Estimating the Variance of the k-step Ahead Predictor for Time-series

Fredrik Tjärnström

Department of Electrical Engineering
Linköping University, S-581 83 Linköping, Sweden
WWW: http://www.control.isy.liu.se
Email: fredrikt@isy.liu.se

March 16, 1999

Report no.: LiTH-ISY-R-2130
Submitted to (Nothing)

Technical reports from the Automatic Control group in Linköping are available by anonymous ftp at the address ftp.control.isy.liu.se. This report is contained in the compressed postscript file 2130.ps.Z.
Estimating the Variance of the k-step Ahead Predictor for Time-series

Fredrik Tjärnström

March 16, 1999

Abstract

This paper considers the problem of estimating the variance of a linear k-step ahead predictor for time series. (The extension to systems including deterministic inputs is straightforward.) We compare the theoretical results with empirically calculated variance on real data, and discuss the quality of the achieved variance estimate.

Keywords: Prediction; Time-series

1 Introduction

We will discuss how to calculate the variance of a k-step ahead predictor for time series. This will give us useful information on how long prediction horizon we possibly should use. If we take the horizon to be even longer we will get less valuable results. (Less valuable in the meaning that the variance is so high in the prediction, so we could as well just have guessed an equally good prediction).

We start by giving a short introduction to prediction error methods in Section 2. Section 3 gives the main results and in Section 4 we apply the results in an example. Finally Section 5 gives the conclusions.

2 Prediction error methods for time-series

In this section we will discuss how conventional linear black-box models are used as predictors for time-series. In connection to this we will measure the goodness of such a predictor in form of estimated variance for the k-step ahead predictor.

To be able to evaluate the performance of a model we will assume that there exists a true model that describes how the output is generated

\[ y(t) = H_0(q)e(t) \]  

(1)
Here the sequence \( \{e(t)\} \) is white noise with variance \( \lambda_0 \). Our aim is to find a model \( H(q) \) that is close to \( H_0(q) \) in some sense. With \( A(q) = 1 + a_1q^{-1} + \cdots + a_nq^{-n} \) and \( C(q) = 1 + c_1q^{-1} + \cdots + c_mq^{-m} \) we can point out three main linear model classes for time-series:

1. The case when \( H(q) = C(q) \) is called a MA\((m)\)-model (moving average).
2. When \( H(q) = 1/A(q) \) the model is called an AR\((n)\)-model (autoregression).
3. Finally the case \( H(q) = C(q)/A(q) \) is called an ARMA\((n,m)\)-model.

Now let \( H(q) = H(q, \theta) \), where the \( \theta \)-vector represents our parameterization of the model structure, e.g., in the case of an ARMA-model we have \( \theta = (a_1 \cdots a_n c_1 \cdots c_m)^T \). Using prediction error methods means that we choose the model that is the best predictor of future outputs. See [1]. Since white noise is unpredictable the best possible predictor of \( y(t) \) given the value of \( \theta \) must be

\[
\hat{y}(t|\theta) = (1 - H^{-1}(q, \theta))y(t)
\]  

If we have collected data, \( y(1), \ldots, y(N) \), the natural estimate of \( \theta \) will be

\[
\hat{\theta}_N = \arg\min_{\theta} V_N(\theta)
\]

\[
V_N(\theta) = \sum_{t=1}^{N} e^2(t, \theta)
\]

\[
e(t, \theta) = y(t) - \hat{y}(t|\theta) = H^{-1}(q, \theta)y(t)
\]

3 The k-step ahead predictor and its variance

The k-step ahead predictor can be found from our estimated model \( \hat{H}(q) = \hat{C}(q)/\hat{A}(q) \) using the following. (See, e.g., [1] for similar expressions.) Let the polynomials \( F_{k-1}(q) = 1 + \cdots + f_{k-1}q^{-k+1} \) and \( G_{n-1}(q) = g_0 + \cdots + g_{n-1}q^{-n+1} \) be given by the Bezout identity

\[
C(q) = A(q)F_{k-1}(q) + q^{-k}G_{n-1}(q)
\]

Then the k-step ahead predictor, \( \hat{y}(t+k|t) \), and the corresponding prediction error will be given by

\[
\hat{y}(t+k|t) = \frac{G_{n-1}(q)}{C(q)}y(t)
\]

\[
y(t+k) - \hat{y}(t+k|t) = F_{k-1}(q)e(t+k)
\]
Note from (8) that the expected value of the prediction error is zero since \( \{e(t)\} \) is a white noise sequence.

Since we only have an estimate of the system from which the data was generated, we will get an increased uncertainty in the predictions. Therefore this lack of knowledge must be included in the expression for the variance of the k-step ahead prediction. How this is done can be seen from the following. If we denote the “true” k-step ahead predictor with \( \hat{y}_0(t + k|t) \) and the one estimated from our \( N \) data points with \( \hat{y}_N(t + k|t) \) we get

\[
e(t + k|t) = y(t + k) - \hat{y}_N(t + k|t) = y(t + k) - \hat{y}_0(t + k|t) + \hat{y}_0(t + k|t) - \hat{y}_N(t + k|t) = F_{k-1}(q)e(t + k) + \left( \frac{G_{n-1}(q)}{C(q)} - \frac{\hat{G}_{n-1}(q)}{\hat{C}(q)} \right) y(t)
\]

(9)

Since \( e(t + s) \), \( s \geq 1 \) is independent of \( y(t) \) we get

\[
\text{Var } e(t + k|t) = (1 + f_1^2 + \cdots + f_{k-1}^2) \lambda_0 + \text{Var} \left[ \left( \frac{G_{n-1}(q)}{C(q)} - \frac{\hat{G}_{n-1}(q)}{\hat{C}(q)} \right) y(t) \right]
\]

(10)

\[
= (1 + f_1^2 + \cdots + f_{k-1}^2) \lambda_0 + \text{Var } f(\hat{\theta})
\]

where \( f(\hat{\theta}) \) is defined by the last equality. We can now apply Gauss’ approximation formula which states that if \( \hat{\theta} \) is sufficiently close to the “true” parameter vector, \( \theta_0 \), we can approximate

\[
\text{Var } f(\hat{\theta}) \approx f'(\theta_0)P[f'(\theta_0)]^T
\]

(11)

where \( f'(\theta_0) = \frac{df(\theta)}{d\theta}\big|_{\theta=\theta_0} \) and \( P \) is the covariance matrix of \( \theta \). Now apply (11) to (10) which results in

\[
\text{Var } e(t + k|t) \approx (1 + f_1^2 + \cdots + f_{k-1}^2) \lambda_0 + f'(\theta_0)P[f'(\theta_0)]^T
\]

(12)

Furthermore we have

\[
f(\theta) = \left( \frac{G_{n-1}(q, \theta)}{C(q, \theta)} - \frac{G_{n-1}(q, \theta)}{C(q, \theta)} \right) y(t)
\]

(13)

which gives

\[
f'(\theta) = -\frac{1}{C(q, \theta)} \frac{dG_{n-1}(q, \theta)}{d\theta} y(t) + \frac{G_{n-1}(q, \theta)}{C^2(q, \theta)} \frac{dC(q, \theta)}{d\theta} y(t)
\]

(14)

Now \( \frac{dC(q, \theta)}{d\theta} = (0 \ldots 0 q^{-1} \ldots q^{-m})^T \) and \( \frac{dG_{n-1}(q, \theta)}{d\theta} \) can be calculated from (6).
When we have estimated the variance of the k-step ahead predictor, we could use it to decide how long the prediction horizon should be. Two simple comparisons could be done. Firstly, we could check if $\text{Var} \varepsilon(t + k|t)$ is lower than, e.g., $\frac{1}{2} \text{Var} y(t)$. Secondly, we could compare $\text{Var} \varepsilon(t + k|t)$ with the mean square error of a simple ad hoc predictor based on some physics or intuition. If we cannot improve on this for some $k$ it would probably be of no use in trying to make longer predictions with this model.

An interesting note is that the covariance matrix $P$ in (12) decays like $\frac{1}{t}$, which implies that if we have a sufficiently large number of data, i.e., $N$ should at least be larger than 100, we could neglect the second term and approximate the variance of the prediction by

$$\text{Var} \varepsilon(t + k|t) \approx (1 + f_1^2 + \cdots + f_{k-1}^2) \lambda_0$$

This expression is possible to calculate at the time the model is estimated since it does not depend on the data we want to predict.

4 Experimental studies on the water demand for Barcelona

To evaluate these methods, we applied them to the water demand data from Barcelona. The data is collected daily and show a strong weekly correlation. An AR(15)-model was fitted to the first 570 data points (after the mean of the data was removed). This model was chosen on due to its better performance on validation data, than the other considered models. Furthermore the variance of the prediction error was calculated for $k = 1, \ldots, 15$ from Equations (12) and (14) on the following 128 data points. As a comparison to these theoretical expressions we computed the empirical prediction error variance, i.e., the variance estimated from the calculated k-step ahead predictions (for $k = 1, \ldots, 15$):

$$\hat{\text{Var}} \varepsilon(t|t - k) = \frac{1}{128} \sum_{t=571}^{608} \left( \bar{y}(t|t - k) - \frac{1}{128} \sum_{s=571}^{608} \bar{y}(s|s - k) \right)^2$$

$$\bar{y}(t|t - k) = y(t) - \bar{y}(t|t - k)$$

These two curves are shown in Figure 1 together with the mean square error of a simple ad hoc predictor (the water demand seven days from now is the same as today). (We have used the mean square error for the ad hoc predictor since it could be biased.) We see that the estimated AR(15)-model has a much lower variance than the simple weekly predictor. Also note that the variance of the prediction error is lower than $\frac{1}{2} \text{Var} y(t) \approx 2.7 \times 10^8$ for $k \leq 14$. By the discussion in the previous section this means that the prediction horizon should probably not be chosen longer than two weeks.
Figure 1: Estimated (thick line) and empirical (dashed line) prediction error variance together with the mean square error for the ad hoc weekly k-step ahead prediction error (thin line) for $k = 1, \ldots, 15$. (The variance of $y$ was approximately $5.4 \cdot 10^8$.)

We also notice that the estimated variance is in close agreement with the empirical prediction error variance.

Finally it would be interesting to see if (15) is a good approximation of (12). In Figure 2 we have plotted both expressions. It is obvious that the last term in (12) can be neglected without much loss of accuracy. This was to be expected since the covariance matrix $P$ decays like $\frac{1}{N}$ and we have
5 Conclusions

We have derived an expression (12) for the prediction error variance of a linear time-invariant model. Its validity has been tested on the water demand data from Barcelona and is shown to be in good agreement with the empirical variance (16) estimated from data. We have also studied an approximation of (12) and found that this (very simple) expression (15) is a good estimate if the model has been estimated on a large data set.

References