Model parameter estimation with trace forecast likelihood

Ivan Svetunkov and Nikolaos Kourentzes

LCF presentation

7th June 2016
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.
The conventional estimation methods is based on MSE:

\[ \text{MSE} = \frac{1}{T} \sum_{t=1}^{T} e_{t+1|t}^2 \]  \hspace{1cm} (1)

where \( e_{t+1|t} = y_{t+1} - \hat{y}_{t+1} \)

MSE – “Mean Squared Error”.

MSE - “Mean Squared Error”.
If the errors in the model are distributed normally, then using (1) is equivalent to maximising the following log-likelihood function (Hyndman et al., 2008):

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

(2)

where $\hat{\sigma}^2$ is the estimated variance of residuals of the model, $\theta$ 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:

\[
\text{MSE}_h = \frac{1}{T} \sum_{t=1}^{T} e_{t+h|t}^2 \tag{3}
\]

or:

\[
\text{MSTFE} = \frac{1}{T} \sum_{t=1}^{T} \sum_{j=1}^{h} e_{t+j|t}^2 \tag{4}
\]

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:
MSEₜ produces robust estimates of parameters.
(???)
The forecast accuracy increases.
(????)
MSTFE is consistent.
(?)

BUT!
The efficiency of estimates of MSEₜ is low.
(?)
? demonstrate on a set of 170 time series that the forecast accuracy using MSEₜ 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 $\text{MSE}_h$ and MSTFE work / don’t work;
- There is no likelihood function for both $\text{MSE}_h$ and MSTFE;
- Model selection using $\text{MSE}_h$ and MSTFE is really tricky (??);
Why they work

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

\[ \text{MSE}_h \text{ is then proportional to variance of } h\text{-step-ahead error.} \]

\[ \text{MSTFE is in fact the sum of } \text{MSE}_h. \]

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

\[ \sigma^2_h = \sigma^2_1 \left( 1 + \sum_{j=1}^{h-1} c^2_j \right). \]  \hspace{1cm} (5)
This means that minimising $\text{MSE}_h$ (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): $c_j = \alpha + \beta j; \sigma_h^2 = \sigma_1^2 \left( 1 + \sum_{j=1}^{h-1} (\alpha + \beta j)^2 \right)$. 
This is root of the problem and main advantage of $\text{MSE}_h$ 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.
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):

$$\ell(\theta, \hat{\Sigma}|Y) = -\frac{T}{2} \left( h \log(2\pi e) + \log |\hat{\Sigma}| \right)$$  \hspace{1cm} (6)

Looks similar to:

$$\ell(\theta, \hat{\sigma}_1^2|Y) = -\frac{T}{2} \left( \log(2\pi e) + \log \hat{\sigma}_1^2 \right)$$  \hspace{1cm} (7)

Model selection can now be done using AIC, AICc, BIC, ...
\( \Sigma \) is covariance matrix that has the structure:

\[
\Sigma = \begin{pmatrix}
\sigma_1^2 & \sigma_{1,2} & \ldots & \sigma_{1,h} \\
\sigma_{1,2} & \sigma_2^2 & \ldots & \sigma_{2,h} \\
\vdots & \vdots & \ddots & \vdots \\
\sigma_{1,h} & \sigma_{2,h} & \ldots & \sigma_h^2
\end{pmatrix},
\]

Note that \( MSE_h \propto \sigma_h^2 \), which makes it a special case of \( \Sigma \).

And MSTFE is just the trace of \( \Sigma \).
What does \( \min \) of \(|\Sigma|\) mean?

Example with \( h = 2 \):

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

Minimising determinant of \(|\Sigma|\) will:

1. decrease variances,
2. increase covariances.
Covariance between $i$ and $j$ errors is equal to:

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

(10)

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

$$\log|\Sigma| = h \log \sigma_1^2 + \log|C|$$

(11)

$C$ 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...
Simulations. ARIMA(0,1,1). Correct model, $\text{MSE}_h$
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 $\text{MSE}_h$ 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
Simulations. ARIMA(0,1,1). Wrong model, $\text{MSE}_h$
Simulations. ARIMA(0,1,1). Wrong model, MSTFE

Model parameter estimation with trace forecast likelihood
Simulations. ARIMA(0,1,1). Wrong model, TFL


