Model Misspecification Tests, Model Building and Predictability in Complex Systems

Radu Manuca
Robert Savit

SFI WORKING PAPER: 1995-09-075

SANTA FE INSTITUTE

Model Misspecification Tests, Model Building and Predictability in Complex Systems.

Radu Manuca and Robert Savit
Physics Department
and
Program for the Study of Complex Systems
University of Michigan
Ann Arbor, MI 48109

ABSTRACT

In this paper we show that significant advantages can be realized by using functions of simple variables, such as time lags, as directions in reconstruction spaces for complicated time series. First, we show that more sensitive model misspecification tests can be constructed when the candidate model is used as one of the directions in the reconstruction space. This naturally results in misspecification tests based on the construction of conditional probabilities which can reveal misspecified models even when other powerful methods, such as the BDS test applied to a sequence of residuals, fails. Second, we show that model building and predictability of the time series can be substantially improved by using an informational criterion to determine the functions with which to associate the directions of the reconstruction space. This criterion is of the form of a conditional probability and is related to a measure of the short term predictability of the time series. We consider specifically an example in which the search space of functions is linear combinations of time-lags, and we compare the resulting model with a linear least squares fit to the data. We also show that this optimized reconstruction space can improve longer-term predictability. We demonstrate this by comparing longer-term predictions made using the optimized space with predictions made using a simple time-lagged reconstruction space.

I. Introduction

A number of methods for the analysis and description of nonlinear processes are based on the study of a data set produced by the system in a reconstruction space. Generally, one studies data sets in a reconstuctions space whose directions are simple lags (if the series is descrete) or derivatives (if the series is continuous), or some other equally simple quantity. Although there has been some work in choosing the best reconstruction directions for different purposes[1], the common practice is to rely on relatively simple quantities for the directions. This attitude is quite understandable, and follows, we believe, in no small part form the influence of the powerful results of Takens and Ruelle[2], which demonstrate that many of the important topological properties of an attractor, are independent of the details of the reconstruction space.

However, when studying real data sets which may be generated by noisy complex processes, one if often interested in aspects of the system other than the topological properties. The problem of extracting information from such data sets, and developing statistically sound conclusions may also be complicated if the system is strongly data-limited. In such cases, it may be advantageous to consider reconstruction spaces whose directions are other than simple variables such as lags or derivatives.

In this paper we describe the benefits of allowing the directions of the reconstruction space to be more complicated functions of the variables in the problem. In any reconstruction space, it is natural to consider probabilities for continuing closeness of sets of variables. Such probabililties have been used by Brock, Dechert and Scheinkman[3] to study residuals of time series for randomness, and by Savit and Green[4] to look for dependencies among sets of variables in a data set. Here we shall consider similar conditional probabilities, but in which some of the arguments are functions rather than simple variables. In particular, we shall describe the application of this technique to two common problems faced in the analysis of complex systems, namely, model misspecification testing and model building. In the case of model misspecification problems, we will show that very powerful discriminants of misspecification can be constructed by allowing one of the directions in the reconstruction space to be the candidate model itself. Conditional probabilities, related to combinations of Grassberger-Procaccia correlation integrals can be constructed which provide very strong and specific tests against model misspecification. In the case of model building, we show that the reconstruction space can be optimized according to an informational measure which is directly related to the predictability of the series. To

demonstrate the efficacy of this reconstruction we shall study a discretized time series deduced from the Lorenz system and we shall show that when overlaid by a simple prediction algorithm, predictions made with this optimized reconstruction are significantly better than those made using a construction based on simple lags.

The rest of this paper is organized as follows: In the next section, we construct the conditional probabilistic indicators we shall use, and describe their meaning, in general. In Section III, we study the problem of model misspecification, by analyzing a time series generated by summing four independent Henon maps. This series was studied previously by Theiler and Eubank[5], who showed that tests of the series of residals from a linear AR fit failed to reject the null hypothesis of IID for the residuals. We show that our appoach is able to reject the linear AR fit as a good explanation of the data. In Section IV we turn to the problem of model building. Here we show that an improved reconstruction space (which we call optimized) can be defined by maximizing a set of conditional probabilities (related to predictability), whose arguments involve functions of the variables in the problem. Using as an example a discretized version of a time series generated by the Lorenz system, to show that predictions made using this optimized reconstruction space are superior to those made using a reconstruction space based on lags of the time series. We also show that, at least in this example, the maximum necessary embedding dimension for the series is reduced by one by using the otpmized variables, rather than simple lags. This is a significant advantage in data-limited situations. Section V consists of a a summary and discussion. Some technical and subsidiary issues are discussed in two appendices. In particular, in Appendix B we show that in a complete search space our method leads to the correct description of the data set, (as do a number of other approaches).

II. Conditional Probabilities and Dynamical Models

In this section we shall introduce the quantities that we shall use to delimit a variable space and to test specific models. For simplicity, we shall first discuss the case of a single time series, and an *a priori* delimited search space of the lags of that time series. Extension of the method to more general cases will be obvious.

Consider a time series x(t). Suppose for the moment that t is discrete. Let σ_x be the standard deviation of the values of the time series x(t), and let μ be real number (typically, $0.05 \leq \mu \leq 2.0$) Define $\varepsilon_x = \mu\sigma_x$. Now consider the d+1-dimensional vector

$$v(i) = (x(i),x(i-1), \ldots, x(i-d)) = (v_0(i),v_1(i), \ldots ,v_d(i)). \qquad (1)$$

Let t_k stand for the statement $|v_k(i)-v_k(j)| \leq \varepsilon_x$.

Suppose we have a dynamical model, f, either deterministic or with additive noise, for the generation of the time series, such that f(i) estimates x(i). We assume that f depends on a set of variables, say lags of the time series, and is an autonomous function in the search space. In particular, we do not intend f to be a simple time-dependent fit to the data. So, for example, if x(t) were sinusoidal, we would seek to test f's which were functions of the first and second lags, x(t-1) and x(t-2), and *not* f's of the form $\sin\omega t$. Let F stand for the statement $|f(i)-f(j)| \leq \varepsilon_f$, where $\varepsilon_f = \mu\sigma_f$, and σ_f is the standard deviation of the values of f(i). We emphasize that f(i) does not necessarily represent the true dynamics of the system, but is in general only a model for those dynamics.

An example of the kind of quantities we shall be concerned with is the conditional probability of the form $P(t_0 \mid F,t_k)$, where $k \leq d$. In the following paragraphs, we shall use this example to introduce the ideas behind the methods we are presenting. Generalizations will be discussed below. In words, this expresses the probability that two values of x are within ε_x if the values of a model function for those x's are within ε_f and if the kth lags are also within ε_x. The *cognoscenti* will recognize in this quantity an underlying picture, in which the usual phase space embedding philosophy for the analysis of complex systems is modified to include embedding spaces in which one of the directions is the model itself, rather than a simpler variable such as a time lag. Depending on the choice of f, many of the powerful geometric and topological theorems of strange attractors may not apply in this space. That, however, is a secondary consideration for us here.

Suppose now that we want to test how well a dynamical model, f, captures the information in the data x(t). We shall present two kinds of measures. The first are global, and are direct measures of the additional explanatory power in f, compared with the absence of a model. The second are more specific measures that indicate additional dependence of the data on a subset of variables in the search space. This second class of measures can be used generically in two ways: First, one may check for additional dependence of the data on a variable that already appears in the model function, f. Used in this way, these measures are tools for helping to tune the parameters of the model. In addition one may check for dependencies on variables that do not appear in f. This application can tell us

something about the relevant search space of variables for our problem. This second application raises a number of interesting problems which we shall relegate to a future work.

A. Global Measures

Consider the quantity

$$\Phi = \frac{P(t_0 \mid F) - P(t_0)}{P(t_0 \mid F)} = 1 - \frac{P(t_0)}{P(t_0 \mid F)} \qquad (2)$$

If the model function f has no explanatory power for the data x, then the conditional probability $P(t_0|F)$ should equal the unconditional probability $P(t_0)$, and so in an appropriate statistical sense, $\Phi = 0$. The extent to which Φ deviates from zero is a measure of the extent to which f is a good explanation for the data. We shall return below to a discussion of the statistics of (2) and a statement of the null hypothesis.

Another quantity which also provides a global characterization of the function f, is

$$\tilde{\Phi} = \frac{P(t_0 \mid F, t_1, t_2, \ldots, t_d) - P(t_0|F)}{P(t_0 \mid F, t_1, t_2, \ldots, t_d)} = 1 - \frac{P(t_0|F)}{P(t_0 \mid F, t_1, t_2, \ldots, t_d)}. \qquad (3)$$

In $P(t_0 \mid F, t_1, t_2, \ldots, t_d)$, we have included all the variables upon which x(i) may depend. Although (2) and (3) are both global measures, they have different interpretations and are in some sense complementary. Φ is a measure of the amount of information contained in f, whereas $\tilde{\Phi}$ is a measure of the amount of information from the search space *not* contained in f. (2) measures the extent to which values of f that are close are associated with values of x that are close. And roughly speaking, (3) is zero if the dependence on the variables in the search space (in our case the first through d^{th} lags) is correctly included in f. Deviations of (3) from zero indicate that the inclusion of other variables in f, or a change in the parametric form of f involving those variables would improve the ability of f to explain the data.

B. Local Measures

As stated above, we shall limit ourselves to measures of dependence on variables that are already contained in the function f. With this in mind, we will consider two objects. The first is

$$\xi_k = \frac{P(t_0 \,|F,t_k) - P(t_0 \,|F)}{P(t_0 \,|F,t_k)} = 1 - \frac{P(t_0 \,|F)}{P(t_0 \,|F,t_k)} \qquad (4)$$

If f contains an accurate parameterization of the dependence on the k-th lag, then we expect ξ_k to be zero in an appropriate statistical sense (see below). In fact, a more precise way to describe the information we can deduce from a calculation of (4) is to note that: 1.) If f is the correct explanation of the data, then surely, ξ_k will be statistically zero, and 2.) if (4) fails to be zero, then we expect that it will have specific power against a misspecification of the dependence on the kth lag in the function f. Equation (4) can also be used to study dependence on lags which are not explicitly contained in f.

An apparently closely related statistic to (4) is the function

$$\rho_k = \frac{P(R_0 \,|t_k) - P(R_0)}{P(R_0 \,|t_k)} = 1 - \frac{P(R_0)}{P(R_0 \,|t_k)} \qquad (5)$$

In (5), R_0 stands for the statement $|r(i)-r(j)| \leq \varepsilon_r$, where $\varepsilon_r = \mu\sigma_r$, and σ_r is the standard deviation of the values of the residuals of the fit: $r(i) = x(i) - f(i)$.

This quantity directly probes the additional dependence of the residuals on the k^{th} lag of the data. Although (4) and (5) are closely related, they have rather different numerical and statistical properties. These different behaviors have important implications for the process of searching for an optimal description of the process. Here we digress briefly to compare the typical behavior of ξ_k and ρ_k in a chaotic system with and without noise. The reader uninterested in these details may skip to the beginning of section III.

Let us suppose we have a model for x(i) which satisfies $x(i) = f(i) + r(i)$ with $\langle r \rangle \sim 0$ so that we may take σ_r as a measure of the typical magnitude of r(i). As an example, consider a set of data generated by the Henon map in the chaotic regime,

$$x(i) = 1 - 1.4x(i-1)^2 + 0.3x(i-2). \qquad (6)$$

Let us calculate ξ_2 and ρ_2 using the Henon map time series with 8,192 points with $\mu = 0.05$ and a function

$$f(i) = x(i-1)^2 - ax(i-2). \qquad (7)$$

(As we shall explain in section IV, the additive constant and an overall multiplicative factor in the model f are irrelevant for our results here.) The only important factor is the ratio of the coefficient of the $x(i-1)^2$ and the x(i-2) terms. This is the value, a. In Fig. 1 we have plotted ξ_2 and ρ_2 as a function of a. Both ξ_2 and ρ_2 are essentially zero at the correct value of $a = a' \approx 0.22$, but their behavior away from a' is quite different. To a first approximation, ρ_2 is nearly constant away from a', while ξ_2 smoothly drops to zero. Notice also that for $0.18 < a < 0.26$, ξ_2 is essentially linear. This range of a corresponds to values of σ_r which are less than ε. Because a is the coefficient of a linear term in f, the linear dependence on a is roughly equivalent to a linear dependence on σ_r. In fact, it is not difficult to show in general (see Appendix A) that if $\sigma_r \sim O(\varepsilon)$, then $\xi_k \sim O(\sigma_r^1)$ and $\rho_k \sim O(\sigma_r^0)$.

Depending on the nature and goal of the search algorithm, ξ_2 or ρ_2 may be preferable as an indicator of additional dependence. For example, in a much larger search space, one might try to use genetic algorithms to find a best model for the data. In that case, ξ_2 might be preferable to ρ_2 as a fitness function, since the sharp dip in ρ_2 could be easily missed. The same comments, of course, apply to other search algorithms in which one is looking coarsely at the space. At the last stages of a search algorithm, where great sensitivity to parameter choices is required, ρ_2 may be the preferred indicator.

It is also useful to see how the sensitivity of our indicators is affected by the presence of noise in the data set. To this end we compute Φ, ξ_2, and ρ_2 for a data set generated by the Henon map with additive noise:

$$x(i) = 1 - 1.4\, x^2(i-1) + 0.3\, x(i-2) + \sigma\eta(i), \qquad (8)$$

where $\eta(i)$ is an IID sequence chosen from a flat probability distribution, centered about zero with a standard deviation equal to that of the Henon map without noise. σ is a numerical factor which scales the strength of the noise, thus $\sigma=1$ means a signal to noise ratio of 1. In Figs. 2-4 we plot Φ, ξ_2, and ρ_2 for various strengths of the noise as a function of a. Although there is degradation in these indicators with increasing noise, we see very clearly, a maximum in Φ and minima in ξ_2, and ρ_2 even for σ on the order of 1/2.

Finally, we comment that we have not yet discussed any of the statistical properties of the indicators presented in this section. An appropriate statistical interpretation for these

quantities depends on the context in which they are used. In particular, one will often wish to compare the values of the indicators computed on a data set, with values of the same quantities, but associated with a data set satisfying a null hypothesis. These issues will be discussed in context in sections III and IV below, and in Appendix B.

III. Model Misspecification

In this section, we demonstrate the sensitivity of the indicators described above to misspecifications of a model. (As before, we have in mind deterministic models or those with additive noise. However, we believe that these techniques can readily be adapted to the study of more general models.) In the following example we make no attempt to deduce the correct model (indeed, in a search space of lags of the time series no simple map describes this process), but rather show that some of the statistics of section II correctly identify a linear fit to the time series as a misspecification, even when other techniques fail to do so. As we shall see, this example is helpful in its clarity, but is unfortunate in that the candidate model is a linear function. Strictly speaking, we are testing for linear misspecification in this example, for which there are other existing methods which may be useful. However, it should be clear that our approach applies equally well to nonlinear candidate models, and is therefore more generally applicable than methods designed to test only for linear misspecification.

With these comments in mind let us consider a time series generated by summing together four independent Henon maps in the chaotic region. This example was discussed in a nice paper by Theiler and Eubank[5], and we shall compare with their results in what follows.

We consider four independent samples of the chaotic Henon map generated by

$$x_j(i) = 1 - 1.4x_j(i-1)^2 + 0.3x_j(i-2)\ ;\ j=1,2,3,4. \qquad (9)$$

The time series to be studied is given by

$$y(i) = \sum_{j=1}^{4} x_j(i). \qquad (10)$$

As pointed out in Ref. 5, if one attempts to model this series (with 512 data points) with a linear auto regressive fit, the best fit, f(i), according to several criteria involves six terms.

The series of residuals of this fit are, according to several tests, statistically indistinguishable from white noise. The residuals also pass a modified version of the BDS statistic as IID. Theiler and Eubank use these results to argue that "bleaching" chaotic data, i.e., studying the residuals of a chaotic series after subtracting a model fit can be misleading. In this case, even a test as sensitive as BDS fails to identify the residual series as non-IID.

In the tables below, we report the results of applying the statistics discussed in the previous section to the analysis of this series. In each of these tables, we report the value of a statistic applied to the data series, as well as the mean, $\hat{S}$, and standard deviation, σ, of that statistic when applied to an ensemble of bootstrapped data sets[6] which are constructed from the best linear AR fit. The k^{th} bootstrapped time series is given by

$$z_k(i) = \tilde{f}(i) + [y(i)-f(i)]_R \qquad (11)$$

Here $\tilde{f}(i)$ is the same functional form as f, the best linear AR fit to the original data, but the arguments here are lags of the bootstrapped data set. The subscript R on the last term indicates that the elements in the square brackets have been randomized to produce an IID series with the same probability distribution as the residuals of the original data set. Thus, the bootstrapped data set, z, are correctly described by the linear AR model with IID residuals. This process is repeated 1000 times, and the mean and standard deviation of the statistic on the collection of bootstrapped data sets is computed. We then compute the significance, Σ

$$\Sigma = \frac{S - \hat{S}}{\sigma}. \qquad (12)$$

This is a measure of the statistical significance of the statistic on the original data set. We look for values of Σ greater than about three as a rough indication of statistical significance. That is, if Σ is greater than about three, we conclude it is unlikely that the original data set could be produced by a linear auto regressive process.[7] As a further measure of statistical significance, we report in the last row of these tables, ν, the number of bootstrapped realizations out of 1000 which have a value of S exceeding the value of S for the original data set. In Tables I-III the statistics are calculated with a value of μ=0.4.

As a first attempt to uncover model misspecification, let us look for structure in the sequence of residuals. In Table I we apply the Savit-Green statistics to the series of

residuals. These statistics are related to the BDS statistics, but are differentially indicative of dependence on specific variables. In this table, δ_j is a measure of the additional dependence of an element of the residual time series on the jth lag of the residual time series, given that the information in the intervening j-1 lags has been used.[8] Looking at the first two columns of this table, we see that the significances for this dependence on the first lag, and on the second lag, given that we have used the information in the first lag are not large, and so, consistent with the general conclusions of Ref. 5, the residuals pass this test as IID. In the third column of the table we also calculate δ_1',the dependence of the residual series on the second lag *without* including the dependence on the first lag. Again, we see no statistically significant dependence.

In Table II, we compute the ρ_k of equation (5), which measures the dependence of the residuals on lags of the original time series. Here we see very significant dependence on the first lag. This is a clear indication that there is significant remaining structure in the residuals and that the linear auto regressive fit is not, therefore, a good explanation of the original time series.

To make the point still clearer, we compute, in Table III, the functions Φ, (eq. (2)) and ξ_k (eq. (4)). The first column in this table compares Φ for the original and bootstrapped data sets. We see that Φ is roughly the same for the original and bootstrapped data sets. That this should be so is not obvious, but neither is it completely unexpected. If the size of the residuals are fairly small, and if ε is not too small, then we might expect that Φ, which is a global measure of the predictability of the linear fit, could be roughly the same for the original and bootstrapped data. Turn now to columns 2-4. Here we compute ξ_k with k=1,2 and 3 for the original data set and its bootstraps. Recall that the ξ_k are differentially indicative of missing dependence in the model function. The significance in column 2 shows that there is additional structure in the original data set which is not captured by the linear auto regressive fit. Thus, we conclude again that there is significant nonlinear structure in this data set that is not captured by the linear auto regressive model. Moreover, the significances in these columns suggest that a description of this nonlinearity could involve, primarily additional dependence on the first lag of the time series. Note that this is consistent with the results of Table II in which we saw that the residuals were primarily dependent on the first lag of the original time series.

There are also other methods that one can use to test for model misspecification that do not require a strict analysis of residuals. For example, one can consider applying the δ's or the

BDS statistic, to the original time series, and then constructing a set of bootstrapped data to the time series as in equation (11) above. One could then compare values of these statistics on the original series and the set of bootstrapped data. Although neither the BDS nor Savit-Green statistics were originally designed to be used in this way, a statistically significant difference between their values on the original and bootstrapped data sets would indicate that the proposed model was inadequate. It is interesting to explore the relative power of this approach and the methods proposed here. As an example, we have computed δ_1 with μ=0.4 on the original time series given in equation (10), and compared it to the value computed on a set of bootstrapped data constructed according to (11). We find that the significance of the statistic in this case is only 1.6. This should be compared with Table III in which we see that the significance of ξ_1 for the same value of μ and similarly constructed bootstrapped data is 3.8. Thus, for this parameter setting ξ_1 is able to distinguish a poor fit to the series (7), while this use of δ_1 is not. We have not made a systematic comparison of these two methods, but it would certainly be interesting to do so.

At this point we digress briefly from our main discussion to describe an interesting observation about this time series. The reader who wishes to focus only on the main points in the paper may skip to Section IV.

In Table IV we present the values of the quantity

$$\xi' = 1 - \frac{P(t_0(\mu)|F(\mu),t_1(\mu'))}{P(t_0(\mu)\,|F(\mu),t_1(\mu'),t_2(\mu))} \qquad (13)$$

where we have explicitly indicated the fact that t_0, F and t_2 are computed with a different value of μ than is t_1. This is an obvious generalization of expression (4), which is sensitive to dependence on the second lag of the time series given the dependence contained in the first lag and in the model function. (For more details about such indicators see Ref. 4.)

Now, by increasing μ' with fixed μ, we can study the additional nonlinear dependence on the second lag in the series, as we lose the information about the nonlinear dependence in the first lag. (Remember that in all of this, the linear auto regressive model is used to "filter" the linear dependence.) As Table IV shows, the dependence on the second lag decreases if we ignore the dependence on the first lag. This is somewhat counter-intuitive: In a more typical case (such as that arising from a chaotic map), one might expect the

remaining information in the second lag to increase if we ignore the first lag. One kind of structure that could give rise to the behavior seen in Table IV is one in which the value of the first lag determines a sector or branch of the process, such that there is clear deterministic dependence on the second lag within each branch.

TABLE I

	δ_1	δ_2	δ_1'
statistic (S)	0.0215	0.0229	0.0020
$\bar{S}$ (bootstrapped data)	-0.0014	-0.0022	-0.0026
standard deviation	0.0145	0.0233	0.0145
significance	1.57	1.08	0.317
ν -value	62/1000	138/1000	350/1000

The Savit-Green statistics applied to the series of residuals. μ=0.4. ν-value is defined in the text.

TABLE II

	ρ_1	ρ_2	ρ_3
statistic (S)	0.0577	-0.020	-0.0036
$\bar{S}$ (bootstrapped data)	-0.0025	-0.0015	-0.0020
standard deviation	0.0146	0.0154	0.0151
significance	4.12	-1.183	-2.104
ν -value	0/1000	108/1000	465/1000

Dependence of the residuals on lags of the original time series. μ=0.4. ν-value is defined in the text.

TABLE III

	Φ	ξ_1	ξ_2	ξ_3
statistic (S)	0.2179	0.0666	-0.0027	-0.0018
$\bar{S}$(bootstrapped data)	0.220	-0.0010	-0.0022	-0.0024
standard deviation	0.0279	0.0178	0.0206	0.0189
significance	-0.078	3.80	-0.0227	0.0293
ν -value	472/1000	1/1000	493/1000	469/1000

Additional dependence of the original time series on lags thereof, given the linear auto regressive fit as a functional filter. μ=0.4. ν-value is defined in the text.

TABLE IV

μ'	ξ'	significance, Σ
0.1	0.189	7.1
0.2	0.139	7.7
0.4	0.098	7.5
0.8	0.050	4.5
1.6	0.036	4.4
3.2	0.035	4.6

Values of the expression (13) for different values of μ'. Here μ=0.1 and a time series of 8,192 points is used.

IV. Model Building

In addition to their use in model misspecification tests, the quantities described in section II can also be used to more efficiently generate a model for the data. To begin the discussion, it is important to distinguish two cases: one in which the search space is complete and the other in which the search space is incomplete. By a complete search space, we mean that the system can, in principle, be completely described by functions contained in the space. In a complete search space, a number of procedures, including ones based on the quantities in section II, will converge to the correct model. An example of model construction in a complete search space is contained in appendix B.

The case of an incomplete search space is considerably more interesting. Unlike the case of the complete search space, our methods lead to a different, and for many purposes superior description of the system than do other procedures, including those designed to minimize residuals such as a least squares approach. We shall show, in particular, that, when restricted to the same search space, our method leads to better medium-term predictions than does a least squares fitting procedure or even a simple Markov approach, as explained below.

The philosophies behind a straight forward data fitting approach and one based on the use of the methods of section II are quite different. Residual minimizing procedures generate models which track the data as closely as possible (in one sense or another), without regard

to the structural differences between the data and the model. The approach we shall describe here is based on the construction of a set of pointers associated with the original data set. These pointers are determined by optimizing the conditional probabilities described in section II. Consequently, the relationship between the set of pointers and the data is dictated by an information criterion, rather than a criterion based solely on the size of residuals. As we shall argue, our method makes efficient informational use of the search space.

In some sense, our approach to modeling is related to a Markov process and may be thought of as an optimized generalization thereof. To specifically demonstrate one sense in which our procedure improves upon a Markov approach, we shall show that using our methods as a basis for a simple prediction algorithm gives superior results over the same prediction algorithm used in conjunction with the Markov description of the system. Moreover, as we shall argue, the use of a model function as a variable in a phase space reduces the dimensionality of the reconstruction, resulting in a more efficient statistical use of the data.

To illustrate our points, it is easiest to consider an example. To this end we consider the Lorenz system generated by the equations

$$\begin{aligned} \frac{dx}{dt} &= 10\,(y - x) \\ \frac{dy}{dt} &= 28x - y - xz \\ \frac{dz}{dt} &= xy - 1.6z \end{aligned} \qquad (14)$$

The Lorenz system is quite simple which is both an advantage and a disadvantage. Pedagogically it is advantageous to have a simple system to study. On the other hand, many methods will do fairly well at describing and predicting such a simple system, so that the marginal advantages of our method are not as pronounced. Nevertheless, the advantages are significant, and should be even more marked on higher dimensional systems.

To construct our time series, we compute the integral time scale for x(t), and sample that series at intervals of 0.1 times the integral time scale to produce a discrete time series, x'(n). This series is then normalized by its standard deviation to produce a time series x(n) whose terms roughly lie in the range (-3,3). Our search space will be defined by functions

which are linear combinations of the lags of x(n). This space is not, in principle, sufficient to completely describe the Lorenz system. This is the incomplete search space to which we shall limit ourselves for purposes of constructing fits to the data, or functions to use as pointers to the data. We now seek a best characterization of the data limiting ourselves to the search space of linear combinations of no more than four lags of x(n).

The most straightforward use of these functions for attempting to model this system is to try to fit the data with a residual minimization procedure. The most common of these is to search for a least squares fit in the space of linear combinations of lags for x(n). For comparative purposes, we present the result of this exercise. It is

$$\hat{x}(n) = 2.02\hat{x}(n-1) - 1.94\hat{x}(n-2) + 1.15\hat{x}(n-3) - 0.35\hat{x}(n-4) \qquad (15)$$

This fit to the data is shown in Fig. 5. One important characteristic of this fit, qualitatively evident from the figure, is that the residuals are large. The rms value of the residuals are about 27% of the value of the time series. This is a rather poor least squares fit, but in this search space is, nonetheless, the optimal one. We shall show below that if the estimates, $\hat{x}(n)$, given by (15) are used as predictors of x(n), the results are also quite poor.

It is clear that the least squares method provides limited information for the system (14) in this search space. We turn now to another method using conditional probabilities. In this approach we will be addressing a somewhat different question than that addressed by a least squares method. We shall be looking for an optimal indicator that is associated with maximal information about x(n) in the following sense: We generate a value $y_1(n)$ which is a linear combination of x(n-1), . . .,x(n-4). We then seek to maximize the conditional probability

$$S_1 = P(t_0 | F) = P(t_0 | Y_1) \qquad (16)$$

where t_0 stands for $|x(i)-x(j)| \leq \varepsilon_x$ and f is the function y_1, and Y_1 stands for $|y_1(i)-y_1(j)| \leq \varepsilon_y$. Notice that with this approach, we do not necessarily require that y_1 be a model for x, in the sense that corresponding values are close. Rather, we seek the y_1 that has strongest informational relationship to x.[9]

Another way to describe this condition of maximal predictability is to imagine that our series is divided up into a training set and a test set. The training set is used to establish a relationship between a set of pointers, in our case, the function $y_1(i)$, corresponding to the i^{th} element of the time series. This set of pointers is constructed by maximizing S_1 defined in (16). Let $\tilde{x}(j)$ be our prediction of x(j), where j is a point in the test set. Let i be a point in the training set. We compute the pointer function, $y_1(j)$, and choose, randomly, a value of $\tilde{x}(j)$ from the set of x(i) in the training set such that $|y_1(i) - y_1(j)| < \varepsilon$. Then, prediction success is measured by computing $P(|x(j) - \tilde{x}(j)| < \varepsilon)$. This prediction procedure is clearly closely related to the statement that $y_1(i)$ is chosen to maximize S_1. Therefore, in the sense of maximizing $P(|x(j) - \tilde{x}(j)| < \varepsilon)$ such predictions should in general be better than a least squares fit, at least for a large class of stationary, nonlinear processes.

For our example we chose $\mu = 0.05$ for both the series x, and the function y_1. We use a simple brute force method for maximizing S in the search space, sequentially maximizing with respect to each parameter, and then iterating to make sure we have identified the global maximum. (See Appendix B for more details.) Maximizing (16) we find for y_1

$$y_1(n) = x(n-1) - 0.27x(n-2) + 0.002x(n-3) + 0.006x(n-4). \qquad (17)$$

where the coefficient of the term x(n-1) is, arbitrarily and for simplicity chosen to be one. (Scaling the y's by an overall factor makes no difference in the analysis.)

Notice that this function is considerably different from (15), generated by a least-squares procedure. Although the search space for the least squares fitting procedure and for the function in (17) are the same, it is important to stress the fundamental differences between these two procedures. In the least squares case, the model is linear with all the attendant limitations thereof. In the case of (17), it is the pointer function that is linear, but the relationship between the pointers and the data does not have to be (and in general is not) linear.

The value of the conditional probability (16) is 0.22 using (17). In Fig. 6 we have plotted x(n) as a function of $y_1(n)$. This graph is qualitatively similar to Fig. 5, but has some different characteristics. Note that the scatter of points is more uniform along the curve in Fig. 5, (the result of least-squares minimization), while in Fig. 6 the distribution is narrower for most of the points and considerably more broad for the few points near the

ends of the lobes. Because of the large spread of these few points the average value of the residuals will be increased, but because most of the points lie in a narrower band, the conditional probability (16) will be larger. We also note that maximizing S1 is equivalent to maximizing Φ, defined in equation (2).

The predictability of this series can be systematically improved by an iterative procedure of which the maximization of (16) is the first step.[10] We can add another function as another condition to the conditional probability to improve the results. That is, we now seek another function, $y_2(n)$, which maximizes $S_2 = P(t_0 \mid Y_1, Y_2)$ using y_1 in (17). We find for y_2

$$y_2(n) = x(n-1) - x(n-2) + 0.24x(n-3) - 0.006x(n-4). \qquad (18)$$

The inclusion of y_2 results in a dramatic improvement in predictability: $S_2 = 0.73$, compared to a value of 0.22 for S_1. The reason for this result can be seen graphically in Figs. 7a-f in which we plot x(n) as a function of $y_1(n)$ for different ranges of values of $y_2(n)$. Although these graphs are not entirely single-valued, it is clear that there is a single or small range of values of x(n) associated with most values of $(y_1(n), y_2(n))$. Notice that although there is a close relationship between x(n) and $(y_1(n), y_2(n))$, we have not constructed a parametric model for x(n) in the usual sense. Rather, the set $(y_1(n), y_2(n))$ should be understood as pointers to an associated value, x(n).[11]

It is possible to improve the predictability still further by adding another functional condition. For our example, we have maximized the value of $S_3 = P(t_0 \mid Y_1, Y_2, Y_3)$ over choices of $y_3(n)$ in our search space, given the previously determined $(y_1(n), y_2(n))$. We find that S_3 is maximized by

$$y_3(n) = x(n-4) \qquad (19)$$

and for this choice, $S_3 = 0.84$. For this case, within the limits of our statistics, we are not able to find a fourth linear combination of the four lags that results in any further improvement in S. Therefore, $S = 0.84$ is the best we can obtain, and that is obtained with three functions. We believe that the sufficiency of three functions for maximal predictability is related to the fact that the correlation dimension of the Lorenz system is between 2 and 3.

Optimizing the S_j in the search space results in a more efficient summary of the information contained in the time series. We can demonstrate that both by comparing the values of short term predictabilities with and without optimization, and by comparing the fidelity of longer-term predictions with and without preliminary optimization.

First, we have computed the conditional probabilities $S'_j = P(t_0 \mid t_1, \ldots t_j)$, using the notation following equation (1). These are just the predictabilities using only the lags of the time series as conditions, rather than the functions y_j. We find $S'_1 = 0.13$, $S'_2 = 0.44$, $S'_3 = 0.58$, and $S'_4 = 0.68$. These values are considerably smaller than the corresponding values using the y_j as conditions. In fact, S'_1 is even smaller than the predictability computed using the least squares fit, (15), which is 0.20.

Success at prediction is one (but certainly not the only) measure of one's ability to extract information from a time series, particularly for stationary series. So we have studied predictions with and without informational optimization in the search space to gauge the ability of various schemes to represent the information in the series. In the case of the Lorenz system, one step-ahead prediction is quite simple. The reason is that, when sampled at intervals of 0.1 times the integral time scale, the series shows considerable continuity with, in general, a non zero first difference. Thus, nearly any sensible prediction method will do reasonably well. In fact, even the linear AR fit, which is certainly a very poor representation of the dynamics is a good guide to the next value of the series. A more sensitive discriminant is longer term predictions to which we now turn.[12]

First we note, parenthetically, that the AR fit (15) decays to zero after several oscillations, and is thus a very poor predictor of the series except for very short times. An AR fit with noise does not decay to zero, but also does a very poor job of tracking the data for more than one or two time steps into the future.

Since our interest here is to demonstrate the added value from optimizing the reconstruction space in the sense of the S_j's, we turn instead to a discussion of longer term predictability using a simple prediction algorithm overlaid on our optimization scheme. We shall compare predictions made using conditional probabilities in which the (optimized) y's are the variables that appear in the conditions, to predictions made using simple lags as variables in the conditional probabilities. Conditional probabilities in which time lags appear as variables are quantities that are naturally associated with reconstruction spaces in which the axes are lagged values of the time series. Such reconstructions are, in turn,

naturally associated with simple Markov processes. It is not clear whether the discretized time series, x(i), can be generated by a simple Markovian process involving only a small set of time lags. Therefore, it is reasonable to seek a representation of the information which leads to an improvement to predictions made using a Markov picture. In what follows we show that our optimization scheme does just that.

Specifically, our procedure is the following: Associated with each value of x(j) from the training set is a value of the set $\{y_k(j)\}$. (Note that the $\{y_k(j)\}$ are combinations of the lags x(j-1), . . .,x(j-4).) To predict x(m+1), we compute the set $\{y_k(m+1)\}$. We then find values $\{y_k(j)\}$ from the training set which are close (within ε) to the $\{y_k(m+1)\}$. As our prediction for x(m+1), $\tilde{x}(m+1)$, we choose, randomly, one of the values of the associated $\{x(j)\}$. Next, to predict x(m+2), we use $\tilde{x}(m+1)$ to compute $\{y_k(m+2)\}$, and the procedure is repeated to obtain a prediction, $\tilde{x}(m+2)$, for x(m+2).

A similar method is used to obtain predictions from the Markov process. In that case, though, a set of k-lags (x(j-1), x(j-2), . . .,x(j-k)) = $\{x_k(j)\}$ replaces the optimized pointer values $\{y_k(j)\}$. Again, the predictions are deduced by choosing randomly from the set of (now Markovian) pointers $\{x_k(j)\}$in an ε-neighborhood of the Markovian pointer set $\{x_k(m+1)\}$. The predicted value of x(m+1), $\tilde{x}(m+1)$, is that value from the training set associated with the chosen pointer. As before, that predicted value is used to compute $\{x_k(m+2)\}$, and thus to predict x(m+2). The process is repeated to obtain predictions further in time.

For our simulations, we used a training set of 4096 points and a value of μ=0.05. We ran 100 realizations of the predictions using the Markov and optimized methods. The results of these simulations are summarized in Fig. 8 and in Table V. In Fig. 8 we show the time series in the test set as well as typical runs for the two prediction algorithms. In Table V we present the standard error for the prediction algorithms based on a simple time lags reconstruction (naturally associated with a Markov process) and based on the optimized reconstruction described above. It is clear from this table that the prediction algorithm based on the optimized reconstruction is consistently more accurate than the same prediction method based on the time lags reconstruction.[13,14] We emphasize again that our purpose here is to use these simple predictions only as indicators of the relative information content of the time lags and optimized reconstructions, which we take as a measure of the added information from optimizing the reconstruction space.

TABLE V

Time Step	time lags (Markov)	optimized
1	.026	.017
2	.053	.025
3	.076	.027
4	.101	.054
5	.133	.085
6	.107	.053
7	.116	.118
8	.191	.130
9	.139	.083
10	.110	.083
15	1.153	.705
20	.474	.314
25	1.593	.963
30	.622	.578

The standard error for the prediction algorithms based on a simple time lags reconstruction and the optimized reconstruction described in the text.

We can also improve our results by expanding the search space. In the example discussed here, we limited ourselves to linear combinations of four lags of x(n). We have also computed S_1 for a search space in which we allow y_1 to have terms of up to third order in the four lags of the time series. In this space, S_1 has a value of 0.32, nearly a 50% improvement over its value computed with only linear combinations of the lags.

Finally, as discussed earlier, we note again that this procedure is a generalization of the formalism of Ref. 4. The difference here is that we include functions of a simple set of variables (say the lags of a time series) as directions in a space for reconstruction of the time series. Other aspects of the formalism described in Ref. 4 can be applied here also. For example, one can define a set of δ's using the y_j's as variables instead of the time lags.

V. Summary and Discussion

In this paper we have introduced new tools for analyzing time series of complex systems. The tools are based on introducing a model function as a variable in the construction of joint and conditional probabilities. We have demonstrated two uses for the tools. First, we showed that they could be used to sensitively test for model misspecification. It is particularly interesting that the use of the test function as a variable in the conditional probabilities results in a considerably more sensitive test (the ξ_k) for misspecification than does an analysis only of the residuals, or even the application of the correlation-integral based statistics (like the δ's) to the original time series. This is true even though all analyses are based on Grassberger-Procaccia-like correlation integrals. Indeed, we also saw in our example in section III that a hybrid statistic (the ρ_k) which asks whether the original time series contained any additional information for the residuals also resulted in a rejection of the IID null for the residuals. In constructing the sequence of residuals one may distort the structure of the time series, making it more difficult to uncover information in an analysis that stays only within the sequence of residuals. This observation is consistent with the warnings of Theiler and Eubank[5] about the dangers of trying to bleach chaotic data. The example in which the δ's applied to the original time series failed to reject a linear IID null while the ξ's did is intriguing and, *a priori*, somewhat more difficult to understand.

As we saw, an analysis based on these tools is also able to directly indicate additional dependence on specific variables after have allowed for the explanatory effect of the model. It can therefore be used constructively to generate model explanations of the data. We showed in Appendix B that in a complete search space this method was able to generate the correct model and parameter values for a time series produced by a noiseless underlying nonlinear map. This is not too surprising, since a least squares method applied to the same search space will also yield the correct model. We also showed that the method continued to be robust with the addition of moderate amounts of dynamical noise. The even more interesting case was that of an incomplete search space, discussed in section IV. There we showed that the description of the system produced by our method was of an entirely different nature than that produced by a residual minimizing procedure. (Even the linear function y_1 differed significantly from linear least squares fit.) Our method is based on the optimization of a set of conditional probabilities. To demonstrate the additional information contained in our reconstruction, we compared predictions made using the optimized basis to those made using a Markovian basis. We found significant improvement in medium-term prediction performance using our optimized method.

From these comments it is clear that optimizing in the search space produces a model with improved predictive power. But, as we discussed earlier, and as is pointed out in Ref. 13, finite time predictability, although important, should not necessarily be regarded as a sufficient criterion for a good model of the dynamics of a system. Particularly in those cases which are not, in principle, data limited, it is not unreasonable to seek models which faithfully reproduce the asymptotic behavior of the system, at least approximately. Both the Markov and our optimized description can reproduce the topological and geometric features of a hyperbolic system (or, in the case of a non-hyperbolic system such as the Lorenz attractor, the hyperbolic sub-set). This is a property not, in general, shared by incomplete "fits" to a data set. In this sense, an information based description of the data, such as a Markov description or our optimized description is asymptotically advantageous, and so theoretically preferable.

Our work raises a number of interesting questions. One interesting issue to address is that of proving the asymptotic properties of our statistics under various null hypotheses. In particular, can one generalize the approach of references 3 and 15 to determine the large sample size behavior of the ρ_k's, ξ_k's and the Φ's under nulls related to the functional forms used in the conditional probabilities?

A second very interesting problem, alluded to above, is that of understanding the difference in performance of the δ's (or the BDS statistic) applied to the original time series, and the ξ's defined in (4). The discussion in section III suggests that the latter are considerably more sensitive at rejecting incomplete models for complex data. It is not clear to what extent this is a property of the example we studied or whether this is more generally true.

A third question which we did not fully address is the question of what to do if there are nearly degenerate extrema in the search space. It is not clear what criteria to use to distinguish among them. It may also be that such degeneracies generically indicate the need to break the degeneracy by expanding the search space.

Several other interesting technical questions present themselves, including questions of convergence in an incomplete search space, and optimal methods for determining an adequate search space. These unresolved questions notwithstanding, it is clear that the tools and methods we have developed in this work are likely to be very flexible and useful in the analysis of time series of complex systems, particularly those for which we have only partial dynamical knowledge, and for which the search space is incomplete.

Acknowledgments

We thank Martin Casdagli for helpful discussions and comments on the manuscript. RS thanks the Santa Fe Institute, where part of this work was done, for their hospitality. This work was supported in part by the US Department of Energy, under grant no. DoE-DE-FG02-85ER45189, and by the National Institutes of Health, under grant no. R01NS31451-02.

APPENDIX A

Here we show that under not too restrictive conditions $\xi_k = O(\sigma_r)$ in the absence of noise for small μ. In what follows, we assume $|r| \le \varepsilon$. If this inequality is violated there may be corrections to the result we demonstrate. First we estimate the conditional probability:

$$P(t_0|F) = \frac{P(\Delta f + \Delta r \in \Omega_{\varepsilon'}(0), \Delta f \in \Omega_\varepsilon(0))}{P(\Delta f \in \Omega_\varepsilon(0))} \qquad (A.1)$$

where $\Omega_\varepsilon(a)$ is the interval of radius ε centered on a, $\varepsilon = \mu\sigma_f$ and $\varepsilon' = \mu\sigma_x$. Now, for a time series generated by an autonomous mapping, generically, the joint probability in the numerator of (A.1) is minimized if r and f are independent. Moreover, in that case $\varepsilon' \ge \varepsilon$, so we can write:

$$P(t_0|F) \ge \frac{\int dP(\Delta r) P(\Delta f \in \Omega_\varepsilon(-\Delta r) \cap \Omega_\varepsilon(0))}{P(\Delta f \in \Omega_\varepsilon(0))} = \qquad (A.2)$$

$$= \frac{\int_{-2\varepsilon}^{2\varepsilon} dP(\Delta r) P(\Delta f \in \Omega_\varepsilon(-\Delta r) \cap \Omega_\varepsilon(0))}{P(\Delta f \in \Omega_\varepsilon(0))} =$$

$$\equiv \int_{-2\varepsilon}^{2\varepsilon} dP(\Delta r)\left(1 - \frac{|\Delta r|}{2\varepsilon}\right) = P(|\Delta r| \le 2\varepsilon) - 2\frac{\int_0^{2\varepsilon} dP(\Delta r)\Delta r}{2\varepsilon}$$

Now it can be seen that if $P(|\Delta r| \le 2\varepsilon) = 1$, the last relation can be written as:

$$P(t_0|F) \ge 1 - 2\langle|\Delta r|\rangle \qquad (A.3)$$

Thus, $P(t_0|F) = 1 - O(\sigma_r)$. If we assume, generically, that by including information on a time lag the conditional probability cannot get lower, $P(t_0|F, t_k) = 1 - O(\sigma_r)$. Thus, under the same conditions $\xi_k = O(\sigma_r)$. Note that this result does not strictly imply the linear behavior of ξ_k near its minimum, although that behavior is consistent with this result.

To show that $\rho_k = O(\sigma_r^0)$ under the same conditions, we observe that if we scale r with α, then ε_r also scales with α, and so:

(A.4)

$$\rho_k(\alpha) = 1 - \frac{P(\alpha\Delta r \in \Omega_{\alpha\varepsilon'}(0))}{P(\alpha\Delta r \in \Omega_{\alpha\varepsilon'}(0) | \Delta x \in \Omega_{\varepsilon'}(0))} = 1 - \frac{P(\Delta r \in \Omega_{\varepsilon'}(0))}{P(\Delta r \in \Omega_{\varepsilon'}(0) | \Delta x \in \Omega_{\varepsilon'}(0))} = \rho_k(1).$$

Thus ρ_k is independent of σ_r.

APPENDIX B

To demonstrate the behavior of our method in a complete search space, we turn to a fairly simple example. For this example, we shall start with a model which is close to, but not correct for the data and we shall show how using the conditional probabilities can lead us to a more accurate model. We shall use a data set generated by the expression

$$x(i) = 1 - 1.4x^2(i-1) + 0.3x(i-2) + 0.2x^3(i-3). \qquad (B.1)$$

This expression is closely related to the Henon map. Using standard methods, we estimate the correlation dimension of the attractor generated by (B.1) at about 1.44, and the second Kolmogorov entropy at 0.37.

The computations described below were carried out on a data set with 8,192 points and were done with μ=0.05. In this section we will present values of the computed indicators. Issues of statistical significance will be discussed below.

In what follows we shall use an iterative procedure to maximize the probabilities introduced in Section II. We shall show that this approach converges in this example to the correct dynamics for this time series. To begin, we postulate, as a model for the data in (B.1),

$$f(i) = x^2(i-1) - ax(i-2). \qquad (B.2)$$

At this point we note that since all of our statistics involve differences of quantities, our techniques are insensitive to overall additive constants. In addition, our results are qualitatively independent of overall normalization factors in the function f. After the correct ratios of parameters in the model are determined, additive constants and overall multiplicative factors can be simply deduced by computing a least squares fit of f(i) to x(i) in a straightforward way.

Our first step is to compute the quantity Φ defined in equation (2), which is a global measure of the goodness of fit of the function f. In Fig. 9 we plot the value of Φ with this f, as a function of a. At this point we can not ascribe a precise meaning to the values of Φ, but we see that the curve has a mild, unpronounced maximum near a=0.1. This suggests that with the form (B.2), a value of a=0.1 may provide a somewhat optimized model for

the data. To further verify this possibility, we compute, in Fig. 10 the function ξ_2 with f given in (B.2), as a function of a. We thus ask for the extra dependence on the second lag in addition to that contained in the function (B.2). We see in this graph a mild minimum at a=0.1, confirming our tentative conclusion that this value of a provides an optimal model for the data, over the class of functions (B.2).

Next, we use this tentative value of a to search for additional dependence on x(i-3). To this end, we compute Φ with

$$f(i) = x^2(i-1) - 0.1x(i-2) - bx^3(i-3). \qquad (B.3)$$

as a function of b. The results are shown in Fig. 11. Here we see two maxima at b=0.05 and 0.15. To further check the optimal value of b, we compute ξ_3 as a function of b, to search for addition dependence on the third lag. In Fig. 12 we see two minima at b=0.05 and b=0.15. To find an optimal fit, we shall fix each of these candidates for b in turn, and return and vary a to search for a better value of the coefficient of the second term. In figures 13 and 14 we plot Φ computed with

$$f(i) = x^2(i-1) - ax(i-2) - bx^3(i-3) \qquad (B.4)$$

as a function of a with b = 0.05 and b = 0.15, respectively. In Fig. 14 we see a rather pronounced maximum at a = 0.22. In Fig. 13 there is a less pronounced maximum at a = 0.15. In addition, the value of Φ at the maximum of Fig. 14 is somewhat higher than the corresponding value of Φ at the maximum of Fig. 13. This leads to the tentative conclusion that b = 0.15 is likely to be a better fit to the data than is b = 0.05. In addition, we are led to a revised best value of a = 0.22. In Figs. 15 and 16, we compute ξ_2 for these two choices of b, as a function of a. Comparing these figures we see a strongly compelling argument for favoring b = 0.15 over b = 0.05. The minimum additional dependence on x(i-2) is much smaller in Fig. 16 than in Fig. 15, and the minimum is much more pronounced. An examination of these curves supports our tentative conclusions of a = 0.22, b = 0.15 as good parameter choices.

As we stated earlier, our calculations are relatively insensitive to additive constants and overall multiplicative factors in f. To fix these, we can calculate a least squares fit of f(i) to x(i). Doing this, we find that f should be multiplied by an overall factor of -1.4, and that

the function should have a constant, 1, added to it. This reproduces the function (10) which generated the data.

In the example above, we showed that the procedure quickly converges to the correct values of the parameters that generated the data if we chose the correct form (i.e., the correct power of the variables) for the monomials in the function, f. It is natural to ask how sensitive the expressions (2) - (5) are if we misspecify the power of the variables. To that end, we compute Φ and ξ_3 with the functions

$$f(i) = x^2(i-1) - 0.22x(i-2) - bx^n(i-3) \qquad (B.5)$$

with n = 1,2,3,4 as a function of b. These results are plotted in Figs. 17 and 18. For Φ plotted in Fig. 17 we see a fairly well-defined maximum for n=3 (the correct value) at the correct value of b, which is missing for the other values of n. In Fig. 18 we have plotted ξ_3 measuring the additional dependence on the third lag. Here we see that the minimum dependence on b for each curve is sharper and deeper for n=3 than for the other values. These results illustrate the sensitivity of these tests to misspecifications of the dependence when searching for the correct parameters of a model. In practice, then, one can distinguish among different functional forms by computing a global measure of information, and comparing the values of the measure among different options. For this purpose it seems more sensible to use Φ as a comparative measure. The reason is that Φ is a more direct measure of information content than are the ξ's. ξ measures additional dependence on a variable, and its value may therefore change depending on the variable upon which one is seeking to determine additional dependence. In particular ξ_3 may be different if one chooses, say, some function of x(i-3) rather than x(i-3) itself as the additional variable in (5). This ambiguity is related to the absence of a natural measure in the space of functions in which one tries to seek an optimum dynamical model for the data. The choice of an additional variable is not present as a source of ambiguity in Φ.

As another example of the sensitivity of these measures to misspecified functional forms, consider Figs. 19 and 20 in which we have plotted Φ and ξ_2 for the function

$$f(i) = x^2(i-1) - ax^n(i-2) - .15x^3(i-3) \qquad (B.6)$$

as a function of a for n=1,2,3,4. In Fig. 20, it is the additional dependence on x(i-2) that is plotted. Here we see results analogous to those in Figs. 17 and 18. In Fig. 19 we see that

as a function of a, the maximum is highest and sharpest for n=1. Moreover that maximum occurs at a = 0.22 and is a global maximum for the four curves. In Fig. 20, the curve with n=1 has a minimum additional dependence on x(i-2) which is considerably more pronounced than any minima in the other three curves. Moreover, that minimum occurs at a=0.22 and is the global minimum for all four curves.

In the procedures for model building described above, we have used the shapes of the computed curves as a guide in choosing parameters in the models. We have said nothing about the statistical significance of these curves. In the examples we have discussed it was clear that the structure in the computed curves contained significant information which led us to a correct model. In other cases, however, particularly if the structure in the model is not very pronounced, or if the indicators hover near zero, it may be useful to compute more precise measures of statistical significance. Such a computation requires a null hypothesis with which to compare.

To illustrate, consider the following example: Suppose we wish to compute Φ as defined in equation (2) for some parameter choices in a model, f(i) for a data set x(i). A commonly chosen null hypothesis is that $x(i) = f(i) + \nu(i)$, where $\nu(i)$ is IID noise. To determine the statistical significance of Φ, one can construct a data set $y_k(i)$,

$$y_k(i) = \tilde{f}(i) + [x(i) - f(i)]_R \qquad (B.7)$$

where $\tilde{f}$ is the same functional form as f, but the arguments of which are lags of the bootstrapped data set, and the subscript R means that the residuals, x(i)-f(i) are randomized, to produce an IID set. An ensemble of such bootstrapped data sets are generated, and the function Φ is computed for the ensemble. If Φ for the original data set is significantly different than Φ for the ensemble of bootstrapped, data sets, then one can reject the null hypothesis that $x(i) = f(i) + \nu(i)$, with $\nu(i)$ IID. Of course, one must also be precise about the meaning of "significantly different"; typically, this will mean a few standard deviations away from the mean of Φ computed on the bootstrapped data set. Depending on the circumstances, this more careful statistical procedure may or may not be useful when using the indicators discussed in this paper in an iterative model fitting procedure.

References

1. Some previous work on using various criteria as guides to a good reconstruction space include A. Fraser and H. Swinney, Phys. Rev. A**33**, 1134 (1986), and A. Fraser Physica D**34**, 391 (1989), who use informational criteria. M. Casdagli, S. Eubank, J. Farmer, and J. Gibson Physica D**51**, 52 (1991) use geometrical criteria such as minimal distortion to indicate a good reconstruction. These references differ from our work, in the criteria used for optimality, and in the range of possibilities for the reconstruction directions.

2. F. Takens in *Dynamical Systems and Turbulence*, D. Rand and L. Young, eds. (Springer, Berlin, 1981), p. 366.

3. W. Brock, W. Dechert, and J. Scheinkman, "A test for independence based on the correlation dimension", Technical Report 8702, SSRI, University of Wisconsin (1987).

4. R. Savit and M. Green, Physica D**50**, 95 (1991); M. Green and R. Savit, Physica D **50**, 521 (1991).

5. J. Theiler and S. Eubank, Chaos, **3**, 771 (1993)

6. B. Efron and R. Tibshirani, Statistical Science **1**, 54 (1986); J. Theiler, B. Galdrikian, A. Longtin, S. Eubank and J. Farmer, "Using surrogate data to detect nonlinearity in time series" in *Nonlinear Modeling and Forecasting*, eds. M. Casdagli and S. Eubank (Addison-Wesley, Reading, MA, 1992), p.163; J. Theiler, S. Eubank, A. Longtin, B. Galdrikian and J. Farmer, Physica D**58**, 77 (1992).

7. We are using these measures of statistical significance in a semi-quantitative way. To ascribe quantitative confidence levels to values of Σ, we would have to study the distribution of the statistics on our bootstrapped data sets in more detail. This is not necessary for our purposes here.

8. Note that the δ_j indicates dependence on the jth lag conditional on dependence on previous lags, unlike the ξ_j and ρ_j. However, this in not the reason for the failure of the δ-statistics to indicate non-IID behavior of the residuals, since δ_1 computed for large time intervals also fails to indicated non-IID behavior.

9. It is possible, of course, that there could be some degeneracy, and that more than one y_1 would optimize the required conditional probability. In such an exceptional case the preferred optimal function could be chosen by imposing other criteria (for example, choosing that function with the best statistics).

10. It is important to note that such an iterative procedure is not within the philosophy of ordinary modeling techniques. Once a search space is chosen, and a criterion for goodness of fit established, there is, in general, one best model for the data. Any improvements must result from an expanded search space.

11. In principle, after determining y_2, one could go back and seek a new y_1 which further optimizes S_2. We have done this on this example, and found that the new y_1 differed little from the original one. A more systematic study of how to iteratively improve the reconstruction space is beyond the scope of this paper.

12. As emphasized elsewhere (see, for example, N. Gershenfeld and A. Weigend, "The Future of Time Series: Learning and Understanding" in *Time Series Prediction: Forecasting the Future and Understanding the Past*, Eds. A. Weigend and N. Gershenfeld, Santa Fe Institute Studies in the Sciences of Complexity, (Addison-Wesley, 1993), especially p. 10) even predictability for finite time has its limitations as a criterion for a good model. It is possible, for instance, for the behavior of the system to be accurately predicted for large but finite times, but to be poorly described asymptotically.

13. We note that for both prediction methods, the standard error oscillates as a function of time. This has to do with the oscillatory nature of the Lorenz system and is not really relevant to our point here.

14. We have also performed a similar calculation in a reconstruction space whose axes are the normal modes associated with four lags. We find no significant improvement in predictions over those made using the simple time lagged reconstruction. The reason for this is two-fold: First, singular value decomposition techniques are useful in the presence noise, which our example does not have. Second, the time delays used in our example are small enough (relative to the integral time scale) so that the distortion is realtively low both in the time-lagged and normal modes reconstuction.

15. K. Wu, R. Savit and W. Brock, Physica D**69**, 172 (1993).

Figure Captions

Fig. 1. Additional dependence on the second time lag indicated by ξ_2 and ρ_2 defined in (4) and (5). The time series is generated by (6) and f(i) is given in (7). The length of the time series, N=8192 and μ=0.05.

Fig. 2. The effect of noise on Φ. The time series is generated by the Henon map with noise (8), and f(i) is given by (7). N=8192 and μ=0.05.

Fig 3. The effect of noise on ξ_2 defined in (4). The time series is generated by the Henon map with noise (8), and f(i) is given by (7). N=8192 and μ=0.05.

Fig. 4. The effect of noise on ρ_2 defined in (5). The time series is generated by the Henon map with noise (8), and f(i) is given by (7). N=8192 and μ=0.05.

Fig. 5. The vectors $(x(i), \hat{x}(i))$ where x(i) is the time series generated by (14), and $\hat{x}(i)$ is our fit to x(i) given by (15).

Fig. 6. The vectors $(x(i),y_1(i))$ where x(i) is the time series generated by (14), and $y_1(i)$ is our improved embedding fit to x(i) given in (17).

Fig. 7. Slices of the phase plot, $(x(i),y_1(i))$ (Fig. 18), for various ranges of $y_2(i)$, given by (18).

Fig. 8. The time series x(i) generated by (14) compared to typical realizations of the predictions using both a simple time lags reconstruction and our optimized reconstruction.

Fig. 9. The function Φ defined in (2). The time series is generated by (B.1) and f(i) is given in (B.2). N=8192 and μ=0.05.

Fig. 10. Additional dependence on the second time lag indicated by ξ_2 defined in (4), as a function of a. The time series is generated by (B.1) and f(i) is given in (B.2). N=8192 and μ=0.05.

Fig. 11. The function Φ defined in (2) as a function of b. The time series is generated by (B.1) and f(i) is given in (B.3). N=8192 and μ=0.05.

Fig. 12. Additional dependence on the third lag indicated by ξ_3 defined in (4), as a function of b. The time series is generated by (B.1) and f(i) is given in (B.3). N=8192 and μ=0.05.

Fig. 13. The function Φ defined in (2) as a function of a with b=0.05. The time series is generated by (B.1) and f(i) is given in (B.4). N=8192 and μ=0.05.

Fig. 14. The function Φ defined in (2) as a function of a with b=0.15. The time series is generated by (B.1) and f(i) is given in (B.4). N=8192 and μ=0.05.

Fig. 15. Additional dependence on the second time lag indicated by ξ_2 defined in (4), as a function of a. The time series is generated by (B.1) and f(i) is given in (B.4) with b=0.05. N=8192 and μ=0.05.

Fig. 16. Additional dependence on the second time lag indicated by ξ_2 defined in (4), as a function of a. The time series is generated by (B.1) and f(i) is given in (B.4) with b=0.15. N=8192 and μ=0.05.

Fig. 17. The function Φ defined in (2) as a function of b. The time series is generated by (B.1) and f(i) is given in (B.5) with n=1,2,3,4. N=8192 and μ=0.05.

Fig. 18. Additional dependence on the third time lag indicated by ξ_3 defined in (4), as a function of b. The time series is generated by (B.1) and f(i) is given in (B.5) with n=1,2,3,4. N=8192 and μ=0.05.

Fig. 19. The function Φ defined in (2) as a function of a. The time series is generated by (B.1) and f(i) is given in (B.6) with n=1,2,3,4. N=8192 and μ=0.05.

Fig. 20. Additional dependence on the second time lag indicated by ξ_2 defined in (4), as a function of a. The time series is generated by (B.1) and f(i) is given in (B.6) with n=1,2,3,4. N=8192 and μ=0.05.

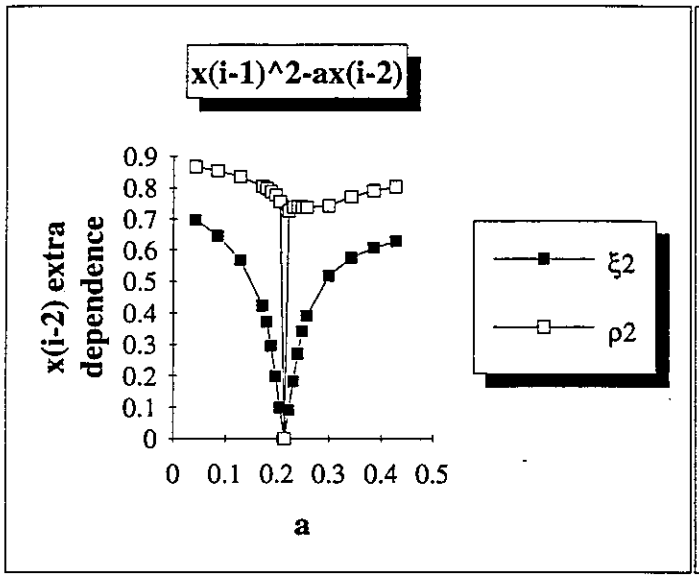

Figure 1

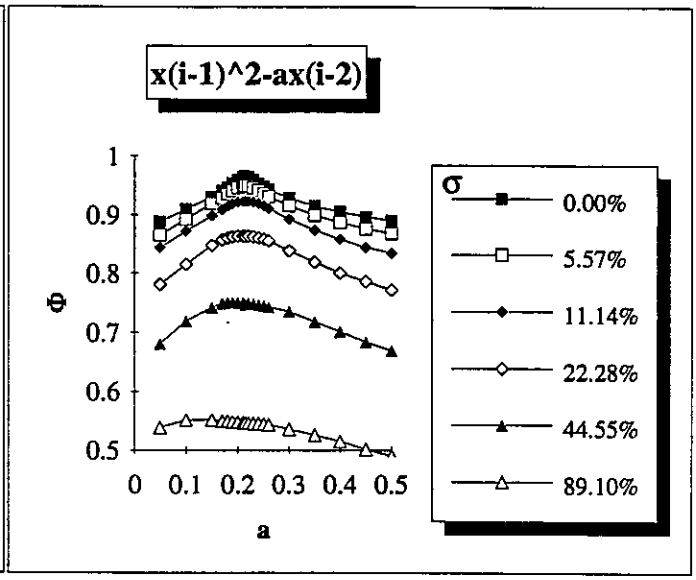

Figure 2

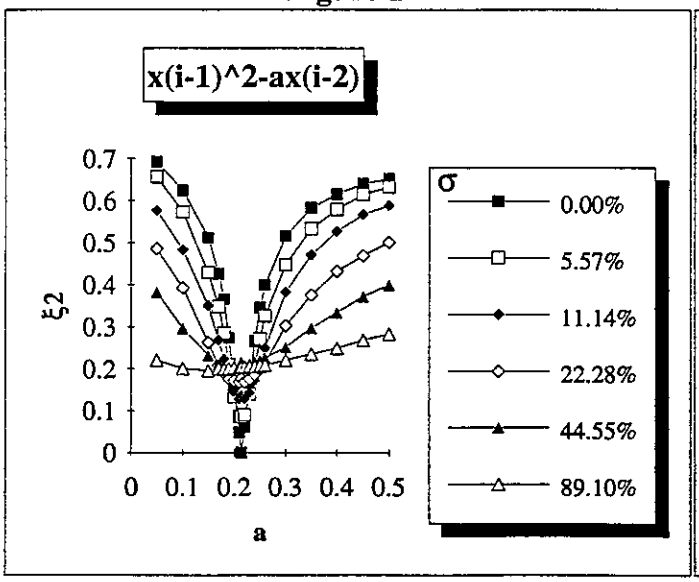

Figure 3

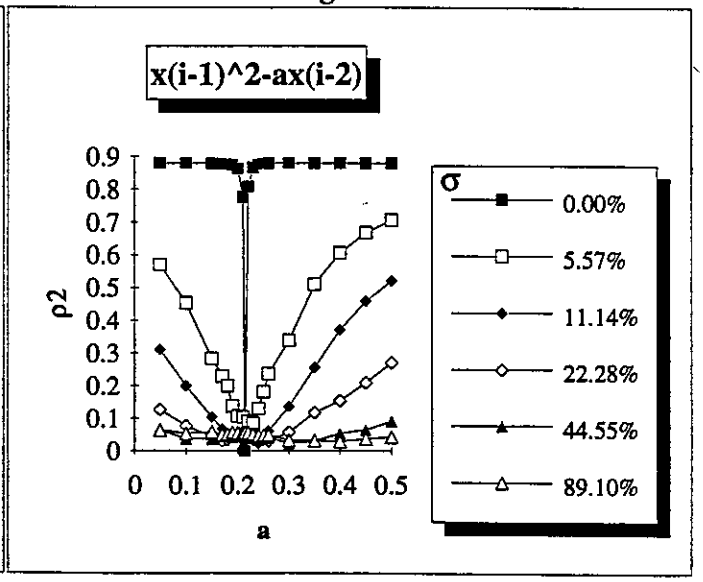

Figure 4

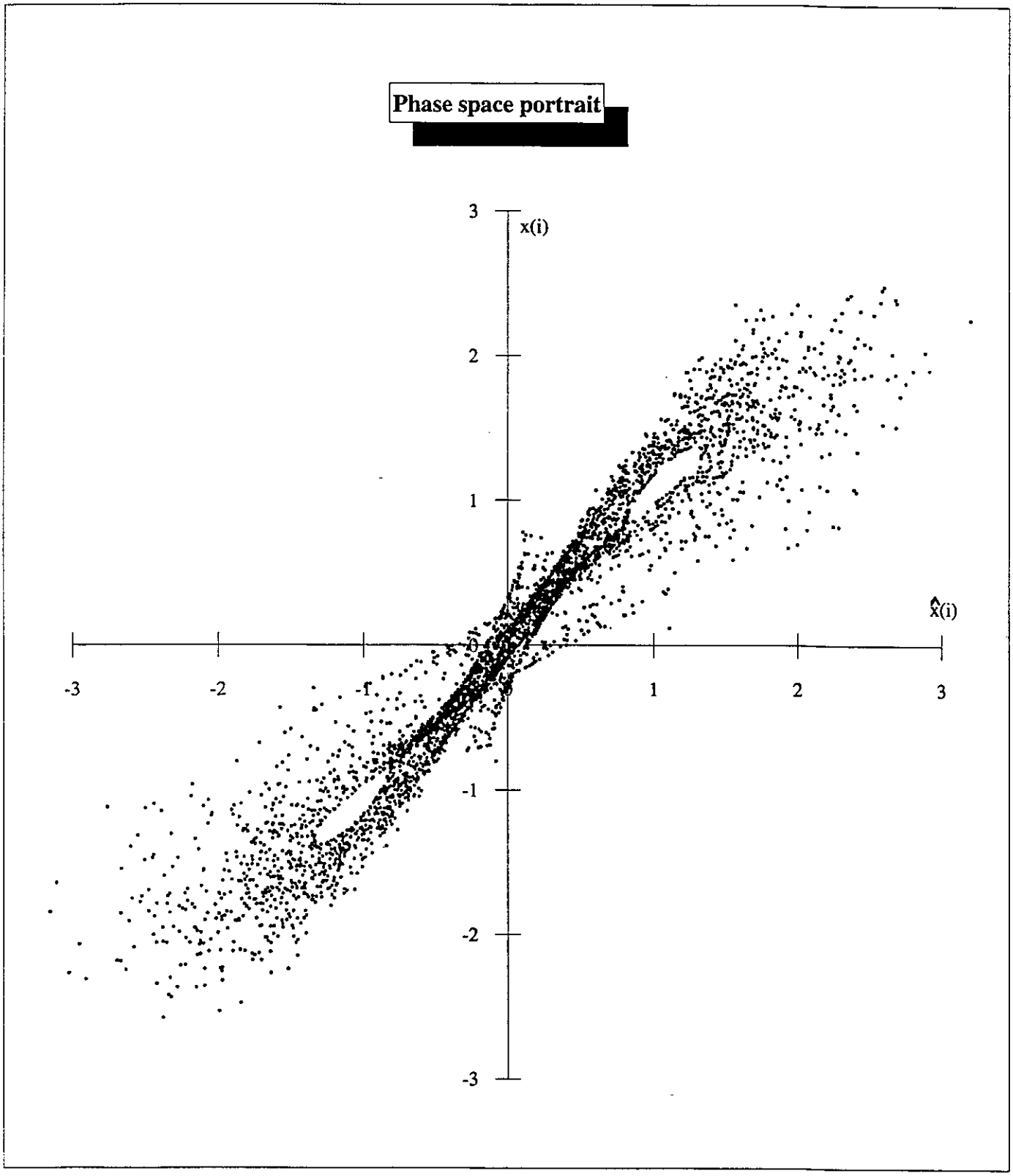

Figure 5

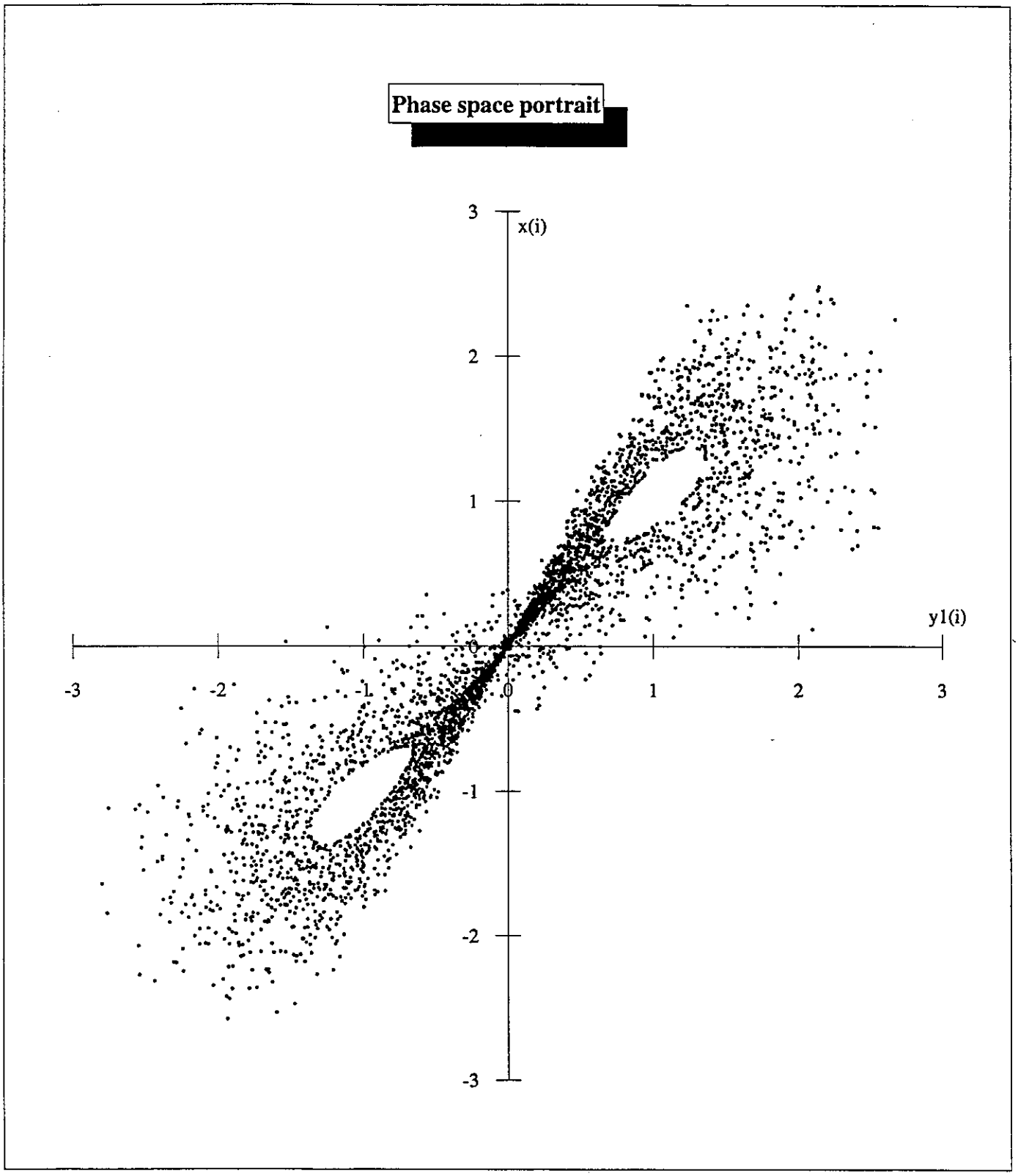

Figure 6

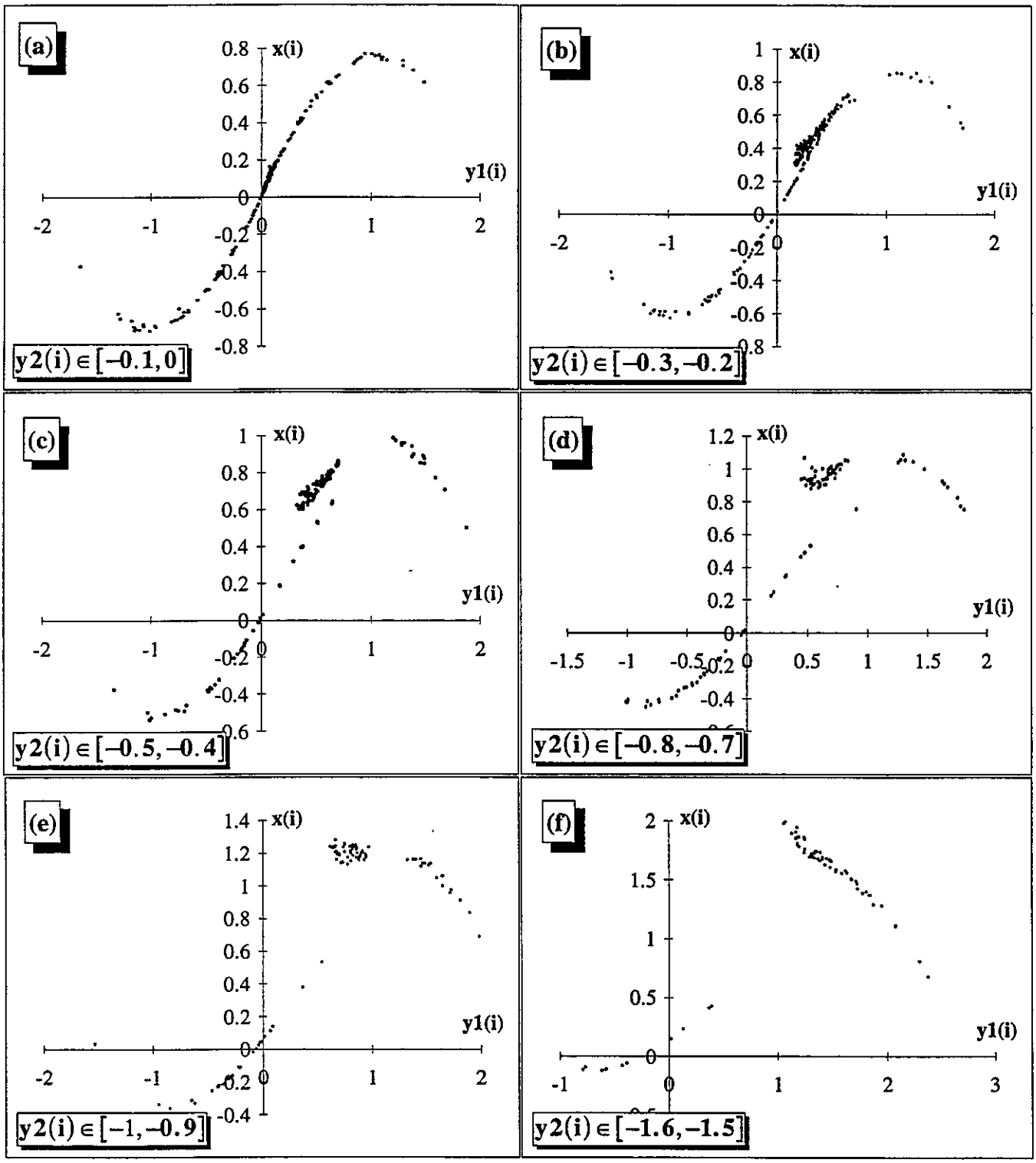
(a)
x(i)
y1(i)
y2(i) ∈ [−0.1, 0]
(b)
x(i)
y1(i)
y2(i) ∈ [−0.3, −0.2]
(c)
x(i)
y1(i)
y2(i) ∈ [−0.5, −0.4]
(d)
x(i)
y1(i)
y2(i) ∈ [−0.8, −0.7]
(e)
x(i)
y1(i)
y2(i) ∈ [−1, −0.9]
(f)
x(i)
y1(i)
y2(i) ∈ [−1.6, −1.5]

Figure 7

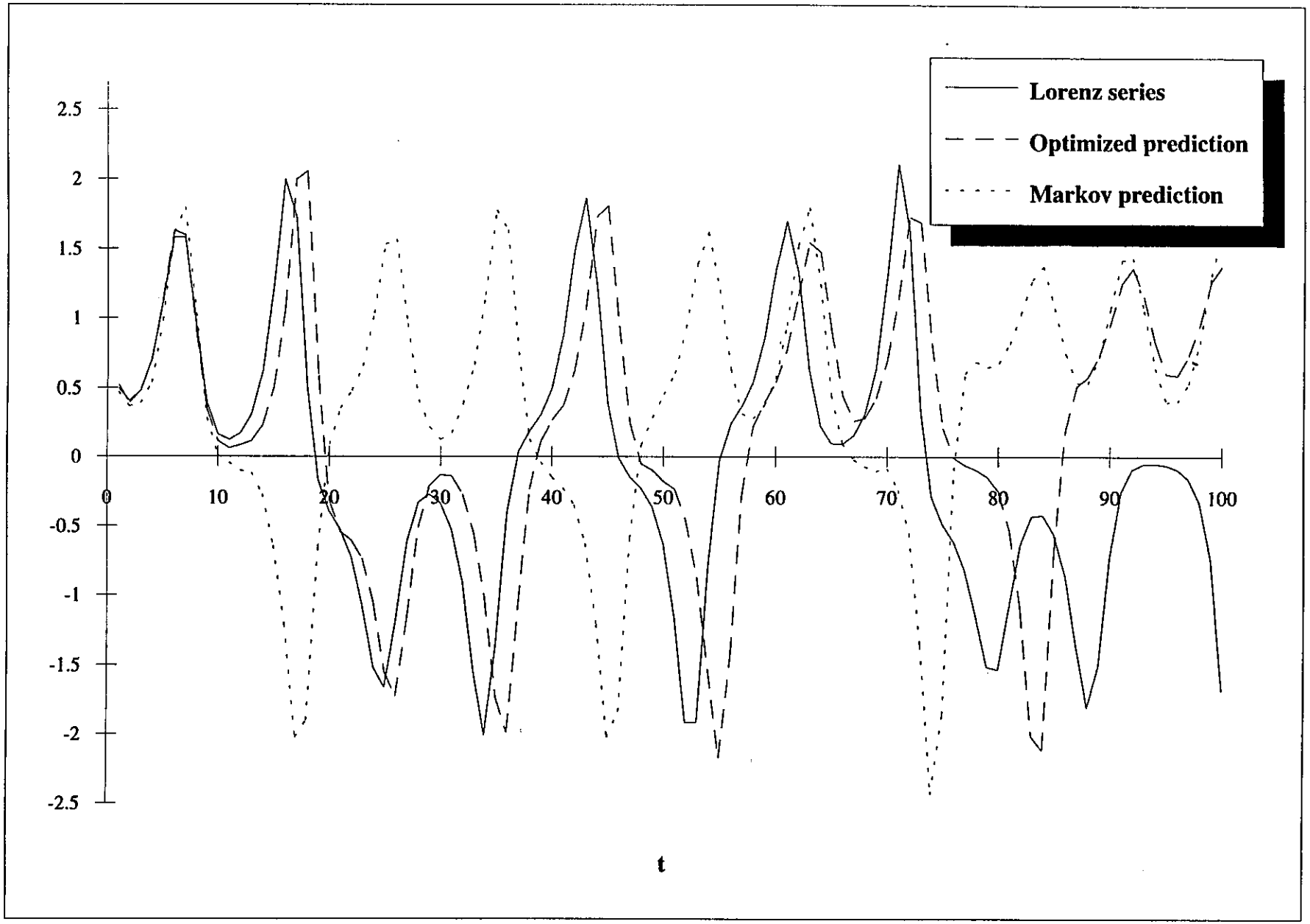
Lorenz series
Optimized prediction
Markov prediction
2.5
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-2.5
0
10
20
30
40
50
60
70
80
90
100
t

Figure 8

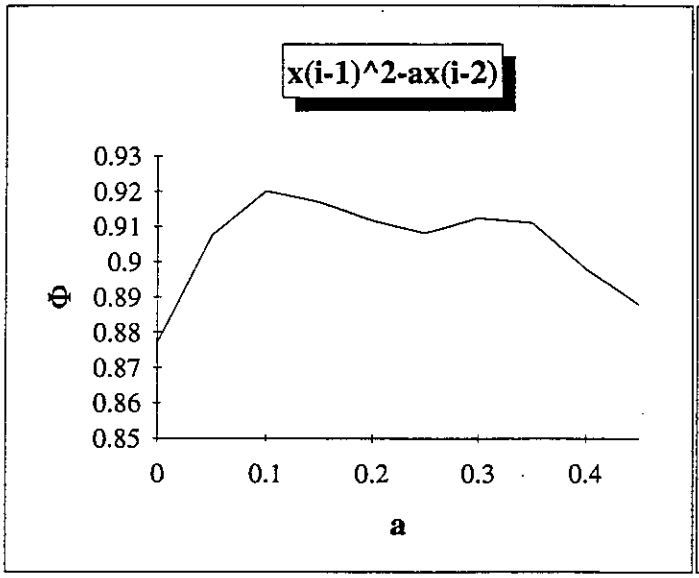

Figure 9

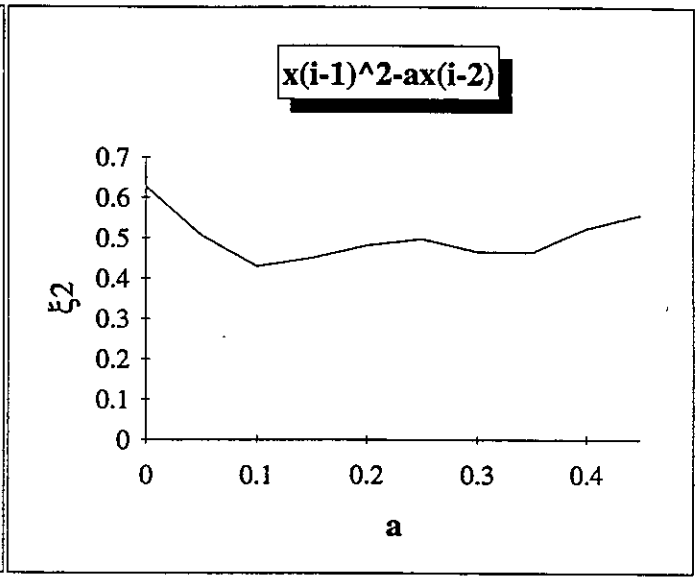

Figure 10

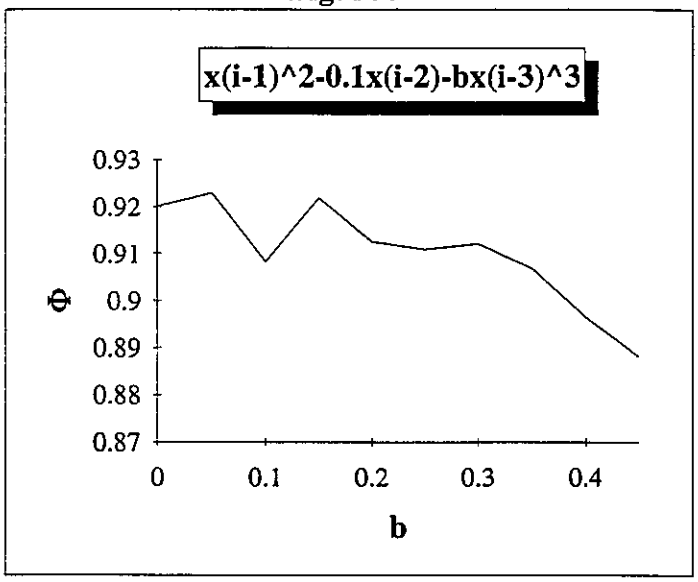

Figure 11

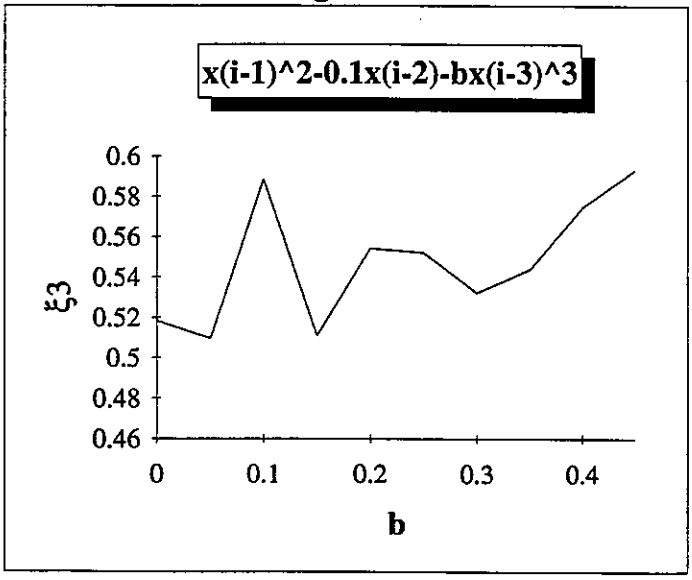

Figure 12

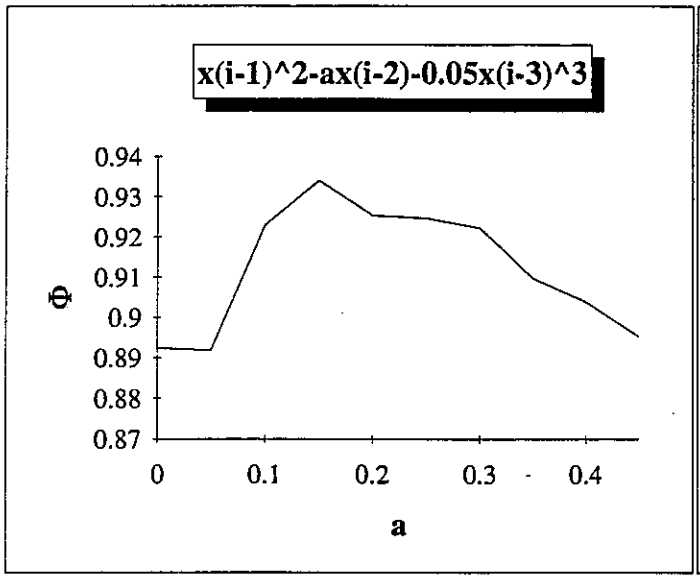

Figure 13

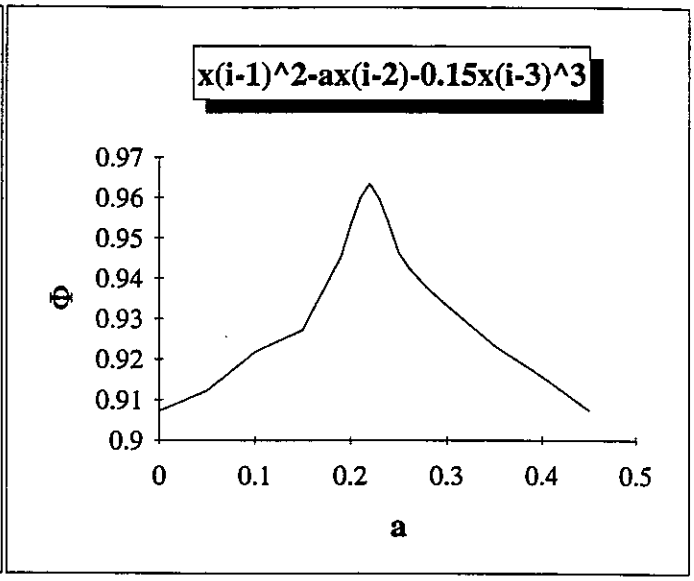

Figure 14

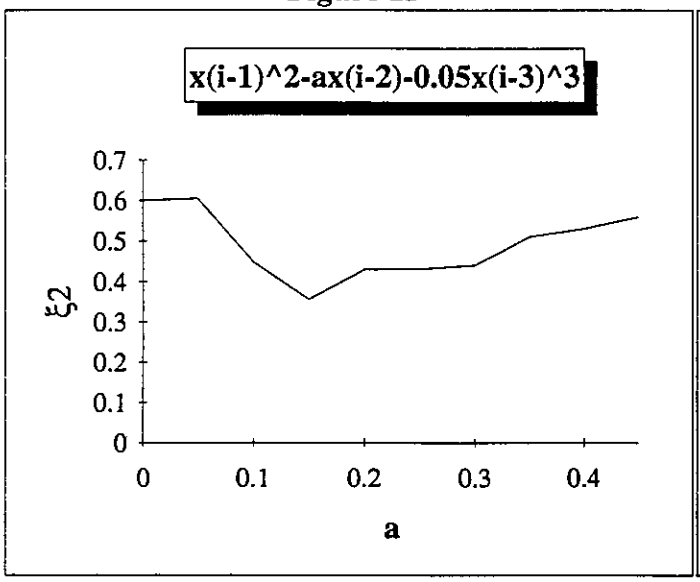

Figure 15

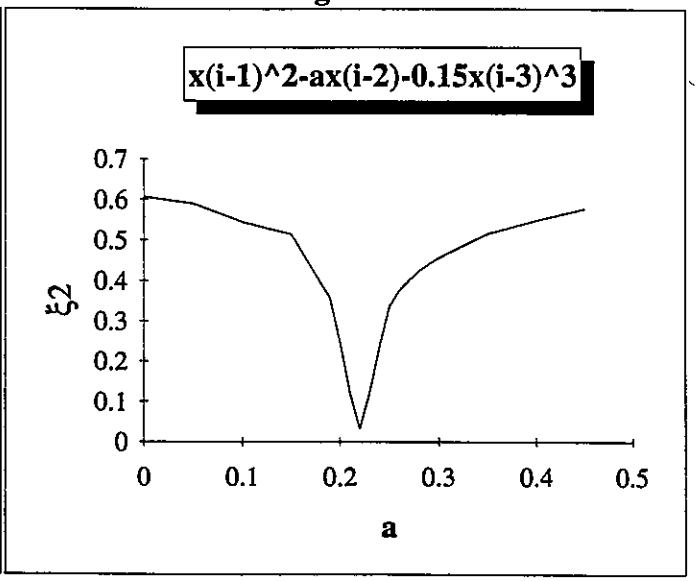

Figure 16

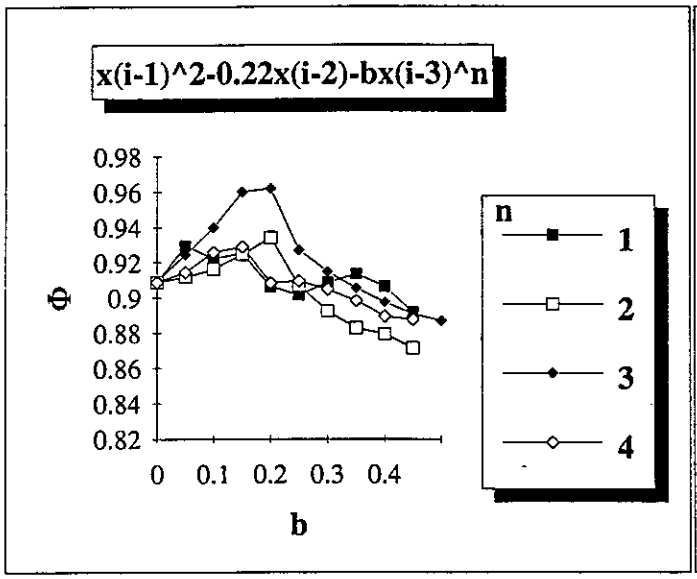

Figure 17

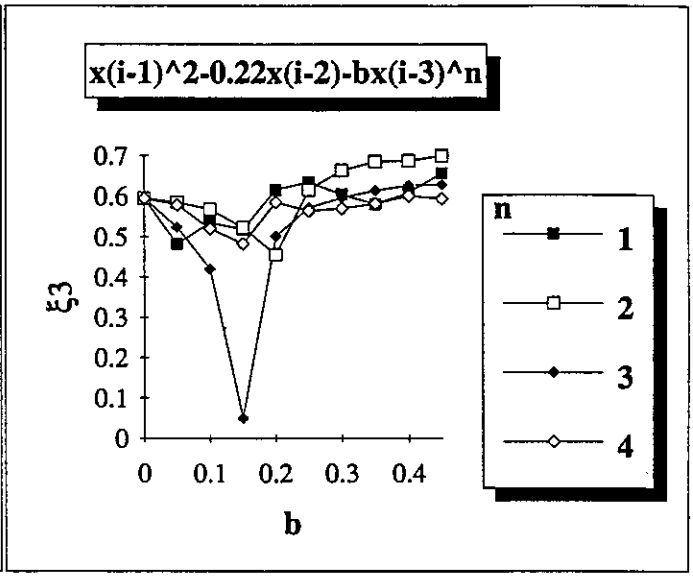

Figure 18

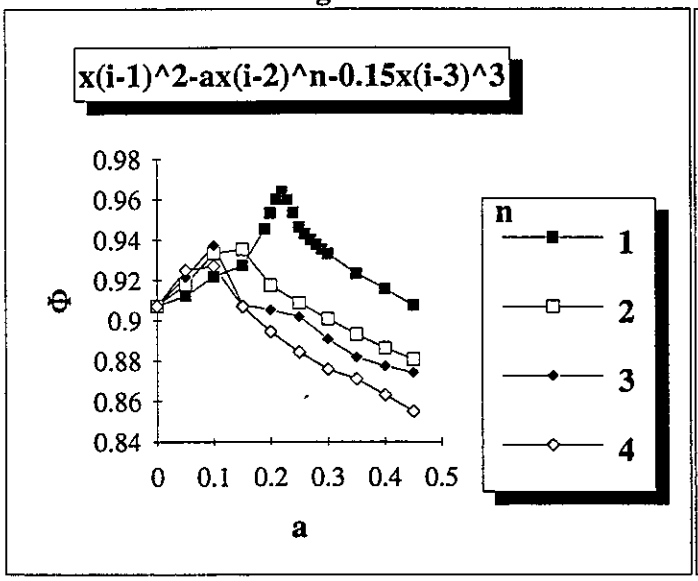

Figure 19

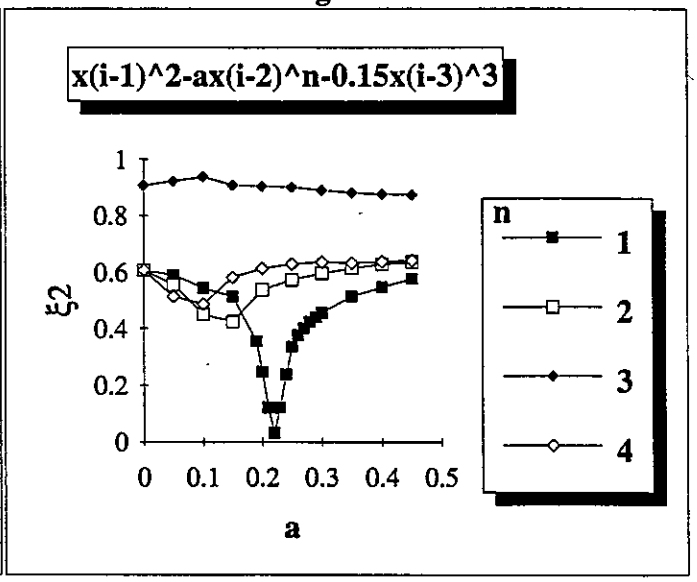

Figure 20