

Trace Forecast Likelihood for Time Series Models Estimation

Ivan Svetunkov

Presentation at Ghent University

28th April 2016



Lancaster University
Management School

Lancaster Centre for
Forecasting



LCF

Motivation

Why should we care about time series modelling and forecasting?

- Meet demand,
- Price optimisation,
- Decrease stock inventory,
- Optimisation of production plan,



How the forecasting is usually done

Extrapolative methods (Exponential smoothing, ARIMA),

Causal methods (linear and non-linear regressions),

Judgemental forecasting.



Simple Exponential Smoothing (SES)

The basic form of SES:

$$\hat{y}_{t+1} = \alpha y_t + (1 - \alpha) \hat{y}_t \quad (1)$$

Substituting $e_t = y_t - \hat{y}_t$, the error correction form is:

$$\hat{y}_{t+1} = \hat{y}_t + \alpha e_t \quad (2)$$

The smoothing parameter can take values: $\alpha \in (0, 2)$

Usually stricter conditions are imposed: $\alpha \in (0, 1)$



Graphically $\hat{y}_{t+1} = \hat{y}_t + \alpha e_t$ means that:

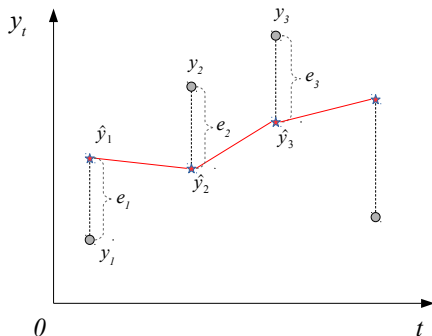


Figure: Simple Exponential Smoothing update mechanism

α regulates the updating mechanism in SES.



LCF

Example of series with a lower value of smoothing parameter:

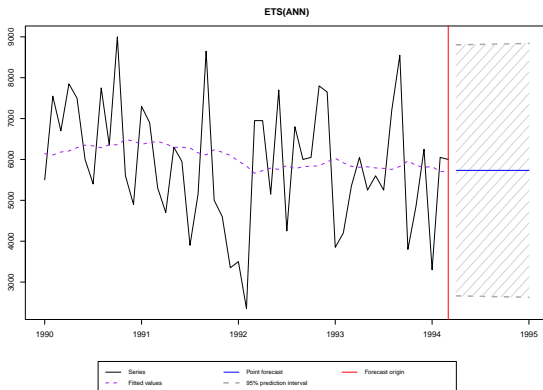


Figure: Forecast of a series with slow changes.



Example of series with a higher value of smoothing parameter:

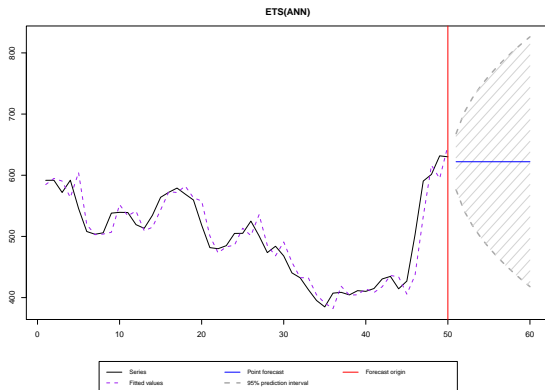


Figure: Forecast of a series with abrupt changes.



Exponential smoothing can also have the following components:

- trend;
- seasonality

Each of them can be either “none”, “additive” or “multiplicative”.



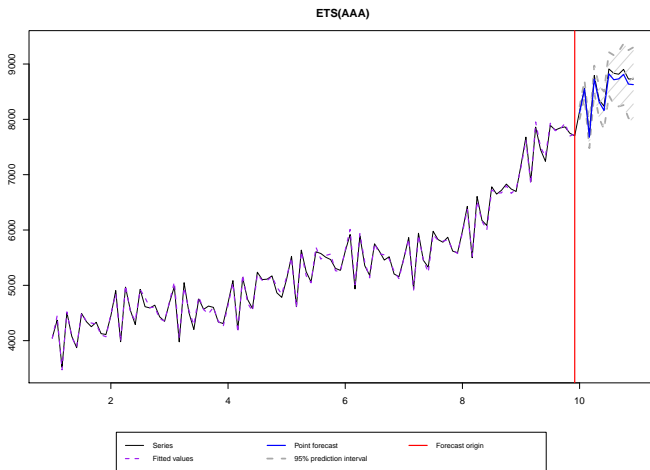


Figure: Time series with additive seasonality.



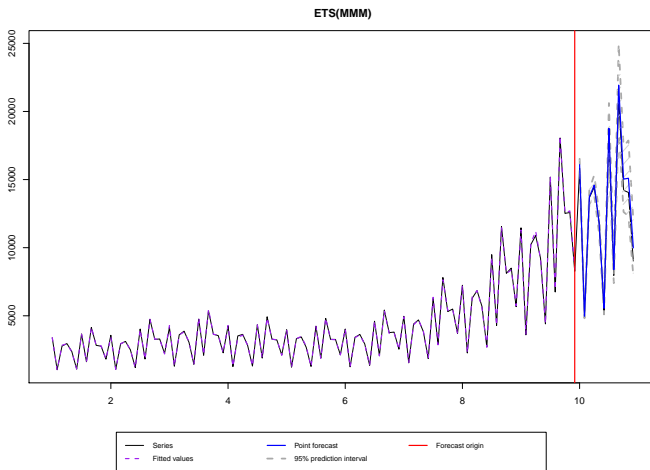


Figure: Time series with multiplicative seasonality.

The error term can be additive and multiplicative as well.

This results in 30 exponential smoothing models.

The same adaptation mechanism is used in all of them.

The short name of these models is ETS.



Estimation of ETS

Smoothing parameters in all the models are usually estimated in sample.

Predefining them is not a good practice.

The simplest way to optimise a model – using MSE:

$$MSE = \frac{1}{T} \sum_{t=1}^T e_{t+1|t}^2 \quad (3)$$

where $e_{t+1|t} = y_{t+1} - \hat{y}_{t+1}$
 MSE – “Mean Squared Error”.



If the errors in the model are distributed normally, than using (3) is equivalent to maximising the following log-likelihood function:

$$\ell(\theta, \hat{\sigma}^2 | Y) = -\frac{T}{2} (\log(2\pi e) + \log \hat{\sigma}^2) \quad (4)$$

where $\hat{\sigma}^2$ is the estimated variance of residuals of the model, θ is a vector of parameters of the model.

This implies that we look at conditional distribution of one-step-ahead forecast error.



This is important, because θ is asymptotically:

- **consistent**, and stabilizes at some value with the increase of T ,
- **efficient**, and doesn't change abruptly with the changes in the sample,
- **unbiased**, and takes values close to the ones it would take in "population".

+ likelihood can be used for information criteria (AIC, BIC etc).
Which allows selecting model that is the most appropriate for data.



The problem of ETS – they do not always produce accurate forecasts for long-term perspective.

This is because optimal α can be too high than needed. For example:

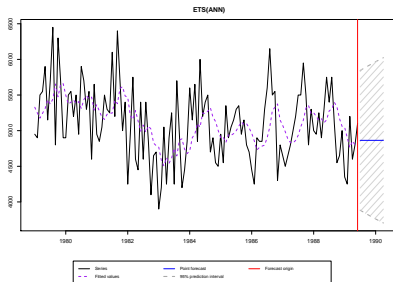


Figure: Series with a higher than needed value of α .



This means that we take short-term information into account more than the long-term.

And we also react on noise too much.

Furthermore, higher smoothing parameter means higher uncertainty.

Higher uncertainty leads to wider prediction intervals.

In inventory it means stocking more than needed.



Advanced estimation methods

To solve this sometimes the forecast task is aligned to the estimation:

$$MSE_h = \frac{1}{T} \sum_{t=1}^T e_{t+h|t}^2 \quad (5)$$

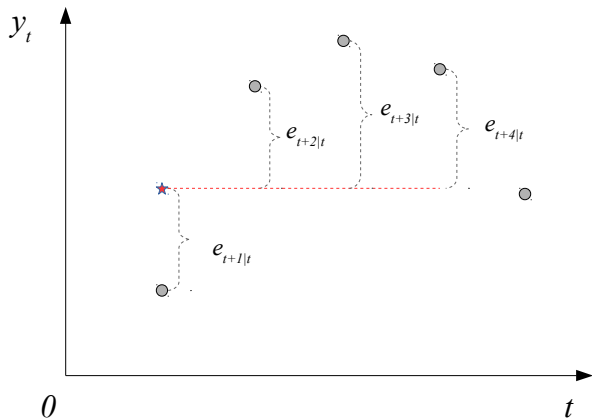
or:

$$MSTFE = \frac{1}{T} \sum_{t=1}^T \sum_{j=1}^h e_{t+j|t}^2 \quad (6)$$

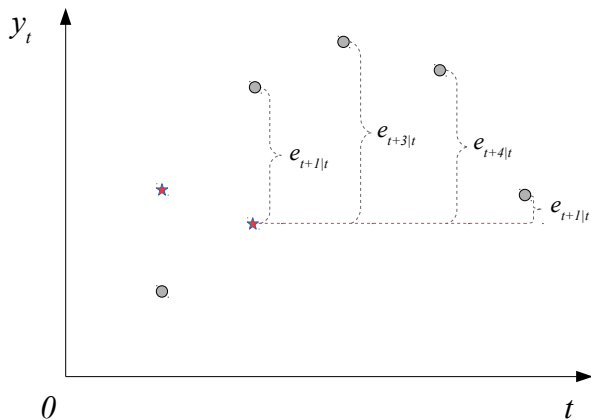
MSTFE – “Mean Squared Trace Forecast Error”.



These cost functions imply that we produce h -steps ahead forecasts from each observation:



These cost functions imply that we produce h -steps ahead forecasts from each observation:



MSE_h produces robust estimates of parameters.

(McElroy and Wildi, 2013, Tiao and Xu, 1993, Clements and Hendry, 2008)

The forecast accuracy increases.

(A. Weiss and Andersen, 1984, A. A. Weiss, 1991, Taylor, 2008, Xia and Tong, 2011)

$MSTFE$ is consistent.

(A. A. Weiss, 1991)

BUT!

The efficiency of estimates of MSE_h is low.

(Tiao and Xu, 1993)

Marcellino, Stock and Watson, 2006 demonstrate on a set of 170 time series that the forecast accuracy using MSE_h is lower than using MSE .



Problems:

- The results are ambiguous;
- Estimates of parameters are inefficient;
- Estimates of parameters could be unstable;
- Nobody has ever explained why MSE_h and $MSTFE$ work / don't work;

- There is no likelihood function for both MSE_h and $MSTFE$;
- Model selection using MSE_h and $MSTFE$ is really tricky;



It can be shown that MSE is proportional to variance of one-step-ahead error.

MSE_h is then proportional to variance of h-step-ahead error.

$MSTFE$ is in fact the sum of MSE_h .

And variance of h-step-ahead error consists of two parts (Hyndman, Koehler, Ord and Snyder, 2008):

1. variance of one-step-ahead error,
2. values of smoothing parameters.

Example: $\sigma_h^2 = \sigma_1^2 \left(1 + \sum_{j=1}^{h-1} \alpha^2 \right)$.



This means that minimising MSE_h (or $MSTFE$) leads to:

1. decrease of variance of one-step-ahead error,
2. shrinkage of values of smoothing parameters towards zero,

Shrinkage in MSE_h is harder than $MSTFE$.

Still the effect will increase with the increase of forecast horizon.



This is the root of the problem and the main advantage of MSE_h and $MSTFE$.

If model is wrong, shrinkage allows to get rid of redundant parameters.

If model is correct, the parameters “overshrink”.

The shrinkage effect becomes stronger when h increases.

It is compensated by the high number of observations.



Examples

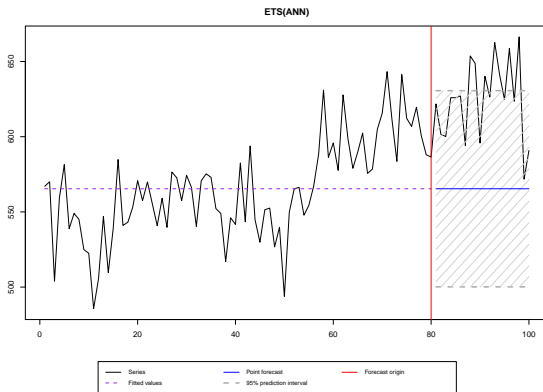


Figure: Series with an obviously wrong α value.

Solution – Trace Forecast Likelihood (TFL)

Let's derive likelihood for multistep cost function.
We need to study multivariate distribution of errors:

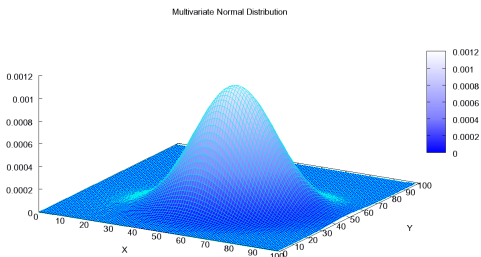


Figure: Multivariate normal distribution.



Based on multivariate normal distribution, we have (skipping derivations):

$$\ell(\theta, \hat{\Sigma}|Y) = -\frac{T}{2} \left(h \log(2\pi e) + \log |\hat{\Sigma}| \right) \quad (7)$$

Looks similar to:

$$\ell(\theta, \hat{\sigma}_1^2|Y) = -\frac{T}{2} \left(\log(2\pi e) + \log \hat{\sigma}_1^2 \right) \quad (8)$$



Σ is covariance matrix that has the structure:

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{1,2} & \dots & \sigma_{1,h} \\ \sigma_{1,2} & \sigma_2^2 & \dots & \sigma_{2,h} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1,h} & \sigma_{2,h} & \dots & \sigma_h^2 \end{pmatrix}, \quad (9)$$

Note that $MSE_h \propto \sigma_h^2$, which makes it a special case of Σ .

And $MSTFE$ is just the sum of diagonals of Σ .



What does $|\hat{\Sigma}|$ mean?

Example with $h = 2$:

$$|\hat{\Sigma}| = \begin{vmatrix} \sigma_1^2 & \sigma_{1,2} \\ \sigma_{1,2} & \sigma_2^2 \end{vmatrix} = \sigma_1^2 \sigma_2^2 - \sigma_{1,2}^2 \quad (10)$$

Minimising determinant of $|\hat{\Sigma}|$ will:

1. decrease variances,
2. increase covariances.



All of them can be rewritten as interaction between variances and parameters

For our example:

$$|\hat{\Sigma}| = \sigma_1^2 \cdot \sigma_1^2(1 + \alpha^2) - (\sigma_1^2\alpha)^2 \quad (11)$$

This means that in general shrinkage effect is weakened.

Finally, model selection is easy:

$$AIC = 2kh - 2\ell(\theta, \hat{\Sigma}|Y) \quad (12)$$



Examples

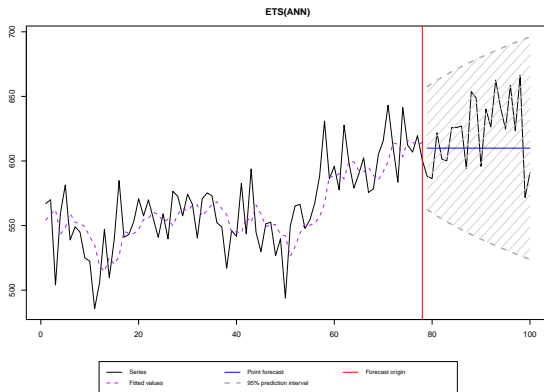


Figure: Model selection and estimation using MSE.



Examples

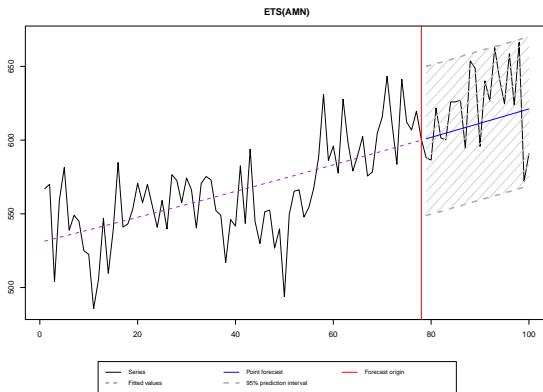


Figure: Model selection and estimation with TFL.



Conclusions

- Multiple steps ahead objective functions imply shrinkage of parameters;
- Parameters of ETS and ARIMA shrink, parameters of regressions do not;
- This gives robustness to models and help in long-term forecasting;
- Parameters may overshrink when estimated using MSE_h and $MSTFE$;



Conclusions

- Trace Forecast Likelihood (TFL) do not overshrink the parameters;
- TFL gives consistent, efficient and unbiased estimates of parameters;
- Using TFL increases long-term forecast accuracy and decreases uncertainty;
- Model selection with TFL is easy.



Thank you for your attention!






Ivan Svetunkov

i.svetunkov@lancaster.ac.uk



Lancaster University
Management School

Lancaster Centre for
Forecasting

-  Clements, M. P. & Hendry, D. F. (2008). *Multi-step estimation for forecasting*.
-  Hyndman, R. J., Koehler, A., Ord, K. & Snyder, R. (2008). *Forecasting with exponential smoothing*. Springer Series in Statistics. Springer Berlin Heidelberg.
doi:10.1007/978-3-540-71918-2
-  Marcellino, M., Stock, J. H. & Watson, M. W. (2006, November). A comparison of direct and iterated multistep ar methods for forecasting macroeconomic time series. *Journal of Econometrics*, 135(1-2), 499–526.
doi:10.1016/j.jeconom.2005.07.020
-  McElroy, T. & Wildi, M. (2013, July). Multi-step-ahead estimation of time series models. *International Journal of Forecasting*, 29(3), 378–394. doi:10.1016/j.ijforecast.2012.08.003
-  Taylor, J. W. (2008). An evaluation of methods for very short-term load forecasting using minute-by-minute british data.



International Journal of Forecasting, 24(4), 645–658.
doi:10.1016/j.ijforecast.2008.07.007



Tiao, G. & Xu, D. (1993). Robustness of maximum likelihood estimates for multi-step predictions: the exponential smoothing case. *Biometrika*, 80(3), 623–641. Retrieved from <http://biomet.oxfordjournals.org/content/80/3/623.short>



Weiss, A. A. (1991, April). Multi-step estimation and forecasting in dynamic models. *Journal of Econometrics*, 48, 135–149.
doi:10.1016/0304-4076(91)90035-C



Weiss, A. & Andersen, A. P. (1984, September). Estimating time series models using the relevant forecast evaluation criterion. *Journal of the Royal Statistical Society. Series A (General)*, 147(3), 484. doi:10.2307/2981579



Xia, Y. & Tong, H. (2011). Feature matching in time series modeling. *Statistical Science*, 26(1), 21–46.
doi:10.1214/10-STS345

