Chapter 13
DNA Sequencing

Lei M. Li

DNA sequencing is one great leap in the advance of life sciences. The research in these two sequencing articles came respectively from two chapters of my PhD thesis at Berkeley, and it is my luck to connect my life with DNA sequencing through my wonderful thesis advisor, Professor Terry Speed.

Several factors motivated the selection of DNA sequencing as my thesis topic. First, I had an ambition to be an applied mathematician when I was young. In my last year of college, however, I was tortured by fatigue and infection. My mood was very low then, and I had no appetite for more mathematics at all. With the help of my family, I gradually recovered with a therapy of Chinese herbs, and I continued to study hard mathematics. From then on, I had a vague yet deep thought in my mind that someday I should apply my mathematical knowledge to the understanding of life and medicine. This was one reason why in graduate school I looked for some applied topic related to life sciences. Second, Terry had been working on statistical genetics and was very enthusiastic about new statistical problems in genomics. He gave me a physical mapping problem as a start-up project. I quickly made some progress that helped me pass the oral exam. Third, in 1995 the Human Genome Project accelerated and researchers from many disciplines such as chemistry, engineering, computer science, mathematics, statistics jumped into the field. And Terry brought me into the adventure with a good will.

Among the interesting mathematical problems associated with genomics, I picked DNA sequencing, or more exactly, DNA base-calling as my thesis topic, as Terry suggested. In 1994-95, DNA sequencing was based on Sanger's dideoxy DNA amplification, fluorescence dye technique and electrophoresis. In the beginning, I knew nothing about molecular biology, and Terry helped me understand the basic ideas with a great patience. His former student, David Nelson, participated in DNA sequencing research at that time too, and provided us with fairly complete background on electrophoresis [8, 9, 10]. Another source of collaboration came from Professor

L.M. Li
Academy of Mathematics and Systems Science, Chinese Academy of Sciences,
and Computational Biology and Bioinformatics, University of Southern California
e-mail: lilei@amss.ac.cn

S. Dudoit (ed.), *Selected Works of Terry Speed*, Selected Works in Probability and Statistics,
DOI 10.1007/978-1-4614-1347-9_13,

Richard Mathies' group in the chemistry department at Berkeley, who were conducting research on capillary DNA sequencing. In a statistical consulting service, of which Terry was in charge for the Statistics Department during one semester in 1995, Dr. Indu Kheterpal, who was a PhD graduate student in Professor Mathies' group, brought in an interesting estimation problem in fluorescence dye technique. Terry set up a good collaboration with them and I learned a lot of chemistry related to DNA sequencing through the interaction.

Sanger DNA sequencing generates a signal trace from each template DNA, and base-calling is the data analysis part of DNA sequencing, aimed at reconstructing the nucleotide sequence with a fair fidelity. We decomposed the problem into three parts: color correction, deconvolution, and base-calling. Then we tried to work out solutions to each of them. In my opinion, the work on color correction and deconvolution is mathematically and statistically more elegant and original, and we put a lot of effort into publishing it. In comparison, the solution to the last step of base-calling is more engineering-like in flavor. Terry introduced me to the technique of the hidden Markov model (HMM), which was not so widely known then as it is now. I was intrigued by the idea and we designed an HMM for base-calling. In genome research, a good idea alone is not sufficient to have an impact, and a good implementation is equally important, if not more so. The implementation of the HMM base-calling requires model-training and a lot of serious software programming. Due to graduation and my limited programming strength, I only tested the idea and did not develop a real software solution. A little later, Dr. Green and his team published their famous work on base-calling. In the meantime, microarray technology gradually caught people's attention. And our HMM base-calling idea was not pursued further [3].

By now our most influential contribution to DNA sequencing is color correction [4]. A few years ago, Terry told me that Solexa, now owned by Illumina, one major next generation sequencing platform, adopted our scheme. This is encouraging and yet not surprising because we have shown, at least in one important perspective, that the color correction scheme we proposed is optimal. In capillary Sanger sequencing, four dyes, which emit different colors as excited by laser, are used to distinguish four kinds of nucleotides. The purpose of color correction is to remove the cross-talk phenomenon of the four dyes' emission spectra. One key idea of our work is that we need to estimate the cross-talk phenomenon adaptively from each experiment. Another key idea of our estimation is that we make use of the "canonical" distribution of data without any cross-talk. As a PhD student, I was enthusiastic about the solution when it was first discovered. In a late afternoon, we walked home down Hearst Avenue, and Terry asked me a serious question, "how do we know our solution is right?" I gave him an answer, "If we estimate the cross-talk matrix properly, the distribution of the corrected data should match the nominal one." Terry agreed. After I graduated, I went through several interesting problems in engineering and science, and realized that they share a common nature with the color correction problem. I wrote an article about this class of blind inversion problems in the festschrift for Professor

Terry Speed's 60th birthday [6], because Terry's question partially inspired the formulation of this notion.

In usual DNA sequencing light intensities at four wavebands are measured, since four dyes are used. Interestingly, Dr. Kheterpal and Professor Mathies asked us if we could instead use only three light intensities for base-calling. After some struggles, we designed a procedure consisting of a series of nonnegative least squares and a model selection scheme [1]. Professor Mathies was very pleased with the result.

The work on deconvolution is also motivated by Sanger sequencing and is more technical than color correction. Each base in a Sanger sequencing trace can roughly be represented by a Gaussian-shaped peak on a continuous scale, and the four kinds of nucleotides, namely A, G, C and T, are represented by four different colors respectively. The motion of DNA molecules in capillary is usually explained by the reptation theory. The aggregation of the molecules of the same size can approximately be described by a Brownian motion. That explains why each peak looks like a normal distribution. In Sanger sequencing, most base-calling errors come from the regions with runs of the same kind of nucleotides, and lead to insertions and deletions, or simply indels. Once an error of this type occurs in base-calling, it often causes more trouble than a substitution error does in an alignment. How to separate these peaks, or in other words, how to count the bases in a run correctly, is a problem that we solved with the deconvolution technique.

The parametric deconvolution [5] was something we worked out without much prior literature knowledge on the topic. Terry suggested that we do a literature survey. In 1995, I searched the key word, deconvolution, on Yahoo (I am sure it was not Google then), and got over one thousand hits, and the early work went back to the nineteenth century. Obviously deconvolution is a common problem in many areas. I read almost all relevant papers I could find, and discussed them with Terry over a long period of time till 2000 in Melbourne. One issue that puzzled us was whether deconvolution is an ill-posed problem — a notion postulated by Hadamard in 1902. Without any constraint on the solution space, deconvolution is an ill-posed problem, and had been classified so in applied mathematics. Nevertheless, in many cases, the signals to be reconstructed are positive and "sparse". In parametric deconvolution, we formulate the unknowns by a mixture of finite Dirac spikes, and we can estimate them well in a regular sense, see Theorem 4.1 and 4.2 in Li and Speed [5], although the dimension of the solution space needs to be estimated too by model selection, see Algorithm 5.2 in Li and Speed [5] and Proposition 3.3 in Li [2]. Thus Terry and I came to the conclusion: if the signal to be reconstructed is positive and sparse, then deconvolution is well-posed.

The well-posedness explains why historically some nonparametric deconvolvers such as the Jansson's method and the folk iteration (5.2) in Li and Speed [7], obtained in different scenarios by either EM algorithms or Bayesian methods in the literature, work quite well in their respective applications. Furthermore, Terry and I did an investigation on the general linear inverse problem with positive constraints (LININPOS) that underlies the folk iteration. We discovered that the iteration in fact minimizes the Kullback-Leibler divergence between the target and the fit, and this result clarifies the core structure of the LININPOS solution.

The work described here has been a source of both enlightenment and enjoyment to me. When I was writing down these words, those scenes when Terry and I walked down Hearst Avenue and chatted on various issues came upon my mind like yesterday. I am sure that Terry's other students and colleagues had their own pleasant study and work experiences with him as well. His spirit is no doubt the source of many good things. In addition to his passion for science and mathematics, his respect of the interests and talents of each student, each collaborator and his own may partially explain his wide research spectrum.

References

[1] I. Kheterpal, L. Li, T. P. Speed, and R. A. Mathies. A three-color labeling approach for DNA sequencing using energy transfer primers and capillary electrophoresis. *Electrophoresis*, 19:1403–1414, 1999.

[2] L. Li. DNA sequencing and parametric deconvolution. *Stat. Sinica*, 12(1): 179–202, 2002.

[3] L. Li. *Statistical Models of DNA Base-calling*. PhD thesis, University of California, Berkeley, 1998.

[4] L. Li and T. P. Speed. An estimate of the crosstalk matrix in four-dye fluorescence-based DNA sequencing. *Electrophoresis*, 20(7):1433–1442, 1999.

[5] L. Li and T. P. Speed. Parametric deconvolution of positive spike trains. *Ann. Stat.*, 28(5):1279–1301, 2000.

[6] L. M. Li. Blind inversion needs distribution (BIND): General notation and case studies. In D. Goldstein, editor, *Science and Statistics: A Festschrift for Terry Speed*, volume 40 of *Lecture Notes – Monograph Series*, pages 275–293. Institute of Mathematical Statistics, Beachwood, OH, 2003.

[7] L. M. Li and T. P. Speed. Deconvolution of sparse positive spikes. *J. Comput. Graph. Stat.*, 13(4):853–870, 2004.

[8] D. O. Nelson. Improving DNA sequence accuracy and throughput. In T. P. Speed and M. S. Waterman, editors, *Genetic Mapping and DNA Sequencing*, volume 81 of *The IMA Volumes in Mathematics and its Applications*, pages 183–206. Springer, 1996.

[9] D. O. Nelson. Introduction of reptation. Technical report, Lawrence Livermore National Laboratory, 1995.

[10] D. O. Nelson and T. P. Speed. Recovering DNA sequences from electrophoresis data. In S. E. Levinson and L. Shepp, editors, *Image Models (and their Speech Model Cousins)*, pages 141–152. Springer-Verlag, 1996.

The Annals of Statistics
2000, Vol. 28, No. 5, 1279–1301

PARAMETRIC DECONVOLUTION OF POSITIVE SPIKE TRAINS

BY LEI LI[1] AND TERENCE P. SPEED[2]

Florida State University and University of California, Berkeley

This paper describes a parametric deconvolution method (PDPS) appropriate for a particular class of signals which we call spike-convolution models. These models arise when a sparse spike train—Dirac deltas according to our mathematical treatment—is convolved with a fixed point-spread function, and additive noise or measurement error is superimposed. We view deconvolution as an estimation problem, regarding the locations and heights of the underlying spikes, as well as the baseline and the measurement error variance as unknown parameters. Our estimation scheme consists of two parts: model fitting and model selection. To fit a spike-convolution model of a specific order, we estimate peak locations by trigonometric moments, and heights and the baseline by least squares. The model selection procedure has two stages. Its first stage is so designed that we expect a model of a somewhat larger order than the truth to be selected. In the second stage, the final model is obtained using backwards deletion. This results in not only an estimate of the model order, but also an estimate of peak locations and heights with much smaller bias and variation than that found in a direct trigonometric moment estimate. A more efficient maximum likelihood estimate can be calculated from these estimates using a Gauss–Newton algorithm. We also present some relevant results concerning the spectral structure of Toeplitz matrices which play a key role in the estimation. Finally, we illustrate the behavior of these estimates using simulated and real DNA sequencing data.

1. Introduction and background. Deconvolution is used in many scientific disciplines, including geophysics, spectroscopy, chromatography and pharmacokinetics. In abstract form, an unobserved signal $x(t)$ is blurred by a known point spread function w, resulting in an observed signal $y(t)$. Mathematically, this can be represented in terms of the convolution operation $*$: $y = w * x$. The task of deconvolution is to reconstruct the unobserved signal x from the observed signal y. The point spread function is assumed to be known throughout this paper. An additive measurement error is also assumed in y in all discrete situations discussed.

Our motivation is the problem of base-calling in DNA sequencing. The Sanger sequencing technique is a combination of enzymatic reactions, electrophoresis and fluorescence-based detection; see [1]. This procedure produces a four-component vector time series. Base-calling is the analysis part of DNA

Received June 1998; revised May 2000.
[1]Supported by NSF Grant DMS-99-71698.
[2]Supported by DOE Grant DE-FG03-97-ER62387.
AMS 1991 *subject classifications*. Primary 62F10; secondary 62F12, 86A22.
Key words and phrases. Deconvolution, spike train, model selection, DNA sequencing, Toeplitz matrix.

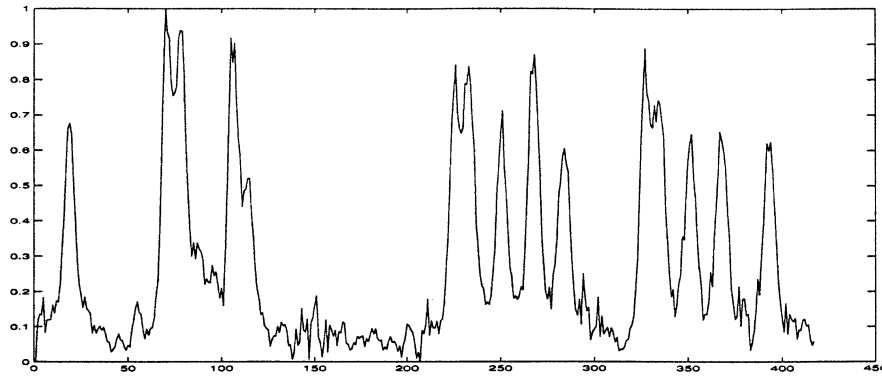

FIG. 1. *A segment slab gel electrophoresis sequencing data (provided by the engineering group at Lawrence Berkeley National Laboratory).*

sequencing, which attempts to recover the underlying DNA sequence from the vector time series. Figure 1 shows a segment of one channel of such a series. (This series is different from the original sequencing data because of the "cross talk" phenomenon. We here pass over this issue, and refer to [21] for details.) Typically, each major peak in the series corresponds to one base. As sequencing progresses, electrophoretic diffusion spreads peaks more and more. In regions where there are multiple occurrences of the same base, several successive peaks may merge into one large block. In this situation, base-calling is far more difficult. A number of studies exist of the errors made by one widely used base-calling system, see [3, 17, 18]. These reports show that errors associated with runs of the same base constitute more than half of the total errors. Furthermore, this kind of error causes more serious difficulties in later analysis than other kinds. For this reason, it seems important to try and do better in resolving peaks, and this is the motivation of the research we report here.

The literature on deconvolution is rich and scattered across a wide variety of fields. Frequently, deconvolution is an ill-posed inverse problem; see [13]. Two techniques, regularization and exploiting bound or nonnegativity constraints on the unknown functions, have been proved to be useful in dealing with ill-posedness. Regularization was introduced by Tikhonov [30], and has been well studied since then. For example, the long-standing iterative deconvolution method of Van Cittert (see [13]) can be viewed as a regularization method. Jansson [13] adjusted this regularization method, added nonnegativity constraints to the unknown function and applied it successfully to problems in spectroscopy.

Maximum entropy deconvolution can also be regarded as a regularization procedure, one which only applies to nonnegative signals; see [10, 11]. Donoho, Johnstone, Hoch and Stern [5] showed this procedure has certain advantages such as signal-to-noise enhancement and superresolution when

the signal is nearly black. Stark and Parker [29] proposed some new algorithms to solve this type of constrained minimization problems.

The deblurring method introduced by Shepp and Vardi [27], and Vardi and Lee [32] uses maximum likelihood estimation under a Poisson or a multinomial model. Again it is assumed that the unknown function is nonnegative. Snyder, Schulz and O'Sullivan [28] obtained a similar algorithm as a solution to a general Fredholm integral equation of the first kind. They derive the formula by minimizing Csiszár's I-divergence, which is closely related to the concept of likelihood. Richardson [26], Kennett, Prestwich and Robertson [14, 15, 16] and Di Gesù and Maccarone [4] obtained the same result from a more intuitive Bayesian point of view and term it as "Bayesian deconvolution." All of these methods could be used in the base-calling problem, and their behavior on DNA sequencing data can be found in [20].

Poskitt, Dogancay and Chung [25] described a double-blind deconvolution method to analyze postsynaptic currents in nerve cells. This analysis is based on an elegant statistical model, and the estimator is derived by minimizing the quasi-profile-likelihood. Though the data from postsynaptic currents in nerve cells are different from DNA seqnencing data, we find that the likelihood structures of the two models proposed for each problem have some similarities. Therefore, we can borrow some ideas from [25] to study the deconvolution problem in DNA sequencing.

Motivated by the DNA sequencing data, we define what we call the spike-convolution model, in which the unknown function is represented as a linear combination of a finite number of positive spikes (Dirac functions) together with a constant baseline. In this model, deconvolution is nothing but a standard parameter estimation problem, where the parameters include the number, locations and heights of the underlying spikes, the baseline and the measurement error variance. Our estimation procedure uses the spectral properties of Toeplitz matrices, least squares and statistical model selection techniques, and we call it parametric deconvolution of positive spikes (PDPS).

The paper is arranged as follows. In Section 2 we introduce the spike-convolution model and some notation, discuss several aspects such as identifiability, and outline the estimation procedure. In Section 3 we study spectral structures of Toeplitz matrices constructed from trigonometric moments and present algorithms and asymptotics of the trigonometric moments estimates. In Section 4 we describe the algorithms and asymptotics of the maximum likelihood estimates. In Section 5 we propose our ultimate method, which is a package including a model fitting procedure and a two-stage model selection procedure. Section 6 contains a simulation study of the proposed methodologies. Finally, we apply PDPS to real DNA sequencing data, and compare the result with that using another nonparametric deconvolution method. Several other practical issues of PDPS are also discussed. The Appendix contains the proofs and the relevant mathematical details.

2. The spike-convolution model. Throughout the paper, the point spread function $w(\cdot)$ is assumed to be known, and to satisfy the following conditions, where $v_k = (1/2\pi)\int_{-\pi}^{\pi} w(t)e^{ikt}\,dt$ are w's Fourier coefficients:

1. It has finite support $(-\kappa_1,\ \kappa_2)$, where $0 < \kappa_1,\ \kappa_2 < \pi$;
2. $w(\cdot) \in C^2[-\pi,\ \pi]$;
3. $v_k \neq 0$, for $k = 0, \pm 1, \ldots, \pm K_0$.

The last assumption is described as requiring that there be no hole in the Fourier transform of the point spread function. It worth mentioning that $w(\cdot)$ is not necessarily nonnegative or causal. There exist studies of the shape of the point spread function in DNA sequencing (see [23] for references), and the determination of the width of the point spread function is another matter worthy of attention. However, these issues are not critical to the present work, and so we pass over them here. The signal to be estimated is assumed to have the following form:

$$x(t) = A_0 + \sum_{j=1}^{p} A_j \delta(t - \tau_j), \tag{1}$$

where $\delta(\cdot)$ is the Dirac delta function. We assume the coefficients A_j of the Dirac functions are positive and refer to them as "heights" of the spikes. Thus the underlying signal $x(t)$ is a linear combination of a finite number of spikes with positive heights, together with a constant baseline. We also assume $-\pi + \kappa_1 < \tau_1 < \cdots < \tau_p < \pi - \kappa_2$. Hence the support of the convolution y of w and x will stay in the range $[-\pi,\ \pi]$. Explicitly, we have

$$y(t) = A_0 + \sum_{j=1}^{p} A_j w(t - \tau_j) = (x * w)(t), \qquad t \in [-\pi,\ \pi], \tag{2}$$

where the time range has been scaled to $[-\pi,\ \pi]$ for convenience. We have assumed there are no peaks near the two ends. In real DNA sequencing, we can always cut the raw data into pieces at valley points and apply the deconvolution to each piece separately. The observations $\{z(t_l)\}$ are a sample of the above model, corrupted by measurement errors which are assumed to be additive:

$$z(t_l) = y(t_l) + \varepsilon(t_l) = A_0 + \sum_{j=1}^{p} A_j w(t_l - \tau_j) + \varepsilon(t_l), \tag{3}$$

where $t_l = 2\pi l/n$, $l = -[n/2], \ldots, 0, \ldots, [n/2] - 1$ if n is even, or $l = -[n/2], \ldots, 0, \ldots, [n/2]$ if n is odd. The $\{\varepsilon(t_l)\}$ are supposed to be i.i.d. with $E(\varepsilon(t_l)) = 0, \mathrm{Var}(\varepsilon(t_l)) = \sigma^2$ and a finite third moment.

Before we proceed, we introduce some notation. We denote the signal in the spike-convolution model (2) by $SC(w; p; \mathbf{A}; \tau)$ where $\mathbf{A}$ and τ, respectively, represent $\{A_j\}$ and $\{\tau_j\}$, and formally denote the signal x in (1) by $SC(\delta; p; \mathbf{A}; \tau)$. We define the inner product of two functions $y_1(t)$ and $y_2(t)$, belonging to $L^2[-\pi, \pi]$, by $\langle y_1, y_2 \rangle = \frac{1}{2\pi}\int_{-\pi}^{\pi} y_1(t)y_2(t)\,dt$. For functions $z_1(t)$ and $z_2(t)$

well defined at the lattice points $t_l = 2\pi l/n$, we also define the following inner product:

$$\langle z_1, z_2 \rangle_n = \begin{cases} \dfrac{1}{n} \displaystyle\sum_{l=-[n/2]}^{[n/2]} z_1(t_l) z_2(t_l), & \text{if } n \text{ is odd,} \\ \dfrac{1}{n} \displaystyle\sum_{l=-[n/2]}^{[n/2]-1} z_1(t_l) z_2(t_l), & \text{if } n \text{ is even.} \end{cases}$$

The norms induced by $\langle \cdot, \cdot \rangle$ and $\langle \cdot, \cdot \rangle_n$ are denoted by $\|\cdot\|$ and $\|\cdot\|_n$, respectively.

The Hilbert–Schmidt theory for Fredholm integral equations of the first kind assumes that the signal $x(t)$ is in $L^2[\pi, \pi]$, and thus excludes the Dirac functions. Consequently, the signal reconstructed by methods within that framework will not contain any Dirac functions. This means that only incomplete deconvolution would be achieved if (2) were the truth. We prefer to regard deconvolution as a problem of parameter estimation, and this can result in a complete deconvolution.

The parameters in a $SC(w; p; \mathbf{A}; \tau)$ model include the number, locations and heights of the peaks, and the baseline. Although these parameters are closely related, they play quite different roles from the perspective of statistical estimation, and it seems very difficult to estimate them all, simultaneously and efficiently, in one step. The difficulty lies in the irregular structure of the the parameter space. For example, suppose we have a spike-convolution model with three positive peaks. If we let the height of one peak tend to zero, then the limiting model can only be regarded as a spike-convolution model with two peaks, for we require peak heights to be positive. Following this limiting process, the dimension of the parameter space changes. However, we have the following result saying that a spike-convolution model of order p cannot arbitrarily well be approximated by models of smaller orders.

THEOREM 2.1. *Let $y(t)$ be a $SC(w; p; \mathbf{A}; \tau)$ model. Then we have: $\inf_{\bar{y}} \|y - \bar{y}\| = d > 0$, where the infimum is taken over all $\bar{y} \in SC(w; l; \bar{\mathbf{A}}; \bar{\tau})$, $l < p$.*

A relating problem in signal processing is to estimate the so-called hidden frequencies ω_j in the following model:

$$s_t = \sum_{j=1}^{p} A_j \cos(t\omega_j + \theta_j) + \varepsilon_t, \tag{4}$$

where the A_j are positive and ε_t is white noise. In fact, the spectral density function of the above process, in the generalized sense, is a constant plus jumps occurring at the ω_j, with corresponding heights A_j. This is exactly the signal to be reconstructed in Model (2). Apart from the point spread function, the two problems are Fourier duals of each other. In order to estimate the hidden frequencies, Pisarenko [24] suggested a method using the Toeplitz matrices constructed from autocovariance functions. We exploit a similar idea as part of our estimation procedure.

Our estimation procedure consists of several steps. In Algorithm 3.1, we estimate the peak locations by connecting deconvolution with the spectral structure of Toeplitz matrices constructed from the Fourier coefficients of the observations. The peak heights of the estimated locations can be estimated either by a trigonometric moment method, Algorithm 3.2, or by least squares, where the latter exploits the connection between spike-convolution models and hypothetical linear regressions; see Algorithm 5.1. This leaves us with the task of estimating the number of peaks, usually regarded as a model selection problem. In our method, models of each candidate order have to be fitted before model selection takes place. The model selection strategy described in Algorithm 5.2 consists of two stages. First we choose a model which should come close to including the true model as a submodel, as overfitting is not completely suppressed at this step. We then use a modified GIC criterion together with a backward deletion procedure to obtain our final model. The resulting estimate can further be tuned by an optional step: maximizing likelihood if the distribution of the measurement errors is assumed known; see Section 4. Under the assumption of normal errors, we calculate the Fisher information matrix of the spike-convolution model, whose inverse gives the nominal standard errors of estimates. Note that these standard errors will not have taken into account the model selection process. The computation of the maximum likelihood estimate or one-step estimate can be carried out by Gauss–Newton algorithm. Please keep in mind that Sections 3 and 4 are about inference given the model order m, the number of spikes included in the spike-convolution model. This is a brief description of PDPS, which will be explained more fully in the rest of the paper.

3. The trigonometric moment estimates. Throughout this section, we assume the model order m is given. The connection between the trigonometric moments and the parameters in a spike-convolution model can be given using the spectral structure of Toeplitz matrices as follows.

THEOREM 3.1. *Let the Fourier coefficients of $y(t)$, which follows $SC(w; m; \mathbf{A}; \tau)$ with $m \le K_0$, be $f_k = \langle y(t), e^{ikt} \rangle$, for $k = 0, \pm 1, \ldots$. First, write $g_0 = f_0$ and $g_k = f_k v_0 / v_k$, $g_{-k} = \bar{g}_k$, for $0 < k \le K_0$. Next, form the Toeplitz matrices $G_m = (g_{j-1})_{i,\, j=0,\ldots,m}$. Finally, write $U^{(m)}(z) = \prod_{j=1}^{m} (z - \exp(i\tau_j)) = \sum_{j=0}^{m} \alpha_j z^j$. Then we have*

(i)

$$
(5) \qquad \begin{cases} f_0 = A_0 + \left(\sum_{j=1}^{m} A_j \right) v_0, \\ f_k = \left(\sum_{j=1}^{m} A_j e^{ik\tau_j} \right) v_k, \qquad k \neq 0. \end{cases}
$$

(ii) *A_0 is the smallest eigenvalue of G_m with multiplicity one and eigenvector $(\alpha_0, \ldots, \alpha_m)^T$.*

PARAMETRIC DECONVOLUTION

(iii) *The* $\{A_j\}$ *satisfy the following linear system:*

$$(6)\qquad v_0 \begin{pmatrix} 1 & 1 & \cdots & 1 \\ \exp(i\tau_1) & \exp(i\tau_2) & \cdots & \exp(i\tau_m) \\ \vdots & \vdots & \ddots & \vdots \\ \exp(i(m-1)\tau_1) & \exp(i(m-1)\tau_2) & \cdots & \exp(i(m-1)\tau_m) \end{pmatrix} \begin{pmatrix} A_1 \\ A_2 \\ \vdots \\ A_m \end{pmatrix} = \begin{pmatrix} g_0 - A_0 \\ g_1 \\ \vdots \\ g_{m-1} \end{pmatrix}$$

Note. When a Toeplitz matrix has distinct eigenvalues, the relation between eigenvalues and eigenvectors is one-to-one. For simplicity, we will use expressions such as "smallest eigenvector" to refer to the eigenvector corresponding to the smallest eigenvalue. The converse of the above theorem is also true in the following sense.

THEOREM 3.2. *Suppose we are given* $2m+1$ *complex numbers* $\{f_j, -m \le j \le m\}$, *where* $m \le K_0$, $\bar{f}_j = f_{-j}$. *Let* $g_0 = f_0$, *and* $g_k = f_k v_0/v_k$, $g_{-k} = \bar{g}_k$, *for* $0 < k \le K_0$. *Assume that the smallest eigenvalue* A_0 *of the Toeplitz matrix* $G_m = (g_{j-i})_{i,\, j=0,\ldots,m}$ *has multiplicity* 1. *Let the smallest eigenvector be* $\alpha = (\alpha_0, \ldots, \alpha_m)^T$, *and* $U^{(m)}(z) = \sum_{j=0}^{m} \alpha_j z^j$. *Then*:

(i) $U^{(m)}(z)$ *has m distinct roots exactly on the unit circle, which are denoted by* $\{\exp(i\tau_j)\}$.

(ii) *Furthermore, if* $-\pi + \kappa_1 < \tau_1 < \cdots < \tau_m < \pi - \kappa_2$, *then there exists a* $SC(w; m; \mathbf{A}; \tau)$ *whose first* $m+1$ *Fourier coefficients are* $\{f_j, 0 \le j \le m\}$. *Its baseline and heights* $\{A_j\}$ *are determined by the linear system (6), and the resulting heights are positive.*

This result is of great significance for practical model fitting from the computational point of view, since the peak locations could be found by restricting the search of roots of $U^{(m)}(z)$ to the unit circle. Starting from data $z(t_l)$, we estimate the trigonometric moments by $\hat{f}_k = \langle z, e^{ikt} \rangle_n$. Based on Theorem 3.2, for any given nonnegative integer $m \le K_0$, we can input these empirical trigonometric moments into the following two algorithms.

ALGORITHM 3.1 (Trigonometric moment estimates of peak locations).

(i) *Deconvolution: let* $\hat{g}_0 = \hat{f}_0$, $\hat{g}_k = \hat{f}_k v_0/v_k$, *for* $k = \pm 1, \ldots, \pm m$.

(ii) *Computing the smallest eigenvalue-vector of the Toeplitz matrix: construct the Toeplitz matrix* $\widehat{G}_m = (\hat{g}_{j-i})$ *and compute its smallest eigenvalue* $\widehat{A}_0^{(m)}$ *(assuming its multiplicity is one) and corresponding eigenvector* $\hat{\alpha}^{(m)} = (\hat{\alpha}_0^{(m)}, \ldots, \hat{\alpha}_m^{(m)})$.

(iii) *Solving a polynomial: on the unit circle, find the m distinct roots of* $\widehat{U}^{(m)}(z) = \sum_{j=0}^{m} \hat{\alpha}_j^{(m)} z^j$, *which we denote by* $\{\exp(i\hat{\tau}^{(m)})\}$, $j = 1, \ldots, m$.

ALGORITHM 3.2 (Trigonometric moment estimates of heights). *Solve the following Vandermonde linear system:*

$$v_0 \begin{pmatrix} 1 & 1 & \cdots & 1 \\ \exp(i\hat{\tau}_1^{(m)}) & \exp(i\hat{\tau}_2^{(m)}) & \cdots & \exp(i\hat{\tau}_m^{(m)}) \\ \vdots & \vdots & \ddots & \vdots \\ \exp(i(m-1)\hat{\tau}_1^{(m)}) & \exp(i(m-1)\hat{\tau}_2^{(m)}) & \cdots & \exp(i(m-1)\hat{\tau}_m^{(m)}) \end{pmatrix} \begin{pmatrix} \widehat{A}_1^{(m)} \\ \widehat{A}_2^{(m)} \\ \vdots \\ \widehat{A}_m^{(m)} \end{pmatrix} = \begin{pmatrix} \hat{g}_0 - \widehat{A}_0^{(m)} \\ \hat{g}_1 \\ \vdots \\ \hat{g}_{m-1} \end{pmatrix}.$$

The output of these two algorithms is a $SC(w; m; \widehat{\mathbf{A}}^{(m)}; \hat{\tau}^{(m)})$ whose first $m+1$ Fourier coefficients are $\hat{f}_k$, $k = 0, \ldots, m$. We make some remarks here. First, we have ignored the case when the multiplicity of A_0 is greater than 1, since the Lebesgue measure of this singular case is zero. Second, strictly speaking, the fitted model makes sense only when $-\pi + \kappa_1 < \hat{\tau}_1^{(m)} < \cdots < \hat{\tau}_m^{(m)} < \pi - \kappa_2$. We thus delete those peaks outside the legitimate range in the regression stage discussed later, and then estimate the heights of the remaining peaks by least squares; see Algorithm 5.1 for more details. Third, our numerical experiments carried out in MATLAB show these two algorithms are robust to round off and noise in the data. For example, the roots of $\widehat{U}^{(m)}(z)$ do indeed lie on the unit circle to the necessary accuracy. Finally, in the case that the observations are generated from a $SC(w; p; \mathbf{A}; \tau)$, if we take $m = p$, then $\hat{\tau}_j$ and $\widehat{A}_j$ are the trigonometric moment estimates, which are consistent. Indeed, we have the following central limit theorem.

THEOREM 3.3.

$$\begin{aligned} &\sqrt{n}\big[(\widehat{A}_0, \widehat{A}_1, \ldots, \widehat{A}_p, \hat{\tau}_1, \ldots, \hat{\tau}_p)^T - (A_0, A_1, \ldots, A_p, \tau_1, \ldots, \tau_p)^T\big] \\ &\quad \xrightarrow{d} N(0, V), \end{aligned} \tag{7}$$

where $V = 4\pi^2 Q^{-1} P Q^{-T}$,

$$P = \frac{\sigma^2}{2} \operatorname{diag}\left(2, \left|\frac{v_0}{v_1}\right|^2, \ldots, \left|\frac{v_0}{v_p}\right|^2, \left|\frac{v_0}{v_1}\right|^2, \ldots, \left|\frac{v_0}{v_p}\right|^2\right) \tag{8}$$

and

$$
(9)\qquad Q=\left(\begin{array}{c|cccc|cccc}
1 & 1 & 1 & \cdots & 1 & 0 & 0 & \cdots & 0 \\
\hline
0 & \cos\tau_1 & \cos\tau_2 & \cdots & \cos\tau_p & -A_1\sin\tau_1 & -A_2\sin\tau_2 & \cdots & -A_p\sin\tau_p \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\
0 & \cos p\tau_1 & \cos p\tau_2 & \cdots & \cos p\tau_p & -pA_1\sin p\tau_1 & -pA_2\sin p\tau_2 & \cdots & -pA_p\sin p\tau_p \\
\hline
0 & \sin\tau_1 & \sin\tau_2 & \cdots & \sin\tau_p & A_1\cos\tau_1 & A_2\cos\tau_2 & \cdots & A_p\cos\tau_p \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\
0 & \sin p\tau_1 & \sin p\tau_2 & \cdots & \sin p\tau_p & pA_1\cos p\tau_1 & pA_2\cos p\tau_2 & \cdots & pA_p\cos p\tau_p
\end{array}\right)
$$

More algebra shows that the asymptotic variances of the $\{A_j\}$ depend only on the $\{\tau_j\}$, while the asymptotic variances of the $\{\tau_j\}$ depend not only on the configuration of the $\{\tau_j\}$, but also on the heights $\{A_j\}$. In fact, if we define A/σ to be the local signal-to-noise ratio, then the asymptotic standard deviation of $\hat{\tau}_j$ is proportional to the reciprocal of its local signal-to-noise ratio.

4. The maximum likelihood estimates. Throughout this section, we assume the model order m is given. In general, trigonometric moment estimates are not as efficient as maximum likelihood estimates. However, starting from trigonometric moment estimates, which are $\sqrt{n}$-consistent, we can construct one-step estimates or find maximum likelihood estimates using Fisher scoring. In either case, we need to specify the error distribution to calculate the Fisher information matrix. Under the assumption of normal errors, the $-2\,loglikelihood$ of the observations generated from Model (3) is

$$
(10)\qquad n\log(2\pi\sigma^2)+\frac{1}{\sigma^2}\sum_l\left\{z(t_l)-A_0-\sum_{j=1}^{p}A_j\,w(t_l-\tau_j)\right\}^2.
$$

More notation is needed in this section. We write $\theta=(A_0,A_1,\ldots,A_p,\tau_1,\ldots,\tau_p)^T$, and sometimes we use $y_\theta(t)$ to denote $SC(w;p;\mathbf{A};\tau)$. Denote the gradient vector by $\nabla_\theta=(\partial\log L/\partial\theta)^T$, and $\nabla_{(\theta,\sigma^2)}=(\nabla_\theta^T,\partial\log L/\partial\sigma^2)^T$. As usual, the Fisher information matrix is defined by $I_{(\theta,\sigma^2)}=\frac{1}{n}E[\nabla_{(\theta,\sigma^2)}\nabla_{(\theta,\sigma^2)}^T]$ and $I_\theta=\frac{1}{n}E[\nabla_\theta\nabla_\theta^T]$.

PROPOSITION 4.1. *Let* $\Psi_\theta=(\psi_{A_0},\psi_{A_1},\ldots,\psi_{A_p},\psi_{\tau_1},\ldots,\psi_{\tau_p})^T$, *where* $\psi_{A_0}=1$, $\psi_{A_j}=w(t-\tau_j)$, $\psi_{\tau_j}=-A_jw'(t-t_j)$; $j=1,\ldots,p$. *Then*

$$
(11)\qquad I_{(\theta,\sigma^2)}=\begin{pmatrix} I_\theta & 0_p^T \\ 0_p & \frac{1}{2\sigma^4}\end{pmatrix},
$$

where 0_p *is a vector with* p *zeros, and*

$$
(12)\qquad I_\theta=\frac{1}{\sigma^2}\langle\Psi_\theta,\Psi_\theta^T\rangle_n\to\frac{1}{\sigma^2}\langle\Psi_\theta,\Psi_\theta^T\rangle.
$$

Here $\langle \Psi_\theta, \Psi_\theta^T \rangle$ *is defined by*

$$\left(\begin{array}{c|ccc|ccc} \langle \psi_{A_0}, \psi_{A_0} \rangle & \langle \psi_{A_0}, \psi_{A_1} \rangle & \cdots & \langle \psi_{A_0}, \psi_{A_p} \rangle & \langle \psi_{A_0}, \psi_{\tau_1} \rangle & \cdots & \langle \psi_{A_0}, \psi_{\tau_p} \rangle \\ \hline \langle \psi_{A_1}, \psi_{A_0} \rangle & \langle \psi_{A_1}, \psi_{A_1} \rangle & \cdots & \langle \psi_{A_1}, \psi_{A_p} \rangle & \langle \psi_{A_1}, \psi_{\tau_1} \rangle & \cdots & \langle \psi_{A_1}, \psi_{\tau_p} \rangle \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \langle \psi_{A_p}, \psi_{A_0} \rangle & \langle \psi_{A_p}, \psi_{A_1} \rangle & \cdots & \langle \psi_{A_p}, \psi_{A_p} \rangle & \langle \psi_{A_p}, \psi_{\tau_1} \rangle & \cdots & \langle \psi_{\tau_1}, \psi_{\tau_p} \rangle \\ \hline \langle \psi_{\tau_1}, \psi_{A_0} \rangle & \langle \psi_{\tau_1}, \psi_{A_1} \rangle & \cdots & \langle \psi_{\tau_1}, \psi_{A_p} \rangle & \langle \psi_{\tau_1}, \psi_{\tau_1} \rangle & \cdots & \langle \psi_{\tau_1}, \psi_{\tau_p} \rangle \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \langle \psi_{\tau_p}, \psi_{A_0} \rangle & \langle \psi_{\tau_p}, \psi_{A_1} \rangle & \cdots & \langle \psi_{\tau_p}, \psi_{A_p} \rangle & \langle \psi_{\tau_p}, \psi_{\tau_1} \rangle & \cdots & \langle \psi_{\tau_p}, \psi_{\tau_p} \rangle \end{array}\right),$$

and similarly for $\langle \Psi_\theta, \Psi_\theta^T \rangle_n$. *Using a similar notation, we compute the gradient vector as follows:*

$$\begin{aligned} \frac{1}{n}\nabla_\theta &= \frac{1}{\sigma^2}\langle \Psi_\theta, \varepsilon \rangle_n \\ &= \frac{1}{\sigma^2}(\langle \psi_{A_0}, \varepsilon_\theta \rangle_n, \langle \psi_{A_1}, \varepsilon_\theta \rangle_n, \ldots, \langle \psi_{A_p}, \varepsilon_\theta \rangle_n, \langle \psi_{\tau_1}, \varepsilon_\theta \rangle_n, \ldots, \langle \psi_{\tau_p}, \varepsilon_\theta \rangle_n)^T, \end{aligned}$$

where $\varepsilon_\theta(t) = z(t) - y_\theta(t)$.

Just as with i.i.d. observations, the MLEs are both consistent and asymptotically efficient under the assumption of normal errors.

THEOREM 4.1. *The maximum likelihood estimates are consistent; indeed, as* $n \to \infty$, *we have*

$$\sqrt{n}[(\tilde{\theta}, \tilde{\sigma}^2)^T - (\theta, \sigma^2)^T] \xrightarrow{d} N(0, I^{-1}_{(\theta, \sigma^2)}).$$

In order to compute the maximum likelihood estimate, we can use Fisher scoring as follows, taking the trigonometric moment estimates as the starting value.

$$\begin{pmatrix} \theta_{\text{new}} \\ \sigma^2_{\text{new}} \end{pmatrix} = \begin{pmatrix} \theta_{\text{old}} \\ \sigma^2_{\text{old}} \end{pmatrix} + \frac{1}{n} I^{-1}_{(\theta, \sigma^2)} \nabla_{(\theta, \sigma^2)}|_{(\theta_{\text{old}}, \sigma^2_{\text{old}})}.$$

Because of the orthogonality of θ and σ^2, we can first use Fisher scoring method to improve the estimate of θ. This leads to the well-known Gauss–Newton algorithm.

ALGORITHM 4.1 (Gauss–Newton).

(i) *Let* θ_{old} *be a* $\sqrt{n}$*-consistent estimate of* θ.

(ii) *Calculate* θ_{new} *by the following:*

$$\begin{aligned} \theta_{\text{new}} - \theta_{\text{old}} &= \frac{1}{n} I_\theta^{-1} \nabla_\theta|_{\theta_{\text{old}}} = [\langle \Psi_{\theta_{\text{old}}}, \Psi^T_{\theta_{\text{old}}} \rangle_n]^{-1} \langle \Psi_{\theta_{\text{old}}}, z - y_{\theta_{\text{old}}} \rangle_n \\ &= \left[\sum_l \Psi_{\theta_{\text{old}}}(t_l) \Psi_{\theta_{\text{old}}}(t_l)^T\right]^{-1} \left[\sum_l \Psi_{\theta_{\text{old}}}(t_l)(z(t_l) - y_{\theta_{\text{old}}}(t_l))\right]. \end{aligned} \tag{13}$$

Although we can iterate the above procedure, the following result shows one step is enough for the consideration of efficiency.

THEOREM 4.2.

$$\sqrt{n}(\theta_{\text{new}} - \theta)^T \xrightarrow{d} N(0, I_\theta^{-1}).$$

Finally, σ^2 can be estimated as

$$\sigma^2_{\text{new}} = \left\| z(t) - A_{0,\text{new}} - \sum_{j=1}^{p} A_{j,\text{new}} w(t - \tau_{j,\text{new}}) \right\|_n^2.$$

5. Hypothetical regressions and model selection. In this section, we deal with the case of m unknown.

5.1. *The least squares estimates and model fitting.* Because of the linear dependence of the signal on the baseline and heights in a $SC(w; p; \mathbf{A}; \tau)$, once the number and locations of the peaks are obtained, we can use least squares to estimate the baseline and peak heights. Combining this idea with Algorithm 3.1, we are led to the following model fitting algorithm.

ALGORITHM 5.1 (Model-fitting). *Starting with the empirical trigonometric moments $\hat{f}_k$ for any given nonnegative integer $m \le K_0$:*

(i) *Use the method of trigonometric moments to estimate the peak locations $\{\bar{\tau}_j\}$ using Algorithm* 3.1.

(ii) *Eliminate those peaks falling outside $[-\pi + \kappa_1, \pi - \kappa_2]$, and denote the locations of the remaining peaks by $\{\bar{\tau}_j, j = 1, \ldots, \bar{m}\}$, where $\bar{m} \le m$.*

(iii) *Estimate the heights $\overline{A}_j$ corresponding to these peaks by minimizing the following:*

$$\left\| z(t) - \overline{A}_0 - \sum_{j=1}^{\bar{m}} \overline{A}_j w(t - \bar{\tau}_j) \right\|_n^2. \tag{14}$$

This results in the least squares estimates of the baseline and heights.

The output of this algorithm is a $SC(w; \bar{m}; \overline{\mathbf{A}}^{(\bar{m})}; \bar{\tau}^{(\bar{m})})$. Next the problem of model selection arises because we can fit a spike-convolution model for each nonnegative integer in a given range.

5.2. *Model selection.* Suppose that the data is generated from a $SC(w; p; \mathbf{A}; \tau)$, and noise is added. In light of Theorem 2.1, models $SC(w; \bar{m}; \overline{\mathbf{A}}^{(\bar{m})}; \bar{\tau}^{(\bar{m})})$ with $\bar{m} < p$ are not good. When $\bar{m} \ge p$, we might expect a subset of $\{\bar{\tau}_j\}$ will be close to the real peak locations $\{\tau_j\}$. This is the basis of our model selection procedure, whose motivation will be clearer following the next result.

PROPOSITION 5.1. *Suppose that a Toeplitz matrix G_m has smallest eigenvalue A_0 with multiplicity $r > 1$.*

(i) *Let $m-r+1 = p$. Then the Toeplitz submatrix G_p has the same smallest eigenvalue A_0 with multiplicity* 1.

(ii) *Let $\{e^{i\tau_j}, j = 1, \ldots, p\}$ be the roots corresponding to the eigenvector of G_p associated with the eigenvector A_0. Then any polynomial whose coefficients form an eigenvector in the invariant space of G_m corresponding to A_0 has $\{e^{i\tau_j}, j = 1, \ldots, p\}$ as a subset of its roots.*

Let us consider Algorithm 3.1 in light of this result. Supposing that $m \geq p$, we explain the situation from two perspectives. From the computational perspective, in the absence of errors, any polynomial (not unique) whose coefficients form an eigenvector in the invariant space of G_m corresponding to the smallest eigenvalue always has $\{e^{i\tau_j}, j = 1, \ldots, p\}$ as a subset of its roots, where τ_j are the real peak locations. But in general it is not true that all its zeros are on the unit circle. In the presence of errors, Theorem 3.2 implies that the multiplicity of the smallest eigenvalue is 1 (apart from a set of Lebesgue measure zero), and the resulting (unique) polynomial has all its roots on the unit circle. This attractive feature, in the computational sense, is the "positive" aspect of noise. From the perspective of spectral structures, the situation here has a close connection to the perturbation theory of matrices, see [8]. In the absence of errors, the distance between the invariant space $\mathcal{S}$ of G_m corresponding to the smallest eigenvalue and other invariant spaces is positive. In other words, it is well separated from other invariant spaces. In the presence of errors and when the sample size is large enough, on the one hand, the last $m - p + 1$ eigenvalues could be close to each other, and their eigenvectors could be "wobbly" (see [8]) under small perturbation of the noise. On the other hand the invariant space $\widehat{\mathcal{S}}$ defined by these possibly "wobbly" eigenvectors is stable. This is true because when the sample size is large enough, $\widehat{\mathcal{S}}$ is close to $\mathcal{S}$, which is well separated from other invariant spaces. Therefore we expect a subset of the roots obtained from Algorithm 3.1 to be close to $\{e^{i\tau_j}, j = 1, \ldots, p\}$ when the sample size is large enough. It is so because the eigenvector from which these roots are obtained belongs to $\widehat{\mathcal{S}}$, and is close to one eigenvector belonging to $\mathcal{S}$ and having the property as shown in Proposition 5.1. Now let us return to model fitting. The above argument means the regressors $1, w(t - \bar{\tau}_1), \ldots, w(t - \bar{\tau}_{\bar{m}})$ will include a subset of "explanatory variables" close to the true regressors $1, w(t - \tau_1), \ldots, w(t - \tau_p)$. For large enough n we can therefore expect model selection criteria to behave as they do in the context of variable selection in regression. Our model selection procedure has two stages. We assume the model order has an upper bound $M(\leq K_0)$.

ALGORITHM 5.2 (Two-stage model selection). (i) *First stage. Among all the $SC(w; \bar{m}; \bar{\mathbf{A}}^{(\bar{m})}; \bar{\tau}^{(\bar{m})})$ models fitted by Algorithm 5.1, choose the one that*

minimizes the following:

$$MGIC_1(r) = \bar{\sigma}(r)^2 + \frac{c_1(n)\log n}{n} r, \tag{15}$$

where $\bar{\sigma}(r)^2$ *is the quantity in* (14)*, and* $c_1(n) \geq 0$ *is a penalty coefficient. Denote this model by* $SC(w; \bar{m}_0; \overline{\mathbf{A}}^{(\bar{m}_0)}; \bar{\tau}^{(\bar{m}_0)})$.

(ii) *Second stage. We regard the model selected in the first stage as a hypothetical regression model, and use a backward deletion procedure to select the final model. That is, starting from* $SC(w; \bar{m}_0; \overline{\mathbf{A}}^{(\bar{m}_0)}; \bar{\tau}^{(\bar{m}_0)})$*, we delete the peak that is least significant in terms of sum of squares. Compare the two models according to the following statistic*:

$$MGIC_2(r) = \check{\sigma}(r)^2 + \frac{c_2(n)\log n}{n} r, \tag{16}$$

where $\check{\sigma}^{(r)2}$ *is the sum of squares fitted by a model with r peaks, and* $c_2(n) > 0$ *is another penalty coefficient possibly depending on n. Choose the one that minimizes* $MGIC_2$*. If one peak can be deleted according to this criterion, then we iterate this procedure until we cannot delete any more peaks.*

We make some remarks about this procedure. First, existing model selection procedures such as AIC and BIC cannot be applied here, for the parameter estimates are not maximum likelihood ones. Second, the penalty term $c_1(n)$ is used in the first stage to compare all models obtained from Algorithm 5.1, and overfitting is not suppressed but encouraged to some extent. In fact, we would like to find the "best overfitting" model in this stage. The penalty term $c_2(n)$ is used in the second stage to eliminate those false peaks in the model obtained in the first stage. This suggests that we impose another restriction $c_1(n) < c_2(n)$. Third, the purpose of this two-stage model selection procedure is not only estimating the model order, but also producing a parameter estimate with much smaller bias and variance than that of the trigonometric moment estimate if the order could be assumed to be known; see the numerical example in Section 6 for details. Fourth, the determination of the two penalty terms needs more investigation in both theory and implementation, though some experience has been gained for the dataset we have been working on. Use of the bootstrap or cross validation is possible. For example, in the analogous problem of estimating the number of hidden frequencies in Model (4), Ulrych and Sacchi [31] chose the number using Kullback divergence as the risk, and a bootstrap method to estimate the risk of each model. Fifth, when applying this methodology to DNA sequencing data, we can set a lower bound as well as an upper bound on the model order, since the numbers of the four kinds of DNA bases in a given range can be estimated. Our experience shows moderate overfitting is not an issue for DNA traces. Finally, in the sequel, we refer to the procedures in Algorithm 3.1, 5.1, 5.2 as PDPS (parametric deconvolution of positive spikes). If the error distribution is known to be normal, Algorithm 4.1 (Gauss–Newton) could be included in PDPS to improve the accuracy of the estimate.

6. Examples and discussion.

6.1. *A simulated example.* Our simulation study is based the following model.

EXAMPLE 6.1.

$$z(t_l) = 0.5 + w(t_l + 1.9) + 1.25w(t_l + 1.6) + 1.25w(t_l + 1.3) + w(t_l) \\ +1.25w(t_l - 0.5) + 1.1w(t_l - 1.0) + 1.25w(t_l - 2.5) + \varepsilon(t_l),$$

where the sample size $n = 1024$, $w(t)$ is a Gaussian function $b/2\pi \times \exp\{-b^2t^2/2\}$ with the scale parameter $b = 8$ being truncated at ± 4 SD. Errors are normally distributed with mean 0 and standard deviation 0.3. Figure 2 shows a simulated sample from this model. The signal contains seven peaks, and the three on the left are quite close to one another. The peak heights are generally similar, which is typical for sequencing data. Simulations in this section are carried out in MATLAB, and are repeated 1000 times for each method we have studied. We apply three estimation procedures to this example. First, we assume the model order is known, and use the method of trigonometric moments. Second, we use PDPS to estimate the parameter. The penalty coefficients $c_1(n)$ and $c_2(n)$ are taken to be 2 and 10, respectively. The upper bound of the model order is taken to be 20. Out of the 1000 replications, all but one of the final-fitted models had order 7, which is the truth. Finally, this result was further tuned by the Gauss–Newton algorithm. Two iterations were used. The statistics of estimates of the peak locations and heights are summarized in Table 1. The estimate tuned by the Gauss–Newton algorithm is almost unbiased, and its standard errors are close to the nominal ones. It

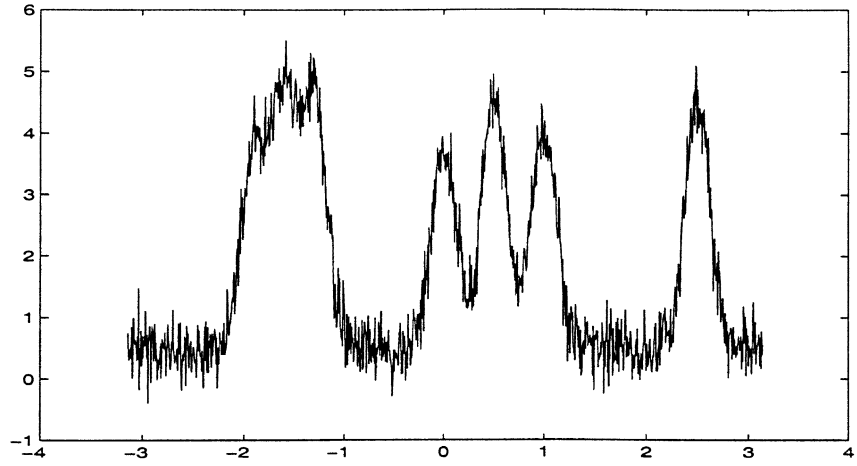

FIG. 2. *A simulated sample of Example* 6.1.

PARAMETRIC DECONVOLUTION

TABLE 1
Statistics of the three estimates of the parameters in Example 6.1 (1000 *replications*)

Parameter truth		Method of trigonometric moments			PDPS without MLE tuning			Gauss–Newton algorithm			
		Bias ($\times 10^3$)	**SD** ($\times 10^2$)	**CV** (%)	**Bias** ($\times 10^3$)	**SD** ($\times 10^2$)	**CV** (%)	**Bias** ($\times 10^3$)	**SD** ($\times 10^2$)	**CV** (%)	**Nominal SE** ($\times 10^2$)
τ_1	−1.9	−211	32		−8	2		0	0.4		0.4
τ_2	−1.6	0	19		3	4		0	0.4		0.4
τ_3	−1.3	315	48		11	2		0	0.3		0.3
τ_4	0.0	64	14		−2	1		0	0.3		0.3
τ_5	0.5	78	16		−2	1		0	0.2		0.2
τ_6	1.0	97	26		−1	1		0	0.3		0.3
τ_7	2.5	2	4		−6	1		0	0.2		0.2
A_1	1.0	−216	67	85	−8	11	11	0	2.1	2	2.1
A_2	1.25	506	20	11	52	6	5	0	2.1	2	2.1
A_3	1.25	−158	68	62	−30	11	9	−1	2.1	2	2.1
A_4	1.0	62	18	17	−1	2	2	0	1.7	2	1.7
A_5	1.25	21	8	6	−1	2	2	−1	1.7	1	1.7
A_6	1.1	−170	36	38	0	2	2	0	1.7	2	1.7
A_7	1.25	−9	13	10	−1	2	1	−1	1.7	1	1.7
A_0	0.5	−6	2	4	−1	2	3	1	1.3	3	1.2

is quite surprising that the accuracy of the trigonometric moment estimate is so poor, even though the order is assumed to be known. In comparison, PDPS has achieved an accuracy much closer to that of the maximum likelihood estimate. This means that modest overfitting in the first stage can greatly control the bias and variance of the estimate. In other words, even when the number of peaks is known, PDPS is better than the direct trigonometric moment estimate in terms of bias and variance. In this case, we set a lower bound by the known order at the first stage, and stop the backward deletion when the number of left peaks equals the known order at the second stage. The frequency of the model orders selected at stage one is shown in Table 2. We see that most models selected at stage one have orders ranging from 9 to 12.

6.2. *Real trace data.* Next we apply PDPS to the sequencing trace shown in Figure 1. The point spread function is taken to be a truncated Gaussian function. The result is shown in Figure 3. (Maximum likelihood estimation was not used.) For a comparison, a nonnegative least squares deconvolution was carried out; see [19]. The results are displayed in Figure 4. The two methods

TABLE 2
Frequency of model orders selected at the first stage

Model orders	8	9	10	11	12	13	14
Frequency	48	248	212	202	244	45	1

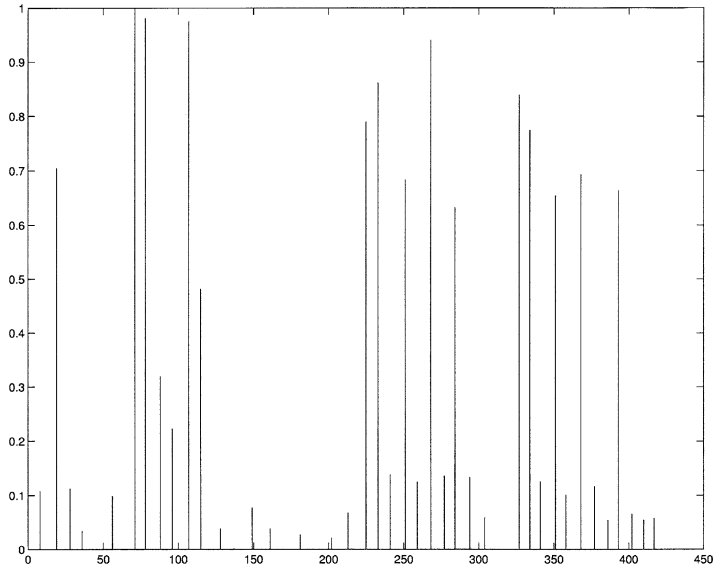

FIG. 3. *Parametric deconvolution.*

yield quite different results. First, that obtained from the parametric deconvolution is cleaner. Second, the relative heights of the major spikes following the parametric deconvolution are more similar to those in the original data. Third, parametric deconvolution is more efficient from the computational point of view. On a Sun Ultra-2 workstation, the Lawson and Hanson algorithm took more than one hour while the parametric method took only two minutes. A more systematic comparison of this parametric deconvolution method with others can be found in [20].

6.3. *Colored noise and reblurring.* The result in this paper can be generalized to situations when the errors are serially correlated. We might approximate such errors by an autoregressive process. That is, we could assume that the errors ε in (2) can be modeled by

$$\varepsilon_t + \sum_{k=1}^{p} \phi_k \varepsilon_{t-k} = \xi_t, \tag{17}$$

where the ξ_t is i.i.d. $N(0, \sigma^2)$. Then we could prewhiten the signal as follows:

$$(\phi * z)(t) = (\phi * w * x)(t) + \xi(t). \tag{18}$$

By replacing $z(t)$ by $(\phi * z)(t)$, we could use our original scheme PDPS to do the deconvolution. This reblurring idea has been used in other similar situations; see [7,13].

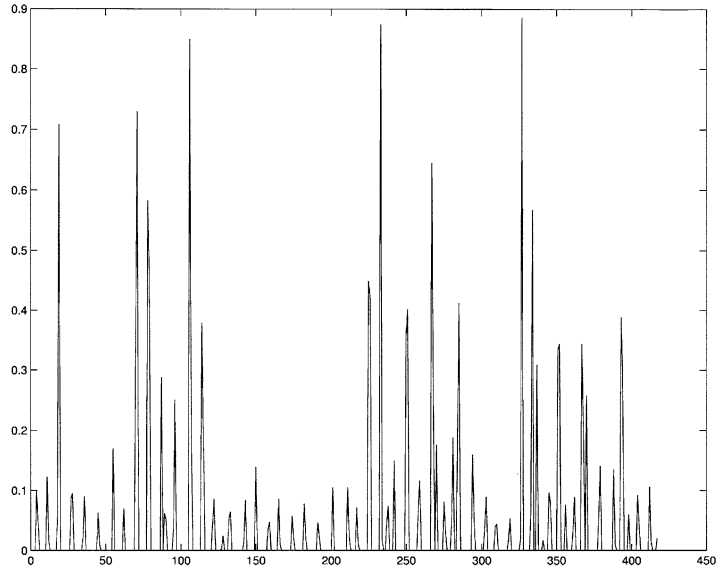

FIG. 4. *NNLS deconvolution (carried out by the Lawson and Hanson algorithm).*

6.4. *Implementations.* The numerical implementation of PDPS hinges on linear regression and computation of smallest eigenvalues and their eigenvectors of Toeplitz matrices. Because of the special structure of Toeplitz matrices, efficient algorithms do exist, and we refer to [12, 22]. Though we have discussed maximum likelihood estimates within the assumption of Gaussian errors in this paper, we may skip the MLE tuning in applications, since either the assumption of normal errors may not be appropriate, or highly accurate estimates of the peak positions and heights may not be necessary.

APPENDIX

This section contains the proofs and relevant mathematical facts. The following theorem is in essence the so-called trigonometric moment problem solved by Carathéodory; see [9]. We present a version under the framework of the spike-convolution model.

THEOREM A.1.

(i) *Given a* $SC(\delta; m; \mathbf{A}; \tau)$, *let* G_m *be the Toeplitz matrix constructed from its Fourier coefficients* $\{g_k\}$. *Then* A_0 *is the smallest eigenvalue of* G_m *with multiplicity one and eigenvector* $(\alpha_0, \ldots, \alpha_m)^T$. *The* $\{A_j\}$ *satisfy the following*

linear system:

$$\frac{1}{2\pi}\begin{pmatrix} 1 & 1 & \cdots & 1 \\ \exp(i\tau_1) & \exp(i\tau_2) & \cdots & \exp(i\tau_m) \\ \vdots & \vdots & \ddots & \vdots \\ \exp(i(m-1)\tau_1) & \exp(i(m-1)\tau_2) & \cdots & \exp(i(m-1)\tau_m) \end{pmatrix}\begin{pmatrix} A_1 \\ A_2 \\ \vdots \\ A_m \end{pmatrix} = \begin{pmatrix} g_0 - A_0 \\ g_1 \\ \vdots \\ g_{m-1} \end{pmatrix}. \tag{19}$$

(ii) *Conversely, suppose we are given* $2m+1$ *complex numbers* $\{g_j, -m \le j \le m\}$, *where* $\bar{g}_j = g_{-j}$. *Assume that the smallest eigenvalue* A_0 *of the Toeplitz matrix* $G_m = (g_{j-i})_{i,\, j=0,\ldots,m}$ *has multiplicity* 1. *Let the smallest eigenvector be* $\alpha = (\alpha_0, \ldots, \alpha_m)^T$, *and* $U^{(m)}(z) = \sum_{j=0}^{m} \alpha_j z^j$. *Then there exists a unique* $SC(\delta; m; \mathbf{A}; \tau)$ *whose first* $m+1$ *Fourier coefficients are* $\{g_j, 0 \le j \le m\}$. *The* $\{\tau_j\}$ *are determined from the* m *distinct roots* $\{e^{i\tau_j}\}$ *of* $U^{(m)}(z)$ *lying exactly on the unit circle. The* $\{A_j\}$ *are determined by the linear system* (19), *and the resulting heights are positive.*

For the proof, the first part is easy to check. As for the second part, an algebraic proof can be found in [9]. Li [20] gave a measure-theoretic proof.

PROOF OF THEOREM 2.1. Let the Fourier coefficients of the corresponding $x(t)$ and $\bar{x}(t)$ be $\{g_k\}, \{\bar{g}_k\}$, respectively. According to the Parseval identity, $\|y - \bar{y}\|^2 = \sum_{k=-\infty}^{\infty} |v_k(g_k - \bar{g}_k)|^2 \ge (\min_{0\le k\le p} |v_k|^2) \sum_{k=0}^{p} |g_k - \bar{g}_k|^2$. Thus if there were a sequence $SC(w; l; \overline{\mathbf{A}}^{(l)}; \bar{\tau}^{(l)})$, $l < p$ which could approach $SC(w; p; \mathbf{A}; \tau)$, then their Fourier coefficients $\bar{g}_k^{(l)} \to g_k$, for $0 \le k \le p$. Therefore their Toeplitz matrices $\overline{T}_p^{(l)}$ will converge to T_p as do their characteristic polynomials. But this is impossible because T_p's smallest eigenvalue has multiplicity 1 according to Theorem 7.1, while $\overline{T}_p^{(l)}$'s smallest eigenvalue has multiplicity larger than 1, as can be easily checked.

PROOF OF THEOREM 3.1. This can be checked by direct calculation.

PROOF OF THEOREM 3.2. By Theorem 7.1, we can find a function x of the form (1) with Fourier coefficients $\{g_k\}$. Next the proof is completed by convolving x with w and using the convolution theorem.

PROOF OF THEOREM 3.3. Without loss of generality, assume n is even. A trigonometric moment $\hat{f}_k$ can be decomposed into three parts,

$$\begin{aligned}\hat{f}_k = f_k &+ \sum_{l=-[n/2]}^{[n/2]-1}\left[\frac{2\pi}{n}y(t_l)\exp(ikt_l) - \int_{t_l}^{t_{l+1}} y(t)\exp(ikt)dt\right] \\ &+\frac{1}{n}\sum_{l=-[n/2]}^{[n/2]-1}\varepsilon(t_l)\exp(ikt_l),\end{aligned} \tag{20}$$

for $k = 0, \pm 1, \ldots$. Denote the second and third terms by $\check{f}_k$ and $\tilde{f}_k$, respectively. Suppose initially the order p of the model is known in advance, and let $\hat{g}_0 = \hat{f}_0$, $\hat{g}_k = \hat{f}_k v_0/v_k$, for $0 < k \le K_0$. Similarly, decompose $\hat{g}_k$ into three parts, g_k, $\check{g}_k$ and $\tilde{g}_k$ corresponding to f_k, $\check{f}_k$ and $\tilde{f}_k$, respectively. Under the smoothness condition on $w(t)$, the effect of $\check{f}_k$ is $O(1/n)$ using Taylor expansion and the bounded property of $w'(\cdot)$ and is thus negligible compared with that of $\tilde{f}_k$. Next, notice that $\tilde{k} = \langle \varepsilon(t), e^{ikt}\rangle_n$. According to the Lyapunov central limit theorem, $\tilde{f}_0, \tilde{f}_1, \ldots, \tilde{f}_p$ are asymptotically normally distributed,

$$\sqrt{n}(\tilde{f}_0, \tilde{f}_1, \ldots, \tilde{f}_p)^T \xrightarrow{d} N(0, \sigma^2 I_{p+1}).$$

Consequently, we have a CLT for the $\hat{f}_k$,

$$\sqrt{n}[(\hat{f}_0, \hat{f}_1, \ldots, \hat{f}_p)^T - (f_0, f_1, \ldots, f_p)^T] \xrightarrow{d} N(0, \sigma^2 I_{p+1}).$$

Thus

$$\sqrt{n}[(\hat{g}_0, \hat{g}_1, \ldots, \hat{g}_p)^T - (g_0, g_1, \ldots, g_p)^T] \xrightarrow{d} N\left(0, \sigma^2 \operatorname{diag}\left(\left|\frac{v_0}{v_k}\right|^2\right)\right).$$

For $k \neq 0$, we can split each term g_k into its real part $g_{k,r}$ and its imaginary part $g_{k,i}$. Then another form of the above CLT is given by

$$\begin{aligned}&\sqrt{n}[(\hat{g}_0, \hat{g}_{1,r}, \ldots, \hat{g}_{p,r}, \hat{g}_{1,i}, \ldots, \hat{g}_{p,i})^T - (g_0, g_{1,r}, \ldots, g_{p,r}, g_{1,i}, \ldots, g_{p,i})^T] \\ &\quad \xrightarrow{d} N(0, P),\end{aligned}$$

where P is given by (8). The mapping between the trigonometric moments and the parameters is continuous. Thus we can find the CLT for the trigonometric moment estimates using the delta method. Now we calculate the Jacobian matrix. From

$$\begin{pmatrix} g_0 \\ g_1 \\ \vdots \\ g_p \end{pmatrix} = \frac{1}{2\pi}\begin{pmatrix} 1 & 1 & 1 & \cdots & 1 \\ 0 & \exp(i\tau_1) & \exp(i\tau_2) & \cdots & \exp(i\tau_p) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \exp(ip\tau_1) & \exp(ip\tau_2) & \cdots & \exp(ip\tau_p) \end{pmatrix}\begin{pmatrix} A_0 \\ A_1 \\ A_2 \\ \vdots \\ A_p \end{pmatrix},$$

we have

$$\frac{\partial(g_0, g_1, \ldots, g_p)}{\partial(A_0, A_1, \ldots, A_p, \tau_1, \ldots, \tau_p)} = \frac{1}{2\pi} \times \left(\begin{array}{c|cccc|cccc} 1 & 1 & 1 & \cdots & 1 & 0 & 0 & \cdots & 0 \\ \hline 0 & \exp(i\tau_1) & \exp(i\tau_2) & \cdots & \exp(i\tau_p) & iA_1\exp(i\tau_1) & iA_2\exp(i\tau_2) & \cdots & iA_p\exp(i\tau_p) \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \exp(ip\tau_1) & \exp(ip\tau_2) & \cdots & \exp(ip\tau_p) & ipA_1\exp(ip\tau_1) & ipA_2\exp(ip\tau_2) & \cdots & ipA_p\exp(ip\tau_p) \end{array}\right).$$

Rewriting this as a square matrix by splitting the derivative into real and imaginary parts, we get

$$\frac{\partial(g_0, g_{1,r}, \ldots, g_{p,r}, g_{1,i}, \ldots, g_{p,i})}{\partial(A_0, A_1, \ldots, A_p, \tau_1, \ldots, \tau_p)} = \frac{1}{2\pi} Q.$$

Therefore $V = 4\pi^2 Q^{-1} P (Q^{-1})^T$. We refer to [24] for the invertibility of Q.

PROOF OF PROPOSITION 4.1. This can be checked by direct calculation.

PROOF OF THEOREM 4.1. (i) Let $l(\theta, \sigma^2)$ be the log likelihood evaluated at (θ, σ^2). To prove the consistency, we need to show that for any fixed $(\theta_0, \sigma_0^2) \neq (\theta, \sigma^2)$, where (θ, σ^2) is the truth,

$$P(l(\theta_0, \sigma_0^2) < l(\theta, \sigma^2)) \to 1 \quad \text{as } n \to \infty. \tag{21}$$

Notice that

$$l(\theta, \sigma^2) - l(\theta_0, \sigma_0^2) = [l(\theta, \sigma^2) - l(\theta, \sigma_0^2)] + [l(\theta, \sigma_0^2) - l(\theta_0, \sigma_0^2)]. \tag{22}$$

Denote $\theta_0 = (A_{0,0}, A_{1,0}, \ldots, A_{p,0}, \tau_{1,0}, \ldots, \tau_{p,0})^T$, and let $\eta(t) = y(t) - [A_{0,0} + \sum_{k=1}^{p} A_{k,0} w(t - \tau_{k,0})]$. Then $\int_{-\pi}^{\pi} \eta(t)^2 dt > 0$ for $\theta_0 \neq \theta$ because of the identifiability of the parameterization. By the law of large numbers, (cf. [2, 6]), the second term in (22) is

$$\begin{aligned} \frac{1}{n}[l(\theta, \sigma_0^2) - l(\theta_0, \sigma_0^2)] &= \frac{1}{2n\sigma_0^2}\left[\sum_l (\varepsilon(t_l) - \eta(t_l))^2 - \sum_l \varepsilon(t_l)^2\right] \\ &= \frac{1}{2n\sigma_0^2}\left[\sum_l (\eta(t_l))^2 - 2\sum_l \eta(t_l)\varepsilon(t_l)\right] \\ &\xrightarrow{p} \frac{1}{2\sigma_0^2}\int_{-\pi}^{\pi} \eta(t)^2\, dt \geq 0. \end{aligned}$$

The first term is

$$\begin{aligned} \frac{1}{n}[l(\theta, \sigma^2) - l(\theta, \sigma_0^2)] &= \left[-\frac{1}{2}\log\sigma^2 - \frac{1}{2n\sigma^2}\sum_l \varepsilon(t_l)^2\right] \\ &\quad - \left[-\frac{1}{2}\log\sigma_0^2 - \frac{1}{2n\sigma_0^2}\sum_l \varepsilon(t_l)^2\right] \xrightarrow{p} \left[-\frac{1}{2}\log\sigma^2 - \frac{\sigma^2}{2\sigma^2}\right] \end{aligned}$$

$$-\left[-\frac{1}{2}\log\sigma_0^2-\frac{\sigma^2}{2\sigma_0^2}\right]\geq 0,$$

since the function $x(s)=-\log s-\sigma^2/s$ has its unique maximum at σ^2. Hence (21) is true.

(ii) Let D be the matrix with entries of second derivatives of the log-likelihood. Since the MLE$(\tilde{\theta},\tilde{\sigma}^2)$ is consistent, the following Taylor expansion can be checked straightforwardly:

$$\frac{1}{\sqrt{n}}\left[\nabla(\tilde{\theta},\tilde{\sigma}^2)-\nabla(\theta,\sigma^2)\right]=\frac{1}{\sqrt{n}}D(\theta,\sigma^2)(\tilde{\theta}-\theta,\tilde{\sigma}^2-\sigma^2)^T + O_p\|(\tilde{\theta}-\theta,\tilde{\sigma}^2-\sigma^2)\|^2. \tag{23}$$

Thus

$$\sqrt{n}(\tilde{\theta}-\theta,\tilde{\sigma}^2-\sigma^2)^T\xrightarrow{d}-\left[\frac{D(\theta,\sigma^2)}{n}\right]^{-1}\frac{\nabla(\theta,\sigma^2)}{\sqrt{n}},$$

where the first term converges to $I^{-1}_{(\theta,\sigma^2)}$ in probability and the second term converges to $N(0,I_{(\theta,\sigma^2)})$ in distribution. The remainder is an application of the Slutsky theorem.

PROOF OF THEOREM 4.2. Notice that by the Taylor expansion,

$$y_{\theta\text{old}}-y=\Psi_\theta^T(\theta_{\text{old}}-\theta)+o_p(\|\theta_{\text{old}}-\theta\|),$$

where the second term is uniform with respect to the variable t, and so

$$z-y_{\theta\text{old}}=\varepsilon-(y_{\theta\text{old}}-y)=\varepsilon-\Psi_\theta^T(\theta_{\text{old}}-\theta)+o_p(\|\theta_{\text{old}}-\theta\|).$$

Inserting this into line 2 of the following, we obtain

$$\begin{aligned}
\sqrt{n}(\theta_{\text{new}}-\theta) &= \sqrt{n}(\theta_{\text{new}}-\theta_{\text{old}})+\sqrt{n}(\theta_{\text{old}}-\theta)\\
&= \sqrt{n}[\langle\Psi_{\theta_{\text{old}}},\Psi_{\theta_{\text{old}}}^T\rangle_n]^{-1}\langle\Psi_{\theta_{\text{old}}},z-y_{\theta_{\text{old}}}\rangle_n+\sqrt{n}(\theta_{\text{old}}-\theta)\\
&= \sqrt{n}[\langle\Psi_{\theta_{\text{old}}},\Psi_{\theta_{\text{old}}}^T\rangle_n]^{-1}\langle\Psi_{\theta_{\text{old}}},\varepsilon\rangle_n\\
&\quad -[\langle\Psi_{\theta_{\text{old}}},\Psi_{\theta_{\text{old}}}^T\rangle_n]^{-1}\langle\Psi_{\theta_{\text{old}}},\Psi_\theta\rangle_n\sqrt{n}(\theta_{\text{old}}-\theta)\\
&\quad +\sqrt{n}(\theta_{\text{old}}-\theta)+o_p(\|\sqrt{n}(\theta_{\text{old}}-\theta\|)\\
\xrightarrow{p}\ &\sqrt{n}[\langle\Psi_\theta,\Psi_\theta^T\rangle_n]^{-1}\langle\Psi_{\theta,}\varepsilon\rangle_n+\sqrt{n}[\langle\Psi_\theta,\Psi_\theta^T\rangle_n]^{-1}\\
&\quad\times\langle\Psi_{\theta_{\text{old}}}-\Psi_\theta,\varepsilon\rangle_n+o_p(1).
\end{aligned}$$

Here we need the $\sqrt{n}$-consistency of θ_{old}. The second term in the last line is $o_p(1)$, since we can apply a Taylor expansion to $\Psi_{\theta_{\text{old}}}-\Psi_\theta$ at θ. Then we complete the proof by applying the central limit theorem to the first term and using the Slutsky theorem.

PROOF OF PROPOSITION 5.1. The dimension of the invariant space corresponding to the smallest eigenvalue A_0 is r. By Gaussian elimination, we can always find a vector in this space with the form $\alpha = (\alpha_0, \ldots, \alpha_p, 0, \ldots, 0)^T$. Of course, $(\alpha_0, \ldots, \alpha_p)^T$ is a eigenvector of G_p with eigenvalue A_0. A_0 is the smallest eigenvalue of G_p because of the monotone property of eigenvalues of nested matrices. Its multiplicity is 1. Otherwise, by repeating the above reasoning, we can find an eigenvector of G_{m-r} with the form of $(\beta_0, \ldots, \beta_{m-r})^T$. Then by extending this vector into $m+1$ Euclidean space by adding zeros in the beginning and the end, we can construct $r+1$ linear independent eigenvectors of G_m corresponding to A_0, which would imply that the multiplicity of A_0 is greater than r, a contradiction. In order to prove the second part, notice that $(\alpha_0, \ldots, \alpha_p, 0, \ldots, 0)^T, (0, \alpha_0, \ldots, \alpha_p, 0, \ldots, 0)^T, \ldots, (0, \ldots, 0, \alpha_0, \ldots, \alpha_p)^T$ is a basis of the invariant space corresponding to A_0. Consequently, any polynomial whose coefficients are an eigenvector corresponding to A_0 is a linear combination of polynomials with common roots $\{e^{i\tau_j}, j = 1, \ldots, p\}$. The rest is obvious.

Acknowledgments. We give our grateful thanks to Prof. J. Rice for his very helpful comments. We owe many debts to the two referees. Their constructive suggestions have greatly enhanced this research and improved the organization of the material.

REFERENCES

[1] ADAMS, M. D., FIELDS, C. and VENTOR, J. C. (eds.). (1994). *Automated DNA Sequencing and Analysis*. Academic Press, London.

[2] BILLINGSLEY. P. (1986). *Probability and Measure*. Wiley, New York.

[3] CHEN, W.-Q. and HUNKAPILLER, T. (1992). Sequence accuracy of larger DNA sequencing projects. *J. DNA Sequencing and Mapping* **2** 335–342.

[4] DI GES, V. and MACCARONE, M. C. (1984). The Bayesian direct deconvolution method: properties and and applications. *Signal Processing* **6** 201–211.

[5] DONOHO, D. L., JOHNSTONE, I. M., HOCH, J. C. and STERN, A. S. (1992). Maximum entropy and the nearly black object. *J. Roy. Statist. Soc. Ser. B* **54** 41–81.

[6] DURRETT, R. (1991). *Probability: Theory and Examples*. Wadsworth and Brooks Cole, Belmont, CA.

[7] FREDKIN, D. R. and RICE, J. A. (1997). Fast evaluation of the likelihood of an HMM: ion channel currents with filetering and colored noise. Dept. Statistics, Univ. California, Berkeley.

[8] GOLUB, G. H. and VAN LOAN, C. F. (1996). *Matrix Computations*, 3rd ed. John Hopkins Univ. Press.

[9] GRENANDER, U. and SZEGÖ, G. (1958). *Toeplitz Forms and Their Applications*. Univ. California Press, Berkeley.

[10] GULL, S. F. (1989). Developments in maximum entropy data analysis. In *Maximum Entropy and Bayesian Methods* (J. Skilling, ed.) Kluwer, Boston.

[11] GULL, S. F. and DANIELL, G. J. (1978). Image reconstruction from incomplete and noisy data. *Nature* **272** 686–690.

[12] HUANG, D. (1992). Symmetric solutions and eigenvalue problems of Toeplitz systems. *IEEE Trans. Acoust. Speech Signal Processing* **40** 3069–3074.

[13] JANSSON, P. A. (ed.). (1997). *Deconvolution of Images and Spectra*. Academic Press, New York.

[14] KENNETT, T. J., PRESTWICH, W. V. and ROBERTSON, A. (1978). Bayesian deconvolution 1. convergence properties. *Nuclear Instrument and Methods* **151** 285–292.
[15] KENNETT, T. J., PRESTWICH, W. V. and ROBERTSON, A. (1978). Bayesian deconvolution 2. noise properties. *Nuclear Instrument and Methods* **151** 293–301.
[16] KENNETT, T. J., PRESTWICH, W. V. and ROBERTSON, A. (1978). Bayesian deconvolution 3. application and algorithm implementation. *Nuclear Instrument and Methods* **153** 125–135.
[17] KOOP, B. F., ROWEN, L., CHEN, W.-Q., DESHPANDE, P., LEE, H. and HOOD, L. (1993). Sequence length and error analysis of sequence and automated *taq* cycle seqeuncing methods. *BioTechniques* **14** 442–447.
[18] LAWRENCE, C. B. and SOLOVYEV, V. V. (1994). Assignment of position-specific error probability to primary DNA sequence data. *Nucleic Acid Research* **22** 1272–1280.
[19] LAWSON, C. L. and HANSON, R. J. (1974). *Solving Least Squares Problems*. Prentice Hall, Englewood Cliff, NJ.
[20] LI, L. (1998). Statistical models of DNA base-calling. Ph.D. dissertation, Univ. California, Berkeley.
[21] LI, L. and SPEED, T. P. (1999). An estimate of the color separation matrix in four-dye fluorescence-based DNA sequencing. *Electrophoresis* **20** 1433–1442.
[22] MAKHOUL, J. (1981). On the eigenvectors of symmetric Toeplitz matrices. *IEEE Trans. Acoust. Speech. Signal Processing* **29** 868–872.
[23] NELSON, D. O. (1995). Introduction of reptation. Technical report, Lawrence Livermore National Lab.
[24] PISARENKO, V. F. (1973). The retrieval of harmonics from a convariance function. *Geophys. J. Roy. Astrophys. Soc.* **33** 347–366.
[25] POSKITT, D. S., DOGANCAY, K. and CHUNG, S.-H. (1999). Double-blind deconvolution: the analysis of post-synaptic currents in nerve cells. *J. Roy. Statist. Soc. Ser. B* **61** 191–212.
[26] RICHARDSON, W. H. (1972). Bayesian-based iterative method of image restoration. *J. Opt. Soc. Amer. A* **62** 55–59.
[27] SHEPP, L. A. and VARDI, Y. (1982). Maximum-likelihood reconstruction for emission tomography. *IEEE Trans. Medical Imaging* **MI-1** 113–121.
[28] SNYDER, D. L., SCHULZ, T. J. and O'SULLIVAN, J. A. (1992). Deblurring subject to nonnegativity constraints. *IEEE Trans. Signal Processing* **40** 1143–1150.
[29] STARK, P. B. and PARKER, R. L. (1995). Bounded-variable least-squares: an algorithm and applications. *Comput. Statist.* **10** 129–141.
[30] TIKHONOV, A. (1963). Solution of incorrectly formulated problems and the regularization method. *Soviet Math. Dokl.* **5** 1035–1038.
[31] ULRYCH, T. J. and SACCHI, M. D. (1995). Sompi, Pisarenko and the extended information criterion. *Geophysical J.* **122** 719–724.
[32] VARDI, Y. and LEE, D. (1993). From image deblurring to optimal investment: maximum likelihood solutions for positive linear inverse problems. *J. Roy. Statist. Soc. Ser. B* **55** 569–612.

DEPARTMENT OF STATISTICS
FLORIDA STATE UNIVERSITY
TALLAHASSEE, FLORIDA 32306-4330
E-MAIL: lilei@stat.fsu.edu

DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA
BERKELEY, CALIFORNIA