Model parameter estimation with trace forecast likelihood

Ivan Svetunkov and Nikolaos Kourentzes

LCF presentation

7th June 2016

Lancaster University Management School | Lancaster Centre for Forecasting

LCF

Motivation

Parameters estimation is a key element of forecasting.

Poor estimation $\rightarrow$ poor forecasts.

Correct estimation leads to more accurate forecast.

It also decreases the uncertainty.

Conventional Estimation methods

The conventional estimation methods is based on MSE:

$$\text{MSE} = \frac{1}{T} \sum_{t=1}^{T} e_{t+1|t}^2 \tag{1}$$

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 (1) is equivalent to maximising the following log-likelihood function (Hyndman et al., 2008):

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

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.

Advanced estimation methods

Sometimes the forecasting task is aligned to estimation:

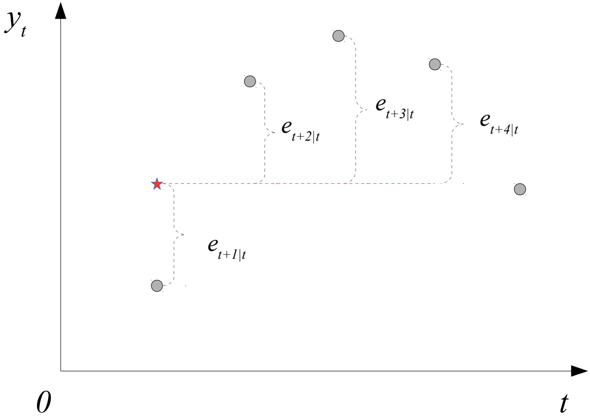
$$\text{MSE}_h = \frac{1}{T} \sum_{t=1}^{T} e^2_{t+h|t} \tag{3}$$

or:

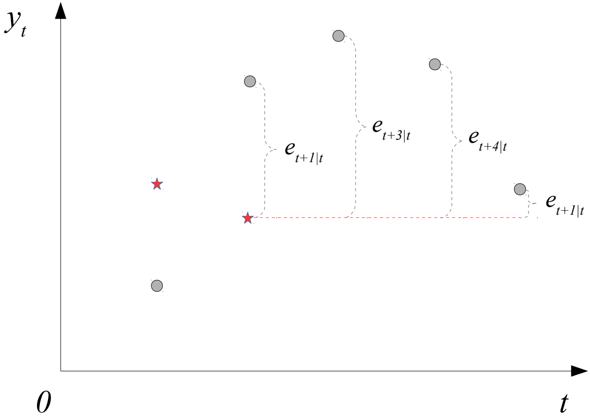
$$\text{MSTFE} = \frac{1}{T} \sum_{t=1}^{T} \sum_{j=1}^{h} e^2_{t+j|t} \tag{4}$$

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.
(**???**)
The forecast accuracy increases.
(**????**)
MSTFE is consistent.
(**?**)

BUT!
The efficiency of estimates of MSE$_h$ is low.
(**?**)
? 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 (??);

Why they work

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 .

Using state-space approach (Snyder, 1985), variance of h-step-ahead error is:

$$\sigma_h^2 = \sigma_1^2 \left(1 + \sum_{j=1}^{h-1} c_j^2 \right) . \tag{5}$$

This means that minimising MSE$_h$ (or MSTFE) in general leads to:

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

Examples:

ETS(A,N,N): $c_j = \alpha$; $\sigma_h^2 = \sigma_1^2 \left(1 + (h-1)\alpha^2\right)$.

ETS(A,A,N): $c_j = \alpha + \beta j$; $\sigma_h^2 = \sigma_1^2 \left(1 + \sum_{j=1}^{h-1} (\alpha + \beta j)^2\right)$.

LCF

This is root of the problem and 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.

Solution – Trace Forecast Likelihood (TFL)

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

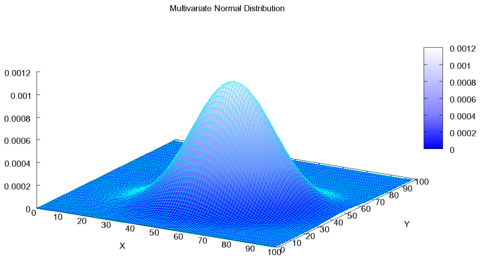


Figure: Multivariate normal distribution.

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

$$\ell(\theta, \hat{\mathbf{\Sigma}}|\mathbf{Y}) = -\frac{T}{2}\left(h \log(2\pi e) + \log|\hat{\mathbf{\Sigma}}|\right) \tag{6}$$

Looks similar to:

$$\ell(\theta, \hat{\sigma_1}^2|\mathbf{Y}) = -\frac{T}{2}\left(\log(2\pi e) + \log \hat{\sigma_1}^2\right) \tag{7}$$

Model selection can now be done using AIC, AICc, BIC, ...

$\boldsymbol{\Sigma}$ is covariance matrix that has the structure:

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_{1,2} & \dots & \sigma_{1,h} \\ \sigma_{1,2} & \sigma_2^2 & \dots & \sigma_{2,h} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1,h} & \sigma_{2,h} & \dots & \sigma_h^2 \end{pmatrix}, \qquad (8)$$

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

And MSTFE is just the trace of Σ.

What does min of $|\mathbf{\Sigma}|$ mean?

Example with $h = 2$:

$$|\mathbf{\Sigma}| = \begin{vmatrix} \sigma_1^2 & \sigma_{1,2} \\ \sigma_{1,2} & \sigma_2^2 \end{vmatrix} = \sigma_1^2 \sigma_2^2 - \sigma_{1,2}^2 \quad (9)$$

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

1. decrease variances,
2. increase covariances.

Covariance between i and j errors is equal to:

$$\sigma_{i,j} = \sigma_1^2 \left(c_{i,j} + \sum_{l=1}^{i-1} c_{l,j} c_{l,i} \right). \quad (10)$$

So $\log |\Sigma|$ can be rewritten as a function of variances and parameters:

$$\log |\mathbf{\Sigma}| = h \log \sigma_1^2 + \log |\mathbf{C}| \quad (11)$$

$\mathbf{C}$ depends on c_j only (thus depends on smoothing parameters).

This means that we shrink parameters...

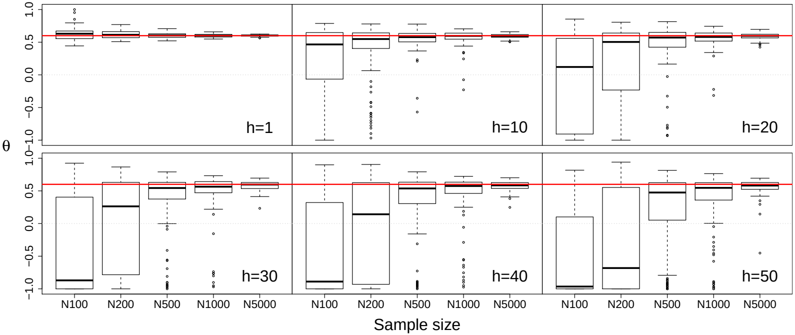
...but shrinkage effect is weakened.

We have conducted a simulation experiment:

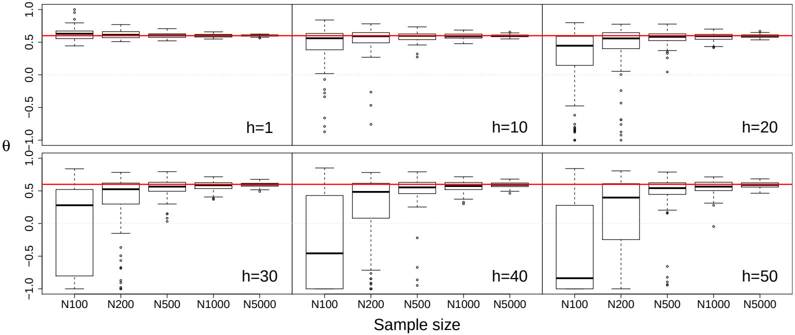
- Generated data using ARIMA(0,1,1).
- Applied correct model and wrong model.
- Applied MSE, MSE$_h$, MSTFE and TFL.
- With h=$\{1, 10, 20, 30, 40, 50\}$.
- Wrote down the parameters...

LCF

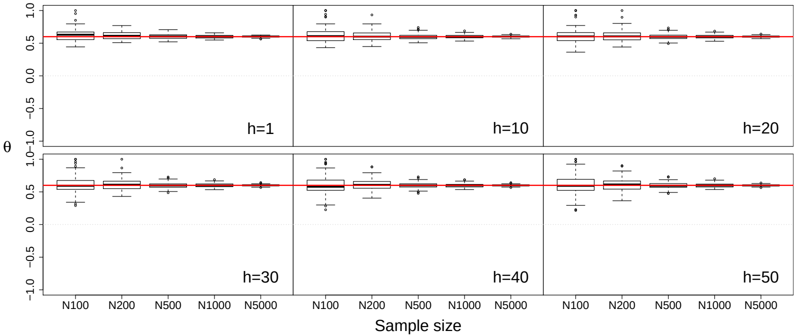
Simulations. ARIMA(0,1,1). Correct model, MSE_h



Simulations. ARIMA(0,1,1). Correct model, MSTFE



Simulations. ARIMA(0,1,1). Correct model, 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;
- Model selection with TFL is possible.

TFL is brilliant in theory!

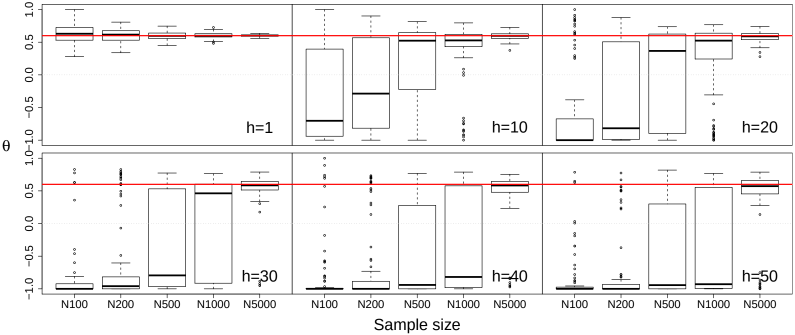
How to make it work in practice?...

Thank you for your attention!

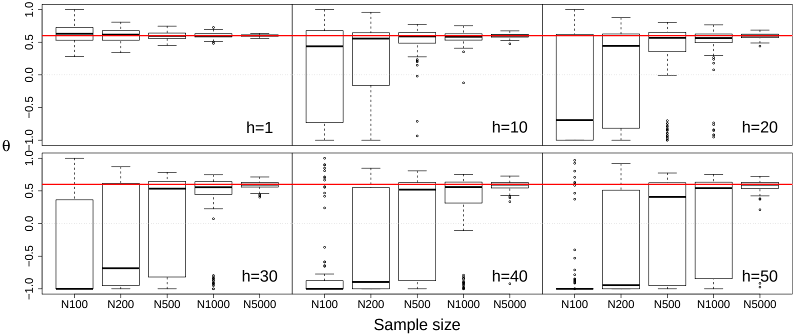
Ivan Svetunkov
i.svetunkov@lancaster.ac.uk

Lancaster University Management School | Lancaster Centre for Forecasting

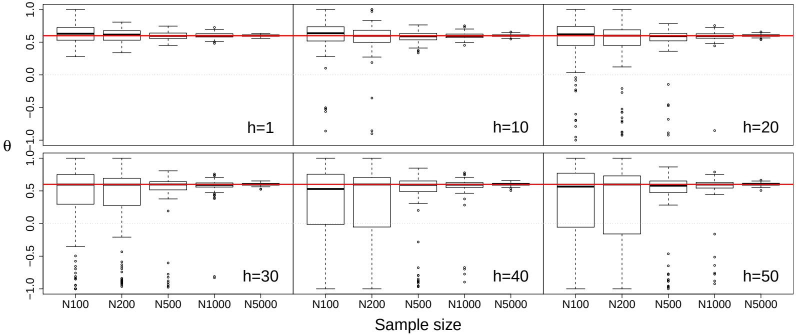
Simulations. ARIMA(0,1,1). Wrong model, MSE_h



Simulations. ARIMA(0,1,1). Wrong model, MSTFE



Simulations. ARIMA(0,1,1). Wrong model, TFL



Bhansali, R., 1996. Asymptotically efficient autoregressive model selection for multistep prediction. Annals of the Institute of Statistical Mathematics 48 (3), 577–602.

Bhansali, R., 1997. Direct autoregressive predictors for multistep prediction: Order selection and performance relative to the plug in predictors. Statistica Sinica 7, 425–449.

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 Berlin Heidelberg.

Marcellino, M., Stock, J. H., Watson, M. W., 2006. A comparison of direct and iterated multistep AR methods for forecasting macroeconomic time series. Journal of Econometrics 135 (1-2), 499–526.

McElroy, T., Wildi, M., 2013. Multi-step-ahead estimation of time

LCF

series models. International Journal of Forecasting 29 (3), 378–394.

Snyder, R. D., 1985. Recursive Estimation of Dynamic Linear Models. Journal of the Royal Statistical Society, Series B (Methodological) 47 (2), 272–276.

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.

Tiao, G., Xu, D., 1993. Robustness of maximum likelihood estimates for multi-step predictions: the exponential smoothing case. Biometrika 80 (3), 623–641.

Weiss, A., Andersen, A. P., 1984. Estimating Time Series Models Using the Relevant Forecast Evaluation Criterion. Journal of the Royal Statistical Society. Series A (General) 147 (3), 484.

Weiss, A. A., 1991. Multi-step estimation and forecasting in dynamic models. Journal of Econometrics 48, 135–149.

Xia, Y., Tong, H., 2011. Feature matching in time series modeling. Statistical Science 26 (1), 21–46.