NRL Release Number 07-1226-1134

Patching C_n^2 Time Series Data Holes using Principal Component Analysis

Mark P. J. L. Chang[1], Haedeh Nazari[2], Carlos O. Font, G. Charmaine Gilbreath and Eun Oh[3]

[1]Physics Department, University of Puerto Rico, Mayagüez, Puerto Rico 00680
[2]234 Calle Bellas Lomas, Mayagüez, Puerto Rico, 00682
[3]U.S. Naval Research Laboratory, Washington D.C. 20375

ABSTRACT

Measurements of C_n^2 time series using unattended commericial scintillometers over long time intervals inevitably lead to data drop-outs or degraded signals. We present a method using Principal Component Analysis (also known as Karhunen-Loève decomposition) that seeks to correct for these event–induced and mechanically–induced signal degradations. We report on the quality of the correction by examining the Intrinsic Mode Functions generated by Empirical Mode Decomposition.

Keywords: Karhunen-Loève Decomposition, Principal Component Analysis, Intrinsic Mode Functions, Empirical Mode Decomposition, Data Holes, Optical Turbulence

1. INTRODUCTION

How much information in a time series record is required for the restoration of a full data set from a partial data set? This question, as it stands, is not answerable. However, if we impose the simplification that the data set belongs to a certain, well defined class of time series data, then the problem becomes somewhat more tractable. Another way to phrase the question is to ask how many data points can be deleted (set to zero) from a 1-D discrete record and still be recoverable? Such a question is prompted by our experimental field work to record the behaviour of C_n^2 over time intervals of several weeks, under different climate conditions.[1–5] The final objective of the field work is to produce an accurate now–casting model of the behaviour of optical turbulence in a littoral environment, based on the local climate record.

The data used in this study are path integrated measures of C_n^2 measured across the visible to near infrared region. The instrumentation are two identical OSI LOA-004 systems, which employ aperture averaging to estimate the value of C_n^2. The LOA-004s have been adapted to function in a completely unsupervised mode, and are generally left unattended over the course of up to a few days during field operation. Naturally spurious events such as wind gusts can lead to misalignments of the receiver/detector pair, causing data gaps. Although for some types of data analysis techniques, data gaps of a limited size can be tolerated, this is not universal. Moreover, there exists a new class of very powerful techniques based on Empirical Mode Decomposition[4,6] that are exceedingly sensitive to lossy datasets. It is for this reason we are exploring different methodologies to synthetically fill the data holes; we describe the results from our studies using Principal Component Analysis in this paper.

2. PRINCIPAL COMPONENT ANALYSIS

The principal components of any ensemble can be used to identify the members of that ensemble. This idea forms the foundation of face recognition and tracking through *eigenfaces* (see Turk and Pentland[7]). We may extend this method to reconstruct the missing data for any data record under given restrictions. The key point is that the gappy data record must have the same, or similar, salient features as all the members of the ensemble.

Further author information: (Send correspondence to M.P.J.L.C.)
M.P.J.L.C.: E-mail: mchang@uprm.edu, Telephone: 1 787 265 3844

Atmospheric Propagation IV, edited by Cynthia Y. Young, G. Charmaine Gilbreath,
Proc. of SPIE Vol. 6551, 65510L, (2007) · 0277-786X/07/$18 · doi: 10.1117/12.724706

The principal components are the eigenvectors of the covariance matrix of the data and represent the features of the dataset. Provided that a reference library can be created, each member of the library will contribute to each eigenvector, more or less. As such, each member can be exactly represented by a linear combination of eigenvectors. Any similar data record external to the library will also be represented by a linear combination of eigenvectors, within a margin of error.

The first step for the filling procedure must therefore be to define the reference library. We may do so by collecting a family of turbulence data series that share certain specific characteristics; a difficult task since the definition of such is an open question. Moreover, since the mean value of the family of reference data plays a key part, all the members of the library would require normalisation. Again how to do this is an open question. Alternatively we may use the neighbouring record around a data hole, sectioning this information to provide the ensemble members. We opt for the latter technique since the record pre and post the data hole (within a certain time interval) ought to be similar in nature to the missing data. The mean value of this type of library would probably not differ greatly from the mean of the missing data, so normalisation would not be so crucial.

How does one determine the principal components of a reference library? Following Sirovich and Kirby,[8] let the M members (each of length N) of the reference ensemble be $\{\varphi_n\}$. Thus, the average data record of this ensemble will be

$$\overline{\varphi} =< \varphi >= \frac{1}{M}\sum_{n=1}^{M}\varphi_n \tag{1}$$

It is very reasonable to assume that departures from the mean record will provide an efficient procedure for extracting the primary features of the data. Therefore, we define

$$\phi_n = \varphi_n - \overline{\varphi} \tag{2}$$

Now, if we consider the dyadic matrix

$$C = \sum_{n=1}^{M}\phi_n\phi_n^T = AA^T \tag{3}$$

where each term of the sum signifies a second rank tensor product, we can recognize this as the ensemble average of the two point correlation of the deviations from the mean. Here, A^T is the transpose of A.

We require eigenvectors u_n of the matrix AA^T. For ensembles whose members have a large number of points $N > M$, matrix AA^T issingular and its order cannot exceed M. To find those eigenvectors of AA^T corresponding to nonzero eigen values, Turk and Pentland used a standard singular value decomposition technique, as described below.

$$\begin{aligned} A^TAv_n &= \mu_n v_n \\ AA^TAv_n &= \mu_n Av_n \\ CAv_n &= \mu_n Av_n \end{aligned} \tag{4}$$

where μ_n are the eigenvalues. This deduction can be equated to

$$Cu_n = \mu_n u_n \tag{5}$$

where $u_n = Av_n$. Thus AA^T and A^TA have the same eigenvalues and their eigenvectors are related through $u_n = Av_n$, provided that $||u_n|| = 1$. The treatment described is recognizable as the Karhunen–Loéve (KL) method.[9]

The implication is that a dataset with holes, ϕ', can be obtained from a limited summation

$$\phi' \approx \sum_{n=1}^{M} a_n u_n \tag{6}$$

where the coefficients a_n are obtained through the inner product

$$a_n = (\phi', u_n) \tag{7}$$

within a certain a priori error bound.

3. PROOF OF CONCEPT

To demonstrate the validity of the assertion of the previous section, we take a perfect data record of C_n^2 measurements over a 7 hour period starting from midnight, smoothed by a forward moving rolling average of 60 data points. The data contain 2492 points in total.

To ensure that all parts of the record are similar in terms of characteristics, we first looked to see if it could be termed self affine. The way to do this is to estimate the fractal dimension D_0, which characterises the roughness of the data, and the Hurst parameter, H, which is a measure of the long range dependence (LRD) within the data. In principle these two quantities are independent of each other, with D_0 being a local measure and H being global in scope. Nevertheless, in the case of self affine (self similar) sets, the two notions are closely linked. Mandelbrot's celebrated relationship between the two variables for self affine sets is[10]

$$D_0 + H = n + 1 \tag{8}$$

where n is the dimensionality of the embedding space; $n = 1$ for our case.

Estimation of both D_0 and H from experimental data is fraught with difficulties due to a host of reasons. We describe the method of estimating D_0 by determining the generalised fractal dimensions $\{D_q; q = 1, 2, 3, ...\}$ in Chang *et al.*.[3]

THE FRACTAL DIMENSION

The values of the generalised fractal dimensions obtained through correlation estimators resulted in $D_0 = 0.9946$, $D_1 = 0.9620$, $D_2 = 0.9348$, $D_4 = 0.8928$. This gives a mean value of 0.9460 and a standard deviation of 0.0431. We therefore consider the data as a monofractal, with D_0 equal to the mean value, rather than a multifractal.[11, 12] This means that the data has only one characteristic local scale exponent, regardless of dilation of the length examined.

THE HURST PARAMETER

There are a slew of methods[13] available to estimate H. For simplicity, we have opted to use the well known Hurst–Mandelbrot R/S technique, which is also the most elementary. The fitting curve for this estimator is shown in Fig. 1. We find for the Hurst parameter a mean $H = 0.99681$ with a standard deviation of 0.0069. Thus $D + H = 1.9428$, which is within 3% of the ideal result of 2 for a self affine set. This suggests that the dataset is appropriate as a test platform for our data hole filling method, since it is very similar at all scales.

4. RESULTS

We split up the test data into 21 sections, where all but 1 are members of the reference library, as shown in Fig. 2. The exclusive section is set to zero, and the algorithm described in Sec. 2 is employed on the 20 library elements. The reconstructions are created from the KL coeffiecients of the eigenvectors equivalent to the sections adjacent to the missing data. Hence we will talk of a prior and posterior reconstruction meaning e.g. for omitted section 5, we use for the prior the KL coefficient equivalent to section 4 and for the posterior reconstruction we use the coefficient equivalent to section 6 of the test data set. Note that this does not reconstruct those sections, since the eigenvectors are generated from the entire reference library.

The reconstructions shown in the left hand set of Fig. 3 represent the best level of error, while the right hand set shows the worst. It is clear that the greater the difference between the (masked off) original data section and its neighbours, the poorer the reconstruction will be. Nevertheless, the reconstruction errors for the full set of test data are all 1 order of magnitude less than the mean value of the reconstruction and the original data segment. Evidently the end points have not been synthesised to be continuous with the adjacent segments of the reference library signal, as can be seen from the error. A "continuity condition" for the function has to be imposed on both ends of the reconstructed segment in order to achieve smoothness. The most effective way to do so is currently being investigated. The eigenvalues determined through this method show a similar distribution in both cases, with minor variations only at the upper end of the spectrum, implying that the KL differences between best and worst section for reconstruction is not very large.

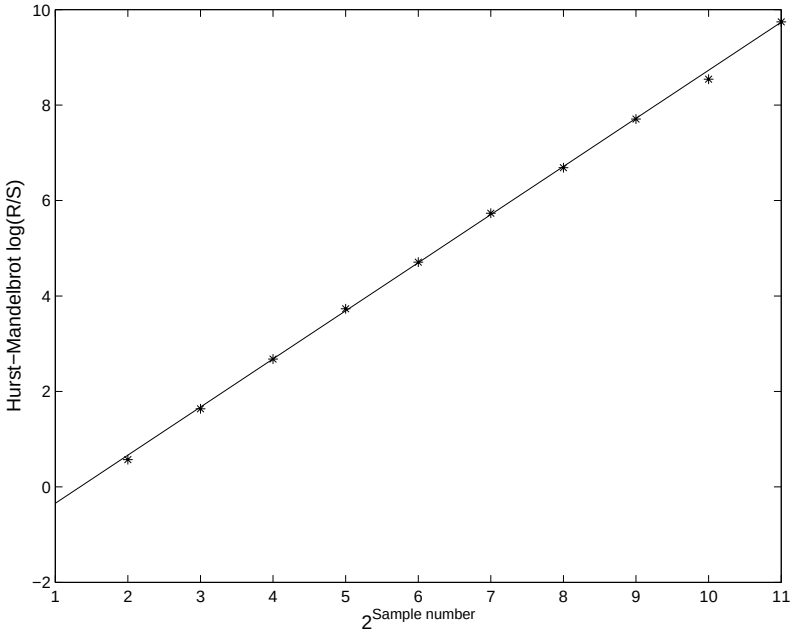

Figure 1. Hurst.

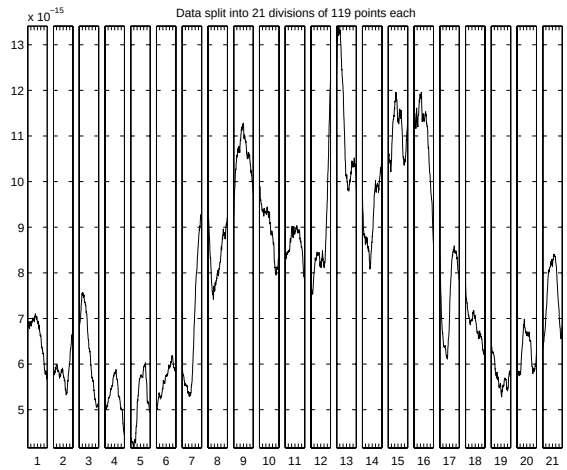

Figure 2. The test data, split into 21 sections. The data are padded before division with a set of points taken from the tail of the time series and mirrored outward.

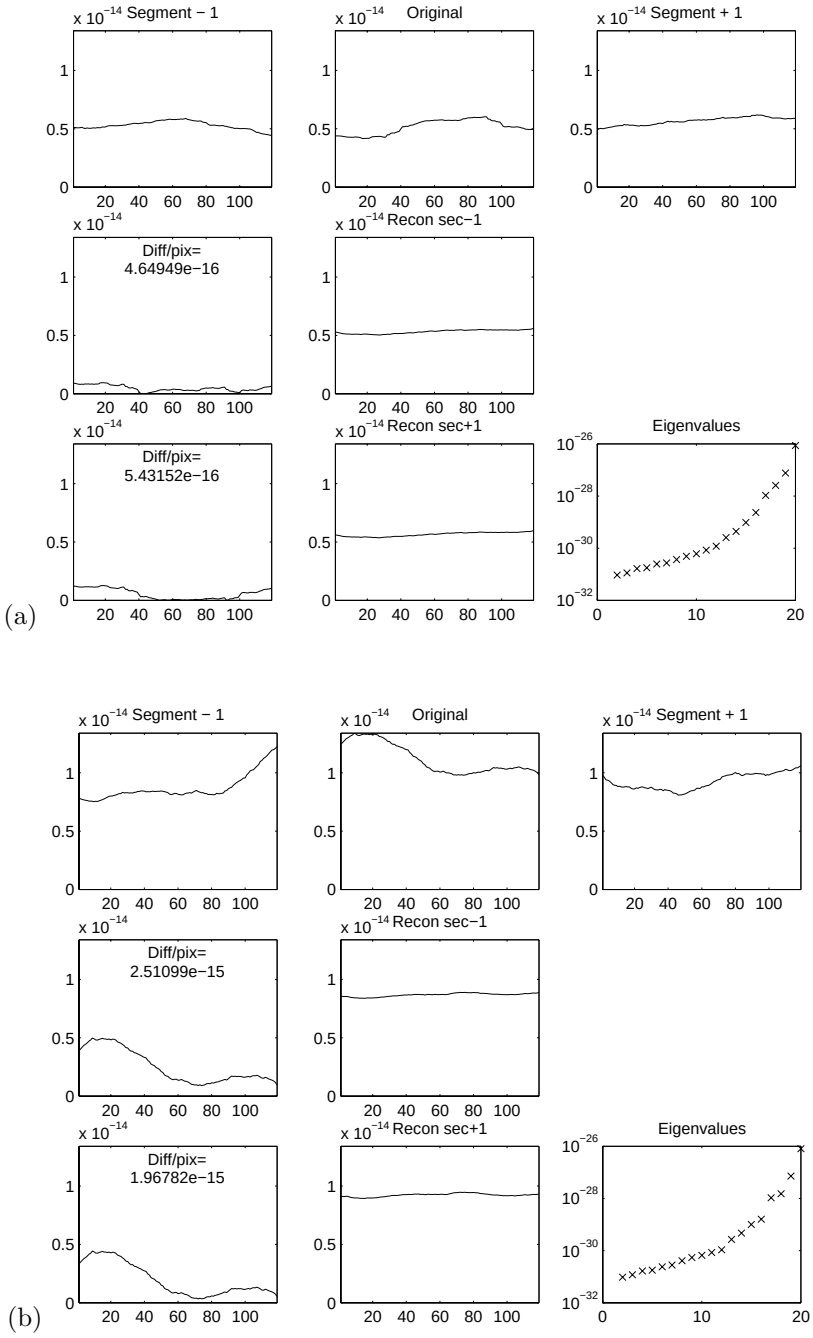

Figure 3. Best and worst reconstruction result using PCA for (a) section 5 and (b) section 13 respectively. The layout in each set is : (Top Left) Segment prior to the selection. (Top Middle) The original selected data. (Top Right) Segment after selection. (Centre Left) The difference between the reconstruction using the coefficient of eigenvector related to the prior segment and the original. The value shown is the mean absolute difference per pixel. (Centre Right) The (prior) reconstruction. (Bottom Left) The difference between the reconstruction using the coefficient related to the posterior segment and the original. (Bottom Middle) The (posterior) reconstruction. (Bottom Right) The eigenvalue spectrum.

4.1. KL and EMD

We present here the effects of patching the data hole with reconstructions determined from the prior and posterior terms with respect to the gap.

We apply the Empirical Mode Decomposition (EMD) algorithm[14] to the reconstructions. EMD is a novel adaptive method for separating a nonlinear time series into components, filtering on instantaneous frequency. The set for the best reconstruction case are illustrated in Figs. 4. We refer to these sets as EMD_L and EMD_R. The original data's intrinsic mode functions (IMFs) and residuals (set EMD_O) are also shown and for comparison, we present the effect of a simple minded linear interpolation between the endpoints of the known data on the IMFs.

Upon visual inspection, we see that the original data generates 9 components: 8 intrinsic modes and 1 residual (the stopping criterion for our EMD implementation is the same as in Huang *et al*[6]). On the other hand, the interpolated data have only 8 components. Numbering the IMFs from highest instantaneous frequency to lowest, starting from IMF 1, we can see by inspection that both EMD_L and EMD_R are strongly similar to EMD_O in IMFs 4,5,6 and 7. The differences appear in the higher frequency components, due to the discontinuity between the inserted segment and the unadulterated data. The endpoint discontinuities evidently modify the variances of IMFs 1,2 and 3, although it seems that they are only affected in the area local to the discontinuity, per IMF.

By way of comparison, a linear interpolant between the edges of the known data show that there is contamination all through the IMFs. It is so strong that IMF 7, which in the other sets clearly distinguishes the baseline rise and fall of the turbulence over the interval under study, is unable to pick out a clean pedestal.

5. CONCLUSIONS

We have discussed a data hole filling method for optical turbulence data, a necessary step to be able to use Empirical Mode Decomposition for the analysis of the time series record. The Principal Components or Karhunen–Loève eigenvectors from an ensemble of neighbouring sections of complete data around a data hole can be used to reconstruct the missing segment to a reasonable degree of accuracy, at least for the purposes of applying EMD. We have shown that the edge continuity is important, although the effect of discontinuities is not universal through all the intrinsic modes of the data. The quality of the reconstruction is much better using PCA than a simplistic linear interpolant (or merely ignoring the data gap).

ACKNOWLEDGMENTS

Part of this work was funded by the Office of Naval Research.

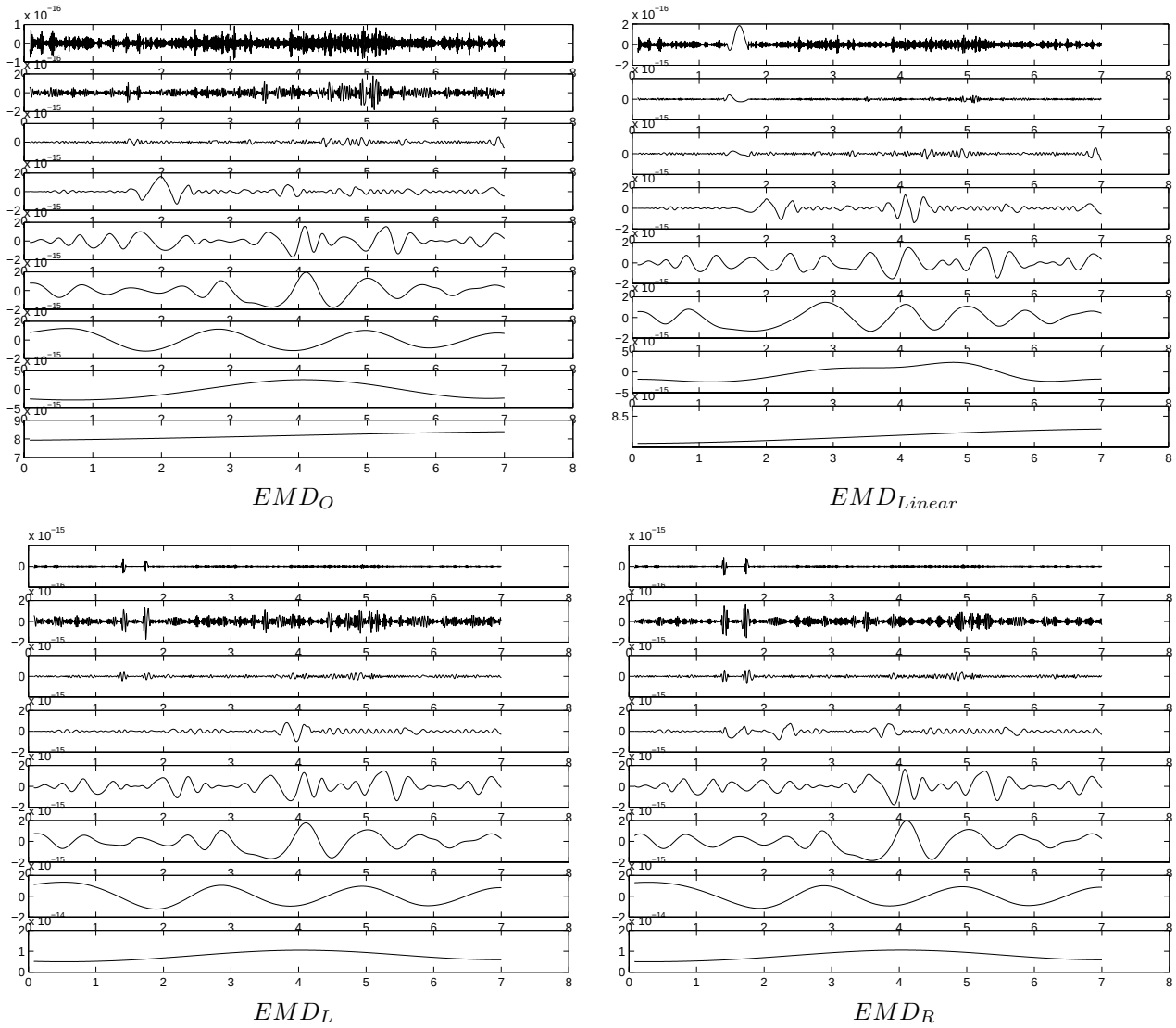

Figure 4. (Top Left) The IMFs 1 to 8 (top to bottom) and the residual trend line of EMD_O, generated by applying Empirical Mode Decomposition to the test data. (Top Right) The IMFs 1 to 7 (top to bottom) and the residue, generated from data with a linear interpolant across section 5 of the test data. (Bottom Left) The IMFs 1 to 7 and the residue of EMD_L, generated from data using the PCA reconstruction method with the coefficient of the prior section. (Bottom Right) The IMFs 1 to 7 and the residue of EMD_R, generated from data using the PCA reconstruction method of the posterior section.

REFERENCES

1. F. Santiago, M. P. J. L. Chang, C. O. Font, E. A. Roura, C. Wilcox, and S. R. Restaino, "Low altitude horizontal scintillation measurements of atmospheric turbulence over the sea: Experimental results," *Proc. SPIE* **6014**, 2005.
2. C. O. Font, M. P. J. L. Chang, E. Oh, and G. C. Gilbreath, "Humidity contribution to the refractive index structure function C_n^2," in *Atmospheric Propagation III*, C. Y. Young and G. C. Gilbreath, eds., *Proc. SPIE* **6215**, 2006.
3. M. P. J. L. Chang, C. O. Font, G. C. Gilbreath, and E. Oh, "Humidity contribution to C_n^2 over a 600m pathlength in a tropical marine environment," *Proc. SPIE* **6457**, 2007.
4. M. P. J. L. Chang, C. O. Font, G. C. Gilbreath, and E. Oh, "Humidity's influence on visible region refractive index structure parameter C_n^2," *Applied Optics (accepted), ArXiv Physics e-prints* **physics/0606075**, June 2006.
5. C. O. Font, "Understanding the atmospheric turbulence structure parameter C_n^2 in the littoral regime," Master's thesis, University of Puerto Rico at Mayagüez, 2006.
6. N. E. Huang, Z. Shen, S. R. Long, M. C. Wu, H. H. Shih, Q. Zheng, N.-C. Yen, C. C. Tung, and H. H. Liu, "The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis," *Proc. R. Soc. Lond. Ser. A* **454**, pp. 903–995, 1998.
7. M. Turk and A. Pentland, "Eigenfaces for recognition," *Journal of Cognitive Neuroscience* **3**, pp. 71–86, 1991.
8. L. Sirovich and M. Kirby, "Low–dimensional procedure for the characterization of human faces," *Journal of the Optical Society of America A* **4**, pp. 519–524, 1987.
9. J. W. Goodman, *Statistical Optics*, Wiley–Interscience, New York, 1996.
10. B. Mandelbrot, *The Fractal Geometry of Nature*, W. H. Freeman and Co., New York, 1983.
11. A. J. Roberts and A. Cronin, "Unbiased estimation of multi–fractal dimensions of finite data sets," *Physica A* **233**, pp. 867–878, 1996.
12. H. E. Stanley, L. A. N. Amaral, A. L. Goldberger, S. Havlin, P. C. Ivanov, and C.-K. Peng, "Statistical physics and physiology: Monofractal and multifractal approaches," *Physica A* **270**, pp. 309–324, 1999.
13. J. Beran, *Statistics for Long–Memory Processes*, Chapman and Hall, New York, 1994.
14. M. P. J. L. Chang, E. A. Roura, C. O. Font, E. Oh, and C. Gilbreath, "Applying the Hilbert-Huang Decomposition to horizontal light propagation C_n^2 data," in *Advances in Stellar Interferometry,* these proceedings, *Proc. SPIE* **6268**, 2006.