Non-linear Prediction

Nora Prean and Peter Lindner

18^{th} January 2011

Contents

Motivation

Basic idea of the presentation

- Get a short introduction to non-linear prediction (Fan & Yao Chapter 10)
- Set-up: Why $\longrightarrow$ Examples $\longrightarrow$ Bits of the theory

Why forecasting? What is it for?

- Who does not want to know the future?
- Policy decisions depend on forecasts (e.g. future development of GDP)
- How to become rich (Financial markets)
- In the end, that is why we do time series analysis

Simple Quadratic Model to understand the graphs and sensitivity to initial values

$$X_t = 0.235 X_{t-1}(16 - X_{t-1}) + \varepsilon_t$$

where

$$\varepsilon_t \sim U[-0.52, 0.52]$$

Now, we look at

- m=2 and 3 step ahead predictor $f_m(\bullet)$
- their conditional variance function
- and a comparison of the conditional variance.

m=2 step ahead predictor

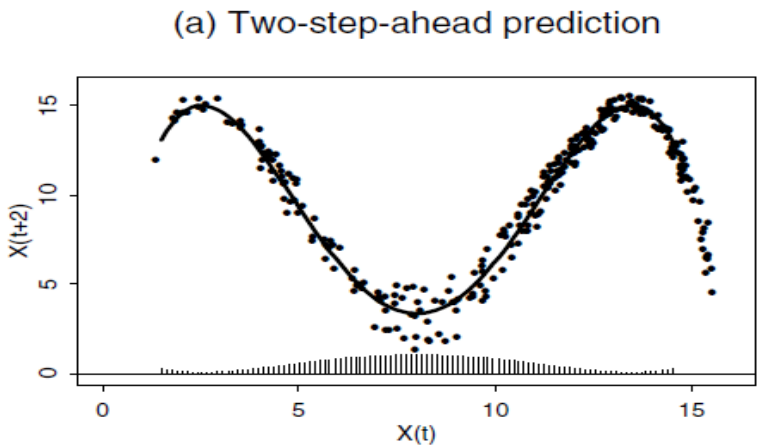

Figure: m=2 step ahead predictor with different initial values

- Predictive error depends on initial value
- Can be seen from the deviation from the dots and the line
- and the conditional variance

Legend

a Dots: Scatter plot X_{t+2} against X_t
b Solid line: 2-step ahead predictor $f_m(\bullet)$
c Impulses: Conditional variance $\sigma_m^2(\bullet)$

m=3 step ahead predictor

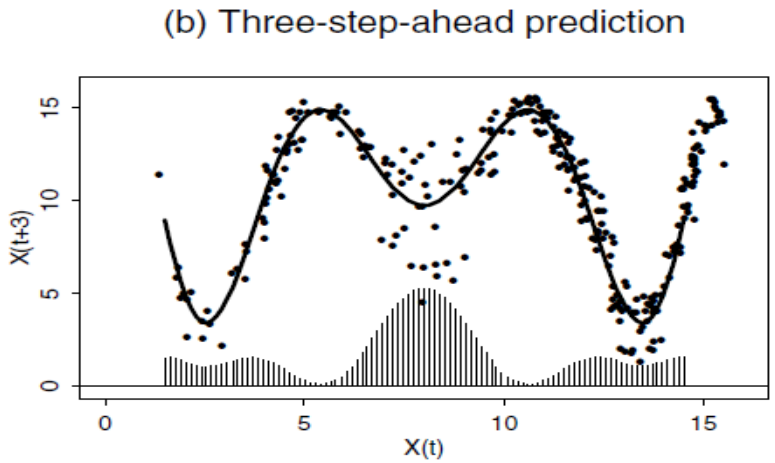

Figure: m=3 step ahead predictor with different initial values

- Predictive error may be larger or smaller than for the m=2 step ahead predictor

Legend

a Dots: Scatter plot X_{t+3} against X_t

b Solid line: 3-step ahead predictor $f_m(\bullet)$

c Impulses: Conditional variance $\sigma_m^2(\bullet)$

Back

Conditional Variance

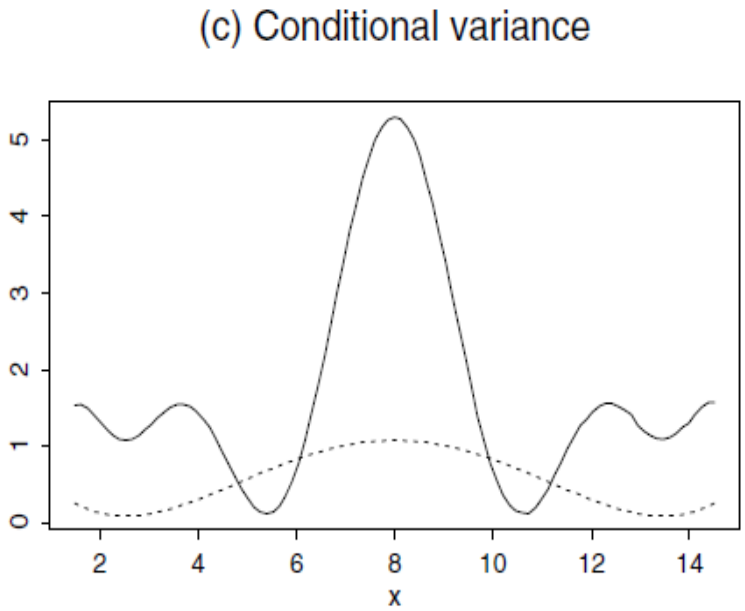

Figure: Conditional variance depending on initial values

- In the ranges where the solid line is below the dotted line, the 3-step predictor is more accurate than the 2-step predictor ▸ More

Legend

a Dotted line: Conditional variance of the 2-step-ahead predictor
b Solid line: Conditional variance of the 3-step-ahead predictor

Least squares predictor

- Observations from time series process $X_1, \ldots X_T$
- Predict X_{T+m} $(m \geq 1)$ based on last p observed values $(X_T, X_{T-1}, \ldots X_{T-p+1})^\tau \equiv \mathbf{X}_T$
- Predictor: $f_{T,m}(\mathbf{X}_T) = \text{arg inf } E\{X_{T+m} - f(\mathbf{X}_T)\}^2$
  - compare OLS: $\hat{\beta} = \text{arg min } S(b) = (X'X)^{-1}X'y$ with $S(b) = \sum_{i=1}^{n}(y_i - x_i'b)^2 = (y - Xb)'(y - Xb)$
- $\longrightarrow f_{T,m}(\mathbf{x}) = E(X_{T+m}|\mathbf{X}_T = \mathbf{x})$

Conditional variances given X_T

- Mean Square Predictive Error of $f_{T,m}$:

$$E\{X_{T+m} - f_{T,m}(\mathbf{X}_T)\}^2$$

$$= E[E\{(X_{T+m} - f_{T,m}(\mathbf{X}_T))^2 | \mathbf{X}_T\}]$$

$$= E\{Var(X_{T+m}|\mathbf{X}_T)\}$$

$\rightarrow$ average of conditional variances of X_{T+m} given $\mathbf{X}_T$

  - Note that for X_T being a linear AR(p) conditional variance is constant:

$$\sigma^2_{T,m}(\mathbf{x}) \equiv Var(X_{T+m}|\mathbf{X}_T = \mathbf{x})$$

- CONDITIONAL Mean Square Predictive Error:

$$E[\{X_{T+m} - f_{T,m}(\mathbf{x})\}^2 | \mathbf{X}_T = \mathbf{x}] = \sigma^2_{T,m}(\mathbf{x}) \qquad (1)$$

Decomposition of conditional mean square error

- Decomposition: $E[\{X_{T+m} - f_{T,m}(\mathbf{x})\}^2 | \mathbf{X}_T = \mathbf{x} + \delta]$

$$= \sigma^2_{T,m}(\mathbf{x} + \delta) + \{\delta^\tau \dot{f}_{T,m}(x)\}^2 + o(||\delta||^2) \qquad (2)$$

  - In non-linear time series we might not neglect the error coming from the drift δ!
  - For non-linear processes both types of errors may be amplified rapidly at some places in the state-space $\rightarrow$ hence it is important for predictions at which point we are!

Noise Amplification

- Assume simple model

$$X_t = f(X_{t-1}) + \epsilon_t$$

with $\{\epsilon_t\} \sim IID(0, \sigma^2)$ and ϵ_t is independent of $\{X_{t-k}, k \geq 1\}$

- From Markov property it follows that

$$f_{T,m}(\mathbf{x}) = E(X_{T+m}|\mathbf{X}_T = \mathbf{x}) \equiv f_m(x)$$

and

$$\sigma^2_{T,m}(\mathbf{x}) = Var(X_{T+m}|\mathbf{X}_T = \mathbf{x}) \equiv \sigma^2_m(\mathbf{x})$$

where x is the first component of $\mathbf{x}$.

- For linear processes, e.g. an AR(1) process, the noise entering at a fixed point in time decays exponentially as m increases. This noise contraction is not necessarily observed in non-linear processes.
- For $X_t = f(X_{t-1}) + \epsilon_t$

$$\sigma_m^2(x) = Var(X_m|X_0 = x) = \mu_m(x)\sigma^2 + O(\zeta^3)$$

with

$$\mu_m(x) = 1 + \sum_{j=1}^{m-1} \left\{ \prod_{k=j}^{m-1} \dot{f}[f^{(k)}(x)] \right\}^2$$

- $\sigma_m^2(x)$ varies with x
- $\mu_m(x)$ dictates noise amplification (for linear processes $\sigma_m^2(x)$ and $\mu_m(x)$ are constant).
- Values of μ_m are determined by those of the derivative $\dot{f}$
- If $|\dot{f}(.)| > 1$ defines large state-space, $\mu_m(.)$ can be large even for small m ▸ See ▸ Maybe Conclusion

Predictive Distribution: Different estimators

- Nadaraya-Watson estimator (NW)
- Local linear regression estimator (LL)
- adjusted Nadaraya-Watson estimator (ANW)
- Local logistic estimator (LG-2)

Compare them using mean absolute deviation error (MADE)

$$MADE = \frac{\sum_i |F_e(y_i|x_i) - F(y_i|x_i)| I\{0.001 \leq F(y_i|x_i) \leq 0.999\}}{\sum_i I\{0.001 \leq F(y_i|x_i) \leq 0.999\}}$$

Example II: Model

$$Y_t = 3.76\,Y_{t-1} - 0.235\,Y_{t-1}^2 + 0.3\varepsilon_t$$

where
ε_t independent with common distribution $U[-0.52, 0.52]$

- Look at conditional distribution function $z = F(y|x)$ for m=2 and 3
- and a compare the predicted distribution using MADE.

Conditional Distribution Function for $m = 2$

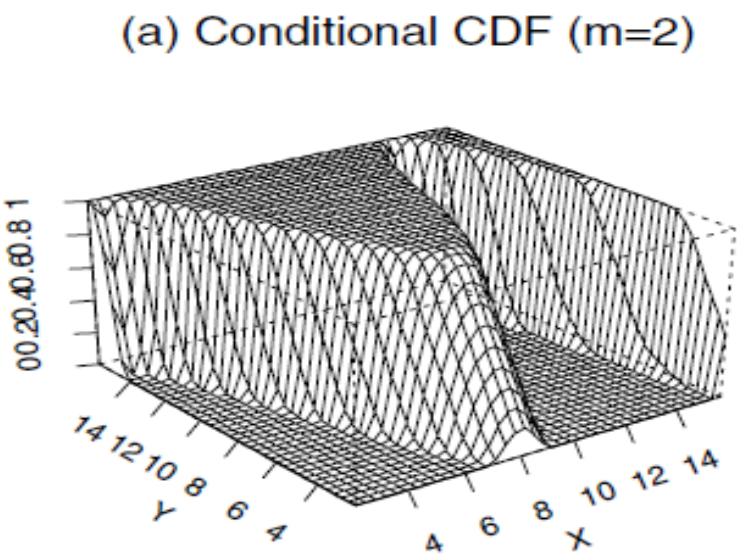

Figure: CDF for $m = 2$

Legend

a X-Axis: past observed values

b Y-Axis: predicted values

c Vertical axis: Probability $z = F(y|x)$

Conditional Distribution Function for $m = 3$

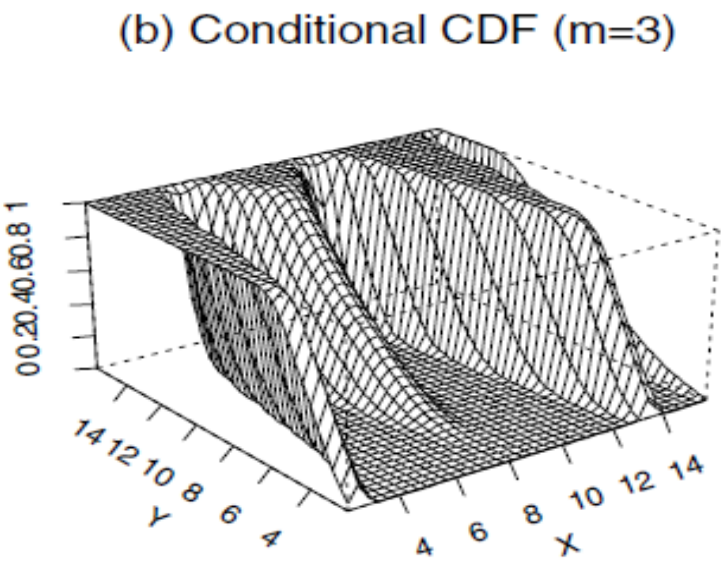

Figure: CDF for $m = 3$

Legend

a X-Axis: past observed values

b Y-Axis: predicted values

c Vertical axis: Probability $z = F(y|x)$

Comparison of the Conditional Distribution Function

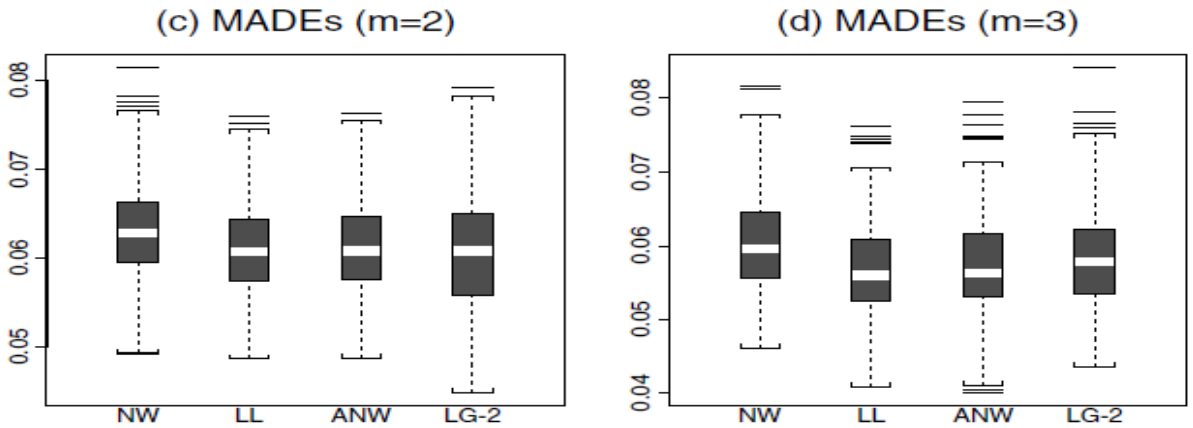

Figure: Comparison of the estimated conditional distribution function

- Authors claim NW is considerable worse than the other estimators
- I do not see this, they look rather similar

Legend

a NW: Nadaraya-Watson estimator
b LL: Local linear regression estimator
c ANW: adjusted Nadaraya-Watson estimator
d LG-2: Local logistic estimator

Estimate predictive distributions

- Generally: forecast a predictive interval/predictive set
  - "All information on the future is [...]contained in a **predictive distribution function**, which is in fact a **conditional** distribution of a **future** variable given the **present** state." (F&Y, p. 454)
- Linear time series: predictive distributions are normal (hence, simply estimate means and variances)
- Non-linear time series: pred. distributions usually **not** normal Furthermore: even if process is generated by **parametric** non-linear model the multiple-step-ahead predictive distributions are of unknown form and may only be estimated in a **non-parametric** manner

Local linear regression estimator (LL)

- Assume: data from str. stat. stochastic process $\{(\mathbf{X}_t, Y_t)\}$ where $\mathbf{X}_t = (X_t, ..., X_{t-p+1})^\tau$ typically denotes a vector of lagged values of $Y_t = X_{t+m}$ for some $m \geq 1$
  - Estimate conditional distribution function

$$F(y|\mathbf{x}) \equiv P(Y_t \leq y|\mathbf{X}_t = \mathbf{x})$$

Rewrite $I(Y_t \leq y) = Z_t$, then

$$F(y|\mathbf{x}) = E(Z_t|\mathbf{X}_t = \mathbf{x})$$

Hence, estimation problem can be viewed as regression of Z_t on $\mathbf{X}_t$ by local linear technique (see F&Y 8.2)

- $\longrightarrow$ yields our **Local Linear regression estimator** (LL)
- Problem: estimator $\hat{F}(y|\mathbf{x})$ is not necessarily a CDF

Adjusted Nadaraya-Watson Estimator (ANW)

- Nadaraya-Watson kernel regression: estimate expectation as a locally weighted average, using a kernel as a weighting function
- Let $p_t = p_t(x)$ denote weights with the following properties: each $p_t \geq 0$, $\sum_t p_t = 1$ and

$$\sum_{t=1}^{T} p_t(x)(X_t - x)K_h(X_t - x) = 0$$

- Now define

$$\tilde{F}(y|x) = \frac{\sum_{t=1}^{T} I(Y_t \leq y)p_t(x)K_h(X_t - x)}{\sum_{t=1}^{T} p_t(x)K_h(X_t - x)}$$

- $\longrightarrow 0 \leq \tilde{F}(y|x) \leq 1$, $\tilde{F}$ is monotone in y

Conclusions

- *Linear* prediction methods still dominant in time series forecasting
- Linear prediction does well, whenever time series is covariance stationary (finite second moments)
- Nevertheless, the best *linear* predictor is not the least squares predictor in general and hence not the best estimator
- Life (real-life generating processes) is not always linear!
- Initial value sensitivity

Literature

Nonlinear Time Searies: Nonparametric and Parametric Methods

Fan, Jianqing and Yoa, Qiwei, Springer Series in Statistics (2003); especially Chapter 10

Quantifying the inference of initial values on nonlinear prediction

Yao, Q. and Tong, H. (1994); Journal of the Royal Statistical Society, Series B, 56, 701-725.

THANKS FOR YOUR ATTENTION

Modified Simple Quadratic Model: Point Prediction

$$X_{it} = 0.23X_{t-1}(16 - X_{t-1}) + 0.4\varepsilon_t$$

where

$$\varepsilon_t \sim iidN[0, 1]$$

on the interval $[-12, 12]$.

- draw a sample of 1,200 data points
- $\sigma_1^2(x) = 0.16$ due to iid normality, thus m=1 is not reported in the book
- look at m=2, 3, and 4 step ahead predictor for sample point 1001 to 1200 and compare them to actual values

m=2 step ahead predictor

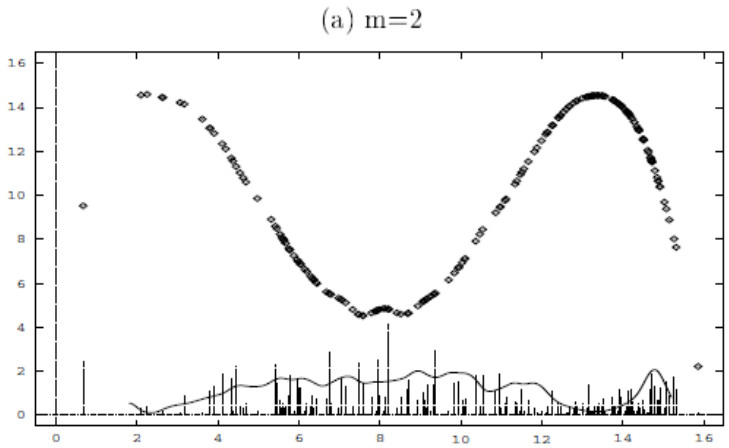

Figure: m=2 step ahead predictor with bandwidth $h = 0.25$

Legend

a Diamonds: Predicted values

b Solid line: Estimated conditional variance $\hat{\sigma}^2_m(\bullet)$

c Impulses: Absolute errors

m=3 step ahead predictor

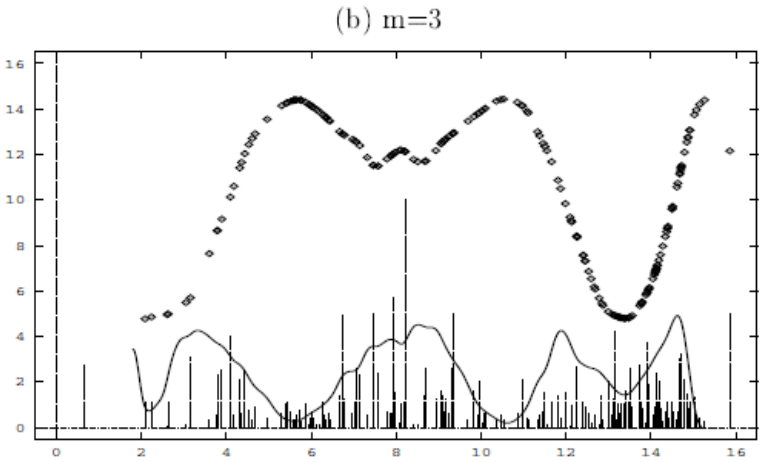

Figure: m=3 step ahead predictor with bandwidth $h = 0.2$

Legend

a Diamonds: Predicted values

b Solid line: Estimated conditional variance $\hat{\sigma}^2_m(\bullet)$

c Impulses: Absolute errors

m=4 step ahead predictor

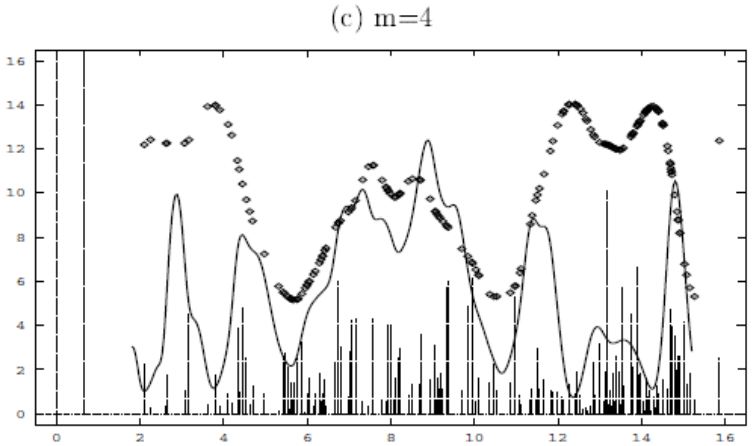

Figure: m=4 step ahead predictor with bandwidth $h = 0.18$

Legend

a Diamonds: Predicted values

b Solid line: Estimated conditional variance $\hat{\sigma}_m^2(\bullet)$

c Impulses: Absolute errors

Back