THE ASTROPHYSICAL JOURNAL, 398:169–176, 1992 October 10
© 1992. The American Astronomical Society. All rights reserved. Printed in U.S.A.

# INTERPOLATION, REALIZATION, AND RECONSTRUCTION OF NOISY, IRREGULARLY SAMPLED DATA

GEORGE B. RYBICKI AND WILLIAM H. PRESS
Harvard-Smithsonian Center for Astrophysics, Cambridge, MA 02138

*Received 1992 January 27; accepted 1992 April 20*

## ABSTRACT

Various statistical procedures related to linear prediction and optimal filtering are developed for general, irregularly sampled, data sets. The data set may be a function of time, a spatial sample, or an unordered set. In the case of time series, the underlying process may be low-frequency divergent (weakly nonstationary). Explicit formulas are given for (i) maximum likelihood reconstruction (interpolation) with estimation of uncertainties, (ii) reconstruction by unbiased estimators (Gauss-Markov), (iii) unconstrained Monte Carlo realization of the underlying process, (iv) Monte Carlo realizations constrained by measured data, and (v) simultaneous reconstruction and determination of unknown linear parameters.

*Subject headings:* methods: analytical — methods: data analysis — methods: numerical

## 1. INTRODUCTION

In two previous papers (Press, Rybicki, & Hewitt 1992a, b; hereafter Papers I and II), we developed a method for determining whether two sets of irregularly sampled data are shifted measurements of the same underlying function, and, if they are, for measuring the time lag between them. Papers I and II were focused on the application to a particular object, gravitational lens 0957+561, and on two particular data sets, the optical data of Vanderriest et al. (1989), and the radio data of Lehár et al. (1992). Those papers thus did not develop the method's more general aspects in any detail.

This paper rectifies that omission. We here discuss, in a unified way, a number of related statistical procedures that can be applied to noisy, irregularly sampled data, including data whose underlying physical process is low-frequency divergent (as, e.g., a random walk process). In fact, it is not necessary that the data be a time series or other ordered one-dimensional set. The methods described here apply equally well to spatial data (for example, image reconstruction) or to data measured on an unordered set.

We are interested, primarily, in the problem of estimating the true values of the underlying physical process at points (" times ", say) which may or may not be associated with measurements. The methods we discuss all involve solving sets of linear equations over all the data. As such, they are closely related to, or generalizations of, a variety of standard techniques in the literature. Estimation at a measured point is usually called Wiener filtering, or optimal filtering. Estimation at a nonmeasured point is often called linear prediction, or least-squares prediction, or minimum variance estimation. We will see that the issue of statistical bias is an important one; the unbiased case that we discuss is usually called Gauss-Markov estimation (see, e.g., Drygas 1970; Malley 1986). We will also be interested in the issue of reconstructing, given a set of measurements, not just the " best " interpolation, but also " typical " realizations, whose statistical properties are as close as possible to the underlying process, which can then be explored with Monte Carlo simulations. This application relates closely to the so-called " missing data " or " data dropout " problem.

To the extent that the methods discussed are standard ones (see, e.g., Rao 1973; Lewis & Odell 1971), this paper should be viewed as primarily pedagogical. However, we have found that the existing literature is in practice so fragmented into special cases, and lacking in unified discussion above-named techniques, as to be almost irrelevant to the emphasis of this paper. While we do not claim anything in this paper as truly " new," neither are we able, in many cases, to recommend anything in the literature as worth consulting (at least by astrophysicists).

We hope also to be clear about where, along the way, certain statistical assumptions need (or do not need) to be made. Do we assume that a process is stationary? Do we assume that it is Gaussian? At what point is the Baysian bargain entered into?

In the interests of a practical emphasis, we will illustrate the discussion by application to a particular (artificial) data set. Figure 1 shows a set of 26 irregularly sampled data points and their error bars. In a nutshell, the question to be answered in this paper is: What is the function that underlies the measurements in Figure 1? An alternative title for this paper might be, " How To Play Connect-the-Dots in a Noisy, Fractal World."

Of course, there also exist other, quite different, approaches to the problem of irregularly sampled data (e.g., Scargle 1989) and the modeling of aperiodic or chaotic processes (e.g., Scargle 1990).

## 2. WIENER FILTERING OR LINEAR PREDICTION

Let $y_i$, $i = 1, \ldots, M$, be a set of $M$ measurements, each of which is equal to the sum of an underlying signal $s_i$ and a noise value $n_i$. It is convenient to represent the quantities as column vectors of length $M$, so

$$\boldsymbol{y} = \boldsymbol{s} + \boldsymbol{n} . \tag{1}$$

Suppose we want to estimate the value of the signal at some particular point which may or may not be one of the $M$ points already measured. Call the true value there $s_*$ (the asterisk can be thought of as taking a value in the range $1, \ldots, M$, or else having a new value $M + 1$). If our estimate is to be linear in the measured $y_i$'s, then we can write

$$s_* = \sum_{i=1}^{M} d_{*i} y_i + x_* , \tag{2}$$

where the $d_{*i}$'s are coefficients that depend on the asterisk. The summation is the linear estimate of $s_*$ and $x_*$ is the *discrepancy*

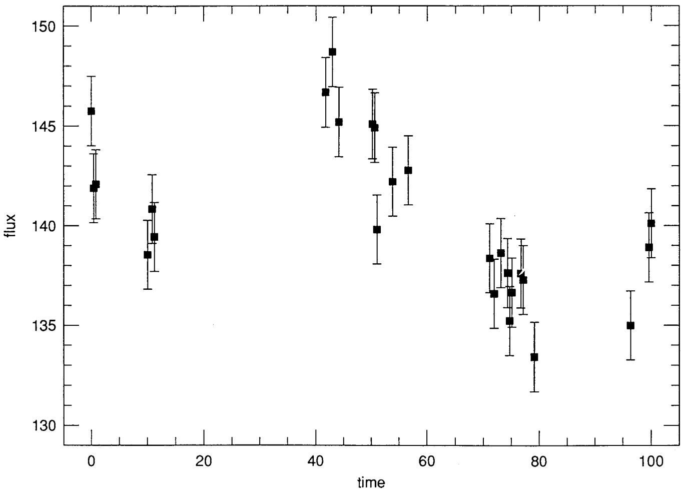

FIG. 1.—Artificial data set used as the example throughout this paper. The data are irregularly sampled, with gaps of various sizes. The process underlying these data is low-frequency divergent, something between $1/f$ noise and random walk.

of the estimate, or equivalently

$$s_* = \boldsymbol{d}_*^T \boldsymbol{y} + x_* . \qquad (3)$$

We obtain equations for $\boldsymbol{d}_*$ by minimizing the discrepancy in the least-squares sense, i.e., minimizing with respect to $\boldsymbol{d}_*$

$$\begin{aligned}\langle x_*^2 \rangle &= \langle (\boldsymbol{d}_*^T \boldsymbol{y} - s_*)^2 \rangle \\ &= \boldsymbol{d}_*^T(\langle \boldsymbol{s}\boldsymbol{s}^T \rangle + \langle \boldsymbol{n}\boldsymbol{n}^T \rangle)\boldsymbol{d}_* - 2\langle s_* \boldsymbol{s}^T \rangle \boldsymbol{d}_* + \langle s_*^2 \rangle . \qquad (4)\end{aligned}$$

Here angle brackets denote statistical ensemble averages, and we have assumed that signal and noise are uncorrelated, $\langle s_i n_j \rangle = 0$. We suppose that we know enough about the underlying process that generates $\boldsymbol{s}$ and $\boldsymbol{n}$ so that the nonvanishing averages can be considered known. (We will discuss this more below.) If we define two symmetric, positive definite correlation matrices and a "correlation vector,"

$$\boldsymbol{S} \equiv \langle \boldsymbol{s}\boldsymbol{s}^T \rangle , \quad \boldsymbol{N} \equiv \langle \boldsymbol{n}\boldsymbol{n}^T \rangle , \quad \boldsymbol{S}_* \equiv \langle (s_*)\boldsymbol{s} \rangle , \qquad (5)$$

then equation (4) can be written ("completing the square") as

$$\begin{aligned}\langle x_*^2 \rangle = (\boldsymbol{d}_* - \widehat{\boldsymbol{d}_*})^T[\boldsymbol{S} + \boldsymbol{N}](\boldsymbol{d}_* - \widehat{\boldsymbol{d}_*}) \\ - \boldsymbol{S}_*^T[\boldsymbol{S} + \boldsymbol{N}]^{-1}\boldsymbol{S}_* + \langle s_*^2 \rangle , \qquad (6)\end{aligned}$$

where

$$\widehat{\boldsymbol{d}_*} = [\boldsymbol{S} + \boldsymbol{N}]^{-1}\boldsymbol{S}_* . \qquad (7)$$

Since $\boldsymbol{S} + \boldsymbol{N}$ is positive definite, the value of $\boldsymbol{d}_*$ that minimizes equation (6) is seen to be $\boldsymbol{d}_* = \widehat{\boldsymbol{d}_*}$. The minimum variance estimate $\widehat{s_*}$ for $s_*$ (eq. [3]) is then

$$\widehat{s_*} = \widehat{\boldsymbol{d}_*}^T \boldsymbol{y} = \boldsymbol{S}_*^T[\boldsymbol{S} + \boldsymbol{N}]^{-1}\boldsymbol{y} \qquad (8)$$

and the mean square residual, i.e., the variance of $s_*$ about $\widehat{s_*}$, is

$$\langle (s_* - \widehat{s_*})^2 \rangle = \langle x_*^2 \rangle_{\min} = \langle s_*^2 \rangle - \boldsymbol{S}_*^T[\boldsymbol{S} + \boldsymbol{N}]^{-1}\boldsymbol{S}_* . \qquad (9)$$

Notice that the correlation quantities $\boldsymbol{S}$, $\boldsymbol{N}$, and $\boldsymbol{S}_*$ do not, in principle, depend on the observed data values $\boldsymbol{y}$, but only on the locations ("times") of the values, and on the underlying process. For example, if the data are an irregularly sampled time series, $y_i$ observed at time $t_i$, then a typical component $S_{ij} = \langle s_i s_j \rangle = \langle s(t_i)s(t_j) \rangle$ depends only on $t_i$ and $t_j$, but not on a particular realization of $s_i$ or $s_j$, since $s$ is the quantity that is ensemble-averaged by the angle brackets.

Note also that, in most practical cases, the noise values are uncorrelated, so that the noise correlation matrix $\boldsymbol{N}$ is diagonal $\boldsymbol{N} = \mathrm{diag}\,(\langle n_i^2 \rangle)$. However, our formulation will allow for the more general case of self-correlated noise, with a general correlation matrix. Inclusion of a known (or estimatable) "signal-noise" correlation can also be done by a simple extension of the present theory, although we shall not give the details here.

One sees in equation (8) the connection with Wiener filtering as it is more conventionally presented (see, e.g., Press et al. 1986), generally in the context of a stationary process with a regularly or continuously sampled time series, analyzed in the Fourier domain. In that case, the formula usually given for the optimal estimator is

$$\widehat{s(\omega)} = \frac{\langle |s(\omega)|^2 \rangle}{\langle |s(\omega)|^2 \rangle + \langle |n(\omega)|^2 \rangle} \, y(\omega) . \qquad (10)$$

This is exactly equation (8) in the special case that the matrices $\boldsymbol{S}$ and $\boldsymbol{N}$ are both diagonal as is indeed the case for the special assumptions made, since correlation matrices of stationary processes on equally spaced grids are diagonal in a Fourier basis.

In practice, one is often in the position of not having independent statistical information about the process $s(t)$ to estimate $\boldsymbol{S}$ a priori. In that case, one may choose to make the additional assumption of *stationarity*, so that $S_{ij}$ is a function only of the time difference $t_i - t_j$, and not of $t_i$ and $t_j$ separately. Then, every $(i, j)$ pair of data points furnishes a one-point estimate of the correlation function $S(t_i - t_j)$, and one can implement various fitting or smoothing procedures to estimate $\boldsymbol{S}$ and $\boldsymbol{S}_*$, the latter depending on the time differences $t_* - t_i$ (see Paper I; Edelson & Krolik 1988; and Hjellming & Narayan 1986; for discussion of the two-dimensional case, see Cressie 1991). Throughout this paper, the only reason to assume stationarity (in the present sense of time-translation invariance) is if there is no other way to estimate the required correlation quantities.

Indeed, there is an algebraic justification for our somewhat casual attitude about how $\boldsymbol{S}$ is estimated: Suppose that small errors in $\boldsymbol{S}$ lead one to use slightly wrong values $\widehat{\boldsymbol{d}_*}$ in the estimation equation (8). Then, the variance of the estimate for $s_*$ is always larger than equation (9). In particular, equation (6) can be rewritten as

$$\langle x_*^2 \rangle = \langle x_*^2 \rangle_{\min} + (\widehat{\boldsymbol{d}_*} - \boldsymbol{d}_*)^T[\boldsymbol{S} + \boldsymbol{N}](\widehat{\boldsymbol{d}_*} - \boldsymbol{d}_*) . \qquad (11)$$

Noting that $\boldsymbol{S}$ and $\boldsymbol{N}$ are both positive definite, one sees that the change in the variance is a pure quadratic form. Thus, first-order errors in $\boldsymbol{N}$ or $\boldsymbol{S}$, leading to first-order errors in $\widehat{\boldsymbol{d}_*}$, lead only to second-order increases in the variance of $s_*$ about $\widehat{s_*}$.

One additional statistical quantity, $\chi^2$, is defined by

$$\chi^2 \equiv \boldsymbol{y}^T[\boldsymbol{S} + \boldsymbol{N}]^{-1}\boldsymbol{y} . \qquad (12)$$

Since we have not made any assumption that the underlying processes are Gaussian, the only knowable property of $\chi^2$ is its expectation value, which is the number of data points $M$:

$$\begin{aligned}\langle \chi^2 \rangle &= \langle \mathrm{tr}\,(\chi^2) \rangle = \langle \mathrm{tr}\,([\boldsymbol{S} + \boldsymbol{N}]^{-1}\boldsymbol{y}\boldsymbol{y}^T) \rangle \\ &= \mathrm{tr}\,([\boldsymbol{S} + \boldsymbol{N}]^{-1}\langle \boldsymbol{y}\boldsymbol{y}^T \rangle) \\ &= \mathrm{tr}\,([\boldsymbol{S} + \boldsymbol{N}]^{-1}[\boldsymbol{S} + \boldsymbol{N}]) = M . \qquad (13)\end{aligned}$$

Here we have used the facts that the trace of a scalar is itself, while the trace of a matrix product is invariant under cyclic permutation of its factors. Without further assumptions, we cannot say anything about the distribution of $\chi^2$ around its mean.

Equations (8) and (9) are the principal results of this section, giving a prescription for estimating the underlying value $s_*$, and an uncertainty of the estimate, at any point. Figure 2 shows the application of these formulas to the data of Figure 1. The estimates $\widehat{s_*}$ are shown as the solid curve, while the 1 $\sigma$ standard deviations (square roots of eq. [9]) are shown as the gray "snake."

The estimates of $\boldsymbol{N}$ and $\boldsymbol{S}$ that underly Figure 2 are obtained as follows: $\boldsymbol{N}$ is taken as diagonal, with components equal to the square of the given error bars, $n_i^2$. For the component $S_{ij}$ of $\boldsymbol{S}$ we write

$$S_{ij} = \langle s_i s_j \rangle = \langle y_i y_j \rangle - n_i^2 \delta_{ij} = \langle y_i^2 \rangle E_i E_j - \tfrac{1}{2}\langle (y_i - y_j)^2 \rangle - n_i^2 \delta_{ij} \,. \quad (14)$$

Here, as a notational convenience, we define a vector $\boldsymbol{E}$ with all unit components, $E_i \equiv 1$. The trick embodied in equation (14) is general, replacing expectation quantities of $s$ (which is unmeasurable) with corresponding quantities on $y$ (which is measurable), and replacing the correlation matrix $\langle y_i y_j \rangle$ by a single population mean square, $\langle y_i^2 \rangle$, and a *structure function*

$$\langle y_i y_j \rangle = \langle y_i^2 \rangle - V_{ij} \quad (15)$$

$$V_{ij} \equiv \tfrac{1}{2}\langle (y_i - y_j)^2 \rangle \,, \quad (16)$$

which is frequently much easier to estimate from the data than is $\langle y_i y_j \rangle$ directly. We then estimate the population mean square by the sample mean square and the population structure function by a fitting method similar to that described in Paper I.

## 3. CONSTRUCTION OF AN UNBIASED ESTIMATOR

There is a quirk in equations (8) and (9) which can sometimes cause difficulties and which is easily remedied. The prediction coefficients $d_{*i}$ produced by equation (7) do not in general sum to 1, but to a value slightly less than 1. The discrepancy from 1 is greatest in gaps far from any data points. This has the peculiar effect of making the minimum variance estimate "sag" slightly, toward the value zero, in the gaps. The reason for this is that, in the absence of information from the data, the value zero is a minimum variance estimate. In fact, being constant, it has zero variance! Formally, the estimate equation (8) is a biased estimator, as can be seen (e.g., in the case of a stationary process) by

$$\langle \widehat{s_*} \rangle = \left\langle \sum_i d_{*i} y_i \right\rangle = \left( \sum_i d_{*i} \right) \langle s_i \rangle \neq \langle s_i \rangle \,. \quad (17)$$

We know of three ways of modifying equation (8) so as to obtain an unbiased estimator, all of which end up giving identical formulas for $\widehat{s_*}$. The first way is to recall the conventional wisdom that one should subtract off the mean of a data set before fitting it (adding the mean back into the fitting predictions at the end). The only question here is *what* mean, i.e., which kind of weighted mean, since we have not only a noise correlation matrix $\boldsymbol{N}$, but also a signal correlation matrix $\boldsymbol{S}$ at our disposal.

A cute way to answer the question is to find that value $\bar{y}$ that, when subtracted off, causes $\chi^2$ (eq. [13]) to be minimized. We seek to minimize with respect to $\bar{y}$

$$\chi^2 = (\boldsymbol{y} - \boldsymbol{E}\bar{y})^T [\boldsymbol{S} + \boldsymbol{N}]^{-1} (\boldsymbol{y} - \boldsymbol{E}\bar{y}) \,. \quad (18)$$

The solution is

$$\bar{y} = \frac{\boldsymbol{E}^T [\boldsymbol{S} + \boldsymbol{N}]^{-1} \boldsymbol{y}}{\boldsymbol{E}^T [\boldsymbol{S} + \boldsymbol{N}]^{-1} \boldsymbol{E}} \,. \quad (19)$$

This is the generalization of the usual "inverse-variance weighted mean" formula. One sees that a data value $y_i$ gets a small weight *either* because its variance is large, *or* because it is highly correlated with other data values (in which case it adds little new information).

In terms of this $\bar{y}$, equation (8) is now replaced by

$$\widehat{s_*} = \boldsymbol{S}_*^T [\boldsymbol{S} + \boldsymbol{N}]^{-1} (\boldsymbol{y} - \bar{y}\boldsymbol{E}) + \bar{y} \,. \quad (20)$$

That is, we subtract $\bar{y}$ from the data, and then add it back to the estimate.

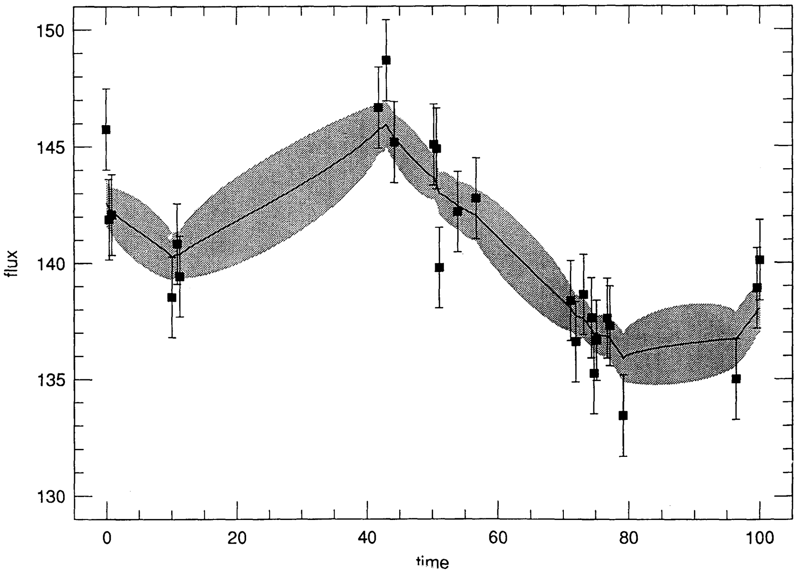

FIG. 2.—Results of applying eqs. (9) and (10) to the data of Fig. 1. The minimum variance prediction for the underlying process is the solid curve. The "snake" indicates 1 $\sigma$ error bars on the prediction. Notice that the snake narrows where the density of data is highest and widens (at a rate determined by the data's correlation function) in data gaps.

The formula for $d_*$ implied by equation (20) is

$$\widehat{d_*} = \left(1 - \frac{[S+N]^{-1}EE^T}{E^T[S+N]^{-1}E}\right)[S+N]^{-1}S_* + \frac{[S+N]^{-1}E}{E^T[S+N]^{-1}E} . \tag{21}$$

One can show that $E^T\widehat{d_*} = 1$, so the estimator is unbiased. With $\bar{y}$ given by equation (19), equation (18) can be shown (again completing the square) to be equivalent to

$$\chi^2 = y^T\left([S+N]^{-1} - \frac{[S+N]^{-1}EE^T[S+N]^{-1}}{E^T[S+N]^{-1}E}\right)y . \tag{22}$$

One here sees a projection operator that renders the value of $\chi^2$ independent of *any* constant value added to all the components of the data vector $y$. It is easy to find the expectation value of $\chi^2$ given by equation (22). Defining $Z = [S+N]^{-1}E$ and using equation (13), we have

$$\begin{aligned}\langle\chi^2\rangle &= M - \mathrm{tr}\left\langle\frac{y^TZZ^Ty}{E^TZ}\right\rangle = M - \frac{1}{E^TZ}\,\mathrm{tr}\,(ZZ\langle yy^T\rangle) \\ &= M - \frac{1}{E^TZ}\,\mathrm{tr}\,(ZE^T) = M - 1 .\end{aligned} \tag{23}$$

A second, completely different, way of getting the same result is often computationally more convenient, and also addresses more directly the issue of low-frequency divergent processes. In a low-frequency divergent (sometimes called *weakly nonstationary*) process like a random walk, the *population* mean square $\langle y^2\rangle$ may be infinite or undefined, while the sample mean square is of course finite. In equation (15) the first term on the right-hand side will thus not, in general, be estimatable, while the second term (structure function) remains well-behaved. The solution is to substitute equation (15) into equation (7) analytically, and then take the limit $\langle y^2\rangle \to \infty$ using the Sherman-Morrison formula (see, e.g., Press et al. 1986, § 2.10) and the fact that the infinite term is a matrix of rank 1. One finds that equation (8) is now transformed exactly to equation (20).

Likewise we find that equation (22) follows directly from equation (12) if the replacement $S \to S + \langle y^2\rangle EE^T$ is made and $\langle y^2\rangle \to \infty$. (The Appendix discusses a generalization of this result that is used in § 7). Armed with this knowledge of equivalence, it is often computationally convenient not to calculate $\bar{y}$ at all, but simply to use the unmodified equations (8) and (9), however substituting for $\langle y_i^2\rangle$ in equation (14) a value sufficiently big as to make the $d_{*i}$'s (as determined by eq. [7]) sum close enough to 1. In practice, it is adequate to choose $\langle y_i^2\rangle$ to be 10 or 100 times the *sample* variance. This simple trick renders most of the rococo matrix formulas in this section supernumerary, while guaranteeing equivalent results.

Third, finally, as Cressie (1991) notes, one can simply constrain the sum of the $d_{*i}$'s to 1 by minimizing not equation (4) but rather

$$\langle x_*^2\rangle = \langle(d_*^Ty - s_*)^2\rangle + 2\lambda(E^Td_* - 1) , \tag{24}$$

where $\lambda$ is a Lagrange multiplier that enforces the desired constraint. One gets straightforwardly

$$\widehat{d_*} = [S+N]^{-1}[S_* - E\lambda] , \tag{25}$$

where

$$\lambda = \frac{E^T[S+N]^{-1}S_* - 1}{E^T[S+N]^{-1}E} . \tag{26}$$

Not surprisingly, these equations are algebraically equivalent to equation (21). This is the approach usually taken in defining so-called Gauss-Markov estimators.

For the data shown in Figures 1 and 2, removal of the bias, while important in principle, makes a negligible effect in practice. In the remainder of this paper, we will assume (cf. eq. [14]) that $S$ always has the form of a rank-one matrix proportional to $EE^T$ minus a structure function $V_{ij}$ (eq. [16]) whose maximum absolute value is very much smaller than the rank-one piece. If $S$ does not start out having this form, it can be forced into this form by the addition of a constant times $EE^T$, as described above. In either case, $S$ will now have one largest eigenvalue whose eigenvector is close to $E$ (corresponding to adding a constant to the process) and the estimator $\widehat{d_*}$ will be close to unbiased.

## 4. GAUSSIAN PROCESSES

There is something disconcerting about the reconstruction shown in Figure 2: it is too smooth. While that reconstruction is in fact "closest to true" in the minimum variance sense, one has the impression that it is not, itself, a very plausible realization of the process that gave the data points that are shown.

Merely thinking this thought involves, however, some additional assumptions about the process $s$. Up to now we have assumed nothing about its full probability distribution, but only knowledge of its second moments, in $S$. To make further statements about "likely" or "unlikely" realizations, we need a full distribution. Absent any additional information, one generally makes the *Gaussian* (*or normal*) *assumption*, that the probability that a vector $s$ of values is generated is

$$P(s) \propto \exp\left[-\tfrac{1}{2}s^TS^{-1}s\right] , \tag{27}$$

where the proportionality constant is determined by normalizing the total probability to unity, and similarly for the noise process,

$$P(n) \propto \exp\left[-\tfrac{1}{2}n^TN^{-1}n\right] . \tag{28}$$

One must also make a stronger assumption about the uncorrelatedness of $s$ and $n$, not just the expectation $\langle sn^T\rangle = 0$, but true independence of probabilities,

$$P(s \text{ and } n) = P(s)P(n) . \tag{29}$$

One now calculates the probability that the two processes will generate a given set of observations $y = s + n$ as

$$\begin{aligned}P(y) &= \int P(s)P(n)\delta[y - (s+n)]d^Ms\,d^Mn \\ &= \int P(s)P(y-s)d^Ms \\ &\propto \int \exp\left\{-\tfrac{1}{2}[s^TS^{-1}s + (y-s)^TN^{-1}(y-s)]\right\}d^Ms \\ &\propto \exp\left\{-\tfrac{1}{2}y^T[S+N]^{-1}y\right\} \\ &\quad\times \int \exp\left\{-\tfrac{1}{2}(s-\hat{s})^T[S^{-1}+N^{-1}](s-\hat{s})\right\}d^Ms ,\end{aligned} \tag{30}$$

where $\hat{s} = S[S+N]^{-1}y$. Changing to $s - \hat{s}$ as the variable of integration, the integral is seen to be independent of $y$, so that finally,

$$P(y) \propto \exp\left\{-\tfrac{1}{2}y^T[S+N]^{-1}y\right\} . \tag{31}$$

Comparing equation (31) with equation (12), one sees explicitly that, as one might expect, the combined process of signal plus noise has a probability density $\propto \exp[-\chi^2/2]$. That is, it has a classical $\chi^2$ distribution, for which all of the usual probability interpretations apply. As we shall see later, this implies that for Gaussian processes the minimum square discrepancy results of the preceding sections are equivalent to maximum likelihood estimation.

## 5. UNCONSTRAINED REALIZATIONS OF THE UNDERLYING PROCESS

Having made the Gaussian assumption, we may now generate random realizations of the process $\boldsymbol{y}$. This is most easily done by first diagonalizing the "covariance" matrix $\boldsymbol{S}+\boldsymbol{N}$ appearing in equation (31). The resulting "normal modes" are then statistically independent, and the problem is reduced to choosing $M$ independent Gaussian random deviates.

We proceed as follows: First, find the eigenvalues $\lambda_1, \ldots, \lambda_M$ and eigenvectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_M$ of the positive definite, symmetric matrix $\boldsymbol{S} + \boldsymbol{N}$, equivalent to the factorization

$$[\boldsymbol{S} + \boldsymbol{N}] = \boldsymbol{V} \operatorname{diag}(\lambda_1, \ldots, \lambda_M)\boldsymbol{V}^T , \tag{32}$$

where $\boldsymbol{V}$ is the orthogonal matrix formed out of the eigenvectors by columns,

$$\boldsymbol{V} = (\boldsymbol{v}_1 \boldsymbol{v}_2 \cdots \boldsymbol{v}_M) . \tag{33}$$

Second, identify the large eigenvalue whose eigenvector is close to $\boldsymbol{E}$ and *set it to zero*. Third, let $\boldsymbol{r}$ be a vector of $M$ independent Gaussian random deviates of zero mean and unit variance. Then a realization of $\boldsymbol{y}$ is

$$\boldsymbol{y} = \boldsymbol{V} \operatorname{diag}(\lambda_1^{1/2}, \ldots, \lambda_M^{1/2})\boldsymbol{r} + \bar{y} , \tag{34}$$

where $\bar{y}$ is any mean value that you wish to give to the realization.

Alternatively, though less efficiently, we could find the eigenvalues $\xi_1, \ldots, \xi_M$, and eigenvectors $\boldsymbol{U}$, of $\boldsymbol{S}$ alone, and find the eigenvalues $\zeta_1, \ldots, \zeta_M$, and eigenvectors $\boldsymbol{Z}$, of $\boldsymbol{N}$ alone (these are trivial when $\boldsymbol{N}$ is diagonal). Setting the large eigenvalue whose eigenvector is close to $\boldsymbol{E}$ to zero, we could then construct

$$\begin{aligned} \boldsymbol{y} = \boldsymbol{s} + \boldsymbol{n} = \boldsymbol{U} \operatorname{diag}(\xi_1^{1/2}, \ldots, \xi_M^{1/2})\boldsymbol{r} \\ + \boldsymbol{Z} \operatorname{diag}(\zeta_1^{1/2}, \ldots, \zeta_M^{1/2})\boldsymbol{r}' , \end{aligned} \tag{35}$$

where $\boldsymbol{r}$ and $\boldsymbol{r}'$ are independent random vectors. The equivalence of procedures (34) and (35) is guaranteed by equation (30).

Figure 3 shows two independent realizations of the process underlying Figures 1 and 2, on a set of $M = 250$ equally spaced points and with $\bar{y} = 140$. For clarity, we have set the noise to zero in these realizations. Adding nonzero noise would simply "fuzz" the curves, independently randomly at each point, by a Gaussian of standard deviation equal to the error bars of Figures 1 or 2. We have not here plotted the data from Figure 1, because these realizations are unconditioned by that data. Such unconditioned realizations are frequently useful in Monte Carlo experiments that address questions of how probable are particular features of an observed data set (see, e.g., Papers I and II).

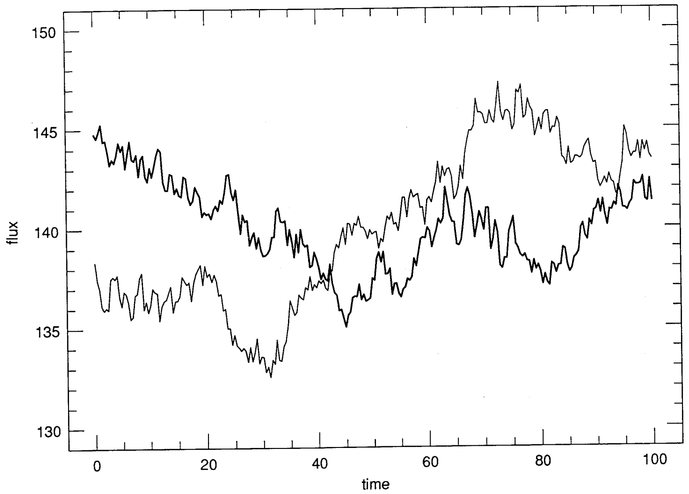

FIG. 3.—Two different unconstrained realizations of the process $s(t)$ underlying the data shown in Figs. 1 and 2. Note that these "typical" realizations do not resemble the minimum variance prediction of Fig. 2, which is "too smooth."

## 6. REALIZATIONS CONSTRAINED BY THE MEASURED DATA

Section 2 (and Fig. 2) constructed minimum variance predictions for values of a process $s$. If we take the step of making the Gaussian assumption, equations (27) and (28), then the minimum variance prediction is also the maximum likelihood prediction. However, the maximum likelihood process is not itself a very typical realization of the process. Section 4 (and Fig. 3) generated realizations that were typical, but were not constrained by the measured data. In this section we show how to combine these techniques and generate an ensemble of realizations, each of which is typical of the underlying process, but also consistent with the measured data. Such an ensemble characterizes ones full knowledge of the actual instance of the process that took place.

Although we could disguise the fact in various ways, our treatment becomes slightly Bayesian at this point (see, e.g., Loredo 1992): "Bayesian," because we are going to assign probabilities (not likelihoods) to different hypotheses about unmeasured quantities; "slightly," because, in making the Gaussian assumption, we have already postulated a set of measurements drawn from a stochastic process with a well-defined probability measure.

Bayes theorem gives the probability of an underlying process $\boldsymbol{s}$ *given* a set of measurements $\boldsymbol{y}$,

$$\begin{aligned} P(\boldsymbol{s} \mid \boldsymbol{y}) &= \frac{P(\boldsymbol{s})P(\boldsymbol{y} \mid \boldsymbol{s})}{P(\boldsymbol{y})} = \frac{P(\boldsymbol{s})P(\boldsymbol{n})}{P(\boldsymbol{y})} \\ &\propto \exp\{-\tfrac{1}{2}[\boldsymbol{s}^T\boldsymbol{S}^{-1}\boldsymbol{s} + (\boldsymbol{y}-\boldsymbol{s})^T\boldsymbol{N}^{-1}(\boldsymbol{y}-\boldsymbol{s})]\} . \end{aligned} \tag{36}$$

Here the second equality follows from the fact that $\boldsymbol{y} = \boldsymbol{s} + \boldsymbol{n}$, so the probability of $\boldsymbol{y}$ conditioned on $\boldsymbol{s}$ is just the probability of $\boldsymbol{n}$. The proportionality uses equations (27) and (28), and the fact that $P(\boldsymbol{y})$ is merely a normalization factor. Once again completing a square, equation (36) can be shown to imply

$$\begin{aligned} P(\boldsymbol{s} \mid \boldsymbol{y}) &\propto \exp\{-\tfrac{1}{2}[(\boldsymbol{s} - \boldsymbol{S}[\boldsymbol{N}+\boldsymbol{S}]^{-1}\boldsymbol{y})^T[\boldsymbol{S}^{-1}+\boldsymbol{N}^{-1}] \\ &\quad \times (\boldsymbol{s} - \boldsymbol{S}[\boldsymbol{N}+\boldsymbol{S}]^{-1}\boldsymbol{y})]\} \\ &\propto \exp\{-\tfrac{1}{2}[\boldsymbol{u}^T\boldsymbol{Q}^{-1}\boldsymbol{u}]\} \end{aligned} \tag{37}$$

where

$$\boldsymbol{u} \equiv \boldsymbol{s} - \boldsymbol{S}[\boldsymbol{N}+\boldsymbol{S}]^{-1}\boldsymbol{y} \tag{38}$$

$$\boldsymbol{Q} \equiv [\boldsymbol{S}^{-1}+\boldsymbol{N}^{-1}]^{-1} = \boldsymbol{S}[\boldsymbol{S}+\boldsymbol{N}]^{-1}\boldsymbol{N} = \boldsymbol{N}[\boldsymbol{S}+\boldsymbol{N}]^{-1}\boldsymbol{S} \tag{39}$$

This derivation assumes square matrices, that is, the same set of $M$ locations for the vectors $\mathbf{s}$ and $\mathbf{y}$. If (in the usual case) the set of measured $y_i$'s is sparser than the desired set of $s_j$'s, then equation (37) still holds on the combined set of points, but one must let $N_{jj} \to \infty$ for any value $j$ where there is no measured $y_j$, signifying infinite uncertainty as to the "measured" value there. Then, the "asterisk" component of equation (38) can be rewritten using equations (5) and (8) as

$$s_* = u_* + \mathbf{S}_*^T[\mathbf{N} + \mathbf{S}]^{-1}\mathbf{y} = u_* + \widehat{s_*}\,, \qquad (40)$$

where, by equation (37), $u_*$ is a Gaussian process with correlation matrix $\mathbf{Q}$ (eq. [38]).

Equation (40) is a powerful and perhaps surprising result. It says that "typical" realizations, in the correct relative probabilities, are obtained by starting with the minimum variance estimator $\widehat{s_*}$ and adding to it a Gaussian process with zero mean and correlation matrix $\mathbf{Q}$ given by equation (39). Computationally, one generates the vector $\mathbf{u}$ by finding the eigenvectors and eigenvalues of $\mathbf{Q}$, and proceeding exactly analogously to the equations (32)–(34).

It is useful to note how this works in the limiting case $\mathbf{N} \to \infty$: then $\widehat{s_*} \to 0$ by equation (8), $\mathbf{Q} \to \mathbf{S}$, and we obtain an unconstrained realization drawn from the probability distribution of equation (27). When (in the typical case) only certain diagonal elements $N_{jj} \to \infty$ (those that have no measured values), then the realization becomes, at these points, constrained only by the information propagated through $\mathbf{S}$. Conversely, if a row and column of $\mathbf{N}$ go to *zero*, then the realization is forced to exactly the measured value $y$ at that point.

Figure 4 shows several independent realizations conditioned on the data of Figure 1, generated by the procedure just described. If we generate a large number of such realizations, they will, at each abscissa $t$ have a Gaussian distribution of ordinates, centered on the minimum variance reconstruction (eq. [8]) and with a standard deviation matching the width of the "snake" in Figure 2 (square root of eq. [9]). However, there is much more information in such an ensemble than there is in Figure 2, since, for each realization, the correlations between different abcissas $t$ are correctly realized. One can thus make use of the ensemble of realizations to answer, via Monte Carlo experiments, many otherwise unaccessible statistical questions. One might ask, for example, "Given the measured data, how often will $s$ be below an upper limit $F = 143$ at time $t = 25$ *and* below an upper limit $F = 145$ at time $t = 30$?" (see Fig. 4).

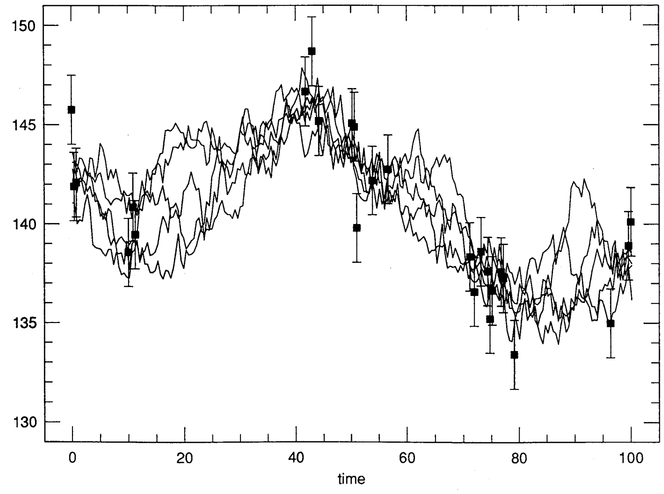

FIG. 4.—Five random realizations conditioned on the measured data points and error bars, generated by eq. (38). Each realization has the statistics of the true underlying process, is consistent with the measurements, and can be viewed as a plausible reconstruction of the actual process that transpired. The ensemble of such realizations can be used for Monte Carlo exploration of additional statistical questions.

### 7. SIMULTANEOUS RECONSTRUCTION AND DETERMINATION OF LINEAR FITTING PARAMETERS

In the derivations of § 2 we relegated $\chi^2$ to a secondary role, since we did not want its Gaussian connotations to be confusing. Now, however, it is safe to point out that equation (8) could have been derived simply as a $\chi^2$ minimization, as follows: Let $\tilde{\mathbf{y}}$ be the "augmented" vector

$$\tilde{\mathbf{y}} = \begin{pmatrix} \mathbf{y} \\ s_* \end{pmatrix}, \qquad (41)$$

whose correlation matrix is

$$\tilde{\mathbf{C}} \equiv \begin{pmatrix} \mathbf{S} + \mathbf{N} & \mathbf{S}_* \\ \mathbf{S}_*^T & \langle s_*^2 \rangle \end{pmatrix} \qquad (42)$$

so that the augmented $\chi^2$ is

$$\chi^2 = \tilde{\mathbf{y}}^T \tilde{\mathbf{C}}^{-1} \tilde{\mathbf{y}}\,. \qquad (43)$$

Then, one can readily verify (using the formula for the matrix inverse of a partitioned matrix) that minimizing $\chi^2$ with respect to the value $s_*$ gives exactly equation (8). This is much more than simply an interesting alternative derivation, however, since exp $(-\frac{1}{2}\chi^2)$ is the probability of a realization of a Gaussian process, we now see that the preceding results using minimum square discrepancy are completely equivalent to *maximum likelihood estimation* for Gaussian processes. Moreover, well-known statistical machinery may be applied to the resulting values of $\chi^2$, leading to confidence limits on the reconstructed signal values, for example.

This notion of generalizing $\chi^2$, and then minimizing it, also yields useful results when applied to the problem of simultaneously reconstructing a process $s$ and fitting for some number $N_q$ of unknown fitting parameters. Suppose that instead of $\mathbf{y} = \mathbf{s} + \mathbf{n}$ we have

$$\mathbf{y} = \mathbf{s} + \mathbf{L}\mathbf{q} + \mathbf{n}\,, \qquad (44)$$

where $\mathbf{q}$ is a vector of unknown parameters (length $N_q$) and $\mathbf{L}$ is an $M \times N_q$ matrix of known coefficients. From the measured values $\mathbf{y}$ we desire to reconstruct a best estimate of $s$ and, simultaneously, best values for the parameters $\mathbf{q}$.

The generalized $\chi^2$ is

$$\chi^2 = (\mathbf{y} - \mathbf{L}\mathbf{q})^T \mathbf{C}^{-1} (\mathbf{y} - \mathbf{L}\mathbf{q})\,, \qquad (45)$$

where $\mathbf{C} \equiv \mathbf{S} + \mathbf{N}$ is now a convenient abbreviation.

A few examples will clarify equations (44) and (45):

1. If $\mathbf{L}$ is the vector $\mathbf{E}$ (viewed as a $M \times 1$ matrix), and $\mathbf{q}$ is the value $\bar{y}$ (viewed as a $1 \times 1$ matrix), then equation (45) is exactly equation (18), which we used to produce an unbiased estimator for $s$.

2. If $\boldsymbol{L}$ and $\boldsymbol{q}$ are of the form

$$\boldsymbol{L} = \begin{bmatrix} 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 0 & 1 \\ \vdots & \vdots \\ 0 & 1 \end{bmatrix} \qquad \boldsymbol{q} = \begin{pmatrix} \bar{y}_1 \\ \bar{y}_2 \end{pmatrix} , \tag{46}$$

then the data consist of two subsets having different (unknown) means. Minimization of equation (45) will, in effect, adjust the subsets to a common offset before determining the best reconstruction of the underlying process $s$. This is precisely what we did in Papers I and II to reduce the two gravitational lens images (which had different, unknown, magnifications) to a common basis.

3. If $\boldsymbol{L}$ and $\boldsymbol{q}$ are of the form

$$\boldsymbol{L} = \begin{bmatrix} 1 & t_1 & t_1^2 \\ 1 & t_2 & t_2^2 \\ \vdots & \vdots & \\ 1 & t_M & t_M^2 \end{bmatrix} \qquad \boldsymbol{q} = \begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \alpha_2 \end{pmatrix} , \tag{47}$$

then minimization of equation (45) has the effect of removing a quadratic trend $\alpha_0 + \alpha_1 t + \alpha_2 t^2$ from the measured data $\boldsymbol{y}$ before reconstructing a model with correlation matrix $\boldsymbol{S}$.

Now properly oriented, we can manipulate equation (45), once again completing its square, to get

$$\chi^2 = [\boldsymbol{q} - \hat{\boldsymbol{q}}]^T(\boldsymbol{L}^T\boldsymbol{C}^{-1}\boldsymbol{L})[\boldsymbol{q} - \hat{\boldsymbol{q}}] + \boldsymbol{y}^T[\boldsymbol{C}^{-1} - \boldsymbol{C}^{-1}\boldsymbol{L}(\boldsymbol{L}^T\boldsymbol{C}^{-1}\boldsymbol{L})^{-1}\boldsymbol{L}^T\boldsymbol{C}^{-1}]\boldsymbol{y} , \tag{48}$$

where

$$\hat{\boldsymbol{q}} = (\boldsymbol{L}^T\boldsymbol{C}^{-1}\boldsymbol{L})^{-1}\boldsymbol{L}^T\boldsymbol{C}^{-1}\boldsymbol{y} . \tag{49}$$

From equations (45) and (48), we can now read off the answers to the simultaneous reconstruction of $s$ and $\boldsymbol{q}$: The parameters $\boldsymbol{q}$ are estimated by $\hat{\boldsymbol{q}}$, which clearly minimizes $\chi^2$ given in equation (48). The covariance matrix of this estimate (whose diagonal elements, e.g., are the standard errors for the fitted parameters) is $(\boldsymbol{L}^T\boldsymbol{C}^{-1}\boldsymbol{L})^{-1}$, which is indeed an $N_q \times N_q$ matrix. Once $\hat{\boldsymbol{q}}$ is calculated, the vector $\boldsymbol{s}$ is estimated by (cf. eq. [8])

$$\hat{\boldsymbol{s}} = \boldsymbol{S}\boldsymbol{C}^{-1}(\boldsymbol{y} - \boldsymbol{L}\hat{\boldsymbol{q}}) , \tag{50}$$

while $s_*$ at any other point is estimated by

$$\widehat{s_*} = \boldsymbol{S}_*^T\boldsymbol{C}^{-1}(\boldsymbol{y} - \boldsymbol{L}\hat{\boldsymbol{q}}) . \tag{51}$$

The variance of $s_*$ about this estimate is

$$\langle(s_* - \widehat{s_*})^2\rangle = \langle s_*^2\rangle - \boldsymbol{S}_*^T[\boldsymbol{C}^{-1} - \boldsymbol{C}^{-1}\boldsymbol{L}(\boldsymbol{L}^T\boldsymbol{C}^{-1}\boldsymbol{L})^{-1}\boldsymbol{L}^T\boldsymbol{C}^{-1}]\boldsymbol{S}_* . \tag{52}$$

In practice, it is often easier to simply use the formulas for the case of no $\boldsymbol{q}$ parameter fitting, but add to $\boldsymbol{S}$ a large scalar multiple of the product $\boldsymbol{L}\boldsymbol{L}^T$. (The validity of this procedure is proved in the Appendix.) This is a generalization of the analogous technique used in preceding sections for subtracting the mean of the process.

## 8. DISCUSSION FOR BAYESIANS ONLY

The Bayesian reader may be squirming with displeasure at the procedure just discussed, since the estimated parameters $\hat{\boldsymbol{q}}$ are seemingly treated quite differently from the estimated signal $\hat{\boldsymbol{s}}$: The former are estimated by maximum likelihood. These maximum likelihood values are then frozen during the estimation of $\hat{\boldsymbol{s}}$. From a frequentist viewpoint, this is the only way to proceed; we *have* made the Gaussian assumption that $\boldsymbol{s}$ has a well-defined probability, but we *have not* made such an assumption about the parameters $\boldsymbol{q}$. In a spirit of statistical ecumenism, however, we can happily report that, for this particular problem, a straightforward Bayesian calculation, treating $\boldsymbol{s}$ and $\boldsymbol{q}$ democratically, gives identical results:

We write the condition probability of $\boldsymbol{s}$ and $\boldsymbol{q}$ given $\boldsymbol{y}$, by Bayes theorem, as

$$P(\boldsymbol{s}, \boldsymbol{q} \,|\, \boldsymbol{y}) = \frac{P(\boldsymbol{s}, \boldsymbol{y} \,|\, \boldsymbol{q})P(\boldsymbol{q})}{P(\boldsymbol{y})} , \tag{53}$$

where the conditional probability $P(\boldsymbol{s}, \boldsymbol{y} \,|\, \boldsymbol{q})$ is clearly given by

$$P(\boldsymbol{s}, \boldsymbol{y} \,|\, \boldsymbol{q}) \propto \exp\{-\tfrac{1}{2}[\boldsymbol{s}^T\boldsymbol{S}^{-1}\boldsymbol{s} + (\boldsymbol{y} - \boldsymbol{s} - \boldsymbol{L}\boldsymbol{q})^T \times \boldsymbol{N}^{-1}(\boldsymbol{y} - \boldsymbol{s} - \boldsymbol{L}\boldsymbol{q})]\} . \tag{54}$$

To use equation (53) to find the most probable estimates of $\boldsymbol{s}$ and $\boldsymbol{q}$, we must first consider the factors $P(\boldsymbol{q})$ and $P(\boldsymbol{y})$. Actually, $P(\boldsymbol{y})$ is irrelevant, since it is merely a constant (recall $\boldsymbol{y}$ *is given*). However, the quantity $P(\boldsymbol{q})$ poses a more serious problem, since this refers to the probability distribution of the parameters $\boldsymbol{q}$, which is, in almost all cases, unknown to us. We resolve this Bayesian dilemma, as is usual in such cases, by making the assumption that $P(\boldsymbol{q})$ is sufficiently broad that it can be considered constant for the purposes of finding the maximum of the probability function $P(\boldsymbol{s}, \boldsymbol{q} \,|\, \boldsymbol{y})$. Then, the most probable estimates of $\boldsymbol{s}$ and $\boldsymbol{q}$ can be found by minimizing the quadratic expression

$$\boldsymbol{s}^T\boldsymbol{S}^{-1}\boldsymbol{s} + (\boldsymbol{y} - \boldsymbol{s} - \boldsymbol{L}\boldsymbol{q})^T\boldsymbol{N}^{-1}(\boldsymbol{y} - \boldsymbol{s} - \boldsymbol{L}\boldsymbol{q}) , \tag{55}$$

with respect to $\boldsymbol{s}$ and $\boldsymbol{q}$. Differentiating with respect to $\boldsymbol{s}$ and $\boldsymbol{q}$ and setting the results equal to zero yields the simultaneous equations for the minimizing values $\hat{\boldsymbol{s}}$ and $\hat{\boldsymbol{q}}$,

$$\begin{aligned} (\boldsymbol{S}^{-1} + \boldsymbol{N}^{-1})\hat{\boldsymbol{s}} + \boldsymbol{N}^{-1}\boldsymbol{L}\hat{\boldsymbol{q}} &= \boldsymbol{N}^{-1}\boldsymbol{y} \\ \boldsymbol{L}^T\boldsymbol{N}^{-1}\hat{\boldsymbol{s}} + \boldsymbol{L}^T\boldsymbol{N}^{-1}\boldsymbol{L}\hat{\boldsymbol{q}} &= \boldsymbol{L}^T\boldsymbol{N}^{-1}\boldsymbol{y} . \end{aligned} \tag{56}$$

The solution of these equations for $\hat{\boldsymbol{s}}$ and $\hat{\boldsymbol{q}}$ can be found straightforwardly, yielding precisely the results of equations (49) and (50), showing that this simultaneous procedure is equivalent to our previous separate minimizations.

## 9. DISCUSSION

Linear estimation is an old subject, and it is perhaps surprising that the principal practical results of this paper (consisting of the progressively more general cases of eqs. [8]–[9]; [19]–[20]; [34]; [39]–[40]; and [49]–[52]) are not standard textbook fare. It seems likely that the reason is one of technology, not mathematics: To use the results of this paper you must be able to solve (and possibly diagonalize) linear systems whose size is *the larger* of the size of your data set and the size of the set on which you want to make estimates. For data sets of any interesting size (hundreds or thousands of points), this capability has only recently become readily available in fast desktop workstations. While there may be little in this paper that could not have been written down in the 1940s (if not 40 years earlier!), it is also true that there is little in this paper which could have been *calculated* before the 1980s–and routinely calculated only in the 1990s.

Our principal conclusion is not an equation but an orientation: One now has the capability to solve many significant problems in "classical" data analysis by *global* manipulation of the full data set. Doing so (particularly in conjunction with Monte Carlo methods) can provide unambiguous answers (as in Papers I and II) to otherwise problematic statistical questions.

We thank Jacqueline Hewitt for introducing us to this subject, and for observational insights. This work was supported in part by the US National Science Foundation, grant PHY-91-06678.

## APPENDIX

We show here how the minimization of the $\chi^2$ given in equation (45) can be expressed as a limit of a certain related $\chi^2$ expression as a parameter approaches infinity. Suppose $\boldsymbol{C}$, $\boldsymbol{L}$, $\boldsymbol{q}$, and $\boldsymbol{y}$ are defined as in § 7, and that $\lambda$ is a real parameter. Our result is then

$$\min_{\boldsymbol{q}} (\boldsymbol{y} - \boldsymbol{Lq})^T \boldsymbol{C}^{-1}(\boldsymbol{y} - \boldsymbol{Lq}) = \lim_{\lambda \to \infty} \boldsymbol{y}^T(\boldsymbol{C} + \lambda \boldsymbol{LL}^T)^{-1}\boldsymbol{y} \, . \tag{57}$$

To prove this, we first note from equation (48) that

$$\min_{\boldsymbol{q}} (\boldsymbol{y} - \boldsymbol{Lq})^T \boldsymbol{C}^{-1}(\boldsymbol{y} - \boldsymbol{Lq}) = \boldsymbol{y}^T[\boldsymbol{C}^{-1} - \boldsymbol{C}^{-1}\boldsymbol{L}(\boldsymbol{L}^T\boldsymbol{C}^{-1}\boldsymbol{L})^{-1}\boldsymbol{L}^T\boldsymbol{C}^{-1}]\boldsymbol{y} \, . \tag{58}$$

Next, from the well-known Woodbury formula (see, e.g., Press et al. 1986) we have

$$(\boldsymbol{C} + \lambda \boldsymbol{LL}^T)^{-1} = \boldsymbol{C}^{-1} - \boldsymbol{C}^{-1}\boldsymbol{L}(\lambda^{-1} + \boldsymbol{L}^T\boldsymbol{C}^{-1}\boldsymbol{L})^{-1}\boldsymbol{L}^T\boldsymbol{C}^{-1} \, . \tag{59}$$

Thus

$$\lim_{\lambda \to \infty} \boldsymbol{y}^T(\boldsymbol{C} + \lambda \boldsymbol{LL}^T)^{-1}\boldsymbol{y} = \boldsymbol{y}^T[\boldsymbol{C}^{-1} - \boldsymbol{C}^{-1}\boldsymbol{L}(\boldsymbol{L}^T\boldsymbol{C}^{-1}\boldsymbol{L})^{-1}\boldsymbol{L}^T\boldsymbol{C}^{-1}]\boldsymbol{y} \, . \tag{60}$$

Comparison of equations (58) and (60) proves the result. Note that this result also applies in the case where $\boldsymbol{q}$ is the scalar $\bar{y}$ and where $\boldsymbol{L} = \boldsymbol{E}$ (see § 3). In that case the parameter $\lambda$ may be interpreted as the variance $\langle y^2 \rangle$ of the data, as explained in the text.

From the interpretation of equation (57) as a fitting of parameters, it seems evident that the result should depend only on the $N_q$-dimensional subspace spanned by the columns of $\boldsymbol{L}$ and not on the particular columns themselves. This can be proved by making the replacement $\boldsymbol{L} \to \boldsymbol{LR}$, where $\boldsymbol{R}$ is an arbitrary, nonsingular $N_q \times N_q$ matrix, which changes the columns of $\boldsymbol{L}$, but maintains the subspace spanned by them. Then

$$\begin{aligned} \boldsymbol{C}^{-1}\boldsymbol{L}(\boldsymbol{L}^T\boldsymbol{C}^{-1}\boldsymbol{L})^{-1}\boldsymbol{L}^T\boldsymbol{C}^{-1} &\to \boldsymbol{C}^{-1}\boldsymbol{LR}(\boldsymbol{R}^T\boldsymbol{L}^T\boldsymbol{C}^{-1}\boldsymbol{LR})^{-1}\boldsymbol{R}^T\boldsymbol{L}^T\boldsymbol{C}^{-1} \\ &= \boldsymbol{C}^{-1}\boldsymbol{LR}(\boldsymbol{R})^{-1}(\boldsymbol{L}^T\boldsymbol{C}^{-1}\boldsymbol{L})^{-1}(\boldsymbol{R}^T)^{-1}\boldsymbol{R}^T\boldsymbol{L}^T\boldsymbol{C}^{-1} \\ &= \boldsymbol{C}^{-1}\boldsymbol{L}(\boldsymbol{L}^T\boldsymbol{C}^{-1}\boldsymbol{L})^{-1}\boldsymbol{L}^T\boldsymbol{C}^{-1} \, , \end{aligned} \tag{61}$$

showing that both results (58) and (60) are invariant under this replacement.

## REFERENCES

Cressie, N. 1991, Statistics for Spatial Data (New York: Wiley)

Drygas, H. 1970, The Coordinate-Free Approach to Gauss-Markov Estimation (Lecture Notes in Operations Research and Mathematical Systems 40) (Berlin: Springer)

Edelson, R. A., & Krolik, J. H. 1988, ApJ, 333, 646

Hjellming, R. M., & Narayan, R. 1986, ApJ, 310, 768

Lehár, J., Hewitt, J. N., Roberts, D. H., & Burke, B. F. 1992, ApJ, 384, 453

Lewis, T. O., & Odell, P. L. 1971, Estimation in Linear Models (Engelwood Cliffs: Prentice-Hall)

Loredo, T. J. 1992, in Statistical Challenges in Modern Astronomy, ed. E. D. Feigelson & G. J. Babu (New York: Springer)

Malley, J. D. 1986, Optimal Unbiased Estimation of Variance Components (Lecture Notes in Statistics 39) (Berlin: Springer)

Press, W. H., Flannery, B. P., Teukolsky, S. A., & Vetterling, W. T. 1986, Numerical Recipes: The Art of Scientific Computing (New York: Cambridge Univ. Press)

Press, W. H., Rybicki, G. B., & Hewitt, J. N. 1992a, ApJ, 385, 404 (Paper I)

———. 1982b, ApJ, 385, 416 (Paper II)

Rao, C. R. 1973, Linear Statistical Inference and its Applications (2d ed.; New York: Wiley)

Scargle, J. D. 1989, ApJ, 343, 874

———. 1990, ApJ, 359, 469

Stuart, A., & Ord, J. K. 1987, Kendall's Advanced Theory of Statistics, Vol. 1 (5th ed.; New York: Griffin); previous editions published as Kendall, M., & Stuart, A., The Advanced Theory of Statistics

Vanderriest, C., Schneider, J., Herpe, G., Chevreton, M., Moles, M., & Wlerick, G. 1989, A&A, 215, 1