Blind Inversion Needs Distribution (BIND): General Notion and Case Studies

Lei Li

Abstract

A class of scientific measurement problems share a common feature which we refer to as "blind inversion." That is, we can regard a module of measurement instruments as a system with quantities to be measured as input and observations as output. In a blind inversion problem, both the effective system and the input are unknown to us. Due to either experimental design or the nature of scientific problem in question, very often the distributional knowledge of the input can be obtained. Given this piece of information, we apply a two-step scheme – abbreviated by BIND – to solve the blind inversion problem. First, we make use of the distributions of the input and output to estimate the system. Second, we reconstruct the value of each individual input using the system obtained in the first step. From this perspective, we have another look at two measurement problems that are part of Professor Speed's recent research in molecular biology. We also connect the idea with the long-standing predictive deconvolution method used in seismology and discuss assessment issues of BIND.

Keywords: blind inversion; color-correction; DNA sequencing; electrophoresis; microarray; seismology

1 Introduction

Scientific discoveries are based on accurate measurements. The innovation of measurement instruments and invention of conceptual models cross each other's track and lead each other's way throughout the history of science. As instrumental techniques advance and the collected information expands, new tools of data analysis emerge along the way. One such famous historical example is Gauss's use of least squares in astronomy and geodesy. Not only have ingenious algorithms been applied to the practice of data analysis, but probabilistic models such as regression models have also been proposed and widely accepted for the purpose of designing and evaluating measurement processes. Nowadays, it has become common sense that uncertainty is the nature of any measurement processes.

In the area of biology, human beings' understanding of life has experienced great breakthroughs at the molecular level since the last century. Based on new understandings, scientists have developed *in vitro* bio-techniques such as cloning and polymerase

chain reaction (PCR). Even more excitingly, by incorporating cutting-edge technologies from physics, chemistry, mechanics, and computer science, modules of genotyping, DNA sequencing, and monitoring of mRNA abundance have been well integrated. As a result of these engineering efforts, many current biological projects have scaled up to the genome level. Consequently, new problems of experimental design, measurement, and analysis arise to challenge researchers in different areas. In this article, we consider two biological measurement problems relating to laser and dye techniques.

A class of scientific measurement problems share a common feature which we refer to as "blind inversion." As shown in Figure 1, we can regard a module of measurement instruments as a system with quantities to be measured as input and observations as output. A full explanation of the figure can be found in Section 2. Even though in some cases we are, at least approximately, able to describe the system structure by a parametric model, it is sometimes difficult to determine the effective parameters because of uncontrollable internal or external factors that affect the performance of the instrument. Thus, both the effective system and the input are unknown in a blind inversion problem. Without further information, the problem is ill-posed because the solution is not unique. In order to define a well-posed problem, more knowledge is required. The nature of the blind inversion problem does not allow us to inquire for either system parameters or values of individual input. It seems that the only choice goes to the distributional knowledge of the input.

Although we formulate the input distribution in statistical language, the knowledge, if there is any, really comes from considerations of the scientific measurement problem in question. As shown later in our examples, because of either the experimental design or the nature of the scientific problem, quantities to be measured usually demonstrate some kind of canonical distributional form.

Given the information of the input distribution, we apply a two-step scheme – abbreviated by BIND – to solve the blind inversion problem. First, we make use of the distributions of the input and output to estimate the system. Second, we reconstruct the value of each individual input using the system obtained in the first step. It is interesting to notice that we use observations twice yet in two different ways. An analog to this dual perspective of the same dataset is the dual nature of light. Sometimes we adopt the perspective of particles – photons – to understand phenomena such as the photo-electric effect. At other times we adopt the perspective of waves to analyze phenomena such as interference, reflection, and refraction. According to the quantum mechanical explanation, the electromagnetic wave is closely associated with a probability distribution; see Fowles [10].

BIND is more a general notion than a precise solution for a specific problem. We realized its value from two recent biological measurement problems; however, it is certain that researchers have already explored similar ideas, consciously or unconsciously, to solve problems in different scenarios. We point out one such example in the discussion section. Still, we would like to spell it out for a broader awareness.

We arrange the materials in this paper as follows. In Section 2 we illustrate the

idea of BIND by an artificial example. In Section 3 we show how to apply the BIND scheme to achieve an adaptive color-correction for DNA sequencing data proposed by Li and Speed [13]. In Section 4 we have another look at the within-slide normalization procedure proposed by Yang, Dudoit, Luu, and Speed [26], from the perspective of BIND. In Section 5 we connect BIND with the predictive deconvolution method in seismology and discuss assessment issues of BIND.

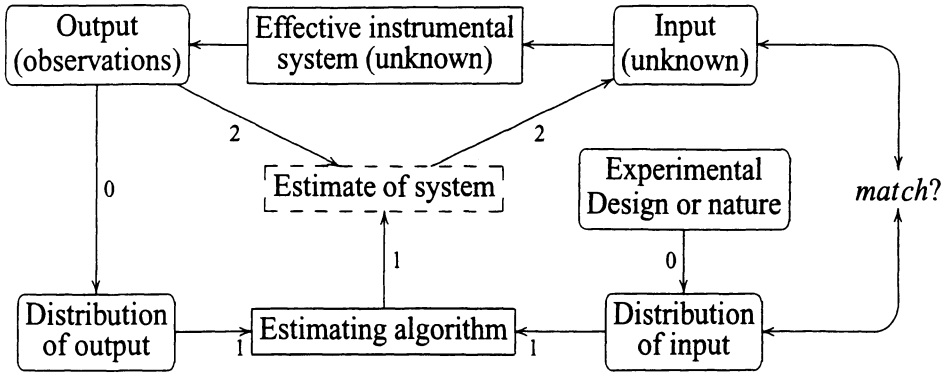

Figure 1: **The schematic representation of the blind inversion problem and BIND.** At the top, each individual input, which is to be measured, goes through the instrumental system and the corresponding output is observed. In a blind inversion problem, both input values and the effective system are unknown. BIND includes three steps. Step 0: identify the distributions of the input and output; Step 1: estimate the system function using the distributional information; Step 2: reconstruct each individual input value.

2 An illustrative example

Consider the following linear system,

$$\begin{bmatrix} y_1(t) \\ y_2(t) \\ y_3(t) \\ y_4(t) \end{bmatrix} = \begin{bmatrix} w_{11} & 0 & 0 & 0 \\ w_{21} & w_{22} & 0 & 0 \\ w_{31} & w_{32} & w_{33} & 0 \\ w_{41} & w_{42} & w_{43} & w_{44} \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \\ x_3(t) \\ x_4(t) \end{bmatrix}, \tag{1}$$

where the input vector $x(t) = (x_1(t), x_2(t), x_3(t), x_4(t))'$ is unknown and to be estimated, the output vector $y(t) = (y_1(t), y_2(t), y_3(t), y_4(t))'$ is observed, and the system matrix $W = [w_{ij}]$ is lower-triangular and non-degenerate. If the system function is given, then the problem is easily solved by inverting the matrix $W = [w_{ij}]$. If the system is not given, then both $[w_{ij}]$ and $x(t)$ are unknown, and this is a blind inversion problem. Without further information, it is an ill-posed problem in the sense that the solution is not unique.

Interestingly, the distributional information of $x(t)$, if it is available somehow, can help solve the blind inversion problem. Let us assume that the distribution of the input

in (1) is **white and normal**, namely, $N(0,\sigma^2 I)$, where I is an identity matrix of order four. Consequently, the distribution of output is normal $N(0,\Sigma)$, where $\Sigma = WW'$. From observations on the output, we construct an empirical estimate of the covariance matrix by the standard method

$$\hat{\Sigma} = \frac{1}{T}\sum_{t=1}^{T}[y(t)-\bar{y}][y(t)-\bar{y}]',$$

where $\bar{y} = \frac{1}{T}\sum_{t=1}^{T} y(t)$ is the average of the observations. Then we estimate the matrix by factorizing $\hat{\Sigma}$. The uniqueness of the factorization is the direct result of the lower-triangular assumption on W. In fact, this is the **Cholesky factorization** for positive definite systems; see [11]. Denote the estimated system matrix by $\hat{W}$. Then for each observation of $y(t)$, we can estimate the corresponding input by $\hat{x}(t) = \hat{W}^{-1}y(t)$.

There is an interesting "cross-talk" interpretation of this model. Suppose a communication system has four channels. The first channel provides perfect transmission and no other channels interfere with it. The second channel is interfered with by the first channel, namely, the first channel leaks some signal to the second. The third channel experiences interference from the first and second channels. The fourth channel is the worst and is interfered with by all the other three channels. This leakage phenomenon can be described by a linear system such as that in (1), in which the signals on the sender's side and receiver's side are respectively represented by $x(t)$ and $y(t)$. The lower triangular cross-talk matrix W is consistent with the above interference structure. In order to reconstruct the original signals from the receiver's side, we need to clear the interference among the channels. If we assume that the signals being transmitted are independent among the four sources and approximately follow a normal distribution, then we can use the above procedure to estimate the signals from the sender's side.

Algorithm 1

(BIND) *The general scheme of blind inversion has three steps as shown in Figure 1:*

Step 0. identify the distributions of the input and output;

Step 1. estimate the system function using the distributional information of the input and output;

Step 2. invert the system and reconstruct each individual input value.

We refer to this idea as BIND (blind inversion needs distribution) hereafter. Measure theory (see Billingsley [3]) sheds some light on the need for the inquiry into the distribution of input. In the absence of singularity, the system function is like the **Radon-Nikodym derivative** of the output distribution with respect to the input distribution. Notice that we have equated terms of distribution and measure in this discussion. The exact meaning of distributional information we refer to here include: first, the support of the measure or the value space; second, the distribution on this space demonstrated by the input. Two general issues ought to be addressed. On the one hand, we expect that the distributional information should be complementary to any partial information

about the system and **sufficient** enough to define a well-posed inversion problem. On the other hand, despite any mathematical formulation, the hypothesis on the input distribution should be based on **scientific considerations** of the problem in question, and we also expect that the hypothesis can be verified to some extent.

3 Color correction of DNA sequencing data

In 1995, I started to do research under the Terry's supervision at UC Berkeley. Around that time, the Human Genome Project was in its accelerating stage; the Lawrence Berkeley and Livermore National Laboratories were part of this joint effort. The key component of this project and of any other genome project is Sanger sequencing; see the book edited by Adams, Fields, and Ventor [1] for background in molecular biology. While working on the crucial problem of physical mapping, David Nelson and Terry [16] initiated research on DNA sequencing and base-calling. The problem interested me and later Simon Cawley. Eventually my thesis [12] and part of Simon's [5] grew out of this research topic. One part of our DNA sequencing work is the correction of the dye cross-talk effect; see Li and Speed [13]. At the time we proposed our algorithm, we did not think much about the underlying principle. Now we explain it according to the BIND scheme. The primary idea of Sanger sequencing lies in its specially-designed **dideoxy enzymatic reactions**. Starting with a target DNA segment, the four dideoxy reactions respectively produce many copies of each possible sub-fragment ending with A, G, C, and T; see Russell [23]. For example, the four kinds of subfragments of a DNA fragment ATTCAGCGT are given by {A, ATTCA}, {ATTCAG, ATTCAGCG}, {ATTC, ATTCAGC} and {AT, ATT, ATTCAGCGT}. These sub-fragments are separated and ordered according to their sizes by electrophoresis, carried out in either a gel or a capillary. A slab gel contains many lanes, yet lane-tracking is required to extract lane signals from raw image data. Capillary electrophoresis, on the other hand, does not require lane-tracking. In order to differentiate the four kinds of sub-fragments from the same electrophoresis lane, each kind of sub-fragment in the enzymatic reaction is labeled with one of four dyes. By design, these four dyes demonstrate different light spectra with respect to a laser of a specific frequency. The problem is to measure the dye concentrations of the four kinds at one region. Excited by the laser, the four dyes emit photons, which are collected in four wavelength bands. However, the observations – four fluorescence intensities – are not direct measurements of the dye concentrations of the four kinds. This is where the complication comes in. The dataset used in this article was from slab gel electrophoresis and was provided by the Human Genome Center at LBNL. In Figure 2 is shown a portion of the fluorescence intensities (top) and the reconstructed dye concentrations (bottom). In the plot of dye concentrations, there is a series of peaks of four colors. The rationale of DNA sequencing and base-calling is: each peak represents one base, and the order of color peaks is consistent with the order of nucleotide bases on the underlying DNA fragment. The color code in Figure 2 is: A – red, G – black, C – green, and T – blue. We notice that adjacent peaks of the same

color overlap and this is where deconvolution is required; see Li and Speed [14]. In comparison with dye concentrations, peaks in the plot of fluorescence intensities are not clean in the sense that they have components in all four colors. Next we explain this phenomenon in some detail.

The spectra of the four dyes used in fluorescence-based DNA sequencing overlap, and thus the cross-talk phenomenon arises. That is, the observed four fluorescence intensities are a transformed version of the four dye concentrations. The transformation is not completely linear because of instrumental limitations. For example, overflow may occur in photon-counters if too many photons are emitted in a short period. Nevertheless, we approximately describe the relationship between the unknowns – four dye concentrations $C(t)$, $G(t)$, $A(t)$ and $T(t)$ – and the observations – four fluorescence intensities $I_1(t), I_2(t), I_3(t), I_4(t)$ – at an electrophoretic time t by the following linear system:

$$\begin{bmatrix} I_1(t) \\ I_2(t) \\ I_3(t) \\ I_4(t) \end{bmatrix} = \begin{bmatrix} 1 & w_{12} & w_{13} & w_{14} \\ w_{21} & 1 & w_{23} & w_{24} \\ w_{31} & w_{32} & 1 & w_{34} \\ w_{41} & w_{42} & w_{43} & 1 \end{bmatrix} \begin{bmatrix} C(t) \\ G(t) \\ A(t) \\ T(t) \end{bmatrix} + \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \end{bmatrix}, \tag{2}$$

where $[w_{ij}]$ is the cross-talk matrix and (b_1, b_2, b_3, b_4) is the baseline. Note that there are only 12 free parameters in the cross-talk matrix because the spectra are determined by relative fluorescence intensities except for scaling. We parameterize the cross-talk matrix in such a way that its diagonal elements are unity, i.e., $w_{ii} = 1$. We simplify the problem by assuming that the baseline is constant, and we support this assumption by the following argument. First, the baseline refers to the fluorescence background of the measured region. It changes slowly along a lane in a relatively small range with respect to signals. Second, although observations are recorded on a time scale as shown in Figure 2, our view of their distribution ignores their time-dependence. This is equivalent to permuting data. According to our simplification, all kinds of variations except for cross-talk are implicitly aggregated into measurement errors.

The goal of color-correction is to reconstruct the dye concentrations using data of fluorescence intensities. If the cross-talk matrix is known, then a straightforward inversion solves the problem. However, the **effective cross-talk matrix is unknown** and needs to be estimated. Thus we are facing a blind inversion problem, or more specifically, an **adaptive color-correction** problem. According to the general scheme of blind inversion, first of all, we need to consider the distribution of the input – dye concentrations.

The following **non-overlapping hypothesis** is crucial for understanding the problem. Although dye concentrations change from lane to lane, and from gel to gel, we discovered that their **distributional pattern changes little across lanes**. The distribution can be graphically displayed by **pairwise scatter plots**; one such example is shown in Figure 3 (the data shown in the figure is explained after Algorithm 2). The first sub-plot only includes concentrations of C and G fragments. Two distinct cluster

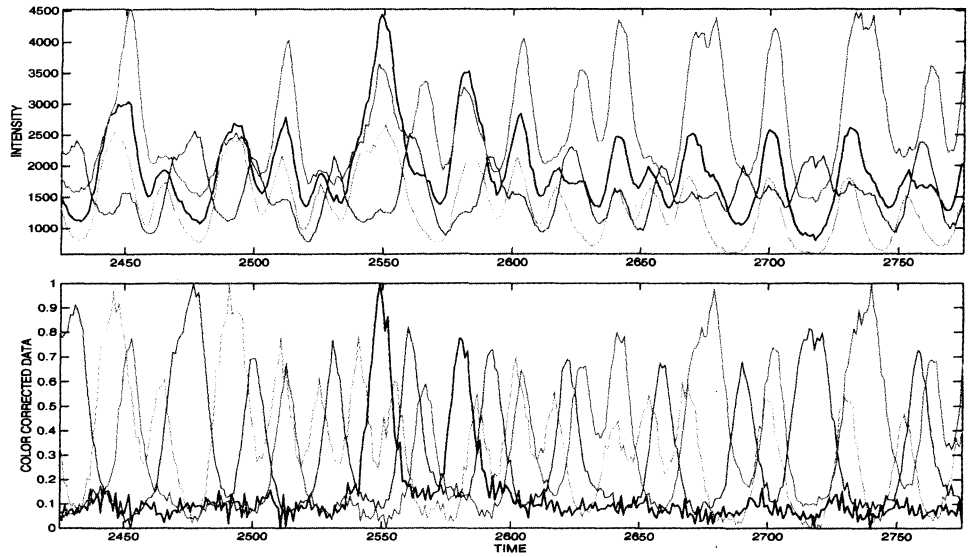

Figure 2: Top: a segment of raw sequencing data from slab gel electrophoresis; Bottom: the color-corrected data, or the estimates of the dye concentrations except for a scale. Color code: A – red, G – black, C – green, and T – blue.

directions are seen along the axes, and almost all other points are in the right-upper quadrant generated by the two cluster directions. We also observe similar patterns in the other five 2-D scatter plots. In fact, such a distributional pattern is determined by the design of Sanger sequencing. Suppose we are in an ideal case by assuming

- **Spectrally non-overlapping hypothesis**: the four dyes are cross-talk free and we observe dye concentrations directly;
- **Spatially non-overlapping hypothesis**: at least in a fairly large range of each trace, the effective mobilities of the four dyes are approximately identical and thus we observe non-overlapping peaks of all four kinds.

In the following, we illustrate how these two hypotheses explain the pattern of scatter plots as shown in Figure 3. For example, we map those observations from non-overlapping C-peaks in Figure 2 (bottom) to points on the cluster directions along the C-concentration axes in the first, second, and third subplots and non-significant points close to the apexes of the fourth, fifth, and sixth in Figure 3. We map those observations from overlapping regions of C-peaks and G-peaks to inner points in the first quadrant, to points on C-concentration axes in the second and third subplots, to points on G-concentration axes in the fourth and fifth subplots, and to non-significant points close to the apex of the sixth quadrant. We map those observations from overlapping regions of more than two kinds of peaks in the same fashion. This key distributional pattern can also be verified empirically using data obtained from a specially designed

experiment. That is, the four differently dye-labeled sub-fragments generated from the four dideoxy reactions are placed into four different yet adjacent lanes of a slab gel. Fluorescence intensities are collected in the same four wavelength bands as those in standard sequencing. This setup uses the same equipment as that in standard sequencing. In this experiment, the four fluorescence intensities obtained from one lane are contributed by only one kind of dye, and their sums are expected to be proportional to the dye concentrations. Here we have ignored the minor baseline issue. The pairwise scatter plots of these "substitutes" of dye concentrations, obtained from one such a cross-talk free experiment, demonstrate the exact pattern in Figure 3; see Figure 1 in [13]. This feature of the distribution provides the basis for our estimation of W and evaluation of color-correction. An appropriate cross-talk matrix is expected to make the distribution of the reconstructed dye concentrations match the pattern shown in Figure 3.

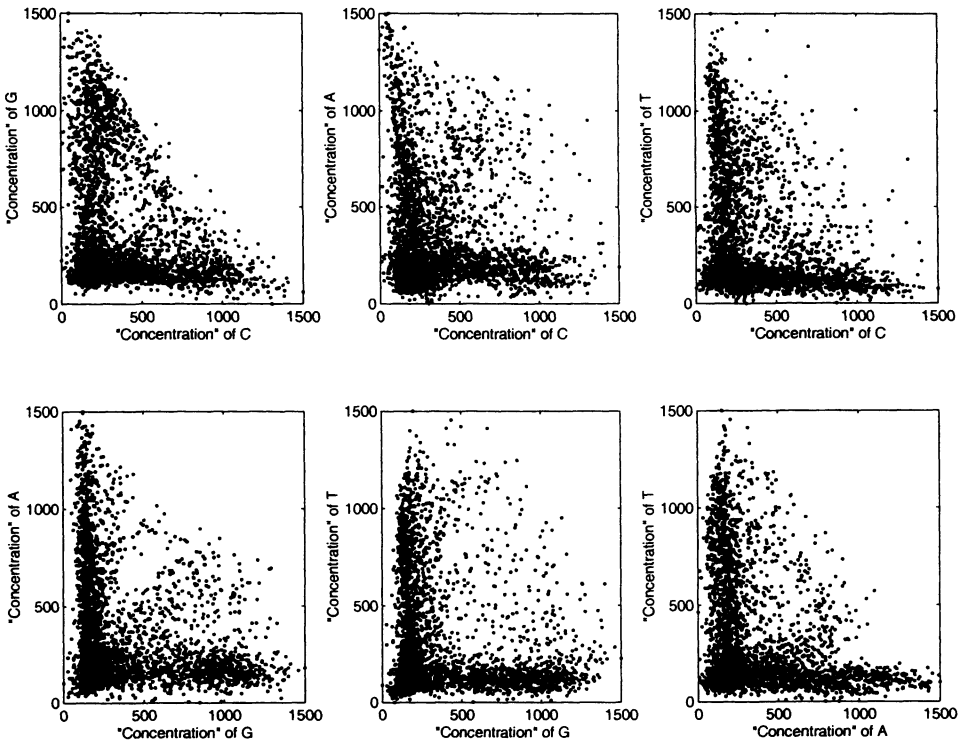

Figure 3: Pairwise scatter plots of the reconstructed dye concentrations.

The non-overlapping hypothesis on distribution of dye concentrations is **sufficient** for estimating the cross-talk matrix. Let us examine the distribution of the raw data – four fluorescence intensities – the output of the system (2). Once again we visualize it with pairwise scatter plots. Figure 4 depicts the six scatter plots of the 3400 observations from one slab gel lane. Let us ignore points in the bottom-left corners of the plots,

which correspond to measurements in valley regions between peaks, or to peaks with low intensities at a pair of wavelength bands; *cf.* Figure 2. Most of the other points lie in a region spanned by two cluster directions – two arms – though they are not as distinct as those in Figure 3. The upper arm of the 3rd, 5th and 6th scatter plots are even more vague because the fourth dye is not as stable as others. If we imagine the complete picture of the distribution in the four-dimensional space, we would find four cluster directions, each corresponding to one column of the cross-talk matrix and almost all other data points lie in the convex cone spanned by them. The pairwise scatter plots are the six 2-dimensional projections of the 4-D scatter plot.

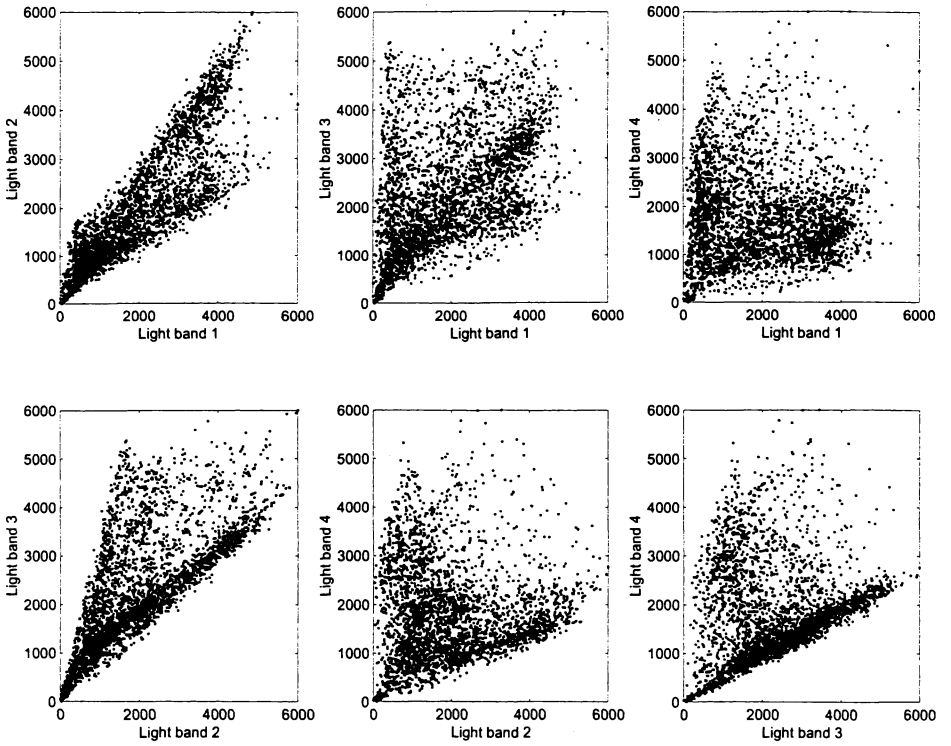

Figure 4: Pairwise scatter plots of the four fluorescence intensities of a slab gel dataset.

The data along each of the 12 arms in the pairwise scatter plot contain the information for estimating one off-diagonal parameter in the cross-talk matrix. Later, we refer to them as "typical points". For example, in the first scatter plot in Figure 4, the slope of the lower boundary should be close to w_{21}, while the slope of the upper boundary should be close to $1/w_{12}$. In other words, the information relevant to the parameter w_{12} can be found in the lower "arm" and that relevant to $1/w_{12}$ can be found in the upper "arm" in the first subplot. Our focus is thus reduced to the 12 slopes. An analog to these "typical points" is the concept of sufficient statistics in statistical modeling. The connection between the data and the parameters leads us to a natural algorithm of esti-

mating the cross-talk matrix: first identify the typical points along the 12 boundaries in the six scatter plots; second, estimate the 12 slopes based on the selected sample points. Technically, we use a **binning** technique to select typical boundary points and a**robust regression** to estimate boundary slopes. These statistical considerations are necessary for handling measurement errors and potential outliers. We notice that some inner points in the 4-D convex cone could possibly mapped to regions close to boundaries when being projected onto 2-D planes, and they would confound with useful information, namely, the typical points. In order to get over this complication, we iterate the above procedure. As iterations are carried out, we expect our estimates to get closer to the target cross-talk matrix. The description of the algorithm is given as follows.

Algorithm 2
(Adaptive color-correction)

0. Initialization. *Let $i = 1$. Set the raw data of fluorescence intensities to be the working dataset and the initial estimate $W^{(0)}$ to be the identity matrix. Also set a small positive number α as the threshold of color-correction and a positive integer M as the maximum number of iterations.*

1. Sampling. *Consider the first component. It is helpful to look at the first scatter plot consisting of the first and second components in Figure 4.*

- **Selecting informative range**. *Choose one quantile for the first component. The two bounds of the informative range for w_{21} are defined by this quantile and the largest value of the first component. Those points in the current working dataset with their first components in the range are selected in this iteration for the estimation of w_{21}. For example, if we choose 50%, then, it says we will use those points whose first coordinates are in the upper half; cf. Figure 2, 3, 4.*
- **Binning**. *Divide the range between these two bounds into bins of the same width.*
- **Selecting extreme points.** *Among those points whose first component falls into a given bin, find the one having the minimum value in the second component.*

2. Robust regression. *Take the points obtained from last step, and run a robust regression of their second components against the first. The estimated slope is taken to be the next estimate of w_{21}. Similarly, the estimate of the slope of the other arm in the same scatter plot is taken to be the next estimate of w_{12}.*

3. Estimating other parameters. *Apply the steps similar to 1 and 2 to estimate the other five pairs $\{w_{13}, w_{31}\}$ $\{w_{14}, w_{41}\}$, $\{w_{23}, w_{32}\}$, $\{w_{24}, w_{42}\}$, $\{w_{34}, w_{43}\}$ and assemble them in $\tilde{W}$.*

4. **Checking the color-correction quality**. *Calculate the maximum of the absolute values of the 12 estimated slopes obtained in step 1, 2 and 3. We hereafter refer to this number as cc-number (initials of color correction). If the cc-number is below the threshold* α, *stop; otherwise, go to step 5.*

5. **Updating**. *Apply the inverse of this matrix to the working dataset (pointwise) and call this the new working dataset. Set* $W^{(i)} = W^{(i-1)} * \tilde{W}$ *and normalize each column of* $W^{(i)}$ *to make the diagonal elements unity. Increase* i *by 1. If* $i > M$, *stop; otherwise, go back to step 1.*

The algorithm is stopped once we recognize a satisfactory color correction by checking the *cc-number*; see [13] for more details. On exit, $W^{(i)}$ is the estimate of the cross-talk matrix and the working dataset contains the reconstructed dye concentrations. Thus, the procedure does bind the two problems: estimating the cross-talk matrix and color-correcting the measurement of fluorescence intensities. In fact, the dye concentrations in Figure 3 were reconstructed using this algorithm from the fluorescence intensities shown in Figure 4, and the results have been examined. We have experimented with different regression methods in step 2. We observe that the samples obtained in step 1 are not always on the boundaries. Least squares does not work well because of its sensitivity to outliers, and in [13] we proposed the use of a robust procedure – least absolute deviations. Later we adopted the **least trimmed squares** method (LTS) because of its high breakdown point and relatively high efficiency; see Rousseeuw and Leroy [22]. Denote the typical samples obtained from step 1 by $(x_1, y_1), \cdots, (x_s, y_s)$. The least trimmed squares method estimate a straight line with intercept b (an equivalent term to the baseline in (2)) and slope w by

$$\min_{b,w} \sum_{k=1}^{q} |y - b - w \cdot x|^2_{(k)},$$

where $|y - b - w \cdot x|^2_{(k)}$ represents the k-th ordered squared residual, and the sum only takes the smallest q squared residuals into account. We have tested LTS with $q = [n/2] + 1$ and found that five iterations offered a satisfactory solution. The cross-talk matrix used in Figure 3 is obtained in this way. **Least median squares** method (LMS) [21] is another robust procedure and is statistically inefficient with a convergence rate $O(1/\sqrt[3]{N})$ under the normal assumption. On the other hand, algorithms requiring only $O(N^2)$ running time do exist to compute its exact solution in our univariate regression case; see Souvaine and Steele [25]. Another remark is that bins in step 1 do not have to be non-overlapping. However, the bin-width, like the width parameter in kernel smoothing, is the most important and sensitive tuning parameter in this algorithm.

4 Within-slide normalization of gene expression data from microarrays

With the rapid progress of genomic-scale sequencing, complete DNA sequences of some organisms are available, and other genomes can be sequenced in a fairly reasonable time period. Genes on DNA sequences – the blueprint of the life – are the basic biological elements. However, understanding genomic information is much more challenging. A further study of functionalities of genes necessitates the tracking of their dynamic expressions in living organisms. The current method to measure the abundance of mRNA for a specific gene makes use of reverse transcription to its complementary DNA (cDNA), followed by hybridization. The cDNA microarray technique prints thousands of genes on a microscope slide and produces snapshots of gene expression profiles at specific times for specific samples; see Schena, Shalon, Davis, and Brown [24]. A comparison strategy is adopted in cDNA microarray; that is, relative gene expression levels of one sample are measured with respect to a reference. The idea is implemented by a dye technique: label cDNAs from a sample and its reference by two different fluorescent dyes, typically Cy3 (green) and Cy5 (red). Our focus is the difference on the logarithm (base 2) scale of every pair of expression levels corresponding to the same spot on a slide (probe). Let us denote the logarithm of expression levels of the sample and reference at the i-th spot by the pair (U_j, V_j), and denote the logarithm of their measured fluorescence intensities by $(\tilde{U}_j, \tilde{V}_j)$. Ideally, we expect that $(\tilde{U}_j, \tilde{V}_j)=(U_j, V_j)$ except for an offset constant. In practice, non-constant measurement bias occurs because of factors such as physical properties of dyes (heat and light sensitivity, relative half-life), efficiency of dye incorporation, experimental variability in probe coupling and processing procedure, and scanner settings at the data collection step. In order to improve the quality of microarray data, a normalization procedure to adjust the measurement is required.

Sources of variability can be classified into two categories: internal and external with respect to each slide. The effects of external factors are potentially detectable and estimable with multiple-slide data if the experiment is well designed. However, the effects of internal factors are confounded with each slide and thus an adjustment procedure adaptive to each slide is indispensable for the reconstruction of the raw expression levels. Yang, Dudoit, Luu, and Speed [26] proposed an ingenious method – within-slide normalization – to solve the problem. In the following, we have another look at the problem and their normalization procedure from the perspective of BIND. Consider a system with (U_j, V_j) as input and $(\tilde{U}_j$ and $\tilde{V}_j)$ as output. Let $\mathbf{h} = (h_1, h_2)$ be the transformation function; namely,

$$\begin{cases} \tilde{U}_j &= h_1(U_j, V_j) \\ \tilde{V}_j &= h_2(U_j, V_j). \end{cases} \tag{3}$$

The goal is to reconstruct the input variables (U_j, V_j) based on the output variables $(\tilde{U}_j, \tilde{V}_j)$. The system function $\mathbf{h} = (h_1, h_2)$ represents the effect caused by all internal

factors. In fact, we can also include external factors if we have no other better way to estimate them. Obviously this is a blind inversion problem. The BIND scheme leads us to the question: what is the distribution of input, the true expression levels? First, let us suppose that the sample and the reference are identical and that the difference of their expression levels is purely caused by random and uncontrolled effect. In this ideal case, we assume that the random variables $\{(U_j, V_j), j = 1, \cdots\})$ are independent among pairs and within each pair, and they are distributed according to $F(u_j - a_j)$ and $F(v_j - a_j)$, where $F(\cdot)$ is a distribution symmetric about zero and a_j is the average expression level of the j-th gene. If we look at their joint distribution by the scatter plot of U versus V, then we should see that the points cluster around the straight line $V = U$. The average deviation of the points from the straight line measures the precision of the experiment. We denote this joint distribution by Ψ. If the effective measurement system $\mathbf{h}$ is not an identity one, then the distribution of the output, denoted by Ψ', could be different from Ψ. This is exactly what is reported by Dudoit, Yang, Callow and Speed; see Figure 2 in [7].

Next we go back to real practice. Nowadays each slide contains more than a few thousand genes. Suppose that only a small proportion α of the genes are differentially expressed while expressions of the other genes are unchanged except for random fluctuations. Consequently, the distribution of the input in the blind inversion story is a mixture of the two components. One component consists of those unchanged genes, and its scatter plot is similar to Ψ. The other component consists of the differentially expressed genes and is denoted by Γ. Although the cloud shape of Γ in its scatter plot is difficult to find out, its contribution to the input is at most α. The scatter plot of the input variables (U_j, V_j) is a superimposition of those of Ψ and Γ weighted respectively by $1 - \alpha$ and α. We assume that the system function $\mathbf{h}$ is a 1-1 transform. Under $\mathbf{h}$, Ψ and Γ are transformed into distributions denoted respectively by Ψ' and Γ'; that is, $\Psi' = \mathbf{h}(\Psi)$, $\Gamma' = \mathbf{h}(\Gamma)$. This implies that the distribution of the output $(\tilde{U}_j, \tilde{V}_j)$ is $(1-\alpha)\Psi' + \alpha\Gamma'$. If we can separate the two components Ψ' and Γ', then the transform $\mathbf{h}$ of some specific form could be estimated from the knowledge of Ψ and Ψ'. An appropriate estimate $\hat{\mathbf{h}}$ of the transform should satisfy the following: the distribution of $\hat{\mathbf{h}}^{-1}(\Psi')$ is similar to that of Ψ, which centers around the line $V = U$. In other words, the right transform straightens out the distribution cloud of Ψ'. Yang, Dudoit, Luu, and Speed [26] first rotate the coordinate system by 45° as follows,

$$\begin{cases} X &= (U+V)/2 \\ Y &= U-V. \end{cases}$$

After the rotation, the conditional distribution of Y given X should be symmetric about zero. In the scatter plot of (X, Y), the cloud should horizontally center around zero. Each measurement pair $(\tilde{X}_j, \tilde{Y}_j)$ is a transformed version of (X_j, Y_j); we denote $\tilde{X}_j = g_1(X_j, Y_j)$ $\tilde{Y}_j = g_2(X_j, Y_j)$ as in (3). To make the system function estimable, we let $g_1(x,y) = x$ and $g_2(x,y) = g_2(x)$, a free function with some kind of smoothness. Thus the problem becomes a regression problem of $\tilde{Y}$ versus $\tilde{X}$, either in a parametric or in

a nonparametric form. Yang, Dudoit, Luu, and Speed [26] proposed to use **lowess**, a robust local linear regression technique (see Cleveland [6]), to remove the component of Γ' and estimate the transform function $\mathbf{g}$. Once it is estimated, we apply its inverse $\hat{\mathbf{g}}$ to the observations and obtain a reconstruction of the expression difference for each probe. Another **stratification** strategy is adopted in combination with **lowess** smoothing in [26]. That is, the data in one microarray are grouped according to the spatial setup of array printing so that data within each group share a more similar bias pattern. Next, the above normalization procedure is applied to each group; this is referred to as within-print-tip-group normalization in [26]. The above argument provides an interpretation of the within-slide normalization from the perspective of BIND. As a consequence, we see that one justification of the procedure lies in **the hypothesis on the joint distribution of the true gene expression levels of a sample and its reference**.

5 Discussion

5.1 A BIND story in seismology: predictive deconvolution

Various cases of blind deconvolution are reported in the literature; see Li [15] for a recent example and for references. We note that they belong to the class of blind inversion problems, and it is the input distribution that comes to help. We briefly discuss one example, the method of predictive deconvolution used in seismic trace processing, because of its scientific merit. The seismic reflection method aims to determine the distance and directions of remote and inaccessible bodies within the Earth, which is of great importance to oil exploration and other geophysical applications. The basic scheme of the seismic data collection process is the following: active sources of energy such as dynamite, air guns, and chirp signals generators at the surface of the Earth are used to produce waves of some form; the waves propagate downward from the sources into the Earth; at the interfaces between geologic layers in the Earth's crust, part of the waves are transmitted while the rest are reflected; eventually some waves propagate upward to the surface of the Earth and can be detected by receivers located at various distances from the source. The recorded traces of the received waves make up the seismogram. More background can be found in Robinson [17, 18], and Robinson and Durrani [19]. One important problem in seismic data processing can be formulated as follows. Denote the seismic traces by a time series $u(k)$, $k = 1, \cdots, T$. At a stage of the processing (after signature deconvolution), we can postulate a convolution model:

$$u(k) = f(k) * v(k) = \sum_{i=0}^{+\infty} f(k-i)v(i), \tag{4}$$

where $f(k)$ is the reverberation waveform and $v(k)$ is the reflectivity function or the reflectivity coefficient at each layer. Let $U(z)$, $F(z)$, and $v(z)$ be the z-transforms of $u(k)$, $f(k)$, and $v(k)$, respectively. Then we have $U(z) = F(z)V(z)$. We can regard Equation (4) as a linear system, in which $v(k)$ and $u(k)$ respectively play the role of unknown

input and known output. According to the **feedback hypothesis**, the reverberation filter $f(k)$, which characterizes the system, is not only causal but also minimum delay. Specifically, the feedback hypothesis assumes that the reverberation effect takes a form of finite feedback filter, namely,

$$F(z) = \frac{1}{1 + \alpha_1 z + \alpha_2 z^2 + \cdots + \alpha_p z^p},$$

whose zeros are outside of the unit circle on the complex plane. The reverberation refers to the fact that waves are successively reflected between two interfaces of a layer. In practice, reverberations are often generated by such a complicated physical situation with many layers that the effective filter is impossible to be obtained by direct measurement. Thus in order to unravel the seismic traces, we need to estimate both the system filter $F(z)$ and the input – the reflection coefficient. It is clear that this is a blind inversion problem. According to the scheme of BIND, we first inquire for the distribution of the reflection coefficients. Fortunately, several studies showed that the **random hypothesis** is approximately valid in many cases; see Robinson [17], page 278 and the references mentioned there. The **random hypothesis** assumes that the reflection coefficients $v(k)$ follow a white noise stochastic process. That is, $E[v(k)] = 0$, $E[v(k)v(j)] = E[v(k)]E[v(j)] = \sigma^2\delta_{j,k}$, if $k \neq j$. As a matter of fact, the second order statistical property of a stationary stochastic process is characterized by its autocorrelation coefficients. If we further assume that $v(k)$ is normal, then the distribution of the input is uniquely determined because the higher-order cumulants of a normal distribution are zero; see Rosenblatt [20]. We note that the normal assumption is not required by the algorithm of predictive deconvolution. A consequence of the **random hypothesis** is that the output, seismic traces, is a zero mean stationary stochastic process whose second order statistical property is characterized by its autocorrelations, denoted by $\gamma_{uu}(k)$, $k = 0, 1, \cdots$. This set of statistical correlations relates to the reverberation filter through the Yule-Walker equation:

$$\begin{pmatrix} \gamma_{uu}(0) & \gamma_{uu}(1) & \cdots & \gamma_{uu}(k-1) \\ \gamma_{uu}(1) & \gamma_{uu}(0) & \cdots & \gamma_{uu}(k-2) \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_{uu}(k-1) & \gamma_{uu}(k-2) & \cdots & \gamma_{uu}(0) \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_k \end{pmatrix} = \begin{pmatrix} \gamma_{uu}(1) \\ \gamma_{uu}(2) \\ \vdots \\ \gamma_{uu}(k) \end{pmatrix}. \quad (5)$$

In practice, we plug into this equation estimates of $\gamma_{uu}(k)$ based on data and apply the Levinson-Durbin algorithm [8] to obtain an estimate of the filter denoted by $\hat{\alpha}_k$. The Burg algorithm alternatively estimates the filter directly from the data. The determination of the order p in the filter is a problem of model selection and the technique of AIC or BIC can be applied. With the estimated filter and starting values $u(1), \cdots, u(q)$ obtained from seismic and numerical considerations, we reconstruct the reflection coefficients using the feedback procedure

$$v(k) = u(k) + \sum_{i=1}^{p} \hat{\alpha}_i u(k-i).$$

This is the so-called predictive deconvolution. Except for the specific time series techniques involved, it is consistent with the BIND philosophy as used in previous examples.

5.2 Statistical assessment

5.2.1 Goodness of match with the input distribution

To assess a BIND procedure, we examine the distribution of the reconstructed input and see if it matches the hypothesis. In the DNA sequencing example, we check the *cc-number* achieved at the exit of Algorithm 2 and recognize a satisfactory color correction if it is below a threshold. A graphical check of the scatter plots of the color-corrected data like Figure 3 might be expected in some cases. For cDNA microarray example, we can similarly define a quality number as the maximum absolute value of the regression line obtained by **lowess**, in which the parameters are chosen as those in [26]. The graphical check could be sufficient for many biologists. For the illustrative example (1), we test the independence of the four components; this can be carried out by the likelihood ratio test or other tests, see Chapter 9, Anderson [2]. For the seismic trace example, we check the whiteness of the reconstructed reflection coefficients by examining the flatness of its periodogram or using other statistical tests; see Brockwell and Davis [4].

5.2.2 System sensitivity analysis by data self-perturbation

The reconstruction of input in the BIND scheme is associated with the problem of system estimation. As in any other estimation problem, it is valuable to assess the accuracy of the estimates of the system. One technique in this regard is the **bootstrap**; see Efron and Tibshirani [9]. In the DNA sequencing example, we can generate bootstrap samples by sampling from the raw dataset with replacement and applying the same color-correction algorithm (with the same set of parameters) to each bootstrap sample. The bias and standard deviation of the bootstrap estimates reflect systematic bias and variability of the algorithm with respect to data self-perturbation. It is possible that these statistics contain some scientific meaning and could provide some guidance for researchers. For the DNA sequencing example, we show the result of a bootstrap study with 200 replicates in Table 1. It includes the bias, standard error, and coefficient of variation for each of the 16 parameter estimates – here we switch from the parameterization of the cross-talk matrix in (2) to the one whose columns sum to unity. The estimates are almost unbiased for all the four dyes. The SDs and CVs measure the stability of the four dyes. For example, the comparatively larger SDs and CVs of the estimates regarding the fourth dye associated with T indicate that its physical and chemical properties are not as stable as others. Our collaborator Dr. Kheterpal (she was with Prof. Mathies' group in the Chemistry Department at UC, Berkeley during our collaboration) verified this observation. This sensitivity analysis by bootstrap applies

to other parametric systems such as the example in (1). However, the technique of data self-perturbation needs special care for systems with a nonparametric form, such as that in the microarray example and for systems with a spatial structure, such as that in the seismology example.

Table 1: Accuracy assessment of the estimate by bootstrap ($\times 10^{-3}$)

	C				G			
	mean	bias	SD	CV	mean	bias	SD	CV
1	333	-12	14	42	203	-1	3	15
2	330	0	7	21	412	1	4	10
3	241	9	10	42	296	0	4	13
4	96	2	6	63	88	0	4	46
	A				T			
	mean	bias	SD	CV	mean	bias	SD	CV
1	70	0	10	143	115	-3	11	96
2	209	5	12	57	139	4	26	187
3	544	-3	16	30	183	-4	17	93
4	176	-2	6	34	562	3	50	89

5.3 Final remarks

Among the many things I learned from Prof. Terry Speed through the years of my study under his supervision, the one that impressed me very much was his conscientious service to the scientific community, especially in genetics and molecular biology, as a statistician. His broad and dynamically-changing research interests are phenomenal. His extensive collaborations with biologists constantly bring research life-blood into his students' study. We also notice that he earned so much respect not only from his statistician colleagues but also from researchers in other fields.

This article is motivated by Terry's advocacy of **considering scientific meanings in mathematical and statistical modeling**. The abstraction of BIND provides a way to think about a scientific problem mathematically as well as a way to think about mathematics scientifically. The BIND scheme hinges on a hypothesis on the distribution of system input. The verification of this hypothesis requires careful consideration, and it varies from one problem to another. No matter how BIND is implemented, either by an algorithm from numerical recipes or by a novel procedure, the bottom line is: the distribution of the reconstructed output should match the hypothesis. It is our hope that the BIND notion can help statisticians apply their toolbox to more scientific measurement problems in the future.

Acknowledgments

Prof. Terence P. Speed and Dr. David O. Nelson provided the author with great help in this work. The research is supported by the NSF grant DMS–9971698, and DOE grant DE–FG03–97ER62387. The author would also like to acknowledge help provided by the Institute of Pure and Applied Mathematics, UCLA.

Lei Li, Molecular and Computational Biology, University of Southern California, 1042 West 36th Place, DRB 289, Los Angeles, CA 90089-1113, lilei@hto.usc.edu

References

[1] M. D. Adams, C. Fields, and J. C. Ventor, editors. *Automated DNA sequencing and analysis*. Academic Press, London, San Diego, 1994.

[2] T. W. Anderson. *An Introduction to Multivariate Statistical Analysis*. John Wiley & Sons, 2nd edition, 1984.

[3] P. Billingsley. *Probability and Measure*. John Wiley & Sons, 1986.

[4] P. J. Brockwell and R. A. Davis. *Time Series: Theory and Models*. Springer-Verlag, 1991.

[5] S. E. Cawley. *Statistical models for DNA sequencing and analysis*. PhD thesis, University of California, Berkeley, 2000.

[6] W. S. Cleveland. Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74:829–836, 1979.

[7] S. Dudoit, Y. H. Yang, M. J. Callow, and T. P. Speed. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica*, 12:111–140, 2002.

[8] J. Durbin. The fitting of time series models. *Rev. Inst. Internat. Statist.*, 28:233–244, 1960.

[9] B. Efron and R. J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall: New York, London, 1993.

[10] G. R. Fowles. *Introduction to Modern Optics*. Halt, Rinehart And Winston, Inc., 2nd edition, 1975.

[11] G. H. Golub and C. F. Van Loan. *Matrix Computations*. The John Hopkins University Press, Baltimore and London, 1996.

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

[13] L. Li and T. P. Speed. An estimate of the color separation matrix in four-dye fluorescence-based DNA sequencing. *Electrophoresis*, 20:1433–1442, 1999.

[14] L. Li and T. P. Speed. Parametric deconvolution of positive spike trains. *Annals of Statistics*, 28:1279–1301, 2000.

[15] T. H. Li. Blind deconvolution of linear system with multilevel nonstationary input. *Annals of Statistics*, 23:690–704, 1995.

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

[17] E. A. Robinson. *Physical Applications of Stationary Time-Series*. Macmillan Publishing, 1980.

[18] E. A. Robinson. *Seismic Inversion and Deconvolution, Part A: Classical Methods*. Geophysical press, 1984.

[19] E. A. Robinson and T. S. Durrani. *Geophysical Signal Processing*. Prentice Hall, 1986.

[20] M. Rosenblatt. *Stationary Sequences and Random fields*. Birhhäuser, 1985.

[21] P. J. Rousseeuw. Least median of squares regression. *Journal of the American Statistical Association*, 79:871–881, 1984.

[22] P. J. Rousseeuw and A. M. Leroy. *Robust Regression and Outlier Detection*. Wiley, New York, 1987.

[23] P. J. Russell. *Genetics*. Harpercollins College Publisher, New York, 1995.

[24] M. Schena, D. Shalon, R. M. Davis, and P. O. Brown. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 270:467–470, 1995.

[25] D. L. Souvaine and J. M. Steele. Time- and space-efficient algorithms for least median of squares regression. *Journal of the American Statistical Association*, 82:794–801, 1987.

[26] Y. H. Yang, S. Dudoit, P. Luu, and T. P. Speed. Normalization for cDNA microarray data. In M. L. Bittner, Y. Chen, A. N. Dorsel, and E. R. Dougherty, editors, *Microarrays: Optical Technologies and Informatics*, Proceedings of SPIE. 2001.