Systematic errors in Hadamard transform optics

N. J. A. Sloane, Martin Harwit, and Ming-Hing Tai

This paper analyzes the systematic errors in Hadamard transform optical instruments caused by moving masks, incorrect mask alignment, faulty mask fabrication, missing data, diffraction, etc. and describes techniques for reducing or eliminating these errors. In a great many cases the behavior of the instrument can be characterized by a single matrix equation of the form $\eta = \mathbf{TWa} + \mathbf{e}$, where the components of η are the measurements, $\mathbf{T}$ is a matrix characterizing the instrument, $\mathbf{W}$ specifies the mask configurations, $\mathbf{a}$ is a vector containing the unknown spectral intensities, and the components of $\mathbf{e}$ are small random errors.

I. Introduction

Hadamard transform optical systems have been widely studied during the past few years. By encoding the light with properly designed masks these instruments benefit from the so-called multiplex advantage and achieve an increased SNR. A variety of multiplexing instruments have been described in the literature, designed to improve images, spectra, or both.[1–10]

Any optical technique that leads to improved performance is likely to have its limitations. Some of these limitations may be inherent in the design of the system, while others may be caused by practical difficulties with the apparatus. In either case once such a limitation is recognized, ways of working with or around it can be examined. The purpose of this paper is to describe a number of sources of error encountered in Hadamard transform optical instruments, to give techniques for reducing or eliminating these errors where such techniques have already been worked out and to suggest some possible ways of dealing with errors that have not yet been fully studied.

We shall mainly describe errors occurring in singly encoded spectrometers, although most of the discussion will apply equally well to imagers. Similar errors occur in doubly encoded systems but have not yet been extensively investigated.

N. J. A. Sloane is with Bell Laboratories, Murray Hill, New Jersey 07974; M. Harwit is with Cornell University, Center for Radiophysics & Space Research, Ithaca, New York 14850; and M. H. Tai is with NASA Langley Research Center, Hampton, Virginia 23665.

Received 7 December 1977.
0003-6835/78/0915-2991$0.50/0.

II. Description of Instrument and Spectrum: Simplest Case

In this section we consider a multiplexing spectrometer in its simplest mode of operation. We give a mathematical description of the instrument under the assumption that no errors occur except random fluctuations in the detector readings. Subsequent sections will consider various departures from this ideal behavior.

A. Description of Instrument

Let us consider then a spectrometer with a single narrow entrance slit and with a mask in the exit focal plane [3,5] It is assumed that the entrance slit is sufficiently narrow so as to be always filled by the incoming radiation. The input to the spectrometer can therefore be completely specified by giving the frequency distribution of the radiation. It is most convenient to express this as a function $F(\nu)$, say, of the wavenumber ν. We further assume that over the operating range in which we are interested the optical system between the entrance slit and the exit mask behaves like a linear system.[11] In particular if the input consists of monochromatic light at wavenumber ν_0 and with unit intensity, so that $F(\nu) = \delta(\nu - \nu_0)$, the distribution of intensity along the exit focal plane can be written as $H(\nu - \nu_0)$. (We assume that distances along the exit focal plane have been calibrated in terms of wavenumber.) Thus $H(\nu - \nu_0)$ is the instrument's spectral display for monochromatic input at wavenumber ν_0. $H(\nu - \nu_0)$ is called the impulse response or point spread function[12] of the instrument. Because of aberrations and diffraction $H(\nu - \nu_0)$ can never be concentrated at a single point, no matter how narrow the entrance slit is. By superposition the distribution of intensity along the exit focal plane produced by an arbitrary input $F(\nu)$ is given by

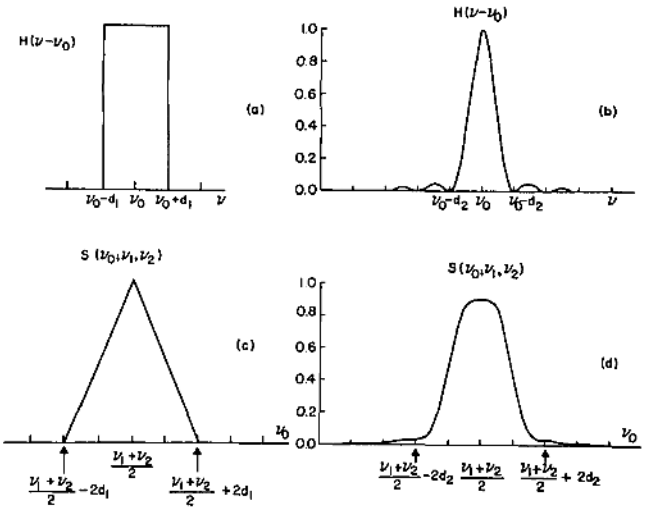

Fig. 1. Impulse response $H(\nu - \nu_0)$ (a) when diffraction can be ignored and (b) when diffraction dominates. (c) and (d) are the corresponding slit functions (see Examples 1 and 2 in Sec. II.D).

$$G(\nu) = \int_{-\infty}^{\infty} F(\nu_0)H(\nu - \nu_0)d\nu_0. \tag{1}$$

In other words $G(\nu)$ is the spectral display along the exit focal plane if the input spectrum is $F(\nu)$.

An ideal instrument, free of aberrations and diffraction, would have impulse response $H(\nu - \nu_0) = \delta(\nu - \nu_0)$ and [from Eq. (1)] $G(\nu) = F(\nu)$. Thus the linear spectral display $G(\nu)$ in the exit focal plane would be identical to the spectral distribution of the input. But in any real instrument $G(\nu) \neq F(\nu)$.

Of course strictly speaking we should write G as a function $G(\varkappa)$ of the position $\varkappa$ along the exit focal plane, rather than as a function of wavenumber. But the latter notation is simpler and more suggestive.

Suppose a single exit slit is used, extending say from wavenumber ν_1 to wavenumber ν_2 along the exit focal plane. The detector output is given by the integral of the exit plane distribution over the slit area:

$$\eta(\nu_1,\nu_2) = \int_{\nu_1}^{\nu_2} G(\nu)d\nu + e,$$

$$= \int_{-\infty}^{\infty} F(\nu_0)d\nu_0 \int_{\nu_1}^{\nu_2} H(\nu - \nu_0)d\nu + e,$$

$$= \int_{-\infty}^{\infty} F(\nu_0)S(\nu_0;\nu_1,\nu_2)d\nu_0 + e, \tag{2}$$

where e is the error due to detector noise, and $S(\nu_0;\nu_1,\nu_2)$ is defined by

$$S(\nu_0;\nu_1,\nu_2) = \int_{\nu_1}^{\nu_2} H(\nu - \nu_0)d\nu. \tag{3}$$

$S(\nu_0;\nu_1,\nu_2)$ is called the slit function[12] of the instrument; it is the total intensity of radiation passing through a slit extending from ν_1 to ν_2 in the exit focal plane, when the input is monochromatic radiation at wavenumber ν_0.

Figures 1(a) and 1(b) show two examples of $H(\nu - \nu_0)$. Figure 1(a) corresponds to an input slit which is wide enough for diffraction to be ignored, while Fig. 1(b) shows the opposite extreme when diffraction effects dominate and $H(\nu - \nu_0) = \text{sinc}^2(\nu - \nu_0)$, where $\text{sinc}\varkappa = (\sin\varkappa)/\varkappa$. The main lobe of $H(\nu - \nu_0)$ has width $2d_1$ in Fig. 1(a), or $2d_2$ in Fig. 1(b). The corresponding slit functions $S(\nu_0;\nu_1,\nu_2)$ are shown in Figs. 1(c) and 1(d), assuming the exit slit has width $\nu_2 - \nu_1$ equal to $2d_1$ or $2d_2$, respectively. For example, in Fig. 1(c) when $\nu_0 = (\nu_1 + \nu_2)/2$ the exit slit extends from $\nu_0 - d_1$ to $\nu_0 + d_1$ and gathers all the radiation, while if $\nu_0 = (\nu_1 + \nu_2)/2 \pm d_1$ only half of the radiation passes through the slit.

Equation (2) is the fundamental equation that relates the output of the detector to the input spectrum $F(\nu)$ when the exit slit extends from $\nu = \nu_1$ to $\nu = \nu_2$.

B Description of Spectrum

It is customary when measuring a spectrum with a dispersion instrument to divide the spectrum into n intervals, estimate the intensity in each interval, and then join the estimates by straight line segments. The result is a piecewise linear curve as shown in Fig. 2. Any spectrum can be approximated as closely as we please by such a curve, if n is taken large enough. We shall therefore start off with the assumption that the input spectrum $F(\nu)$ is a piecewise linear curve as shown in Fig. 2: the spectrum is divided into n segments $\theta'_0\theta'_1$, $\theta'_1\theta'_2, \ldots, \theta'_{n-1}\theta'_n$, corresponding to positions $\theta_0\theta_1$, $\theta_1\theta_2$, $\ldots, \theta_{n-1}\theta_n$ of the slits in the mask, and a_i is the intensity at the center of $\theta_i\theta_{i+1}$. (We assume that the scales have been chosen so that θ'_i and θ_i are related by an equation of the form $\theta'_i = \theta_i + A\theta_i^2 + \ldots$, where A is a constant.)

The goal is to determine the n unknowns $a_0, a_1, \ldots a_{n-1}$ as accurately as possible. In order to have an unambiguous expression for $F(\nu)$ in the first and last segments we assume that $F(\nu)$ is periodic outside the interval $\theta_0\theta_n$, as shown by the broken lines in Fig. 2. Thus we define $a_{-1} = a_{n-1}$ and $a_n = a_0$ as the intensities in the segments $\theta_{-1}\theta_0$ to the left of $\theta_0\theta_1$ and $\theta_n\theta_{n+1}$ to the right of $\theta_{n-1}\theta_n$, respectively. Analytically we are assuming that $F(\nu)$ is given by

$$F(\nu) = a_i + (a_{i+1} - a_i)(\nu - \theta_{i+1} + b)/2b \tag{4}$$

for $\theta_{i+1} - b \leqslant \nu \leqslant \theta_{i+1} + b$ where $2b = \theta_{j+1} - \theta_j$ is the length of each segment.

C. Operation as a Multiplexed Spectrometer

Let us consider what happens when the instrument is operated as a multiplexed spectrometer by allowing light from several slits to fall simultaneously onto the detector during each measurement. Suppose there are n slit positions in all: $\theta_0\theta_1, \theta_1\theta_2, \ldots, \theta_{n-1}\theta_n$. The pat-

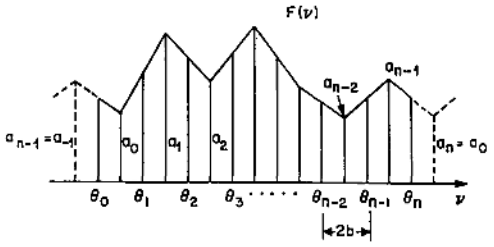

Fig. 2. A piecewise linear input spectrum $F(\nu)$. a_i is the intensity at the midpoint of the interval $\theta_i\theta_{i+1}$.

tern of open and closed slits is specified by means of the configuration matrix $W = (w_{ij})$, where $w_{ij} = 1$ if the jth slit is open during the ith measurement and $w_{ij} = 0$ otherwise, for $0 \leq i,j \leq n-1$. In statistical terminology W specifies a weighing design.[13-15] The mechanical design of the instrument is simplified (as will appear in the next section) if W is either a (right) circulant matrix, with $w_{i,j} = w_{i-j}$, or a left circulant matrix, with $w_{i,j} = w_{i+j}$ (and subscripts taken modulo n in both cases). From now on we assume that W is a left circulant; similar results hold in the other case.

If the jth slit is open during a particular measurement the total intensity of light passing through that slit is

$$\int_{\theta_j}^{\theta_{j+1}} G(\nu)d\nu. \tag{5}$$

Since $G(\nu)$ is a linear function of $F(\nu)$ [Eq. (1)] and $F(\nu)$ is a linear function of the a_k [Eq. (4)] this intensity may be written as

$$\sum_{k=0}^{n-1} t_{j-k}a_k, \tag{6}$$

where the t_{j-k} are constants characterizing the instrument. Thus t_{j-k} specifies the amount of radiation that is transferred through an open slit at position j for unit intensity at the kth spectral element. We assume that this is only a function of the separation $j-k$, independent of the actual values of j and k. To include more general situations (for example, misaligned masks, see Secs. III and IV), we allow these constants to vary from measurement to measurement. Thus we assume that if the jth slit is open during the ith measurement, the total intensity of light passing through that slit (i.e., the contribution of this slit to the ith detector reading) is given by

$$\tau_j^i = \sum_{k=0}^{n-1} t_{j-k}^i a_k, \tag{7}$$

where again the t_{j-k}^i are constants depending on the particular instrument. The ith detector reading η_i is given by the sum of all τ_j^i for which the corresponding slit is open:

$$\eta_i = \sum_{j=0}^{n-1} \tau_j^i w_{ij} + e_i, \tag{8}$$

$$= \sum_{k=0}^{n-1}\sum_{j=0}^{n-1} t_{j-k}^i w_{ij}a_k + e_i,$$

$$= \sum_{k=0}^{n-1} c_{ik}a_k + e_i, \tag{9}$$

where e_i is the error in the ith reading, and the constants c_{ik} are given by

$$c_{ik} = \sum_{j=0}^{n-1} t_{j-k}^i w_{ij},$$

$$= \sum_{j=0}^{n-1} t_{j-k}^i w_{i+j} \text{ (since } W \text{ is a left circulant)},$$

$$= \sum_{r=0}^{n-1} t^i{}_{r-i} w_{r+k} \text{ (where } r = i+j-k),$$

$$= \sum_{r=0}^{n-1} t^i{}_{r-i} w_{r,k}. \tag{10}$$

This is simpler in matrix notation. Let

$$\boldsymbol{\eta} = (\eta_0, \eta_1, \ldots, \eta_{n-1})^T,$$

$$\mathbf{a} = (a_0, a_1, \ldots, a_{n-1})^T,$$

$$\mathbf{e} = (e_0, e_1, \ldots, e_{n-1})^T,$$

denote column vectors of measurements, unknowns, and errors, respectively. From Eqs. (9) and (10) we obtain

$$\boldsymbol{\eta} = C\mathbf{a} + \mathbf{e} \tag{11}$$

$$= TW\mathbf{a} + \mathbf{e}, \tag{12}$$

where C is the matrix with (i,k)th entry c_{ik}, and

$$T = \begin{bmatrix} t_0^0 & t_1^0 & t_2^0 & \ldots & t_{n-1}^0 \\ t_{n-1}^1 & t_0^1 & t_1^1 & \ldots & t_{n-2}^1 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ t_1^{n-1} & t_2^{n-1} & t_3^{n-1} & \ldots & t_0^{n-1} \end{bmatrix}. \tag{13}$$

T is called the transfer matrix of the instrument. In an ideal instrument in which distortion, aberrations, and diffraction were negligible, T would be the unit matrix I. In a real instrument, however, radiation that should be exiting through a given mask slit often spills over into neighboring mask positions, so that diagonal elements of T are reduced in value and off-diagonal elements of the matrix grow. Since radiant energy incident on a detector cannot be negative, $0 \leq t_{j-k} \leq 1$. Conservation of radiant energy requires that the sum of the matrix elements in any row or column of T equals unity, unless dissipative processes such as absorption or scattering play a role. If dissipation is significant these sums,

$$\sum_i t_{j-k}^i \text{ or } \sum_k t_{j-k}^i,$$

may be less than unity and greater than zero. A zero value for any one of these sums would imply a systematic blockage in the spectrometer—either by design or through faulty construction. If no light is lost, T satisfies $TJ = T^{-1}J = J$, where J is a square matrix of ones.

Equations (11) and (12) are the basic equations describing the performance of this multiplexing spectrometer. They relate the unknowns $a_0, \ldots, a_{n-1}$ of Fig. 2 to the measurements $\eta_0, \ldots, \eta_{n-1}$ via the matrix $C = (c_{ik})$ or via the matrices T and W. Here $W = (w_{ij}) = (w_{i+j})$ is the configuration matrix that describes which slits are open and which are closed during the measurements, and the transfer matrix T [Eq. (13)] characterizes the particular instrument but is independent of the multiplexing. The entries in T are determined by Eq. (7) and depend on the impulse response $H(\nu)$ and on the positions $\theta_0, \ldots, \theta_{n-1}$. We shall see below that by suitably choosing the transfer matrix T in Eq. (12), we can use the same equation to describe the distortion introduced by a number of different sources of error.

D. Example 1: No Diffraction

As a first example we assume a wide entrance slit and no diffraction, so that the impulse response is as shown in Fig. 1(a). We take $d_1 = \frac{1}{2}$, so that $H(\nu - \nu_0) = 1$ for

$-\frac{1}{2} \leq \nu - \nu_0 \leq \frac{1}{2}, H(\nu - \nu_0) = 0$ elsewhere. We also assume that the exit slits have width 1, so that $\theta_{j+1} - \theta_j = 1$ and $b = \frac{1}{2}$ in Eq. (4) and Fig. 2. The slit function $S(\nu_0;\theta_j,\theta_{j+1})$ is given in Fig. 1(c) or analytically by

$$S(\nu_0;\theta_j,\theta_{j+1}) = \begin{cases} \nu_0 - \theta_j + \frac{1}{2} & \text{if } \theta_j - \frac{1}{2} \leq \nu_0 \leq \theta_j + \frac{1}{2}, \\ -\nu_0 + \theta_j + \frac{3}{2} & \text{if } \theta_j + \frac{1}{2} \leq \nu_0 \leq \theta_j + \frac{3}{2}, \\ 0 & \text{otherwise.} \end{cases} \tag{14}$$

If the jth slit is open during the ith measurement, the total intensity of light passing through that slit is found by substituting $S(\nu_0;\theta_j,\theta_{j+1})$ and $F(\nu)$ into Eq. (2) to give

$$\begin{aligned} \tau_j^i &= \int_{\theta_j - 1/2}^{\theta_j + 1/2} (\nu_0 - \theta_j + \tfrac{1}{2})[a_{j-1} + (a_j - a_{j-1})(\nu_0 - \theta_j + \tfrac{1}{2})]d\nu_0 \\ &+ \int_{\theta_j + 1/2}^{\theta_j + 3/2} (-\nu_0 + \theta_j + \tfrac{3}{2})[a_j + (a_{j+1} - a_j)(\nu_0 - \theta_j - \tfrac{1}{2})]d\nu_0, \\ &= \tfrac{1}{6}(a_{j-1} + 4a_j + a_{j+1}). \end{aligned} \tag{15}$$

Therefore the constants t_{j-k}^i in Eq. (7) are given by

$$t_0^i = \tfrac{4}{6}, \quad t_{-1}^i = t_1^i = \tfrac{1}{6},$$

$$t_j^i = 0 \text{ otherwise.}$$

(Note that in this instrument the t_j^i are independent of i.) The transfer matrix T for this instrument is therefore

$$T = \tfrac{1}{6}\begin{bmatrix} 4 & 1 & 0 & \dots & 0 & 1 \\ 1 & 4 & 1 & \dots & 0 & 0 \\ 0 & 1 & 4 & \dots & 0 & 0 \\ \cdot & \cdot & \cdot & \dots & \cdot & \cdot \\ 0 & 0 & 0 & \dots & 4 & 1 \\ 1 & 0 & 0 & \dots & 1 & 4 \end{bmatrix} \tag{16}$$

E. Example 2: Full Diffraction

The second example assumes a very narrow entrance slit, so that the impulse response is

$$H(\nu - \nu_0) = 2\,\mathrm{sinc}^2[2\pi(\nu - \nu_0)], \tag{17}$$

as in Fig. 1(b) with $d_2 = \frac{1}{2}$. [The constant 2 is chosen so that the total area under $H(\nu - \nu_0)$ is one.] Again we assume the slits have width 1, $\theta_{j+1} - \theta_j = 1$, $b = \frac{1}{2}$. Now the contribution of the jth slit to the ith measurement is [from Eqs. (2), (4), and (17)]

$$\begin{aligned} \tau_j^i = \sum_{r=0}^{n-1} &\int_{\theta_{r+1} - 1/2}^{\theta_{r+1} + 1/2} [a_r + (a_{r+1} - a_r)(\nu_0 - \theta_{r+1} + \tfrac{1}{2})]d\nu_0 \\ &\cdot \int_{\theta_j}^{\theta_j + 1} 2\,\mathrm{sinc}^2 2\pi(\nu - \nu_0)d\nu. \end{aligned}$$

After some algebra this becomes $\tau_j^i = t_0^i a_j + (t_1^i a_{j-1} + t^i{}_{-1}a_{j+1}) + \dots$ [i.e., Eq. (7)], where the coefficients t_r^i are given by

$$t_r^i = 2\int_{-3/2}^{3/2} f(x)\,\mathrm{sinc}^2 2\pi(x - r)dx, \tag{18}$$

and

$$\begin{aligned} f(x) &= \tfrac{1}{2}(x + \tfrac{3}{2})^2 \quad \text{if} \quad -\tfrac{3}{2} \leq x \leq -\tfrac{1}{2}, \\ f(x) &= \tfrac{3}{4} - x^2 \quad \text{if} \quad -\tfrac{1}{2} \leq x \leq \tfrac{1}{2}, \\ f(x) &= \tfrac{1}{2}(x - \tfrac{3}{2})^2 \quad \text{if} \quad \tfrac{1}{2} \leq x \leq \tfrac{3}{2}. \end{aligned} \tag{19}$$

Numerical evaluation of the integrals (18) shows that the transfer matrix T of this instrument is a symmetric circulant matrix with the first row equal to

$$\begin{aligned} &(0.6667,\ 0.1482,\ 0.0080,\ 0.0031,\ 0.0017, \\ &0.0010,\ 0.0007,\ 0.0005,\ 0.0004,\ 0.0003, \\ &0.0003,\ 0.0002,\ \dots,\ 0.0017,\ 0.0031, \\ &0.0080,\ 0.1482), \end{aligned} \tag{20}$$

correct to four decimal places. Notice the very slow decay of elements off the main diagonal, illustrating the pronounced spreading effect of this impulse response.

In general whenever $H(\nu)$ is an even function, i.e., satisfies $H(-\nu) = H(\nu)$, T is a symmetric circulant—a matrix of the form

$$\begin{bmatrix} u_0 & u_1 & u_2 & \dots & u_3 & u_2 & u_1 \\ u_1 & u_0 & u_1 & \dots & u_4 & u_3 & u_2 \\ u_2 & u_1 & u_0 & \dots & u_5 & u_4 & u_3 \\ & \cdot & \cdot & \cdot & \dots & \cdot & \cdot & \cdot \\ u_1 & u_2 & u_3 & \dots & u_2 & u_1 & u_0 \end{bmatrix}. \tag{21}$$

F. Operation as a Monochromator

If the same instrument is operated as a conventional monochromator, n measurements are made, where in the ith measurement a single exit slit extends from θ_i to θ_{i+1}. The basic equation is now

$$\eta = T\mathbf{a} + \mathbf{e}, \tag{22}$$

obtained by setting $W = I$ in Eq. (12).

G. Recovery of Spectrum

To estimate the spectrum from the measurements we proceed as follows. We assume (again this is the ideal case) that e_i, the detector error in the ith measurement, is a random variable with mean zero:

$$E\{e_i\} = 0, \tag{23}$$

that the errors in distinct measurements are uncorrelated:

$$E\{e_i e_j\} = 0, \qquad i \neq j, \tag{24}$$

and that

$$E\{e_i^2\} = \sigma^2, \tag{25}$$

where the variance σ^2 is independent of the amount of radiation falling on the detector. Under these assumptions if Eq. (12) holds then the best estimate of the a_i's is given by

$$\hat{\mathbf{a}} = W^{-1}T^{-1}\eta \tag{26}$$

$$= \mathbf{a} + W^{-1}T^{-1}\mathbf{e}. \tag{27}$$

This is best in the sense of being that linear unbiased estimate which minimizes the average mean square error

$$\epsilon = \frac{1}{n}\sum_{i=0}^{n-1} E[(\hat{a}_i - a_i)^2]$$

$$= \frac{\sigma^2}{n}\,\text{Trace}\,\{(TW)^{-1}[(TW)^{-1}]^T\}$$

$$= \frac{\sigma^2}{n} \times [\text{sum of squares of elements of } (TW)^{-1}]; \qquad (28)$$

see Refs. 3 and 15. Then ϵ is a measure of the accuracy of the experiment and should be made as small as possible.

H. Computing the Spectrum. Inverse Matrices

The spectrum is estimated by Eq. (26). Usually W^{-1} is known explicitly,[3] but problems may arise in finding T^{-1}, since in general the inverse of say a 255×255 matrix is both difficult to compute and awkward to store.

If T is a circulant matrix, as is often the case, the inverse is fairly easy to find and to store (see below). If T is Toeplitz (constant along diagonals) but not a circulant, the inverse can still be found but with more difficulty.[16-21] But usually if T is not a circulant, the best method of computing the spectrum is not to attempt to find T^{-1} but to solve directly the system of equations

$$\eta = TW\mathbf{a} \qquad (29)$$

for $\mathbf{a}$. This is particularly straightforward in the important case when W^{-1} is known and T is a band matrix, i.e., the only nonzero entries in T are those within a fixed distance of the main diagonal.[22]

The following are two methods for finding the inverse of an invertible circulant matrix C with first row $c_0, c_1, \ldots, c_{n-1}$. Method (I)[16,17] often gives an explicit formula for the inverse: Let $c(x) = c_0 + c_1x + \ldots + c_{n-1}x^{n-1}$. Find the unique polynomial $d(x) = d_0 + d_1x + \ldots + d_{n-1}x^{n-1}$ such that $c(x)d(x) \equiv 1 \pmod{x^n - 1}$. Then C^{-1} is the circulant with first row $d_0, \ldots, d_{n-1}$. Method II (which may be new) is an efficient algorithm for finding the inverse of a symmetric circulant with a computer: For large n the mth entry d_m of the first row of the inverse approaches

$$\frac{1}{\pi}\int_0^\pi \frac{\cos m\theta\, d\theta}{c_0 + 2\sum_{r=1}^{\infty} c_r \cos r\theta}. \qquad (30)$$

In calculating Eq. (26) it is useful to remember that any two right circulants commute, i.e., satisfy $AB = BA$, any two left circulants commute, and any left circulant commutes with any symmetric right circulant.

1. *Example 1 (cont.)*

The inverse of Eq. (16) was found by Method I. First define the sequence of integers $b_0, b_1, b_2, \ldots$ by

$$b_0 = 1, \quad b_1 = -4$$

$$b_n = -4b_{n-1} - b_{n-2}, \quad n \geq 2. \qquad (31)$$

This is Sequence 1420 in Ref. 23. The first few terms are

b_0	b_1	b_2	b_3	b_4	b_5
1	−4	15	−56	209	−780

Table I. First Row of Inverse Matrix of (16)

i	u'_i
0	1.732
1	−0.464
2	0.124
3	−0.033
4	0.009
5	−0.002
6	0.001

An explicit formula is

$$b_n = \frac{1}{6}[(3 - 2\sqrt{3})(-2 + \sqrt{3})^n + (3 + 2\sqrt{3})(-2 - \sqrt{3})^n] \qquad (32)$$

and also

$$b_{n-1}^2 - b_n b_{n-2} = 1, \quad n \geq 1. \qquad (33)$$

Then the inverse of Eq. (16) is given by the symmetric circulant (21) with

$$u_0 = \frac{3b_{n-1}}{2b_{n-1} + b_{n-2} + 1}, \qquad (34)$$

$$u_i = \frac{3(b_{n-i-1} + b_{i-1})}{2b_{n-1} + b_{n-2} + 1}, \quad i \geq 1. \qquad (35)$$

(This may be verified by using Method I: we omit the rather lengthy details.) For example, when $n = 4$

$$T = \tfrac{1}{6}\begin{bmatrix} 4 & 1 & 0 & 1 \\ 1 & 4 & 1 & 0 \\ 0 & 1 & 4 & 1 \\ 1 & 0 & 1 & 4 \end{bmatrix}, \quad T^{-1} = \tfrac{1}{4}\begin{bmatrix} 7 & -2 & 1 & -2 \\ -2 & 7 & -2 & 1 \\ 1 & -2 & 7 & -2 \\ -2 & 1 & -2 & 7 \end{bmatrix}.$$

For large n the second term in Eq. (32) dominates, and we can approximate u_i by

$$u_i \approx u'_i = (-1)^i \sqrt{3}(2 - \sqrt{3})^i \qquad (36)$$

for $0 \leq i \leq n/2$, independent of n. The approximation is very accurate, in fact $|u_i - u'_i| < 10^{-5}$ if $n \geq 20$. Table I gives the values of u'_i correct to three decimal places. Note that u'_i approaches 0 rapidly as i increases. Also

$$\det T \approx \left(\frac{2 + \sqrt{3}}{6}\right)^n. \qquad (37)$$

Therefore, for example, if the instrument is being operated as a monochromator, so that Eq. (22) applies, the best estimate of the a_i's is given by

$$\hat{\mathbf{a}} = T^{-1}\eta, \qquad (38)$$

and for $n \geq 20$ we can write Eq. (38) as

$$\hat{a}_i = u'_0\eta_i + u'_1(\eta_{i-1} + \eta_{i+1}) + \ldots + u'_6(\eta_{i-6} + \eta_{i+6})$$

with an error of less than 10^{-3}, where the u'_j are given by Table I, and subscripts are read modulo n. The average mean square error ϵ for this mode of operation can be found from Eqs. (28) and (36) and is

$$\epsilon \approx 2\sqrt{3}\sigma^2. \qquad (39)$$

Incidentally, if the correction matrix T^{-1} is not applied, the result is a coarsening of the spectrum. Consider an input spectrum with a single sharp line, corresponding

to $\mathbf{a} = (0,0,0,1,0,0,\ldots)$ as shown in Fig. 3(a). Then from Eq. (22), ignoring $\mathbf{e}$, $\eta = T\mathbf{a} = (0,0,\frac{1}{6},\frac{2}{3},\frac{1}{6},0,0,\ldots)$. If we ignore T^{-1} and estimate $\hat{\mathbf{a}}$ by η [instead of Eq. (38)] the resulting spectrum is shown in Fig. 3(b).

Suppose the same instrument is operated as a multiplexed spectrometer, so that Eq. (12) applies. The argument used in the Appendix of Ref. 15 can be modified to show that for any (0,1)-matrix W we have

$$\epsilon \geq \frac{10.34\sigma^2}{n}, \quad n \text{ large} \tag{40}$$

[using Eq. (37)]. If W is an S-matrix,[3,15] it can be shown that

$$\epsilon \approx \frac{13.86\sigma^2}{n}, \quad n \text{ large}, \tag{41}$$

which is reasonably close to (40) and may well be the smallest ϵ that can be achieved. Comparing Eq. (39) with Eq. (41) we see that

$$\frac{\text{m.s.e. with multiplexing}}{\text{m.s.e. without multiplexing}} \approx \frac{4.00}{n}, \tag{42}$$

and in fact for this choice of W it can be shown that Eq. (42) holds for any matrix T.

2. Example 2 (cont)

The inverse of Eq. (20) was found by Method II and is a circulant with first row which approaches

$$(1.6683,\ -0.3820,\ 0.0705,\ -0.0183,\ 0.0017,\ -0.0017,\ -0.0006,\ -0.0006,\ -0.0004,\ -0.0003,\ldots) \tag{43}$$

correct to four decimal places for large n. Again the elements off the main diagonal decay very slowly. Without multiplexing the average mean square error for this instrument is $\epsilon \approx 2.93\sigma^2$, while multiplexing with an S-matrix gives $\epsilon \approx (11.74\sigma^2)/n$, and again (42) holds.

Of course for any instrument we can ask the following question. The impulse response of $H(\nu)$ determines the transfer matrix T, as in Eq. (13). Then which configuration matrix W minimizes the average mean square error (28)? In Ref. 15 we studied the performance of various W's in the very special case when T is the identity matrix [thus ignoring the spreading effect of $H(\nu)$]. The more general question for an arbitrary $H(\nu)$ or T remains unsolved.

III. Errors Occurring with a Mask which is Moved in Steps

This section deals with some classes of errors that can occur with a mask which is stepped to the next position between measurements. Most of these can be described by a suitable modification of the transfer matrix T in the basic Eq. (12). These errors may also arise in instruments with a continuously moving mask (see Sec. IV). The resulting T is then the product of the matrices described in Sec. IV and those below. The echo effects produced by slits which are uniformly too narrow (or too wide) were described by Tai *et al.*[24]

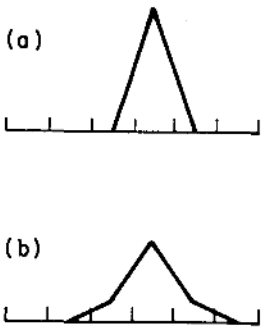

Fig. 3. How a sharp input (a) is coarsened (b) if the correction matrix T^{-1} is not used.

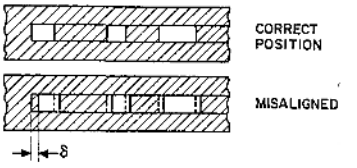

Fig. 4. Faulty mask alignment.

A. Faulty Mask Alignment

Sometimes the position of an encoding mask is systematically displaced by some small distance δ from its correct position (see Fig. 4). Plankey *et al.*[9] have described the effects produced by deliberate misalignment of a spectral encoding mask by 0.5 and 2.5 slit widths. As one would expect, the main result is an apparent spectral shift by 0.5 and 2.5 resolution elements, respectively, and an increase in spectral noise. The exact form of this noise may be determined as follows. We assume that δ is small compared to the slit width. Then the contribution to the ith detector reading from light passing through the jth slit is now given by

$$\tau_j^i = \int_{\theta_j-\delta}^{\theta_{j+1}-\delta} G(\nu)d\nu \tag{44}$$

instead of Eq. (5). For the instrument of Example 1, this becomes

$$\begin{aligned}
\tau_j^i = &\int_{\theta_j-1/2-\delta}^{\theta_j-1/2} (-\theta_j + \tfrac{1}{2} + \delta + \nu_0)[a_{j-2} \\
&+ (a_{j-1} - a_{j-2})(\nu_0 - \theta_j + \tfrac{3}{2})]d\nu_0 \\
&+ \int_{\theta_j-1/2}^{\theta_j+1/2-\delta} (-\theta_j + \tfrac{1}{2} + \delta + \nu_0)[a_{j-1} \\
&+ (a_j - a_{j-1})(\nu_0 - \theta_j + \tfrac{1}{2})]d\nu_0 \\
&+ \int_{\theta_j+1/2-\delta}^{\theta_j+1/2} (\theta_j + \tfrac{3}{2} - \delta - \nu_0)[a_{j-1} + (a_j - a_{j-1})(\nu_0 - \theta_j + \tfrac{1}{2})]d\nu_0 \\
&+ \int_{\theta_j+1/2}^{\theta_j+3/2-\delta} (\theta_j + \tfrac{3}{2} - \delta - \nu_0)[a_j + (a_{j+1} - a_j)(\nu_0 - \theta_j - \tfrac{1}{2})]d\nu_0 \\
&= \tfrac{1}{6}(a_{j-1} + 4a_j + a_{j+1}) + \tfrac{1}{2}\delta(a_{j-1} - a_{j+1}) \\
&+ \tfrac{1}{2}\delta^2(a_{j-1} - 2a_j + a_{j+1}) + \tfrac{1}{6}\delta^3(a_{j-2} - 3a_{j-1} + 3a_j - a_{j+1}).
\end{aligned} \tag{45}$$

Hence the transfer matrix T is an asymmetric circulant with first row equal to

$$\tfrac{1}{6}[4 - 6\delta^2 + 3\delta^3, (1-\delta)^3, 0, \ldots, 0, \delta^3, 1 + 3\delta + 3\delta^2 - 3\delta^3]. \tag{46}$$

Thus a sharp spectral line is dispersed asymmetrically into the adjoining resolution elements. Furthermore a sharp line at one end of the spectrum affects elements at the opposite end of the spectrum.

To deal with this problem the mask should first be repositioned so as to minimize the misalignment. The remaining error can then be removed by using the correct T^{-1} in Eq. (26).

B. Differences Between Slit Width and Step Size

When a mask contains a large number of slits, a similarly large number of steps must be taken in order to moved the encoding mask from one extreme position to the other. Over such a range of steps, small differences between the step size and the encoding slit width can accumulate. For example, a typical spectral mask might have a slit width of 0.1 mm and be 255 slits wide. Over this total width of 2.55 cm, the mask must be stepped with sufficient precision that its final position is precisely one step short of cycling into the initial configuration. Thus the systematic error in the mask motion has to be less than 4×10^{-5} mm per step for the final position to be within one-tenth of a slit width from the intended location.

If such precision cannot be attained, sharp spectral lines will take on a broadened appearance. Figure 5 shows a computer simulation of this effect. The analysis of this distortion is very similar to that in Sec. III.A.

Suppose the initial position of the mask is correct, and the final position is $n\Delta$ behind the correct position. Then Eq. (5) must be replaced by

$$\tau_j^i = \int_{\theta_j - i\Delta}^{\theta_{j+1} - i\Delta} G(\nu) d\nu. \tag{47}$$

The transfer matrix T for the instrument of Example 1 has ith row equal to

$$\begin{aligned} &\tfrac{1}{6}(0, \ldots, 0, (i\Delta)^3, 1 + 3i\Delta + 3(i\Delta)^2 - 3(i\Delta)^3, \\ &4 - 6(i\Delta)^2 + 3(i\Delta)^3, (1 - i\Delta)^3, 0, \ldots, 0), \end{aligned} \tag{48}$$

with the entry $4 - 6(i\Delta)^2 + 3(i\Delta)^3$ on the main diagonal.

C. Excessive Gap Between Encoding and Blocking Masks

If the blocking mask and encoding mask are mounted too close together, there is a danger that one of the opaque elements of the encoding mask may catch the edge of the blocking mask. To avoid this there must be a gap between the two masks. If this gap is too large, however, some of the radiation that strikes the blocking mask obliquely will pass into an encoding slit that actually was meant to be obscured by the blocking mask and therefore inaccessible to any incident radiation. Then a spectral line imaged on either extreme end of the blocking mask will produce a false spectral line at the opposite end of the spectrum. This effect is most pronounced in fast optical systems (with small focal ratios).

As long as the encoding (rather than the blocking) mask is placed in the plane where the sharpest focus is obtained, this effect is not very serious because it only affects the extreme elements at the two ends of the spectrum. In contrast the mask misalignment mentioned in Sec. III.B produces distortion over the entire spectrum.

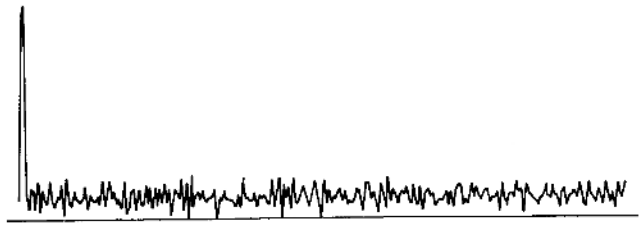

Fig. 5. Computer simulation of the errors produced by a systematic difference between slit width and step size. The final position of the mask after 255 steps is displaced by one slit width from the correct position. An input spectrum with a single sharp spectral line has been broadened to a width of two spectral elements, and noise has appeared across the whole spectrum.

D. Nonlinearities

Any of the versions of Eq. (12) derived in this paper will only be valid as long as there is a linear relation between the input and the detector readings. Eventually this linear relation will fail, as the number of spectral elements increases, or as the resolution is increased, for example. Equation (12) must then be replaced by a system of simultaneous nonlinear equations. Although numerical techniques for solving systems of nonlinear equations are available,[25,26] little work has been done so far to determine the exact form of the equations that will be needed to replace Eq. (12).

Some nonlinearities can be compensated for very simply:

(a) A nonlinear response of the detector to increasing amounts of radiation can be eliminated through judicious calibration which permits nonlinear (compensating) scaling of the actual detector output before further data processing is undertaken.

(b) A nonlinear wavelength or wavenumber response—when $G(\nu)$, the spectral display function along the exit plane, is not a linear function of ν—can be compensated for by suitably plotting the final spectrum in a way that takes the actual wavelength calibration of the instrument into account.

(c) However, if the detector response at one wavenumber is a nonlinear function of the other wavenumbers present, compensation becomes very difficult, and the best procedure may be to solve for the spectrum iteratively by successive approximations.

IV Distortion Introduced by a Continuously Moving Mask

In this section we analyze the distortion introduced when the slits move continuously across the exit focal plane instead of being discretely stepped between measurements. (An alternative analysis, which applies also to imagers, has been given by Gottlieb.[1] However, the matrix approach given here seems simpler.) We

shall see that in many cases the basic Eq. (12) still holds, with T given by Eq. (13), where the entries t^i_{j-k} are obtained from Eq. (7). Again the spectrum is estimated by $\hat{\mathbf{a}} = W^{-1}T^{-1}\eta$, and the effect of T^{-1} is (theoretically at least) to eliminate the distortion caused by the moving slits.

A. Derivation of Basic Equation

As mentioned in Sec. II.C we assume that the configuration matrix is a left circulant:

$$W = \begin{bmatrix} w_0 & w_1 & \cdots & w_{n-1} \\ w_1 & w_2 & \cdots & w_0 \\ w_2 & w_3 & \cdots & w_1 \\ \cdot & \cdot & \cdots & \cdot \\ w_{n-1} & w_0 & \cdots & w_{n-2} \end{bmatrix} \tag{49}$$

In many cases the S-matrices[3,15] can be arranged so as to have this form. An example with $n = 7$ is given in Eq. (50):

$$S_7 = \begin{bmatrix} 1 & 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 & 0 & 1 \\ 0 & 1 & 1 & 1 & 0 & 1 & 0 \end{bmatrix} \tag{50}$$

Then instead of using n separate masks of length n—one for each row of Eq. (49)—we can use one long mask of length $2n$ which is moved continuously across the exit plane. Figure 6 shows the mask of length 14 corresponding to Eq. (50). The extra half-slit at each end of the mask is necessary to avoid errors at the ends of the spectrum. If $w_{n-1} = 0$ in Eq. (49) the extra half-slits are opaque. The mask is periodic with period n.

Ideally the slits have width $2b = \theta_{j+1} - \theta_j$. Let T_0 be the time for each measurement. When operated correctly the mask should move at a constant velocity $2b/T_0$, so as to move one slit width in the time taken to make a measurement.

In general, as in the previous section, we let τ^i_j denote the contribution toward the ith detector reading from light passing through an open slit which crosses the jth segment $\theta_j\theta_{j+1}$ (see Fig. 7). We may write

$$\tau^i_j = \int_{\nu_{i,j}}^{\nu_{i+1,j+2}} l(\nu)G(\nu)d\nu, \tag{51}$$

where $\nu_{i,j}$, $\nu_{i,j+1}$ are the endpoints of this slit at the start of the ith measurement, $\nu_{i+1,j+1}$, $\nu_{i+1,j+2}$ are the endpoints of the slit at the conclusion of the ith measurement, and $l(\nu)$ is the length of time during which light at wavenumber ν can pass through the slit. Thus $l(\nu)$ can be read off Fig. 7: it is the length of the intersection of a vertical line at position ν with the shaded region which is the path swept out by the slit. By combining Eqs. (8) and (51) we again get the basic Eq. (12).

B. Ideal Case

Figure 8 shows the ideal case, when the mask has the correct dimensions, is properly aligned, and moves at the right speed. In this case

$$l(\nu) = (\nu - \theta_i + b)T_0/2b, \quad \theta_i - b \leqslant \nu \leqslant \theta_i + b,$$
$$l(\nu) = (\theta_{i+1} + b - \nu)T_0/2b, \quad \theta_{i+1} - b \leqslant \nu \leqslant \theta_{i+1} + b. \tag{52}$$

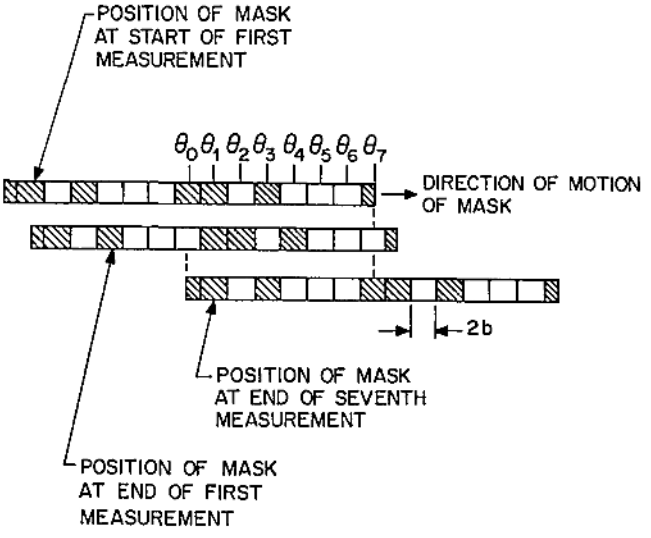

Fig. 6. Mask designed to be moved continuously. This is a mask of length 14 corresponding to the W matrix of Eq. (50). Elements 1–6 are repeated as elements 8–13, respectively. The extra half slit at each end, corresponding to element 7, is needed to avoid errors at the ends of the spectrum.

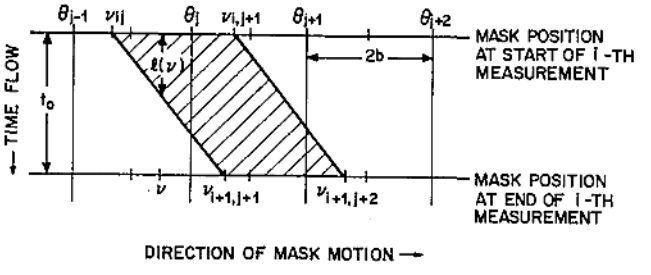

Fig. 7. Path swept out by an open slit crossing the jth segment $\theta_j\theta_{j+1}$ during the ith detector reading. Time runs downward in this diagram, and the mask moves to the right.

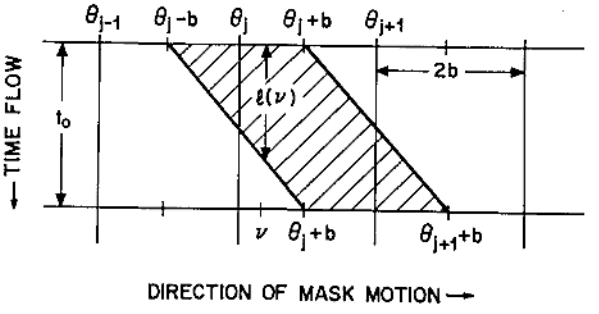

Fig. 8. The ideal case when the mask has the correct dimensions, is properly aligned, and moves at the right speed. Again time runs downward, and the mask moves to the right.

As an example, suppose the instrument of Example 1 is being operated in this ideal mode. It follows from Eqs. (4), (51), and (52) that

$$\tau_j^i = \frac{1}{384}(a_{j-2} + 76a_{j-1} + 230a_j + 76a_{j+1} + a_{j+2}). \quad (53)$$

Since this is of the same form as Eq. (7) we see that Eq. (12) still holds, where the transfer matrix T is now a circulant with first row equal to

$$\frac{1}{384}(230,76,1,0,0,\ldots,0,0,1,76). \quad (54)$$

Actually this is an approximation. Since the input spectrum $F(\nu)$ is in fact not periodic, as we have assumed, Eq. (53) must be modified for $i = 0,1,2,n-3,n-2,n-1$, and the top right-hand and bottom left-hand corners of T are missing. Thus T is a Toeplitz[19] matrix, not a circulant. However, it seems worthwhile changing T to a circulant and accepting the resulting distortion in the ends of the spectrum in order to obtain a matrix which has a manageable inverse. The inverse of the circulant (54) is easily obtained, e.g., by Method II of Sec. II.H. The inverse is a circulant with first row which approaches

$$(2.213,\ -0.826,\ 0.299,\ -0.108,\ 0.039,\ -0.014,\ 0.006,\ -0.002,\ 0.001,\ 0,\ 0,\ 0,\ldots) \quad (55)$$

correct to three decimal places, as n increases. For $n = 30$, (55) is already valid to this order of accuracy.

Again the effect of ignoring T is to broaden the spectrum, but now the broadening is more pronounced than in Fig. 3.

C. If the Mask Velocity or Slit Width is Wrong

It is not difficult to modify the above analysis to determine the distortion introduced if the mask is misaligned or if the mask velocity or slit widths are incorrect. It can be shown that, as long as each slit has the same (possibly incorrect) width, Eqs. (12) and (13) still hold, e.g., for the instrument of Example 1 operating with a misaligned mask moving with incorrect velocity we find that

$$\tau_j^i = t_2^i a_{j-2} + t_1^i a_{j-1} + t_0^i a_j + t_{-1}^i a_{j+1} + t_{-2}^i a_{j+2} \quad (56)$$

for certain constants t_r^i (again with suitable modifications for the first and last few measurements).

It does not seem worthwhile giving further examples, since each instrument will have its own transfer matrix T to be used in Eq. (12), and T should be determined when the instrument is calibrated.

D. Rotating 2-D Masks

The above effects also occur in imagers,[6,10] when the encoding mask may be a 2-D mask mounted on a rotating disk. For example, a slight radial eccentricity will cause an effective radial motion of the mask. The result of this run-out should be a superposition of three effects. There should be a widening of the image along the radial direction, similar to the effect discussed in Sec. III.B. The over-all appearance of the image should become noisier, because of the continuous motion, as in Sec. IV.A and IV.B. Third, there should be a single cycle of roughly sinusoidal change in intensity along the unfolded 2-D chain of elements, somewhat similar to an effect described in Sec. V. This is because the eccentric motion of the mask will move an opaque portion of the wheel into the open frame in the blocking mask, reducing the intensity during that portion of the cycle.

E. Imaging a Moving Source

If one of the sources being observed moves slightly during the measurements, the results will be similar to those described in Secs. III.B. and IV.D., depending on whether the motion is along the direction of the mask's motion or across it. By and large, the result of a small displacement, of the order say of one spatial element during the frame time (or complete cycling time) of the mask, will be to stretch the image along the direction of motion and to add noise to the entire scene.

V. Effect of Drift in Background Level

If the intensity of background radiation incident on a spectrometer or imaging system varies during a spectral run, the derived spectrum will necessarily be affected. Suppose, for example, that the ir spectrum of a star is to be observed. If the foreground atmospheric emission drifts during the course of the spectral run, the instrument will record a corresponding drift, and the final spectrum obtained for the star will contain a component which can be directly attributed to the drift. Because the spectrum is estimated by a linear operation [Eq. (26)], the drift component is simply added to the true stellar spectrum. In this section we analyze the magnitude of this effect for several different types of drift, assuming for simplicity that $T = I$. (The analysis in the general case is much more complicated, and we do not go into it here.)

Let η denote the vector of measurements with no drift present [Eq. (12)], let d_j be the drift in the jth measurement, $\mathbf{d} = (d_0,d_1,\ldots,d_{n-1})^T$, and let

$$\eta' = \eta + \mathbf{d} \quad (57)$$

be the actual measurements. The estimate of the spectrum is

$$\hat{\mathbf{a}} = W^{-1}\eta' = W^{-1}\eta + W^{-1}\mathbf{d}, \quad (58)$$

and we wish to analyze the drift component

$$\mathbf{D} = W^{-1}\mathbf{d}. \quad (59)$$

For concreteness we take W to be a symmetric left circulant S-matrix,[3,15] with

$$W^{-1} = S^{-1} = [2/(n+1)](2S - J). \quad (60)$$

Case 1 = Constant Offset.

Suppose

$$\mathbf{d} = \delta\mathbf{1}, \quad (61)$$

where $\mathbf{1} = (1,1,\ldots,1)^T$. Then

$$\mathbf{D} = [(2\delta)/(n+1)]\mathbf{1}. \quad (62)$$

Case 2: Sinusoidal Drift

Suppose

Table II. Noise in Spectrum Produced by Drift of Unit Amplitude

Type of drift	R.m.s. drift in data	Noise in spectrum	R.m.s. spectrum noise	Ratio of r.m.s. spectrum noise to r.m.s. drift
Constant offset	1	Constant offset	$2/n$	$2/n$
Sinusoid	$1/\sqrt{2}$	Sinusoid	$\sqrt{2}/n$	$2/\sqrt{n}$
Single spike	$1/\sqrt{n}$	Eq. (69)	$2/n$	$2/\sqrt{n}$
Random	-	Random	-	$2/\sqrt{n}$

$$d_j = \cos 2\pi a(j-\phi), \quad j = 0,1,\ldots,n-1, \tag{63}$$

where a,ϕ are constants, ϕ being a phase shift. The kth component of the drift is

$$D_k = \sum_{j=0}^{n-1} \xi_j \cos 2\pi a(j-k-\phi), \tag{64}$$

where

$$(\xi_0,\xi_1,\ldots,\xi_{n-1}) \tag{65}$$

is the first row of W^{-1}. Simplifying Eq. (64) we obtain

$$D_k = A\cos 2\pi a(k+\phi-\beta), \tag{66}$$

where the amplitude is given by

$$A^2 = \frac{4n}{(n+1)^2} + 2\sum_{i<j} \xi_i\xi_j \cos 2\pi a(i-j) \approx \frac{4}{n} \tag{67}$$

for large n, since Eq. (65) is a pseudo-random sequence,[27] and the phase shift β is given by

$$\tan(2\pi a\beta) = \frac{\sum_{j=0}^{n-1} \xi \sin 2\pi aj}{\sum_{j=0}^{n-1} \xi_j \cos 2\pi aj}. \tag{68}$$

We conclude that a sinusoidal drift voltage adds to the true spectrum a sinusoid of the same frequency as the drift, with amplitude multiplied by $2/\sqrt{n}$ and with a phase shift given by Eq. (68). Notice that a sinusoidal drift affects the spectrum more strongly than a constant offset of equal amplitude.

Case 3: A Single Noise Spike.

Suppose $d_k = \delta, d_j = 0$ for $j \neq k$. Then $\{D_j\} = \{\delta\xi_{j+k}\}$ is a pseudo-random sequence of rms value

$$\left(\frac{1}{n}\sum_j D_j^2\right)^{1/2} \approx 2\delta/n. \tag{69}$$

Thus a single noise spike of amplitude δ produces fairly random noise over the whole spectrum with rms value $2\delta/n$.

Case 4

Conversely, if $d_j = \delta\xi_{j+k}$, **D** has a single spike of amplitude δ in the kth component.

These results have been collected in Table II, where the different drifts have been normalized so as to have unit amplitude. The last line of the table shows the improvement in SNR obtained by considering random noise, obtained from Eq. (42). It is interesting that this is approximately equal to the improvement obtained when the noise is coherent. Only a constant offset in the data gives rise to a considerably smaller mean square error in the final spectrum.

Similar results are obtained in Fourier spectroscopy. A noise spike in the data will again produce a wide distribution of error signals in the final spectrum. A sinusoidal noise component, on the other hand, will produce a single spectral line at a frequency corresponding to that of the noise. (The analogous noise in the Hadamard instrument is that considered in Case 4.)

VI. Singular Designs

A. More Measurements Than Unknowns

It is sometimes desirable to design the experiment so that the number of measurements p exceeds the number of unknowns n. The purpose in doing this is to ensure that in case some measurements are lost (for example, if a cloud passes over an observatory during an astronomical observation), the spectrum can still be calculated.[28]

The basic Eq (12) relating the vector of measurements $\eta = (\eta_0,\eta_1,\ldots,\eta_{p-1})^T$ to the vector of unknowns **a** becomes

$$\eta = TW\mathbf{a} + \mathbf{e}, \tag{70}$$

where W is a $p \times n(0,1)$ configuration matrix describing the experiment, and T is a $p \times p$ transfer matrix characterizing the instrument. The (i,j) th entry of T is t_{j-i}^i, $\leqslant i,j \leqslant p-1$ [see Eq. (7)]. The best estimate for **a** is now[3,18]

$$\hat{\mathbf{a}} = (TW)^+\eta, \tag{71}$$

where $(TW)^+$ is the generalized inverse, given in this case by

$$(TW)^+ = (W^TT^TTW)^{-1}(TW)^T. \tag{72}$$

The average mean square error is then[3]

$$\epsilon = \frac{\sigma^2}{n}\,\mathrm{Trace}[(W^TT^TTW)^{-1}]. \tag{73}$$

A mask of this type suitable for use in a spectrometer can be obtained by taking W to be the first n columns of a $p \times p$ circulant S-matrix. This can be accomplished by means of a blocking mask which only exposes a frame of n slits at a time.

B. More Unknowns Than Measurements

Suppose an experiment has been designed to make n measurements in order to determine a spectrum with n unknown components, via Eqs. (12) and (26), but is prematurely terminated after only $p < n$ measurements have been made. In some cases it is still possible to say something about the unknown spectrum. For example, suppose $W = S_n$ and $T = I$. Then $\eta = S_n\mathbf{a}$, and the sum of all n measurements is equal to $(n+1)/2$ times the sum of the unknown spectral components:

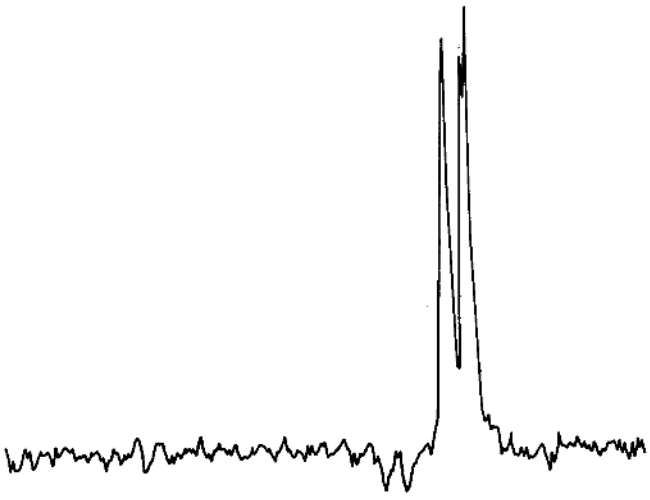

Fig. 9. Uncorrected 255-element spectrum of a laboratory source showing the 1.7-μm mercury vapor lines. There has been no data processing other than applying the inverse Hadamard transformation.

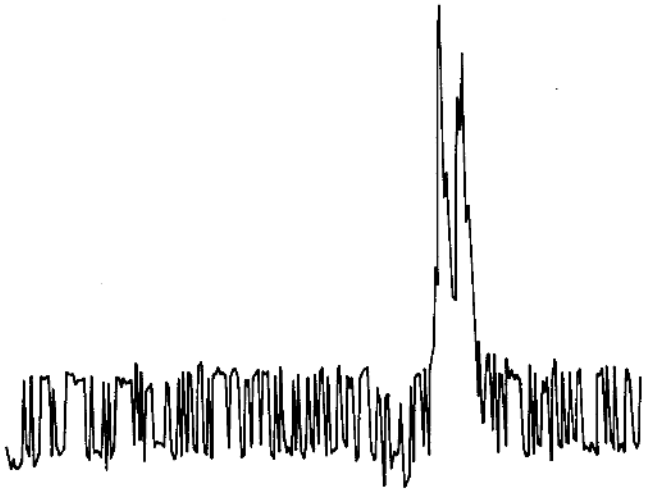

Fig. 10. Same as Fig. 9 except that the last data point, of height 1.16 $\times 10^3$, has been replaced by a noise spike of height 4×10^3. The main spectral features remain, but noise is added to the entire spectrum.

$$\sum_{j=1}^{n} \eta_j = \frac{1}{2}(n+1) \sum_{i=1}^{n} a_i. \tag{74}$$

Suppose the last measurement η_n is missing: what can be said about its possible value? Since η_n is the sum of those a_i for which there is a 1 in the last row of S_n, we have

$$\eta_n \leqslant \sum_{i=1}^{n} a_i. \tag{75}$$

From Eq. (74) and (75) we obtain the bounds

$$0 \leqslant \eta_n \leqslant \frac{2}{n-1} \sum_{j=1}^{n-1} \eta_j. \tag{76}$$

A similar result can be given if two measurements are missing. In some cases the lower bound in (76) can be improved by considering the differences $\eta_i - \eta_j$ for suitably chosen i and j.

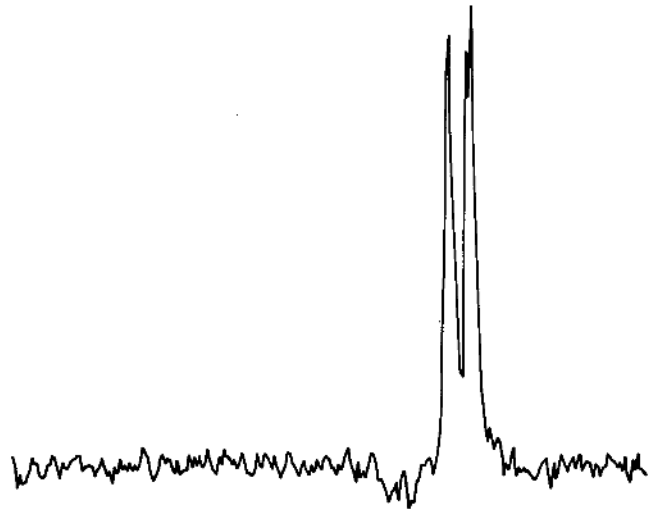

Fig 11. Same as Fig. 9, except that we have simulated losing the last five data points and replacing them by a straight line joining the 250th and 1st data points.

An analysis of this type can also be given for other multiplexing schemes, including Fourier spectroscopy. A single missing measurement never entails complete loss of information about the spectrum. The uncertainty however increases rapidly as the number of missing data points increases.

C. Correction Procedures

A similar situation arises when one or more measurements are lost because of a burst of noise. A single noise spike of amplitude δ produces random noise over the whole spectrum with rms value $2\delta/n$, as we saw in Sec. V.

A large noise spike can often be recognized by examining the other measurements obtained in the same run. When the spectrum is continuous, or contains a large number of intense lines, the individual data values do not greatly deviate from each other. Only if one or two spectral lines dominate do wide deviations occur. For a polychromatic spectrum, then, large noise spikes tend to appear as well defined extraordinary points.

In the laboratory we have tended to remove these spikes and replace them by the mean value of adjacent data points on either side. The theoretical justification for this procedure is somewhat questionable, but we find in practice that the spectra obtained by means of such corrections are rather good approximations to the expected forms. The procedure is illustrated in Figs. 9–12. Figure 9 shows a spectrum obtained in the usual way with a 255×255 S-matrix. Figure 10 shows the effect of a noise spike in the last measurement, while Fig. 11 shows the result of replacing this and four adjacent data points by a straight line joining the 250th and 1st data points. Similarly in Fig. 12 the last 15 measurements have been replaced by 15 points linearly interpolated between the 240th and 1st measurements.

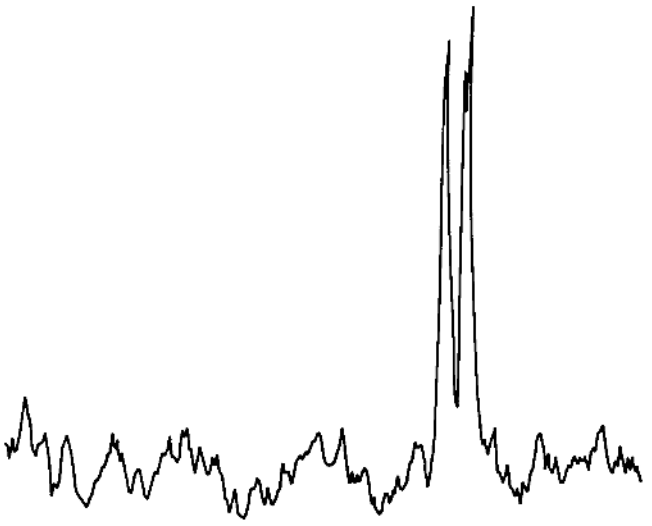

Fig. 12. Same as Fig. 11, but now the last 15 measurements have been replaced by 15 points linearly interpolated between the 240th and 1st data points.

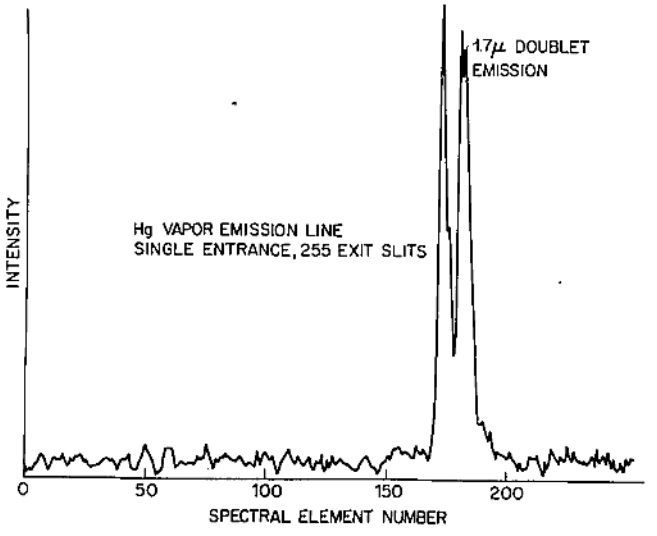

Fig. 13. The data used here are the same as in Fig. 9, except that a correction has been inserted to remove the negative echo caused by faulty slit deposition (or etching) as in Tai *et al.*[24]

A comparison of Figs. 9 and 11 suggests that linear interpolation between values spanning the gap of missing data entries produces relatively little distortion. A somewhat better procedure might be to use a quadratic or higher degree curve for the interpolation.

Figure 13 uses the same data as in Fig. 9, except that a correction has been inserted to remove the negative echo caused by faulty slit deposition (or etching), as in Tai *et al.*[24]

Thanks are due to B. F. Logan, C. L. Mallows, and S. K. Park for helpful discussions and to the MACSYMA[29] and PORT[25] computer systems for helping with some of the calculations. One of us (M.-H.T.) acknowledges the award of a NASA/NRC postdoctoral fellowship. M. H. thanks P. Mezger for his hospitality at the Max Planck Institute for Radioastronomy in Bonn and the Alexander-von-Humboldt Stiftung for a U.S. Senior Scientist award in West Germany. Work on Hadamard transform techniques at Cornell University has been supported by NASA grants NSG-1263 and NGR-33-010-210.

References

1. P. Gottlieb, IEEE Trans. Inf. Theory **IT-14**, 428 (1968).
2. R. N. Ibbett, D. Aspinall, and J. F. Graniger, Appl. Opt. **7**, 1089 (1968).
3. N. J. A. Sloane, T. Fine, P. G. Phillips, and M. Harwit, Appl. Opt. **8**, 2130 (1969).
4. E. D. Nelson and M. L. Fredman, J. Opt. Soc. Am. **60**, 1664 (1970).
5. J. A. Decker, Jr., Appl. Opt. **10**, 510 (1971).
6. M. Harwit, Appl. Opt. **10**, 1415 (1971); **12**, 285 (1973).
7. M. Harwit, P. G. Phillips, L. W. King, and D. A. Briotta, Jr., Appl. Opt. **13**, 2669 (1974).
8. C. J. Oliver and E. R. Pike, Appl. Opt. **13**, 158 (1974).
9. F. W. Plankey, T. H. Glenn, L. P. Hart, and J. D. Winebordner, Anal. Chem. **46**, 1000 (1971).
10. R. D. Swift, R. B. Wattson, J. A. Decker, Jr., R. Paganetti, and M. Harwit, Appl. Opt. **15**, 1595 (1976).
11. B. P. Lathi, *Signals, Systems and Communication*, (Wiley, New York, 1965).
12. J. E. Stewart, *Infrared Spectroscopy: Experimental Methods and Techniques* (Marcel Dekker, New York, 1970).
13. D. Raghavarao, *Constructions and Combinatorial Problems in the Design of Experiments*, (Wiley, New York, 1971).
14. A. G. Marshall and M. B. Comisarow, Anal. Chem. **47**, 491A (1975).
15. N. J. A. Sloane and M. Harwit, Appl. Opt. **15**, 107 (1976).
16. A. C. Aitken, *Determinants and Matrices* (Oliver and Boyd, Edinburgh, 1959).
17. F. J. MacWilliams and N. J. A. Sloane, *The Theory of Error-Correcting Codes* (North-Holland, Amsterdam, 1977), pp. 500, 501.
18. V. V. Federov, *Theory of Optimal Experiments* (Academic New York, 1972).
19. U. Grenander and G. Szegö, *Toeplitz Forms and Their Applications* (U. California Press, Berkeley, 1958).
20. A. Calderón, F. Spitzer, and H. Widom, Ill. J. Math. **3**, 490 (1959).
21. R. M. Gray, Toeplitz and Circulant Matrices: II, Stanford University, Stanford Electronics Laboratories, Stanford U., Stanford, Calif, Technical Report 6504-1 (April 1977).
22. G. E. Forsythe and C. B. Moler, *Computer Solution of Linear Algebraic Systems* (Prentice-Hall, Englewood Cliffs, N.J., 1967).
23. N. J. A. Sloane, *A Handbook of Integer Sequences* (Academic, New York, 1973).
24. M.-H. Tai, M. Harwit, and N. J. A. Sloane, Appl. Opt. **14**, 2678 (1975).
25. P. A. Fox, A. D. Hall, Jr., and N. L. Schryer, "The PORT Mathematical Subroutine Library," Bell Laboratories Computer Science Technical Report 47 (Bell Laboratories, Murray Hill, N.J., 1976).
26. R. P. Brent, SIAM J. Numer. Anal. **10**, 327 (1973). J. L. Blue, *Solving Systems of Nonlinear Equations*, preprint (1977).
27. F. J. MacWilliams and N. J. A. Sloane, Proc. IEEE **64**, 1715 (1976).
28. Some recent papers in the statistical literature deal with questions similar to those discussed in Secs. VI.A and VI.B, but do not appear to be directly applicable. See, for example, E. M. L. Beale and R. J. A. Little, J. R. Stat. Soc. B **37**, 129 (1975); P. W. M. John, Ann. Stat. **4**, 960 (1976); A. M. Herzberg and D. F. Andrews, J. R. Stat. Soc. B **38**, 284 (1976); D. B. Rubin, Biometrika **63**, 581 (1976).
29. Mathlab Group, *MACSYMA Reference Manual*, Version 8 (M.I.T. Laboratory for Computation, Cambridge, Mass., 1975).