Time Series

Viktor

December 22, 2020

Introduction. Time series are commonly thought as a superposition of several components

- *Trend*: long-term direction of time series
- *Seasonal*: pattern that repeats with known and fixed periodicity
- *Cycle*: pattern that repeats but with unknown und changing periodicity.
- *Autocorrelation*:
- *Error*: unpredictable component of time series.

Many time-series models can be classified as additive, multiplicative, or a combination of these components [1]. The aim to time series modelling is to capture all effects until only error (=noise) is left. More precisely if there ist still come uncaptured trend, seasonal, cylce or autocorrelation then we can be pretty sure that we can improve the model. However, this does not imply that all models that just exhibit unexplained noise perform equally.

Cross validation. For time series a split in test and train data is performed along the temporal axis. Observations in the past –on which the model is trained – are called train set and observations in the futures –for which the accuracy is determined – are called test set.

A more sophisticated split of data is called cross validation. In this approach there is a series of train/test sets. Each of which contains only one observation y_l in the test set and the training set are the observations *prior* to this observation with an offset ("step") s $y_1, \ldots, y_{l-s}$. For fixed s, l induces a series of test/train sets.The forecast accuracy is then measured by averaging over the test sets. This is also called *evaluation on a rolling foracast origin.* $s = 1$ is called 1-step-forecast. In practice the accuracy as a function of s is investigated.

Model diagnostics. Assume that there is model that makes predictions $\hat{y}_t$ at time t and the corresponding observed values y_t are known. Then there are several methods that help diagnosing the model: residuals, autocorrlation function, and portmonaie tests. **Residuals** of a time series is the difference between predicted values $\hat{y}_t$ and observed values y_t at time t,

$$e_t = y_t - \hat{y}_t. \tag{1}$$

For a "good" models residuals have *zero mean* are *uncorrelated*. If the mean is not zero then the model is biased and if there is correlation then there is still some information left in the residuals that should go in the model [4]. These criteria do not assist in choosing a model as they may hold for several models simultaneously (model selection is performed via the error on the test set). In order to ease the computation of prediction intervals (for certain models) it is also beneficial that residuals are normally distributed with constant variance. Fixing the correlation and the distributional shape of the residual is challenging whereas zero mean can be achieved by adding a constant (the mean) to the model. A Box-Cox transformation may achieve a normal constant variance distribution.

The **autocorrelation function (ACF).** is the Pearson correlation of a time dependent random variable Y_t with itself shifted in time by τ. The ACF is typically computed on the raw data (to observed lag relationships) and residuals (for model diagnostic).

$$\rho(\tau) = \frac{\sum_{t=\tau+1}^{T} (y_t - \bar{y})\,(y_{t-\tau} - \bar{y})}{\sum_{t=1}^{T} (y_t - \bar{y})^2}, \tag{2}$$

where $\bar{y}$ is the (empirical) mean. The autocorrelation function depends on τ. For white noise 95 % of its (τ-dependent) values are within the interval $\pm 2/\sqrt{T}$. Thus the and is used to identify white noise [4].

Portmonaie tests for ACF perform hypothesis tests for the ACF as a point wise analysis of $\rho(\tau)$ for each τ could yield false positives (claiming erroneous autocorrelation). The **Box-Pierce test** is based on the statistics

$$Q = T \sum_{\tau=1}^{\ell} \rho_\tau^2 \tag{3}$$

for fixed ℓ. In practice, the choice $\ell = 10$ for non-seasonal data and $\ell = 2p$ where p is the period of seasonality is common. For large ℓ (compared to T) the test deteriorates, which suggests the rule of thumb $\ell = \min(10, T/5)$ ($\ell = \min(2p, T/5)$) for non-seasonal (seasonal) data [4]. The more accurate

Ljung-Box test is based on the statistics

$$Q = T(T+2)\sum_{\tau=1}^{\ell}\frac{\rho_\tau^2}{T-\tau} \tag{4}$$

For both tests, large values of Q indicate that the ACF does *not* come from white noise. More precisely, if the autocorrelations are from white noise then Q would have a χ^2 distribution with $\ell - K$ degrees of freedom, where K is the number of parameters in the model (so, if Q is is calculated on the raw data rather than the residuals $K = 0$).

Loss functions and errors

Definition 1 (Loss functions)**.** *Let y_t be the observed time series and $\hat{y}_t$ the corresponding predictions at timestep t of T timesteps.*

- *Absolute loss*
- *Squared loss*
- *Huber loss*

$$\begin{cases} \frac{1}{2}(y_t - \hat{y}_t)^2 & if |y_t - \hat{y}_t| \leq \delta, \\ \delta\,|y_t - \hat{y}_t| - \frac{1}{2}\delta^2 & otherwise. \end{cases} \tag{5}$$

- *ρ-risk metric (quantile loss) [5]*

Definition 2 (Forecast errors)**.** *Let y_t be the observed time series and $\hat{y}_t$ the corresponding predictions at timestep t of T timesteps.*

- *Mean absolute error (MAE):* $\frac{1}{T}\sum_{t=1}^{T}|y_t - \hat{y}_t|$
- *Mean squared error:* $\frac{1}{T}\sum_{t=1}^{T}(y_t - \hat{y}_t)^2$.
- *Root mean squared error (RMSE):* $\sqrt{\frac{1}{T}\sum_{t=1}^{T}(y_t - \hat{y}_t)^2}$

A forecast method that minimizes the MAE (RMSE) will lead to forecasts of the median (expectation value) of the underlying forecast density. Thus, the RMSE is widely used although the MAE is easier to interpret. Both, the MAE and RMSE dependent on the physical units.

In order to compare algorithms among various problems unit independent measures are needed. However this is a delicate task and quantities such as the the (symmetrized) mean abolute percentage error may diverge. Rather, errors scaled by a simple forecast method are more appropriate [4, 2].

Transformations.

- **Calender adjustments.** Sometimes pattern occur that are due to different number of days in a month or if there is some additional known contraint in the data (data points are not generated on sundays). It is much easier to remove this structure by changing to an appropriate time unit than trying to let the model learn this.

- **Correction for known temporal adjustments.** If data is affected by e.g. the population or inflation adjust it in order to account for these effects on the training data level instead of trying to learn these effects. E.g. consider quantities per capita (population) or use inflation corrected data (financial times series).

- **Fourier transformation.** Include the dominant Fourier contributions in order to help the model dealing with periodicity. In practice these components my be found by an FFT and sometimes only the corresponding basis functions are included.

- **Box-Cox transformation** on target values [4],

$$y_t \mapsto \begin{cases} \ln y_t & \text{if } \lambda = 0 \\ (y_t^\lambda - 1)/\lambda & \text{otherwise.} \end{cases} \tag{6}$$

  This transformation aims to render the seasonal effects to be about the same across the whole time series, which dictates the choice of λ. In practice forecasting results (mean point estimates) are relatively insensitive to λ in contrast prediction intervals Once a prediction has been made on the transformed space it needs to be back-transformed into the original scale. The inverse transformation

$$y_t \mapsto \begin{cases} e^{y_t} & \text{if } \lambda = 0 \\ (\lambda y_t + 1)^{\frac{1}{\lambda}} & \text{otherwise,} \end{cases} \tag{7}$$

  is sometimes used. However, this point wise inverse of the Box-Cox transformation is *not* the inverse of the underlying forecast distribution [1]. If only the mean as point estimate is of interest then the *biased adjusted* back-transformed mean can be approximated by [3],

$$y_t \mapsto \begin{cases} e^{y_t}\left[1 + \frac{\sigma^2}{2}\right] & \text{if } \lambda = 0 \\ (\lambda y_t + 1)^{\frac{1}{\lambda}}\left[1 + \frac{\sigma^2(1-\lambda)}{2(\lambda y_t + 1)^{\frac{1}{2}}}\right] & \text{otherwise,} \end{cases} \tag{8}$$

[1] More precisely, the back transformation (8) is only correct for the median if the transformed forecast distribution is symmetric [4]

where by assumption y_t is the mean on the transformed scale and σ the variance.

Prediction intervals Prediction intervals are given by the forecast distribution, in particular its standard deviation. However these are for most models not available and one uses the residual distribution instead. In fact, if the model is parameter-free, then the standard deviations of the model forecast distribution and the residual distribution are identical. If the model has parameters they would increase the standard deviation (due to uncertainty on the model parameters), which is often ignored.

Further assuming that the residuals are **normally distributed** with constant variance the 95 % one step prediction interval is a approximated by $\hat{y}_{T+1|T} \pm 1.95\,\hat{\sigma}$. Here 1.95 is the 95 %-width of the standard normal distribution.

This can be generalized to multi step prediction interval $h > 1$,

$$\hat{y}_{T+h|T} \pm c\hat{\sigma}_h, \tag{9}$$

where $\hat{\sigma}_h$ is the (model dependent) standard deviation at step h and c is the width of the standard normal distribution for a given prediction interval (e.g., $c = 1.95$ for the 95 % prediction interval).

If no analytical standard deviations are available then prediction intervals may be obtained from **bootstrapped residuals**.

Baseline Methods [4]. Let $\{y_1, \ldots, y_T\}$ be historic data. The aim is to predict a future value $y_{T+h|T}$ based on the historic data. For these models also estimates for the standard deviations $\hat{\sigma}_h$ are available, which can be used to estimate prediction intervals according to eq. (9).

- **Average method:** forecasts future values as mean of historical data,

$$\hat{y}_{T+h|T} = \frac{1}{T}\sum_{t=1}^{T} y_t \tag{10}$$

$$\hat{\sigma}_h = \hat{\sigma}\sqrt{1 + \frac{1}{T}}. \tag{11}$$

- **Naïve method:** forecasts future value as last observed value,

$$\hat{y}_{T+h|T} = y_T \tag{12}$$

$$\hat{\sigma}_h = \hat{\sigma}\sqrt{h}. \tag{13}$$

- **Seasonal naïve method:** forecasts last observed value of the same season,

$$\hat{y}_{T+h|T} = y_{T+h-np}, \tag{14}$$
$$\hat{\sigma}_h = \hat{\sigma}\sqrt{n}, \tag{15}$$

  where p is the seasonal period and $n = \left\lfloor \frac{h-1}{p} \right\rfloor + 1$.

- **Drift method:** adds to naïve method an overall gradient,

$$\hat{y}_{T+h|T} = y_T + \frac{h}{T-1}\sum_{t=2}^{T}(y_t - y_{t-1}) \tag{16}$$
$$= y_T + \frac{h}{T-1}(y_T - y_1), \tag{17}$$
$$\hat{\sigma}_h = \hat{\sigma}\sqrt{h\left(1 + \frac{h}{T}\right)} \tag{18}$$

In the estimates for the $h-$step standard deviations the quantity $\hat{\sigma}$ is the empirical standard deviation of the residuals. The average and naïve methods give constant predictions.

Neural networks. There are several neural network (NN) based approaches, RNN, LSTM, 1D CNN, deep NN. The basic idea behind deep NN is to translate the problem into a supervised learning problem by predicting y_t with features that are the last l target values $y_{t-1}, y_{t-2}, \ldots y_{t-l}$

Overview of methods

- Regression models
- Exponential smoothing methods
- Arima
- Dynamic Regression
- Hierarchical forecasting
- Vector autoregression
- Neural networks
- survival analysis

Comments

- Temporal cutoffs (e.g. returns are allowed only for 30 days)
- How many data points for the prediction of returns (data from yesterday are not relevant for returns for today). ← evaluation on rolling forecast origin (?)
- What shall we do with new products? Consider cohorts maybe.

References

[1] David J. Hand. Forecasting with exponential smoothing: The state space approach by rob j. hyndman, anne b. koehler, j. keith ord, ralph d. snyder. *International Statistical Review*, 77(2):315–316, 2009.

[2] Rob J Hyndman and Anne B Koehler. Another look at measures of forecast accuracy. *International Journal of Forecasting*, pages 679–688, 2006.

[3] Robin John Hyndman. The forecast mean after back-transformation, 2014.

[4] Robin John Hyndman and George Athanasopoulos. *Forecasting: Principles and Practice*. OTexts, Australia, 2nd edition, 2018.

[5] Matthias W Seeger, David Salinas, and Valentin Flunkert. Bayesian intermittent demand forecasting for large inventories. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 4646–4654. Curran Associates, Inc., 2016.