A STATISTICAL ANALYSIS OF DIFFRACTION-LIMITED IMAGING

Peyman Milanfar and Ali Shakouri

Electrical Engineering Department
University of California
Santa Cruz, CA 94065
{milanfar, ali}@ee.ucsc.edu

ABSTRACT

The Rayleigh criterion is generally regarded as a fundamental limit and due to its practical accuracy in predicting the performance of optical imaging systems, it has unfortunately become accepted as a de-facto physical law. In this work, we will show that this limit is simply a very good rule of thumb, which under proper conditions typically related to the signal-to-noise (SNR) of the sensor, can be overcome. While we will not be discussing specific methodology (e.g. [1]) for improving resolution in this paper, it is the aim of this work to explore how far beyond the Rayleigh limit one can go, and to identify the *theoretical limits* to such resolution enhancement.

1. INTRODUCTION

For the sake of clarity and focus in the initial presentation, we carry out the analysis in one dimension, which we will later extend to 2 dimensions. To begin, let us assume that the signal of interest is the sum of two impulse functions separated by a small distance d as follows:

$$r(x;d) = \delta(x - \frac{d}{2}) + \delta(x + \frac{d}{2}) \quad (1)$$

When this signal is measured through an incoherent optical imaging system, the measured signal is the incoherent sum of two sinc functions, that represent the effect of the diffraction. As a result, the measured signal will be

$$f(x;d) = \text{sinc}^2(x-\frac{d}{2})+\text{sinc}^2(x+\frac{d}{2})+w(x) = s(x;d)+w(x), \quad (2)$$

where $w(x)$ is assumed to be a zero-mean Gaussian white noise process[1] with variance σ^2, and we recall that

$$\text{sinc}(x) = \frac{\sin(\pi x)}{\pi x}.$$

This work was supported in part by NSF Grant CCR-9984246, and the Packard Foundation

[1]Clearly for a photon-limited imaging system this assumption is inappropriate and our analysis will eventually take this into account

According to the *Rayleigh criterion* of resolution [2], the two point sources defined above are "barely resolved" when the center of one of the sinc function falls exactly on the first zero of the second sinc function. Accordingly, given the above definition, the Rayleigh limit to resolution is given by $d = 1$. That is, sources that are closer together than $d = 1$ are presumed to be *not* resolvable.

2. A STATISTICAL ANALYSIS OF RESOLUTION

The question of whether one or two peaks are present in the measured signal $f(x;d)$ can be formulated in statistical terms. Specifically, for the proposed model, the equivalent question is whether the parameter d is equal to zero or not. If $d = 0$, then we only have one signal, and if $d > 1$, then there are two "well-resolved" peaks according to Rayleigh's rule. So the problem of interest revolves around the values of d in the range $0 \leq d < 1$. To be more precise, let us define two hypotheses, which will form the basis of our work for understanding resolution in the statistical sense. Namely, let $\mathcal{H}_0$ denote the null hypothesis that $d = 0$ (i.e. one peak present) and let $\mathcal{H}_1$ denote the alternate hypothesis that $d > 0$ (i.e. two peaks present).

As we have indicated, since the range of interest are the values of $d < 1$, these representing resolution beyond the classical Rayleigh limit ($d < 1$), it is appropriate for the purposes of the following analysis to consider linearizing the model of the signal around $d = 0$. Specifically, consider the Taylor series expansion of $s(x;d)$ around $d = 0$, for any fixed x. (Note that this is a linearization about the parameter d and not the variable x.) We have

$$s(x;d) = g(x) + d^2 h(x) + O(d^3) \approx g(x) + d^2 h(x) \quad (3)$$

where terms of order d^3 and higher (which are quite small for $0 < d < 1$) can be ignored and where

$$g(x) = 2\frac{\sin^2(\pi x)}{\pi^2 x^2} \quad (4)$$

$$h(x) = \frac{\left(2\pi^2 x^2 - 3\right)\cos(2\pi x) - 4\pi x \sin(2\pi x) + 3}{4\pi^2 x^4} \quad (5)$$

It is interesting to note that due to the symmetry in the position of the two sinc functions with respect to the origin, no linear terms in d appear in the above approximation. This simplifies the hypothesis testing problem, and in what follows we will denote $d^2 = D$. With this definition, the hypothesis testing problem can now be rephrased as follows:

Given samples x_k ($k = 1, \cdots, N$) of the function $f(x_k)$, decide between the two hypotheses:

$$\begin{aligned} \mathcal{H}_0 : \quad & f(x_k) = g(x_k) + w(x_k), \\ \mathcal{H}_1 : \quad & f(x_k) = g(x_k) + D\, h(x_k) + w(x_k), \end{aligned}$$

where the parameter D is unknown.

Before continuing with the development of a detector structure and studying its performance, let us make a definition of signal to noise. In practice, with measurements of the values $f(x_k)$, we can make the definition of measured SNR *per sample* as

$$\text{SNR}_m = \frac{1}{N\sigma^2} \sum_{k=1}^{N} f^2(x_k) \approx \frac{1}{N\sigma^2} \sum_{k=1}^{N} \left(g(x_k) + D\, h(x_k)\right)^2 \tag{6}$$

On the other hand, since the function $g(x_k)$ is independent of the parameter D, we may further simplify the formulation of the detection problem by defining

$$y(x_k) = f(x_k) - g(x_k), \tag{7}$$

which yields

$$\begin{aligned} \mathcal{H}_0 : \quad & y(x_k) = w(x_k), \\ \mathcal{H}_1 : \quad & y(x_k) = D\, h(x_k) + w(x_k), \end{aligned}$$

For this simpler model, the natural definition of SNR per sample is given by

$$\text{SNR}_s = \frac{1}{N\sigma^2} \sum_{k=1}^{N} D^2\, h^2(x_k) \tag{8}$$

It is easily seen that

$$\text{SNR}_m = \frac{1}{N\sigma^2} \sum_{k=1}^{N} g^2(x_k) + \frac{2D}{N\sigma^2} \sum_{k=1}^{N} g(x_k)h(x_k) + \text{SNR}_s. \tag{9}$$

In what follows, we will deal mostly with the simplified model but we mention the relationship between the two different definitions of SNR because it will be convenient, and more intuitive, later to plot the results in terms of the measured SNR_m instead of SNR_s.

Returning to the detection problem posed in terms of $y(x_k)$, we observe that this is a problem of detecting a deterministic signal with an unknown parameter. With an explicit prior knowledge as to the likely values of D (i.e. a prior model), we can take a Bayesian approach to this detection problem. However, in general, there is no such prior information available. Therefore, we resort to the method of maximum likelihood (ML) for the estimation of the parameter D, and use this estimated value to form the standard Neyman-Pearson detector. This widely-used approach is known as *Generalized Likelihood Ratio Testing* or GLRT.

It is readily shown that the ML estimate for the parameter D is given by

$$\widehat{D} = \frac{\sum_{k=1}^{N} y(x_k)h(x_k)}{\sum_{k=1}^{N} h^2(x_k)} = (\mathbf{h}^T\mathbf{h})^{-1}\mathbf{h}^T\mathbf{y} \tag{10}$$

where

$$\begin{aligned} \mathbf{y} &= \left[y(x_1), \cdots,\, y(x_N)\right]^T \\ \mathbf{h} &= \left[h(x_1), \cdots,\, h(x_N)\right]^T \end{aligned}$$

The test-statistic resulting from the Neyman-Pearson likelihood ratio, with $D = \widehat{D}$ is given by [3]

$$\mathbf{T}(\mathbf{y}) = \frac{\widehat{D}^2}{\sigma^2}\mathbf{h}^T\mathbf{h}. \tag{11}$$

We note that the expression for the test-statistic is essentially and energy detector with the condition that the value of D is in fact estimated from the data itself. For any given data set $\mathbf{y}$, we decide $\mathcal{H}_1$ if the statistic exceeds a specified threshold:

$$\mathbf{T}(\mathbf{y}) > \gamma. \tag{12}$$

The choice of γ is motivated by the level of tolerable false alarm P_f (or false-positive) in a given problem. Typically, P_f is kept very low. In any event, the standard Neyman-Pearson detector is designed to produce the largest detection rate (P_d) for a specified P_f.

The detection and false-alarm rates for this detector are related as

$$\begin{aligned} P_d &= Q(Q^{-1}(P_f) - \sqrt{\lambda}) && (13) \\ &= Q(Q^{-1}(P_f) - \sqrt{N\ \text{SNR}_s}) && (14) \end{aligned}$$

where the parameter λ is:

$$\lambda = \frac{D^2}{\sigma^2}\mathbf{h}^T\mathbf{h} = \frac{1}{\sigma^2} \sum_{k=1}^{N} D^2\, h^2(x_k) = N\,\text{SNR}_s, \tag{15}$$

and where Q is the right-tail probability function for a standard Gaussian random variable (mean zeros, and unit variance.); and Q^{-1} is the inverse of this function. Recalling from (9) that

$$\text{SNR}_s = \text{SNR}_m - \frac{1}{N\sigma^2}\mathbf{g}^T\mathbf{g} - \frac{2D}{N\sigma^2}\mathbf{h}^T\mathbf{g} \qquad (16)$$

with

$$\mathbf{g} = [g(x_1), \cdots, \ g(x_N)]^T .$$

we observe that the detection rate can be written a function of the pre-specified false alarm rate, and the measured SNR per sample. It is worth noting here that the detector structure, due to our knowledge of the sign of the unknown distance parameter, is in fact a Uniformly Most Powerful (UMP) detector in the sense that it produces the highest detection probability for all values of the unknown parameter, and for a given false-alarm rate.

A particularly intriguing and useful relationship we have studied is the behavior of the smallest peak separation d, which can be detected with very high probability (say 0.99), and very low false alarm rate (say 10^{-6}) at a given SNR_m. We have examined this question by setting the values of $P_d = 0.99$ and $P_f = 10^{-6}$ in (13), and studying the resulting implicit curve which relates the variables SNR_m and d. Specifically, Figure 1 shows the SNR_m in units of decibles against the minimum detectable d at detection probability of at least 0.99 and at false-alarm rate of 10^{-6}. The samples x_k for this case were acquired over the range $[-10, \ 10]$ at just above the Nyquist rate. In this plot a fit to the curve is also shown. This very good fit has the following functional form:

$$d_{\text{min}} = \alpha SNR_m^{-1/4}, \qquad (17)$$

or equivalently,

$$d^2_{\text{min}} = \alpha^2 \frac{1}{\sqrt{SNR_m}}, \qquad (18)$$

where for the specific case shown in Figure 1, a best fit in the least-squares sense yielded the value of $\alpha = 1.27$. The relation (17) is a neat and rather intuitive one which can be used to, for instance, understand the required SNR per sample to achieve a particular resolution level of interest below the diffraction limit.

Figure 2 shows the same curve for different sampling rates. Namely, curves are shown for Nyquist rate, twice Nyquist, four times Nyquist, and eight times Nyquist. The minimum detectable d becomes smaller as the number of samples increases, but it does not do so at a very fast rate. In fact, upon closer examination, we find that the exponent of SNR_m in relation (17) does not change with increasing number of samples, but instead, as the number of samples N is increased, the coefficient α is reduced accordingly. We will quantify the rate of decrease of α with N in our ongoing work.

3. SUMMARY

The main conclusion that can be drawn from the above analysis is that in deciding the minimum resolvable separation between two sources from sampled data, two main factors enter into play; first, and foremost, is the SNR per sample of the imaging array. We have shown that, at least in the 1-D example here, the minimum detectable d behaves as the inverse of the SNR figure raised to the fractional power of $1/4$, indicating that with sufficiently high SNR, resolution beyond the diffraction limit is indeed possible. A second parameter of importance is the sampling rate, the increase of which can also improve performance. Though several earlier papers (e.g. [4] and [5]) carried out somewhat similar calculations, none have addressed nor posed the question of resolution in the particular context of the minimum resolvable distance and in particular the GLRT framework.

While the results we have obtained are intuitively pleasing, significant work remains to be done. For instance, we are presently carrying out the analysis for the scenario where the two sources are of unequal amplitude. Also, it would be interesting to further study whether the particular position of the samples (instead of, or in addition to, simply increasing the number of samples) would improve performance. We suspect this to be the case. We are also studying statistical lower bounds on the estimation error covariance for the amplitude and separation parameters. Finally, in the continuation of our work, we plan to extend these results to multiple dimensions.

4. REFERENCES

[1] N. Nguyen, P. Milanfar, and G. Golub. A computationally efficient superresolution image reconstruction algorithm. *IEEE Trans. on Image Processing*, vol. 10, no. 4, pp. 573-583, April 2001

[2] Joseph W. Goodman Introduction to Fourier Optics, Second Edition McGraw-Hill, 1996

[3] Steven Kay Fundamentals of Statistical Signal Processing, Vol. II: Detection Prentice Hall, 1993

[4] Arnold J. den Dekker Model-based Optical Resolution *IEEE Trans. on Instrumentation and Measurement*, vol. 46, no. 4, pp. 798-802, August, 1997

[5] A.J. den Dekker, A. van den Bos Resolution: A Survey *J. of the Optical Society of America: A*, vol. 14, no. 3, pp. 547-557, March, 1997

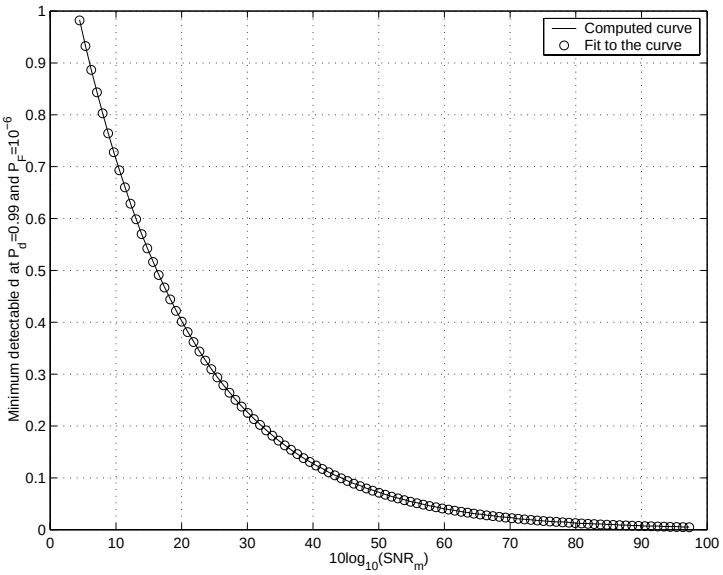

Fig. 1. Minimum detectable distance d between two point sources of equal amplitude as a function of the measured SNR (in units of dB) per sample. Note that $d = 1$ corresponds to the diffraction limit. It is assumed that the probability of detection is $P_d = 0.99$ and probability of false alarm is $P_f = 10^{-6}$. The fit shows remarkable accuracy. Sampling is just above Nyquist rate.

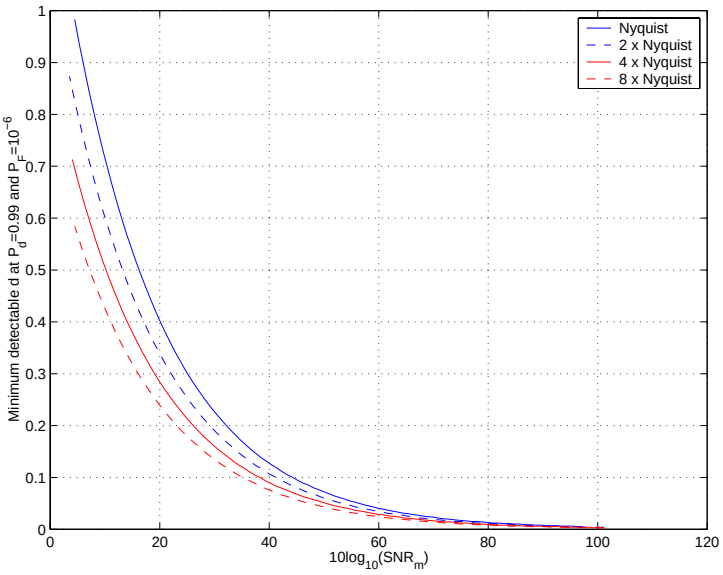

Fig. 2. Minimum detectable d at $P_d = 0.99$ and $P_f = 10^{-6}$ as a function of measured SNR (in units of dB) per sample. Different curves correspond to various number of samples in the image.