MODELING NONLINEAR TIME SERIES USING IMPROVED LEAST SQUARES METHOD

TOMOMICHI NAKAMURA* and MICHAEL SMALL
Department of Electronic and Information Engineering,
The Hong Kong Polytechnic University, Hung Hom,
Kowloon, Hong Kong
*entomo@eie.polyu.edu.hk

Received July 16, 2004; Revised February 28, 2005

We improve the least squares (LS) method for building models of a nonlinear dynamical system given finite time series which are contaminated by observational noise. When the noise level is low, the LS method gives good estimates for the parameters, however, the models selected as the best by information criteria often tend to be over-parameterized or even degenerate. We observe that the correct model is not selected as the best model despite belonging to the chosen model class. To overcome this, we propose a simple but very effective idea to use the LS method more appropriately. We apply the method for model selection. Numerical studies indicate that the method can be used to apply information criteria more effectively, and generally avoid over-fitting and model degeneracy.

Keywords: Description length; model degeneracy; model selection; least squares method; nonlinear time series modeling; over-fitting.

1. Introduction

We consider the problem of model selection for a nonlinear deterministic dynamical system given only a finite time series of observations contaminated by observational noise. A particularly convenient and general class of nonlinear models is the pseudo-linear models [Judd & Mees, 1995], which are linear combinations of arbitrary nonlinear functions. Pseudo-linear models have large numbers of parameters and, hence, the models must be tuned via parameter estimation. Usually, the least squares (LS) method is applied for the estimation. In practice, noisy data is used as a proxy of the true state (noise free data) in the method, because the true state is unknown. The LS method effectively assumes noise only affects the “response” variables at time \( t + 1 \) and not the “regressor” variables \( t \), which is clearly false. That is, the LS method is usually used \textit{inappropriately} in practice. We will describe this in more detail in the next section. As a result, despite the parameters being good estimates when the noise level is low, the models obtained often tend to be over-fitted, even if an information criterion such as description length is applied.

We demonstrate that the inappropriate application of least squares can cause the significant problem of model degeneracy. That is, the correct model is not selected as the best model from the viewpoint of information criteria, despite belonging to the chosen model class. We propose a simple but very effective idea to use the LS method more appropriately, which does not need any prior knowledge about the model and data nor require the use of any filter. The idea also has the added benefit of improving the application of information criteria for model selection. To demonstrate the problem of model degeneracy, we use an artificial system described by polynomial functions. However, for building models, it is usually not recommended to use only polynomial basis functions [Judd, 2003].
solves the optimization problem

\[ \text{model is identical to the system.} \]

is perfect, that is, there is a correct value of \( x \),

becomes colored regression noise in the regression noise. White observational noise at the output that LS is biased in the presence of observational assumptions hold and the noise is Gaussian and independent. Even when these which makes the assumption that the noise is fitting. This is a maximum likelihood method, where \( s \) is the true state (noise free data) at time \( t \).

Hence, as a more practical example, we build models using radial basis functions from artificial data.

2. The Least Squares Method and Information Criteria for Modeling

For building pseudo-linear models, we perform parameter estimation and model selection. For parameter estimation the LS method is applied, and for model selection an information criterion is applied. In Sec. 2.1 we describe how the LS method is applied in practice and the drawbacks, and how the LS method should be applied. In Sec. 2.2 we describe the purpose of using information criteria and description length which are adopted as an information criterion.

2.1. The least squares method

We now consider the problem of estimating the parameters \( \lambda \in \mathbb{R}^k \) of a model \( x_{t+1} = f(x_t, \lambda) \), \( x_t \in \mathbb{R}^d \) of a nonlinear deterministic dynamical system given only a finite time series of observations \( s_t \) contaminated by observational noise \( \epsilon_t \) (that is, \( s_t = x_t + \epsilon_t \)), where \( \epsilon_t \) is Gaussian with zero mean and the data comprise a set of \( n \) scalar measurements. We will assume that the model class \( f(x_t, \lambda) \) is perfect, that is, there is a correct value of \( \lambda \), where the model is identical to the system.

A commonly and practically used LS method solves the optimization problem

\[
\min_{\lambda} \sum_{t=1}^{n-1} ||s_{t+1} - f(s_t, \lambda)||^2, \tag{1}
\]

where only the noisy observations \( s_t \) are used for the fitting. This is a maximum likelihood method, which makes the assumption that the noise is Gaussian and independent. Even when these assumptions hold and the noise is white, it is known that LS is biased in the presence of observational noise. White observational noise at the output becomes colored regression noise in the regression equation which LS cannot handle.

The parameter estimates would be much less biased if we could solve the optimization problem

\[
\min_{\lambda} \sum_{t=1}^{n-1} ||s_{t+1} - f(x_t, \lambda)||^2, \tag{2}
\]

where \( x_t \) is the true state (noise free data) at time \( t \). But of course we cannot know \( x_t \). So in Eq. (1) noisy data \( s_t \) is used as a proxy for the noise free data \( x_t \). This is clearly not a good thing to do because \( s_t \) is corrupted by noise.

When using Eq. (1) for nonlinear models, the fitting error (in other words, prediction error) is usually not the same as observational noise and also not normally distributed, even if the observational noise is Gaussian. However, when using Eq. (2), the fitting error is almost identical to the observational noise and normally distributed. It should be noted that the normal distribution is very important because we usually use the assumption that the fitting error is Gaussian when using the LS method and deriving the information criteria formulae. For more details see [Ljung & Ljung, 1999].

We consider that the usage of Eq. (1) is “inappropriate” and that of Eq. (2) is “appropriate” as the LS method.

For finding the best model among many, some information criteria are usually employed. Hence, we briefly review the basic concept for information criteria and Rissanen’s Description Length modified by Judd and Mees (DL) [Judd & Mees, 1995], which we adopt here. Then, we show a significant example: when using Eq. (1), by applying DL, the correct model cannot be found as the best model, even if the noise level is low and the parameters estimated are good.

2.2. Information criteria

For building pseudo-linear models, the model size\(^1\) is important. Time series available to us are imperfect and insufficient, because they are usually contaminated by observational noise, and numerical accuracy is also limited. Using these unfavorable data, we have to build models, and we hope these models can reflect the underlying system and the influence of noise will be removed as much as possible. Hence, the model should not be fitted to the data either too closely or too poorly, this is called over-fitting and under-fitting, respectively. To find the model that best balances model error against model size so as to prevent over-fitting or under-fitting of the data, information criteria are often applied. Figure 1 shows the rough sketch of the relationship between model size and fitting error (prediction error) for a generic information criterion. Hence, it is considered that the minimum of the information criterion corresponds to the

\(^1\)For pseudo-linear models, the model size refers to the number of basis functions, which is the same as the number of parameters of the model, because the only parameters used to fit the data are the linear weights.
best (optimal) model size to be used and the smaller the value, the better the model.

Another important reason for using information criteria is to avoid unnecessary increase in the model size, which occurs when a model is built that follows a nested, self-iterated, form of the original system. Such models are called degenerate. A simple example is the following. Let the original model be

$$x_t = a_1 x_{t-1} + a_3 x_{t-3},$$  \hspace{1cm} (3)

which has a model size 2. This model can be rewritten as

$$x_{t-1} = a_1 x_{t-2} + a_3 x_{t-4},$$  \hspace{1cm} (4)

which is essentially the same as the original model. Using expression (4) to replace the basis function $(1/2)x_{t-1}$, for example, in the original model gives

$$x_t = \frac{1}{2} a_1 x_{t-1} + \frac{1}{2} a_1 x_{t-1} + a_3 x_{t-3}$$

$$= \frac{1}{2} a_1 x_{t-1} + \frac{1}{2} a_1 (a_1 x_{t-2} + a_3 x_{t-4}) + a_3 x_{t-3}$$

$$= \frac{1}{2} a_1 x_{t-1} + \frac{1}{2} a_1^2 x_{t-2} + a_3 x_{t-3} + \frac{1}{2} a_1 a_3 x_{t-4}$$  \hspace{1cm} (5)

Although model (5) is identical to the original model (3), its size is 4, which is larger than that of the original model. We refer to this kind of model as degenerate. If such an operation is done continuously, the model size increases infinitely. Hence, it is important to remove the above mentioned nesting effect and determine the smallest model size which can substantially model the system, because such models are clearly over-fitting.

Some information criteria have already been proposed for these purposes. The best known is the Akaike information criterion (AIC) [Akaike, 1974], but it is known to fail to provide statistically consistent estimates [Rissanen, 1989]. Hence, the criterion we use for determining the best (optimal) model is the Rissanen’s Description Length modified by Judd and Mees [1995]. The minimum description length (MDL) principle states that the best model is the one that attains the greatest compression. Under the assumption that the fitting error (prediction error) is normally distributed, Judd and Mees showed that the description length $DL_{\gamma}$ can be approximated by

$$DL_{\gamma}(k) = \left(\frac{n}{2} - 1\right) \ln \frac{e^T e}{n} + (k + 1) \left(\frac{1}{2} + \ln \gamma\right)$$

$$- \sum_{i=1}^{k} \ln \delta_i$$  \hspace{1cm} (6)

where $k$ is the model size, $\gamma$ is related to the scale of the data (see below), and the variables $\delta$ can be interpreted as the relative precision to which the parameters $\lambda$ are specified. See more details in [Judd & Mees, 1995].

A more precise interpretation of $\gamma$ is that it is the exponent in a floating point representation of the parameters scaled relative to some fixed amount, and it is supposed that $1 \leq \gamma \leq 32$. There are good arguments that $\gamma$ should be 1, but in practice larger values seem to be necessary to prevent over-fitting [Judd, 2003; Nakamura, 2003]. We use $\gamma = 32$ (that is, $DL_{32}$) throughout this paper.

Here, we note that we do not use DL as a strict criterion for model selections, rather we use it as a method for screening alternative formulations in order to produce a set of candidate models for further consideration.

### 3. Degeneracy and Model Selection

In this section, to investigate if the correct model is selected as the best model or not, we use a simple well-known polynomial model, the Henon map as a nonlinear AR model. We note that the “correct model” means that only all basis functions included in the actual system model are selected. Also, the “best model” means that the model has the smallest value of DL. We then expect that the DL becomes the smallest when the model built is the correct model. The levels of observational noise are 20 dB, 40 dB, 60 dB and 80 dB, the numbers of data points...
used are 1000 and 10 000, and Eq. (1) is used where only noisy data is used.

When building models, many candidate basis functions are usually prepared in the form of a dictionary, and the basis functions are selected from the dictionary by a selection method. Selecting the optimal subset of basis functions is typically an NP-hard problem which is usually solved by heuristic methods [Judd & Mees, 1995]. There is always an embarrassing problem in model selection, and this is the possibility of finding only a local minima. Generally speaking, it is very difficult to find a global best model. Hence, to obtain the truly best model, we calculate all possible combinations (in other words, an exhaustive search). However, it should be noted that such a procedure is only practical for fairly small dictionaries.

3.1. The Henon map as a nonlinear AR model

The Henon map [Henon, 1976]

\[ x_t = A_0 + A_2 x_{t-2} + A_4 x_{t-1}^2, \]
\[ s_t = x_t + \epsilon_t, \]  (7)

where \( A_0 = 1.0, A_2 = 0.3 \) and \( A_4 = -1.4 \) and \( \epsilon_t \) is Gaussian observational noise. We use \( s_t \) as observational data and will use time delay linear polynomial models. Choosing lag = 3 and degree = 3 gives 20 candidate basis functions in the dictionary. The basis functions corresponding to the parameters are, \( A_0 \): constant, \( A_1 \): \( x_{t-1} \), \( A_2 \): \( x_{t-2} \), \( A_3 \): \( x_{t-3} \), \( A_4 \): \( x_{t-1}^2 \), \( A_5 \): \( x_{t-2}^2 \), \( A_6 \): \( x_{t-3}^2 \), \( A_7 \): \( x_{t-1}^3 \), \( A_8 \): \( x_{t-2}^3 \), \( A_9 \): \( x_{t-3}^3 \), \( A_{10} \): \( x_{t-1} x_{t-2} \), \( A_{11} \): \( x_{t-1} x_{t-3} \), \( A_{12} \): \( x_{t-2} x_{t-3} \), \( A_{13} \): \( x_{t-2}^2 x_{t-3} \), \( A_{14} \): \( x_{t-1}^2 x_{t-3} \), \( A_{15} \): \( x_{t-2}^2 x_{t-1} \), \( A_{16} \): \( x_{t-3}^2 x_{t-1} \), \( A_{17} \): \( x_{t-2}^3 x_{t-1} \), \( A_{18} \): \( x_{t-3}^2 x_{t-2} \) and \( A_{19} \): \( x_{t-1} x_{t-2} x_{t-3} \). From the dictionary we can build a model \( x_{t-1} = A_0 + A_2 x_{t-3} + A_4 x_{t-2}^2 \), from which we can build the degenerate models.

Table 1 shows the model size of the best model selected at different noise levels when using 1000 and 10 000 data points. In all cases, \( DL_{32} \) (description length with \( \gamma = 32 \)) is not the smallest when the model size is 3 (the correct model size). However, it should be noted that the correct model is selected at the correct size. That is, the correct model is not the best model (i.e. not the MDL-best model).

Figure 2 shows the results when the noise levels are 20 dB and 40 dB, and the number of data points is 10 000. When the noise level is 20 dB, \( DL_{32} \) is the smallest at model size 10, and when the noise level is 40 dB, \( DL_{32} \) is the smallest at model size 8. It should be noted that this is not a particular phenomenon of \( DL_{32} \). Although we show the results using only \( DL_{32} \), we note that the results using other information criteria such as SIC [Schwarz, 1978] and NML [Rissanen, 2000] are essentially the same.

This is an important issue because we cannot know the correct model even if the model is selected at the correct size, and the best model is always larger than the correct model, which means that the best model is over-fitted.

It is usually considered that larger incorrect models may predict a given noisy time series more effectively than the correct model, because they exploit correlation or information contained in the noise, and larger incorrect models based on higher dimensional embedding may benefit from some additional averaging effect. That is, the best model is the best predictor of the time series, while the correct model may not be the best predictor (when the data used are noisy). See more details in Sec. 3.2. As this is a general phenomenon, the reason that the size of the best models was larger than that of the correct model is likely due to the high noise level.

If the above mentioned reason is true, the correct model should be selected as the best model when the noise level is low. However, even when the noise levels are 40 dB, 60 dB and 80 dB (that is, when the noise level is lower), the correct model was not selected as the best model in all cases. The size of the best model selected is larger than that of the correct model. These results imply that this phenomenon is more complicated than simply reducing the influence of noise.

We investigate this phenomenon. The formula obtained as the best model when the noise level is 40 dB and the number of data points is 10 000 is

\[ x_t = 0.7781 - 0.4539 x_{t-1} + 0.3004 x_{t-2} - 0.0663 x_{t-3} - 0.7235 x_{t-1}^2 + 0.3097 x_{t-2}^2 - 0.2032 x_{t-1} x_{t-3} + 0.9460 x_{t-1} x_{t-2}^2. \]  (8)
Fig. 2. A plot of description length against size of model for the Henon map when the noise levels are 20 dB and 40 dB and the number of data points is 10,000. Panels (a)–(c) are the results when the noise level is 20 dB, where (b) and (c) are enlargements of (a) framed by the dashed lines. Panels (d)–(f) are the results when the noise level is 40 dB, where (e) and (f) are enlargements of (d) framed by the dashed lines.
This is, in fact, a degenerate,\textsuperscript{2} approximation to the correct model. In the Appendix, we will show the model can be reduced to essentially the correct model. When the noise levels are 40 dB, 60 dB and 80 dB, we can find such good but degenerate approximations\textsuperscript{3} of sizes 6, 8 and 11. When the noise

\textsuperscript{2}We note that degeneracy does not always occur. When the model is a linear AR model [Small & Judd, 1998] where there are ten candidate basis functions, constant, \(x_{t-1}, x_{t-2}, \ldots, x_{t-9}\) in the dictionary, we can build degenerate models. The correct model is selected as the best model. That is, we can avoid unnecessary increase in the model size.

\textsuperscript{3}The parameters of basis functions selected in the degenerate models are, model size 6: \((A_0, A_1, A_2, A_4, A_{11}, A_{15})\), model size 8: \((A_0, A_1, A_2, A_3, A_4, A_5, A_{11}, A_{15})\), and model size 11: \((A_0, A_1, A_2, A_3, A_4, A_5, A_8, A_{10}, A_{11}, A_{12}, A_{15})\).
level is 20 dB, we also find degenerate models when the model sizes are 6 and 8. Furthermore, as Table 1 shows, these degenerate models are selected as the best models in most cases.

3.2. Some observations about degeneracy

The problems highlighted in the previous section indicate that the best one-step-predictor is not necessarily the perfect model. That is, even with a perfect model class, one can obtain a MDL model which is larger than the perfect model [Judd, 2003].

Equation (7) of the Henon map gives the identities

\[
\begin{align*}
x_{t-1} &= A_0 + A_2 x_{t-3} + A_4 x_{t-2}^2, \quad (9) \\
x_{t-2} &= A_0 + A_2 x_{t-1} + A_4 x_{t-2}^2, \quad (10)
\end{align*}
\]

where \(A_0 = 1.0, A_2 = 0.3\) and \(A_4 = -1.4\). If the \(x_t\) are the noisy data, the original equation and the identities would be no longer satisfied exactly for any \(t\). Hence, the information on the right-hand side of Eqs. (9) and (10) could be used to obtain a better estimate of the true value of \(x_{t-1}\) and \(x_{t-2}\) in the original model, Eq. (7), for example, by averaging these calculated estimates with the observed values. These better estimates could be used in the original model to obtain better predictions of \(x_t\) from past data. Hence, this operation gives a larger model which is a better predictor than the correct model when using noisy data. Smaller DL (information criteria) automatically implies a smaller prediction error (fitting error) and hence a degenerate model of larger size than the correct model must use additional basis functions to decrease the prediction error [Judd, 2003].

These difficulties can be circumvented somewhat if noise free data could be used for system state. Unfortunately, in practice, this is not possible and so in Sec. 4 we suggest an alternative and quite effective technique to generate better proxy for the state.

3.3. When using noisy and noise free data

Here, we use the previous true state (noise free data) and the current noisy datum to investigate the performance of the optimization problem, Eq. (2), with noise levels 20 dB, 40 dB, 60 dB and 80 dB and number of data points 1000 and 10,000, where the best models selected were the degenerate models in most cases when using Eq. (1). We again calculate all possible combinations to obtain the truly best model.

In all cases the correct model is selected as the best model (MDL model). Figure 3 shows the results when the noise levels are 20 dB and 40 dB, and the number of data points is 10,000. The figure shows that there is no local minima and \(DL_{32}\) is the smallest when the model size is 3 which is the correct model size. This indicates that it is very useful to use the noise free data for not only parameter estimation but also to take advantage of information criteria more effectively.

However, we must usually use the noisy data as proxy for the noise free data, it is clear that it is very difficult work to obtain the true state from noisy data. In the next section, we suggest a technique to address these problems.

4. An Idea to Use the Least Squares Method More Appropriately

Sections 3.1 and 3.3 show some problems when applying the LS method and that noise in the time series requires modeling and this cannot be accomplished by the LS method. An alternative is to model the noise and use adequate estimators, for example, generalized LS (GLS). The method is introduced to improve upon parameter estimation efficiency. However, a drawback of the GLS method is that it is difficult to implement. More detailed discussions of the GLS theory can also be found elsewhere [Amemiya, 1985; Greene, 2000].

In Sec. 4.1 we describe a basic idea to use the LS method more appropriately without using true state and theoretical analysis, where we call the method the noisy least squares (NLS) method. In Sec. 4.2 we apply the NLS method to the Henon map again used in Sec. 3.1 and show the NLS method works well.

4.1. How to improve the least squares method

As highlighted in the previous section, it is very useful to use the true state (noise free data) for the LS method to avoid over-fitting (degeneracy) (in other words, to take advantage of information criteria effectively). To obtain the estimates of true state from the noisy state for nonlinear models, there are a number of methods, for example, the extended Kalman filter [Walker & Mees,
We can consider three cases:

- When $\frac{\sigma^2_{t+1}}{\epsilon^2_t} \gg \frac{\sigma^2_{t+1}}{\epsilon^2_{t+1}}$ (this is a case where the additional noise level is too large), the situation becomes worse, because $\epsilon_{t+1}$ dominates Eq. (12), where $E$ is the expectation.

- When $\frac{\sigma^2_{t+1}}{\epsilon^2_{t+1}} \approx \frac{\sigma^2_{t+1}}{\epsilon^2_{t+1}}$ (this is a case where the level of the additional noise is almost the same as that of noise in the original data), Eq. (12) becomes $s_{t+1}^* - f(x_t) + \eta_t$, where $E(\eta_t^2) \approx E(\epsilon_{t+1}^2)$. Hence, almost the same situation of Eq. (2) can be obtained.

Hence, when the level of the additional noise is appropriate, we expect Eq. (11) to be a good approximation to Eq. (2). We refer to the method as the “noisy least squares (NLS)” method. Equation (12) indicates that Gaussian noise is not the only possible noise. However, we usually assume that the observational noise is Gaussian and we also usually apply the LS method on the assumption that the noise is Gaussian and independent to achieve maximum likelihood. Hence, we add Gaussian noise in the NLS method.

Table 2 shows the parameter estimates using the NLS method (that is, Eq. (11)) for the Henon map at different additional noise levels, 0 dB, 20 dB, 40 dB, 60 dB and 80 dB (that is, $\epsilon_{t+1}$), where the number of data point used is 1000 and the original noisy data (that is, $s_t$) is the noise free data contaminated by 40 dB noise. This result shows that the NLS method does not always give worse estimates for the parameters. Especially, the method gives the best estimates when the additional noise level is 40 dB, which is the same noise level as that in the original data. However, we also note that the difference is not so significant, because the parameters estimated by the LS method are good estimates. This indicates that we can still use the parameters estimated by

The situation depends on the additional noise level. We can consider three cases:

- When $E(\epsilon_{t+1}^2) \gg E((DF)\epsilon_t^2)$ (this is a case where the additional noise level is too large), the situation becomes worse, because $\epsilon_{t+1}$ dominates Eq. (12), where $E$ is the expectation.

<table>
<thead>
<tr>
<th>Additional Noise Level</th>
<th>$A_0$</th>
<th>$A_2$</th>
<th>$A_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 dB</td>
<td>1.04660051</td>
<td>0.25949298</td>
<td>-1.48132109</td>
</tr>
<tr>
<td>20 dB</td>
<td>1.00389082</td>
<td>0.29662514</td>
<td>-1.40752412</td>
</tr>
<tr>
<td>40 dB</td>
<td>0.99961985</td>
<td>0.30033836</td>
<td>-1.39984743</td>
</tr>
<tr>
<td>60 dB</td>
<td>0.99919276</td>
<td>0.29662514</td>
<td>-1.39910676</td>
</tr>
<tr>
<td>80 dB</td>
<td>0.99915005</td>
<td>0.30074681</td>
<td>-1.39903269</td>
</tr>
<tr>
<td>Original noisy data</td>
<td>0.99914530</td>
<td>0.30075093</td>
<td>-1.39902446</td>
</tr>
</tbody>
</table>

We propose an idea to use the LS method more appropriately like Eq. (2) without using the true state. The method is very simple and we do not need any prior knowledge for the dynamics and the model. However, we do still have to assume that the observational noise is Gaussian. The reason is that when we apply the LS method and information criteria, we always assume that the noise is Gaussian. However, our method should work well even if the noise is not Gaussian (see below).

When the observational noise level is low, the noisy data clearly can be regarded as a good proxy for the true state. Also, in Eq. (2), the $s_{t+1}$ term has more noise than $x_t$. Hence, we consider that we can achieve a proxy of Eq. (2) using only noisy data when the noise in the part of $s_t$ is low compared to that of $s_{t+1}$ in Eq. (1). Then, we propose to add noise to the part of $s_{t+1}$.

Let the additional noise be $\epsilon_t$ and $s'_{t+1} = s_{t+1} + \epsilon_{t+1}$. Then, we obtain the new optimization problem

$$
\min_\lambda \sum_{t=1}^{n-1} \| s'_{t+1} - f(s_t, \lambda) \|^2. 
$$

(11)

This idea can be supported in a theoretical sense as follows.

$$
x_{t+1} = f(x_t),
$$

$$
f(s_t) = f(x_t + \epsilon_t) \approx f(x_t) + (DF)\epsilon_t,
$$

where $DF$ is the Jacobian matrix. Hence, we can obtain

$$
s_{t+1} - f(s_t) \approx s_{t+1} - \{f(x_t) + (DF)\epsilon_t\}.
$$

We add noise $\epsilon_t$, then we obtain

$$
s'_{t+1} - f(s_t) \approx s_{t+1} + \epsilon'_{t+1} - \{f(x_t) + (DF)\epsilon_t\},
$$

$$
= (s_{t+1} - f(x_t)) + \{\epsilon'_{t+1} - (DF)\epsilon_t\}. 
$$

(12)

The situation depends on the additional noise level. We can consider three cases:

- When $E(\epsilon_{t+1}^2) \gg E((DF)\epsilon_t^2)$ (this is a case where the additional noise level is too large), the situation becomes worse, because $\epsilon_{t+1}$ dominates Eq. (12), where $E$ is the expectation.

- When $E(\epsilon_{t+1}^2) \approx E((DF)\epsilon_t^2)$ (this is a case where the level of the additional noise is almost the same as that of noise in the original data), $E(\epsilon_{t+1}^2)$ becomes $\frac{s_{t+1} - f(x_t) + \eta_t}{\epsilon_t}$, where $E(\eta_t^2) \approx E(\epsilon_{t+1}^2)$. Hence, almost the same situation of Eq. (2) can be obtained.

Hence, when the level of the additional noise is appropriate, we expect Eq. (11) to be a good approximation to Eq. (2). We refer to the method as the “noisy least squares (NLS)” method. Equation (12) indicates that Gaussian noise is not the only possible noise. However, we usually assume that the observational noise is Gaussian and we also usually apply the LS method on the assumption that the noise is Gaussian and independent to achieve maximum likelihood. Hence, we add Gaussian noise in the NLS method.
the LS method when the observational noise level is low. We expect that the NLS method is rather effective for model selection. In the next section, we apply the NLS method to the same example used in Sec. 3.1 for model selection, where the correct model was not selected as the best model.

Here, to distinguish clearly between the best models obtained by applying the LS and NLS methods, we refer them as “the LS-best model” and “the NLS-best model”, respectively.

4.2. Verification of model selection using the case of degenerate models

Consider the Henon system again used in Sec. 3.1. We apply the NLS method for system identification.

![Graphs](image1.png)

**Fig. 4.** A plot of description length against size of model for the Henon map when the noise level in the original noisy data used is 40 dB and the number of data points used is 10 000. Panels (a) and (b) are the results when the additional noise level is 40 dB for the NLS method, where (b) is an enlargement of (a) framed by the dashed lines. Panels (c) and (d) are the results when the additional noise level is 10 dB for the NLS method, where (c) is an enlargement of (d) framed by the dashed lines.
Fig. 5. A plot of the selected basis functions in the NLS-best model against additional noise levels where the noise level in the original noisy data used is 40 dB and 60 dB, the number of data points used is 1000 and 10,000, and the number in the parentheses is the number of parameters selected (the model size) in the NLS-best model. Panel (a) 40 dB and 1000 data points, (b) 40 dB and 10,000 data points, (c) 60 dB and 1000 data points, and (d) 60 dB and 10,000 data points.
We again calculate all possible combinations to obtain the truly best model. We increase the additional noise level every 10 dB up to 0 dB from 80 dB.

It should be noted that when we create noisy data from noise free data, where the noisy data is used as an observational data, we calculate the additional noise levels relative to the noise free data. We call such noisy data “the original noisy data”. But to create noisy data for applying the NLS method, we calculate the additional noise levels using the original noisy data, because we do not know the noise free data in practice. Hence, even if we use “60 dB” for the additional noise level to the original noisy data and the noise level included in this data, the level itself for each data is not the same. Actually, the level of additional noise is slightly larger than that in the original noisy data.

Figure 4 shows typical behavior of $DL_{32}$ against model size, where the noise level included in the original noisy data used is 40 dB, the number of data points used is 10,000, and the additional noise levels are 10 dB and 40 dB for the NLS method. Panels (a) and (b) are similar to Figs. 2(d) and 2(e), where Eq. (1) was used (that is, only noisy data was used for the LS method) and the local minima in the description length corresponded to degenerate models. However, panels (c) and (d) are similar to Fig. 3, where Eq. (2) was used (that is, noisy and noise free data were used for the LS method).

Figure 5 simultaneously shows the basis functions in the NLS-best models and the additional noise level, where Y-axis index corresponds to that of the parameters given in Sec. 3.1. The figure indicates that the selected basis functions do not change when the additional noise level is lower than that included in the original noisy data. However, as the noise level becomes larger, the basis functions selected are identical to those in the correct model: larger than 40 dB in panel(a), 20 dB in panel(b), 60 dB in panel(c) and 40 dB in panel(d). Although we show only four cases in Fig. 5, we obtain essentially the same results for other conditions except for the case when the original data noise level is 20 dB and the number of data points used is 10,000.

We expect that the NLS-best model can be improved as the additional noise level increases within some intervals. As a matter of fact, the NLS-best model is always the correct model except when the additional noise level is over 0 dB, the noise level in the original data is 20 dB and the number of data points used is 10,000. Also, when the number of data points is 1000, the correct model is selected when the additional noise level is larger than 10 dB, which is large noise. The reason is that over 0 dB noise is too much and the data contaminated by 20 dB noise is not a good proxy for the true state.

These results indicate that when the observational noise level is low, by applying the NLS method, we can avoid unnecessary increase in the model size and some kind of over-fitting, and we can select the correct model. That is, by applying the NLS method we can take advantage of information criteria more effectively.

5. Application

The NLS method can be used to avoid degeneracy and to select the correct model. However, the examples presented so far are somewhat unrealistic. Firstly, correct basis functions for the system are in the dictionary. When using real-world time series, we usually do not know the system or the model. Hence, all models we can build are not true [Mees, 2000]. Secondly, when building models, we usually cannot calculate all possible combination sets (an exhaustive search) to obtain the truly best model. We usually apply a method for selecting basis functions from a dictionary [Judd & Mees, 1995; Nakamura et al., 2003, 2004]. Thirdly, the basis functions (and correct models) used were all polynomial. It is not recommended to build models using only polynomials [Judd, 2003]. Hence, we investigate how the NLS method works in practical cases.

In practice, for building models, radial basis functions (RBF) are often used for building pseudolinear models. For building good or strong models, we should use basis functions which can extract the peculiarities of the time series as much as possible. However, there is no simple way to generate such best basis functions and to know the best combination set of basis functions for a model beforehand. Hence, many candidate basis functions are generated. We hope that some of the candidate basis functions will be able to describe any peculiarity and nonlinearity of the data well. One technique for selecting basis function is a greedy method, for example, orthogonal least squares method [Chen et al., 1991]. A more powerful method is described by Judd and Mees [1995], who developed an

---

4We call these basis functions, although they need not actually form a basis.
algorithm for selecting a good parsimonious sub-set from a large dictionary of basis functions. The method has proven to be effective [Judd & Mees, 1995, 1998; Small & Judd, 1998a, 1998b] in modeling nonlinear dynamics. The improved method is proposed in [Nakamura et al., 2003, 2004] and we adopt this method; the up-and-down method.

In the earlier examples, we could always obtain clear results for any noise level even when the noise level in the original data was 20 dB, which is a relatively large noise level. This was possible, because there were the correct basis functions in the dictionary and all possible combination sets were always calculated. In practice, as we have to select a basis function from a dictionary using a selection method, it can be assumed that this selected function comes under the influence of the features of training data. A general explanation for selecting basis functions is that they should be adequate for extracting the features of the training data, when the training data is very clean. However, when this is not the case, such basis functions would not be selected, and those reflecting noise would be added. That is, noisy data is not appropriate as training data for building models, and we should avoid the influence of noise in training data for selecting basis functions as much as possible. We wish to use clean data in practice.

Hence, to save time and trouble, we will use the following idea to apply the NLS method.

1. We build models using the LS method, a selection method and training data as usual.
2. Keep using the model obtained (that is, the same basis functions selected) at each model size in step 1, $DL_{32}$ of these models is calculated again using the fitting error obtained by applying the NLS method.
3. We find the NLS-best model at each additional noise level.

To investigate the quality and performance of models obtained by applying the idea, we examine long-term free-run data, because one needs to get the dynamics right to obtain good long-term free-run data.

It should be noted that we apply the NLS method for model selection not for parameter estimation. We have shown in Table 2 that the NLS method can improve parameter estimates when the additional noise is adequate. However, we do not know the correct values of parameters in practice. Also, when the noise level is low, the LS method, Eq. (1), can give good parameter estimates. Hence, the models for generating free-run data are models obtained when using the original training data. That is, parameters used are with the LS method.

We apply the idea to an artificial data set generated by the differential equation of Chua’s circuit [Matsumoto et al., 1985]. The data is well known as an archetypal nonlinear time series and hence, often used for modeling. See elsewhere [Timmer et al., 2000; Aguirre et al., 2001].

5.1. The differential equation of Chua’s circuit

This model is a electronic circuit proposed by [Matsumoto et al., 1985]. The circuit dynamics is described by

$$\begin{align*}
C_1 \frac{dv_1}{dt} &= g(v_2 - v_1) - g(v_1) \\
C_2 \frac{dv_2}{dt} &= g(v_1 - v_2) + i_L \\
L \frac{di_L}{dt} &= -v_2
\end{align*}$$

(13)

where $v_{c1}$, $v_{c2}$ and $i_L$ denote the voltage across $C_1$, the voltage across $C_2$ and the current through $L$, respectively, and $g(v)$ is the piecewise-linear function:

$$g(v) = m_0 v + \frac{1}{2} (m_1 - m_0) |v + B_p|$$

$$+ \frac{1}{2} (m_0 - m_1) |v - B_p|.$$  

(14)

The parameters we use for both the equations are $1/C_1 = 9.0$, $1/C_2 = 1.0$, $1/L = 7.0$, $G = 0.7$, $m_0 = -0.5$, $m_1 = -0.8$ and $B_p = 1.0$. When calculating this model we use the fourth order Runge-Kutta method with sampling interval 0.1, and we use the $v_{c1}$ component of the model. We contaminate it with 60 dB Gaussian noise and use it as observational data. The radial basis functions used are Gaussian basis functions $f_i(x) = \exp(-(1/2)\|x - c_i\|^2/r_i^2)$, where $c_i$ are called centers and $r_i$ are the radii. Many basis functions are generated by random assignment of the centers and radii using time series data points to form a dictionary [Judd & Mees, 1995, 1998].

For building a model using radial basis functions, 5000 data points are used as the training data and the up-and-down method using marginal error is applied [Nakamura et al., 2003, 2004]. The data is
embedded using \( (t - 1, t - 5, t - 9) \) in three dimensions [Takens, 1981] with the aim of predicting a value at time \( t \). The time-lag is equal to one-quarter of the major period (17 points) [Atten et al., 1984] and the embedding dimension is selected by applying the false nearest neighbor method [Kennel et al., 1992].

When using the original noisy data, the size of the LS-best model is 67. Figure 6 shows the training data (that is, 60 dB noisy data) of the circuit equations, a free-run time series of the model, and their corresponding reconstructed attractors. The figure shows that although there are some small differences, they are very similar, indicating that the model obtained can get the dynamics almost right.

From our knowledge of degeneracy, we expect the LS-best model of size 67 is over-fitted. Hence, we apply the NLS method, where we use five different Gaussian noise realizations in the NLS method. We add noise from 80 dB every 5 dB, calculated using the data. Then, we expect to find a better model. Table 3 shows the mode of the NLS-best model size at each additional noise level. As the noise level increases, the size becomes smaller. When the noise levels are 55 dB and 50 dB, the model size is 63, which is 4 model sizes smaller than the original size, size 67. As described in Sec. 4.1, the NLS method

![Fig. 6. Time series of training data of the Chua's circuit equations and a free-run of the LS-best model, along with the reconstructed attractors of these time series. In panels (c) and (d), 5000 data points are plotted. Panel (a) training data, (b) free-run data, (c) reconstructed attractor of training data and (d) reconstructed attractor of free-run data.](image-url)
Table 3. The mode of the NLS-best model size obtained at each additional noise level.

<table>
<thead>
<tr>
<th>Additional Noise Level</th>
<th>Model Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>20 dB</td>
<td>4</td>
</tr>
<tr>
<td>25 dB</td>
<td>6</td>
</tr>
<tr>
<td>30 dB</td>
<td>14</td>
</tr>
<tr>
<td>35 dB</td>
<td>20</td>
</tr>
<tr>
<td>40 dB</td>
<td>39</td>
</tr>
<tr>
<td>45 dB</td>
<td>53</td>
</tr>
<tr>
<td>50 dB</td>
<td>63</td>
</tr>
<tr>
<td>55 dB</td>
<td>63</td>
</tr>
<tr>
<td>60 dB</td>
<td>64</td>
</tr>
<tr>
<td>65 dB</td>
<td>67</td>
</tr>
<tr>
<td>70 dB</td>
<td>67</td>
</tr>
<tr>
<td>75 dB</td>
<td>67</td>
</tr>
<tr>
<td>80 dB</td>
<td>67</td>
</tr>
<tr>
<td>Original model</td>
<td>67</td>
</tr>
</tbody>
</table>

In Fig. 8, we plot the Fourier power spectrum of the original training data and the free-run data of the models obtained, where the number of data points used is 4096. When the frequency is smaller than 100, all the power spectra of free-run data are very similar. However, as the frequency increases, the power except model size 63 is clearly larger than that of the training data (particularly towards higher frequency, over 100). Only the power of model size 63 shows similar behavior to that of the original training data for all frequencies.

From these results, we conclude that the model obtained by applying the NLS method is better than that by applying the LS method and avoids some kind of over-fitting.

6. Summary and Conclusion

This paper has shown that in unexpected situations, models tend to over-fit: even if we use description length (DL). Degeneracy is one example of this significant problem. We have described an idea to use the LS method more appropriately to overcome these problems.

We showed that a possible reason why models tend to over-fit (that is, the model size is larger than enough) is that noisy data is used as a proxy of the true state for the LS method. However, we usually do not know the true state in practice. Hence, we have proposed a method to improve the LS method without using true state. The idea is

---

5Even when we plot 20,000 data, the empty spaces persist.
Fig. 7. The reconstructed attractors of time series of Chua’s circuit equations and a free-run of the models obtained. In each panel, 5000 data points are plotted. Panel (a) training data, (b) model size 67, (c) model size 63, (d) model size 53 and (e) model size 41.
Fig. 8. Power spectrum of time series of Chua’s circuit equations and a free-run of the models obtained. Panel (a) training data, (b) model size 67, (c) model size 63, (d) model size 53 and (e) model size 41.
to add noise in the LS method, which we call the noisy least squares (NLS) method. We have provided theoretical arguments that the method can find a better model when the additional noise level is optimal. Applying the NLS method can avoid selecting a degenerate model as the best model and find the correct model as the best model. That is, to apply this idea DL can be used more effectively and generally avoid over-fitting (find smaller model).

As a more practical example (for building models using radial basis functions), the NSL method was applied to Chua’s circuit equation data set. The results indicate that the method obtained is a better one.

One of our future works is to investigate when the system has dynamic noise, because we have applied here the NSL method only to systems which do not have dynamic noise. The work is currently under investigation. In this paper, we have shown a problem, model degeneracy, and the model can be reduced to the correct model. This brings up two further questions, what is a better model and what is true dynamics? We now consider that if the model is degenerate, the model would have the same dynamics as the correct model, and if not, the model dynamics would differ. In a subsequent paper we shall continue a study of model size and dynamics.

Acknowledgments

The author would like to thank reviewers for their valuable comments. Tomo Nakamura would like to acknowledge valuable discussions with Associate Professor Kevin Judd (The University of Western Australia) concerning degenerate models. The authors would also like to thank Dr. David Walker (The Macaulay Institute) for his fruitful discussions on the NLS method. This research is supported by a Hong Kong University Grants Council Competitive Earmarked Research Grant (CERG) number PolyU 5235/03.

References


Appendix

In this appendix, we show that the degenerate models can be reduced to essentially the correct model. In Sec. 3.1, it was shown that the size of the LSE-best model was much larger than that of the correct model. Here, we show that the model is essentially the same as the correct model.

The Henon map (actual system model) is the 3-basis functions model

$$x_t = 1.0 + 0.3x_{t-2} - 1.4x^2_{t-1}.$$  

When the noise level was 40 dB and the number of data points used was 10 000, the model of size 3 was

$$x_t = 0.9995 + 0.3001x_{t-2} - 1.3992x^2_{t-1}.$$  

This is the correct model and not much different from the actual system. However, the model was
Fig. 9. (Continued)
not the LS-best model. The LS-best model was the 8-basis functions model. The model was

\[
x_t = 0.7781 - 0.4539x_{t-1} + 0.3004x_{t-2} - 0.235x_{t-1}^2 - 0.0663x_{t-3} - 0.7235x_{t-1}^2 + 0.3097x_{t-2}^2 - 0.2032x_{t-1}x_{t-3} + 0.9460x_{t-1}x_{t-2}^2.
\]

This model can be shown to be a good approximation of the correct model as follows. Rearrangement gives

\[
x_t = 0.7781 - 0.4539x_{t-1} + 0.3004x_{t-2} - 0.7235x_{t-1}^2
- 0.2211\left(\left(1.0 + 0.0663\frac{x_{t-1}}{0.2211} - 0.3097\frac{x_{t-2}^2}{0.2211}\right) - 1.0 \right)\frac{x_{t-1}}{0.2032}\frac{x_{t-3}}{0.6765}
- 0.9460x_{t-2}^2 - x_{t-1}\right)
= 0.7781 - 0.4539x_{t-1} + 0.3004x_{t-2} - 0.7235x_{t-1}^2
- 0.2211\left(\left(1.0 + 0.2999x_{t-3} - 1.4007x_{t-2}^2\right) - 1.0 \right) - 0.6765\left(x_{t-1}(1.0 + 0.3004x_{t-3}
- 1.3984x_{t-2}^2) - x_{t-1}\right).
\]

Since 1.0 + 0.2999x_{t-3} - 1.4007x_{t-2}^2 and 1.0 + 0.3004x_{t-3} - 1.3984x_{t-2}^2 are almost the same as the right-hand side of the correct model \(x_{t-1} = 1.0 + 0.3x_{t-3} - 1.4x_{t-2}^2\), these can be approximated by \(x_{t-1}\), giving

\[
x_t = 0.7781 - 0.4539x_{t-1} + 0.3004x_{t-2} - 0.7235x_{t-1}^2
- 0.2211(x_{t-1} - 1.0) - 0.6765(x_{t-1}^2 - x_{t-1})
= (0.7781 + 0.2211) + 0.3004x_{t-2}
+ x_{t-1}(-0.4539 - 0.2211 + 0.6765)
= 0.9992 + 0.0015x_{t-1} + 0.3004x_{t-2} - 1.4x_{t-1}^2
\approx 1.0 + 0.3x_{t-2} - 1.4x_{t-1}^2.
\]

Approximations such as model (A.1) are described as being degenerate. The other degenerate models of sizes 6 and 11 at other observational noise levels also have a structure like the above mentioned model.

Next, we compare the dynamics of the correct model and the degenerate models obtained of sizes 3, 6, and 8. To find the embedding parameters for each model, we apply the auto-correlation function and the false nearest neighbor method. The time-lag selected is 1 and dimension selected is 2 for embedding in all cases. Figure 9 shows the reconstructed attractors using data generated by these models. These figures show that these attractors seem to be identical. Also, we estimate the Lyapunov spectrum and the Lyapunov dimension of scalar time series. For this, we apply the algorithm proposed by Sano and Sawada [1985], where the time-lag and dimension for embedding is 1 and 2 as selected above and the number of data points is 10000. Table 4 shows that these values are almost the same.

<table>
<thead>
<tr>
<th>Model Size</th>
<th>Lyapunov Spectrum</th>
<th>Lyapunov Dimension</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>0.421760, -1.548846</td>
<td>1.272906</td>
</tr>
<tr>
<td>6</td>
<td>0.419802, -1.544187</td>
<td>1.271859</td>
</tr>
<tr>
<td>8</td>
<td>0.425957, -1.548973</td>
<td>1.274993</td>
</tr>
<tr>
<td>11</td>
<td>0.429385, -1.548963</td>
<td>1.272908</td>
</tr>
<tr>
<td>Correct model</td>
<td>0.425723, -1.555807</td>
<td>1.273635</td>
</tr>
</tbody>
</table>

Table 4. Comparison of the Lyapunov spectrum and Lyapunov dimension estimates for the correct model and models obtained of sizes 3, 6, 8 and 11.