Validation of Complex Representations – Is This Worthwhile?

So far we have addressed some inherent properties of complex processes, such as complex nonlinearity (Chapter 4), augmented complex statistics (Chapter 12), and topological properties of complex mappings (Chapter 11). It seems clear that complex valued models will be more advantageous the greater the coupling between the real and imaginary components of a process—that is, the more ‘complex’ the process. This is borne out by empirical evidence (Chapter 12) where it is shown that the relative benefit of complex valued modelling is related to the degree of coupling between the speed and direction components of the wind profile. In addition, signal dynamics and the degree of averaging, which affect the component coupling, will have a major influence on the choice of an appropriate signal model. Again wind data provide an example; the areas denoted by A, B and C in Figure 18.1 correspond respectively to ‘high’, ‘medium’ and ‘low’ dynamics.

It has been shown in Chapter 13 that the use of widely linear model is justified only if there is a statistical evidence that the signal in hand is not second-order circular. However, the pseudocovariance matrix is estimated from the data available, and such estimate will, in general, be nonzero, although the actual source is circular. Since we are mostly interested in ‘complex by convenience of representation’ signals (see Chapter 2), it would appear vital to establish a rigorous statistical testing framework which would reveal whether the complex representation is worthwhile—that is, does it offer theoretical and practical advantages over the bivariate\(^1\) or dual univariate signal models?

Following on from the Delay Vector Variance (DVV) technique for statistical testing for signal nonlinearity [83, 86], one such statistical test for the ‘complex nature’ of real-world processes [85] is the ratio of statistical differences between realisations under the null hypothesis of ‘linear bivariate’ and ‘linear circular.’

\(^1\)For convenience, we use the term ‘bivariate’ to denote the ‘real valued bivariate’ signals.
Signal Modality Characterisation in $\mathbb{R}$

Signal modality characterisation is becoming an increasingly important area of multidisciplinary research and considerable effort has been put into devising efficient algorithms for this purpose. Research in this area started in physics in the mid 1990s [303], but its applications in machine learning and signal processing are only recently becoming apparent [84]. As changes in the signal nature between, say, linear and nonlinear and deterministic and stochastic can reveal information (knowledge) which is critical in certain applications (e.g. health conditions), the accurate characterisation of the nature of signals is a key prerequisite to choosing a signal processing framework.

By the ‘nature’ of a signal we refer to the following fundamental properties: [82, 83, 265]:

- **P1. Linear** (strict definition) – a linear signal is generated by a linear time-invariant system, driven by white Gaussian noise.
- **P2. Linear** (commonly adopted) – property P1 is relaxed somewhat by allowing the amplitude distribution of the signal to deviate from the Gaussian distribution (a linear signal from P1 is measured by a static, possibly nonlinear, observation function).
- **P3. Nonlinear** – a signal that does not meet the criteria P1 or P2 is considered nonlinear.
- **P4. Deterministic** (predictable) – a signal is considered deterministic if it can be precisely described by a set of equations.
- **P5. Stochastic** – a signal that is not deterministic.²

²The Wold decomposition theorem [314] states that any discrete stationary signal can be decomposed into its deterministic and stochastic (random) component, which are uncorrelated. This theorem forms the basis for many prediction models, since the presence of a deterministic component imposes a bound on the performance of these models.
Figure 18.2 A variety of signal modalities spanned by the properties ‘stochastic’ and ‘nonlinear’. Classes of signals for which the generating mechanisms are well understood are ‘Chaos’, ‘ARMA’, and ‘NARMA’.

The range of real world signals spanned by their linear vs nonlinear and deterministic vs stochastic natures is illustrated in Figure 18.2 (modified from [263]). It is interesting that the classes of signals which are well established and understood, such as the linear stochastic autoregressive moving average (ARMA) models and nonlinear deterministic chaotic signals, are at the opposite corners of Figure 18.2. Real world signals, however, are likely to belong to the areas denoted by (a), (b), (c) or ‘?’, since they are recorded in noisy environments and by nonlinear sensors; these are major signal classes about which we know little or nothing.

18.1.1 Surrogate Data Methods

The concept of ‘surrogate data’ was introduced by Theiler et al. [287], and has been extensively used in the context of statistical testing for signal nonlinearity; more detail on surrogate data methods can be found in [144, 265, 288, 290]. Hypothesis testing assesses a fundamental property of signal (say nonlinearity) by generating a large number, say 100, of independent linear realisations of the original signal (surrogates) and comparing the ‘test statistic’ for the surrogates against that of the original signal.

The basic principle of hypothesis based statistical testing for signal nonlinearity can be summarised in the following steps:

\textbf{N1.} Establish a null hypothesis \( H_0 \), e.g. the signal is generated by a linear stochastic system driven by white Gaussian noise.

\textbf{N2.} Generate a number of independent surrogates, which are linear realisations of the original signal.

\footnote{The environment is also typically statistically nonstationary, and the signal modality changes with time, say from (a) to (b), or from ARMA to Chaos (heart rates, epileptic seizures), which may, e.g. indicate a health hazard.}

\footnote{The analysis of the nonlinearity of a signal can often provide insights into the nature of the underlying signal production system. However, care should be taken in interpreting the results, since the assessment of nonlinearity within a signal does not necessarily imply that the underlying signal generation system is nonlinear: the input signal and system (transfer function) nonlinearities are confounded.}
N3. Establish a discriminating criterion between the linear and nonlinear signal, the so-called test statistic.

N4. Based on the number of rejections of the null hypothesis from N1, the nature of the original signal is judged linear or nonlinear.

Since surrogate data are linear realisations of the original signal, they can be generated in many ways, for instance by ARMA modelling. It is, however, much more desirable to have nonparametric generation methods for surrogate data. By definition (P1 and P2), the property of signal linearity is derived from the second-order statistics (mean, variance, autocorrelation or equivalently amplitude spectrum), and hence for a linear signal the phase spectrum and higher-order statistics (HOS) are irrelevant; one simple way to generate a number of surrogates is based on the Fourier transform (FT surrogates), as illustrated in Figure 18.3. The FT surrogates are generated by simply performing the inverse Fourier transform of a signal generated from the original amplitude spectrum and the randomised phase spectrum; this method, however, is not suitable for signals described by P2.

A reliable surrogate data method, capable of generating surrogates for data observed through a static nonlinearity, is the ‘iterative Amplitude Adjusted Fourier Transform’ (iAAFT) method [264]; it has been shown to produce superior results compared with other available surrogate data generation methods [160, 265].

The iAAFT method can be summarised as follows:

S1. Let \(|S|\) be the Fourier amplitude spectrum of the original time series \(s\), and \(|c|\) the amplitude sorted version of the original time series

Repeat:

S2. At every iteration \(j\) generate two additional series:

(i) \(r^{(j)}\), which has the same distribution as the original signal
(ii) \(s^{(j)}\), which has the same amplitude spectrum as the original signal

Starting with \(r^{(0)}\), a random permutation of the time samples of the original time series:

1. Compute the phase spectrum of \(r^{(j-1)} \rightarrow \{\phi\}\)
2. Compute $s^{(j)}$ as the inverse transform of $\{|S_k| \exp(j \phi_k)\}$
3. Compute $r^{(j)}$ as obtained by rank-ordering $s^{(j)}$ so as to match $\{c_k\}$

Until error convergence.

18.1.2 Test Statistics: The DVV Method

A convenient test statistic, which makes use of some notions from nonlinear dynamics and chaos theory (embedding dimension and phase space) is the Delay Vector Variance (DVV) method [83]. It is based upon examining the local predictability of a signal in phase space, which when combined with the surrogate data methodology allows one to examine simultaneously the determinism and nonlinearity within a signal.

The signal flow within the DVV method is illustrated in Figure 18.4. For a given embedding dimension $m$, the DVV method can be summarised as follows [83, 84, 86]:

**D1.** The mean, $\mu_d$, and standard deviation, $\sigma_d$, are computed over all pairwise Euclidean distances between delay vectors (DVs), $\|x(i) - x(j)\|$ $(i \neq j)$.

**D2.** The sets of ‘neighbouring’ delay vectors $\Omega_k(r_d)$ are generated such that $\Omega_k(r_d) = \{x(i) | \|x(k) - x(i)\| \leq r_d\}$, that is, sets which consist of all DVs that lie closer to $x(k)$ than a certain distance $r_d$, taken from the interval $[\max\{0, \mu_d - n_d \sigma_d\}; \mu_d + n_d \sigma_d]$.

5Apart from the surrogate methods, other established methods for detecting the nonlinear nature of a signal include the Deterministic versus Stochastic (DVS) plot [44] and $\delta$–$\epsilon$ Method [145].
for example, \( N_v \) uniformly spaced distances, where \( n_d \) is a parameter controlling the span over which to perform the DVV analysis.

**D3.** For every set \( \Omega_k(r_d) \), the variance of the corresponding targets, \( \sigma^2_k(r_d) \), is computed. The average over all sets \( \Omega_k(r_d) \), normalised by the variance of the time series, \( \sigma_x^2 \), yields the target variance \( \sigma^{*2}(r_d) \):

\[
\sigma^{*2}(r_d) = \frac{1}{N} \sum_{k=1}^{N} \frac{\sigma^2_k(r_d)}{\sigma_x^2} \quad (18.1)
\]

As a rule of thumb, we only consider a variance measurement valid, if the set \( \Omega_k(r_d) \) contains around \( N_v = 30 \) DVs, since having too few points for computing a sample variance yields unreliable estimates of the true (population) variance.

As a result of the standardisation of the distance axis, the resulting ‘DVV plots’ (target variance, \( \sigma^{*2}(r_d) \) as a function of the standardised\(^6\) distance, \( (r_d - \mu_d)/\sigma_d \)) are straightforward to interpret:

- The presence of a strong deterministic component will lead to small target variances \( \sigma^{*2}(r_d) \) for small spans \( r_d \).
- The minimal target variance, \( \sigma^{*2}_{\text{min}} = \min_{r_d}[\sigma^{*2}(r_d)] \), is a measure for the amount of noise which is present in the time series (the prevalence of the stochastic component).
- At the extreme right, the DVV plots smoothly converge to unity, since for maximum spans, all DVs belong to the same universal set, and the variance of the targets is equal to the variance of the time series.

To illustrate the operation of the DVV method, consider a linear AR(4) signal [190], given by

\[
x(k) = 1.79 x(k - 1) - 1.85 x(k - 2) + 1.27 x(k - 3) - 0.41 x(k - 4) + n(k) \quad (18.2)
\]

and a benchmark nonlinear signal [216], given by

\[
z(k) = \frac{z(k - 1)}{1 + z^2(k - 1)} + x^3(k) \quad (18.3)
\]

where \( x(k) \) denotes the AR(4) signal defined above and \( n(k) \sim N(0,1) \).

Averaged DVV plots, computed over 25 iAAFT-based surrogates for these two benchmark signals are shown respectively in Figure 18.5(a) and (b). Since the AR(4) is linear, and surrogates are also linear by design, the DVV curves for the original and averaged surrogates are very close (Figure 18.5a). This is not the case with the DVV curves for the original and averaged surrogates for the nonlinear signal in Figure 18.5(b), which are far apart, indicating that the linear surrogates were not able to model the nonlinear nature of the signal.

Due to the standardisation of the distance axis, these plots can be conveniently combined in a scatter diagram, where the horizontal axis corresponds to the DVV plot of the original time

\(^6\)Note that we use the term ‘standardised’ in the statistical sense, namely as having zero mean and unit variance.
Testing for the Validity of Complex Representation

When testing for the suitability of the complex valued signal representation, one convenient null hypothesis $H_0$ would be that the time series is generated by a linear circular complex valued process, followed by a (possibly nonlinear) static observation function, $h(\cdot)$ which operates on the moduli of the complex valued time samples.

Figure 18.5 Signal modality characterisation. The DVV plots for the original and surrogates are obtained by plotting the target variance as a function of standardised distance.

Figure 18.6 The DVV scatter diagrams obtained by plotting the target variance of the original data against the mean of the target variances of the surrogate data.

series, and the vertical to that of the surrogate time series. If the averaged surrogate time series yield DVV plots similar to that of the original time series, the DVV scatter diagram coincides with the bisector line, and the original time series is judged to be linear, as illustrated in Figure 18.6(a). Conversely, as shown in Figure 18.6(b), the DVV scatter diagram for a nonlinear signal deviates from the bisector line.

18.2 Testing for the Validity of Complex Representation

When testing for the suitability of the complex valued signal representation, one convenient null hypothesis $H_0$ would be that the time series is generated by a linear circular complex valued process, followed by a (possibly nonlinear) static observation function, $h(\cdot)$ which operates on the moduli of the complex valued time samples.
To cater for complex valued signals, a straightforward extension of the (real valued) bivariate iAAFT-method [265] would be to match the amplitude spectrum of the surrogate and the amplitude spectrum of the original complex valued signal. Then, the signal distribution of the original needs to be imposed on the surrogates in the time domain (step $S3$ in the iAAFT-procedure). This can be achieved by rank-ordering the real and imaginary parts of the complex valued signal separately, as this is in line with the notion of circularity. However, in practice, for complex valued signals it is more important to impose equal empirical distributions on the moduli of the samples, rather than on the real and imaginary parts separately. This way, the empirical distribution of the moduli is (approximately) identical to that of the original signal, thus retaining signal circularity (if present). This approach will be adopted in the derivation of the statistical test for the validity of complex valued representations.

Similar to the real valued iAAFT case, during the computation of complex iAAFT (CiAAFT) surrogates, $|S_k|$ denotes the Fourier amplitude spectrum of the original time series $s$; for every iteration $j$, three additional time series are generated:

- the ‘modulus sorted’ version of the complex valued original time series $s$, denoted by $c_k$;
- a time series with the same distribution (in terms of moduli) as the original complex valued time series $s$, denoted by $r^{(j)}$;
- a time series which has the amplitude spectrum identical to that of the original time series $s$, but not necessarily the same distribution of the moduli, denoted by $s^{(j)}$.

The CiAAFT procedure is summarised below; the iteration starts with $r^{(0)}$, a random permutation of the original complex valued time samples.

**Repeat:**

**C1.** Compute the phase spectrum of $r^{(j-1)} \rightarrow \{\phi_k\}$

**C2.** Compute $s^{(j)}$ as the inverse transform of $|S_k| \exp(j\phi_k)$

**C3.** Rank-order the real and imaginary parts of $r^{(j)}$ to match the real and imaginary parts of $\{c_k\}$

**C4.** Rank-order the moduli of $r^{(j)}$ to match the corresponding modulus distribution of $\{c_k\}$

**Until** error convergence

Figure 18.7 illustrates the results of the statistical testing for the complex nature of (fully complex by design) Ikeda map. Figure 18.7(b) shows that the bivariate approach (although two-dimensional) was not able to preserve the state space properties of Ikeda map, whereas the surrogate realisation based on CiAAFT (Figure 18.7c) was much better suited for this purpose.

### 18.2.1 Complex Delay Vector Variance Method (CDVV)

The extension of the Delay Vector Variance method into the complex domain is straightforward. For a given embedding dimension $m$ and a resulting time delay embedding representation (i.e., a set of complex valued delay vectors $(DV)$, $s(k) = [s_k-m, \ldots, s_k-1]^T$), a measure of unpredictability, $\sigma^2(r_d)$, is computed for a standardised range of degrees of locality, $r_d$, similar to the real case described in Section 18.1.2. Since for both the bivariate and complex time
Figure 18.7 Judging the validity of complex representation of the Ikeda map time series denoted by $z = \Re(z) + j \Im(z)$: (a) the original signal; (b) realisation using bivariate iAAFT surrogates; (c) realisation using complex iAAFT surrogates; (d) DVV plots for the Ikeda map (thick solid line), iAAFT surrogate (think dashed line), and CiAAFT surrogate (thin solid line)

series, a delay vector is generated by concatenating time delay embedded versions of the two dimensions (real and imaginary), the complex valued and real valued bivariate versions of the DVV method provide equivalent results, and the variance of such variables is computed as the sum of the variances of each variate, that is\(^7\)

$$\sigma^2_s(r_d) = \sigma^2_{s,r}(r_d) + \sigma^2_{s,i}(r_d)$$

where $\sigma^2_{s,r}(r_d)$ denotes the target variance for the real part of the original signal $s$, and $\sigma^2_{s,i}(r_d)$ denotes that for the imaginary part.

\(^7\)Note that by augmenting the delay vectors within the DVV method, that is, by computing the Euclidean distances between the augmented delay vectors, would not make any difference. This is due to the deterministic relationship between a complex delay vector $\mathbf{z}(k)$ and its complex conjugate $\mathbf{z}^*(k)$ (see Chapter 12).
Validation of Complex Representations – Is This Worthwhile?

Test for the validity of complex valued representation. To test for the potential benefits of complex valued representations, rather than comparing the original time series to the surrogates, it is convenient to compare the surrogates generated under the fundamentally different null hypotheses of:

- a linear bivariate time series, denoted by $H_0^b$, for which the surrogates are generated using the (real valued) bivariate iAAFT [265];
- a linear and complex valued time series, denoted by $H_0^c$, for which the surrogates are generated using the CiAAFT method [85].

The DVV method can be used to characterise the natures of the two different realisations. A statistically different characterisation means that the two null hypotheses lead to different realisations of the original signal. Since the difference between the null hypotheses is the property of circularity in the linearisations of the original signal, a statistical difference can be interpreted as an indication of the presence of circularity in the original signal, which is retained in the CiAAFT realisations, and not in the iAAFT realisations. This test can be used to justify a complex valued representation over a dual univariate one.

The complex DVV based statistical test for the validity of the complex valued representation can now be summarised as follows:

**T1.** Generate $M$ CiAAFT surrogates and produce the averaged DVV plot denoted by $D^0$;

**T2.** Generate $N$ bivariate iAAFT surrogates and produce the corresponding DVV plots, denoted by $\{D^b\}$;

**T3.** Generate $N$ CiAAFT surrogates and produce the corresponding DVV plots, denoted by $\{D^c\}$;

**T4.** Compare $(D^0 - \{D^b\})$ and $(D^0 - \{D^c\})$.

To perform T4 in a statistical manner, the (cumulative) empirical distributions of root-mean-square distances between $\{D^b\}$ and $D^0$, and between $\{D^c\}$ and $D^0$, are compared using a Kolmogorov–Smirnoff (K-S) test. This way, the different types of linearisations (bivariate $\{D^b\}$ from T2, and complex valued $\{D^c\}$ from T3) are compared with the ‘reference’ linearisation, that is, $D^0$ from T1. If the two distributions of test statistics are significantly different at a certain confidence level $\alpha$, say 95%, the original time series is judged complex valued [85].

A DVV plot of a complex signal is obtained by plotting the target variance, $\sigma^*^2(\tau_d)$, as a function of the standardised distance $(\tau_d - \mu_d)/\sigma_d$. The DVV plots for a 1000 sample realisation of the Ikeda map and the iAAFT and CiAAFT surrogates (using $m = 3$ and $n_d = 3$) are shown in Figure 18.7(d). Figure 18.8 illustrates the convergence of CiAAFT surrogates when modelling the Ikeda map, and their behaviour in the presence of noise. The convergence curve is uniform and the CiAAFT iteration settles after about 30 iterations, as shown in Figure 18.8(a). Since the complex white Gaussian noise used in simulations was ‘doubly white’ (see Section 13.2), bivariate iAAFT surrogates provided a suitable model. Figure 18.8(b) shows the rejection ratios of the CDVV method for the Ikeda map contaminated with complex WGN (CWGN) over a range of power levels. The complex Ikeda map and CWGN had equal variances, and the noisy signal $\text{Ikeda}_{\text{noisy}}$ was generated according to

\[ \text{Ikeda}_{\text{noisy}} = \text{Ikeda}_{\text{original}} + \gamma_n \times \text{CWGN} \quad (18.4) \]
18.3 Quantifying Benefits of Complex Valued Representation

To test for the benefits of complex valued representation of real valued processes, consider wind data recorded over 24 hours, shown in Figure 18.1. Areas denoted by A, B, and C correspond respectively to the ‘high’, ‘medium’ and ‘low’ dynamics of wind. It is expected that the larger the changes in wind dynamics (‘high’) the greater the advantage obtained by the complex valued representation of wind (see also Chapter 12). For relatively mild and slowly changing wind dynamics, it is expected that the complex valued modelling should not exhibit significant performance advantage over, say, the dual univariate one. Also, it is expected that the complex (and nonlinear) nature of wind would be less pronounced with increased averaging of the raw data.\(^8\)

Table 18.1 illustrates the rejection ratios\(^9\) for the null hypothesis of ‘no difference between bivariate and complex linearisations’, that is, for the bivariate nature of wind data from Figure 18.1, for different wind regimes and degrees of averaging.\(^10\) Also, there are stronger indications of a complex valued nature when the wind is averaged over shorter intervals, as represented

---

\(^8\)This is closely related to circularity and augmented complex statistics addressed in Chapter 12.

\(^9\)Each result was obtained by performing the complex DVV test 100 times, and by counting the number of rejections of the null hypothesis.

\(^10\)As expected (see Figure 2.6), there is a significant component dependence within the complex valued wind signal representation, as indicated by the rejection ratio of the null hypothesis (of a bivariate nature) being significantly greater than zero.
Table 18.1  Rejection rates (the higher the rejection rate the greater the benefit of complex valued representation) for the wind signal from Figure 18.1

<table>
<thead>
<tr>
<th>Wind signal</th>
<th>Region A%</th>
<th>Region B%</th>
<th>Region C%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Averaged over 1 s</td>
<td>96</td>
<td>80</td>
<td>71</td>
</tr>
<tr>
<td>Averaged over 10 s</td>
<td>90</td>
<td>74</td>
<td>62</td>
</tr>
<tr>
<td>Averaged over 60 s</td>
<td>83</td>
<td>69</td>
<td>58</td>
</tr>
</tbody>
</table>

by the respective percentage values of the ratio of the rejection of the null hypothesis of a real bivariate nature.

18.3.1 Pros and Cons of the Complex DVV Method

This chapter illustrates the importance and usefulness of statistical testing for the fundamental nature of data (nonlinearity, determinism, complex valued nature). This can provide additional knowledge which can be exploited when choosing a signal processing model best suited to the data. Indeed, following from the signal modality characterisation for real valued data from Figure 18.2, the complex surrogates and complex DVV method enable us to test whether the complex valued representation of real world data is worthwhile.

Despite being well founded mathematically and extremely useful, these tests suffer from drawbacks, such as:

- Surrogate and DVV tests can only be performed in an off-line block manner;
- The DVV and surrogate data based tests are only applicable for quasistationary data segments (since we need to calculate the embedding parameters and compute the Fourier transform, these operation only apply to stationary data);
- Due to the requirement of piece-wise stationarity, it is possible to mistake the property of nonstationarity for nonlinearity;
- Hypothesis testing methods are not readily suitable for real time mode of operation, they can, however, operate in near real time on overlapping data windows.

Some of these issues have been addressed in Chapter 16, where online tests for the characterisation of the nature of complex valued data are introduced.