Class of Fast Methods for Processing Irregularly Sampled or Otherwise Inhomogeneous One-Dimensional Data

George B. Rybicki and William H. Press
Harvard-Smithsonian Center for Astrophysics, 60 Garden Street, Cambridge, Massachusetts 02138
(Received 20 May 1994)

With the *ansatz* that a data set's correlation matrix has a certain parametrized form (one general enough, however, to allow the arbitrary specification of a slowly varying decorrelation distance and population variance), the general machinery of Wiener or optimal filtering can be reduced from $O(n^3)$ to $O(n)$ operations, where n is the size of the data set. The implied vast increase in computational speed can allow many common suboptimal or heuristic data analysis methods to be replaced by fast, relatively sophisticated, statistical algorithms. Three examples are given: data rectification, high- or low-pass filtering, and linear least-squares fitting to a model with unaligned data points.

PACS numbers: 06.50.–x, 02.50.Rj, 02.50.Vn

In a previous analysis of irregularly spaced observations of the gravitationally lensed quasar 0957+561 [1,2] we have used to good effect the full machinery of Wiener (or optimal) filtering in the time domain, including the use, where appropriate, of unbiased ("Gauss-Markov") estimators [3]. These, and related, techniques are general enough to be applicable to data that are not only irregularly sampled (including the notoriously common case of "gappy" data), but also to data that are highly inhomogeneous in its error bars, or with point-to-point correlations (so that errors are not independent). The principal reason that these methods are not better known and more widely used seems to be the fact that their use entails the numerical solution of sets of linear equations as large as the full data set. Only recently have fast workstations allowed application to data sets as large as a few hundred points; sets larger than a few thousand points are currently out of reach even on the largest supercomputers, since the computational burden for n data points scales as n^3. As an example, the analysis in [2] (leading to a measurement of the offset in time of the two radio images of the lensed quasar) required overnight runs on a fast workstation.

In this context, we were therefore quite surprised recently to notice that the introduction of a particular simplifying assumption (essentially the *ansatz* of a certain parametrized form of the data's correlation function) allows all the calculations already mentioned, and many more, to be done in *linear* time, that is, with only a handful of floating operations per data point. In fact, we have verified that we are able to obtain results substantially identical to [2] in less than 2 sec of computer time for ~ 160 data points, about 10^4 times faster than the previous analysis.

Speed increases of 10^4 or greater (that is, from $O[n^3]$ to $O[n]$ for n data points) are not merely computer time savers. Such increases are *enabling* for the application of sophisticated statistical techniques to data sets that hitherto have been analyzed only by heuristic and *ad hoc* methods. The fast Fourier transform (FFT) is a previous example of a numerical algorithm whose raw speed caused it to engender a considerable universe of sophisticated applications. By their nature, FFT methods are generally not applicable to irregularly sampled, or otherwise inhomogeneous, data sets (though see [4]). Although the methods we describe here are not related to the FFT in a mathematical sense, we think they have the potential to be comparably significant in engendering new and powerful techniques of data analysis. In the interest of making such new methods available to the widest possible community, we outline, in this Letter, the mathematical foundation of the class, and give three examples of early applications. We are also making available, via the Internet [5], a "developer's kit" of FORTRAN-90 code, fully implementing the examples given here.

We begin with the observation that many, if not most, one-dimensional processes of interest (e.g., measurements as a function of time t) have a characteristic decorrelation time (which may itself vary with time), so that a set of measurements y_i, $i = 1, \ldots, n$ at the ordered times $t_1 < t_2 < \cdots < t_n$, have an expected (population) correlation matrix Φ_{ij} that is peaked on the diagonal $i = j$ and decays away from the diagonal in both directions. We consider the case where this decay can be modeled, even if only roughly, by the form

$$\Phi_{ij} = \begin{cases} \exp[-\int_{t_i}^{t_j} w(t)\,dt], & t_i < t_j, \\ \exp[-\int_{t_j}^{t_i} w(t)\,dt], & t_i > t_j, \end{cases} \tag{1}$$

where $w(t)$, the reciprocal of the decorrelation time, can be thought of as slowly varying with time (or constant). Although this represents only one special type of correlation matrix, it can be applied to quite a large class of problems.

All our results derive from the remarkable fact that the inverse of the matrix (1), $\Phi_{ij}^{-1} \equiv T_{ij}$, is tridiagonal with

$$T_{ij} = \begin{cases} 1 + r_1 e_1, & i = j = 1, \\ -e_1, & 1 \le i = j - 1 \le n - 1, \\ 1 + r_i e_i + r_{i-1} e_{i-1}, & 1 < i = j < n, \\ -e_j, & i \le j = i - 1 \le n - 1, \\ 1 + r_{n-1} e_{n-1}, & i = j = n, \\ 0, & \text{otherwise}, \end{cases} \tag{2}$$

 0031-9007/95/74(7)/1060(4)$06.00

where

$$r_i \equiv \exp\left[-\int_{t_i}^{t_{i+1}} w(t)\,dt \right] \tag{3}$$

and

$$e_i \equiv (r_i^{-1} - r_i)^{-1}. \tag{4}$$

It is well known [6,7] that tridiagonal systems can be solved in linear time, requiring $8n$ arithmetic operations. A first surprising result is, therefore, that the operation of multiplying the *nonsparse* matrix Φ by any data vector y,

$$I_i = \sum_j \Phi_{ij} y_j , \tag{5}$$

can be rendered fast by instead solving (for I) the tridiagonal problem

$$\sum_j T_{ij} I_j = y_i . \tag{6}$$

The matrix Φ by itself is not a very good correlation matrix, since it is normalized to unity on the diagonal. This is easily remedied by the introduction of another (generally slowly varying or constant) positive function $V(t)$ to represent the population variance of the process, that is, the typical or *a priori* mean square amplitude for a measurement y at time t. Our correlation matrix *ansatz* is then

$$C_{ij} = [V(t_i)V(t_j)]^{1/2} \Phi_{ij} , \tag{7}$$

which has the (tridiagonal) inverse

$$[C^{-1}]_{ij} = [V(t_i)V(t_j)]^{-1/2} T_{ij} . \tag{8}$$

(The function V is not to be confused with the error or noise in a single measurement, which need not be slowly varying, see below.)

In fact, Eqs. (1)–(6) represent just one example that derives from the more general fast evaluation of "forward plus backward" integrals with the form

$$I_i = F_i + B_i , \qquad i = 1,\dots,n , \tag{9}$$

where the forward and backward pieces have individual exponentially decaying recurrences,

$$\begin{aligned} F_{i+1} &= F_i r_i + f_i , \\ B_{i-1} &= B_i r_{i-1} + b_{i-1} , \end{aligned} \tag{10}$$

where the "initial values" F_1 and B_n are specified, and where the f_i (or b_i) is the forward (or backward) increment across the single interval (t_i, t_{i+1}). The key observation is that I_j's can be found from the f's and b's (without the intermediate calculation of F's and B's) by solving the tridiagonal system

$$\sum_j T_{ij} I_j = \begin{cases} F_1 + e_1(b_1/r_1 - f_1), & i = 1 , \\ e_{i-1}(f_{i-1}/r_{i-1} - b_{i-1}) + e_i(b_i/r_i - f_i), & 2 \le i \le n-1 , \\ e_{n-1}(f_{n-1}/r_{n-1} - b_{n-1}) + B_n , & i = n . \end{cases} \tag{11}$$

The special case of Eq. (6) is recovered by the choices

$$\begin{aligned} f_i = (y_{i+1} + r_i y_i)/2 , \qquad & b_i = (y_i + r_i y_{i+1})/2 , \\ F_1 = y_1/2 , \qquad & B_n = y_n/2 . \end{aligned} \tag{12}$$

However, other choices for these quantities, e.g., corresponding to quadrature formulas across the interval (t_i, t_{i+1}), have other uses, as we will see below.

This is all the machinery we need for the three examples that we now give. The idea that certain special correlation matrices can have simple inverses is not new [8–10], but we are not aware of the previous use in data processing of tridiagonally fast forms as general as Eqs. (1)–(4), (7) and (8), or (9)–(12). Stochastic processes satisfying Eq. (7) are generalizations of the so-called Ornstein-Uhlenbeck process [11–13]

Example 1: Wiener filtering and data rectification.— Here, one is given an irregularly sampled data set $y_i = y(t_i)$, $i = 1,\dots,n$, with error estimates σ_i. The σ_i's may be highly variable, with well-measured values intermixed with poorly measured ones. One desires best estimates (in the sense, with some technical assumptions, of minimum variance, see [3]) of the underlying signal $s(t)$ either at the measured times t_i (Wiener filtering), or else at some different, usually equally spaced, set of times t'_j, $j = 1,\dots,m$ (data rectification). In practice, data rectification is now often accomplished by linear interpolation between nearest measured points. Such interpolation can give highly suboptimal results, since it can use a poorly measured near point instead of a much better measured point only negligibly farther away. The procedure described here uses, in effect, an optimal combination of points, weighting them appropriately by their combination of nearness to the desired point and smallness of their measured error.

One proceeds as follows. Step 1.1: Estimate the inverse decorrelation length $w(t)$ and the population variance $V(t)$. These estimates need not be very accurate, since the error in the result of Wiener filtering is second order in any error in the filter. It often suffices to use constant values w and V.

Step 1.2: Form an "augmented" vector of times t_i, $i = 1,\dots,N$, consisting of the union of the times at which data are measured and the times at which output is desired. (If there are no overlaps, then $N = m + n$.) Form a corresponding augmented data vector $\mathbf{y}_*$, with measured data values in the appropriate slots, an arbitrary constant value (generally the mean of the measured values) in the other slots. Form augmented "reciprocal errors" by similarly combining the measured $1/\sigma_i$ values (in the measured slots) with a value zero in the unmeasured slots.

Step 1.3: Calculate the output of the Wiener filter, by using the right-hand form of the matrix equation

$$\hat{\mathbf{s}} = \mathbf{S}^T[\mathbf{S} + \mathbf{N}]^{-1}\mathbf{y}_* = [\mathbf{N}^{-1} + \mathbf{S}^{-1}]^{-1}\mathbf{N}^{-1}\mathbf{y}_* \tag{13}$$

[see, e.g., Eq. (8) of Ref. [3]]. Here $\mathbf{N}^{-1}$ is the diagonal matrix formed from the square of the reciprocal errors, $N_{ij}^{-1} = (1/\sigma_i^2)\delta_{ij}$, while $\mathbf{S}$ is the correlation matrix [called $\mathbf{C}$ in Eq. (7), above], fully specified by the sets V_i, w_i, and t_i. In particular, since $\mathbf{S}^{-1}$ is tridiagonal, the right-hand term in brackets is also tridiagonal, and $\hat{\mathbf{s}}$ can be obtained by a single fast tridiagonal solution.

Step 1.4: Unpack the desired results from $\hat{\mathbf{s}}$, either the values at the measured t_i's or the values at the rectified t_i''s, or both.

Example 2: Low- or high-pass filtering of irregularly spaced values.—Here, for brevity, we assume that the data are error free. (A more complicated example would combine the optimal estimation of example 1 with the filtering described here.) At first sight, a matrix like Eq. (1) does not look like a very good low-pass filter, because the discontinuity in slope on the diagonal introduces significant high-frequency leakage. An important trick, however, is to allow w to be complex, but take the filter to be the real part of the result. Then, there is a *unique* exponential filter with all the following properties: (i) no discontinuity in the derivative of the impulse response, (ii) frequency response unity at $f = 0$ and flat through the third derivative, and (iii) response falls off as f^{-4} (amplitude), or 24 dB per octave (power), for $f > f_c$, where f_c is the 3 dB cutoff frequency. Using this exponential, the filter can be derived from Eq. (11) by the application of a quadrature rule (here, derived simply by representing the underlying signal as piecewise linear between its measured points) to the intervals between the points.

Step 2.1: Calculate the complex values

$$W_j = K(1 + i)f_c(t_{j+1} - t_j), \qquad j = 1,\ldots,n-1, \tag{14}$$

where $K = \sqrt{2}\pi(\sqrt{2} - 1)^{-1/4} = 5.538\,07$ for a low-pass filter and $K = \sqrt{2}\pi(\sqrt{2} - 1)^{1/4} = 3.564\,27$ for a high-pass filter. (The different values simply scale the respective 3 dB points to f_c.) For each W_j, calculate $r_j \equiv \exp[-W_j]$ [replacing Eq. (3)]. Note that the accuracy condition for the piecewise-linear quadrature is that $W_j \ll 1$ for all j.

Step 2.2: Solve the complex tridiagonal system

$$\sum_j T_{ij}u_j = \frac{1}{2} \times \begin{cases} (s_1 - s_2)/W_1, & i = 1, \\ (s_i - s_{i+1})/W_i + (s_i - s_{i-1})/W_{i-1}, & i = 2,\ldots,n-1 \\ (s_n - s_{n-1})/W_{n-1}, & i = n, \end{cases} \tag{15}$$

where the s_i's are the input values.

Step 2.3: The outputs of the high- or low-pass filters, at the locations of the input data, are given, respectively, by

$$\mathcal{H}\mathbf{s} = \mathrm{Re}(\mathbf{u}), \qquad \mathcal{L}\mathbf{s} = \mathbf{s} - \mathrm{Re}(\mathbf{u}). \tag{16}$$

(It is worth remarking that, by a different choice of complex exponential, one can similarly approximate Gaussian convolution and deconvolution.)

Example 3: Linear least squares fitting a model, where the data points and model points are not available at the same positions.—This is just one of many possible extensions of example 1. One desires the coefficient vector $\mathbf{q}$ that best fits a linear combination of model basis functions. For further explanation, see Ref. [3].

Step 3.1 is the same as step 1.1, above. Step 3.2 is similar to step 1.2. Model (as opposed to data) slots should, however, be filled with zero values for both $\mathbf{y}_*$ and $\mathbf{N}$. Additionally form an augmented matrix $\mathbf{L}_*$, each of whose columns contains the values of a single model basis function (in rows corresponding to the model slots) or zero (in rows corresponding to the data slots).

For step 3.3, as in step 1.3, note that the inverse

$$\mathbf{C}^{-1} \equiv (\mathbf{S} + \mathbf{N})^{-1} = \mathbf{S}^{-1}[\mathbf{1} + \mathbf{N}\mathbf{S}^{-1}]^{-1} \tag{17}$$

can be calculated by a single tridiagonal solution, followed by a (also fast) tridiagonal left multiplication by $\mathbf{S}^{-1}$. Thus the coefficient matrix, which is given by

$$\mathbf{q} = [\mathbf{L}_*^T(\mathbf{C}^{-1}\mathbf{L}_*)]^{-1}\mathbf{L}_*^T(\mathbf{C}^{-1}\mathbf{y}_*), \tag{18}$$

can be calculated by performing fast operations on the columns of $\mathbf{L}_*$ and $\mathbf{y}_*$.

Step 3.4: Using the inverses already calculated for $\mathbf{q}$, the χ^2 for the fit is calculated as

$$\chi^2 = (\mathbf{y}_* - \mathbf{L}_*\mathbf{q})^T[(C^{-1}\mathbf{y}_*) - (C^{-1}\mathbf{L}_*)\mathbf{q}]. \tag{19}$$

By repeating the calculations of Eqs. (18) and (19), this quantity can be minimized with respect to any additional parameters in the fit that enter nonlinearly. [We have applied this procedure with great success to a problem where the nonlinear parameter is a time scaling of the data, so that the positions of the measured data points are shifted and compressed or expanded with each iteration. Use of Eqs. (18) and (19) avoids the necessity of recalculating the model at each iteration.]

The examples given here are only the simplest of many related techniques that share the underpinnings of (i) nonsparse matrices with tridiagonal (or, more generally, band-diagonal or other fast) inverses, and (ii) judicious factorization to maintain fast evaluation at each step, as in Eqs. (13) and (17). Single exponential forms are only the simplest of a larger family. For example, correlation matrices that are approximated by the sum of *two* exponential forms with tridiagonal inverses can be treated by the factorization

$$(\mathbf{T}_1^{-1} + \mathbf{T}_2^{-1} + \mathbf{N})^{-1} = \mathbf{T}_1(\mathbf{T}_2 + \mathbf{T}_1 + \mathbf{T}_2\mathbf{N}\mathbf{T}_1)^{-1}\mathbf{T}_2, \qquad (20)$$

where $\mathbf{T}_1$ and $\mathbf{T}_2$ are each tridiagonal, and $\mathbf{N}$ is diagonal. The term in parentheses on the right is pentadiagonal, therefore also admitting fast inversion. We conjecture that analogous formulas for a sum of k exponential forms can be found and will involve no more than band-diagonal inversion with bandwith $2k + 1$.

This work was suppported in part by the National Science Foundation (PHY-91-06678).

[1] W. H. Press, G. B. Rybicki, and J. N. Hewitt, Astrophys. J. **385**, 404 (1992).
[2] W. H. Press, G. B. Rybicki, and J. N. Hewitt, Astrophys. J. **385**, 416 (1992).
[3] G. B. Rybicki and W. H. Press, Astrophys. J. **398**, 169 (1992).
[4] W. H. Press and G. B. Rybicki, Astrophys. J. **338**, 277 (1989).
[5] Code can be obtained by anonymous ftp to cfata4.harvard.edu. The ascii file to be downloaded is "pub/fast-method-toolkit".
[6] G. H. Golub and C. F. Van Loan, *Matrix Computations* (Johns Hopkins University Press, Baltimore, 1989), 2nd ed., Sec. 4.3.7.
[7] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes in FORTRAN* (Cambridge University Press, New York, 1992), 2nd ed., pp. 42-46.
[8] S. N. Roy and A. E. Sarhan, Biometrika **43**, 227 (1956).
[9] S. N. Roy, B. G. Greenberg, and A. E. Sarhan, J. R. Statist. Soc., Series B **22**, 348 (1960).
[10] F. A. Graybill, *Introduction to Matrices with Applications to Statistics* (Wadsworth Publishing Co., Belmont, 1969), Sec. 8.3.
[11] G. E. Uhlenbeck and L. S. Ornstein, Phys. Rev. **36**, 823 (1930).
[12] J. L. Doob, Ann. Math. **43**, 351 (1942).
[13] E. Parezen, *Stochastic Processes*, (Holden-Day, San Francisco, 1962), Sec. 3.4, p. 96.