2004 International Symposium on Nonlinear
Theory and its Applications (NOLTA2004)
Fukuoka, Japan, Nov. 29 - Dec. 3, 2004

Modelling for 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
Email: entomo@eie.polyu.edu.hk, ensmall@polyu.edu.hk

Abstract—We consider the problem of using the least squares method for building models which have large numbers of parameters. When using the least squares method for parameter estimation in practice, although we want to use the true state (noise free data), we usually use noisy data as a proxy of the true state in the formula, because we do not know the true state. When the noise level is low, although this gives good estimates for the parameters, the models selected as the best model by information criteria tend to over-fit because of the proxy. We show a significant example that the correct model is not selected as the best model, propose an idea to overcome the problem and demonstrate the ability to find better models.

1. Introduction

In this paper we consider the problem of using the least squares method for building pseudo-linear models [1] of a nonlinear dynamical system. The models have large numbers of parameters and information criteria are applied to find the best (optimal) model. For parameter estimation, the least squares method is usually applied, where although we want to use the true state (noise free data), we usually use noisy data as a proxy of the true state as the common usage, because we do not know the true state. Even when the noise level is low, although this gives good estimates for the parameters, because of the proxy, the models selected as the best model by information criteria tend to be over-fitted, that is, the model size becomes unnecessarily larger. However, it should be noted that the usage of the least squares method is not particular but rather ordinary.

2. The least squares method and parameter estimation

We now consider the problem of estimating the parameters $\lambda \in \mathbf{R}^k$ of a model $x_{t+1} = f(x_t, \lambda)$, $x \in \mathbf{R}^d$ of a nonlinear deterministic dynamical system given only a finite time series of observations s_t contaminated by observational noise, where the data comprises 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 λ, where the model is identical to the system. A commonly used method to estimate λ is by least squares, that is, to solve the optimization problem

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

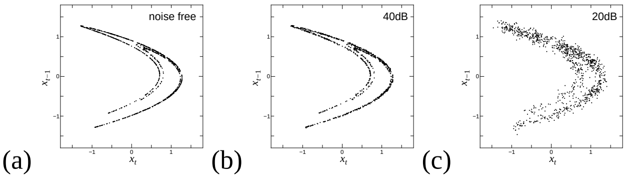

Figure 1: The Henon map attractors reconstructed with different observational noise levels. (a) noise free, (b) 40dB and (c) 20dB.

where only the noisy observations s_t are used for the fitting. This is a maximum likelihood method that makes the assumption that the noise is Gaussian and independent. When the noise level is low this gives good estimates for the parameters. However, when not so, it is well known that even when these assumptions hold, least squares does not provide good estimates for the parameters, because the estimates show significant bias. The parameter estimates would be much less biased if we could solve the optimization problem

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

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. We consider that the usage of Eq. (1) is "inappropriate" and that of Eq. (2) is "appropriate" as the least squares method.

We compare parameter estimations and fitting errors when using both the equations. We use the Henon map [1] as one example of nonlinear models. The Henon map, when formulated as a second order difference equation, is given by $x_t = A_0 + A_1 x_{t-2} + A_2 x_{t-1}^2$, where $(A_0, A_1, A_2) = (1.0, 0.3, -1.4)$. Figure 1 shows the reconstructed attractors of noise free, 40dB and 20dB noisy data. Panel (b) is very similar to panel (a) where the noise free data are plotted. However, panel (c) shows that the shape of the reconstructed attractor is very fuzzy. When the number of data points used is 1,000, the parameters estimated (A_0, A_1, A_2) are (0.9646, 0.3038, -1.3316) using Eq. (1) and (0.9966, 0.3044, -1.3968) using Eq. (2) at the noise level 20dB, and the parameters estimated are (0.9991, 0.3008, -1.3990) using Eq. (1) and (0.9997, 0.3004, -1.3997) using Eq. (2) at

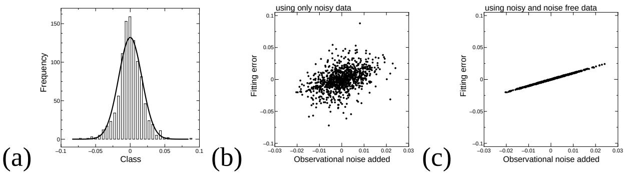

Figure 2: The fitting error for the Henon map when the noise level is 40dB and the number of data points used is $1,000$, where Eq. (1) is used for (a) and (b), and Eq. (2) is used for (c). (a): The distribution of the fitting error. (b) and (c): The fitting errors and observational noise added.

the noise level 40dB. When the noise level is 20dB, the parameters estimated using Eq. (2) are good estimates, but those using Eq. (1) are not, and the difference between both the parameters are very large. However, when the noise level is 40dB, although the parameters estimated using Eq. (2) are better estimates than those using Eq. (1), the difference between both the parameters are not significant and both the parameters are good estimates.

However, there is a problem to use Eq. (1), even when the noise level is low. We show it in Figure 2. Panel (a) shows the distribution of the fitting error (prediction error), where the histogram is obtained from the fitting errors and the solid line is a theoretical Gaussian with the same standard deviation as the fitting errors. As the figure shows clearly, the distribution of the fitting error is not normal. We also investigate the relationship of the fitting errors and Gaussian observational noise added. Panel (b) shows that the fitting errors obtained using Eq. (1) and the observational noise added are uncorrelated: the correlation coefficient is 0.42889. We expect that the fitting error is the same as the observational noise included in the data. However, this seems to indicate that there is no connection between them. Also, although the standard deviation of the 40dB observational noise added is 0.00724, that of the fitting errors is 0.01678, which is much larger than that of the observational noise added. Panel (c) shows that the fitting errors obtained using Eq. (2) and the observational noise added are almost identical. These results show adverse affects due to using Eq. (1) and that it is very useful to use Eq. (2). However, we cannot know the noise free data in the real world. We show a problem due to that noisy data is used as a proxy of the true state in the least squares method in the next section.

3. Degeneracy of time series models

In this section we will observe a significant example as the correct model is not selected as the best model when using Eq. (1).

When building pseudo-linear models, many candidate basis functions are first prepared in the form of a dictionary, which one hopes will be able to describe any likely nonlinearity and feature of the data. Then, basis functions are selected from the dictionary by a selection method, and models are obtained by taking a linear combination of these to form the model. It is considered that the minimum value of the information criterion yields the best (optimal) model. Hence, selection algorithms usually employ some information criteria to find a model that balances the model error against model size so as to prevent over-fitting and under-fitting data.

Another important reason for using information criteria is to avoid unnecessary increase in the model size, which occurs when a model is build that models a nested, that is, self-iterated form of the original system. A simple example is the following. Let the original model be $x_t = a_1 x_{t-1} + a_3 x_{t-3}$, which has model size 2. This model can be rewritten as $x_{t-1} = a_1 x_{t-2} + a_3 x_{t-4}$, which is essentially the same as the original model. Using the latter expression to replace the basis function $\frac{1}{2}x_{t-1}$ in the original model gives $x_t = \frac{1}{2}a_1 x_{t-1} + \frac{1}{2}a_1^2 x_{t-2} + a_3 x_{t-3} + \frac{1}{2}a_1 a_3 x_{t-4}$. Although the model is identical to the original model, 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 model the system.

Some information criteria have already been proposed for these purposes. The criterion we use is the Rissanen's Description Length modified by Judd and Mees (*DL*) [1].

3.1. Degenerate models

We consider the Henon map using multivariate polynomial models, and choosing lag=3 and degree=3 gives 20 candidate basis functions in the dictionary. From the dictionary we can build a model $x_{t-1} = A_0 + A_1 x_{t-3} + A_2 x_{t-2}^2$, from which we can build the degenerate models. The levels of observational noise added are 20dB, 40dB, 60dB and 80dB, the number of data points used are 1,000 and 10,000, and Eq. (1) is used. We calculate all possible combinations (an exhaustive search) to obtain the truly best model.

Table 1 shows the model size of the best model selected at different noise levels. In all cases, *DL* (description length) 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 model size. That is, the correct model is not the best model. It should be noted that this is not particular phenomenon using *DL*. Although we show the results using only *DL*, we have confirmed that the results using other information criteria, for example the Normalized Maximum Likelihood criterion [1], are essentially the same.

It is usually considered that larger incorrect models may predict a given noisy time series more effectively than the correct model. If this is correct, the reason that the size of the best models was larger than that of the correct model was expected that this would be due to the high noise level.

Table 1: The size of the best model for the Henon map at different noise levels.

Noise level	Number of data points	
	1,000	10,000
20dB	8	10
40dB	8	8
60dB	6	8
80dB	6	8

However, even when the noise levels are lower, the correct model is 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 the above mentioned reason.

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

$$\begin{aligned} x_t &= 0.7781 - 0.4539x_{t-1} + 0.3004x_{t-2} - 0.0663x_{t-3} \\ &\quad - 0.7235x_{t-1}^2 + 0.3097x_{t-2}^2 - 0.2032x_{t-1}x_{t-3} \\ &\quad + 0.9460x_{t-1}x_{t-2}^2. \end{aligned} \tag{3}$$

This is, in fact, a very good but *degenerate*, approximation to the correct model. That is, the model can be reduced to the essentially correct model. We can find such good but degenerate approximations of sizes 6, 8 and 11 in most cases. Furthermore, these degenerate models are selected as the best models in most cases.

The results show that we cannot avoid unnecessary increase in the model size and we cannot select the correct model as the best model, even when the noise level is low and truly the best model is obtained by calculating all possible combinations. This is a very significant problem for building models.

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), under the same condition. We again calculate all possible combinations to obtain the truly best model. In all cases the correct model is selected as the best model. This indicates that it is very useful to use the noise free data for not only parameter estimation and but also take advantage of information criteria effectively. However, we usually must use the noisy data as proxy for the noise free data, because we cannot know the noise free data.

4. An idea to use the least squares method more appropriately

As shown in previous section, it is very useful to use the true state (noise free data). However, very large problem is that it is difficult work to obtain true state from noisy state. Hence, we propose an idea to use the lease squares method more appropriately without using the true state. The only assumption we use is that the observational noise is Gaussian.

As Figure 1 (b) shows, the reconstructed attractor with 40dB noise is very similar to that with noise free, and the parameters estimated are almost the same as the correct values. These facts indicate that the noisy data can be regarded as a good proxy for the true state when the noise level is low. Hence, to achieve a proxy of Eq. (2) using only noisy data, we propose the addition of larger Gaussian noise to the part of s_{t+1} in Eq. (1).

Let the added Gaussian noise be ϵ'_t and $s'_{t+1} = s_{t+1} + \epsilon'_{t+1}$. Then we obtain new optimization problem

$$\min_{\lambda} \sum_{t=1}^{n-1} \left\| s'_{t+1} - f(s_t, \lambda) \right\|^2 . \tag{4}$$

In Eq. (2), the s_{t+1} term has more noise than x_t. Hence, when the level of the added noise is large enough relative to the noise included in the original noisy data s_t, we expect that the Eq. (4) can be good approximation to the Eq. (2). We refer to the method as the "additional Gaussian noise least squares (AGLS)" method.

We apply the idea to the same example used in section 3.1. We again calculate all possible combinations to obtain the truly best model. We add the noise level up to 0dB from 80dB every 10dB, and see what happen. The selected basis functions does not change when the noise level added is lower than that included in the original noisy data. However, as the noise level added becomes larger, only the basis functions in the correct model are selected. That is, the correct model is selected as the best model. This result indicates that applying the idea can avoid over-fitting and degeneracy.

5. Application

In the earlier examples, we always could obtain clear results in any noise level even when the noise level was 20dB, which is relative large noise level. This would be possible, because there were the correct basis functions in the dictionary and all possible combination sets were calculated. However, it is very unrealistic, because in practical cases, there is no correct basis functions, instead of calculating all possible combination sets, selection methods are applied, and only polynomial basis functions are not recommended for modelling [1]. Hence, we investigate how our idea works in practical cases. For this purpose, we build models using radial basis functions and apply a selection method, the up-and-down method using marginal error [1]. For applying the proposed idea, we use the following idea. We first build models using a selection method and a training data as usual. When we apply the proposed idea for the least squares method, we do not build models again using the noise added data. We keep using the original models obtained using the original training data, but we calculate the description length using Eq. (4), that is, the original training data and noise added data. Then we find the best model at each noise level added. The reason why we apply the above idea is that selecting basis functions is much influenced by noise. Hence, we do not want to use very noisy data as the training data.

5.1. The model of the differential equation of the double scroll circuit equations

The model we use is a electronic circuit proposed by Chua *et al* [2]. We contaminate the data by 60dB noise and use the data as observational data. For building a model 5,000 data points are used as the training data, and the data is embedded using uniform embedding $(t-1, t-5, t-9)$ with the aim of predicting a value at time t.

The size of the model obtained as the best model is 67. We apply the idea used to avoid degeneracy and find the best model again at the each noise level added, where we use 5 different Gaussian noise realizations in the AGLS method. Table 2 shows the mode of the model size. From 60dB to 50dB, the model size is almost the same, size 63. Hence, we regard the model as the global best model.

To investigate the quality and performance of the models obatined, we use long-term free-run data of the models, because one needs to get the dynamics right to obtain good long-term free-run data. Figure 3 shows the reconstructed attractors of the training data and those of the ubiquitous behaviours of the free-run data of the models. Panel (a) shows that there are empty spaces around the centres in the left and right sides. Panel (b) shows that the empty spaces are not clear. However, the behaviour on other areas are very similar to panel (a). Panel (c) shows that the empty space in the right side is very clear and that in the left side is more clear than that using the model of the size 67. Also, the behaviour on other areas are similar to panel (a) as well as the model of size 67. Panel (d) shows that although the empty spaces in both the sides are clear, the behaviour on other areas is getting periodic. Also, panel (d) shows that the behaviour in the middle section (between the two unstable focii) is much simpler than others. This result indicates that the model of the size 63 shows the best behaviour. Also, it indicates that the model of size 67 is over-fitted.

6. Summary and Conclusion

We have described that in unexpected situations, some models tend to over-fit; degeneracy is one example of this significant problem. To overcome these problem, we proposed an idea to use the least squares method more appropriately without using the true state, the AGLS method. The results indicate that applying the proposed method can take advantage of information criteria more effectively and generally avoid over-fitting.

Table 2: The mode of the best model size obtained at different noise level added for the models of the double scroll circuit equations

Noise level	Model size	Noise level	Model size
20dB	4	50dB	63
25dB	6	55dB	63
30dB	14	60dB	64
35dB	20	65dB	67
40dB	39	70dB	67
45dB	53	Original model	67

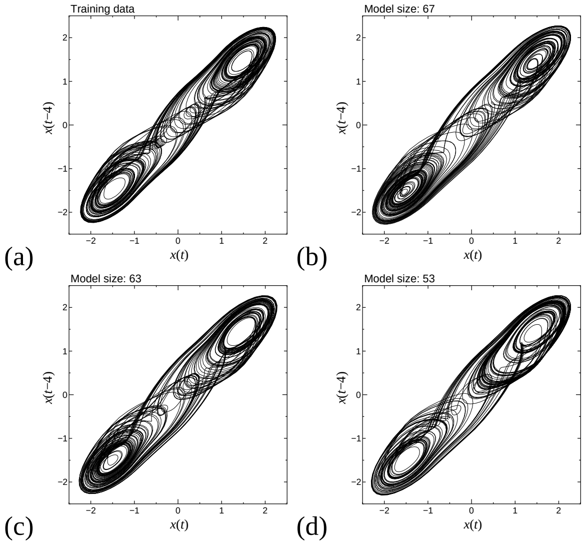

Figure 3: The reconstructed attractors of time series. 5000 data points are plotted. Panel (a) training data, (b) model size 67, (c) model size 63, and (d) model size 53.

Acknowledgments

Tomo Nakamura would like to acknowledge valuable discussions with Associate Professor Kevin Judd (The University of Western Australia) concerning degenerate models. This research is supported by a Hong Kong University Grants Council Competitive Earmarked Research Grant (CERG) number PolyU 5235/03.

References

[1] K. Judd, “Building Optimal Models of Time Series.” In: *Chaos and its Reconstruction*, ed. G. Gouesbet, S. Meunier-Guttin-Cluzel and O. Menard, Nova Science Pub Inc, pp. 179–214, 2003.

[2] T. Matsumoto, L. O. Chua and M. Komuro, “The Double Scroll,” *IEEE Trans. Circuits Syst*, vol. 32, pp. 797–818, 1985.