Nonparametric Inference for Local Extrema with Application to Oligonucleotide Microarray Data in Yeast Genome

Peter X.-K. Song,1,* Xin Gao,2 Rui Liu,3 and Wen Le4

1Department of Statistics and Actuarial Science, University of Waterloo, 200 University Avenue W., Waterloo, Ontario N2L 3G1, Canada
2Department of Mathematics and Statistics, York University, Toronto, Ontario M3J 1P3, Canada
3Department of Biostatistics, Columbia University, New York, New York 10032, U.S.A.
4Department of Public Health Sciences, University of Toronto, Toronto, Ontario M5S 1A1, Canada
*email: song@math.uwaterloo.ca

SUMMARY. Identifying local extrema of expression profiles is one primary objective in some cDNA microarray experiments. To study the replication dynamics of the yeast genome, for example, local peaks of hybridization intensity profiles correspond to putative replication origins. We propose a nonparametric kernel smoothing (NKS) technique to detect local hybridization intensity extrema across chromosomes. The novelty of our approach is that we base our inference procedures on equilibrium points, namely those locations at which the first derivative of the intensity curve is zero. The proposed smoothing technique provides both point and interval estimation for the location of local extrema. Also, this technique can be used to test for the hypothesis of either one or multiple suspected locations being the true equilibrium points. We illustrate the proposed method on a microarray data set from an experiment designed to study the replication origins in the yeast genome, in that the locations of autonomous replication sequence (ARS) elements are identified through the equilibrium points of the smoothed intensity profile curve. Our method found a few ARS elements that were not detected by the current smoothing methods such as the Fourier convolution smoothing.

KEY WORDS: Equilibrium point; Fourier convolution smoothing; Kernel smoothing; Local fitting; Replication origin; Yeast genome.

1. Introduction

Local extrema of hybridization intensity profiles are often of interest to biologists, because they pertain closely to some important biological features. For example, in a study of the replication dynamics of the yeast genome by Raghuraman et al. (2001), local maxima of an intensity profile curve, denoted as %HL(·), across a chromosome correspond to the putative origins of replication that are coded by the autonomous replication sequence (ARS) elements. Here, variable %HL(x) presents a summary of the normalized intensity measurements at chromosomal coordinate x, which corresponds to a region of approximately 10 kb.

Let θ(x) denote an underlying profile curve (e.g., %HL curve across a chromosome) in a study, where variable x stands for chromosomal coordinate. Data collected from chromosome c are given in the form of \((y_{ci}, x_{ci}), i = 1, \ldots, nc\), where \(y_{ci}\) is an observed value of the profile variable θ subject to some noise, say \(ε_{ci}\), at coordinate \(x_{ci}\). That is, we assume that the observed outcomes \(y_{ci}\) are generated by the true underlying signal θ(·) that is perturbed by experimental noise. In this article, we focus on statistical inference on local extrema as a special kind of functional feature of θ(x). When values \(y_{ci}\)’s are plotted across the coordinates of chromosome c, local peaks or valleys can be easily visualized. See Figure 1, in which the solid line displays the observed %HL curve across yeast chromosome VI. However, not all observed peaks or valleys in the plot are the true extrema of θ(x). This is because the true underlying features of interest can be obscured or distorted by noise from a variety of sources such as probe gene selection variation and measurement errors in microarray experiments. As seen in Figure 1, the trajectory of the observed profile variable %HL in fact appears to be very wiggly, and therefore statistical methods are called for help in extracting and inferring the true underlying patterns such as local maximum.

Identifying local extrema of hybridization intensity profiles across a chromosome is a challenging task, simply because microarray data are typically noisy even if the raw data have been extensively pre-processed to remove random noise (Yang and Speed, 2003). Smoothing has been proved to be a powerful technique in the statistical literature to remove noise. For example, both moving-average smoothing (MAS) and Fourier convolution smoothing (FCS) have been used in Raghuraman et al. (2001) to smooth away noise-made wiggles. However, these two methods have a few drawbacks that limit their use in statistical inference. In particular, some of the drawbacks with the FCS are: (i) the FCS is essentially a deterministic smoothing technique, in the sense that it does
The sample path of summary variable %HL for chromosome VI. The vertical bars indicate the estimated 95\% confidence regions. The inverted triangles mark the origins detected by both FCS and the proposed NKS method, and the diamonds mark the origins detected only by the proposed NKS method.

not account for the randomness or uncertainty arising from a variety of noise sources in microarray experiments, (ii) the FCS does not have a data-driven algorithm to determine the optimal smoothing parameter, and (iii) the FCS does not lead to any statistical conclusions, such as confidence interval estimation, hypothesis testing, and p-values for statistical significance. The MAS suffers the same shortcomings as those of the FCS. In addition, the MAS essentially carries out equally weighted local averages based on bins of the same size, so the resultant smoothing cannot provide consistent fine-scale detection of local features across a chromosome. To overcome these issues, nonparametric smoothing techniques provide a natural arsenal in the detection of local features of the hybridization intensity data. Ghosh (2004) outlined three different approaches, namely Song et al.’s (1995) kernel estimation, locally weighted least squares estimation, and penalized smoothing spline estimation, with applications to the analysis of replication origins. In the paper, he also developed a hypothesis-testing procedure for the randomness of replication locations via Kolmogorov–Smirnov and scan statistics. As pointed out in his paper, however, assessing the variability of the identified replication origins remains an outstanding issue.

In this article, we propose a nonparametric kernel smoothing (NKS) technique that is naturally bonded with statistical inference for local extrema. The proposed method gives rise to a full package of nonparametric inference, including procedures of point estimation, confidence interval estimation, and hypothesis testing for positions of local maxima or minima. More importantly, we address the issue of assessing the variability of the identified replication origins, and provide unbiased estimation of the asymptotic variance–covariance matrices of the estimated local extrema. Such a new development is crucial to conduct a proper statistical analysis and to draw sensible conclusions in the microarray data analysis. As shown in Section 5, our NKS approach was able to detect a few new ARSs that were not found by the FCS method. Also, in most cases, our prediction of the ARS position appears to be more accurate than that given by the FCS method. This article shows another example that smoothing techniques can be appealing to make efficient inference on local features of hybridization intensity profiles. Simulation studies in Section 4 have revealed that the proposed inference procedure is powerful and robust to different error distributions with either heavier tails or longer tails than the normal as well as to cusp-type peak points.

The proposed NKS method is different from Song et al.’s (1995) kernel estimation approach that utilizes the derivatives of kernel functions to formulate the estimators of derivatives of the nonparametric regression function. We prefer the NKS method, because as a local fitting approach it has a few advantages: (i) it is easy to compute, as it requires effectively a weighted least squares procedure; (ii) it is easy to derive and assess bias and standard errors of the estimators; (iii) it is easy to conduct a local bandwidth selection. We programmed the NKS method in R/SPs package, and the source codes are available, together with a user’s manual, at www.stats.uwaterloo.ca/~song/Yeast_Genome.

Also we would like to point out that although the methodological development presented in this article is primarily focused on microarray data analysis, the proposed inference package is general and applicable to deal with a wide range of problems, such as proteomic analysis where the detection of local peaks in protein mass spectra is needed for selecting individual protein features (Tibshirani et al., 2003) or protein pattern clustering (Petricoin et al., 2002).

2. Yeast Expression Data
Our development in this article is largely motivated by the microarray data analysis given by Raghuraman et al. (2001) in which oligonucleotide microarrays were used to map the detailed topograph of chromosome replication in the budding yeast *Saccharomyces cerevisiae*. The replication of eukaryotic chromosomes is highly regulated, and it is known that replication is limited to the S phase of the cell cycle, and within S phase, initiation of replication is controlled with respect to time and location.

To map the dynamics of replication in the genome, culture samples were produced over time in a synchronized S phase. The outline of the experiment is described as follows. The yeast cells were first grown in isotopically heavy medium for many generations, and then were moved to isotopically light medium, so that the newly replicated DNA was labeled exclusively with the light isotopes. The replicated DNA was signified as heavy–light (HL) chromosomal DNA, as opposed to unreplicated heavy–heavy (HH) chromosomal DNA. The replicated HL DNA was also separated from the unreplicated HH DNA by density gradient centrifugation. This procedure was repeated eight times (0, 10, 14, 19, 25, 33, 44, and 60 minutes). Each DNA sample was then labeled and hybridized to a high-density oligonucleotide array, and each array contained a total of 157,112 different 25mer probes that covered 21.8\% of the nonrepetitive regions of the yeast genome. In order to identify origins, the average of the hybridization intensities for the surrounding 10 kb (5 kb on each side) was calculated for
points located every 500 bases along the genome from both HL and HH DNA samples. Note that hybridization intensity characterizes the relative abundance of specific genome sequences present in the unreplicated versus replicated DNAs at different times in S phase.

Aggregating the raw hybridization intensity values over the eight timed samples from both HH and HL DNA at each chromosome coordinate gives a pooled HL value, \( b_{\text{HL},i} \), and a pooled HH value, \( b_{\text{HH},i} \). Moreover, Raghuraman et al. (2001) suggested a summary variable defined as a percentage of \( 100 \times b_{\text{HL},i}/(b_{\text{HL},i} + b_{\text{HH},i}) \) at each chromosomal coordinate. This summary measure, referred to as the pooled %HL value, is useful to indicate the stage of replication, and a higher value of %HL is indicative of earlier replication.

To proceed with the statistical analysis, we suppose the observed pooled %HL values at chromosomal coordinate \( x \)'s are generated by the true underlying replication profile function \( \theta(x) \) that is perturbed by certain random noise. Figure 1 displays the observed pooled %HL values for chromosome VI at 538 coordinates, where each chromosomal coordinate represents approximately a region of 10 kb.

The objective is to infer the local maxima (peaks) and local minima (valleys) of \( \theta(x) \) across the yeast chromosomes. The local extrema relate to important biological features. That is, regions of the genome that replicate before their neighboring sequences should have higher %HL values than their neighbors. This implies that an origin of replication would appear as a local peak and a termination zone as a local valley of the pooled %HL curve \( \theta(x) \).

Results given by the MAS and FCS in Raghuraman et al. (2001) are not satisfactory simply because both methods did not provide any statistical inference. For example, they lack any suitable precision assessment as to how likely an identified local extremum would correspond to the true origin of replication. In addition, when a given chromosomal coordinate \( x_0 \) is suspected as an origin of replication, how would a hypothesis-testing procedure be invoked to confirm whether the corresponding \( \theta(x_0) \) is a true local extremum at 5% significance level?

The proposed NKS method in this article provides a full package of statistical tools that enable us to answer all these questions. Refer to Section 5 for the reanalysis of the yeast hybridization intensity data, with a detailed comparison to the Raghuraman et al. FCS method as well as to Ghosh’s (2004) findings.

3. Smoothing Technique

For each chromosome, let \( \theta(x) \) be a (locally) smooth function of a summary profile variable at chromosomal coordinate \( x \), with a finite number of local maxima and/or minima. Mathematically, by a local maximum or minimum at \( x^* \), we mean the value \( \theta(x^*) \) at the \( x^* \) with zero first-order derivative \( \theta(x^*) = 0 \). Sometimes, such an \( x^* \) is referred to as the equilibrium point. Because the proposed smoothing method is essentially a local inference, the assumption of smoothness is only required locally in the neighborhood of each equilibrium point \( x^* \).

A problem of interest is to detect the collection of all equilibrium points, namely a set

\[ E = \{ x : \hat{\theta}(x) = 0 \}, \]

which contains a finite number of elements. Statistical inference is needed simply because the function \( \theta(\cdot) \) is not known, and nor is its first-order derivative function \( \hat{\theta}(\cdot) \). To estimate \( \theta(x) \) or \( \hat{\theta}(\cdot) \), we assume the observed data \( (y_i, x_i), i = 1, \ldots, n \) (here the chromosome index \( c \) being suppressed for convenience), are generated according to the following nonparametric regression model:

\[ y_i = \theta(x_i) + \varepsilon_i, \quad i = 1, \ldots, n, \]

where \( \varepsilon_i \)'s are independent noise terms with mean zero and variance \( \sigma^2 \). Note that we do not assume the error terms \( \varepsilon_i \) are normally distributed. Later in Section 4.2, we investigate the robustness of the proposed inference against the distributional assumption, where three distributions, normal, \( \varepsilon_i \), and log normal, are used in a simulation study.

3.1 Detection

Our first goal is to estimate function \( \theta(\cdot) \) and its first-order derivative function \( \hat{\theta}(\cdot) \) nonparametrically using the local quadratic fitting method (Wand and Jones, 1995, Chapter 5).

Instead of directly identifying local peaks or valleys from the estimated \( \hat{\theta}(\cdot) \), we used the estimated derivatives to find set \( E \) that contains all equilibrium points (or roots) of \( \hat{\theta}(x) = 0 \). At peak points the sign of \( \hat{\theta}(x) \) will change from positive to negative (or \( \hat{\theta}(x) < 0 \)), and the opposite direction of sign change (or \( \hat{\theta}(x) > 0 \)) suggests valley points. More importantly, this approach allows us to easily establish statistical inference by a straightforward application of the existing nonparametric kernel regression theory.

The local quadratic fitting method requires to minimize the following objective function in the form of kernel weighted least squares at a given \( x \), with respect to a parameter vector \( \alpha = (\alpha_0, \alpha_1, \alpha_2)^T \):

\[ U(\alpha) = \sum_{i=1}^n \mathcal{K}_h(x_i - x) \{ y_i - \alpha_0 - \alpha_1(x_i - x) - \alpha_2(x_i - x)^2 \}^2, \]

with \( \mathcal{K}_h(u) = h^{-1}\mathcal{K}(u/h) \), where \( h \) is the bandwidth and \( \mathcal{K}(\cdot) \) is the standard kernel function (symmetric and variance 1). In this article, we always choose the normal kernel, namely \( \mathcal{K}(\cdot) \), as the density of the standard normal, which is the most commonly used kernel in the literature. It follows immediately that at a given target coordinate \( x, \hat{\theta}(x) = \hat{\alpha}_0, \hat{\theta}(x) = \hat{\alpha}_1, \hat{\theta}(x) = \hat{\alpha}_2 \), where

\[ \hat{\alpha} = (\hat{\alpha}_0, \hat{\alpha}_1, \hat{\alpha}_2)^T = \arg\min_{\alpha} U(\alpha). \]

Note that here we adopt the local quadratic fit, rather than local linear fit or local cubic fit, in order to reduce the edge effect (especially in estimating the derivative function \( \theta \)) known to the local polynomial fitting method, while retaining the minimal requirement of the model assumptions for the development of the smoothing technique. The explicit expression of the estimator \( \hat{\alpha} \) is given by equation (A.1). Moreover, \( \hat{\theta}(x) \) and \( \hat{\theta}(x) \) are given, respectively, by (A.2) and (A.3).
Following Wand and Jones (1995, Section 5.7), we obtain the asymptotic bias and variance of the estimator $\hat{\theta}(x)$ given by (A.3), that
\[
\text{bias}(\hat{\theta}(x)) = \frac{1}{6} \mu_4(K) \hat{\theta}(x) h^2 + o_p(h^2)
\]
\[
\text{var}(\hat{\theta}(x)) = n^{-1} h^{-3} \nu_2(K) \sigma^2 \frac{1}{f(x)} + o_p(n^{-1} h^{-3}),
\]
where $\hat{\theta} (\cdot)$ denotes the third-order derivative of $\theta (\cdot)$. Moreover, under some mild regularity conditions, when $n \to \infty$ and $h \to 0$ at rate of $nh^3 \to \infty$,
\[
\sqrt{nh^3} \left\{ \hat{\theta}(x) - \theta(x) - \frac{1}{6} \mu_4(K) \hat{\theta}(x) \sqrt{\frac{h}{n}} \right\} \to N \left( 0, \nu_2(K) \sigma^2 \frac{1}{f(x)} \right), \tag{3}
\]
where $\mu_4(K) = \int w^4 K(u) \, du$, $j = 1, \ldots, 4$ and $\nu_2 = \int w^2 K^2(u) \, du$, $j = 1, 2.$

The result of asymptotic normality in (3) would be immediately applied to construct the confidence interval or to determine the asymptotic rejection critical value if $f(x)$, the density function of the chromosomal coordinates, were known. Instead, we will use the estimated variance–covariance matrix of $\hat{\alpha}$, given by (A.5). Note that the asymptotics given above require the number of coordinates ($n$) going to infinity, rather than the number of arrays going to infinity.

### 3.2 Inference

The result of asymptotic normality in (3) would be immediately applied to construct the confidence interval or to determine the asymptotic rejection critical value. Unfortunately, as shown in (3), the asymptotic variance of the $\hat{\theta}(x)$ is not directly usable because of the involvement of the unknown function $f(x).$ Here, $f(x)$ represents the probability density of the coordinate variable $x$ across a chromosome. Instead, we use the variance of the weighted least squares estimator $\hat{\alpha}$ to derive the variance of $\hat{\theta}(x)$, given by equation (A.5). Then, the standard error of $\hat{\theta}(x)$, denoted by $SE[\hat{\theta}(x)]$, is available for statistical inference. See equation (A.6).

Consequently, the $95\%$ pointwise confidence band for $\hat{\theta}(x)$ is
\[
\hat{\theta}(x) \pm 1.96 \, SE[\hat{\theta}(x)].
\]

Denote the upper and lower confidence limits, respectively, by
\[
\text{UCL}_\theta(x) = \hat{\theta}(x) + 1.96 \, SE[\hat{\theta}(x)],
\]
\[
\text{LCL}_\theta(x) = \hat{\theta}(x) - 1.96 \, SE[\hat{\theta}(x)].
\]

Let $x_0$ be an (identified) equilibrium point. In a small neighborhood $\delta(x_0)$ around $x_0$ on a chromosome, both $\text{UCL}_\theta(x)$ and $\text{LCL}_\theta(x)$, $x \in \delta(x_0)$, usually appear to be monotonic, so inverting them will lead to the (approximate) $95\%$ confidence interval, $[\text{LCL}_\theta(x_0), \text{UCL}_\theta(x_0)]$, for the given chromosomal location $x_0$. However, this inverting procedure may fail to work, although very rarely, in the situation where the $\hat{\theta}(x)$ itself does not cross zero at $x_0$ but the confidence interval $[\text{LCL}_\theta(x_0), \text{UCL}_\theta(x_0)]$ contains zero. In such a case, the inverting procedure only gives one limit, but the other one can be estimated by the symmetry. See more details in the supplementary documents, including the R/S+ code, available at the first author’s web site as given at the end of Section 1.

Let us now consider the hypothesis-testing problem at a single coordinate. Suppose one wants to test whether a suspected location $x_0$ is a true equilibrium point. Then, the null and alternative hypotheses are
\[
H_0 : \theta(x_0) = 0 \quad \text{versus} \quad H_1 : \theta(x_0) \neq 0.
\]

The rejection rule is that at this location $x_0$, we reject the null at $\alpha = 5\%$ significance level, when the $95\%$ confidence interval $[\text{LCL}_\theta(x_0), \text{UCL}_\theta(x_0)]$ does not contain zero. Moreover, if the second derivative $\hat{\theta}''(x_0)$ is negative, then we declare $x_0$ as a peak point. It is worth mentioning that it is the null hypothesis that interests us, as it gives rise to a discovery. So, in the testing problem, the locations with $p$-values bigger than 0.05 will be selected.

### 3.3 Multiple Equilibrium Points

The problem of testing multiple locations requires generally more sophisticated statistical tools, which are currently still under development in the statistical literature. For the analysis considered in the present article, we suggest the following two approaches that seem appealing in the context of detecting replication origins through the method of confidence bands. Depending on which type of error is controlled, these two simultaneous approaches lead to the respective rejection rules.

The first approach is established when the type I error is controlled at level $\alpha$. It seems natural here to adopt the method of simultaneous confidence band for a nonparametric function, which has been studied extensively in the statistical literature. See, for example, Loader (1999, Chapter 9), where an R-software function $\text{locfit}()$ is available to obtain a $(1 - \alpha)100\%$ simultaneous confidence band, $\{(L(x), U(x)), x \in \mathcal{X}\}$, satisfying
\[
\sup_{\theta \in \Theta} P_{\theta}(L(x) \leq \theta(x) \leq U(x), \forall x \in \mathcal{X}) \geq 1 - \alpha,
\]
where $\mathcal{X}$ is the domain of the underlying function $\theta(x)$ and $\Theta$ is a suitable class of smooth functions.

Our experience with the method of simultaneous confidence band indicates that it often comes up with undersmoothed bands. This is because in this case the bandwidth has to be selected globally and the procedure of determining an optimal bandwidth is computationally intensive. In addition, this simultaneous method assumes global smoothness for the underlying intensity function $\theta(\cdot)$, a stronger assumption than local smoothness that is required in the method of pointwise confidence interval. Therefore, in the application of the simultaneous confidence band for the detection of multiple equilibrium points, one needs to validate the global smoothness assumption and to obtain a certain global optimal bandwidth that has to accommodate local curvatures well.

The other approach is to control the experimentwise type II error. Suppose there are $m$ locations across a chromosome, which are potentially equilibrium points. To perform the $m$
hypothesis tests simultaneously, we establish the needed rejection rule by controlling the type II error. To proceed, we relax the statement of the null hypothesis at each of the coordinates as $H_0 : |\hat{\theta}(x)| < \delta$, where $\delta$ is a pre-specified tolerance level. In fact, the idea of tolerance is commonly used in the field of bioequivalence to establish equivalence, which shares great similarity to our problem here.

Let $x_i$, $i = 1, \ldots, m$ be the $m$ suspected coordinates. At each $x_i$, reject the $i$th single null if $|\hat{\theta}(x_i)| \geq C_i$, for a certain value $C_i$. Thus, the corresponding type II error takes the form

$$\beta_i = P(|\hat{\theta}(x_i)| < C_i |\hat{\theta}(x_i)| \geq \delta), \quad i = 1, \ldots, m.$$ 

The utility of well-known Bonferroni correction simply leads to $\beta_i = \beta/m$, under which we can determine the critical value $C_i$ based on the theory of the asymptotic normality and hence the acceptance rule $|\hat{\theta}(x_i)| < C_i$.

Alternatively, if a less stringent multiple hypothesis control is preferred, one can follow the stepwise procedures of the false discovery rate approach proposed by, for example, Benjamini and Hochberg (1995) and Dudoit et al. (2002).

### 3.4 Bandwidth Selection

Bandwidth $h$ is an important parameter in the smoothing algorithm as it determines the degree of smoothness. An advantage of the NKS technique is that the bandwidth can be determined automatically by the observed data under a certain optimality criterion, rather than being picked up artificially by a data analyst. The criterion that we use in this article is minimum asymptotic mean integrated squared error (AMISE). To calculate the AMISE optimal $h$, we adopt Ruppert, Sheather, and Wand’s (1995) direct plug-in (DPI) rule, which has been shown to possess attractive theoretical and practical properties.

### 4. Simulation Experiments

We conducted three simulation studies to examine the power and robustness of the proposed NKS procedure, assessing the effects of cusp points, of error distribution assumptions, and of the existence of multiple equilibrium points. The significance level $\alpha$ was set at 0.05.

#### 4.1 Effects of Cusps

The first simulation focused on the effects of cusps on the robustness of the proposed method. By a cusp location, we mean the coordinate at which two linear functions intersect. Cusps might be regarded as one main type of violation of the smoothness assumption. We consider a simple simulation to provide some numerical evidence of how the ignorance of smoothness condition affects the detection of equilibrium points. The data were generated from the following model with a single cusp point at coordinate $x_0 = 2$:

$$y_i = -2|x_i - 2| + \epsilon_i, \quad i = 1, \ldots, 100,$$

where $\epsilon_i$’s are independently simulated from $N(0, \sigma^2)$. The performance of the NKS procedure was assessed at different variance values, each based on 1000 replicates. Table 1 reports the results.

Based on Table 1, we make the following conclusions: (i) Evidently, because the coverage rate for $\theta(2)$ is close to the nominal level, the asymptotic normality seems valid even if the cusp model (4) violates the smoothness; (ii) the coverage rate for the cusp location ($x_0 = 2$) drops when the variance increases, but at a reasonable noise level, say $\sigma \leq 2$, the performance of the NKS method remains robust against such deviation from the smoothness. Through careful low-level data pre-processing with random noise being greatly reduced, the proposed NKS method is expected to work well.

We also comment that if one knows that data are generated from the cusp model (4), the empirical least squares method proposed by Lazzeroni (1998) can be directly used to infer such a cusp location.

#### 4.2 Effects of Error Distributions

The second simulation is aimed to assess the power and robustness of the proposed NKS technique to different error distributions. According to Carey (2004), microarray (normalized) expression data may follow different distributions such as heavy tailed, skewed, and multimodal distributions. Therefore a statistical inference that does not bear on the normality assumption, like the one proposed in this article, is typically appealing in the microarray data analysis. In particular, we considered three error distributions in data generation of model (1), namely, normal $N(0, \sigma^2)$, heavy tailed $\sigma_3$, (with variance $2\sigma^2$), and right-skewed log normal $LN(0, \sigma^2)$.

Because the proposed NKS technique was an inference method for local extrema, it is sufficient to consider a simple function $\theta(x) = -3(x - 2)^2 + 5$ with only an equilibrium point at $x_0 = 2$. The reason that we considered a quadratic function was that its third-order derivative $\hat{\theta}(x) \equiv 0$, which implies that the bias given in (3) is zero, regardless of bandwidth $h$. The null hypothesis of interest was $H_0 : \hat{\theta}(2) = 0$, with a two-sided alternative. The testing procedure proposed in Section 3.2 was employed to test for this $H_0$, and the power function, which was calculated as the percentage of rejecting $H_0$ among 1000 simulations at each of grid locations $x = 2 \pm 0.05q$, $q = 1, 2, \ldots, n$, was displayed in Figure 2 for all three error distributions. In particular, at $x = 2$ with $q = 0$ (i.e., $H_0$) the rejection percentage effectively is equal to the empirical type I error rate, and at any fixed location different from $x = 2$ (i.e., $H_1$) the rejection percentage corresponds to the empirical power. In addition, we specified noise variance $\sigma^2$ at two values 1 and 0.5, and in each replication we used the DPI method to compute the optimal bandwidth. The empirical type I error rates were (0.050, 0.054) for normal, (0.052, 0.047) for $t_2$, and (0.050, 0.053) for log normal, with the first value in the parentheses corresponding to the $\sigma = 0.5$ and the second to the $\sigma = 1$.

### Table 1

| SD | Bandwidth Coverage Coverage for $\theta(2)$ Coverage for $x_0 = 2$ |
|----|-----------------|-----------------|-----------------|
| 0.5| 0.21            | 0.956           | 0.988           |
| 1.0| 0.29            | 0.952           | 0.976           |
| 2.0| 0.37            | 0.943           | 0.910           |
| 3.0| 0.41            | 0.940           | 0.775           |
Figure 2. The empirical power function based on 1000 simulations, for normal (top), $t_3$ (middle), and log-normal (bottom) error distributions. Two functions in one panel correspond to two noise variances $\sigma^2 = 1$ and 0.5. The lowest position of each curve gives the empirical size of the testing procedure.

Figure 2 clearly indicates that the proposed NKS technique works very well for various error distributions, and for all three cases the power gain accelerates rapidly when $x$ moves away from the true equilibrium point $x_0 = 2$. As expected, a better designed experiment with smaller noise variance $\sigma^2$ would definitely give rise to the significant gain of testing power. It can be seen that in the case of log-normal error distribution (a right-skewed distribution) the variance parameter will affect the power gain the most among the three distributions considered, because the gap between the two power functions appears to be the largest in this case. Note that in all three panels the first added horizontal line from the bottom is at the height of the significance level $\alpha = 0.05$. This indicates that the testing procedure attained proper empirical sizes in all three distribution cases.

4.3 Effects of Multiple Equilibrium Points

The objective of the third simulation was to illustrate that the proposed NKS technique is indeed a local inference approach. To make the simulation setup slightly more general, we considered a $\theta(x)$ function with two peak points, at one of which the local smoothness was satisfied but was not at the other location. The function $\theta(x)$ in model (1) used to generate the data was specified as

$$\theta(x) = \begin{cases} -2|x - 1.5| & \text{if } x < 2.5 \\ -3(x - 3.5)^2 + 1 & \text{if } x \geq 2.5. \end{cases}$$

Clearly, two peaks locate at $x_1 = 1.5$ and $x_2 = 3.5$. Again, 100 errors in model (1) were simulated from a normal distribution with variance $\sigma^2$. Based on 1000 replications, we summarize the results in Table 2.

We learned from Table 2 that the asymptotic normality was local and valid at each of the two local peak locations, as the empirical coverage rates for the respective first derivatives were close to the nominal 0.95 level. However, the coverage rate for each true peak location clearly depended on the magnitude of the variance (or noise) $\sigma^2$. The proposed method worked well when the variance was low, but could lose its power substantially when the noise level was high.

In the yeast data analysis presented in Section 5, the variance of noise is found mostly low for the 16 chromosomes. For example, for the chromosome VI data, the estimated $\sigma$ is 0.6, and hence the proposed NKS technique should remain powerful.

5. Application

This section presents a reanalysis of the yeast genome data that have been previously analyzed by Raghuraman et al. (2001) using the FCS method. In this analysis, the primary interest was to infer local maxima that correspond to the putative origins of replications in the cell cycle. Refer to Section 2 for details regarding the experiment as well as data collection. Due to space limitation, we here only report our findings on chromosome VI (see Figure 1 for the observed %HL values). As pointed out in Raghuraman et al. (2001), chromosome VI has been studied extensively in the literature (Friedman, Brewer, and Fangman, 1997; Yamashita et al., 1997) and is a good candidate for assessment of the proposed approach. The rest of the 15 chromosome data were analyzed in an almost identical manner, and interested readers can find the complete results in the supplementary documents available at www.stats.uwaterloo.ca/~song/Yeast_Genome.html.

One most remarkable finding in the analysis of chromosome VI is that the proposed NKS method detected a peak at location 202.042, with the 95% confidence interval [200.540, 203.544]. This location is very close to the region, 199.383–199.493, of an ARS element labeled as ARS607 in the yeast genome database (www.yeastgenome.org). The distance from
The re-
hanate 202.042 passed. More discussions will follow later in this
stringent criterion given in the multiple testing, while coordi-
nate 255.620 to be a real discovery as it did not pass a more
ther. However, we are cautious to announce the edge coordi-
which was not previously detected by the FCS method, ei-
omosome. It is also notable that we found another new location
has been shown to be the earliest-activated origin on the chro-
cence, because according to Friedman et al. (1997), ARS607
between two positions 196.033 and 235.591 that were found
dected by our NKS method is a new discovery, which locates
196.033, which was also detected by our NKS method. As a
matter of fact, the local peak located at coordinate 202.042
detected by our NKS method is a new discovery, which locates
found by the FCS method. This new finding has biological signifi-
cance, because according to Friedman et al. (1997), ARS607
shown to be the earliest-activated origin on the chro-
section.
We found the optimal bandwidth value by the DPI rule,
$h_{opt} = 1.895$. Figure 3 shows the fitted curve of the first-order
derivatives, $\hat{\theta}(x)$, and the 95% pointwise confidence band in-
dicated by the shadowed belt (in brown color in the supple-
mentary documents available at the first author’s web site
as given at the end of Section 1). On the basis of the 95% confidence intervals, we found 14 equilibrium points in the
interior region and 2 points at the two end edges, all of which
are marked by crosses (×) along the x-axis in Figure 3. These
locations are also marked by the inverted triangles and di-
amonds in Figure 1, where the diamonds mark the new co-
ordinates found by our NKS method. Note that we ignored
these two edge points in our further analysis, simply because
the NKS method is known often to produce biased estimation
at the edges. The resulting confidence regions for all 14
identified locations are plotted by the vertical gray bars in
Figure 1.

The detailed numerical results are summarized in Table 3.
The first column reports the actual coordinates of the 14
sites found to have significant local peaks, with their approx-
imate 95% confidence intervals being listed in the sec-
ond column. It is worth pointing out that our NKS method
used p-values to judge statistical significance for each of these
sites, as seen in the third column. In contrast, Raghuraman
et al.’s (2001) FCS method relied on a measure of frequency
defined as the number of times that a local peak appeared to
be a local maximum over a small number of chosen smoothing
values. See column 5, in which Raghuraman et al. (2001)
used nine different smooth parameters. This frequency can
only be thought of intuitively as a kind of persistence of a
local maximum against smoothing, and this frequency mea-
ure has no proper probability interpretation associated with

<table>
<thead>
<tr>
<th>SD</th>
<th>Bandwidth</th>
<th>Coverage for location $\hat{\theta}(1.5)$</th>
<th>Coverage for location $\hat{\theta}(3.5)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5</td>
<td>0.17</td>
<td>0.960</td>
<td>0.981</td>
</tr>
<tr>
<td>1.0</td>
<td>0.22</td>
<td>0.942</td>
<td>0.895</td>
</tr>
<tr>
<td>2.0</td>
<td>0.27</td>
<td>0.940</td>
<td>0.940</td>
</tr>
<tr>
<td>3.0</td>
<td>0.31</td>
<td>0.938</td>
<td>0.940</td>
</tr>
</tbody>
</table>

The estimated first-order derivative curve with 95% confidence band. The equilibrium points are marked by crosses.
Table 3
Data analysis results. Column 1 lists the coordinates (in kb) of the 14 identified equilibrium points and the corresponding 95% confidence intervals and p-values are listed in columns 2 and 3 by the proposed NKS method. For comparison, columns 4 and 5 list the coordinates of the sites and their frequencies of persistence against smoothing by the FCS method. The rest lists the available ARSs and ACSs found from the respective databases.

<table>
<thead>
<tr>
<th>Coordinate</th>
<th>NKS [LCL, UCL]</th>
<th>p-value</th>
<th>FCS Coordinate</th>
<th>Freq.</th>
<th>Name</th>
<th>Region</th>
<th>ARS</th>
<th>Region</th>
<th>ACS</th>
<th>Region</th>
</tr>
</thead>
<tbody>
<tr>
<td>36.302</td>
<td>(34.800, 39.308)</td>
<td>0.91*</td>
<td>37.304</td>
<td>9/9</td>
<td>ARS603</td>
<td>68.691 - 68.869</td>
<td>ARS4</td>
<td>68.927 - 68.837</td>
<td></td>
<td></td>
</tr>
<tr>
<td>47.819</td>
<td>(46.314, 49.346)</td>
<td>0.79*</td>
<td>46.346</td>
<td>9/9</td>
<td>ARS604</td>
<td>73.746 - 74.836</td>
<td>ARS6</td>
<td>73.827 - 74.836</td>
<td></td>
<td></td>
</tr>
<tr>
<td>96.890</td>
<td>(95.888, 97.891)</td>
<td>0.46</td>
<td>NA</td>
<td>0/9</td>
<td>ARS605</td>
<td>135.980 - 136.080</td>
<td>ARS3</td>
<td>135.976 - 136.080</td>
<td></td>
<td></td>
</tr>
<tr>
<td>122.427</td>
<td>(121.926, 122.927)</td>
<td>0.41*</td>
<td>122.427</td>
<td>9/9</td>
<td>ARS606</td>
<td>167.607 - 168.041</td>
<td>ARS1</td>
<td>167.716 - 168.041</td>
<td></td>
<td></td>
</tr>
<tr>
<td>138.951</td>
<td>(138.450, 139.451)</td>
<td>0.17</td>
<td>138.951</td>
<td>6/9</td>
<td>ARS607</td>
<td>199.380 - 199.493</td>
<td>ARS2</td>
<td>199.404 - 199.493</td>
<td></td>
<td></td>
</tr>
<tr>
<td>146.462</td>
<td>(145.460, 147.463)</td>
<td>0.88</td>
<td>NA</td>
<td>0/9</td>
<td>ARS608</td>
<td>256.264 - 256.418</td>
<td>ARS5</td>
<td>256.372 - 256.418</td>
<td></td>
<td></td>
</tr>
<tr>
<td>168.493</td>
<td>(167.492, 169.996)</td>
<td>0.78*</td>
<td>168.994</td>
<td>9/9</td>
<td>ARS609</td>
<td>279.803 - 280.053</td>
<td>ARS7</td>
<td>279.803 - 280.053</td>
<td></td>
<td></td>
</tr>
<tr>
<td>195.533</td>
<td>(194.532, 196.533)</td>
<td>0.51</td>
<td>196.033</td>
<td>9/9</td>
<td>ARS610</td>
<td>320.532 - 320.732</td>
<td>ARS8</td>
<td>320.532 - 320.732</td>
<td></td>
<td></td>
</tr>
<tr>
<td>251.113</td>
<td>(250.112, 252.114)</td>
<td>0.58*</td>
<td>NA</td>
<td>0/9</td>
<td>ARS611</td>
<td>354.532 - 354.732</td>
<td>ARS9</td>
<td>354.532 - 354.732</td>
<td></td>
<td></td>
</tr>
<tr>
<td>255.620</td>
<td>(254.618, 256.621)</td>
<td>0.09</td>
<td>NA</td>
<td>0/9</td>
<td>ARS612</td>
<td>378.532 - 378.732</td>
<td>ARS10</td>
<td>378.532 - 378.732</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*See text below.

Table 4
Comparison between the proposed NKS and the existing FCS methods on the basis of the match between known efficient origins and local peaks in replication profiles. Locations are in kb from the left end of each chromosome.

<table>
<thead>
<tr>
<th>Chromosome</th>
<th>Known origins</th>
<th>FCS</th>
<th>NKS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Name</td>
<td>Left end</td>
<td>Right end</td>
</tr>
<tr>
<td>II</td>
<td>SH1022</td>
<td>61.200</td>
<td>61.500</td>
</tr>
<tr>
<td>III</td>
<td>ARS305</td>
<td>39.132</td>
<td>39.682</td>
</tr>
<tr>
<td>III</td>
<td>ARS306</td>
<td>74.428</td>
<td>74.648</td>
</tr>
<tr>
<td>III</td>
<td>ARS307</td>
<td>108.775</td>
<td>108.975</td>
</tr>
<tr>
<td>III</td>
<td>ARS309</td>
<td>131.582</td>
<td>131.928</td>
</tr>
<tr>
<td>III</td>
<td>ARS310</td>
<td>165.211</td>
<td>165.928</td>
</tr>
<tr>
<td>IV</td>
<td>ARS1</td>
<td>462.611</td>
<td>462.648</td>
</tr>
<tr>
<td>IV</td>
<td>SH1023</td>
<td>750.000</td>
<td>752.500</td>
</tr>
<tr>
<td>V</td>
<td>ARS501</td>
<td>549.644</td>
<td>549.680</td>
</tr>
<tr>
<td>VI</td>
<td>ARS603</td>
<td>65.000</td>
<td>65.000</td>
</tr>
<tr>
<td>VI</td>
<td>ARS603.5</td>
<td>117.101</td>
<td>120.038</td>
</tr>
<tr>
<td>VI</td>
<td>ARS606</td>
<td>167.652</td>
<td>167.750</td>
</tr>
<tr>
<td>VI</td>
<td>ARS607</td>
<td>199.051</td>
<td>199.554</td>
</tr>
<tr>
<td>X</td>
<td>ARS121</td>
<td>683.653</td>
<td>683.702</td>
</tr>
<tr>
<td>XIV</td>
<td>ARS1411</td>
<td>168.000</td>
<td>170.100</td>
</tr>
<tr>
<td>XIV</td>
<td>ARS1412</td>
<td>195.652</td>
<td>196.883</td>
</tr>
<tr>
<td>XIV</td>
<td>ARS1413</td>
<td>250.800</td>
<td>250.849</td>
</tr>
<tr>
<td>XIV</td>
<td>ARS1414</td>
<td>280.001</td>
<td>280.049</td>
</tr>
</tbody>
</table>
to low quality allocation of probes and limited samples collected from the experiment. It is remarkable that the key new discovery located at 202.042 kb passed this more stringent criterion.

In the comparison of our NKS method to the FCS method on the other chromosomes, similar to Table 1 in Raghuraman et al. (2001), Table 4 presents a summary regarding the match between known efficient origins and local peaks in replication profiles for both methods. Clearly, the NKS method was able to rediscover all reported locations by the FCS method which were matched with the known replication origins. When the 95% confidence interval estimation is used, the NKS method is obviously a more favorable approach to detecting locations of local peaks, because it gives not only an estimated location but also the precision of the estimation.

It is noted that the data of the known origins are whole copied from Raghuraman et al. (2001), which are effectively different from those given in the yeast genome database, www.yeastgenome.org. For instance, the coordinates of chromosome VI are different in Tables 3 and 4, because Table 3 was created using the updated data from this yeast genome database. Also, we found the coordinates of chromosome XIV in Table 4 are not available in this yeast genome database, and we list them here merely for comparison purpose.

6. Concluding Remarks
We present a powerful nonparametric technique to infer local extrema. The results from both simulation studies and application to chromosome VI suggest that the proposed NKS method is preferred to the FCS method. In the reanalysis of chromosome VI data, we were able to find three new locations, at two of which exist ARS elements and at the other is an ACS element. However, accepting the edge coordinate 255.620 as a new discovery requires more cautious analysis. The proposed method gives p-values to declare statistical significance in the detection of local extrema, conclusions can be drawn in the proper probability sense. In addition, the related computational burden is marginal because the NKS method essentially is built based on the weighted least squares approach. The determination of the optimal bandwidth appears to be more computationally extensive, but for all 16 chromosomes we analyzed, the computation with the DPI rule was fast in R software.

Residual analysis is always important to assess the goodness of fit for the model invoked. In this particular application, the presence of serial correlation is not uncommon due to the nature of the replication procedure itself. So, when the serial correlation is present in residuals, it is generally difficult to judge whether the correlation is due to the data themselves or due to a wrong model. The reason that we did not include serial correlation in the estimation is that Lin and Carroll (2000) show that the local polynomial regression estimation attains the highest efficiency under the independence correlation when a traditional kernel (such as the normal kernel) is used. Moreover, with regard to the estimation of the mean function, namely the $\theta(x)$ function in this article, consistency is always valid even if the serial correlation is misspecified or not incorporated.

The proposed method gives p-values to reflect the evidence of significance, but it is known that the p-value has its own limitations, including its reliance on sample size. In the application presented in this article, we emphasize more the use of confidence intervals to detect peaks of interest, and the use of the p-value is limited to determining whether particular regions exhibit peaks or not. We comment that in general, the detection of peaks is more interesting than testing, in the context of local feature extraction considered in this article.

Finally, we remark that although the NKS method was mostly developed to analyze a special kind of microarray data, it is a general smoothing technique that can be applied to solve similar problems in other fields such as proteomics. See, for example, Tibshirani et al. (2003) and Petricoin et al. (2002), in which the analysis of protein mass spectroscopy relates to the identification of local peaks.

ACKNOWLEDGEMENTS
The authors are very grateful to the three referees for their valuable comments and constructive suggestions that led to a substantial improvement of the article. The authors also thank the editor (Mike Kenward) for his suggestions for improving the exposition of the article. This research is supported by Natural Science and Engineering Research Council of Canada grants held by the first two authors (Song and Gao).

REFERENCES


APPENDIX

Local Quadratic Fitting

This appendix is given solely for the purpose of self-containedness of our technical presentation in the article. Here, we follow the notation in Section 3. The estimator $\hat{\alpha}$ can be obtained by the formula of weighted least-squares estimation (WLSE), given as follows. Let

\[
Y = [y_1, \ldots, y_n]^T, \quad X_i(x) = [1, (x_i - x), (x_i - x)^2]^T,
\]

\[
X(x) = \begin{bmatrix}
X_1(x)^T \\
\vdots \\
X_n(x)^T
\end{bmatrix} = \begin{bmatrix}
1 & x_1 - x & (x_1 - x)^2 \\
\vdots & \vdots & \vdots \\
1 & x_n - x & (x_n - x)^2
\end{bmatrix},
\]

and $K_h(x) = \text{diag}[K_h(x_1 - x), \ldots, K_h(x_n - x)],$ a diagonal matrix with the $i$th diagonal element being $K_h(x_i - x).$ Then, the estimator takes the WLSE form,

\[
\hat{\alpha} = [X(x)^T K_h(x) X(x)]^{-1} X(x)^T K_h(x) Y, \quad (A.1)
\]

where $K_h(x)$ is effectively the weight matrix. In particular,

\[
\hat{\theta}(x) = \hat{\alpha}_0 = e_1^T \hat{\alpha}
\]

\[
= e_1^T [X(x)^T K_h(x) X(x)]^{-1} X(x)^T K_h(x) Y, \quad (A.2)
\]

\[
\hat{\theta}(x) = \hat{\alpha}_1 = e_2^T \hat{\alpha}
\]

\[
= e_2^T [X(x)^T K_h(x) X(x)]^{-1} X(x)^T K_h(x) Y, \quad (A.3)
\]

\[
\hat{\theta}(x) = 2\hat{\alpha}_2 = 2e_3^T \hat{\alpha}
\]

\[
= 2e_3^T [X(x)^T K_h(x) X(x)]^{-1} X(x)^T K_h(x) Y, \quad (A.4)
\]

where $e_1 = [1, 0, 0]^T,$ $e_2 = [0, 1, 0]^T,$ and $e_3 = [0, 0, 1]^T.$ Repeating the above estimation procedure at different coordinates across a chromosome leads to estimated functions $\hat{\theta}(\cdot)$ and $\hat{\theta}(\cdot).$

We now present the estimation for the asymptotic variance of $\hat{\theta}(x).$ Under the assumption $\varepsilon_i$'s being white noise with mean zero and variance $\sigma^2,$ we have $\text{var}(Y) = \sigma^2 I_n,$ where $I_n$ is the $n$-dimensional identity matrix. It follows immediately from (A.1) that the variance–covariance matrix of $\hat{\alpha}$ is

\[
\text{var}(\hat{\alpha}) = \sigma^2 A_n^{-1} B_n A_n^{-1},
\]

and the second diagonal element of this matrix gives the variance of $\hat{\theta}(x),$

\[
\text{var} [\hat{\theta}(x)] = \sigma^2 e_1^T A_n^{-1} B_n A_n^{-1} e_2, \quad (A.5)
\]

where

\[
A_n = \frac{1}{n} X(x)^T K(x) X(x),
\]

\[
B_n = \frac{1}{n^2} X(x)^T K(x) X(x).
\]

In addition, an unbiased estimator of the variance parameter $\sigma^2$ is given as follows:

\[
\hat{\sigma}^2 = \frac{1}{n-v} \sum_{i=1}^n [y_i - \hat{\theta}(x_i)]^2
\]

with “degrees of freedom”

\[
v = 2 \sum_{i=1}^n w_{ii} - \sum_{i=1}^n \sum_{j=1}^n w_{ij}^2
\]

where

\[
w_{ij} = e_1^T [X(x_i)^T K_h(x_i) X(x_i)]^{-1} X(x_i)^T K_h(x_i) e_j
\]

and

\[
K_h(x_i) = \text{diag}[K_h(x_1 - x_i), \ldots, K_h(x_n - x_i)]
\]

based on a bandwidth $h'$ that may be different from the $h$ used in the estimation $\hat{\theta}(x).$

Therefore, the estimated standard error is

\[
\text{SE}[\hat{\theta}(x)] = \hat{\sigma} \left\{ e_1^T A_n^{-1} B_n A_n^{-1} e_2 \right\}^{1/2}. \quad (A.6)
\]