Nonlinear Prediction

Chapter 10 – Fan/Yao book

Monika Turyna and Ulrich Gunter
Department of Economics, University of Vienna

January 12th, 2010

Features of Nonlinear Prediction (Section 10.1) – Decomposition of Mean Square Predictive Errors

Least squares m-step ahead predictor of time-series process $\{X_t\}$ taken over all measurable functions of $\mathbf{X}_T$ is defined as:

$$f_{T,m}(\mathbf{X}_T) = \arg\inf_f E\{X_{T+m} - f(\mathbf{X}_T)\}^2 \tag{1}$$

where T denotes forecast origin, m $(m \geq 1)$ denotes forecast horizon, and $\mathbf{X}_T$ denotes last p observed values of available data $X_1, ..., X_T$ only

Let $\mathbf{x}$ denote observed value of $\mathbf{X}_T$:

$$\Rightarrow f_{T,m}(\mathbf{x}) = E(X_{t+m}|\mathbf{X}_T = \mathbf{x}) \tag{2}$$

Corresponding *mean square predictive error* (average of conditional variances) is given by:

$$E\{X_{T+m} - f(\mathbf{X}_T)\}^2 = E\{Var(X_{T+m}|\mathbf{X}_T)\} \quad (3)$$

If $\{X_t\}$ were linear AR(p) process, conditional variance $\sigma^2_{T,m} \equiv Var(X_{T+m}|\mathbf{X}_T = \mathbf{x})$ would be constant

For nonlinear processes, this is not true in general:

- ⇒ *Conditional mean square predictive error* more relevant measure of predictive performance
- ⇒ Goodness of prediction depends on where we are
- ⇒ Prediction from a nonlinear point of view "one-step closer to reality"

Conditional mean square predictive error reads:

$$E[\{X_{T+m} - f_{T,m}(\mathbf{x})\}^2 | \mathbf{X}_T = \mathbf{x}] = \sigma^2_{T,m}(\mathbf{x}) \quad (4)$$

True and unobserved value of $\mathbf{X}_T = \mathbf{x} + \delta$, where δ denotes a small drift due to measurement error, experimental error and/or so on

Hence, for least squares m-step ahead predictor $f_{T,m}(\mathbf{X}_T)$ subsequent decomposition of conditional mean square predictive error holds (see Fan/Yao 2003, pp. 442-443 for a proof):

$$\begin{aligned} &E[\{X_{T+m} - f_{T,m}(\mathbf{x})\}^2 | \mathbf{X}_T = \mathbf{x} + \delta] \\ &= \sigma^2_{T,m}(\mathbf{x} + \delta) + \{\delta^\tau \dot{f}_{T,m}(\mathbf{x})\}^2 + o(||\delta||^2) \quad (5) \end{aligned}$$

where $\dot{f}_{T,m}$ denotes gradient vector of $f_{T,m}$

As shown by YAO/TONG (1998), conditional variance $\sigma^2_{T,m}(\mathbf{x} + \delta)$ is not necessarily dominant term in case of nonlinear processes

$\Rightarrow$ Error due to drift δ no longer negligible

Noise Amplification

For a linear AR(1) process with coefficient b ($|b| < 1$) mean square predictive error reads:

$$\sigma^2 \sum_{j=0}^{m-1} b^{2j} = \sum_{j=0}^{m-1} b^{2j} Var(\varepsilon_{T+1+j}) \tag{6}$$

where noise entering at a fixed time exponentially decays as m increases

For a time-series process $\{X_t\}$ (not necessarily stationary) generated by nonlinear AR model

$$X_t = f(X_{t-1}) + \varepsilon_t \tag{7}$$

with $\{\varepsilon_t\} \sim IID(0, \sigma^2)$, ε_t independent of $\{X_{t-k}, k \geq 1\}$, and $|\varepsilon_t| \leq \zeta$ ($\zeta > 0$) o.c.s (see FAN/YAO 2003, p. 444):

$$\sigma_m^2(x) = Var(X_m|X_0 = x) = \mu_m(x)\sigma^2 + O(\zeta^3) \tag{8}$$

where

$$\mu_m(x) = 1 + \sum_{j=0}^{m-1} \left\{ \prod_{k=j}^{m-1} \dot{f}[f^{(k)}(x)] \right\}^2 \tag{9}$$

- For linear processes $\dot{f}(\cdot)$ is constant and therefore $\mu_m(x)$ and $\sigma_m^2(x)$ are constant
- If, however, $|\dot{f}(\cdot)| > 1$ on a large part of the state space, $\mu_m(x)$ and $\sigma_m^2(x)$ can be very large for even very small m

$\Rightarrow$ Only very short-range prediction is practically meaningful

Sensitivity to Initial Values

Divergence of conditional expected values of two trajectories based on different initial values ($x + \delta$ versus x) is given by:

$$E\{X_m(x+\delta)|X_0 = x+\delta\} - E\{X_m(x)|X_0 = x\} = \delta \dot{f}_m(x) + o(||\delta||) \quad (10)$$

where

$$\dot{f}_m(x) = E\left\{\prod_{k=1}^{m} \dot{f}(X_{k-1})|X_0 = x\right\} \quad (11)$$

If again $|\dot{f}(\cdot)| > 1$ on a large part of the state space, $\dot{f}_m(x)$ can be very large for even very small m

Multi-Step Prediction versus a One-Step Plug-in Method

One-step plug-in predictor for X_{T+m} based on model (7) is given by $f^{(m)}(X_T)$, which differs from least square m-step ahead predictor $f_m(X_T) = E(X_{T+m}|X_T)$ unless $f(\cdot)$ is linear

Hence,

$$E[\{X_{T+m} - f^{(m)}(X_T)\}^2|X_T] \geq E[\{X_{T+m} - f_m(X_T)\}^2|X_T] \qquad (12)$$

⇒ One-step plug-in method not desirable in principle

⇒ Suggestion to stick to least square m-step ahead predictor

Nonlinear versus Linear Prediction

- Empirical studies suggest that linear prediction methods often perform well despite their simplicity and that gains from nonlinear prediction are not always statistically significant (see CHATFIELD 2001)
- Linear prediction methods can be applied to any time series as long as it has finite second moments

Let $\{X_t\}$ be a covariance-stationary time-series process and let us seek best linear predictor (predictor that is a linear combination of $\{X_{t-k}, k \geq 1\}$) such that mean square error is minimized

Wold decomposition theorem yields:

$$X_t = e_t + \sum_{j=1}^{\infty} \psi_j e_{t-j} + V_t \tag{13}$$

where $\{e_t\} \sim N(0, \sigma^2)$ and

$$e_t = X_t - \sum_{i=1}^{\infty} \varphi_i X_{t-i} \tag{14}$$

with V_t purely deterministic and $\{\psi_j\}, \{\varphi_i\}$ each square-summable coefficients

Hence,

$$E(X_t|X_{t-k}, k \geq 1) \neq \sum_{i=1}^{\infty} X_{t-i} \equiv \hat{X}_t \tag{15}$$

- $\hat{X}_t$ is *best linear predictor* as it minimizes $E\{X_t - \sum_{i=1}^{\infty} b_i X_{t-i}\}$ for square-summable coefficients $\{b_i\}$
- Mean square error of $\hat{X}_t$ is $E(X_t - \hat{X}_t)^2 = E(e_t^2) = \sigma^2$
- However, best linear predictor is not least squares predictor in general and therefore not best estimator

Point Prediction (Section 10.2) – Local Linear Predictors

$f(\cdot)$ and $\dot{f}(\cdot)$ can be estimated by applying *local linear regression*, which is a nonparametric regression technique (see FAN/YAO 2003, pp. 314-317)

Let $\hat{f}_m(\mathbf{x}) = \hat{a}, \hat{\dot{f}}_m(\mathbf{x}) = \hat{\mathbf{b}}$, and $(\hat{a}, \hat{\mathbf{b}})$ be minimizer of subsequent sum:

$$\sum_{t=p}^{T-m} \{X_{T+m} - a - \mathbf{b}^{\tau}(\mathbf{X}_T - \mathbf{x})\} K\left(\frac{\mathbf{X}_T - \mathbf{x}}{h(T)}\right) \tag{16}$$

where $K(\cdot)$ is a kernel function and $h(T)$ a bandwidth

Calculation yields:

$$\hat{f}_m(\mathbf{x}) = \frac{T_0(\mathbf{x}) - S_1^\tau(\mathbf{x})S_2^{-1}(\mathbf{x})T_1(\mathbf{x})}{S_0(\mathbf{x}) - S_1^\tau(\mathbf{x})S_2^{-1}(\mathbf{x})S_1(\mathbf{x})} \quad (17)$$

$$\hat{\dot{f}}_m(\mathbf{x}) = \frac{S_1(\mathbf{x})T_0(\mathbf{x})/S_0(\mathbf{x}) - T_1(\mathbf{x})}{S_2(\mathbf{x}) - S_1(\mathbf{x})S_1^\tau(\mathbf{x})/S_0(\mathbf{x})} \quad (18)$$

where $S_0(\mathbf{x}), S_1(\mathbf{x}), S_2(\mathbf{x}), T_0(\mathbf{x}), T_1(\mathbf{x})$ are given in FAN/YAO (2003, p. 451)

- $\hat{f}_m(\mathbf{x})$ is mean square consistent since $E[\{f_m(\mathbf{x}) - \hat{f}_m(\mathbf{x})\}^2|\mathbf{X}_T = \mathbf{x} + \delta] \to 0$ as $T \to \infty$
- Decomposition of conditional mean square predictive error (5) still holds asymptotically

Predictive distributions – Introduction

- For linear time series with normally distributed errors, the predictive distributions are normal – predictive intervals are easily obtained
- Mean $\pm$ a multiple of standard deviation
- Used also for some non–linear models (e.g. threshold autoregressive models)
- Skewed distributions occur even if errors have symmetric distributions
- Most generally we want to estimate $F(y|\mathbf{x}) \equiv P(Y_t \leq y|\mathbf{X_t} = \mathbf{x})$
- If we write $Z_t = I(Y_t \leq y)$ then $E(Z_t|\mathbf{X_t} = \mathbf{x}) = F(y|\mathbf{x})$ and estimation may be seen as regression of Z_t on $\mathbf{X_t}$

Estimators for $F(\cdot|\mathbf{x})$

- Local logistic estimator:
  - A generalized local logistic model for $P(x)$ has the form

$$L(x;\theta) \equiv \frac{A(x;\theta)}{\{1 + A(x;\theta)\}}$$

    where $A(x;\theta)$ denotes a nonnegative function that depends on a vector of parameters $\theta = (\theta_1, \ldots, \theta_r)$ that represents the values of $P(x), P^{(1)}(x), \ldots, P^{(r-1)}(x)$
  - Fitting this model locally to indicator–function data leads to an estimator $\hat{F}(y|\mathbf{x}) \equiv L(0;\hat{\theta})$ where $\hat{\theta}$ minimizes

$$R(\theta;x;y) = \sum_{t=1}^{T} \{I(Y_t \le y) - L(X_t - x, \theta)\}^2 K_h(X_t - x)$$

- Adjusted Nadaraya–Watson estimator:
  - Let $p_t = p_t(x)$ for $1 \leq t \leq T$, denote weights with the property that $p_t \geq 0$, $\sum_t p_t = 1$ and

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

  - Estimator:

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

  - $\tilde{F}$ is first–order equivalent to local linear estimator

Minimum–Length Predictive Sets

- $\{Y_t, \mathbf{X_t}\}$ is a strictly stationary process
- $Y_t = X_{t+m}$ for some $m \geq 1$ and $\mathbf{X_t} = (X_t, \cdots, X_{t-p+1})$
- General form of the predictive set is $P\{X_{T+m} \in \Omega_m(x) | X_T = x\} = \alpha$
- We restrict attention to $\mathcal{C}$ a class of measurable subsets of R (usually $\mathcal{C}$ consists of all intervals in R)
- Define: $\mathcal{C}_\alpha(x) = \{C \in \mathcal{C} : F(C|x) \geq \alpha\}$
- Minimum–Length Predictor: *The set in $\mathcal{C}_\alpha(x)$ with the minimum Lebesgue measure is called the minimum length predictor for Y_t based on $\mathbf{X_t} = \mathbf{x}$ in $\mathcal{C}$ with coverage probability α, which is denoted $M_\mathcal{C}(\alpha|\mathbf{x})$.*
- If the conditional density $g(y|\mathbf{x})$ of Y_t given $\mathbf{X_t} = \mathbf{x}$ exists than the minimum–length predictor is given by

$$\{y : g(y|\mathbf{x}) \geq \lambda_\alpha\}$$

Estimation of Minimum–Length Predictors

- Three steps:
  - Estimating the conditional distribution $F(\cdot|\mathbf{x})$
  - Specifying the set $\mathcal{C}$
  - Searching for $M_{\mathcal{C}}(\alpha|\mathbf{x})$ with F replaced by its estimator
- Illustration with Nadaraya–Watson estimator:

$$\hat{F}(C\mathbf{x}) = \frac{\sum_{t=1}^{T} I(Y_t \in C) K\left(\frac{\mathbf{X_t}-x}{h}\right)}{\sum_{t=1}^{T} K\left(\frac{\mathbf{X_t}-x}{h}\right)}$$

- We replace then F with $\hat{F}$ to obtain a minimum–length predictor

$$\hat{M}_{\mathcal{C}}(\alpha|\mathbf{x}) = \arg\min_{C \in \mathcal{C}} \{Leb(C) : \hat{F}(C|\mathbf{x} \geq \alpha)\}$$

with true coverage probability

$$\hat{\alpha} \equiv F\{\hat{M}_{\mathcal{C}}(\alpha|\mathbf{x})|\mathbf{x}\}$$

which converges to α

Predictive Sets based on conditional density

- Let $g(\cdot|\mathbf{x})$ be the conditional density of Y_t given $\mathbf{X_t} = \mathbf{x}$
- The minimum–length predictor may be defined as

$$M(\alpha|x) = \{y : g(y|\mathbf{x}) \leq \lambda_\alpha\}$$

where λ_α is the maximum value for which

$$\int_{\{y:g(y|\mathbf{x})\leq\lambda_\alpha\}} g(y|\mathbf{x})dy \leq \alpha$$

- Does not require specification of candidate $\mathcal{C}$