Deconvolution of sparse positive spikes: is it ill-posed?

Technical Report No. 586

Lei Li*
Florida State University
Institute for Pure and Applied Mathematics, UCLA

Terence P. Speed†
University of California, Berkeley
Walter and Eliza Hall Institute

October 10, 2000

Abstract

Deconvolution is usually regarded as one of the so called ill-posed problems of applied mathematics if no constraints on the unknowns can be assumed. In this paper, we discuss the idea of well-defined statistical models being a counterpart of the notion of well-posedness. We show that constraints on the unknowns such as non-negativity and sparsity can help a great deal to get over the inherent ill-posedness in deconvolution. This is illustrated by a parametric deconvolution method based on the spike-convolution model. Not only does this knowledge, together with the choice of the measure of goodness of fit, help people think about data (models), it also determines the way people compute with data (algorithms). This view is illustrated by taking a fresh look at two familiar deconvolvers: the widely-used Jansson method, and another one which is to minimize the Kullback-Leibler distance between observations and fitted values. In the latter case, we point out that a counterpart of the EM algorithm exists for the problem of minimizing the Kullback-Leibler distance in the context of deconvolution. We compare the performance of these deconvolvers using data simulated from a spike-convolution model and DNA sequencing data.

Abbreviated title: deconvolution

AMS 1991 subject classification: Primary 62F10; Secondary 62F12, 86A22.

Key words and phrases: ill-posed, deconvolution, parametric deconvolution, Kullback-Leibler, EM, Kuhn-Tucker

*Supported by NSF grant DMS-9971698
†Supported by DOE grant DE-FG03-97ER62387

1 Introduction and background

Many signals from natural phenomena and advanced scientific instruments can be approximately described by the following model,

$$y(t) = w(t) * x(t) = \int_{-\pi}^{\pi} w(t-s)\, x(s)\, ds\,, \tag{1}$$

where $x(t)$ is the unknown function of interest, $w(t)$ is a known point spread function, and the observed function $y(t)$ is recorded over a finite period $[-\pi, \pi]$. The task of deconvolution is to reconstruct x from the observations on y, the knowledge of w and any prior knowledge concerning x. We make two remarks here. First, the convolution relation is frequently an approximation to a true but more complex relationship. Second, data are always observed in the presence of measurement errors. Different variants of the deconvolution problem arise in many areas such as spectroscopy, chromatography, tomography, geophysics, seismology and pharmacokinetics. We also see it in the form of equalization in communication theory, and deblurring in image analysis.

Our work on deconvolution was motivated by the problem of base-calling in DNA sequencing. The current 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 sequencing, which attempts to recover the underlying DNA sequence from the vector time series. Figure 1 and 2 illustrate two segments of a signal from a single DNA sequencing run. The data shown here have been preprocessed by color-correction to eliminate the so called "cross talk" phenomenon, see [26, 19] for recent progress in this issue. In each of the two figures, several peaks can be observed within each channel. Typically, each major peak in the series corresponds to one base. Channels 1, 2, 3 and 4 correspond to bases C, G, A and T respectively. If we superimpose the four channels, then we obtain a train of peaks. The rationale of base calling is that the order of peaks from the four channels should agree with the order of bases on the underlying DNA fragment. In Figure 1, the order of the underlying bases is: CGTAGGACTTAGATGTTCTGTGATATCGCCTGGGT. This segment is easy to handle, since it comes from the beginning of a sequencing trace. 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. This can be seen in Figure 2. Some research showed that errors associated with runs of the same base constitute more than half of the total errors in base-calling made by one widely-used system, see [2, 20, 23]. Furthermore, this kind of error is more difficult to deal with than other kinds in the assembling and editing step of DNA sequencing. For this reason, it is worth making a greater effort in the regions with runs of the same peaks.

The point spread function in (1) usually represents a blurring effect. If we assume that a sparse virtual spike train corresponds to occurrences of one of the four kinds of bases, then the observations in the the corresponding channel can be approximated by a convolution of this positive spike train, denoted by x, with a fixed point spread function denoted by w. Thus the convolution relation shown in (1) may be suitable for modeling DNA sequencing data, with different point spread functions being appropriate in different regions of the run.

Deconvolution is one of the so called ill-posed problems of applied mathematics if no further constraints on the unknowns can be assumed. In this paper, we discuss the general idea of well-defined statistical models being a counterpart of the notion of well-posedness. We show that constraints on the unknowns such as non-negativity and sparsity can help a great deal to get over the inherent

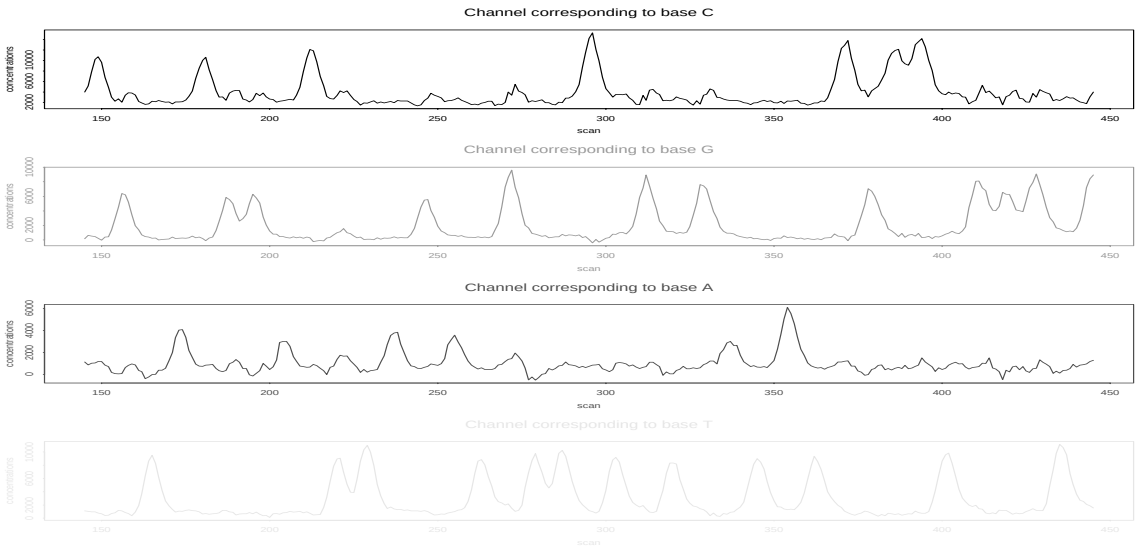

Figure 1: A segment of slab gel electrophoresis sequencing data from near the beginning (provided by the engineering group at Lawrence Berkeley National Laboratory). Each sub-plot of a run contains a component of the vector time series. Since the data has been color-corrected, the vertical scales are concentrations.

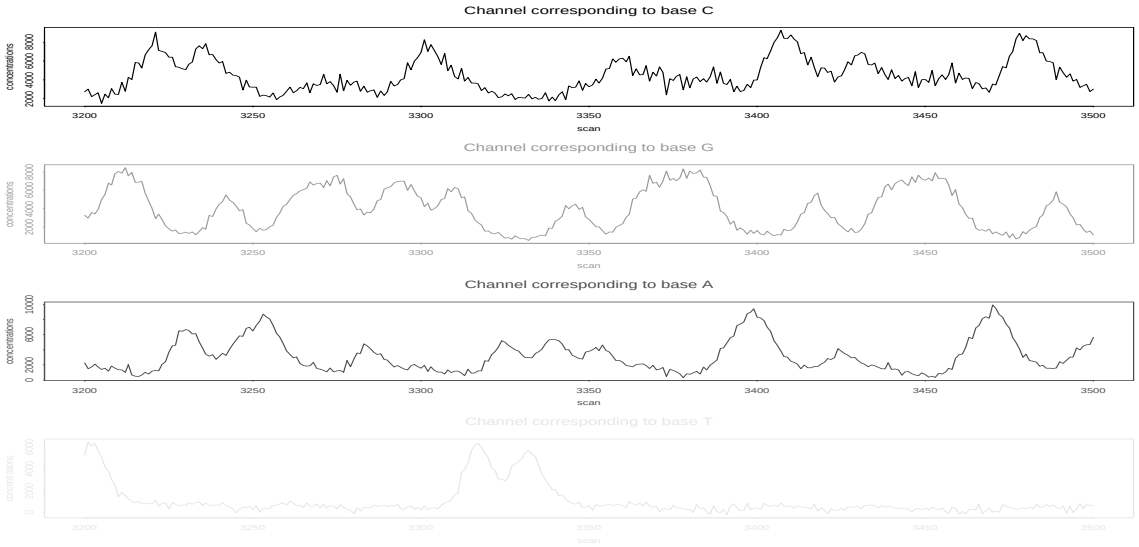

Figure 2: A segment of slab gel electrophoresis sequencing data from near the end of a run (cf. figure 1 above).

ill-posedness in deconvolution. This is illustrated by a parametric deconvolution method based on the spike-convolution model. Not only does this knowledge, together with the choice of the measure of goodness of fit, help people think of data (models), it also determines the way people compute with data (algorithms). This view is illustrated by taking a fresh look at two familiar deconvolvers: the widely-used Jansson method, and another one which is to minimize the Kullback-Leibler distance between observation and fitted values. In the latter case, we point out that a counterpart of the EM algorithm, to which we refer by the term synchronization and minimization, exists for the problem of minimizing the Kullback-Leibler distance in the context of deconvolution. We arrange the material as follows. In section 2, we discuss the general philosophy, and give our point of view on the problem of deconvolution of positive sparse spikes. In section 3, we discuss one concrete version of this view, namely, the spike-convolution model and the associated parametric deconvolution method. In Section 4 and 5, we look at Jansson's method, and deconvolution by minimizing the Kullback-Leibler distance via the synchronization-minimization algorithm, respectively. In section 6, we compare the performance of various deconvolvers using two numerical examples, one based on the data simulated from a spike-convolution model, and one based on DNA sequencing data.

2 General philosophy

2.1 A statistician's counterpart of the notion of well-posedness

Technically, we assume the following throughout this section in (1): the point spread function $w(\cdot)$ has finite support $(-\kappa, \kappa)$, where $0 < \kappa < \pi$ and $w(\cdot) \in C^2[-\pi, \pi]$; $x(t)$ has support in $[-\pi+\kappa, \pi-\kappa]$. In DNA sequencing, we can cut the raw data into pieces at valley points, and deconvolve each piece separately. By doing so, we can not only make the assumption that there are no significant signals near the two ends as mentioned above, but we can also assume the point spread functions are more or less homogeneous in terms of their spread.

Let us ignore measurement errors and focus on the signal structure for the moment The continuous version of the deconvolution problem (1) can be regarded as one of solving a Fredholm integral equation of the first kind as follows:

$$\int_{-\pi}^{\pi} w(t,s)x(s)\,ds = y(t)\,, \tag{2}$$

if the kernel $w(t,s)$ is taken to be $w(t-s)$. The structure of Fredholm equations of the first kind is revealed by the singular value decomposition theorem. If $\int_{-\pi}^{\pi}\int_{-\pi}^{\pi} w^2(t,s)\,dtds < \infty$, then the integral operator $\mathcal{W}$ as defined in the equation (2) is a compact operator from $L^2[-\pi,\pi]$ to $L^2[-\pi,\pi]$. Its adjoint operator $\mathcal{W}^*$ defined by $\int_{-\pi}^{\pi} w(s,t)x(s)\,ds = y(t)$, is also compact. In this case, there exist two orthogonal sequences $\{\phi_k\}$ and $\{\psi_k\}$ in $L^2[-\pi,\pi]$ such that $\mathcal{W}\phi_k = \mu_k\psi_k$, $\mathcal{W}^*\psi_k = \mu_k\phi_k$ for all positive integers n, where $\mu_1 \geq \mu_2 \geq \cdots > 0$ are the singular values of $\mathcal{W}$. If there are infinite many nonzero μ_k's, then zero is their only cluster point. Let $\mathcal{Q}$ be the orthogonal projection from $L^2[-\pi,\pi]$ onto the null space $N(\mathcal{W})$. Then for any $x(\cdot) \in L^2[-\pi,\pi]$, we have the following decomposition:

$$x = \sum_{k=1}^{\infty} < x, \phi_k > \phi_k + \mathcal{Q}x, \qquad\qquad \mathcal{W}x = \sum_{k=1}^{\infty} \mu_k < x, \phi_k > \psi_k\,.$$

The equation of the first kind $\mathcal{W}x = y$ is solvable if and only if $y \in N(\mathcal{W}^*)^{\perp}$ and

$\sum_{k=1}^{\infty} \frac{1}{\mu_k{}^2} | < y, \psi_k > |^2 < \infty$. In this case, a solution is given by

$$x = \sum_{k=1}^{\infty} \frac{1}{\mu_k} < y, \psi_k > \phi_k \,. \tag{3}$$

This is in fact the result of Picard Theorem, see [21, 5]. A reproducing kernel view of this structure can be found in Wahba [42] and references mentioned there. As can be seen, the problem of solving a Fredholm equation of the first kind is ill-posed because one of the three well-posedness conditions is violated. According to Hadamard [12], a problem in mathematical physics is well-posed if it satisfies the following three properties: the existence of a solution, the uniqueness of the solution, and the continuous dependence of the solution on the data; otherwise, it is called ill-posed. The deconvolution problem (1) as a special case of (2) is generally an ill-posed one if no further knowledge can be assumed. To see this more clearly, let us denote the Fourier coefficients of x, w and y by $\{\check{x}_k\}$, $\{\check{w}_k\}$ and $\{\check{y}_k\}$ respectively. By the convolution theorem, we have $\check{y}_k = \check{w}_k \check{x}_k$. Now the Fourier base $\{e^{ikt}\}$ plays the roles of ϕ_k and ψ_k in the singular decomposition as described above. We expect one of the following would happen.

1. For some k, $\check{w}_k = 0$, but $\check{y}_k \neq 0$. In this case, the solution does not exist.

2. For some k, $\check{w}_k = 0$, and $\check{y}_k = 0$. In this case, the solution is not unique for we cannot reconstruct the information contained in the corresponding term of $\check{x}_k$.

3. None of $\check{w}_k$ is zero, but $\check{w}_k \to 0$ as $k \to 0$ according to the smoothness condition we impose on w. In this case, the solution x is not stable with respect to small perturbation of y.

The point spread functions in deconvolution are usually low pass linear filters, and do indeed have Fourier coefficients which are zero or close to zero in the high frequency range.

The method of regularization introduced by Tikhonov [37], also see [38], is widely used in numerical analysis to deal with ill-posed problems. One major idea underlying the regularization method is to try to damp out the influence of the factor $1/\mu_k$ in (3). O'Sullivan [31] provided a statistical perspective of the method of regularization, and identified the tools for assessing the performance characteristics of an inversion algorithm.

The convolution structure in (1) is frequently an approximation to a true but more complex relationship. Besides, data are always observed in the presence of measurement errors. Thus the idea of statistical modeling could be helpful in these problems. In a well defined statistical model, assumptions are made about the elements involved in a system which reflect the best of our knowledge. Inference can then be made based on data and the model structure. In fact, statisticians have the counterpart of concepts of Hadamard's well-posedness. In a well defined statistical model, usually we can find: 1. identifiability of the unknowns; 2. existence of an unbiased estimate or something close to this; 3. reasonable stability of estimates in the statistical sense—variance, robustness, etc. We make several remarks here. First, the unknowns, either in the form of parameters or random variables, are not necessarily points in Euclidean space. They could be defined as equivalence classes whose elements share some common feature defined by researchers, and they could be represented by typical elements in the classes or by some other mathematical device. Second, in many cases, especially in nonparametric statistical problems, the bias and variance cannot be controlled simultaneously, and a trade-off has to be made when constructing estimates. Third, we can measure the degree of well-posedness or ill-posedness using

the achievable rate of convergence of the estimates towards the unknowns, one such example can be found in Fan [9]. Finally, it is quite possible that we cannot propose legitimate or well defined statistical models in the early stages of some piece of scientific research. There are two possibilities. It could be that we have not found the appropriate mathematical tool. Or it could be that we do not have enough information to build up a model for the purpose of the research. In the latter case, more "shoe leather" seems to be the only way to carry on the work, though usually this is a hard job, see Freedman [10]. Statisticians can still play an useful role here for we can provide guidance as what information could be relevant and where the labor should be spent. Confidence in a hypothetical model relies on both a good understanding of the relevant mathematical structures and the solid scientific evidence supporting the model's assumptions.

2.2 Deconvolution of positive sparse spikes

Sometimes we consider a discretized version of (1):

$$y(t_k) = \int_{-\pi}^{\pi} w(t_k - s)\, x(s)\, ds \approx \sum_{j=-[n/2]+l}^{[(n-1)/2]-l} w(t_k - s_j)\, x(s_j), \tag{4}$$

where y is observed at t_k, and (t_k) and (s_k) are the lattice points $2\pi k/n$, $k = -[n/2], \cdots, [(n-1)/2]$, $l = [2\pi\kappa/n]$. Let $\mathbf{y} = (y(t_k))'$, $\mathbf{x} = (x(s_j))'$, and $\mathbf{W} = (w(t_k - s_j))$. Then we can rewrite (4)

$$\mathbf{y} = \mathbf{W}\,\mathbf{x}. \tag{5}$$

Note that $\mathbf{W}$ is only a n by $(n - 2l)$ matrix, for we have removed from $\mathbf{x}$ the zero elements near the ends of the interval. Later we also use the notation y_k, x_j, $w_{k,j}$, where $k = 1, \cdots, n$, $j = 1, \cdots, n - 2l$, and it is understood that they are the elements of $\mathbf{y}$, $\mathbf{x}$ and $\mathbf{W}$. Usually the design matrix $\mathbf{W}$ is very-ill conditioned, and so direct least squares method does not apply. One immediate modification is the ridge regression estimate given by $(\mathbf{W}'\mathbf{W} + \lambda\mathbf{I})^{-1}\mathbf{W}'\mathbf{y}$, which is the minimizer of $(\mathbf{y} - \mathbf{W}\,\mathbf{x})'(\mathbf{y} - \mathbf{W}\,\mathbf{x}) + \lambda\mathbf{x}'\mathbf{x}$. That is, we impose a penalty on the "energy" of the unknowns while minimizing the residual sum of squares. This is one example of the regularization method. It is known in statistics that the ridge estimate can give a considerably smaller variance at the price of a slightly larger bias.

In order to have a better understanding of this problem, we ask ourselves a few fundamental questions: How should we measure the goodness of a solution? Is there any other information available on the unknowns x? If so, how can be use it? One way of answering the first question is to select a measure of distortion between $\mathbf{y}$ and $\mathbf{W}\,\mathbf{x}$. Candidates include the L_2 norm, L_1 norm, or in general a L_p norm with $1 \le p \le \infty$. A deeper yet difficult question relevant to the discussion here is: what is the distortion measure used by human perception? As for the second question, in many situations the unknowns represent some kind of "energy" produced by nature or a relevant experiment. Thus we might describe them by a positive spike train, possibly with various inter-arrival times and amplitudes. In addition, the spikes—the "signal" to be detected or monitored—are sparse in many applications, though they could be quite close to each other. Roughly speaking, sparsity in the formulation (5) means that the number of nonzero elements in the unknown is comparatively small in relation to the number of observations. We note that the above two questions are related to one another. For example, if we can assume that all the elements of $\mathbf{x}$, $\mathbf{y}$ and $\mathbf{W}$ are nonnegative, then we can use the Kullback-Leibler divergence to measure the

distance between $\mathbf{y}$ and $\mathbf{W}\,\mathbf{x}$. In this exposition, our focus is on the deconvolution of positive sparse spikes motivated by the DNA sequencing data. We feel an effective deconvolver should take into account known features of unknowns and appropriately address issues such as the distortion measure. Consideration of these issues determines not only the way we think about data, but also the the way we compute with the data. From the perspective of implementation, a deconvolver is essentially a suite of algorithms, and algorithms that are good in terms of convergence, complexity, use of memory, and ease of automation are extremely important in applications such as DNA sequencing. Indeed, in the literature some researchers go directly to the design of algorithms based on their previous experience. By contrast other researchers, especially statisticians, build statistical models, and propose algorithms based on them. But even in the first case, the algorithms used in a deconvolver should be consistent with the researchers' experience, and this experience might be regarded as a kind of "model", though not one rigorously formulated in mathematical terms. In the next three sections, we study several deconvolvers which are quite distinct in their mathematical formulations and algorithms. The spike-convolution model described in Section 3, which explicitly formulates the positivity and sparsity by assuming the unknowns have the form of a positive Dirac train, is a well-defined statistical model in the sense that it is identifiable and there exists consistent estimates of its parameters. If the spike-convolution model were a good approximation to the problem being studied, this suggests that the problem of deconvolution of positive sparse trains be not ill-posed.

3 The spike-convolution model and parametric deconvolution

With DNA sequencing data in mind, we [27] proposed a specific form for the unknown signal x as follows.

$$x(t) = A_0 + \sum_{j=1}^{p} A_j\,\delta(t-\tau_j), \tag{6}$$

where $\delta(\cdot)$ is the Dirac delta function, and the coefficients A_j, referred as "heights" of the spikes, are positive. 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 denote the signal x in (6) by $SC(\delta; p; \boldsymbol{A}; \boldsymbol{\tau})$, and refer its convolution with w as in (1) by $SC(w; p; \boldsymbol{A}; \boldsymbol{\tau})$ We assume what can be observed in practice is a sample of the $SC(w; p; \boldsymbol{A}; \boldsymbol{\tau})$ corrupted by additive measurement errors:

$$z(t_k) = y(t_k) + \epsilon(t_k) = A_0 + \sum_{j=1}^{p} A_j\,w(t_l-\tau_j) + \epsilon(t_k), \tag{7}$$

where t_k are the lattice points $2\pi k/n$, $k = -[n/2], \cdots, [(n-1)/2]$. Here $\{z(t_k)\}$ denotes the observations, and the $\{\epsilon(t_k)\}$ are i.i.d. with $E(\epsilon(t_k)) = 0$, $Var(\epsilon(t_k)) = \sigma^2$, and a finite third moment. The features of the unknowns—non-negativity and sparsity, as discussed in subsection 2.2—are included in this parameterization. Besides, the model is defined on a continuous scale, and the use of Dirac deltas leaves the room for a high resolution deconvolution.

Within the setting of this model, deconvolution is a standard parameter estimation problem. The parameters in a spike-convolution model include the baseline, the error variance, and the number, locations and amplitudes of the spikes. This is why we term the deconvolution procedure proposed in [27] as parametric deconvolution of positive spikes (abbreviated by PDPS later). The notation of two inner products is useful. We define the inner product of two functions $y_1(t)$ and

$y_2(t)$, belonging to $L^2[-\pi,\pi]$, by $< y_1, y_2 >= \frac{1}{2\pi}\int_{-\pi}^{\pi} y_1(t)\, y_2(t)\, dt$. For functions $z_1(t)$ and $z_2(t)$ well defined at the lattice points t_k, we also define the following inner product $< z_1, z_2 >_n = \frac{1}{n}\sum_{k=-[\frac{n}{2}]}^{[\frac{n-1}{2}]} z_1(t_k)\, z_2(t_k)$. PDPS does not require that the point spread function be nonnegative, but does require that there be no hole in the Fourier transform of the point spread function. That is, the Fourier coefficients of w, $v_k =< w, e^{ikt} >$ are not zero for $k = 0\ , \pm 1,\ \cdots, \pm L_0$, where L_0 is the upper bound on the number of spikes.

The spike-convolution model $SC(w; p; \boldsymbol{A}; \boldsymbol{\tau})$ is identifiable, but as pointed in [27] it does have a irregular structure. It is very difficult to estimate all the parameters in one step because of their different roles in the model. PDPS uses residual sum of squares as the goodness of fit, and combines several ideas such as Toeplitz structures, regression and model selection techniques. The core of PDPS consists of two parts: model fitting and model selection, which are sketched in the following.

Algorithm 3.1 Model-fitting.
Starting with the empirical trigonometric moments $\hat{f}_k =< z, e^{ikt} >_n$, for any given nonnegative integer $m \le L_0$,

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

2. *Trigonometric moment estimates of spike locations: construct the Toeplitz matrix $\hat{G}_m = (\hat{g}_{j-i})$, and compute its smallest eigenvalue $\hat{A}_0^{(m)}$ (assuming its multiplicity is one), and corresponding eigenvector $\hat{\alpha}^{(m)} = (\hat{\alpha}_0^{(m)}, \cdots, \hat{\alpha}_m^{(m)})$. On the unit circle of the complex plane, find the m distinct roots of $\hat{U}^{(m)}(z) = \sum_{j=0}^{m} \hat{\alpha}_j^{(m)}\, z^j$, which we denote by $\{e^{i\hat{\tau}_j^{(m)}}\}$, $j = 1, \cdots, m$.*

3. *Eliminate those $\{\hat{\tau}_j^{(m)}\}$ falling outside $[-\pi+\kappa, \pi-\kappa]$, and denote the locations of the remaining spikes by $\{\bar{\tau}_j, j = 1, \cdots, \bar{m}\}$, where $\bar{m} \le m$.*

4. *Estimate the heights $\bar{A}_j$ corresponding to these spikes by minimizing*

$$\| z(t) - \bar{A}_0 - \sum_{j=1}^{\bar{m}} \bar{A}_j\, w(t - \bar{\tau}_j)\ \|_n^2\ . \tag{8}$$

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

The output of this algorithm is a $SC(w; \bar{m}; \bar{\boldsymbol{A}}^{(\bar{m})}; \bar{\boldsymbol{\tau}}^{(\bar{m})})$. The model selection procedure has two stages.

Algorithm 3.2 Two-stage model selection.

1. *First stage. Among all the $SC(w; \bar{m}; \bar{\boldsymbol{A}}^{(\bar{m})}; \bar{\boldsymbol{\tau}}^{(\bar{m})})$ models fitted by Algorithm 3.1, choose the one that minimizes the following*

$$MGIC_1(s) = \bar{\sigma}(s)^2 + \frac{c_1(n)\log n}{n} s\,, \tag{9}$$

where $\bar{\sigma}(s)^2$ is the quantity in (8), and $c_1(n) \ge 0$ is a penalty coefficient. Denote this model by $SC(w; \bar{m}_0; \bar{\boldsymbol{A}}^{(\bar{m}_0)}; \bar{\boldsymbol{\tau}}^{(\bar{m}_0)})$.

2. *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. Namely, starting from* $SC(w; \bar{m}_0; \bar{\boldsymbol{A}}^{(\bar{\boldsymbol{m}}_0)}; \bar{\boldsymbol{\tau}}^{(\bar{\boldsymbol{m}}_0)})$*, we delete the peak that is least significant in terms of a penalized sum of squares. Compare the two models according to the following statistic*

$$MGIC_2(s) = \breve{\sigma}(s)^2 + \frac{c_2(n)\log n}{n} s\,, \tag{10}$$

where $\breve{\sigma}^{(s)\,2}$ *is the sum of squares fitted by a model with* s *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.*

This two-stage model selection procedure is a bit subtle. Its goal 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, This is achieved by finding a "best overfitting" model in the first stage, and eliminating the false spikes in the second stage. Usually we impose smaller penalty in the first stage, i.e. $c_1(n) < c_2(n)$. The estimate obtained from PDPS is $\sqrt{n}$-consistent if the spike number is estimated correctly from the model selection procedure. Under the assumption of normal errors, we can further use Gauss-Newton algorithm to construct one-step estimate to improve the accuracy. This estimate is asymptotically normal and efficient in the sense that the variance of its asymptotic distribution is the inverse of the Fisher information matrix given by

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

where $\psi_{A_0} = 1$, $\psi_{A_j} = w(t - \tau_j)$, $\psi_{\tau_j} = -A_j w'(t - \tau_j)$, $j = 1, \cdots, p$. The likelihood structure in the spike-convolution model is similar to that of the model used in Poskitt et al. [32]. We may skip the maximum likelihood estimate or the one-step estimate tuning in applications, but the MLE, as a benchmark in terms of asymptotics, helps us to evaluate other procedures. For example, we found, in the simulation study in [27], that the bias and variance of PDPS parameter estimates are much closer to those of MLE compared with a direct trigonometric moment estimate, even though the number of spikes in the latter method is assumed to be known. Asymptotic efficiency theory also describes what resolution we can achieve based on this spike-convolution model. From the computational point of view, PDPS requires the computation of minimum eigenvectors of Toeplitz matrices with sizes determined by the possible number of spikes involved (usually up to 20 in DNA sequencing), the calculation of roots of polynomials restricted on the unit circle, and regressions. It is obviously more computationally intensive than the two methods discussed in the next two sections, but is not a problem at all with current computing power. In practice, we could even skip the second round of model selection, and all we need is to set a upper bound (and a lower bound if possible) for the number of spikes and the penalty constant in MGIC.

The identifiability of the spike-convolution model and the consistency of the estimates generated by PDPS imply that this model is well-defined. Hence the problem of deconvolution associated with this model is really not ill-posed. If this model is a reasonable one for the real problem, this analysis sheds lights on how far we can go. As for implementation, algorithms other than PDPS do exist in the literature, and we now look at two of them.

4 Jansson's method

Jansson's method is very widely used in applied sciences, and works quite well. It has its roots in the long standing Van Cittert iterative method, see [15], which can be regarded as a regularization method. We here look at it from the perspectives of the last section. First we digress and study a simpler but illuminating method. Assume the observations $\mathbf{y}$ in (5) are corrupted by additive measurement errors $\boldsymbol{\epsilon}$. Under the assumption that the measurement errors $\boldsymbol{\epsilon}$ are i.i.d. Gaussian, the maximum likelihood estimate of $\mathbf{x}$ is

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}\geq 0} \|\mathbf{y} - \mathbf{W}\,\mathbf{x}\|_2\,. \tag{11}$$

In this case, we are naturally led to the L_2 norm as a measure of goodness of fit. The computation here is the minimizing a convex functional over a convex set. Conditions characterizing legitimate solutions are provided by the Kuhn-Tucker theorem, see [29]. According to this theorem, we can divide the set $1, 2, \cdots, n$ into two subsets: $\mathcal{E}$ and $\mathcal{S}$, see [24]. A necessary and sufficient condition for $\hat{x}$ to be a solution of (11) is the following:

$$\begin{cases} \hat{x}_k = 0, & \mathbf{W}'(\mathbf{W}\hat{\mathbf{x}} - \mathbf{y})_k > 0, \quad for \quad k \in \mathcal{E}, \\ \hat{x}_k > 0, & \mathbf{W}'(\mathbf{W}\hat{\mathbf{x}} - \mathbf{y})_k = 0, \quad for \quad k \in \mathcal{S}. \end{cases} \tag{12}$$

The last quantity $\mathbf{W}'(\mathbf{W}\hat{\mathbf{x}} - \mathbf{y})$ is in fact the derivative of the L_2 loss function with respect to $\hat{\mathbf{x}}$. Lawson and Hanson ([24], page 159-164) provided an algorithm solving the general problem of nonnegative least squares, and prove its finite convergence. In this algorithm, a series of least squares problems without constraints are solved sequentially according to the tentative status of the slack set $\mathcal{S}$, whose initial state is an empty set. A positive component is added to $\mathcal{S}$ in the main loop, followed by a possible deletion of some components from $\mathcal{S}$ in the inner loop, and the loop is terminated when the Kuhn-Tucker condition is satisfied. Other algorithms such as the interior point algorithm have been developed for the optimization problem with constraints, see [29]. Motivated by the Kuhn-Tucker condition, we describe the following algorithm, taking into account the non-negativity constraints, to compute an approximate solution of (11).

Algorithm 4.1

1. *Let* $\mathbf{x}^{old} = \mathbf{y}$, *i.e. take the observations as our starting point. Go to step 2.*
2. *Let* $\mathbf{x}^{new} = \mathbf{x}^{old} - r\,\mathbf{W}'(\mathbf{W}\hat{\mathbf{x}}^{old} - \mathbf{y})$, *where* r *is a positive relaxation number. Go to step 3.*
3. *Let* $\mathbf{x}^{new} = \max\{\mathbf{x}^{new}, 0\}$. *Go to step 4.*
4. *Check condition (12). If satisfied, stop; otherwise, set* $\mathbf{x}^{old} = \mathbf{x}^{new}$, *and go to step 2.*

This algorithm guarantees that once we come to a point satisfying a Kuhn-Tucker condition, no further iterations will move it away. In contrast with Lawson and Hanson's algorithm, each iteration of this algorithm could change the tentative slack set $\mathcal{S}$ by adding some components and deleting others simultaneously. Though the matrix $\mathbf{W}$ appears in the description, because of its special structure the matrix multiplications in step two are in fact carried out by convolution. No storage and inversion of large matrices is needed here. Another prominent feature of this algorithm is the fact that the more iterations we carry out, the more complete the deconvolution, or, visually, the sharper are the reconstructed spikes. We use this algorithm as a bridge to understand the core algorithm in Jansson's method.

In Jansson's method, the unknowns are not assumed to be zero at the two ends. With a slight abuse of notation, from now throughout this section, we assume $\mathbf{x}$ has the same length n as $\mathbf{y}$, and so $\mathbf{W}$ is a n by n square matrix. For some point spread functions such as the Gaussian, which is commonly seen in spectroscopy, the matrix $\mathbf{W}$ is almost nonnegative. That is, $\mathbf{W}_\eta = \mathbf{W} + \eta\mathbf{I}$ can be a positive matrix even for a small η, though $\mathbf{W}_\eta$ could still be ill-conditioned to some extent. Notice that the solution $\mathbf{y} = \mathbf{W}_\eta\,\mathbf{x}$ minimizes $\mathbf{x}'\mathbf{W}_\eta\,\mathbf{x} - 2\mathbf{y}'\mathbf{x}$. This observation suggests the use of the following weighted L^2 norm as a measure of goodness of fit:

$$\min_{\mathbf{x}\geq 0} \left[(\mathbf{W}_\eta\,\mathbf{x} - \mathbf{y})'{\mathbf{W}_\eta}^{-1}(\mathbf{W}_\eta\,\mathbf{x} - \mathbf{y})\right]. \tag{13}$$

The Kuhn-Tucker condition for the solutions of this is the following:

$$\begin{cases} \hat{x}_k = 0, & (\mathbf{W}_\eta\,\hat{\mathbf{x}} - \mathbf{y})_k > 0, \qquad for \quad k \in \mathcal{E}, \\ \hat{x}_k > 0, & (\mathbf{W}_\eta\,\hat{\mathbf{x}} - \mathbf{y})_k = 0, \qquad for \quad k \in \mathcal{S}. \end{cases} \tag{14}$$

In comparison with (12), $\mathbf{W}_\eta$ appears only once here. Now we provide a version of Jansson's method.

Algorithm 4.2 (Jansson's method)

1. *Let* $\mathbf{x}^{old} = \mathbf{y}$. *Go to step 2.*
2. *Let* $\mathbf{x}^{new} = \mathbf{x}^{old} - r_0\, r(\mathbf{x}^{old})\,(\mathbf{W}\mathbf{x}^{old} - \mathbf{y}))$, *where* $r(\cdot)$ *is a positive relaxation function, and* r_0 *is the relaxation number. Go to step 3.*
3. *For the first few iterations, apply a smoother to the estimate. Otherwise go to step 4.*
4. *Let* $\mathbf{x}^{new} = \max\{\mathbf{x}^{new}, 0\}$.
5. *To continue the iteration, set* $\mathbf{x}^{old} = \mathbf{x}^{new}$, *go to step 2. Otherwise stop.*

Jansson's method does not require convergence, but stops the iterations when the result is satisfactory. Now let us compare this algorithm with Algorithm 4.1. If we skip step 3, ignore the relaxation function $r(\cdot)$, and replace $\mathbf{W}$ by a slight different version $\mathbf{W}_\eta$, then it is easy to see that the goal of the iteration is to minimize the weighted L_2 distance as in (13), whose derivative with respect to $\mathbf{x}$ is given by $\mathbf{W}_\eta\,\mathbf{x}^{old} - \mathbf{y}$. Now let us look at the role of the relaxation function. This nonnegative function takes the value zero at the boundary. One such example is shown in Figure 3, which will be used later in our numerical examples. Let us adopt the notation, $\mathcal{E}_\mathcal{J}$ and $\mathcal{S}_\mathcal{J}$, analogous to $\mathcal{E}$ and $\mathcal{S}$ in the Kuhn Tucker condition (12). Whenever a component of x is set to zero at some point in the iterations, it will remain in the zero component set $\mathcal{E}_\mathcal{J}$ because of the special

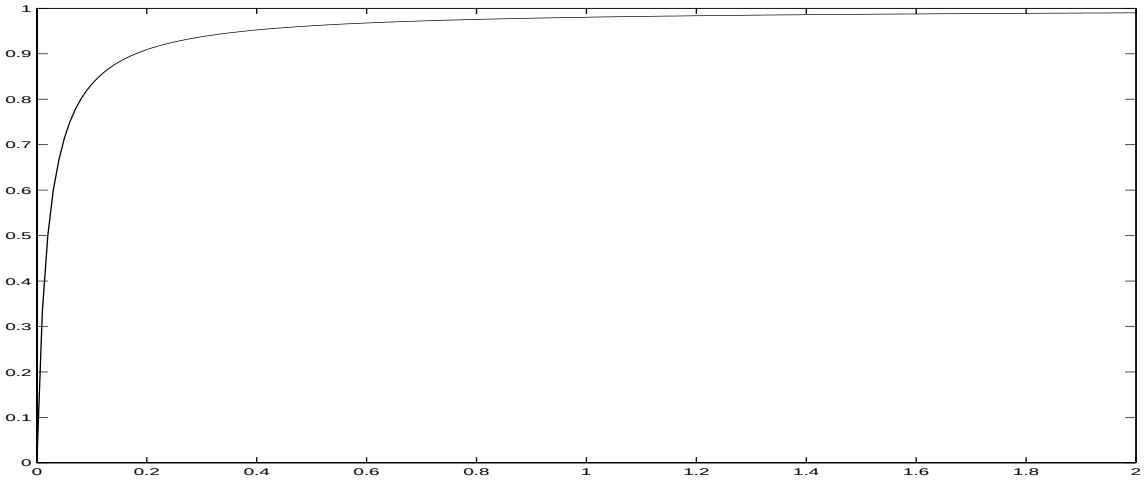

Figure 3: A relaxation function.

feature of the relaxation function. Thus the relaxation function introduces a backward deletion procedure with respect to the nonzero component set $\mathcal{S}_{\mathcal{J}}$. If we skip the smoothing steps, then we observe from our numerical experiments that each signal peak could split into several closely standing spikes as iterations progress. A similar phenomenon occurs in Algorithm 4.1, see Figure 11. Moderate smoothing could dampen this effect, and this is inconsistent with the assumption that the objects to be recovered are sparse. Keep in mind that over-smoothing could result in the loss of resolution. From the computational point of view, the main algorithmic operations involved in Jansson's method are multiplication and convolution of vectors, which are easily implemented with simple computing facilities. But we do need to determine the relaxation function and the smoothers.

5 Deconvolution by minimizing the Kullback-Leibler distance via the synchronization-minimization algorithm

In this section, we assume that all the elements of $\mathbf{x}$, $\mathbf{y}$ and $\mathbf{W}$ are nonnegative. It should be noticed that our treatment to the deconvolution problem in this section applies to the general linear inverse problems with positivity constraints discussed in Vardi and Lee, [40], and Vardi [39], and referred to as LININPOS problems. With this generalization in mind, we denote the index sets of $\mathbf{x}$ and $\mathbf{y}$ by J and K respectively. Without loss of generality, we normalize them such that $\sum_{j\in J} x_j = 1$, $\sum_{k\in K} y_k = 1$, and $\sum_{k\in\mathcal{K}} w_{k,j} = 1$. Now $\mathbf{y}$ and $\mathbf{W}\,\mathbf{x}$ can be regarded as two probability mass functions on K, and we can find an estimate of $\mathbf{x}$ by minimizing their Kullback-Leibler divergence $D(\mathbf{y}\|\mathbf{W}\,\mathbf{x})$, see [3] for definition. At this moment, it is convenient to consider two probability measures on $J \times K$ denoted by $\beta(j,k)$ and $\alpha(j,k)$ respectively. The two marginal probabilities and the two conditional probabilities of $\beta(j,k)$ are denoted by $\beta_1(j)$, $\beta_2(k)$, $\beta_{1|2}(j|k)$, $\beta_{2|1}(k|j)$. Similar notation are defined for $\alpha(j,k)$. The two sets of notation are linked as follows: the marginal probability $\beta_2(k)$ is given by y_k; the conditional probability $\alpha_{2|1}(k|j)$ is given by $w_{k,j}$. According to the chain rule for the Kullback-Leibler distance, which can be easily checked or found

in [3],

$$D(\beta(j,k)\|\alpha(j,k)) = D(\beta_2(k)\|\alpha_2(k)) + D(\beta_{1|2}(j|k)\|\alpha_{1|2}(j|k))\,,$$

and the problem of minimizing the Kullback-Leibler distance can be stated as a double minimization problem

$$\begin{aligned} \min_{\alpha_1(j)} D(\beta_2(k)\|\alpha_2(k)) &= \min_{\alpha_1(j)} \min_{\beta_{1|2}(j|k)} [D(\beta(j,k)\|\alpha(j,k)) - D(\beta_{1|2}(j|k)\|\alpha_{1|2}(j|k))] \qquad (15) \\ &= \min_{\alpha_1(j)} \min_{\beta_{1|2}(j|k)} [D(\beta_{1|2}(j|k)\beta_2(k)\|\alpha_{2|1}(k|j)\alpha_1(j)) - D(\beta_{1|2}(j|k)\|\alpha_{1|2}(j|k))]\,. \end{aligned}$$

Algorithm 5.1 (Synchronization-minimization)

1. *Initialization: let $\alpha_1(j)$ be the uniform distribution. Go to step 2.*
2. *Synchronization: compute the reverse conditional probability $\alpha_{1|2}(j|k)$ from $\alpha_{2|1}(k|j)$ using Bayes' theorem. Let $\beta_{1|2}(j|k) = \alpha_{1|2}(j|k)$. Go to step 3.*
3. *Minimization: minimize $D(\beta_{1|2}(j|k)\beta_2(k)\|\alpha_{2|1}(k|j)\alpha_1(j)$ over $\alpha_1(j)$. Go to step 2.*

Combining the solutions in step 2 and 3, which are $\alpha_{1|2}(j|k) = \alpha_{2|1}(k|j)\alpha_1(j)/(\sum_j \alpha_{2|1}(k|j)\alpha_1(j))$ and $\alpha_1(j) = \sum_k \beta(j,k)$, we get the iterative formula,

$$\alpha_1^{new}(j) = \alpha_1^{old}(j) \sum_k \frac{\alpha_{2|1}(k|j)\beta_2(k)}{\sum_j \alpha_{2|1}(k|j)\alpha_1^{old}(j)}\,. \qquad (16)$$

The Kullback-Leibler distance on the left hand side of (15) is non-increasing after each iteration. In the synchronization step, we make the second term on the right hand side zero by synchronizing the two reverse conditional probability measures while keeping the difference of the two terms unchanged. In the minimization step, the first term on the right hand side is minimized while the second term cannot go down at all. In fact, the synchronization and minimization under the framework of Kullback-Leibler distance are the analogs of the expectation and maximization steps in the E-M algorithm. The two terms on the right hand side of (15) correspond to the Q and H in the study of the convergence of E-M algorithm, see Dempster et al. [6], and also Little and Rubin [28], Wu [43], McLachlan and Krishnan [30]. The great power of E-M algorithm in statistical modeling lies in the appropriate introduction of missing data. It opens modelers' minds while keeping the computational job under control. The reason that we are taking another look at it in this deconvolution or more generally LININPOS context is to point out that a parallel idea, which we refer to by the terms of synchronization and minimization, exists for the problem of minimizing the Kullback-Leibler distance. For similar ideas, see Csiszár and Tusnády, [4]. If we put our argument in the framework of the majorization-minimization theory, see Lange et al. [22] and Hunter and Lange [13, 14], the Kullback-Leibler distance between the joint measures plays the role of the majorizing function.

Going back to the old set of notation, the synchronization-minimization algorithm leads to the iterative formula,

$$x_j^{new} = x_j^{old} \sum_k \frac{w_{k,j} y_k}{\sum_j w_{k,j} x_j^{old}}\,. \qquad (17)$$

In comparison to (12), the Kuhn-Tucker condition of a solution $\hat{\mathbf{x}}$ for this minimizing Kullback-Leibler distance problem is given by

$$\begin{cases} \hat{x}_j = 0, & \sum_k \frac{w_{k,j} y_k}{\sum_j w_{k,j} \hat{x}_j} < 1, \qquad for \quad j \in \mathcal{E}, \\ \hat{x}_j > 0, & \sum_k \frac{w_{k,j} y_k}{\sum_j w_{k,j} \hat{x}_j} = 1 \qquad for \quad j \in \mathcal{S}. \end{cases} \tag{18}$$

Vardi and Lee [40] obtained the same algorithm by maximizing the likelihood of data with "infinitely large sample size" and using the E-M algorithm. We originally formulated deconvolution of DNA sequencing in [25] as a missing data problem in a multinomial model with a two way indices. But the determination of the effective sample size remains a problem, and needs further investigation for the purpose of inference. As a matter of fact, the sample size does not appear in (17) at all, and the idea of the minimizing the Kullback-Leibler divergence suffices for deriving the algorithm. Vardi, et al. [41] proved that the the sequence generated by (17) converges to a global minimizer of (15), a convex functional over a convex set, using the Csiszár's three point inequality on Kullback-Leibler divergence, [4, 3]. The condition that ensures the uniqueness of the global maximum can also be found there. It is easy to see that the limit of the sequence generated in (16) is a solution to the self-consistency equation,

$$\alpha_1(j) = \alpha_1(j) \sum_k \frac{\alpha_{2|1}(k|j)\beta_2(k)}{\sum_j \alpha_{2|1}(k|j)\alpha_1(j)},$$

and that it satisfies $D(\beta_2(k)\|\alpha_2(k)) = D(\beta(j,k)\|\alpha(j,k))$, which says the Kullback-Leibler divergence of the two marginals equals that of the joint measures, and that the synchronization of the two reverse conditional probabilities is achieved when the minimum is reached.

The iterative formula (17) is really a folk algorithm, for many writers have come up with it in different scenarios. Shepp and Vardi [34] and Vardi, et al. [41], proposed a model for positron emission tomography (PET), in which emission occurs according to a spatial Poisson distribution, and use this formula to obtain the maximum likelihood estimates of the emission intensities. Later, Vardi and Lee [40] and Vardi [39] found this algorithm is applicable to a wide class of linear inverse problem with positive constraints. Snyder et al. [35] obtained the algorithm as a solution to a general Fredholm integral equation of the first type, and their derivation used the E-M algorithm. Richardson [33], Kennett et al. [16, 17, 18], and Di *Gesù* et al. [7] obtained the formula from an intuitive Bayesian point of view, and termed it "Bayesian deconvolution". It is worth mentioning that the 2-D image data studied by Shepp and Vardi [34], Vardi, et al. [41], Snyder et al. [35] are quite distinct from the DNA sequencing data, the focus of this research, though the same algorithm is derived. In the first case, if we project image data onto 1-D as curves, then the purpose of deblurring is to sharpen the edges of a kind of "step function"s. In deconvolution, the signal to be recovered are spikes. Thus the performance of the algorithm in the 2-D deblurring case, although studied by many people, cannot be transplanted directly to the later case. From the computational point of view, this algorithm is extremely easy to code and implement. But we need to determine when to stop the iterations.

6 Numerical examples and discussion

In this section, we show the performance of various deconvolvers discussed in this paper using two numerical examples. First let us look at one data set generated from a spike-convolution model.

Example 6.1

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

where the sample size $n = 1024$, $w(t)$ is a Gaussian function $\frac{b}{\sqrt{2\pi}} \exp\{-\frac{b^2\, t^2}{2}\}$ with the scale parameter $b = 8$ being truncated at ± 4 SD. Errors are normally distributed with mean 0 and standard deviation 0.3. Figure 4 shows a simulated sample from this model. The seven spikes contained in this signal have similar heights, and this is typical for sequencing data. The three on the left are quite close to one another, and this is the tough part. The baseline is assume to be known in all methods except PDPS. In Figure 5, we show the result of PDPS by depicting the estimated locations and heights of the spikes, which is essentially the truth. A systematic simulation study of PDPS using the same model can be found in [27]. Figure 6 is a result of applying Jansson's method. The spikes are easily identified by not so sharp but well separated peaks, though the relative heights are slight different from the truth. The synchronization-maximization algorithm, which is used to minimize the Kullback-Leibler distance, is quite slow. Figure 7 is the result after 200,000 iterations. The Kuhn-Tucker condition (18) is checked to be approximately true up to the the fifth-digit to the left of the decimal point. The convergence rate of this algorithm is obviously not high. As the iterations approach a minimum point, which satisfies the Kuhn-Tucker condition, we found there exists a kind of spike splitting phenomenon around the true ones. But it is also observed that this sensitive phenomenon occurs gradually, which could be due to the the monotonicity property of the Kullback-Leibler distance being minimized. Thus solutions obtained part way to convergence make more sense if we prefer to have sparse peaks, especially when we more or less know how many peaks there should be. Figure 8 shows a result after 1000 iterations. At the early stage of the synchronization-maximization algorithm, the boundary effect is strong. This could be dampened out by padding some number of zeros to the two ends of the raw data. We also apply Lawson and Hanson's algorithm to minimize the L_2 distance, and the solution is shown in Figure 9. The sharp spikes being reconstructed by this method is impressive. One the other hand, the ridge regression solution is unsatisfactory, no matter how the λ parameter is selected. Figure 10 shows the result with $\lambda = 100$. Algorithm 4.1, is much less computationally intensive and storage-demanding than Lawson and Hanson's algorithm, and it is interesting to look at one of its solution in Figure 11. By applying a smoother to it, we can generate a signal very similar to that in Figure 6. In fact, there is no significant difference if we replace step 2 in Jansson's method by that of Algorithm 4.1.

Example 6.2 *Real sequencing data.*

We now extend our comparison of these deconvolvers to a segment of typical DNA sequencing data, depicted in Figure 12. It was provided by the engineering group at Lawrence Berkeley National Laboratory. Deconvolution is carried out separately for each channel, and so only one channel is presented here. The point spread function is chosen to be a truncated Gaussian function. A result of Jansson's method and a solution of minimizing the Kullback-Leibler distance after 600 iterations are shown in Figure 14 and 15 respectively. The results of applying Lawson and Hanson's method and Algorithm 4.1 can are in Figure 16 and 17. For this data set, we also tried the method of minimizing the L_1 distance between $\mathbf{y}$ and $\mathbf{W}\,\mathbf{x}$ via the linear programming algorithm. It can be seen that the results obtained by minimizing L_1 or L_2 are similar to each other. We also tried the maximum entropy deconvolution method, see Gull and Daniel [11], which is to maximize the

entropy of the unknown x subject to the L_2 distance, namely,

$$\min_{x \geq 0} \sum_k x_k \log x_k \qquad \text{subject} \quad \text{to} \quad \|Wx - y\|_2 \leq E, \tag{19}$$

where E is a suitable positive number which controls the goodness of fit of the model. A regularized form of the maximum entropy deconvolution is given by

$$\min_{x \geq 0} [\|Wx - y\|_2 + \lambda \sum_k x_k \log x_k],$$

where λ is a positive number. In fact, there is a one-to-one correspondence between the E in (19) and λ if we can make it data-dependent, see [29]. Donoho et al. [8] show that this maximum entropy deconvolution procedure has certain advantages such as the enhancement of signal-to-noise and super-resolution when the signal is nearly-black. However, note that the estimate obtained by maximizing the entropy functional is not scale-invariant. One result from such a procedure is shown in Figure 19. Stark and Parker [36] proposed new algorithms for solving this kind of problem. In comparison with these " non-parametric" methods, the result obtained by parametric deconvolution as shown in Figure 12 is much cleaner and more more accurate in estimating the locations and heights of the major spikes.

To summarize, if non-negativity constraints can be identified as prior knowledge on the unknowns, this is a great help in a deconvolution problem. The gains are greatly enhanced if a parsimonious parametric model such as the spike-convolution model can be proposed for the unknowns. In the application of deconvolution to DNA sequencing, Jansson's method, which is easy in implementation, is a good practical choice. The method of minimizing the Kullback Leibler distance, which has been used widely in similar contexts, provides a fair solution. The fairly new parametric deconvolution method based on the spike-convolution model seems promising based on our analytic and numerical experience.

References

[1] M. D. Adams, C. Fields, and J. C. Ventor, editors. *Automated DNA sequencing and analysis.*

[2] W.-Q. Chen and T. Hunkapiller. Sequence accuracy of larger DNA sequencing projects. *J. DNA Sequencing and Mapping*, 2:335–342, 1992.

[3] T. M. Cover and J. A. Thomas. *Elements of Information Theory.* Wiley, 1991.

[4] I. Csiszár and G. Tusnády. Information geometry and alternating minimization procedures. *Statistics and Decisions, Supplementary Issue No. 1*, pages 205–237, 1984.

[5] L. M. Delves and J. Walsh. *Numerical Solution of Integral Equations.* Clarendon Press, Oxford, 1974.

[6] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39:1–22, 1977.

[7] V. Di Gesù and M. C. Maccarone. The Bayesian direct deconvolution method: properties and and applications. *Signal Processing*, 6:201–211, 1984.

[8] D. L. Donoho, I. M. Johnstone, J. C. Hoch, and A. S. Stern. Maximum entropy and the nearly black object. *Journal of the Royal Statistical Society, B*, 54(1):41–81, 1992.

[9] J. Fan. On the optimal rates of convergence for nonparametric deconvolution problem. *Annals of Statistics*, 1257-1272, 1991.

[10] D. Freedman. Statistical models and shoe leather (with discussion). in *Sociological Methodology*, page 291-358, American Sociological Association, Washington, D.C. 1991.

[11] S. F. Gull and G. J. Daniell. Image reconstruction from incomplete and noisy data. *Nature*, 272:686–690, 1978.

[12] J. Hadamard. *Lectures on Cauchy's problem in linear partial differential equations.* Yale University Press, New Haven, 1923.

[13] D. R. Hunter, K. Lange. Quantile regression via an MM algorithm. *J. Comput. Graphical Stat.* 9:60–77, 2000.

[14] D. R. Hunter, K. Lange. Computing estimates in the proportional odds models. *Ann Inst Stat Math* (in press).

[15] P. A. Jansson, editor. *Deconvolution of Images and Spectra.* Academic Press, New York, 1997.

[16] T. J. Kennett, W. V. Prestwich, and A. Robertson. Bayesian deconvolution 1: convergence properties. *Nuclear Instrument and Methods*, 151:285–292, 1978.

[17] T. J. Kennett, W. V. Prestwich, and A. Robertson. Bayesian deconvolution 2: Noise properties. *Nuclear Instrument and Methods*, 151:293–301, 1978.

[18] T. J. Kennett, W. V. Prestwich, and A. Robertson. Bayesian deconvolution 3: application and algorithm implementation. *Nuclear Instrument and Methods*, 153:125–135, 1978.

[19] 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.

[20] B. F. Koop, L. Rowen, W.-Q. Chen, P. Deshpande, H. Lee, and L. Hood. Sequence length and error analysis of sequence and automated *taq* cycle sequencing methods. *BioTechniques*, 14(3):442–447, 1993.

[21] R. Kress. *Linear Integral Equations.* Applied Mathematical Sciences, 82. Springer-Verlag, Berlin, New York, 1989.

[22] K. Lange, D. R. Hunter, I. Yang. Optimization transfer using surrogate objective functions (with discussion). *J. Comput. Graphical Stat.* 9:1–59, 2000.

[23] C. B. Lawrence and V. V. Solovyev. Assignment of position-specific error probability to primary DNA sequence data. *Nucleic Acid Research*, 22(7):1272–1280, 1994.

[24] C. L. Lawson and R. J. Hanson. *Solving Least Squares Problems.* Prentice-Hall, 1974.

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

[26] 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.

[27] L. Li and T. P. Speed. Parametric deconvolution of positive spike trains. *to appear in Annals of Statistics*, 2000.

[28] R. J. A. Little and D. B. Rubin. *Statistical Analysis with Missing Data.* John Wiley & Sons: New York, Chichester, Brisbane, Toronto, Singapore, 1987.

[29] D. G. Luenberger. *Linear and Nonlinear Programming.* Addison-Wesley Publishing Company, 1984.

[30] G. J. McLachlan and T. Krishnan. *The EM Algorithm and Extensions.* John Wiley & Sons: New York, Chichester, Brisbane, Toronto, Singapore, Weinheim, 1997.

[31] Finbarr O'Sullivan. A statistical perspective on ill-posed inverse problems. *Statistical Science*, 1:502–527, 1986.

[32] D. S. Poskitt, K. Dogancay, and S-H Chung. Double-blind deconvolution: the analysis of post-synaptic currents in nerve cells. *Journal of Royal Statistical Society, B*, 61:191–212, 1999.

[33] W. H. Richardson. Bayesian–based iterative method of image restoration. *J. Opt. Soc. Am.*, 62(1):55–59, 1972.

[34] L. A. Shepp and Y. Vardi. Maximum-likelihood reconstruction for emission tomography. *IEEE Transaction on Medical Imaging*, MI-1:113–121, 1982.

[35] D. L. Snyder, T. J. Schulz, and J. A. O'Sullivan. Deblurring subject to nonnegativity constraints. *IEEE Transactions on signal processing*, 40(5):1143–1150, 1992.

[36] P. B. Stark and R. L. Parker. Bounded-variable least-squares: an algorithm and applications. *Computational Statistics*, 10:129–141, 1995.

[37] A. Tikhonov. Solution of incorrectly formulated problems and the regularization method. *Soviet Math. Dokl.*, 5:1035–1038, 1963.

[38] A. N. Tikhonov and A. V. Goncharsky. *Solutions of Ill-posed Problems.* Wiley & Sons: New York, 1977.

[39] Y. Vardi. Applications of the em algorithm to linear inverse problems with positive constraints. In S. E. Levinson and L. Shepp, editors, *Image Models (and their Speech Model Cousins).* Springer, 1996.

[40] Y. Vardi and D. Lee. From image deblurring to optimal investment: maximum likelihood solutions for positive linear inverse problems. *J. R. Statist. Soc. B*, 55(3):569–612, 1993.

[41] Y. Vardi, L. A. Shepp, and L. Kaufman. A statistical model for positron emisson tomography. *Journal of the American Statistical Association*, 80, 1985.

[42] G. Wahba. *Spline Models for Observational Data*. SIAM, 1990.

[43] C. F. Jeff Wu. On the convergence properties of the EM algorithm. *Annals of Statistics*, 11:95–103, 1983.

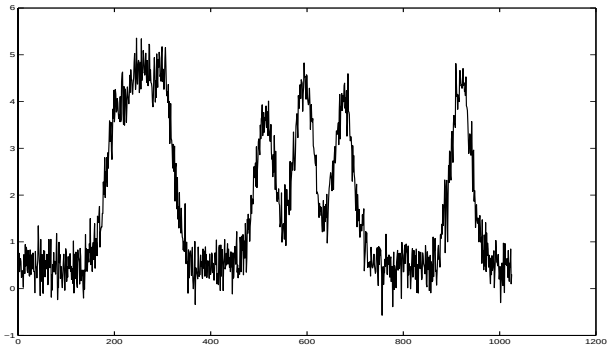

Figure 4: Data.

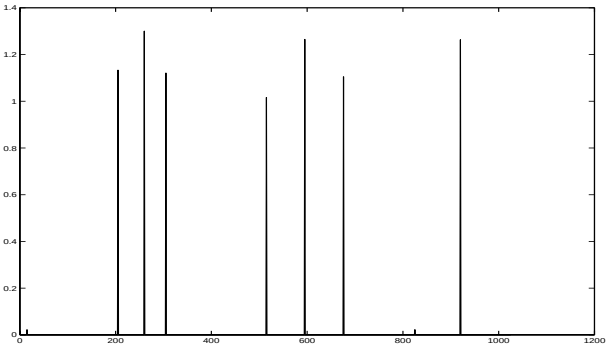

Figure 5: PDPS.

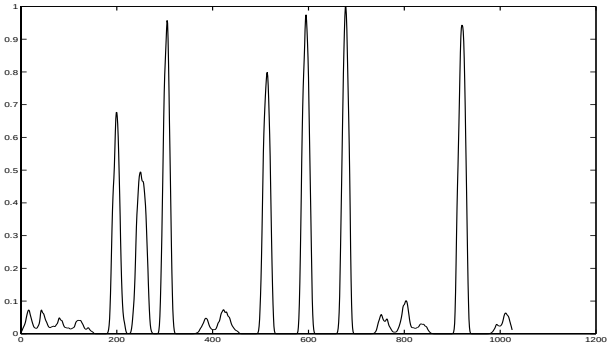

Figure 6: Jansson's method.

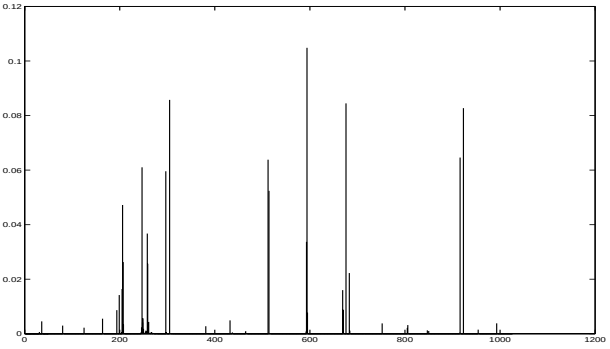

Figure 7: Minimum K-L distance

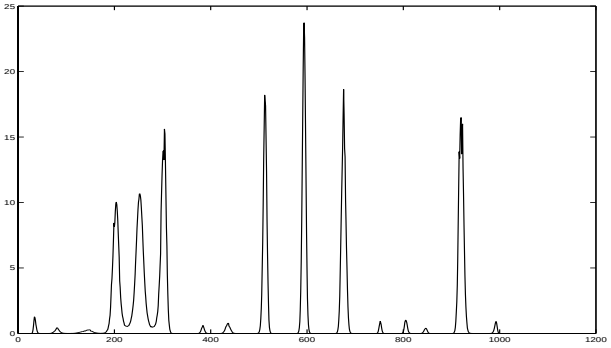

Figure 8: Halfway of Minimizing K-L distance.

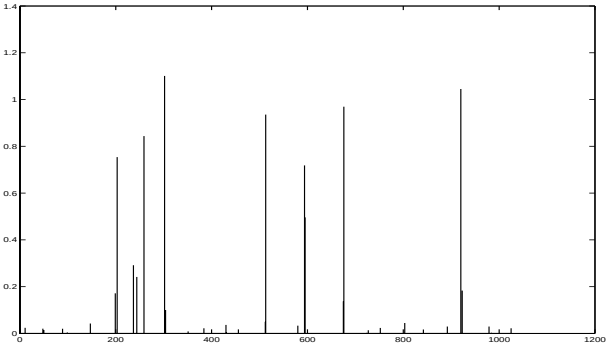

Figure 9: Minimum L_2 norm.

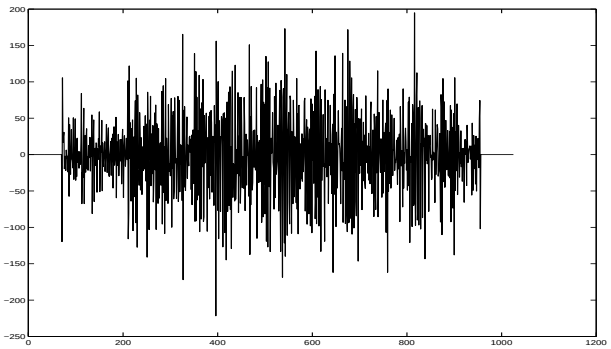

Figure 10: Ridge regression.

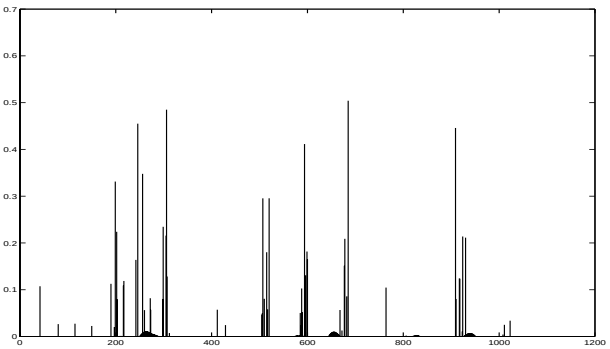

Figure 11: Algorithm 4.1.

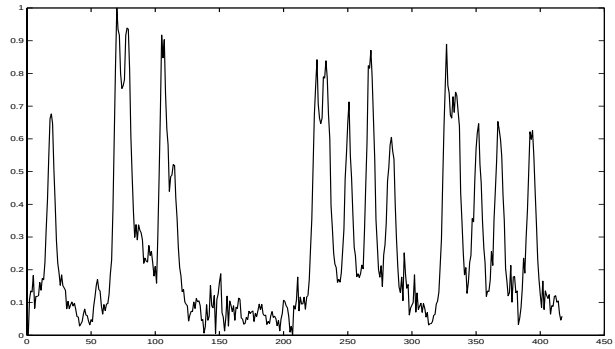

Figure 12: Raw data.

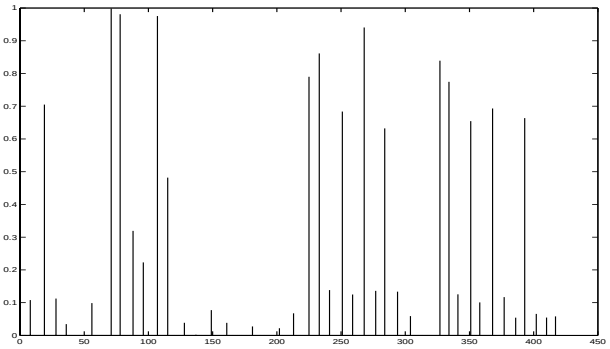

Figure 13: PDPS.

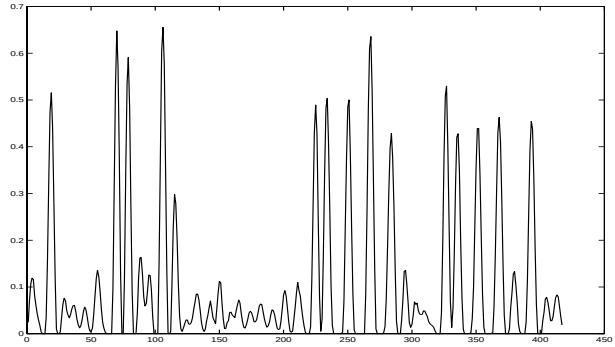

Figure 14: Jansson's method.

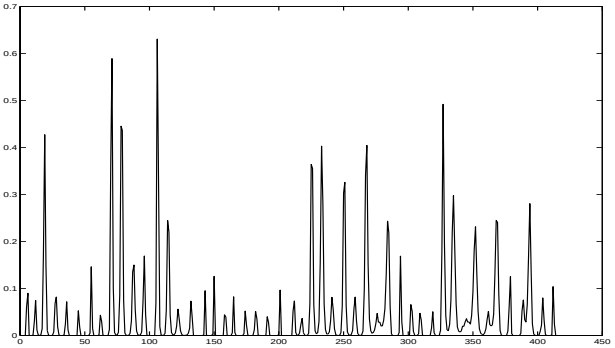

Figure 15: Minimizing K-L distance.

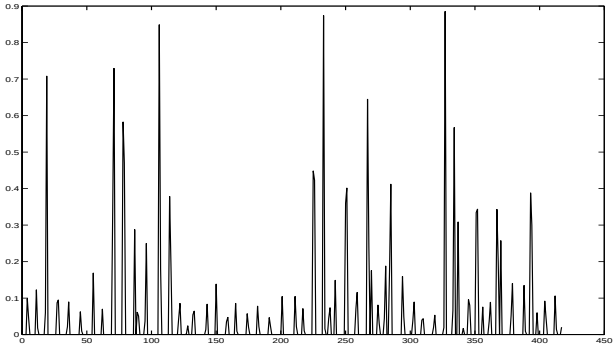

Figure 16: Minimum L_2 norm.

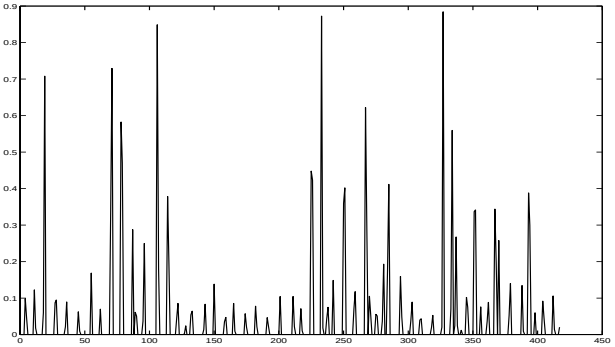

Figure 17: Algorithm 4.1.

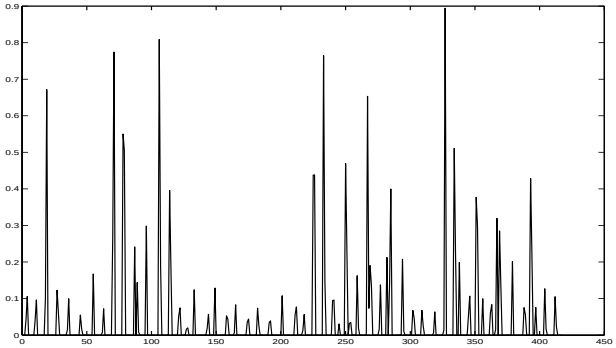

Figure 18: Minimum L_1 norm.

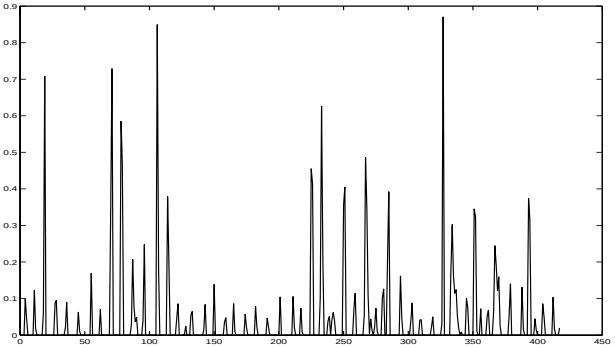

Figure 19: Maximum entropy.