\Omega^{(1)},\widehat{\Omega}^{(1)}}$, and $\widehat{\text{AFII}}$. The former performs better than the latter with regard to $d^{(+)}_{\Omega^{(1)},\widehat{\Omega}^{(1)}}$ in cases (v)–(vii), when the spot is "donut"-shaped or when the noise distribution is skewed; in other cases, the former performs a little worse in terms of $d^{(+)}_{\Omega^{(1)},\widehat{\Omega}^{(1)}}$. Compared to the Hist procedure, the proposed procedure outperforms in all cases in terms of $d_{\Omega^{(1)},\widehat{\Omega}^{(1)}}$, $d^{(+)}_{\Omega^{(1)},\widehat{\Omega}^{(1)}}$, and $d^{(-)}_{\Omega^{(1)},\widehat{\Omega}^{(1)}}$.

The CV score defined in equation (13) equals the sum of two regional CV scores standardized by $\widehat{\sigma}_1^2$ and $\widehat{\sigma}_2^2$. Since noise variances in foreground and background could be very different in a microarray image, standardizations by $\widehat{\sigma}_1^2$ and $\widehat{\sigma}_2^2$ are important, which is demonstrated by the next example. Assume that the foreground has a single boundary curve with parameters $a = b = .12$, $\sigma_1 = 1$ and $\sigma_2 = .5$, as in Figure 2(a). Selected bandwidths by the CV procedure (13) with and without variance standardizations are presented in Table 3, along with the corresponding $d_{\Omega^{(1)},\widehat{\Omega}^{(1)}}(\widehat{h}_N)$ and $d_{\Omega^{(1)},\widehat{\Omega}^{(1)}}(\widehat{h}^*_N)$ values. It can be seen that the detected boundary curves $\widehat{\Gamma}$ and $\widehat{\Gamma}^*$ are closer to the true boundary curve Γ if the CV score with variance standardizations is used.

Performance of the related image segmentation procedures discussed above may depend on correct specification of the center and border of a given grid cell. To investigate this issue, we consider the following example, in which the spot has a single circular boundary curve, as in Figure

2(a), with $a = b = .12$, $\sigma_1 = 1$, and $\sigma_2 = .5$. We then let the center of the grid cell move to positions $(0, .02)$, $(0, .04)$, and $(.04, .04)$, respectively. In each case, the related image segmentation procedures are applied, as in Table 2. Table 4 presents their averaged values of $d_{\Omega^{(1)},\widehat{\Omega}^{(1)}}$ in the three cases, based on 100 replications. For convenience of comparison, their averaged values of $d_{\Omega^{(1)},\widehat{\Omega}^{(1)}}$ when the center does not move are also reported in the table. From the table, it can be seen that, besides the adaptive circle procedure, the remaining three procedures are quite robust to correct specification of the grid cell center. Since the robustness to correct specification of the grid cell border has a similar nature, it will not be discussed separately.

Table 4: Averaged values of $d_{\Omega^{(1)},\widehat{\Omega}^{(1)}}$ and their standard errors (in parentheses) when the grid cell center moves to various positions.

Center Position	Kernel	SRG	Hist	Circle
(0,0)	.222 (.007)	.292 (.004)	.492 (.011)	.044 (.015)
(0,.02)	.216 (.007)	.291 (.004)	.490 (.011)	.418 (.024)
(0,.04)	.221 (.007)	.287 (.004)	.489 (.011)	.512 (.021)
(.04,.04)	.229 (.007)	.280 (.004)	.490 (.012)	.659 (.020)

5 An Application

In this section, we apply the proposed procedure based on criterion $M_N^*(r, \theta)$ and the three existing procedures discussed in Section 4 to a real microarray image, which is from a study by van't Wout *et al* (2003) about the biochemical changes that occur during HIV-1 infection. In the study, expression levels of 4,608 cellular RNA transcripts were assessed in $CD4^+$-T-Cell lines, at different times after infection with HIV virus type 1 strain BRU (HIV-1_{BRU}), using DNA microarrays. For each of the infection conditions considered, duplicate slides were hybridized with probes generated from the same RNAs, and another duplicate slides were hybridized with probes in the same way except that fluorescent labels were reversed to control for dye-specific effects. Therefore, we have four replicated images which shared the same DNA samples. For ease of presentation, here we only consider 4 replicated subarrays, each consisting of $32 \times 12 = 384$ genes, corresponding to the first block of each image. The whole images can be downloaded from http://expression.microslu.washington.edu/expression/vantwoutjvi2002.html.

Before image segmentation, the automatic image addressing procedure described in Bergemann *et al.* (2004) is used for specifying the centers and borders of grid cells of each replicated image. By

this approach, peak positions of column/row totals of image intensities are used for specifying the grid cell centers, after the column/row totals are pre-smoothed by a smoothing operator. Similarly, the borders of grid cells are estimated by the valley positions of the smoothed column/row totals. The size of a typical grid cell of these images is about 30×30 pixels. Then, the four segmentation procedures are applied to each grid cell. In the proposed procedure, the foreground noise variance is estimated in the region within a circle of radius 4 pixels centered at the origin, and the background noise variance is estimated in the region outside a circle of radius 12 pixels centered at the origin. In the SRG procedure, pixels located at the border of the grid cell are used as background seeds and pixels in a square of size 5×5 pixels centered at the origin are used as foreground seeds. In the Hist procedure, the circular mask is centered at the origin and has radius of 12 pixels; its threshold is chosen in the same way as we did in Table 2. In the adaptive circle procedure, the circle radius is chosen by the approach used in the package *Dapple* (2000), described in Section 4.

Segmentation results of the four procedures for the first replicated image are shown in the four plots of Figure 4, respectively. Results for the other three replicated images are similar. In the plots, white pixels denote detected foreground pixels for the Hist procedure; they denote detected boundary curves for the other three procedures. It can be seen that the proposed procedure detects most boundary curves, including the ones of some "donut" spots (cf., e.g., the (24,11)-th spot), reasonably well. The SRG procedure can not handle "donut" spots well, and its detected boundary curves are quite noisy in some grid cells (cf., e.g., the (1,11)-th spot). The Hist procedure can handle some "donut" spots well, but its detected foreground pixels indeed do not form connected regions in some cases (cf., e.g., the (5,10)-th spot), due to the fact that it does not make use of any spatial information of the image. The adaptive circle procedure can not handle "donut" spots at all (cf., e.g., the (24,11)-th spot), and its results are affected much if the grid cell center is not specified well (cf., e.g., the (8,1)-th spot).

To further compare the four segmentation procedures in this example, we compute the averaged foreground (fg) intensity and background (bg) intensity for each spot, based on the segmentation results of each procedure. These averaged intensities are further averaged over four replications. Then, the ratios of averaged foreground intensities based on the proposed procedure to averaged foreground intensities based on the SRG procedure are plotted with corresponding ratios of averaged background intensities in Figure 5(a) by a scatter plot. Figures 5(b) and 5(c) present similar results for the pair of (proposed procedure, Hist procedure) and the pair of (proposed procedure, adaptive

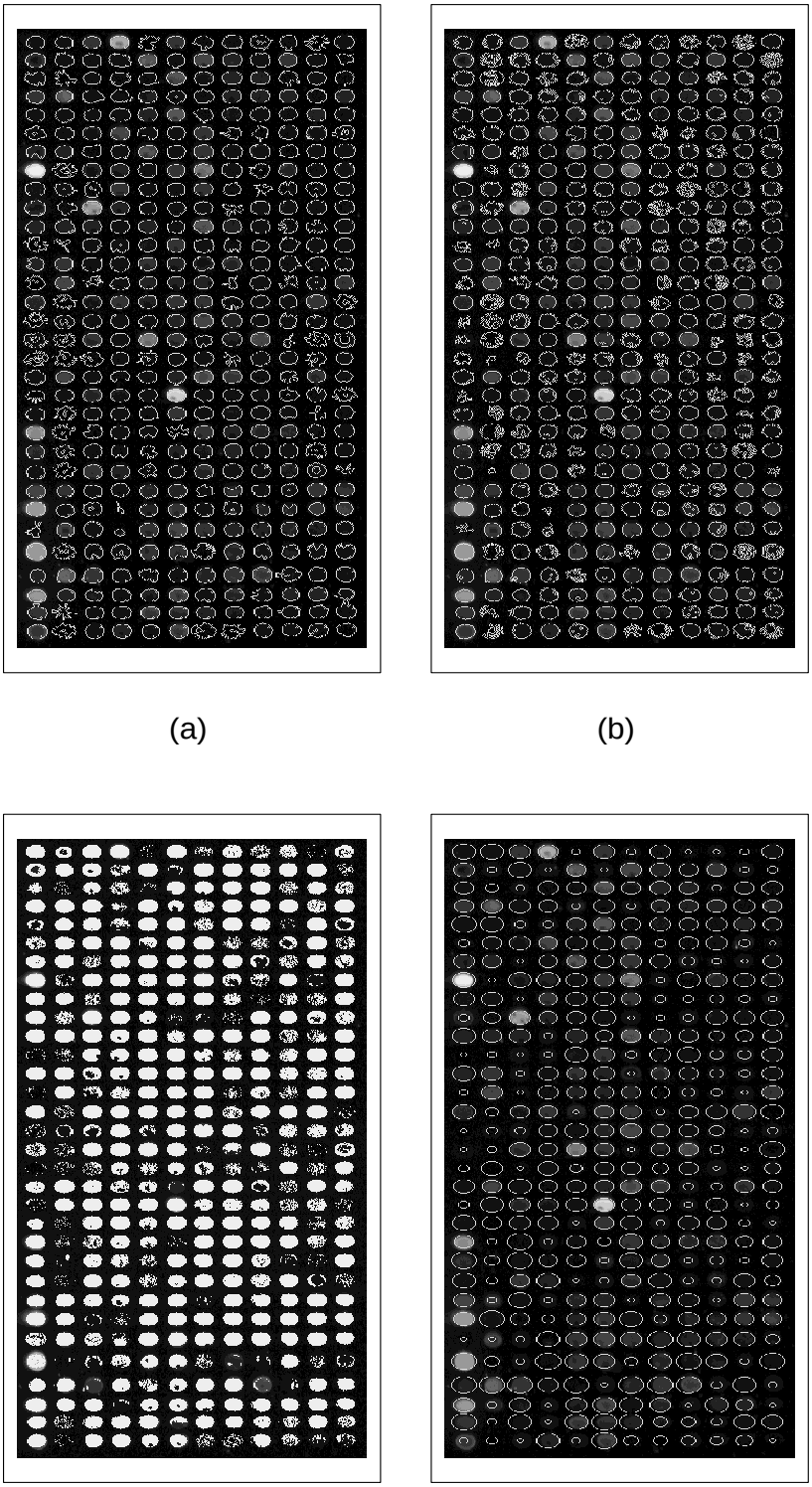


Figure 4: (a) Detected boundary curves by the proposed procedure. (b) Detected boundary curves by the SRG procedure. (c) Detected foreground pixels by the histogram thresholding procedure. (d) Detected boundary curves by the adaptive circle procedure.

circle procedure), respectively. It can be seen from the plots that the SRG procedure tends to give larger averaged background intensities, compared to the proposed procedure, due to the fact that it often misses some foreground pixels, especially when a given spot is "donut"-shaped (cf., e.g., the (8,2)-th spot in Figure 4(b)). Compared to the proposed procedure, the Hist procedure tends to provide larger foreground and larger background intensities, partly because it misses many foreground pixels when the corresponding image intensities are low (cf., e.g., the (1,5)-th spot in Figure 4(c)); finally, the adaptive circle procedure tends to provide smaller foreground intensities and larger background intensities, because it misses foreground pixels when a given spot is "donut"-shaped, and misclassifies many background pixels as foreground pixels when the grid cell center is not specified properly (cf., e.g., the (8,2)-th spot in Figure 4(d)).

Then, for each replicated image, we compute the gene expression level, defined by the $\log_2$ ratio of the averaged foreground intensity of the Cy5 image to the averaged foreground intensity of the Cy3 image, for each spot, based on segmentation results of each of the four procedures. For each segmentation procedure, the standard deviation (SD) and mean of the four computed gene expression levels, corresponding to four replicates, are computed, for each spot. According to Bergemann *et al.* (2004), a good segmentation procedure is expected to produce reliable gene expression data, and thus, the SDs of the gene expression levels computed from replicated images based on its segmentation should be small, without reducing the corresponding means. In Figures 5(d)–(f), we present the SDs of the gene expression levels in three scatter plots, corresponding to three pairs of segmentation procedures considered above. It can be seen from the plots that the proposed procedure seems to have smaller SDs, compared to the SRG and adaptive circle procedures; it has smaller SDs compared to the Hist procedure as well if several points in the upper-right part of Figure 5(e), which correspond to some spots with weak signals, can be down weighted. As a reference, the mean results are presented in Figures 5(g)–(i), from which it can be seen that the mean gene expression data based on the proposed procedure are comparable to those based on the other three procedures. By the way, the three spots shown in the lower-left parts of Figures 5(g)–(i) are the ones included intentionally by the researchers for varification purposes. They correspond to three highly expressed genes; thus, they are far away from the other spots in the plots. By one referee's suggestion, for each segmentation procedure, the within-spot SD computed from four replicates of each spot is then divided by the between-spot SD, defined as the SD of the mean gene expression levels (each mean is computed from four replicates of each spot) of all

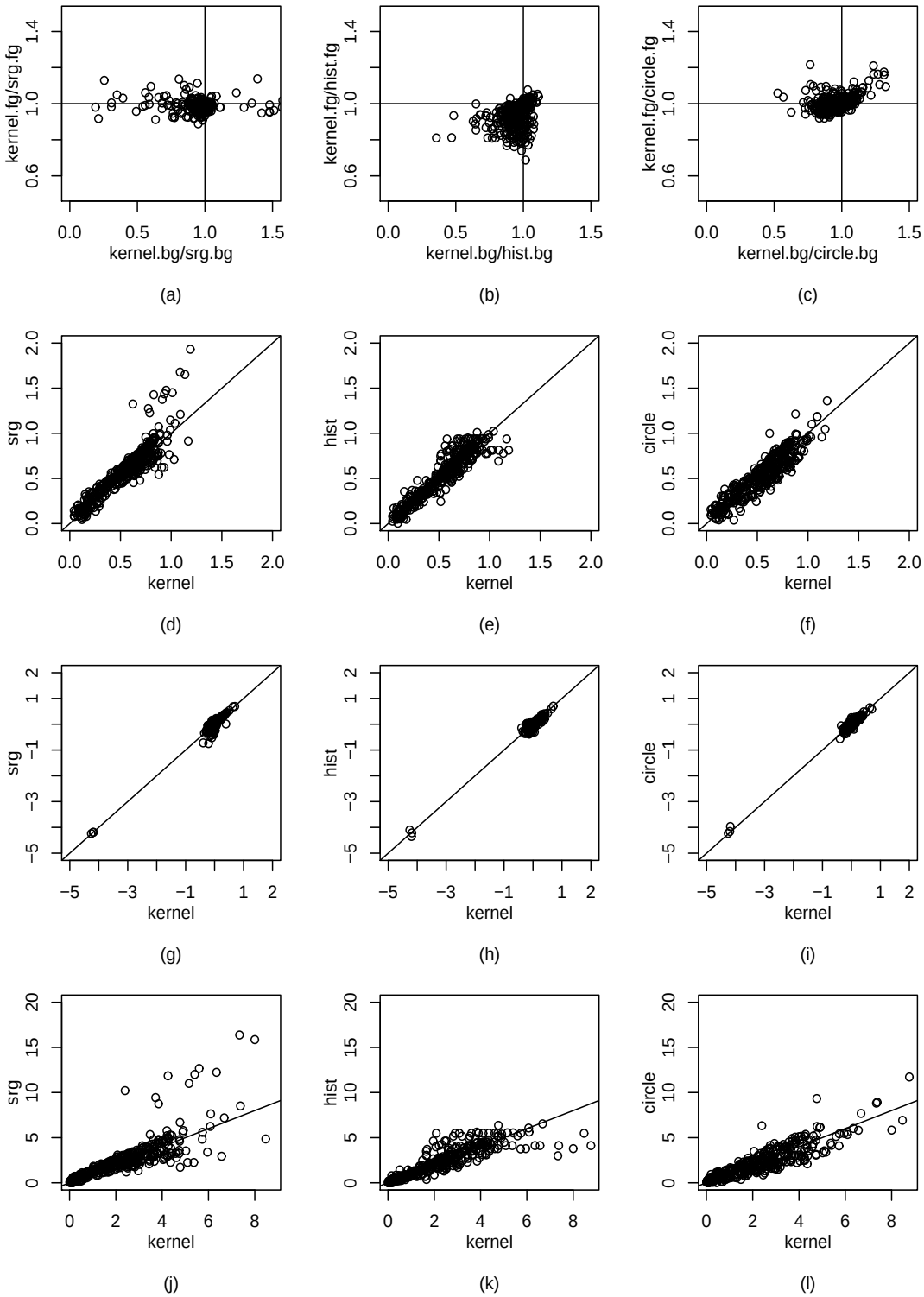


Figure 5: (a)–(c): Scatter plots of ratios of averaged foreground intensities based on the proposed procedure to averaged foreground intensities based on one of the three existing procedures versus corresponding ratios of averaged background intensities. (d)–(f): Within-spot SDs of the gene expression levels computed from four replicates based on the proposed procedure versus corresponding results based on one of the three existing procedures. (g)-(i): Corresponding spotwise means. (j)–(l): Ratios (within-spot SD)/(between-spot SD) based on the proposed procedure versus corresponding ratios based on one of the three existing procedures.

the spots. In Figures 5(j)–(l), we present the ratios (within-spot SD)/(between-spot SD) in three scatter plots, corresponding to three pairs of segmentation procedures, as in Figures 5(d)–(f). It can be seen that these plots have similar patterns to those in Figures 5(d)–(f). It should be noticed that both within-spot SDs and ratios of (within-spot SD)/(between-spot SD) are mainly for measuring replicability of a segmentation method. They are not good for measuring performance of image segmentation. For instance, let us consider some obviously bad segmentations based on the fixed circle segmentation procedure (cf., Section 1 for introduction), with radius of circles chosen to be a too small number (e.g., 2-pixels), or, a too large number (e.g., 14-pixels). It is easy to check that such segmentations would have much better replicability than the segmentation methods commonly used in practice (e.g., methods "srg", "hist", and "circle").

6 Concluding Remarks

We have presented an image segmentation procedure for analyzing microarray images. This procedure is based on local polynomial kernel smoothing. For a given squared unit of a microarray image, its center and border are determined in the image addressing stage. In the case the foreground has a single boundary curve Γ, which is a continuous closed curve, it can be expressed by (1) in polar coordinate system. Estimation of Γ is equivalent to estimation of its radius function $r(\theta)$ for any $\theta \in [0, 2\pi)$. To estimate $r(\theta)$ at a given $\theta \in [0, 2\pi)$, we consider a half-line starting from the center of the squared unit and forming an angle θ with the positive x-axis. At each point $(r\cos(\theta), r\sin(\theta))$ on this half-line, we search for a direction along which the absolute difference between two one-sided local constant kernel estimators reaches the maximum. This maximal absolute difference $M_N(r, \theta)$ is then used as a criterion for boundary curve estimation, and the maximizer of $M_N(r, \theta)$ with respect to r is used as an estimator of $r(\theta)$. A corresponding procedure is suggested for handling "donut" spots. To simplify computation, a modified version is suggested, in which the "direction search" step mentioned above is replaced by gradient estimation. We also propose a cross-validation procedure for choosing bandwidths used in our image segmentation procedures. Both theoretical arguments and numerical examples show that the proposed image segmentation procedures work well in applications. It should be pointed out that the proposed procedures still have much room for improvement. For instance, one referee mentioned that the procedures might be improved by searching for $\widehat{r}(\theta)$ in equation (5) at several different θ values simultaneously.

Acknowledgments: The authors thank the editor, the associate editor, and four reviewers for many constructive comments and suggestions which greatly improved the quality of the paper. Peihua Qiu's research was partially supported by an National Security Agency grant and an National Science Foundation grant.

References

Adams, R., and Bischof, L. (1994), "Seeded region growing," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16, 641–647.

Angulo, J., and Serra, J. (2003), "Automatic analysis of DNA microarray images using mathematical morphology," *Bioinformatics*, 19, 553–562.

Bergemann, T.L., Laws, R.J., Quiaoit, F., and Zhao, L.P. (2004), "A statistically driven approach for image segmentation and signal extraction in cDNA microarrays," *Journal of Computational Biology*, 11, 695–713.

Bozinov, D., and Rahnenführer, J. (2002), "Unsupervised technique for robust target separation and analysis of DNA microarray spots through adaptive pixel clustering," *Bioinformatics*, 18, 747–756.

Chen, Y., Dougherty, E.R., Bitter, M.L. (1997), "Ratio-based decisions and the quantitative analysis of cDNA microarray images," *Journal of Biomedical Optics*, 2, 364–374.

Chu, S., DeRisi, J., Eisen, M., Mulholland, J., Botstein, D., Brown, P.O., and Herskowitz, I. (1998), "The transcriptional program of sporulation in budding yeast," *Science*, 282, 699–705.

Dapple (2000), *Dapple: Image Analysis Software for DNA Microarrays,* http://www.cs.wustl.edu/ jbuhler/research/dapple/.

Eisen, M.B. (1999), *ScanAlyze,* http://rana.Stanford.EDU/software/.

Fan, J., and Gijbels, I. (1996), *Local Polynomial Modelling and Its Applications,* London: Chapman & Hall.

Ferea, T.L., Botstein, D., Brown, P.O., and Rosenzweig, R.F. (1999), "Systematic changes in gene expression patterns following adaptive evolution in yeast," *Proceedings of the national Academy of Science*, 96, 9721–9726.

GenePix (1999), *GenePix 4000A Users' Guide,* Axon Instruments, Inc.

Gijbels, I., Goderniaux, A.C. (2004), "Bandwidth selection for changepoint estimation in nonparametric regression," *Technometrics*, 46, 76–86.

Gijbels, I., Hall, P., and Kneip, A. (1999), "On the estimation of jump points in smooth curves," *Annals of the Institute of Statistical Mathematics*, 51, 231–251.

Glasbey, C.A., and Ghazal, P. (2003), "Combinatorial image analysis of DNA microarray features," *Bioinformatics*, 19, 194–203.

Hall, P., and Rau, C. (2000), "Tracking a smooth fault line in a response surface", *The Annals of Statistics*, 28, 713–733.

Hall, P., and Rau, C. (2002), "Likelihood-based confidence bands for fault lines in response surfaces," *Probability Theory and Related Fields*, 124, 26–49.

Hall, P., Peng, L., and Rau, C. (2001), "Local likelihood tracking of fault lines and boundaries," *Journal of the Royal Statistical Society - B*, 63, 569–582.

Hastie, T., and Loader, C. (1993), "Local regression: automatic kernel carpentry (with discussion)," *Statistical Science*, 8, 120–143.

Lashkari, D.A., DeRisi, J.L., McCusker, J.H., Namath, A.F., Gentile, C., Hwang, S.Y., Brown, P.O., and davis, R.W. (1997), "Yeast microarrays for genome wide parallel genetic and gene expression analysis," *Proceedings of the national Academy of Science*, 94, 13057–13062.

Qiu, P. (1997), "Nonparametric estimation of jump surface," *Sankhya (Series A)*, 59, 268–294.

Qiu, P. (2002), "A nonparametric procedure to detect jumps in regression surfaces," *Journal of Computational and Graphical Statistics*, 11, 799–822.

Qiu, P. (2005), *Image Processing and Jump Regression Analysis,* New York: John Wiley & Sons.

Qiu, P., and Bhandarkar, S.M. (1996), "An edge detection technique using local smoothing and statistical hypothesis testing," *Pattern Recognition Letters*, 17, 849–872.

Qiu, P., and Sun, J. (2006), "Local smoothing image segmentation for spotted microarray images," *Technical Report #650*, School of Statistics, University of Minnesota, Minneapolis, MN 55455.

Qiu, P., and Yandell, B. (1997), "Jump detection in regression surfaces," *Journal of Computational and Graphical Statistics*, 6, 332–354.

Steinfath, M., Wruck, W., Seidel, H., Lehrach, H., Radelof, U., and O'Brien, J. (2001), "Automated image analysis for array hybridization experiments," *Bioinformatics*, 17, 634–641.

van't Wout, A.B., Lehrman, G.K., Mikheeva, S.A., O'Keeffe, G.C., Katze, M.G., Bumgarner, R.E., Geiss, G.K., and Mullins, J.I. (2003), "Cellular gene expression upon human immunodeficiency virus type 1 infection of CD4(+)-T-cell lines," *Journal of Virology*, 77, 1392–1402.

Yang, Y.H., Buckley, M.J., Dudoit, S., and Speed, T. (2002), "Comparison of Methods for Image Analysis on cDNA Microarray Data," *Journal of Computational and Graphical Statistics*, 11, 108–136.

Appendix

A Proof Of Theorem 3.1

For any point $(x, y) \in \Omega_{h_N}$, it can be checked that

$$\begin{aligned} E(\widehat{b}(x,y)) &= f'_x(x,y) + \frac{1}{\Delta}\sum_{i=1}^{n}\sum_{j=1}^{n}\{\gamma_2 + \gamma_4(x_i - x) + \gamma_5(y_j - y)\}\{\frac{1}{2}[f''_x(x,y)(x_i - x)^2 + \\ & 2f''_{xy}(x,y)(x_i - x)(y_j - y) + f''_y(x,y)(y_j - y)^2]\}K\left(\frac{x_i - x}{h_N}, \frac{y_j - y}{h_N}\right) \\ & + o(h_N^2), \end{aligned} \tag{19}$$

where $N = n^2$, and

$$\Delta = \begin{vmatrix} \widetilde{K}_{00} & \widetilde{K}_{10} & \widetilde{K}_{01} \\ \widetilde{K}_{10} & \widetilde{K}_{20} & \widetilde{K}_{11} \\ \widetilde{K}_{01} & \widetilde{K}_{11} & \widetilde{K}_{02} \end{vmatrix},$$

$$\gamma_2 = \widetilde{K}_{01}\widetilde{K}_{11} - \widetilde{K}_{10}\widetilde{K}_{02},\ \gamma_4 = \widetilde{K}_{00}\widetilde{K}_{02} - \widetilde{K}_{01}\widetilde{K}_{01},\ \gamma_5 = \widetilde{K}_{01}\widetilde{K}_{10} - \widetilde{K}_{00}\widetilde{K}_{11},$$

$$\widetilde{K}_{s_1 s_2} = \sum_{i=1}^{n}\sum_{j=1}^{n}(x_i - x)^{s_1}(y_j - y)^{s_2}K\left(\frac{x_i - x}{h_N}, \frac{y_j - y}{h_N}\right),\ \text{for } s_1, s_2 = 0, 1, 2.$$

Since K is radially symmetric, we have $\widetilde{K}_{s_1 s_2}/(N h_N^{s_1+s_2+2}) = K_{s_1 s_2} + o(1)$ for $s_1, s_2 = 0, 1, 2$, $K_{00} = 1, K_{01} = K_{10} = K_{11} = 0$, and $K_{02} = K_{20} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} u^2 K(u, v)\, dudv$. Therefore,

$$\frac{\gamma_2}{N^2 h_N^7} = o(1);\ \frac{\gamma_4}{N^2 h_N^6} = K_{02} + o(1);\ \frac{\gamma_5}{N^2 h_N^6} = o(1);\ \frac{\Delta}{N^3 h_N^{10}} = (K_{02})^2 + o(1). \tag{20}$$

After combining (19) and (20), we have

$$\begin{aligned} E(\widehat{b}(x,y)) &= f'_x(x,y) + \frac{h_N}{K_{20}} \frac{1}{Nh_N^5} \sum_{i=1}^{n}\sum_{j=1}^{n} \{\frac{1}{2}[f''_x(x,y)(x_i-x)^3 + 2f''_{xy}(x,y)(x_i-x)^2(y_j-y) + \\ & \quad f''_y(x,y)(x_i-x)(y_j-y)^2]\} K\left(\frac{x_i-x}{h_N}, \frac{y_j-y}{h_N}\right) + o(h_N) \\ &= f'_x(x,y) + o(h_N). \end{aligned}$$

It can be checked that this expression is uniformly true for $(x,y) \in \Omega_{h_N}$. So,

$$\|E(\widehat{b}) - f'_x\|_{\Omega_{h_N}} = o(h_N). \tag{21}$$

On the other hand, for any $(x,y) \in \Omega_{h_N}$,

$$\begin{aligned} \widehat{b}(x,y) - E(\widehat{b}(x,y)) &= \frac{1}{\Delta} \sum_{i=1}^{n}\sum_{j=1}^{n} \{\gamma_2 + \gamma_4(x_i-x) + \gamma_5(y_j-y)\} \epsilon_{ij} K\left(\frac{x_i-x}{h_N}, \frac{y_j-y}{h_N}\right) \\ &= \frac{1}{Nh_N^3 K_{20}} \sum_{i=1}^{n}\sum_{j=1}^{n} \left(\frac{x_i-x}{h_N} + o(1)\right) \epsilon_{ij} K\left(\frac{x_i-x}{h_N}, \frac{y_j-y}{h_N}\right), \end{aligned} \tag{22}$$

where ϵ_{ij} is the random error involved in Z_{ij} with mean 0 and variance σ^2, for $i,j = 1,2,\ldots,n$. Let

$$\widetilde{\epsilon}_{ij} = \epsilon_{ij} I(|\epsilon_{ij}| \le t_{(i-1)n+j}),$$

where $t_m = \sqrt{m \log(m)(\log\log(m))^{(1+\delta)}}$, for $m \ge 3$, $t_1 = t_2 = t_3$, and $\delta \in (0,1)$ is any constant. Then we define

$$\begin{aligned} g_N(x,y) &= \frac{1}{Nh_N^2} \sum_{i=1}^{n}\sum_{j=1}^{n} \left(\frac{x_i-x}{h_N}\right) \epsilon_{ij} K\left(\frac{x_i-x}{h_N}, \frac{y_j-y}{h_N}\right) \\ \widetilde{g}_N(x,y) &= \frac{1}{Nh_N^2} \sum_{i=1}^{n}\sum_{j=1}^{n} \left(\frac{x_i-x}{h_N}\right) \widetilde{\epsilon}_{ij} K\left(\frac{x_i-x}{h_N}, \frac{y_j-y}{h_N}\right). \end{aligned}$$

It's obvious that

$$\|g_N\|_{\Omega_{h_N}} \le \|g_N - \widetilde{g}_N\|_{\Omega_{h_N}} + \|\widetilde{g}_N - E(\widetilde{g}_N)\|_{\Omega_{h_N}} + \|E(\widetilde{g}_N)\|_{\Omega_{h_N}}. \tag{23}$$

By some similar arguments to those in the proof of Theorem 3.1 of Qiu (2002), we have

$$\|g_N - \widetilde{g}_N\|_{\Omega_{h_N}} = o\left(\frac{\sqrt{\log(N)}}{N^{1/2}h_N}\right), \text{ a.s.} \tag{24}$$

$$\|\widetilde{g}_N - E(\widetilde{g}_N)\|_{\Omega_{h_N}} = O\left(\frac{\sqrt{\log(N)}}{N^{1/2}h_N}\right), \text{ a.s.} \tag{25}$$

$$\|E(\widetilde{g}_N)\|_{\Omega_{h_N}} = o\left(\frac{\sqrt{\log(N)}}{N^{1/2}h_N}\right). \tag{26}$$

By (22)-(26), we have

$$\|\widehat{b} - E(\widehat{b})\|_{\Omega_{h_N}} = O\left(\frac{\sqrt{\log(N)}}{N^{1/2}h_N^2}\right), \text{ a.s.} \tag{27}$$

Then, by equations (21) and (27), we have equation (14) in the theorem. Equation (15) can be proved similarly.

For a point $(x, y) \in \Omega$, if (x_τ, y_τ) is the closest point in Γ from (x, y), and their Euclidean distance is τh_N where $0 \le \tau < 1$ a constant, then by equations (19) and (20), we have

$$\begin{aligned} E(\widehat{b}(x,y)) &= \frac{1}{Nh_N^4 K_{02}} \left(\sum_{(x_i,y_j)\in Q_N^{(1)}(x,y)} + \sum_{(x_i,y_j)\in Q_N^{(2)}(x,y)} \right) (x_i - x) f(x_i, y_j) K\left(\frac{x_i - x}{h_N}, \frac{y_j - y}{h_N}\right) \\ &\quad + o(\frac{1}{h_N}) \\ &= \frac{1}{K_{02}h_N} f_-(x_\tau, y_\tau) \int\int_{Q^{(1)}} uK(u,v)\, dudv \\ &\quad + \frac{1}{K_{02}h_N} f_+(x_\tau, y_\tau) \int\int_{Q^{(2)}} uK(u,v)\, dudv + o(\frac{1}{h_N}) \\ &= \frac{C_\tau}{K_{02}h_N} \int\int_{Q^{(2)}} uK(u,v)\, dudv + o(\frac{1}{h_N}). \end{aligned} \tag{28}$$

Similar to equation (27), it can be shown that $\widehat{b}(x,y) - E(\widehat{b}(x,y)) = O\left(\frac{\sqrt{\log(N)}}{N^{1/2}h_N^2}\right)$, a.s. Equation (16) can be obtained by combining this result with equation (28). Equation (17) can be proved similarly.

B Proof Of Theorem 3.2

For a point $(x, y) \in \Omega_{h_N}$ and for any $\tilde{\theta} \in [0, \pi)$, we have

$$\begin{aligned} \widehat{a}^{(\ell)}(x,y,\tilde{\theta}) &= \frac{\sum_{(x_i,y_j)\in O_N^{(\ell)}(x,y,\tilde{\theta})} Z_{ij} K\left(\frac{x_i-x}{h_N}, \frac{y_j-y}{h_N}\right)}{\sum_{(x_i,y_j)\in O_N^{(\ell)}(x,y,\tilde{\theta})} K\left(\frac{x_i-x}{h_N}, \frac{y_j-y}{h_N}\right)}, \\ \widehat{a}^{(\ell)}(x,y,\tilde{\theta}) - f(x,y) &= \frac{\sum_{(x_i,y_j)\in O_N^{(\ell)}(x,y,\tilde{\theta})} (Z_{ij} - f(x,y)) K\left(\frac{x_i-x}{h_N}, \frac{y_j-y}{h_N}\right)}{\sum_{(x_i,y_j)\in O_N^{(\ell)}(x,y,\tilde{\theta})} K\left(\frac{x_i-x}{h_N}, \frac{y_j-y}{h_N}\right)}, \text{ for } \ell = 1, 2. \end{aligned}$$

By similar arguments to those in the proof of Theorem 3.1, we have

$$\begin{aligned} \frac{1}{Nh_N^2} \sum_{(x_i,y_j)\in O_N^{(\ell)}(x,y,\tilde{\theta})} (Z_{ij} - f(x,y)) K\left(\frac{x_i-x}{h_N}, \frac{y_j-y}{h_N}\right) &= o(h_N) + O\left(\frac{\sqrt{\log(N)}}{N^{1/2}h_N}\right), \text{ a.s.} \\ \frac{1}{Nh_N^2} \sum_{(x_i,y_j)\in O_N^{(\ell)}(x,y,\tilde{\theta})} K\left(\frac{x_i-x}{h_N}, \frac{y_j-y}{h_N}\right) &= \frac{1}{2} + o(1), \end{aligned}$$

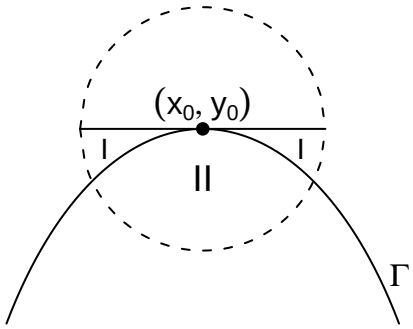


Figure 6: Dashed circle denotes the circular neighborhood $O_N(x_0, y_0, \arctan(\phi\prime(x_0)))$.

and these results are uniformly true for $(x, y) \in \Omega_{h_N}$ and $\tilde{\theta} \in [0, \pi)$. So,

$$\|\widehat{a}^{(\ell)} - f\|_{\Omega_{h_N} \times [0,\pi)} = o(h_N) + O\left(\frac{\sqrt{\log(N)}}{N^{1/2} h_N}\right), \text{ a.s., for } \ell = 1, 2,$$

and

$$\|\widehat{a}^{(1)} - \widehat{a}^{(2)}\|_{\Omega_{h_N} \times [0,\pi)} = o(h_N) + O\left(\frac{\sqrt{\log(N)}}{N^{1/2} h_N}\right), \text{ a.s.}$$

Hence, for any $\theta_0 \in [0, 2\pi)$, we have

$$\lim_{N\to\infty} \max_{r' \in [r(\theta_0)+h_N, R_{\theta_0}-h_N]} M_N^\star(r', \theta_0) = 0 \tag{29}$$

$$\lim_{N\to\infty} \max_{r' \in [0, r(\theta_0)-h_N)} M_N^\star(r', \theta_0) = 0. \tag{30}$$

Without loss of generality, let us assume that Γ has the expression $y = \phi(x)$ around the point $(x_0, y_0) = (r(\theta_0)\cos(\theta_0), r(\theta_0)\sin(\theta_0))$. The circular neighborhood $O_N(x_0, y_0, \arctan(\phi'(x_0)))$ of the point (x_0, y_0) is displayed in Figure 6, which consists of two parts separated by the tangent line of Γ at (x_0, y_0). One part is outside Γ, and the other one is divided by Γ into two subparts, denoted by I and II in the plot, with subpart I outside Γ and subpart II inside Γ. Suppose that $\widehat{a}^{(1)}(x_0, y_0, \arctan(\phi'(x_0)))$ is the one-sided local constant kernel estimator constructed from the part of $O_N(x_0, y_0, \arctan(\phi'(x_0)))$ with two subparts I and II. Then

$$\begin{aligned}
\widehat{a}^{(1)}\left(x_0, y_0, \arctan(\phi'(x_0))\right) &= \frac{2}{N h_N^2} \sum{}' Z_{ij} K\left(\frac{x_i - x_0}{h_N}, \frac{y_j - y_0}{h_N}\right) \\
&+ \frac{2}{N h_N^2} \sum{}'' (Z_{ij} - C_{ij}) K\left(\frac{x_i - x_0}{h_N}, \frac{y_j - y_0}{h_N}\right) \\
&+ \frac{2}{N h_N^2} \sum{}'' C_{ij} K\left(\frac{x_i - x_0}{h_N}, \frac{y_j - y_0}{h_N}\right) + o(1),
\end{aligned} \tag{31}$$

where $C_{ij} = f(x_i, y_j) - f_+(x_0, y_0)$, $f_+(x_0, y_0)$ denotes the limit of f at (x_0, y_0) from I, $\sum'$ denotes summation of the terms with design points inside I, and $\sum''$ denotes summation of the terms with design points inside II.

Since the radius function $r(\theta)$ is assumed to have continuous second-order derivatives on $[0, 2\pi)$, $\phi(x)$ also has second-order derivatives around x_0. So,

$$\phi(x) = \phi(x_0) + \phi\prime(x_0)(x - x_0) + \frac{\phi\prime\prime(x_0)}{2!}(x - x_0)^2 + o((x - x_0)^2)$$

from which we can see that the area of part I equals $\int_{-h_N/2}^{h_N/2} \left(\frac{\phi\prime\prime(x_0)}{2!}(x - x_0)^2 + o((x - x_0)^2) \right) dx = O(h_N^3)$. So, the third term of (31) converges to $f_-(x_0, y_0) - f_+(x_0, y_0)$, where $f_-(x_0, y_0)$ is the limit of f at (x_0, y_0) from II.

From Theorem 3.1, the estimated tangent direction $\widehat{\theta}_T(x_0, y_0) = (\widehat{b}(x_0, y_0), \widehat{c}(x_0, y_0))$ converges to $\arctan(\phi'(x_0))$ almost surely. So,

$$\lim_{N\to\infty} \widehat{a}^{(1)}(x_0, y_0, \hat{\theta}_T(x_0, y_0)) = f_-(x_0, y_0), \text{ a.s.}$$

Similarly,

$$\lim_{N\to\infty} \widehat{a}^{(2)}(x_0, y_0, \hat{\theta}_T(x_0, y_0)) = f_+(x_0, y_0), \text{ a.s.}$$

So,

$$\lim_{N\to\infty} M_N^\star(r(\theta_0), \theta_0) = f_-(x_0, y_0) - f_+(x_0, y_0) > 0. \tag{32}$$

By (29), (30), and (32), we have

$$r(\theta_0) - h_N \le \widehat{r}^\star(\theta_0) \le r(\theta_0) + h_N,$$

and this conclusion is uniformly true for $\theta_0 \in [0, 2\pi)$. So equation (18) is proved.