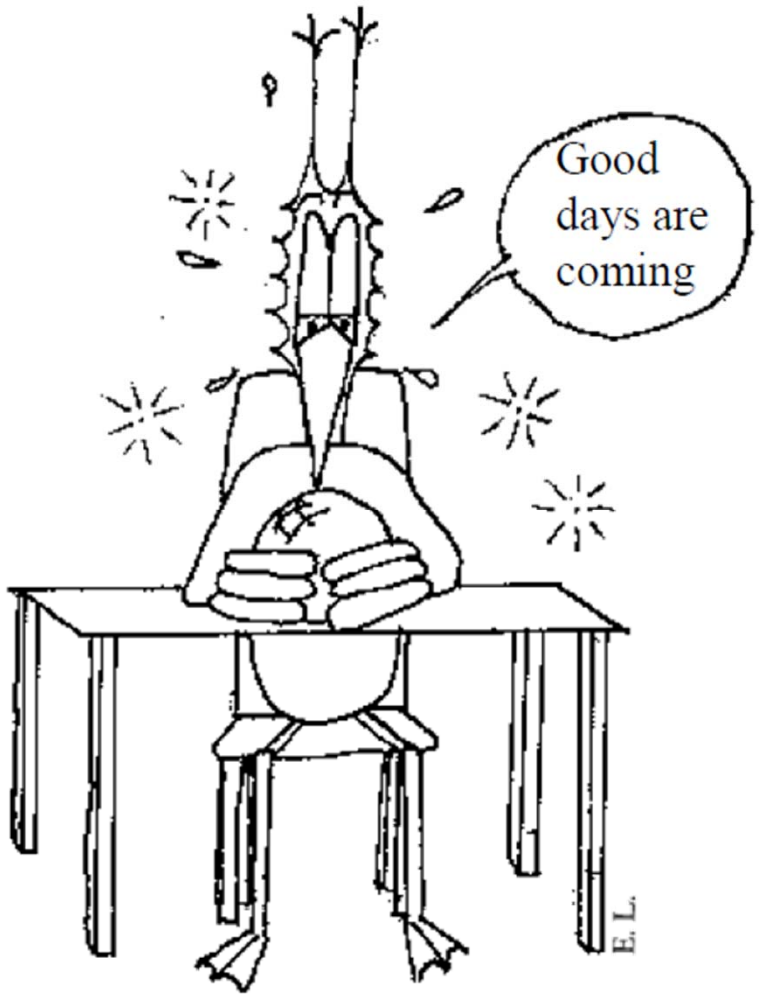
Forecasting
Part 1

JY Le Boudec



Contents

1. What is forecasting ?

Assume you have been able to define the *nature* of the load for your study

It remains to have an idea about its *intensity*

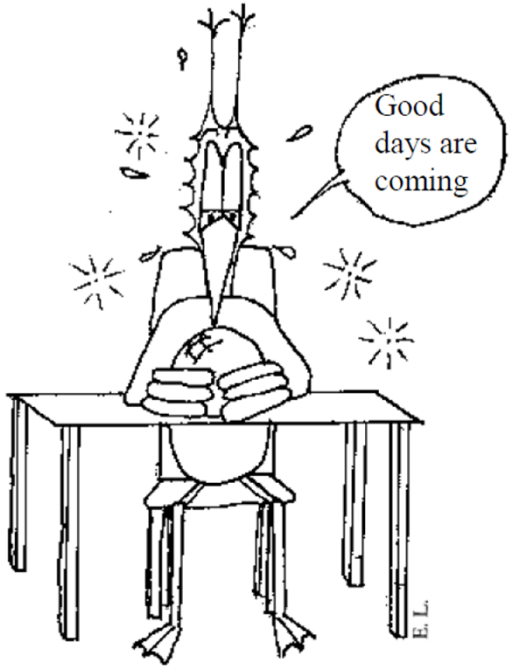
It is impossible to forecast without error

The good engineer should

Forecast *what can be forecast*

Give *uncertainty* intervals

The rest is outside our control



Forecasting = finding conditional distribution of future given past

Assume we observe some data $Y_t, t = 1,2,3 \ldots$

We have observed $Y_1, \ldots, Y_t$ and want to forecast $Y_{t+\ell}$
A full forecast is the conditional distribution of $Y_{t+\ell}$ given $Y_1, \ldots, Y_t$

A point forecast is (e.g.) the mean, i.e. $\widehat{Y}_t(\ell) = E(Y_{t+\ell}|Y_1, \ldots, Y_t)$
(or the median)

A prediction interval $[A; B]$ at level 95% is such that

$$P(A \leq Y_{t+\ell} \leq B|Y_1, \ldots, Y_t) \geq 0.95$$

2. Use of Regression Models

Simple, often used

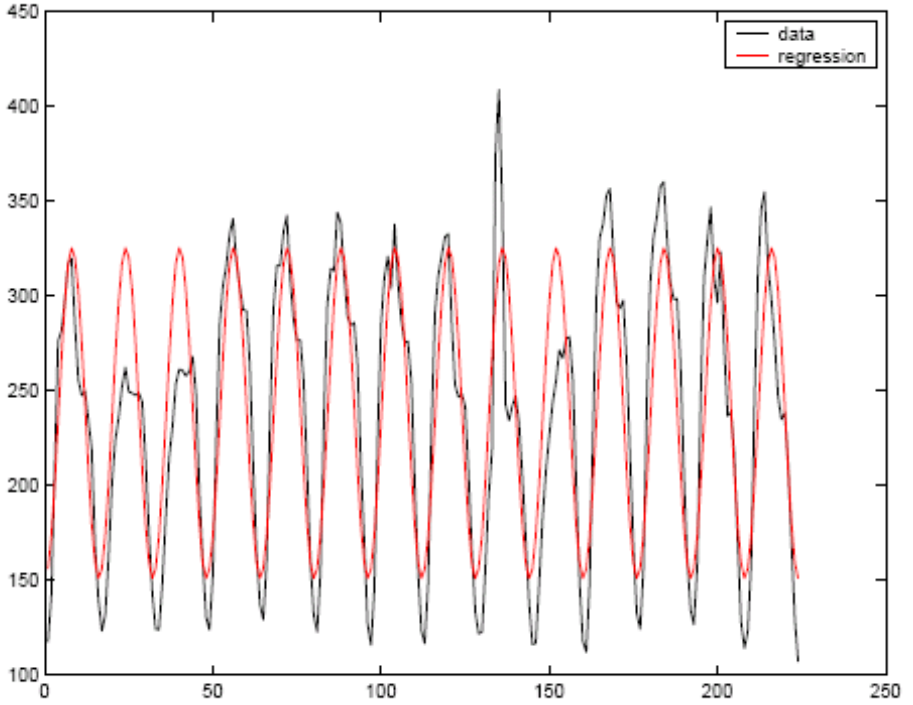
Based on a model fitted over the past, assumed to hold in the future

Example:

$$Y_t = 238.2475 - 871876\cos\left(\frac{\pi}{8}t\right) - 4.2961\sin\left(\frac{\pi}{8}t\right) + \epsilon_t$$

with $\epsilon_t \sim iid\ N(0, \sigma^2)$,

and $\sigma = 38.2667$



Prediction

We have obtained the model

$$Y_t = 238.2475 - 871876\cos\left(\frac{\pi}{8}t\right) - 4.2961\sin\left(\frac{\pi}{8}t\right) + \epsilon_t$$

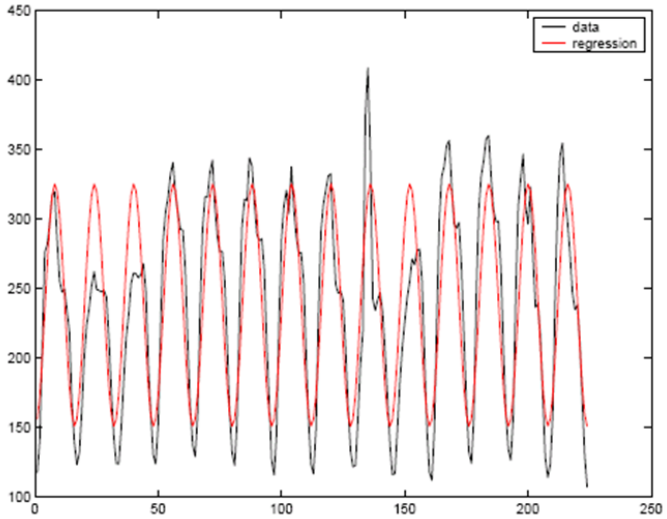
with $\epsilon_t \sim iid\ N(0, \sigma^2), \sigma = 38.2667$

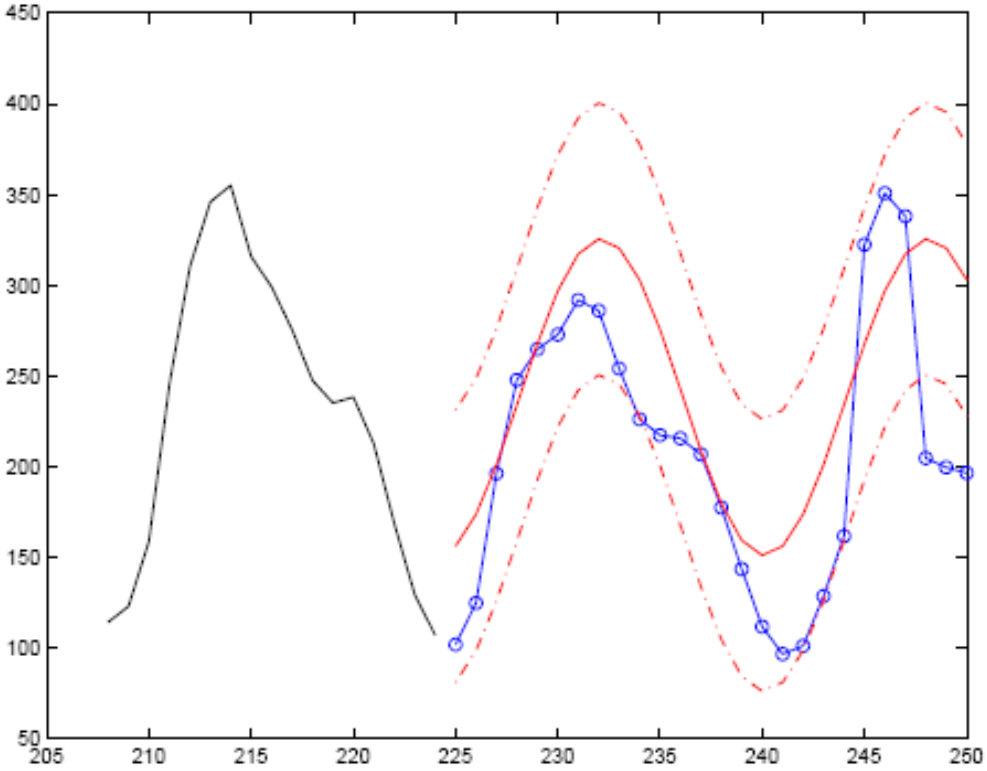
The conditional distribution of $Y_{t+\ell}$ given $Y_1, \ldots, Y_t$ is

$$Y_{t+\ell} = 238.2475 - 871876\cos\left(\frac{\pi}{8}(t+\ell)\right) - 4.2961\sin\left(\frac{\pi}{8}(t+\ell)\right) + \epsilon_{t+\ell}$$

with $\epsilon_{t+\ell} \sim N(0, \sigma^2), \sigma = 38.2667$

because $\epsilon_{t+\ell}$ is independent of $Y_1, \ldots, Y_t$ (iid assumption)





A point prediction is:

$$\hat{Y}_t(\ell) = \sum_{j=1}^{3} \beta_j f_j(t+\ell) = 238.2475 - 87.1876\cos(\frac{\pi}{8}(t+\ell)) - 4.2961\sin(\frac{\pi}{8}(t+\ell))$$

and a 95%-prediction interval can be approximated by $\hat{Y}_t(\ell) \pm 1.96\sigma$.

Virus Growth Data

We have obtained the model

$$\log Y_t = \log a + \alpha t + \epsilon_t$$

with $\epsilon_t \sim iid$ Laplace(λ), λ =6.2205

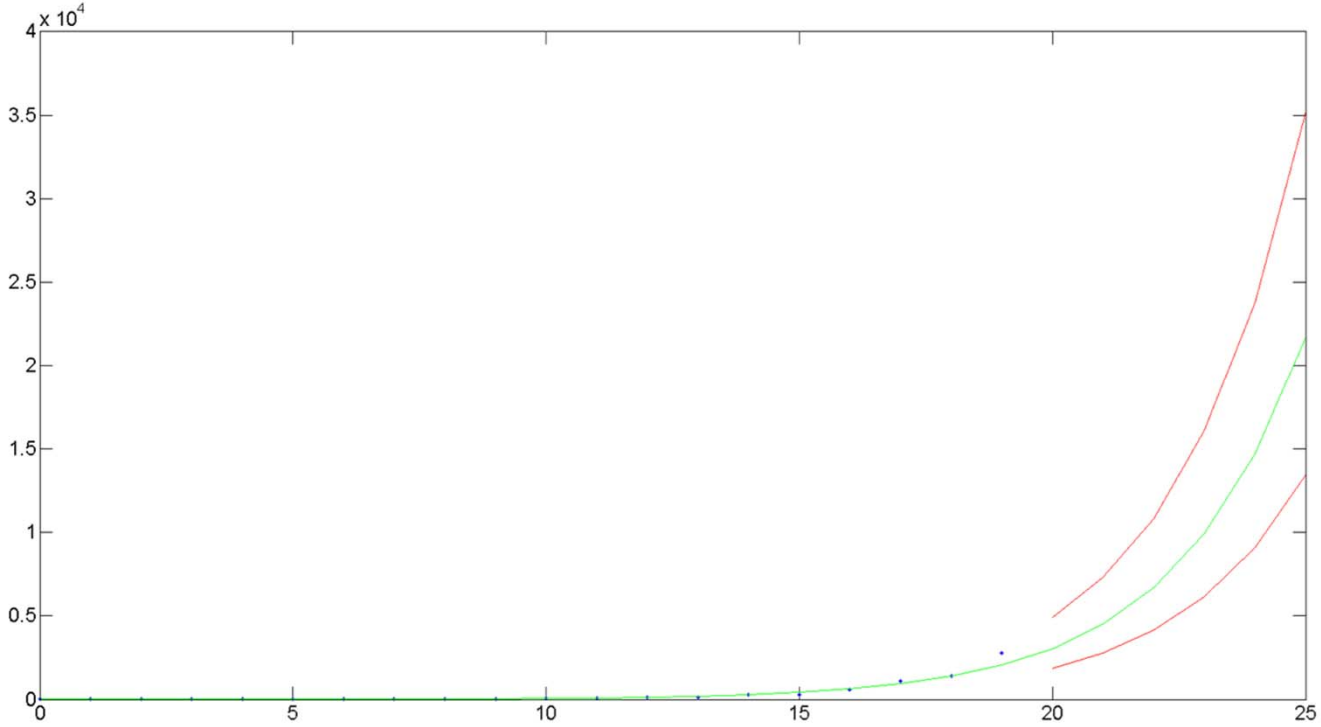
A 95%-prediction interval is

$$\log Y_{t+\ell} = \log a + \alpha(t+\ell) \pm \eta$$

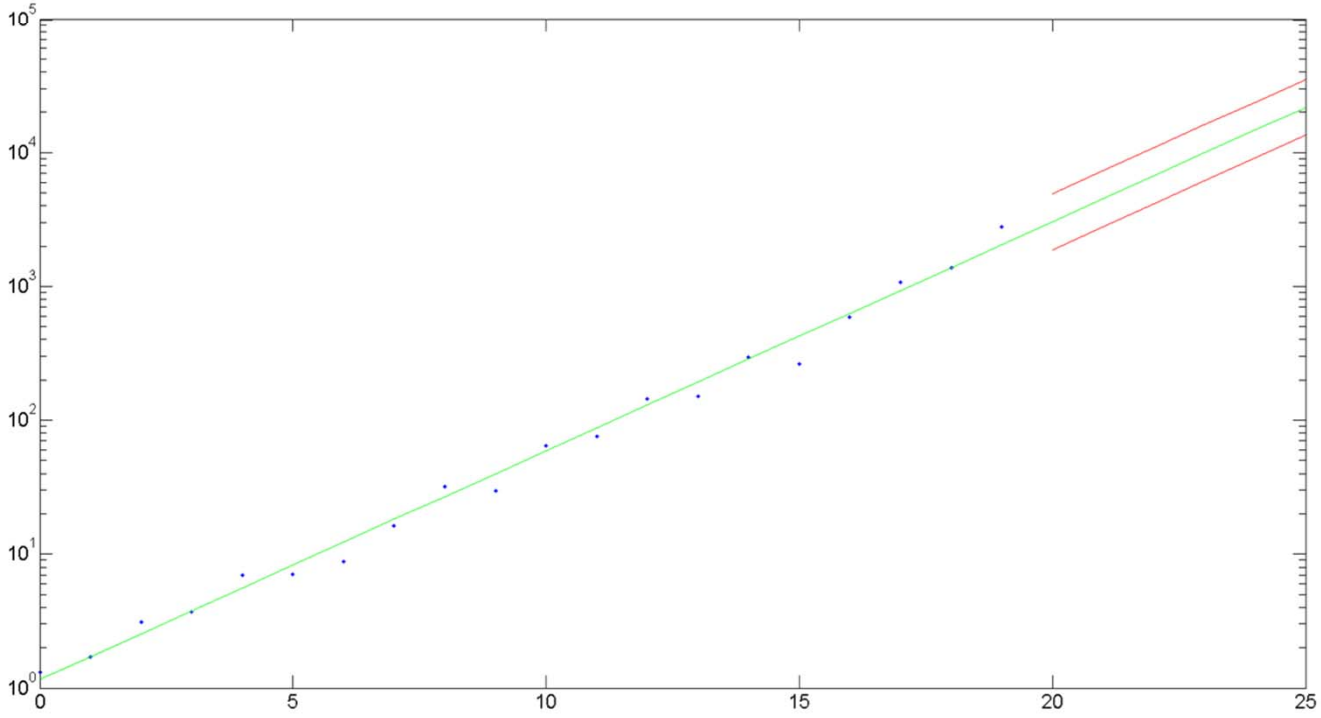
where η is the 97.5% quantile of the Laplace(λ) distribution;

In natural scale: Point prediction: $\widehat{Y}_{t+l} = ae^{\alpha(t+\ell)}$

95%-prediction interval: $[ae^{\alpha(t+\ell)}e^{-\eta}; ae^{\alpha(t+\ell)}e^{\eta}]$



Natural scale



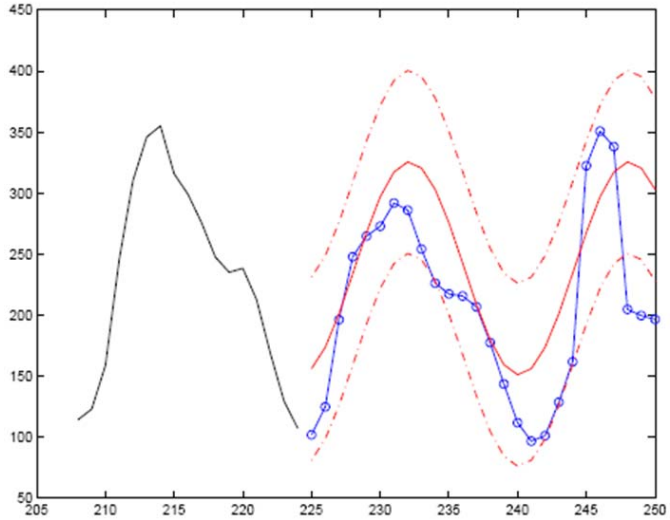
Log scale

$\lambda = 6.2205$

Prediction interval at time 25

PI = [19942 ; 52248]

Say what is true, for this model →



A. The width of prediction interval is constant and equal to $2 \times 1.96\sigma$
B. A is true and σ is the root mean square of the residuals up to time $t = 224$
C. A is true and σ is the root mean square of the forecast errors if we apply the model up to time $t = 224$
D. B and C
E. None of the above
F. I don't know

3. How about the estimation error ?

In practice we estimate the model parameter θ from $y_1, \ldots, y_t$

When computing the forecast, we pretend θ is known, and thus make an estimation error (ie we ignore confidence intervals on θ– it is hoped that the estimation error is much less than the prediction interval).

Let us return to an example we already saw. Assume we observe $X_1, \ldots X_n$ and want to forecast X_{n+1}. Assume that we believe in the model $X_i = \mu + \epsilon_i, \epsilon_i \sim iid\ N(0, \sigma^2)$. We estimate and obtain $\hat{\mu}, \hat{\sigma}$.

Point prediction for X_{n+1} if we ignore estimation uncertainty: $\hat{\mu}$;

if we account for estimation uncertainty, $\hat{\mu} \pm 1.96 \frac{\hat{\sigma}}{\sqrt{n}}$

95%-prediction interval for X_{n+1} if we ignore estimation uncertainty: $\hat{\mu} \pm 1.96\ \hat{\sigma}$

THEOREM 2.6 (Normal IID Case). *Let $X_1, ..., X_n, X_{n+1}$ be an iid sequence with common distribution N_{μ,σ^2}. Let $\hat{\mu}_n$ and $\hat{\sigma}_n^2$ be as in Theorem 2.3. The distribution of $\sqrt{\frac{n}{n+1}}\frac{X_{n+1}-\hat{\mu}_n}{\hat{\sigma}_n}$ is Student's t_{n-1}; a prediction interval at level $1-\alpha$ is*

$$\hat{\mu}_n \pm \eta' \sqrt{1+\frac{1}{n}}\hat{\sigma}_n \tag{2.33}$$

where η' is the $\left(1-\frac{\alpha}{2}\right)$ quantile of the student distribution t_{n-1}.
For large n, an approximate prediction interval is

$$\hat{\mu}_n \pm \eta\hat{\sigma}_n \tag{2.34}$$

where η is the $\left(1-\frac{\alpha}{2}\right)$ quantile of the normal distribution $N_{0,1}$.

Thm 2.6 says that (for $n = 100$) an exact interval that accounts for estimation uncertainty is $\hat{\mu} \pm 1.99\,\hat{\sigma}$
– compare to $\hat{\mu} \pm 1.96\,\hat{\sigma}$

The estimation error decays in $\frac{1}{\sqrt{n}}$ and is small for large n

Exact Formulas exist for Linear Regression with LS

THEOREM 5.1. *Consider a linear regression model as in Eq.(5.1) with p degrees of freedom for $\vec{\beta}$. Assume that we have observed the data at n time points $t_1, ..., t_n$, and that we fit the model to these n observations using Theorem 3.3. Assume that the model is regular, i.e. the matrix X defined by $X_{i,j} = f_j(t_i)$, $i = 1, ..., n$, $j = 1, ..., p$ has full rank. Let $\hat{\beta}_j$ be the estimator of β_j and s^2 the estimator of the variance, as in Theorem 3.3.*

1. *The point prediction at time $t_n + \ell$ is $\hat{Y}_{t_n}(\ell) = \sum_{j=1}^{p} \hat{\beta}_j f_j(t_n + \ell)$*
2. *An exact prediction interval at level $1 - \alpha$ is*

$$\hat{Y}_{t_n}(\ell) \pm \xi \sqrt{1 + g}\, s \tag{5.3}$$

with

$$g = \sum_{j=1}^{p} \sum_{k=1}^{p} f_j(t_n + \ell) G_{j,k} f_k(t_n + \ell)$$

where $G = (X^T X)^{-1}$ and ξ is the $(1 - \frac{\alpha}{2})$ quantile of the student distribution with $n - p$ degrees of freedom, or, for large n, of the standard normal distribution.

3. *An approximate prediction interval that ignores estimation uncertainty is*

$$\hat{Y}_{t_n}(\ell) \pm \eta s \tag{5.4}$$

where η is the $1 - \alpha$ quantile of the standard normal distribution.

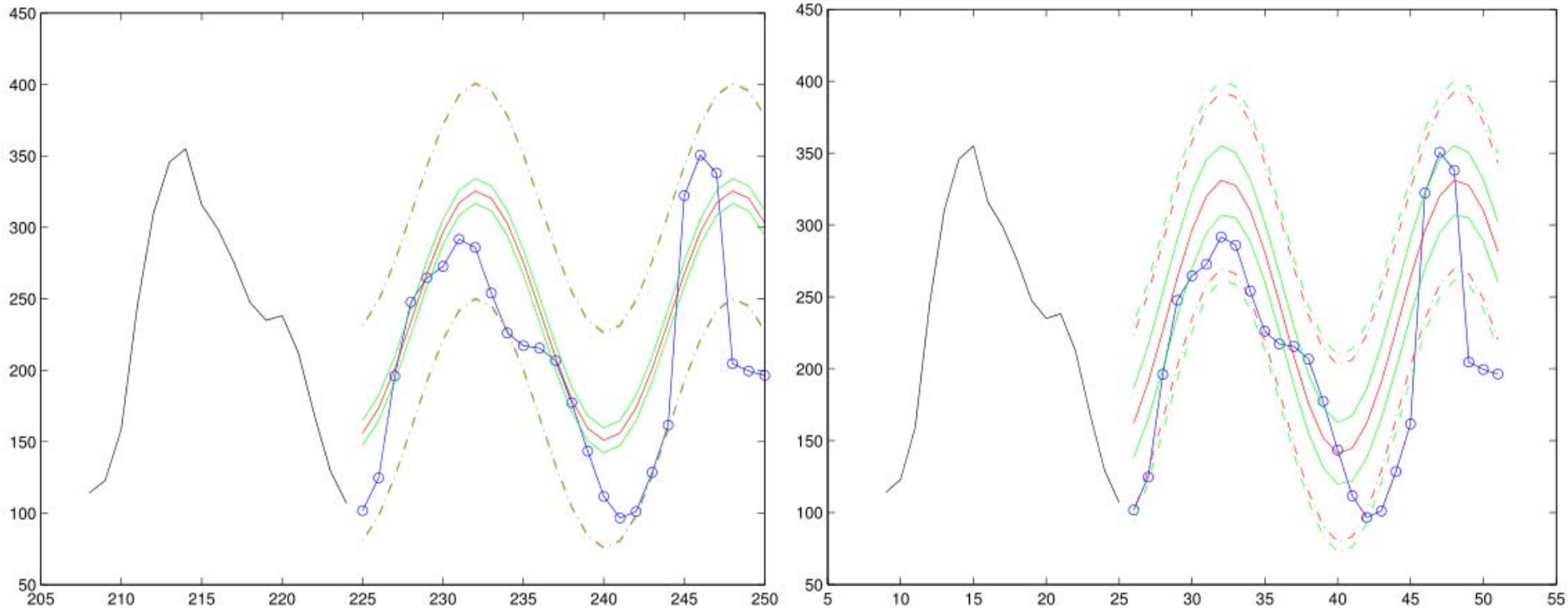


Figure 5.2: Left: Same example as Figure 5.1, showing the prediction interval computed by Theorem 5.1(dot-dashed lines) and the confidence interval for the point prediction (plain lines around center values). The predictions intervals computed by Eq.(5.3) and Eq.(5.4) are indistinguishable. Right: same except only the last 24 points of the past data are used to fitting the model (instead of 224). The confidence interval for the point prediction is slightly larger than in the left panel; the exact prediction interval computed from Theorem 5.1 is only slightly larger than the approximate one computed from Eq.(5.4).

Take-Home Message

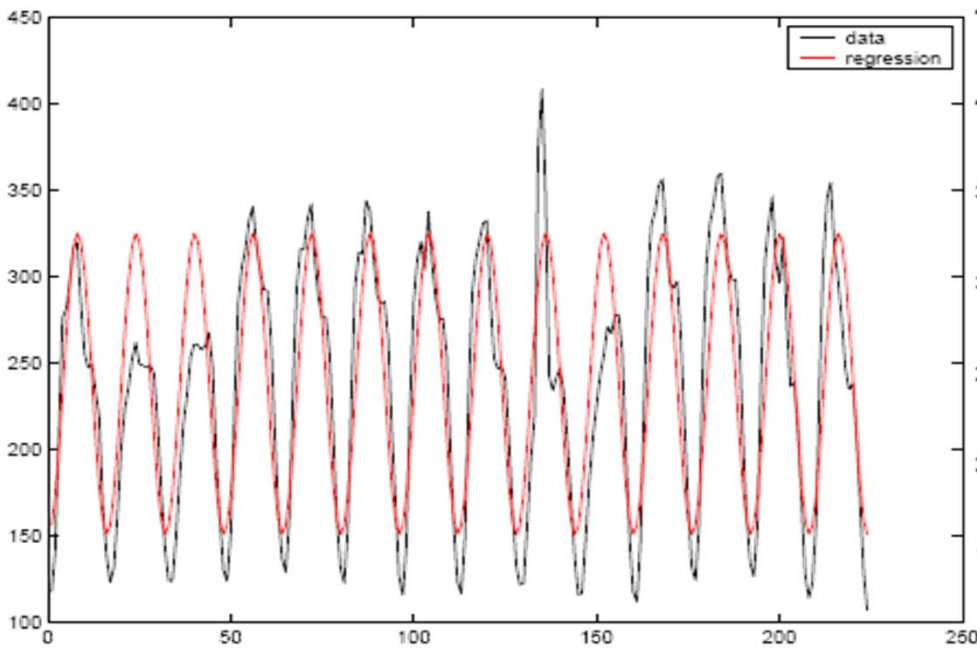
When we use a fitted model there is some uncertainty that adds to the prediction intervals

In most cases we can ignore the model uncertainty because it impacts the prediction intervals only marginally

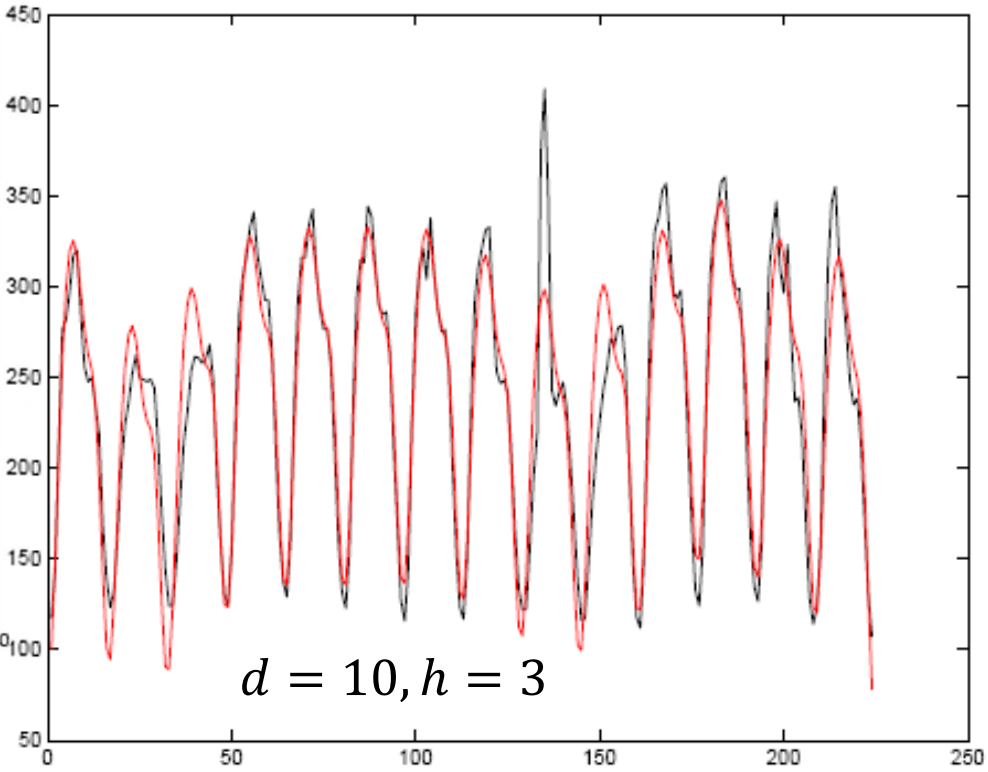
In some rare cases (e.g. linear regression with gaussian errors) there are exact formulas

4. The Overfitting Problem

Assume we want to improve our model by adding more parameters: add a polynomial term + more harmonics



$d = 0, h = 1$



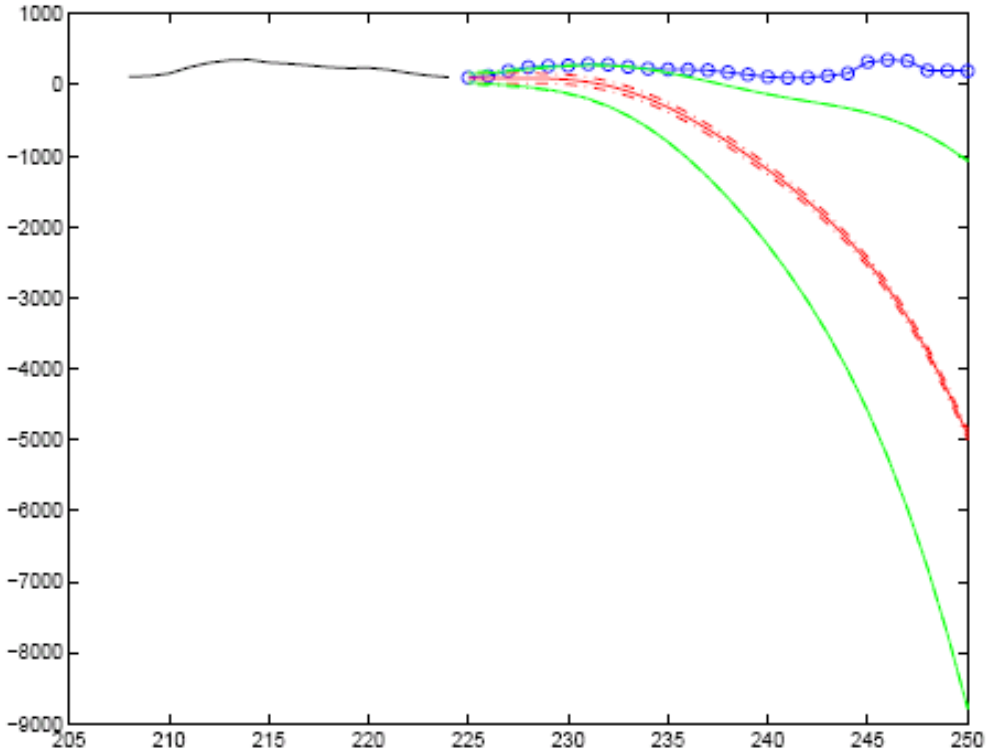
$d = 10, h = 3$

harmonics), with the hope of improving the fit, thus the prediction. The new model has the form

$$Y_t = \sum_{i=0}^{d} a_i t^i + \sum_{j=1}^{h} \left(b_j \cos \frac{j\pi t}{8} + c_j \sin \frac{j\pi t}{8} \right) \qquad (6.5)$$

Prediction for the better model

Figure 6.4 shows the resulting fit for a polynomial of degree $d = 10$ and with $h - 1 = 2$ harmonics. The fit is better ($\sigma = 25.4375$ instead of 38.2667), however, the prediction power is ridiculous. This

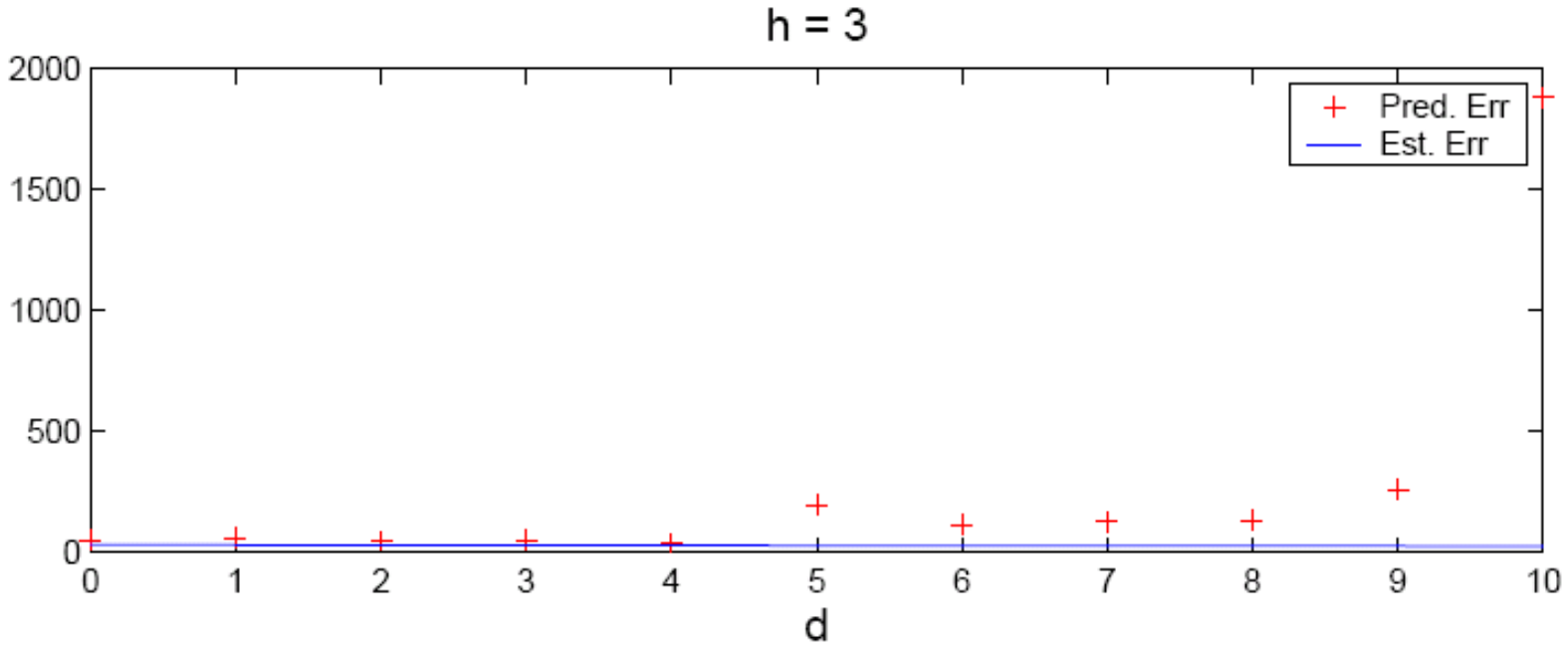


This is the *overfitting* problem: a better fit is not the best predictor – in the extreme case, a model can fit exactly the data and is unable to model it

How to avoid overfitting

Method 1: reserve some data for testing

The idea is to reserve a small fraction of the data set to test the model prediction. Consider for example Figure 6.5. We fitted the model in Eq.(6.5) with $h - 1 = 2$ harmonics and a polynomial of degree $d = 0$ to 10. The prediction error is defined here as the mean square error between the true values of the data at t=225 to 250 and the point predictions given by Theorem 6.2.1. The estimation error is the estimator s of σ. The smallest prediction error is for $d = 4$. The fitting error decreases with d, whereas the prediction error is minimal for $d = 4$. This method is quite



Method 2: Information Criteria

The log-likelihood $\log f_Y(y)$ was used to derive a score function to be minimized for model fitting. E.g. for a linear homoscedastic model with n data points: score $= -\log f_Y(y) = n \log \hat{\sigma} +$ constant,

with $\hat{\sigma} =$ maximum likelihood estimator of σ

To avoid overfitting, add a penalty term to the score

Akaike's Information Criteria: AIC $= -2 \log f_Y(y) + 2k$ where k is the number of (continuous) free parameters to be fitted.

E.g. for a linear homoscedastic model with parameter $\beta \in \mathbb{R}^p$ we have $k = p + 1$ and AIC $= 2n \log \hat{\sigma} + 2p +$ constant

AIC can be interpreted as the amount of information required to describe a new hypothetical sample when we estimate the model from one sample.

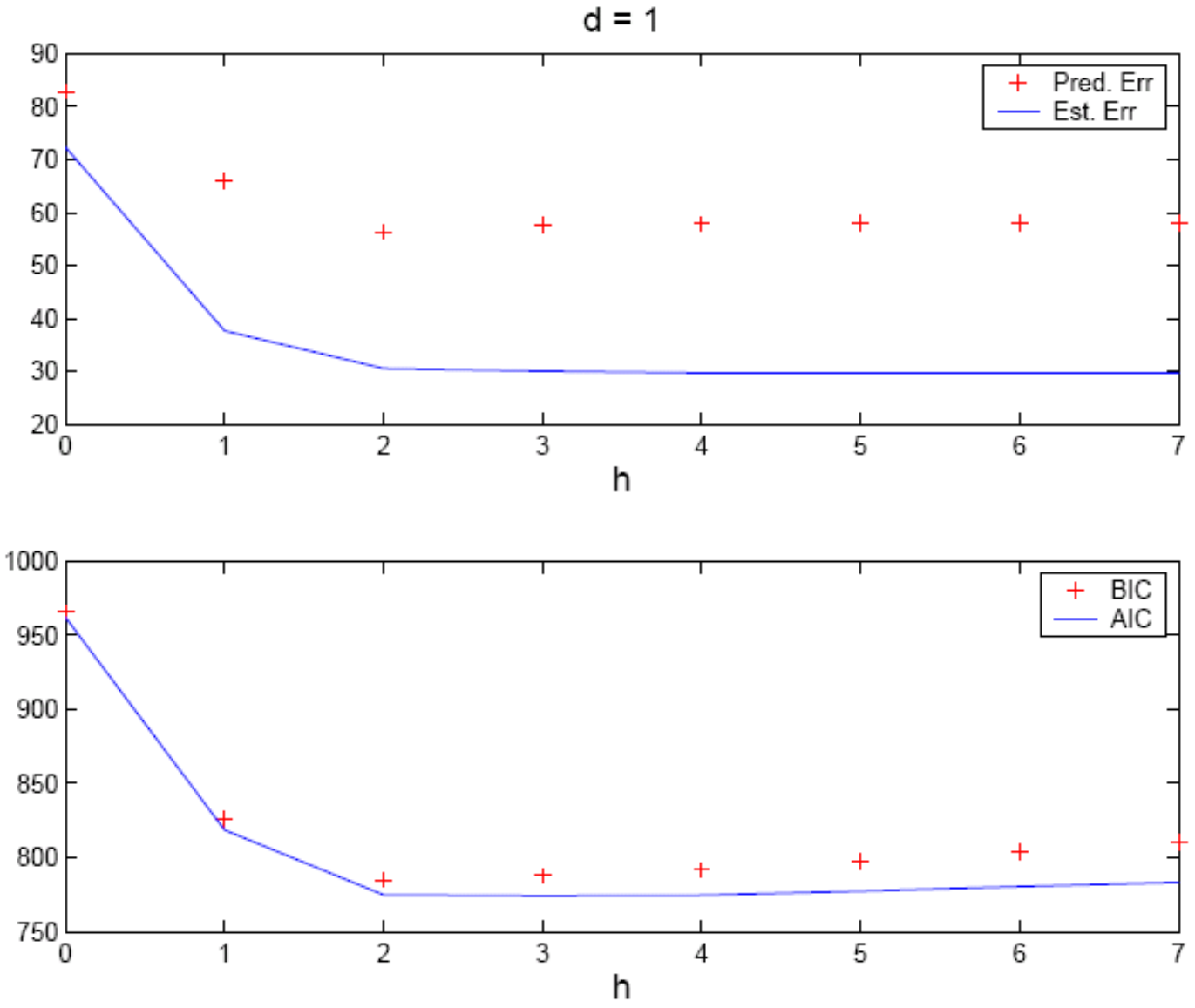
Other information criteria are also used. They are defined empirically.

For example, the Bayesian Information Criterion (BIC) is defined for linear regression models

$$\text{BIC} = 2n \log \hat{\sigma} + 2p \log n + \text{constant}$$

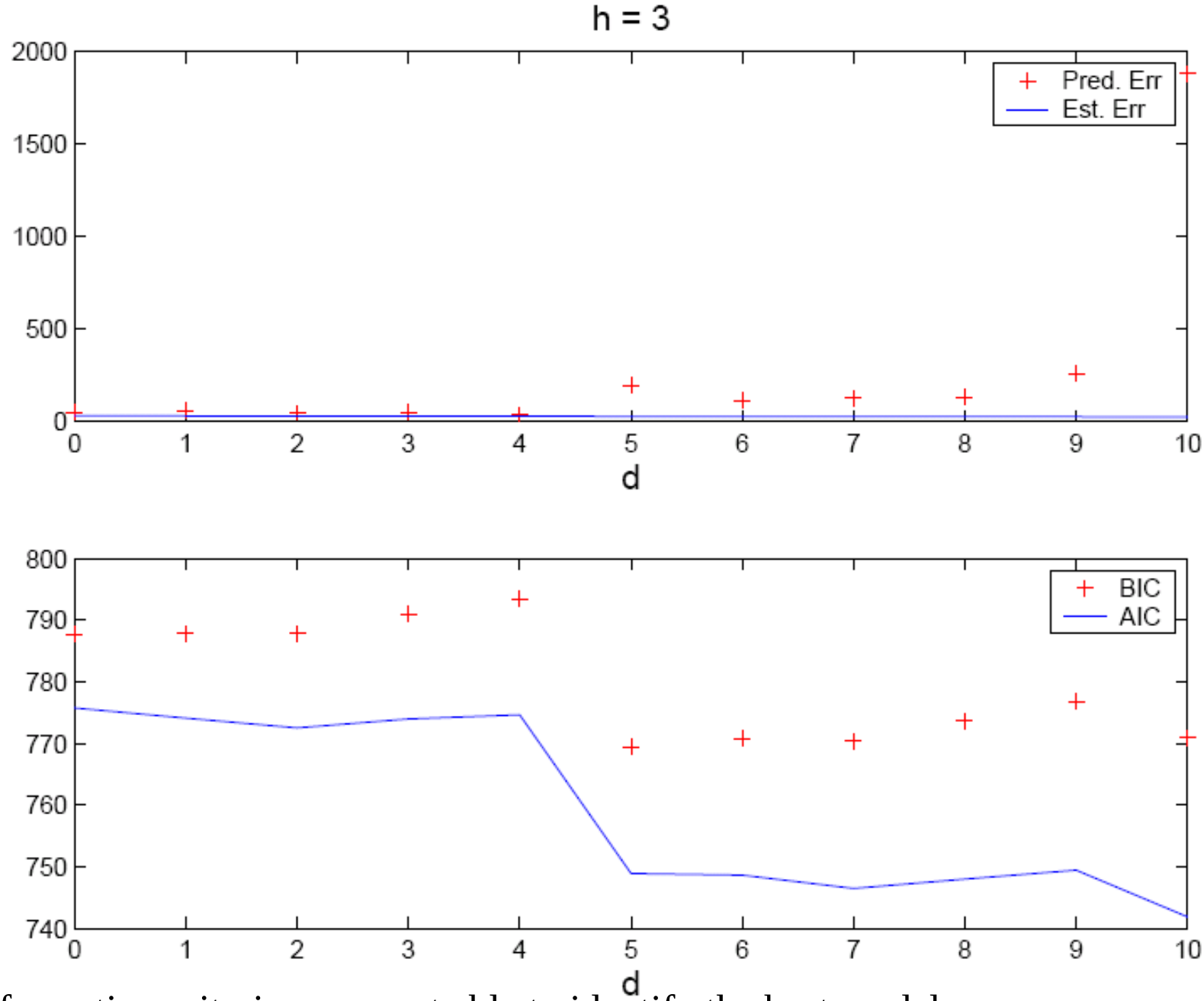
(gives more weight than AIC to model dimension p when the sample size n is large)

Best Model for Internet Data, $d = 1$, h up to 10



Information criterions are able to identify the best model

Best Model for Internet Data, $h = 3$, d up to 10



Information criterions are not able to identify the best model;
the polynomial models are not a good class of models

Say what is true

A. When doing the fit and if we use an information criterion, we can use all data available up to time t
B. When doing the fit and if we use a score + test data we can use all data available up to time t
C. A and B
D. None
E. I don't know

5. Use of Bootstrap

Assume we have a prediction model $Y_t = f_t(\beta) + \epsilon_t$

The estimation of β is done assuming some distribution for ϵ_t;

Assume this distribution is only approximately known; we can improve the prediction intervals if we use a better approximation of this distribution.

For example, we can use the principle of the Boostrap, i.e. estimate the distribution of ϵ_t by its empirical distribution.

THEOREM 2.5 (General IID Case). *Let $X_1, ..., X_n, X_{n+1}$ be an iid sequence and assume that the common distribution has a density. Let $X^n_{(1)}, ..., X^n_{(n)}$ be the order statistic of $X_1, ..., X_n$. For $1 \le j \le k \le n$:*

$$\mathbb{P}\left(X^n_{(j)} \le X_{n+1} \le X^n_{(k)}\right) = \frac{k-j}{n+1} \tag{2.32}$$

thus for $\alpha \ge \frac{2}{n+1}$, $[X^n_{(\lfloor (n+1)\frac{\alpha}{2} \rfloor)}, X^n_{(\lceil (n+1)\left(1-\frac{\alpha}{2}\right) \rceil)}]$ is a prediction interval at level at least $\gamma = 1-\alpha$.

Assume $Y_t = f_t(\beta) + \epsilon_t$ and apply theorem 2.5 to

$$X_1 = \epsilon_1, \dots, X_n = \epsilon_t, X_{n+1} = \epsilon_{t+\ell}$$

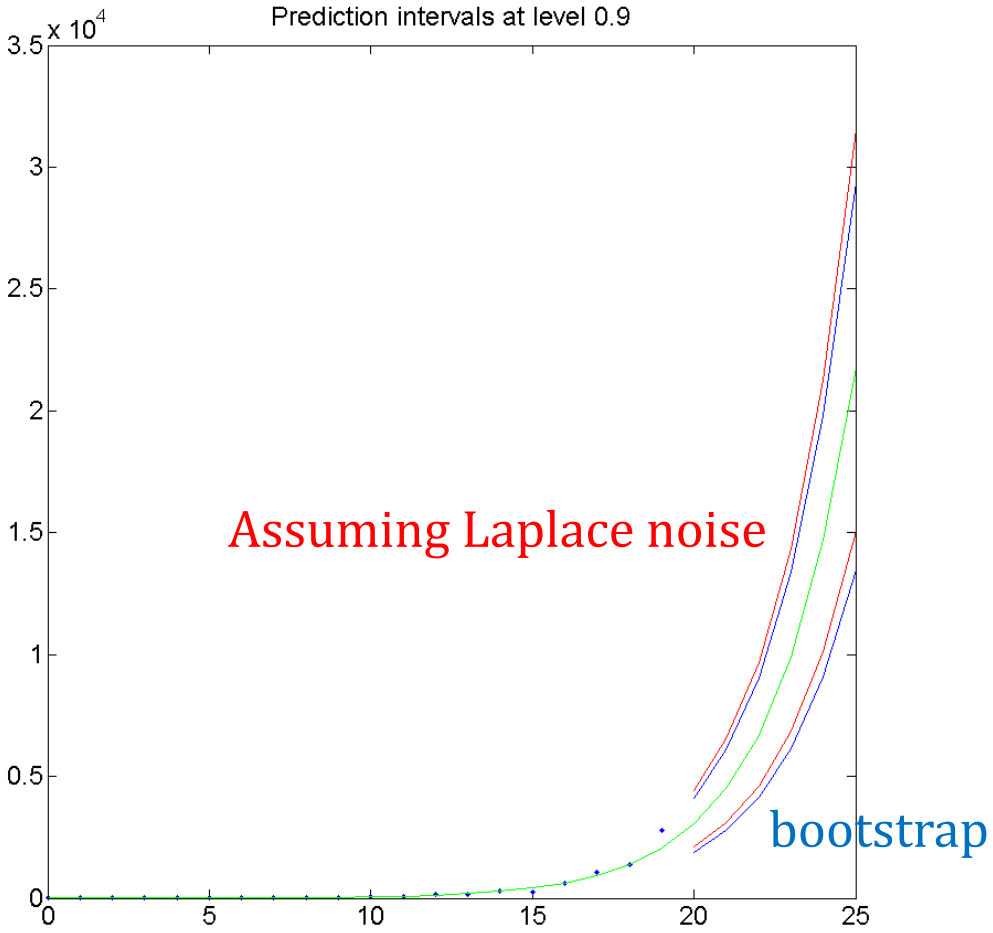
This gives the algorithm:

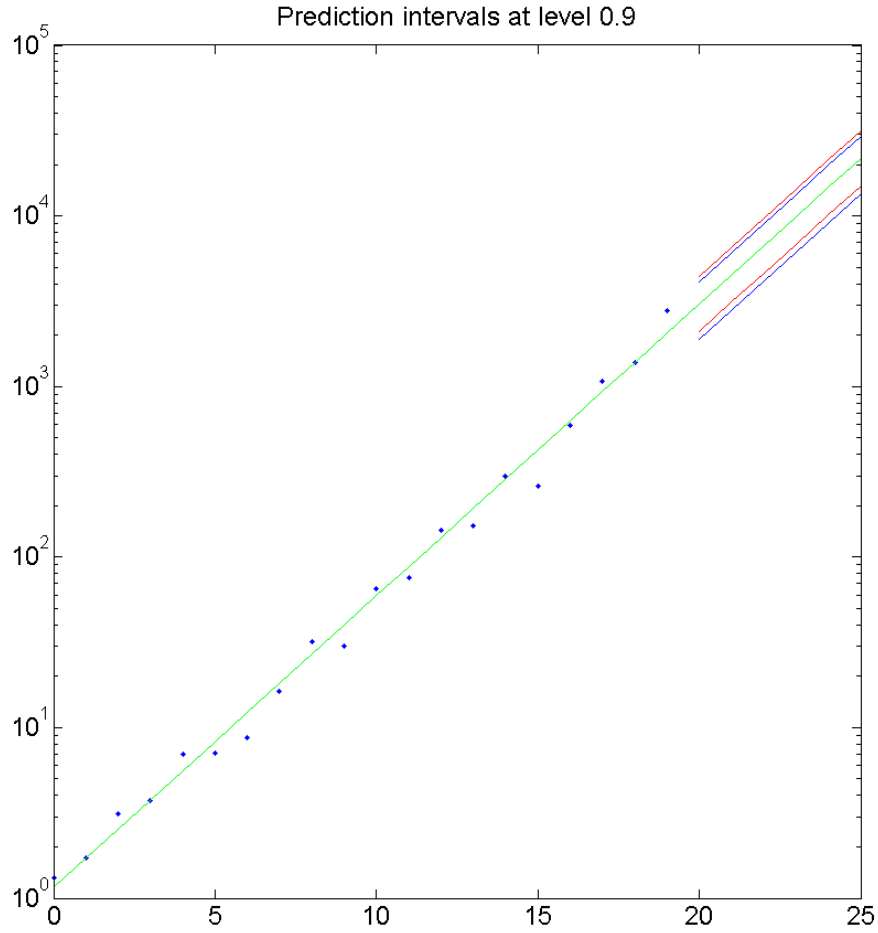
1. Estimate $\hat{\beta}$ by some method
2. Estimate residuals $e_t = Y_t - f_t\left(\hat{\beta}\right)$
3. (Thm 2.5) $\eta = e_{\lfloor \frac{(t+1)\alpha}{2} \rfloor}, \xi = e_{\lfloor (t+1)(1-\frac{\alpha}{2}) \rfloor}$

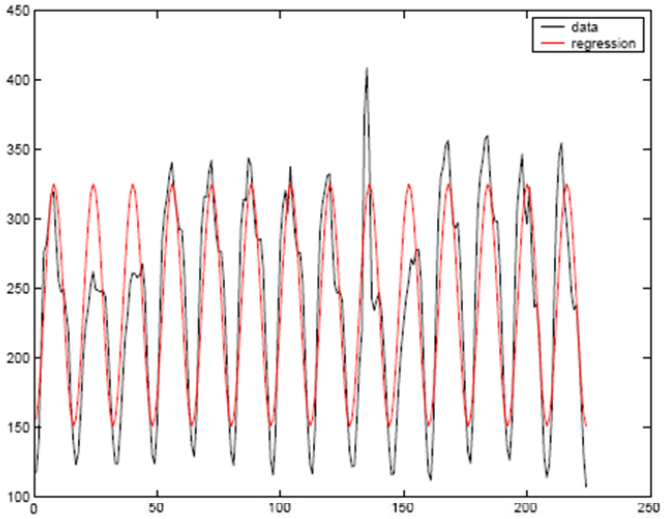
   (for $\alpha = 5\%, t = 100$: $\eta = e_{(2)}, \xi = e_{(99)}$)
4. Prediction interval for $Y_{t+\ell}$: $[f_{t+\ell}\left(\hat{\beta}\right) + \eta, f_{t+\ell}\left(\hat{\beta}\right) + \xi]$

Example

For this example, the bootstrap (done in log scale) gives asymmetric prediction interval







For this example, the bootstrap gives slightly smaller intervals than the ones based on gaussian noise