Machine Learning

Chapter 4: Forecasting III

April 2021

Contents

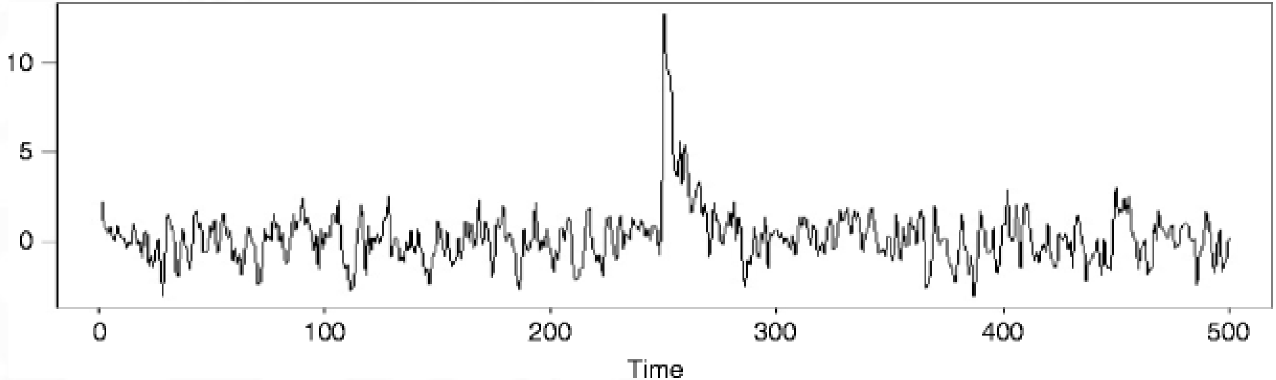
1

Intervention analysis

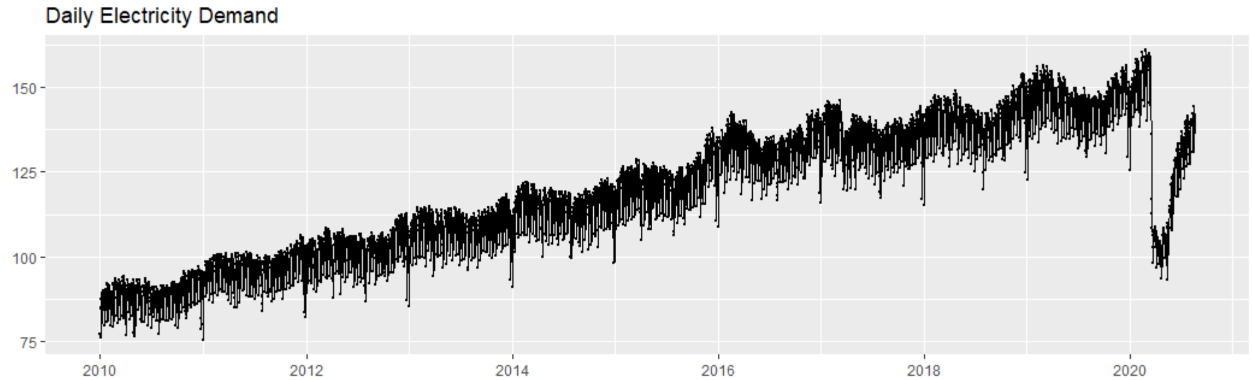
Intervention analysis (1)

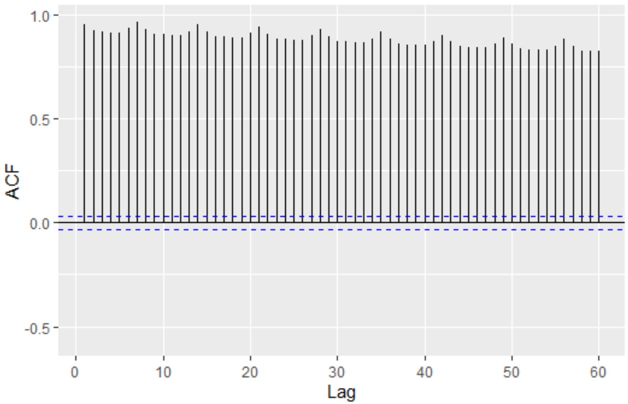
- Time series are very often affected by external events that happen at very specific dates (strikes, accidents, sales promotions, change in legislation, …).

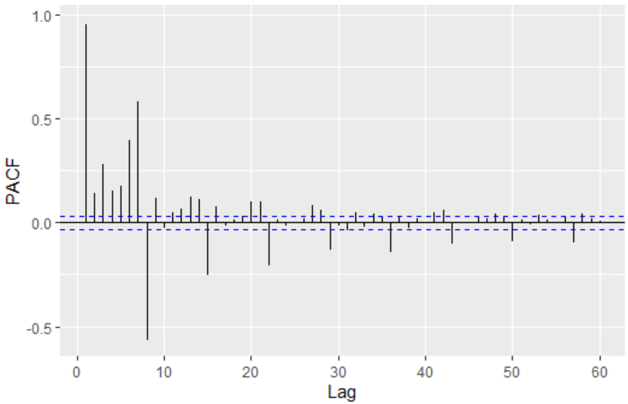
- If we model the effect of these events, we will improve the accuracy of the parameters and therefore the forecasts.



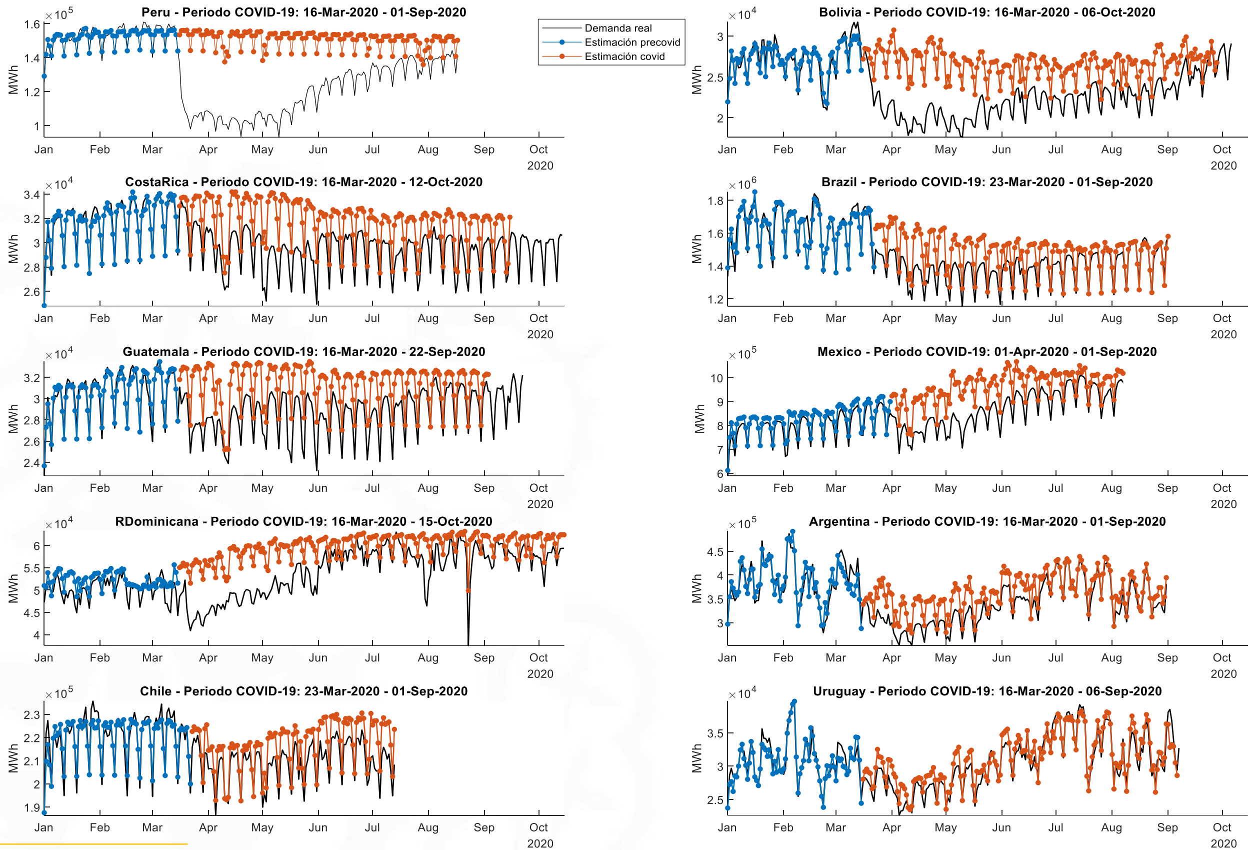
Intervention analysis (2)



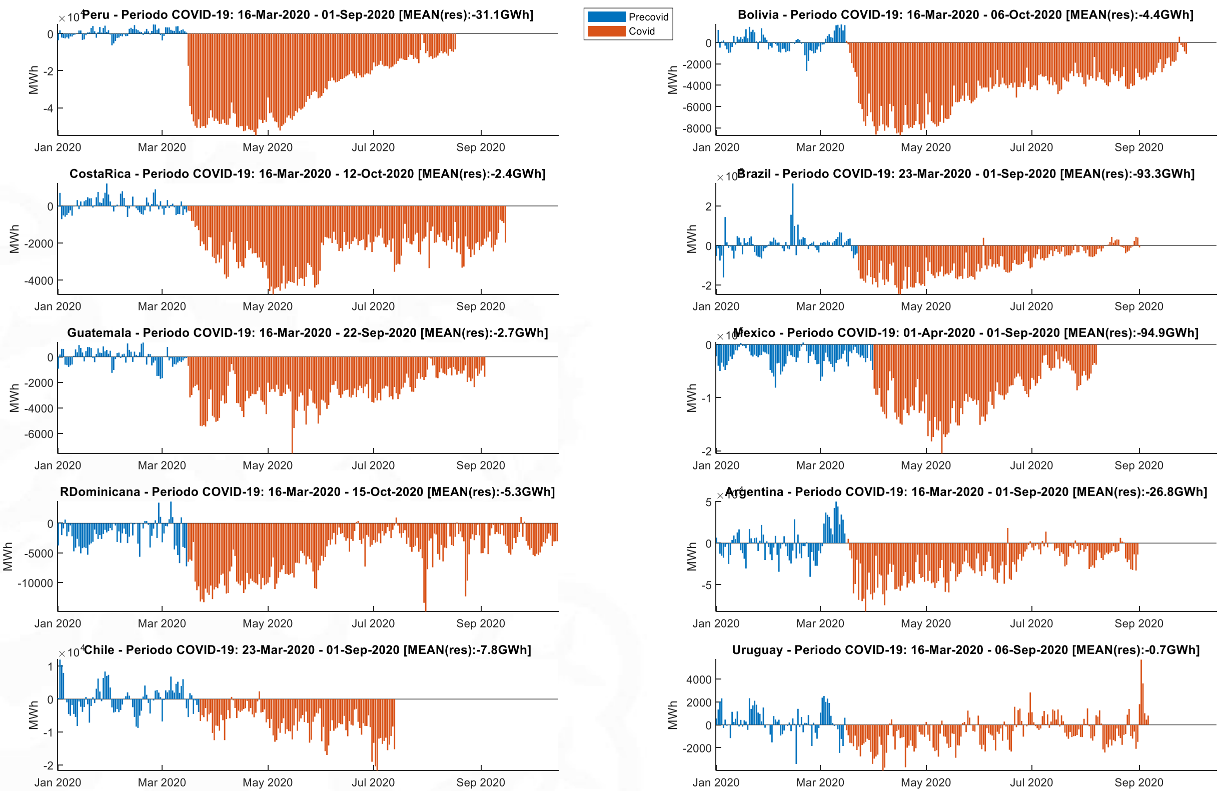




Intervention analysis (3)



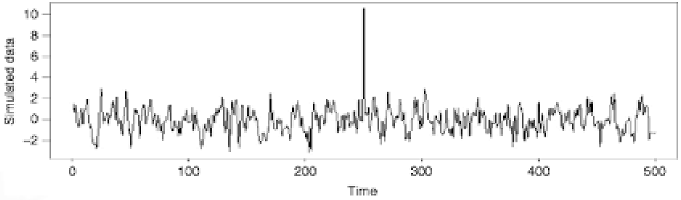
Intervention analysis (4)



Intervention analysis (5)

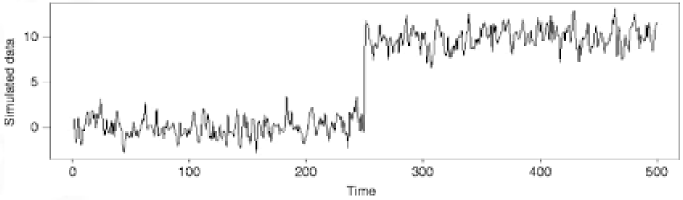
- For example, to model the effect of a strike in a daily production time series, we can build an impulse variable *I[t]* that takes the value 1 the day of the strike and 0 otherwise, and use the model:

$$y[t] = \mu + w \cdot I[t] + v[t]$$



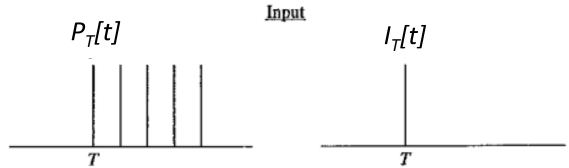
- To model the effect of a change in legislation we can build a step variable *S[t]* that takes the value 0 before the change and the value 1 after it, and use the model:

$$y[t] = \mu + w \cdot S[t] + v[t]$$



Intervention analysis (6)

- The dummy variables which are most frequently used for modeling the deterministic effect of an event on a time series are the impulse and the step functions.

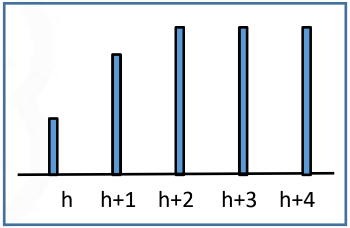


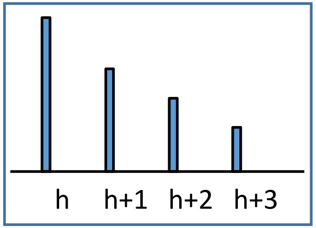
- Impulse variables are used to model the effect of events that occur at a given time, such as a strike or an accident.

- Step variables represent events that start at a given time and that remain from that moment, as a change in regulation or a change in the basis of an index.

- The response may require a Dynamic regression model:

$$\boxed{y[t] = \omega(B) I_h[t] + \psi(B)\varepsilon[t]}$$





2

Nonlinear time series models

Nonlinear Time Series Models
The nonliear regression approach

- General model:

$$y[k] = f(y^{\{k-1\}}, \boldsymbol{x}^{\{k\}}, \boldsymbol{\varepsilon}^{\{k-1\}}) + \boldsymbol{\varepsilon}[k]$$

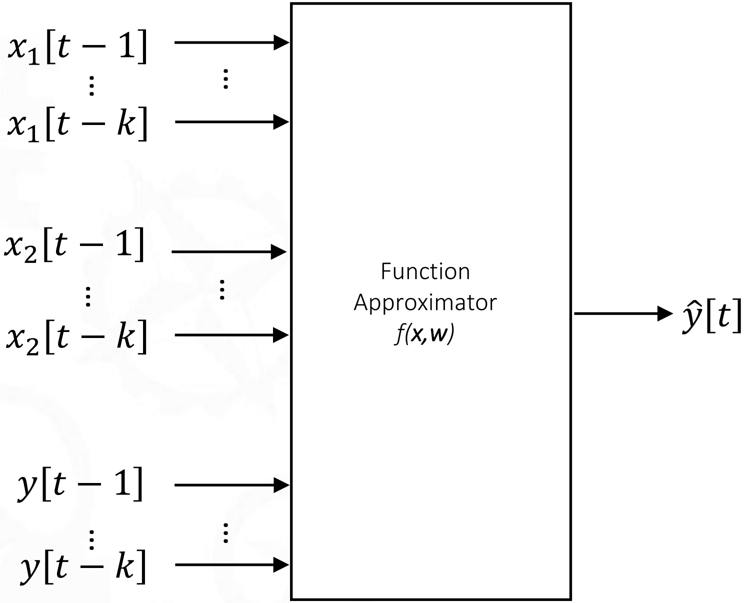
where:

- f: nonlinear function
- $\boldsymbol{y}[k] \in \Re^m$: outputs at time k
- $\boldsymbol{x}[k] \in \Re^n$: inputs at time k
- $\boldsymbol{\varepsilon}[k] \in \Re^m$: white noise process
- $\boldsymbol{v}^{\{k-1\}} = [\boldsymbol{v}[k-1], \boldsymbol{v}[k-2], \ldots]^T$

comillas.edu

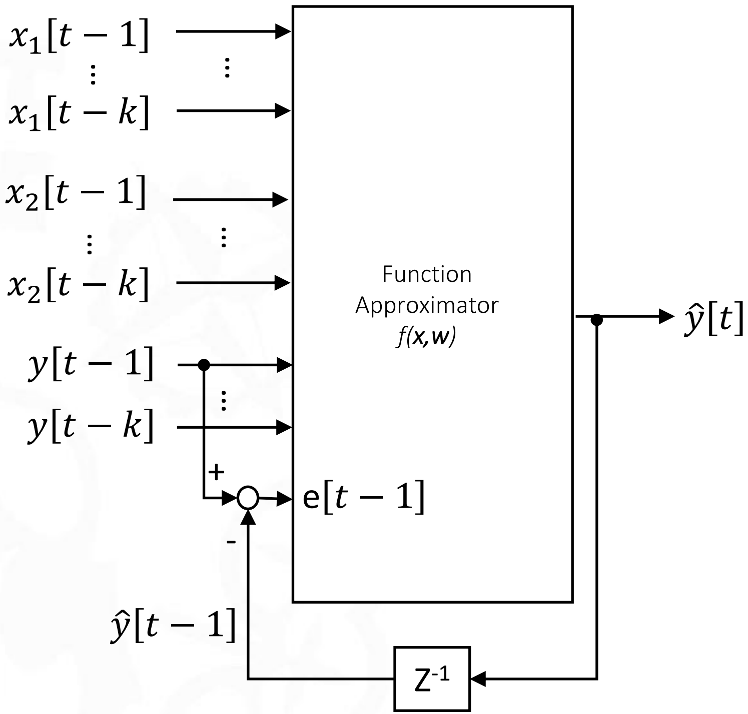
Nonlinear Time Series Models
The nonliear regression approach

- NARX

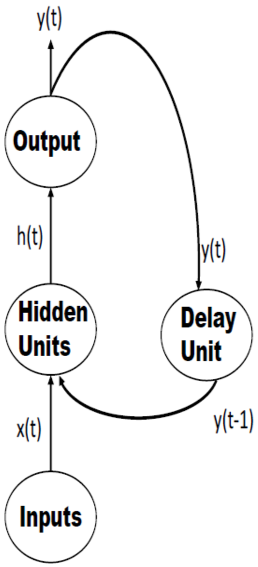


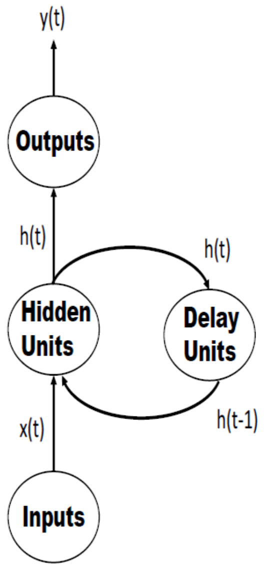
Nonlinear Time Series Models
The nonliear regression approach

- NARMAX

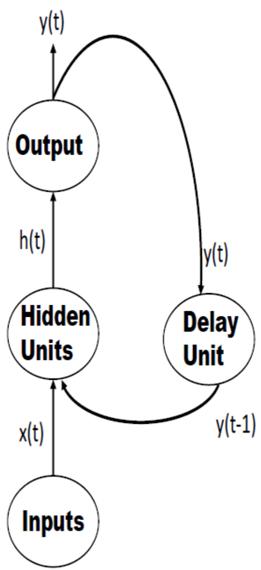


Recurrent Neural Networks
Jordan & Elman Neural Networks

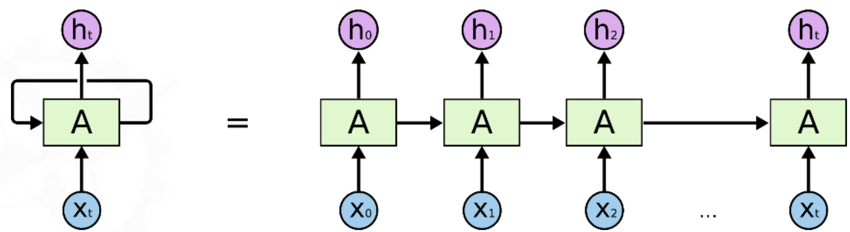




Recurrent Neural Networks
Jordan & Elman Neural Networks



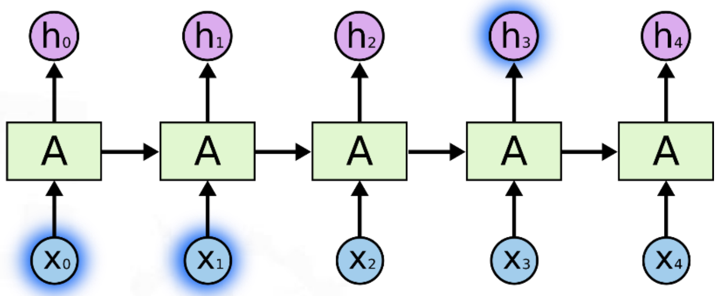
=

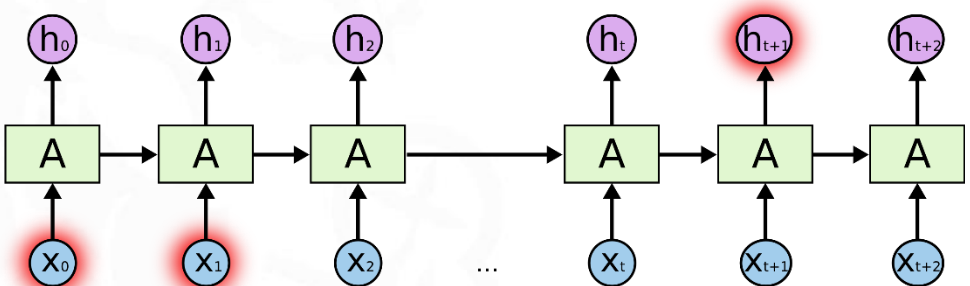


An unrolled recurrent neural network

See: https://colah.github.io/posts/2015-08-Understanding-LSTMs/

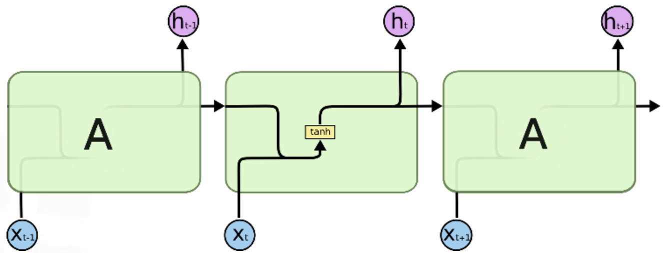
Recurrent Neural Networks
LSTM Neural Networks



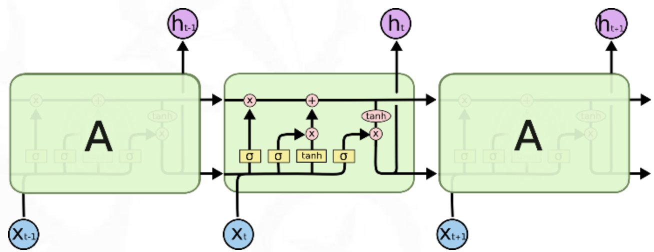


See: https://colah.github.io/posts/2015-08-Understanding-LSTMs/

Recurrent Neural Networks
LSTM Neural Networks



The repeating module in a standard RNN contains a single layer



The repeating module in an LSTM contains four interacting layers

See: https://colah.github.io/posts/2015-08-Understanding-LSTMs/

comillas.edu

3

Combination of forecasts

Introduction

- Any time series can be modeled with different methods, so that forecasts could benefit from the advantages of each of them to get a better prediction in terms of error than any of the individual predictors.

- Bates & Granger (1969) analysed the linear combination of two predictors:

$$\hat{y}^c[t+h\,|\,t] = k_1\hat{y}_1[t+h\,|\,t] + k_2\hat{y}_2[t+h\,|\,t]$$

- If both are unbiased predictors, we take $k_2=1-k_1$ for an unbiased combination.

Optimal combination of two predictors (1)

- The error of the combination of forecasts is given by:

$$e^c[t+h|t] = y[t+h] - \hat{y}^c[t+h|t]$$

whose variance is given by:

$$\begin{aligned} Var(e^c[t+h|t]) &= Var(y[t+h] - \hat{y}^c[t+h|t]) \\ &= Var(k \cdot e_1^c[t+h|t] + (1-k) \cdot e_2^c[t+h|t]) \\ &= k^2\sigma_1^2 + (1-k)^2\sigma_2^2 + 2k(1-k)\rho\sigma_1\sigma_2 \end{aligned}$$

where $\sigma_1{}^2$ and $\sigma_2{}^2$ are the error variances of the two predictors and ρ the correlation coeficient between the two errors.

Optimal combination of two predictors (2)

- If we minimize the variance of the combined error then we obtain as optimal weight :

$$k^{*} = \frac{\sigma_2^2 - \rho\sigma_1\sigma_2}{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2}$$

and the minimum variance:

$$\sigma_C^2 = Min\,Var(e^c[t+h\,|\,t]) = \frac{\sigma_1^2\sigma_2^2(1-\rho^2)}{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2}$$

which is less than or equal to the minimum variance of the two predictors.

$$\sigma_1^2 - \sigma_C^2 = \frac{\sigma_1^2(\sigma_1 - \rho\sigma_2)^2}{(\sigma_1 - \rho\sigma_2)^2 + \sigma_2^2(1-\rho^2)} \geq 0$$

Optimal combination of two predictors (3)

- Assuming $\sigma_1 < \sigma_2$:
  - If $\rho = \sigma_1 / \sigma_2$, then: $MinVar(e^c[t+h\,|\,t]) = \sigma_1^2$
  - If $\rho = 0$, then: $MinVar(e^c[t+h\,|\,t]) = \dfrac{\sigma_1^2 \sigma_2^2}{\sigma_1^2 + \sigma_2^2}$
  - If $\rho \to -1$, then: $MinVar(e^c[t+h\,|\,t]) \to 0$
  - If $\rho \to +1$, then: $MinVar(e^c[t+h\,|\,t]) \to 0 \quad \text{if} \quad \sigma_1^2 \neq \sigma_2^2$

- Clearly the optimal situation is obtained when the correlation is negative and high, but even with positive correlations improvements are obtained.

Optimal combination of two predictors (4)

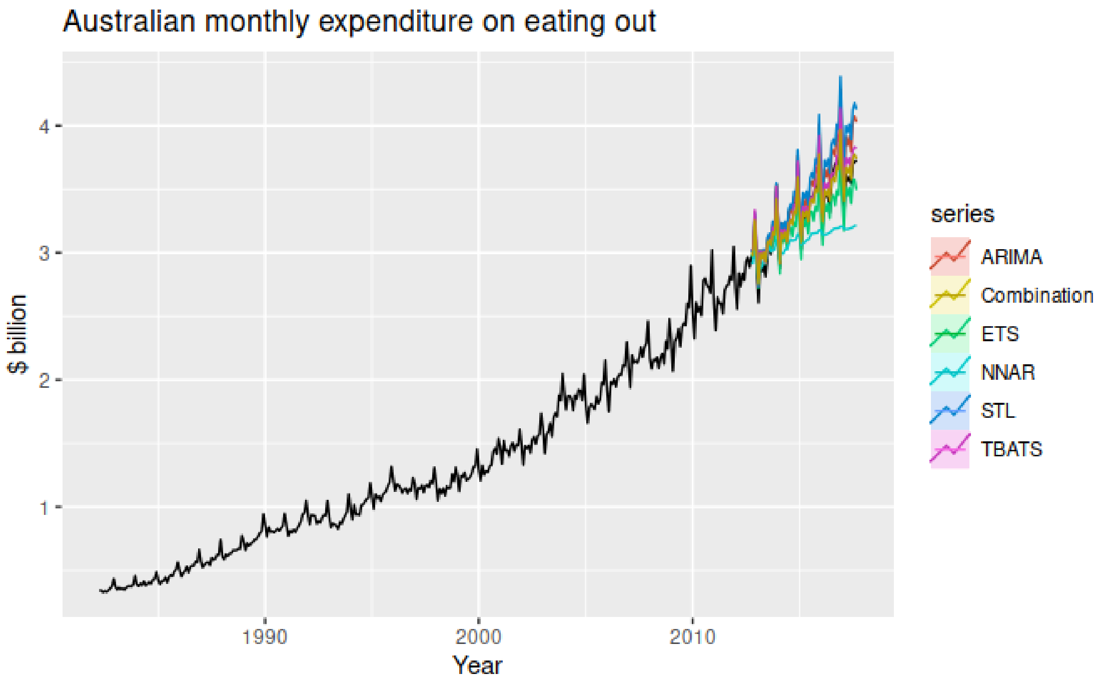
- Example:

  σ_1=20, σ_2=40, ρ=-0.6 then: $\sigma_C = 11.76$

  σ_1=20, σ_2=40, ρ=+0.6 then: $\sigma_C = 19.65$

- This method can be extended to *n* predictors.

Example



RMSE					
ETS	ARIMA	STL-ETS	NNAR	TBATS	COMBINATION
0.137	0.159	0.214	0.318	0.094	**0.072**

AFTER algorithm (Zou & Yang, 2004)

- The AFTER algorithm in an adaptive algorithm that adjusts the weights of the combination based on the evolution of the errors of each predictor.

- The weights of the linear combination: $\hat{y}^c[t] = \sum_{j=1}^{J} k_j \hat{y}_j[t]$ are given by:

$$k_j[n] = \frac{\prod_{i=1}^{n-1} \hat{\sigma}_j^{-1}[i] \exp\left(-\frac{1}{2} \sum_{i=1}^{n-1} \frac{\left(y[i] - \hat{y}_j[i]\right)^2}{\hat{\sigma}_j^2[i]} \right)}{\sum_{j'=1}^{J} \prod_{i=1}^{n-1} \hat{\sigma}_{j'}^{-1}[i] \exp\left(-\frac{1}{2} \sum_{i=1}^{n-1} \frac{\left(y[i] - \hat{y}_{j'}[i]\right)^2}{\hat{\sigma}_{j'}^2[i]} \right)}$$

where $\hat{\sigma}_j[n]$ is an estimation of the error conditional variance of the predictor *j* at time *n*.

AFTER algorithm (Yang, 2001)

- A recursive version of the AFTER algorithm is given by:

$$\hat{y}^c[t] = \sum_{j=1}^{J} k_j \hat{y}_j[t]$$

$$k_j[n] = \frac{k_j[n-1]\hat{\sigma}_j^{-1}[n-1]\exp\left(-\frac{\left(y[n-1]-\hat{y}_j[n-1]\right)^2}{2\hat{\sigma}_j^2[n-1]}\right)}{\sum_{j'=1}^{J} k_{j'}[n-1]\hat{\sigma}_{j'}^{-1}[n-1]\exp\left(-\frac{\left(y[n-1]-\hat{y}_{j'}[n-1]\right)^2}{2\hat{\sigma}_{j'}^2[n-1]}\right)}$$

- Its name comes from Aggregated Forecast Through Exponential Reweighting (AFTER)

4

Bibliography

Bibliography

- J. M. Bates & C.W.J. Granger (1969). *The combination of forecasts*. Operational Research Quarterly, 20(4), 451–468. https://doi.org/10.1057/jors.1969.103
- C. Bishop (2007). *Pattern Recognition and Machine Learning*. Springer.
- R. Clemen, R. (1989). *Combining forecasts: A review and annotated bibliography with discussion*. International Journal of Forecasting, 5, 559–608. https://doi.org/10.1016/0169-2070(89)90012-5
- P. H. Franses & R. Paap (2004). *Periodic Time Series Models*. Oxford University Press.
- K. Greff, R. K. Srivastava, J. Koutník, B. R. Steunebrink, & J. Schmidhuber (2017). *LSTM: A Search Space Odyssey.* IEEE Transactions on Neural Networks and Learning Systems, vol. 28, n.o 10, pp. 2222-2232.
- S. Haykin (2009). *Neural Networks and Learning Machines*. 3rd Ed. Pearson.
- A. Mateo, A. Muñoz, J. García-González (2005). *Modeling and forecasting electricity prices with input/output hidden Markov models*. IEEE Transactions on Power Systems. vol. 20, no. 1, pp. 13-24.
- A. Muñoz and T. Czernichow (1998). *Variable selection using feedforward and recurrent neural networks*. Engineering Intelligent Systems for Electrical Engineering and Communications. vol. 6, no. 2, pp. 91-102.
- A. Pankratz (1991). *Forecasting with Dynamic Regression Models*. Wiley-Interscience.
- D. E. Rumelhart, G. E. Hinton, & R. J. Williams (1986). *Learning representations by back-propagating errors*. Nature, vol. 323, n.o 6088, pp. 533-536.
- H. Tong (1983). *Threshold Models in Nonlinear Time-Series Analysis*. Springer-Verlag, New York.
- H. Zou & Y. Yang (2004). *Combining time series models for forecasting*. International Journal of Forecasting, 20, 69–84.

Alberto Aguilera 23, E-28015 Madrid - Tel: +34 91 542 2800 - http://www.iit.comillas.edu