Lecture 4: Optimal prediction and Holt-Winters forecasting

Josep Vives

March 3, 2024

1 Introduction

Assume we have observed n variables $X_1, \cdots, X_n$. Denote the realized data by $x_1, \ldots, x_n$. Assume we want to predict the value of a non observed random variable Y. In the Time Series framework, this means, for example, to predict X_{n+1}, X_{n+p} or $f(X_{n+p})$ for a certain function f. The question is: given $x_1, \ldots, x_n$, what can we say about Y?

In the second order framework, the natural answer is to search for a function

$$f(X_1, \ldots, X_n)$$

such that

$$\mathbb{E}[(Y - f(X_1, \ldots, X_n))^2]$$

is minimal.

It is well-known, see for example [1] (Section 2.7), that the solution of this problem is the conditional expectation, that is,

$$f(X_1, \cdots, X_n) = E[Y \mid X_1, \ldots, X_n].$$

This function is called, usually, the optimal predictor of Y in terms of $X_1, \ldots, X_n$.

But to compute the conditional expectation is sometimes difficult. So, frequently, the search for a predictor of Y is restricted to the class of affine functions

$$f(X_1, \ldots, X_n) = b_0 + b_1 X_1 + \cdots + b_n X_n$$

and so, the problem reduces to search for the coefficients $b_0, b_1, \ldots, b_n$ such that

$$\mathbb{E}[(Y - b_0 - b_1 X_1 - \cdots - b_n X_n)^2]$$

is minimal. The solution is the so-called optimal linear predictor of Y in terms of $X_1, \ldots, X_n$. It is written as

$$P(Y \mid X_1, \ldots, X_n) := \widehat{b}_0 + \widehat{b}_1 X_1 + \cdots + \widehat{b}_n X_n.$$

Naturally, quantities

$$\mathbb{E}[(Y - E(Y \mid X_1, \ldots, X_n))^2]$$

and

$$\mathbb{E}[(Y - P(Y \mid X_1, \ldots, X_n))^2]$$

are the corresponding prediction errors.

The abstract idea behind the concept of prediction under a second order framework is the following. Consider a Hilbert space (or an Euclidean space) H. Associated to this space we have a scalar product and the concept of orthogonality. Consider a Hilbert subspace $G \subseteq H$. It is well-known that the best prediction of $h \in H$ by elements of G is the element $g := P_G h \in G$, that is, the orthogonal projection of h onto G. It is known that this projection exists and is unique.

The optimal prediction is the projection over the subspace generated by the random variables that are measurable with respect the $\sigma-$algebra generated by $X_1, \ldots, X_n$, and the optimal linear prediction is the projection over the linear subspace generated by $X_1, \ldots, X_n$.

2 Example

Let's begin with a simple but illustrative example. Throw two dices, one after another. Assume we are interested in the following two variables:

$$X: \text{result of the first dice}$$

and

$$Y: \text{square of the sum of the two results.}$$

The question we are interested in is the following: after observing the first result, the result of X, what is the best prediction of Y?

Let X and X' the results of the two dices. Then, $Y = (X + X')^2$.

Compute first of all the optimal predictor. We have

$$\begin{aligned} E[Y \mid X = x] &= E[(X + X')^2 \mid X = x] \\ &= E[(x + X')^2 \mid X = x] \\ &= x^2 + E[(X')^2 \mid X = x] + 2xE[X' \mid X = x] \\ &= x^2 + \mathbb{E}[(X')^2] + 2x\mathbb{E}[X']. \end{aligned}$$

Recall we have $\mathbb{E}(X') = 3.5$ and $\mathbb{E}((X')^2) = 15.167$. Therefore,

$$E[Y \mid X = x] = x^2 + 7x + 15.167$$

and, as a random variable, we have

$$E[Y \mid X] = X^2 + 7X + 15.167.$$

Compute now the optimal linear predictor. We have

$$P(Y \mid X) = b_0 + b_1 X$$

and we search for $\widehat{b}_0$ and $\widehat{b}_1$ such that

$$\mathbb{E}[(Y - \widehat{b}_0 - \widehat{b}_1 X)^2]$$

is minimal.

The traditional method to solve this problem consist in consider the function φ defined on $\mathbb{R}^2$ such that

$$\begin{aligned}\varphi(b_0,b_1) &= \mathbb{E}[(Y-b_0-b_1X)^2]\\ &= \mathbb{E}(Y-b_0)^2+b_1^2\mathbb{E}(X^2)-2b_1\mathbb{E}[(Y-b_0)X]\\ &= \mathbb{E}(Y^2)+b_0^2-2b_0\mathbb{E}(Y)+b_1^2\mathbb{E}(X^2)-2b_1\mathbb{E}(YX)+2b_0b_1\mathbb{E}(X),\end{aligned}$$

to compute its partial derivatives

$$\begin{aligned}(\partial_{b_0}\varphi)(b_0,b_1) &= 2b_0-2\mathbb{E}(Y)+2b_1\mathbb{E}(X)\\ (\partial_{b_1}\varphi)(b_0,b_1) &= 2b_1\mathbb{E}(X^2)-2\mathbb{E}(XY)+2b_0\mathbb{E}(X)\end{aligned}$$

and equate them to 0 in order to find the point $(\widehat{b}_0,\widehat{b}_1)\in\mathbb{R}^2$ that minimizes φ. Compute the matrix of second derivatives and check it is positive definite, that is, its determinant is positive. Concretely, it is equal to $4\mathbb{V}(X)$.

We have

$$\begin{aligned}2b_0+2\mathbb{E}(X)b_1 &= 2\mathbb{E}(Y)\\ 2\mathbb{E}(X)b_0+2\mathbb{E}(X^2)b_1 &= 2\mathbb{E}(XY),\end{aligned}$$

that is,

$$\begin{bmatrix}1 & \mathbb{E}X\\ \mathbb{E}X & \mathbb{E}(X^2)\end{bmatrix}\begin{bmatrix}b_0\\ b_1\end{bmatrix}=\begin{bmatrix}\mathbb{E}Y\\ \mathbb{E}(XY)\end{bmatrix}.$$

Writing $\mu_X=\mathbb{E}X$, $\mu_Y=\mathbb{E}Y$, $\sigma_X^2+\mu_X^2=\mathbb{E}(X^2)$ and $\sigma_{XY}+\mu_X\mu_Y=\mathbb{E}(XY)$ we have

$$\begin{bmatrix}\widehat{b}_0\\ \widehat{b}_1\end{bmatrix}=\begin{bmatrix}1 & \mu_X\\ \mu_X & \mu_X^2+\sigma_X^2\end{bmatrix}^{-1}\begin{bmatrix}\mu_Y\\ \sigma_{XY}+\mu_X\mu_Y\end{bmatrix}.$$

The inverse matrix is

$$\frac{1}{\sigma_X^2}\begin{bmatrix}\mu_X^2+\sigma_X^2 & -\mu_X\\ -\mu_X & 1\end{bmatrix}.$$

Therefore,

$$\widehat{b}_0=(1+\frac{\mu_X^2}{\sigma_X^2})\mu_Y-\frac{\mu_X}{\sigma_X^2}(\sigma_{XY}+\mu_X\mu_Y)=\mu_Y-\frac{\sigma_{XY}}{\sigma_X^2}\mu_X$$

and

$$\widehat{b}_1=-\frac{\mu_X\mu_Y}{\sigma_X^2}+\frac{\sigma_{XY}}{\sigma_X^2}+\frac{\mu_X\mu_Y}{\sigma_X^2}=\frac{\sigma_{XY}}{\sigma_X^2}.$$

Note that

$$\widehat{b}_0=\mu_Y-\widehat{b}_1\mu_X.$$

In our case we have $\mu_X=3.5$, $\sigma_X^2=15.167-3.5^2=2.917$,

$$\begin{aligned}\mathbb{E}(Y) = \mathbb{E}[(X+X')^2] &= \mathbb{E}(X)^2 + \mathbb{E}((X')^2) + 2\mathbb{E}(X)\mathbb{E}(X') \\ &= 2(\mu_X^2 + \sigma_X^2) + 2\mu_X^2 \\ &= 4\mu_X^2 + 2\sigma_X^2 \\ &= 4 \cdot 3.5^2 + 2 \cdot 2.917 \\ &= 54.834\end{aligned}$$

and

$$\begin{aligned}\sigma_{XY} &= \mathbb{E}(XY) - \mu_X\mu_Y \\ &= \mathbb{E}[X(X+X')^2] - \mu_X\mu_Y \\ &= \mathbb{E}(X^3) + \mathbb{E}[X(X')^2] + 2\mathbb{E}[X^2X'] - \mu_X\mu_Y \\ &= \mathbb{E}(X^3) + \mu_X\mathbb{E}(X^2) + 2\mu_X\mathbb{E}(X^2) - \mu_X\mu_Y \\ &= \mathbb{E}(X^3) + 3\mu_X(\sigma_X^2 + \mu_X^2) - \mu_X\mu_Y \\ &= 73.5 + 3 \cdot 3.5 \cdot (2.917 + 3.5^2) - 3.5 \cdot 54.834 \\ &= 40.835.\end{aligned}$$

Finally

$$\begin{aligned}\widehat{b}_1 &= 13.999 \\ \widehat{b}_0 &= 5.837\end{aligned}$$

and therefore

$$P(Y \mid X) = 13.999X + 5.837.$$

Prediction errors are

$$\begin{aligned}\mathbb{V}(Y - E(Y \mid X)) &= \mathbb{E}[(Y - X^2 - 7X - 15.167)^2] = 611.908 \\ \mathbb{V}(Y - P(Y \mid X)) &= \mathbb{E}[(Y - 13.999X - 5.837)^2] = 1152.652\end{aligned}$$

Note that these errors can be written as polynomials of the four first moments of X and then, the computation is done using $\mathbb{E}(X) = 3.5$, $\mathbb{E}(X^2) = 15.167$, $\mathbb{E}(X^3) = 73.5$ and $\mathbb{E}(X^4) = 379.167$.

Take for example $X = 3$. The variable Y, conditioned to $X = 3$, takes values in the set $\{16, 25, 36, 49, 64, 81\}$ with equal probability. Therefore, its expected value is 45.167. The optimal prediction is exactly the same and the linear best prediction is 47.834. Take now $X = 1$. We have that Y takes values in $\{4, 9, 16, 25, 36, 64\}$, its expectation is 23.667, equal of course to the optimal prediction. Its linear best prediction is 19.836. It would be interesting to compute all cases and to represent graphycally both predicting functions.

3 The optimal linear predictor

Note that, in general, we have the following expression for the optimal linear predictor:

$$P(Y \mid X) = \mu_Y + \frac{\sigma_{XY}}{\sigma_X^2}(X - \mu_X).$$

Note that we have seen that to compute $\widehat{b}_0$ and $\widehat{b}_1$ consist in searching Z such that $Z = b_0 + b_1 X$ and $E[(Y - Z)^2]$ is minimal. It is known that the solution of this problem is the orthogonal projection of Y over the linear subspace of $L^2(\Omega)$ generated by 1 and X. So, Y has to be orthogonal to 1 and X. Therefore, it satisfies

$$\begin{aligned} \mathbb{E}[(Y - Z) \cdot 1] &= 0 \\ \mathbb{E}[(Y - Z) \cdot X] &= 0, \end{aligned}$$

and so,

$$\begin{aligned} \mathbb{E}[(Y - b_0 - b_1 X)] &= 0 \\ \mathbb{E}[(Y - b_0 - b_1 X)X)] &= 0 \end{aligned}$$

and finally

$$\begin{aligned} b_0 + \mathbb{E}(X) b_1 &= \mathbb{E}(Y) \\ \mathbb{E}(X) b_0 + \mathbb{E}(X^2) b_1 &= \mathbb{E}(XY). \end{aligned}$$

Note that this is exactly the same system of equations obtained in the previous section.

4 Exponential smoothing and Holt-Winters forecasting

A particular case of filtering that is used to predict is the so-called exponential smoothing, a method related with the Holt-Winters prediction algorithm.

Imagine we have a times series $x_1, x_2, \ldots, x_n$ with no systematic trend or seasonal effects and we want to predict x_{n+h} with $h \geq 1$. For example, the series could be a series of sales of a well-established product in a stable market.

The method of exponential smoothing proposes to estimate a trend process m recursively, defining $m_1 := x_1$ and

$$m_n = a x_n + (1 - a) m_{n-1},\ n \geq 2,$$

with $a \in (0, 1)$.

Note that this implies on one hand,

$$m_n = m_{n-1} + a(x_n - m_{n-1}),\ n \geq 2,$$

and on other hand,

$$m_n := \sum_{j=0}^{n-2} a(1 - a)^j x_{n-j} + (1 - a)^{n-1} x_1.$$

Observe that the exponential smoothing is a weighted average of the previous data, with decreasing weights in the past.

The parameter a can be chosen or estimated. Note that if a is close to 1, the trend is almost equal to the series. On the contrary, if a is close to 0 the estimated trend m changes very slowly from x_1. The closer is a to 0 the smoother is the estimated trend m. A typical value is $a = 0.2$.

By default, the best estimation of x_n is m_{n-1}. This fact suggests that parameter a can be estimated by minimizing the error

$$E_n(a) := \sum_{j=2}^{n}(x_j - P_{j-1}x_j)^2 = \sum_{j=2}^{n}(x_j - m_{j-1})^2.$$

In general, $P_n x_{n+h} = m_n$ for any $h \geq 1$.

Holt, in 1957, generalized the method to include times series with trends. The formulas are

$$m_n = ax_n + (1-a)(m_{n-1} + t_{n-1})$$

and

$$t_n = b(m_n - m_{n-1}) + (1-b)t_{n-1}$$

for $n \geq 2$ and with $m_1 = x_1$ and $t_1 = 0$.

In this case, the prediction is given by $P_n x_{n+h} = m_n + t_n$ for any $h \geq 1$.

The error, as before, is given by

$$E_n(a,b) = \sum_{j=2}^{n}(x_j - P_{j-1}x_j)^2 = \sum_{j=2}^{n}(x_j - m_{j-1} - t_{j-1}))^2.$$

Finally, Winters, in 1960, made an extension to the case the series has seasonal effects.

In this case, the equations are given by

$$m_n = a(x_n - s_{n-p}) + (1-a)(m_{n-1} + t_{n-1}),$$

$$t_n = b(m_n - m_{n-1}) + (1-b)t_{n-1}$$

and

$$s_n = c(x_n - m_n) + (1-c)s_{n-p}$$

for $n \geq 2$ and with $m_1 = x_1$ and $t_1 = s_1 = \cdots = s_p = 0$.

In this case, the prediction is given by

$$P_n x_{n+h} = m_n + ht_n + s_{n+h-p}$$

and the error by

$$E_n(a,b,c) = \sum_{j=2}^{n}(x_j - P_{j-1}x_j)^2 = \sum_{j=2}^{n}(x_j - m_{j-1} - t_{j-1} - s_{j-p})^2.$$

References

[1] P. J. Brockwell and R. A. Davis (1991): *Time Series: Theory and Methods*. Springer.

[2] P. J. Brockwell and R. A. Davis (1991): *Introduction to Time Series and Forecasting*. Springer.

[3] P. S. P. Cowpertwait and A. V. Metcalfe (2009): Introductory Time Series with R. Springer.