Degeneracy of ARMA Time Series Models and The Arrow of Time

Jeff Scargle

NASA Ames Research Center, Space Science Division, Moffett Field, CA 94035, USA

Jeffrey.D.Scargle@nasa.gov

Contents

1. Introduction

This discussion is meant to clarify some issues with autoregressive-moving average (ARMA) modeling that were discussed at the recent SAMSI workshop. The basic point is that conventional algorithms yield ARMA models that – while "fitting" the data exactly – may not actually be good representations of the underlying astronomical process. Some procedures aimed at resolving the degeneracy underlying this problem are known, but I do not think any of them provide practical, robust, and accurate models in the signal-to-noise setting of most astronomical time series. I hope statisticians will contribute to this discussion, especially if they know of algorithms addressing these issues. A useful working group task might be to explore these issues.

2. Brief Overview of ARMA Modeling

Let's start with a concise overview to set notation and describe some aspects of ARMA modeling that may not be widely appreciated among astronomers. Much of this discussion probably applies also to non-stationary models such as ARIMA and ARFIMA, but for simplicity this note is limited strictly to the ARMA context.

2.1. An Existence Theorem for MA Models: The Wold Decomposition

A remarkable theorem due to econometrician Herman Wold guarantees the existence of a moving average (MA) model for any stationary time series process. Simply stated, any such process X can be represented as the sum of a deterministic and a random part; in turn the latter can be further decomposed into the convolution of a purely random and a purely deterministic component. For data sampled at an evenly spaced set of times n we have:

$$X_n = \sum_i C_{n-i}\ R_i + D_n = C * R + D \quad \text{(The Wold Decomposition)} \tag{1}$$

where $*$ stands for the convenient convolution notation. Here D is (linearly) deterministic, R is a random white noise process, and C is a constant array (similar in operation to an electronic filter responding to an input voltage R). Often D is an unimportant trend; here we ignore it, leaving just the part known as a moving average (MA) model, $X = C * R$. Properties of this representation (see Scargle 1981b, and many textbooks) include:

1. This MA represents the data as a superposition of impulsive events.
2. These events are random in amplitude: R_n is the amplitude at time n.
3. R is uncorrelated (white noise): $\mathrm{E}(R_n R_m) = \sigma_n^2\ \delta_{n,m}$
4. Each event generates a *pulse*[1] of shape C with very special properties, namely that it is:
   (a) *Stable*, also called *convergent*: $\sum_k C_k \leq \infty$
   (b) *Causal*: $C_k = 0$ for $k < 0$; response only at or after input at $k = 0$
   (c) *Minimum delay*: this restriction on the shape of C is familiar in electrical engineering
   (d) *Constant*: all pulses have the same shape
5. There is a unique autoregressive (AR) model equivalent to this MA.

[1]Throughout we use the term *pulse* for the events. Alternative terms in astrophysics are *flare, burst, outburst, flash* or *flicker*; in mathematics *filter* and in the early literature *wavelet*.

A few notes about this remarkable theorem are in order. The random pulse picture evoked in item 1 is a natural way to describe variable astronomical sources such as gamma-ray bursts, x-ray flares from black holes, and gamma-ray flaring of the Crab Nebula – in short "shot noise" or "flicker noise" in general. But this is not the only possible view. Sometimes one is interested in the short-term or long-term "memory" aspects of time series that AR models directly express. What seems common to almost all settings is that R_i can be viewed as bringing new information into the process at time i, and is therefore called the *innovation.* Note that if R is sparse, e.g. zero for most times, one can think of the impulses as occurring at random times. Practically speaking this picture can represent pulses as random in both amplitude and time.

With regard to item 4 it is important to note that a pulse shape that violates 4a, 4b or 4c is not necessarily unphysical; these are merely mathematical characterizations of pulse shape. Item 4 also poses a conundrum: How can stationary MAs with pulse shapes C that do not have these properties be represented as one that does? Part of this puzzle connected with items 4a, 4b and 4c was unraveled in Scargle (1981b) but unsolved issues remain, especially with regard to 4d.

2.2. An Extension of the Wold Decomposition

The Wold Theorem states that for each stationary X there is a MA representation with unique innovation R and pulse shape C. (Its constructive proof even shows how to compute this model.) The restrictions of the properties of C are not essential physically – or mathematically for that matter. Accordingly a useful extension of the Wold theorem results from simply allowing more general pulse shapes – namely ones that violate one or more of the conditions in items 4a, 4b and 4c. We define a family of MA representations, in the same form as Eq. (1) but with different innovations and pulse shapes, in this extension of the basic Wold theorem:

For any stationary process X there exist:

1. a purely deterministic process D
2. a family of uncorrelated, zero-mean noise processes $\{R^{(k)},\ k = 1, 2, \ldots, N_{EW}\}$,
3. each with a different two-sided moving average filter $C^{(k)}$,

such that $X = D + C^{(k)} * R^{(k)}$ for each member of the family.

This theorem establishes the existence of a family of MA models, all of which are equivalent representations of the data X but with different innovations and pulses. The pulse shapes consist of all those which share X's autocorrelation function, the length of which determines their actual number, N_{EW}. One of them is minimum delay, one is maximum delay, and the rest are mixed delay. All of these shapes are well-defined mathematically and are possible descriptions of impulsive astrophysical events. See (Scargle 1981b) for more details and Section 4 for some ideas for selecting the member of the family that best represents the underlying process.

Thus, as already presaged by item 5 in the list above in Section 1, ARMA models are a large umbrella under which lie many ways to separate a process into combinations of random and deterministic components. These are equivalent in their exact representation of the data. Each provides equivalent representations of the data; distinctions between them are matters of preference that depend on context and scientific goals.

3. The ARMA Degeneracy Problem

We see that an important problem with ARMA modeling of astronomical time series is that given data can be represented by any one of a large set of models. The member of this model family which conventional algorithms select may be a poor representation of the astrophysical process underlying the data. In this regard, the quality of the model is judged in terms of its random and deterministic components – somewhat paradoxically not through a "goodness-of-fit" measure, say based on the residuals between model and data. We will now see why this is.

3.1. A Fundamental Degeneracy

To set the stage we first consider an ARMA degeneracy that goes even beyond model estimation. The following applies as well to MA models, but let's frame the discussion for autoregressive (AR) models:

$$\sum_i A_{n-i}\ X_i = R_n \tag{2}$$

$$A * X = R \tag{3}$$

where X is the random variable representing the data, A is the set of AR coefficients essentially describing memory aspects of the model, and R is a random noise input. The estimate $\hat{\mathrm{R}}$ may be obtained by replacing A with some estimate $\hat{\mathrm{A}}$, and X with the actual data $\hat{\mathrm{X}}$, in Equation (3):

$$\hat{R} = \hat{A} * \hat{X} \tag{4}$$

The model of the data is found by multiplying this equation by the convolutional inverse of $\hat{A}$:

$$\hat{X} = \hat{A}^{-1} * \hat{R} = \hat{C} * \hat{R} \tag{5}$$

where $\hat{C} = \hat{A}^{-1}$ is the model's estimate of the pulse shape. Note that Equation (5) is none other than the MA equivalent to the AR model (cf. item 5 in the list in Section 1). Computationally it is very convenient to translate between AR and MA representations using FFTs; e.g. $C = IFFT(1/FFT(A))$, where IFFT is the inverse transform, as this method automatically yields convergent forms with no worries about locations of zeros of the Z-transform in the complex plane.

But suppose $\hat{A}$ were chosen arbitrarily, not from any modeling procedure, but simply "out of a hat." Inserting the expression for $\hat{R}$ from Equation (4) into Equation (5) we have

$$\hat{X} = \hat{A}^{-1} * (\hat{A} * \hat{X}) = (\hat{A}^{-1} * \hat{A}) * \hat{X} = \hat{X} \tag{6}$$

By the definition of the convolutional inverse $A^{-1} * A$ is a delta function, so this expression exactly recovers the time series data! The point is that **ARMA models cannot be judged by how well they fit the input time series**, since any such model (at least with this scheme for computing $\hat{R}$) exactly reproduces the data. Caveat: Procedures for estimating R alternative to Equation (4) may circumvent this result.

3.2. The ARMA Modeling Degeneracy

The basic problem is that this non-uniqueness, or degeneracy, holds even if the model is the unique output of a modeling algorithm. For given time series data:

1. Conventional algorithms yield a unique ARMA model.
2. There is a family of ARMA models that have the same second-order statistics as this one.
3. The astrophysical process may be presumed to correspond to one member of this family.
4. Model (1), in spite of exactly fitting the data, may be very different from the desired one (3).

This is the situation made explicit in the Extended Wold theorem in Section 2.2. To get a feeling for size of the family in item 2 note that MA(p), a moving average of order p, has an autocorrelation of length p. The size of the family referenced in (2) is then 2^p – quite large in many realistic cases.

In summary, because of unnecessary restrictions **conventional algorithms yield MA models that may not faithfully represent the process underlying the time series data. This misrepresentation ranges from slight to gross.** This situation is largely due to misconceptions about the nature of physical innovations and pulses: that the former need only be uncorrelated and that the latter must be stable, causal and minimum delay. Imposing independence on the innovation and relaxing conditions on the pulses are key to achieving unique astrophysically faithful ARMA-type models, as formalized in the Extended Wold Decomposition discussed in Section 2.2. Use of modeling algorithms without regard to these limitations pervades the literature largely because most algorithms are based on the autocorrelation function, which does not contain the information needed to resolve the degeneracy. (NB: the coefficients of the Yule-Walker equations often used to estimate AR models, are precisely estimates of the autocorrelation at various lags.)

3.3. Example: A Tale of two Autoregressive Processes

Consider two toy examples. Shown in Figure 1 is a synthetic time series from a causal AR of order 1 with coefficient $a = 0.96$, plus some additive Gaussian noise. This case is exactly the same as MA(∞) with a causal pulse shape of the fast-rise-exponential-decay type that gamma-ray burst people call FREDs. Here *causal* refers to the fact that the pulse is viewed as the response to an impulse occurring at a fixed moment, with no response before that time. This pulse rises instantaneously to its maximum (here 1) and then decays by the factor a at each successive time step.

The first panel of Figure 1 is a plot of the data, with five such pulses at random times with little overlap. These data were fit in two ways: (a) with the MatLab ARMA routine constrained to yield an AR(1) model, and (b) with the L_1 optimization procedure from Scargle (1977, 1981b). The latter simply replaces least-squares (i.e. L_2) with least-absolute-value (i.e. L_1) optimization, which for this simple case is a straightforward minimization of a function of a single variable. The pulse shape in these data is invertible, causal and minimum-delay, so these two fitting procedures should yield approximately the same models. They do indeed. Panel 2 shows the *innovation* and panel 3 the corresponding pulse shapes. Both give nearly the same pulse shape and innovation, although the L_1 pulse is somewhat more accurate. All is well here.

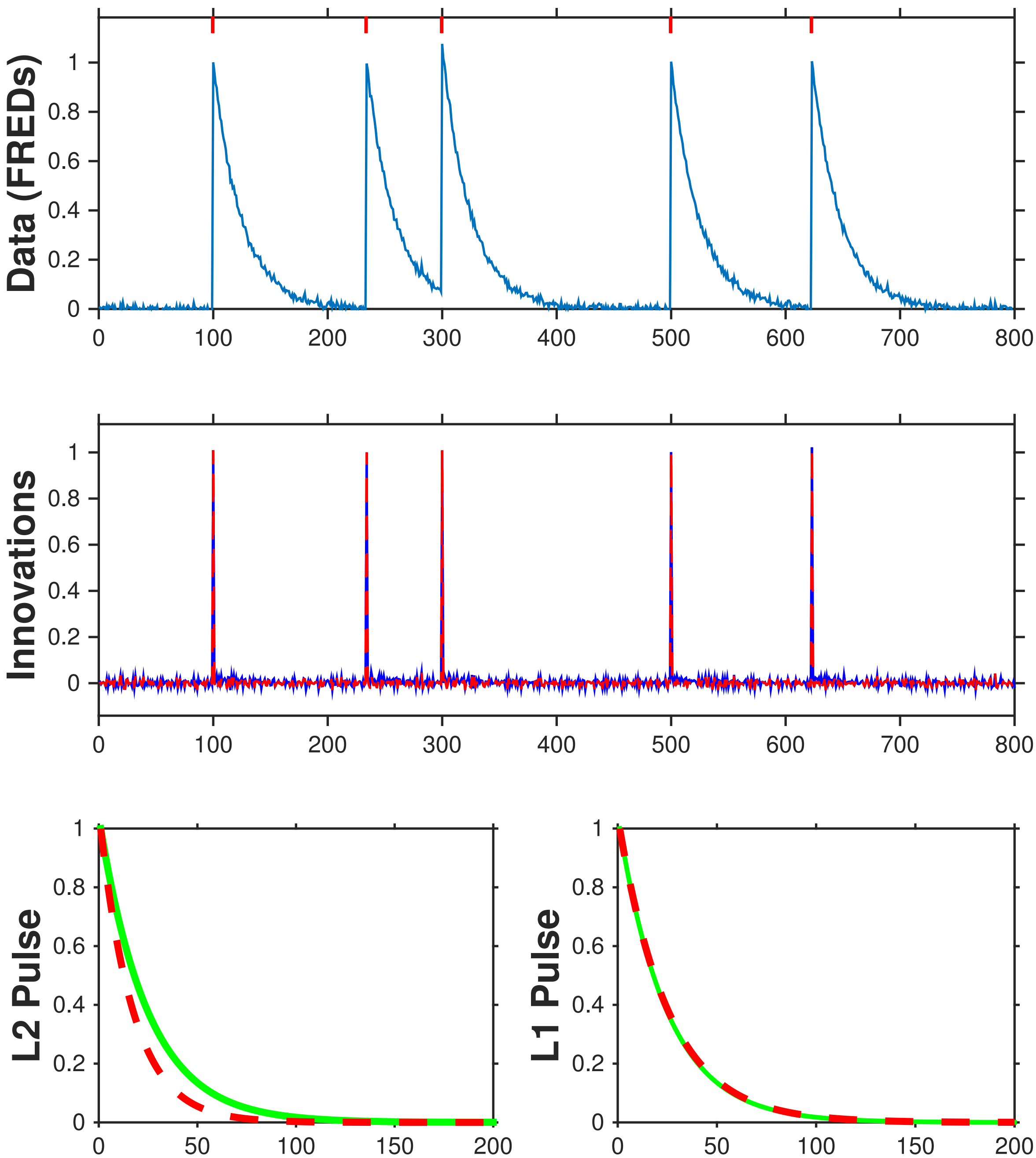

Fig. 1.— Panel 1: data from a causal invertible model AR(1) with coefficient $a = 0.96$ and additive noise. Panel 2: innovations from conventional model (blue) and L_1 procedure (red). Despite much overlap the blue line can be seen behind dashes in the red one. Panel 3: Estimated pulse shapes (conventional left; L_1 right) superimposed on the true pulse shape (green).

The second example is also first order AR with the same coefficient but now acausal[2]. This term means the body of the pulse occurs before the peak associated with the origin of time for the pulse, a relation that's in no way unphysical. The pulse grows by a factor $1/a$ at each time step, dropping instantaneously to zero after its peak. The innovations in panel 2 are significantly different, but perhaps more serious is the complete misidentification of the pulse shape (panel 3) by the conventional model – **it finds the time reverse of the true shape!** All is not well here.

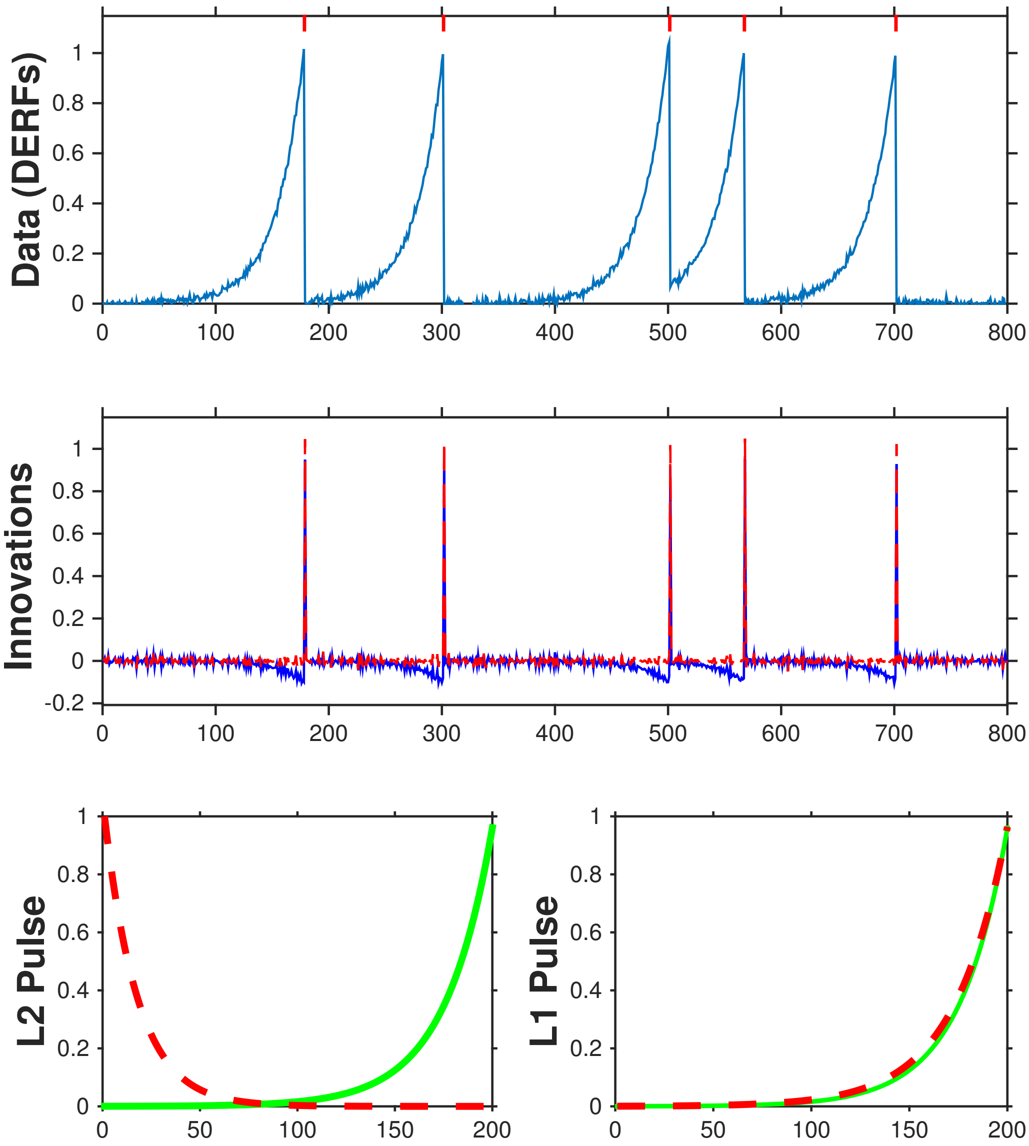

Fig. 2.— As in Figure 1 for the same model but backwards in time.

[2]I like the notation AR(p,q) where p is the causal order and q is the acausal order, so AR(1,0) and AR(0,1) are the two examples here. This exponential-rise-fast-decay pulse, the reverse of a FRED. is whimsically called a DERF.

4. What is the Solution? Time-asymmetry Estimation Techniques

It should now be clear that the twin issues – (a) first estimating ARMA models from data, and then (b) judging their quality – are intertwined in ways far different from what happens in standard model fitting. In particular, metrics (data vs. model) are useless and must be replaced by evaluation of the separate factors R and C in terms of astrophysical appropriateness.

If a key astrophysical desideratum is independence (IID) of R, then minimization of some measure of its independence could be the crux of a modeling procedure (Scargle 1977, 1981a,b, 1990) as well as defining a measure of ... let's call it "goodness-of-representation" in lieu of the usual "goodness-of-fit." Invoking independence is not an arbitrary choice, but dovetails with the assumption that impulsive events making up the lightcurve arise from regions of the astronomical source which do not communicate with each other – thus generating events that are probabilistically independent. Alternatively the astronomer might want to impose the condition of causality and/or minimum delay on C – thus driving quite different estimation and evaluation procedures. Other possible criteria are model simplicity (parsimony) and uncertainty.

A number of techniques capable of resolving ARMA degeneracy and detecting time-asymmetry, thereby estimating physically realistic models, have been suggested. Additions to references given already include Olsen and Ruzinsky (1989); Breidt and Davis (1992); Nemiroff, Norris et al. (1994); Breidt, Davis and Trindade (2000); Andrews, Davis and Breidt (2006); Bauer, Scholköpf and Peters (2016). Other approaches, possibly relevant here include (Scaringi, Maccarone and Middleton 2014; Maccarone 2013) using bi-spectrum techniques, (Press and Rybicki 1997) using 3-point statistics, and (Kawaguchi, Mineshige, Umemura and Turner 1998) using structure functions computed separately for rising and falling parts of the light curve.

At the SAMSI meeting Javier Pascual-Granado, Rafael Garrido, and Lourdes Verdes-Montenegro introduced the *connectivity* statistic that may well be useful in this regard. It measures the difference between forward and backward linear predictions – the former based on the data just before the prediction point and the latter just after. Hence it is sensitive to the distinction between causal (good forward predictions) and acausal (good backward predictions). Some preliminary tests of this concept are quite encouraging.

It is not clear if any of these approaches can yield practical, efficient, robust algorithms for the problems of astronomical time series analysis (in both little- and big-data settings), including models of high order. Many of the techniques, including the L_1 method demonstrated in Section 3.2, seem to break down relatively rapidly when the noise level, pulse overlap, or other aspects of the complexity of the signal increase. These issues may be good topics for a working group.

REFERENCES

Andrews, B., Davis, R., and Breidt, F. 2006, Maximum likelihood estimation for all-pass time series models, Journal of Multivariate Analysis, 97, 1638-1659.

Bauer, S., Scholköpf , B. and Peters, J. 2016, The Arrow of Time in Multivariate Time Series, in Proceedings of the 33rd International Conference on Machine Learning, New York, NY, USA, 2016. JMLR: W&CP volume 48
`http://jmlr.org/proceedings/papers/v48/bauer16.pdf`

Breidt, F., and Davis, R. 1992, Time-reversibility, identifiability and independence of innovations for stationary time series, Journal of Time Series Analysis, 13, 377-390.

Breidt, F., Davis, R., Trindade, A. 2000, Least Absolute Deviation Estimation for All Pass Time Series Models, `http://www.stat.columbia.edu/~rdavis/papers/lad5.pdf`

Kawaguchi, T., Mineshige, S., Umemura, M. and Turner, E. 1998, Optical Variability in Active Galactic Nuclei: Starbursts of Disk Instabilities?, ApJ504, 671-679
`https://arxiv.org/pdf/astro-ph/9712006v1.pdf`

Maccarone, T. 2014, The biphase explained: understanding the asymmetries in coupled Fourier components of astronomical timeseries, MNRAS435, 1546
`http://arxiv.org/abs/1308.3150`

Nemiroff, R., Norris, J., Kouveliotou C., Fishman, G., Meegan, C. and Paciesas, W. 1994, Gamma-Ray Bursts are Time-Asymmetric, ApJ423, 432-435

Olsen, E. and Ruzinsky, S. 1989, Characterization of the LAD (L,) AR Parameter Estimator When Applied to Stationary ARMA, MA, and Higher Order AR Processes, IEEE Transactions on Acoustics, Speech and Signal Processing, 37, 1451

Press, W. and Rybicki, G. 1997, Desperately Seeking Non-Gaussianity, The Light Curve of 0957+561, in Astronomical Time Series, Proceedings of the Florence and George Wise Observatory Symposium
`http://www.lanl.gov/DLDSTP/SeekNonGauss.pdf`

Scargle, J. 1977, Absolute Value Optimization to Estimate Phase Properties of Stochastic Time Series,' 1977, *IEEE Transactions on Information Theory,* **IT-23**, pp. 140-143.

Scargle, J. 1981, Phase sensitive deconvolution reference to astronomical data, In Applied Time Series Analysis II, 549-564, D. F. Findley, ed., Academic Press New York

Scargle, J. 1981, Studies in Astronomical Time Series Analysis. I. Modeling Random Processes in the Time Domain, *Astrophysical Journal Supplement Series*, **45**, pp. 1-71.

Scargle, J. 1990, Studies in astronomical time series analysis. IV - Modeling chaotic and random processes with linear filters ApJ, 359, 469.

Scaringi, S., Maccarone, T., and Middleton, M. 2014, Reversibility of time series: revealing the hidden messages in X-ray binaries and cataclysmic variables, MNRAS, 445, 1031-1038 https://arxiv.org/abs/1409.1223