Trace forecast likelihood

Ivan Svetunkov[1] Nikos Kourentzes[1]

[1]Lancaster Centre for Forecasting
Lancaster University

Ask Stephan Kolassa
(ASK, 2016)

Before we start

What is “trace forecast”? (Weiss and Andersen, 1984)?
Here it is:

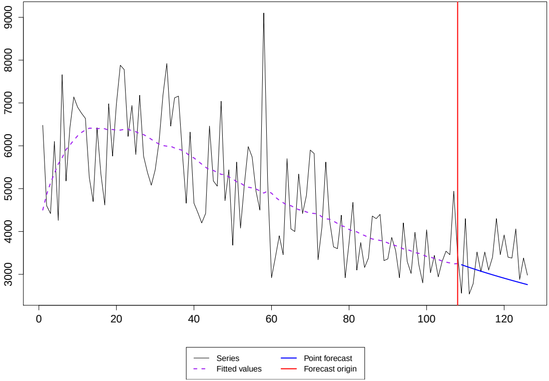

Figure: Trace forecast of a random model.

Introduction

There is a lot of different estimation methods.

Sometimes in practice simple heuristics is used. For example:

$$MSE_h = \frac{1}{T-h} \sum_{t=1}^{T-h} e^2_{t+h|t}, \tag{1}$$

or Trace Forecast MSE:

$$TFMSE = \frac{1}{h} \sum_{j=1}^{h} \frac{1}{T-j} \sum_{t=1}^{T-j} e^2_{t+j|t}, \tag{2}$$

where $e_{t+j|t}$ is a forecast error, produced from observation t.

Literature

Theory gives insights on the work of these (only for ARIMA):

1. Using 1-step ahead forecast error may lead to bias even when model is specified correctly due to finite sample (Clements and Hendry, 2008). Using MSE_h minimises this bias.
2. If model is wrong then the usage of MSE_h leads to the convergence of parameters to “pseudo-true” values (McElroy and Wildi, 2013).
3. Using *TFMSE* increases accuracy of ARIMA (Weiss and Andersen, 1984; Weiss, 1991). Estimates are consistent and asymptotically normal.
4. Using *TFMSE* reduces bias and leads to more robust parameters estimation (Tiao and Xu (1993), Xia and Tong (2011)).

Problems

- Standard one-step ahead cost function is derived from the likelihood.

- Cost functions (1) and (2) do not have a proper statistical rationale.

- Cost functions (1) and (2) work in practice, but nobody knows exactly why.

Solution

Estimate the joint distribution of 1 to h steps ahead conditional errors.

$$Y_t = \begin{pmatrix} y_{t+1} \\ y_{t+2} \\ \vdots \\ y_{t+h} \end{pmatrix} \text{ and } Y = \{Y_1, Y_2, \ldots, Y_T\} \tag{3}$$

So we want to estimate likelihood:

$$L(\theta, \Sigma | Y_t) = P(Y_t | \theta, \Sigma), \tag{4}$$

which for the whole sample T will be:

$$L(\theta, \Sigma | Y) = \prod_{t=1}^{T} P(Y_t | \theta, \Sigma). \tag{5}$$

Solution

For the multivariate normal distribution likelihood (5) transforms into:

$$L(\theta, \Sigma | Y) = \prod_{t=1}^{T} \left[(2\pi)^{-\frac{h}{2}} \left|\Sigma\right|^{-\frac{1}{2}} \exp\left(-\frac{1}{2} E_t{}' \Sigma^{-1} E_t \right) \right] \qquad (6)$$

$$\text{where } E_t = \begin{pmatrix} e_{t+1|t} \\ e_{t+2|t} \\ \vdots \\ e_{t+h|t} \end{pmatrix} \text{ and } \hat{\Sigma} = \frac{1}{T-h} \sum_{t=1}^{T-h} E_t E_t'$$

Solution

Concentrated log-likelihood using estimated $\hat{\Sigma}$ is much simpler:

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

So:

- Model selection can be performed using any information criteria (for example, AIC);
- Maximisation of (7) is equivalent to minimisation of the generalised variance (GV):

$$GV = |\hat{\Sigma}| \tag{8}$$

Solution

What is Σ?

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

$$\log|\hat{\Sigma}| = \sum_{j=1}^{h} \log \sigma_j^2 + \log|R|, \tag{10}$$

where $R = \begin{pmatrix} 1 & r_{1,2} & \dots & r_{1,h} \\ r_{1,2} & 1 & \dots & r_{2,h} \\ \vdots & \vdots & \ddots & \vdots \\ r_{1,h} & r_{2,h} & \dots & 1 \end{pmatrix}$ is the correlation matrix.

Solution

Minimising GV means the decrease of log variances and increase of correlations between some errors.

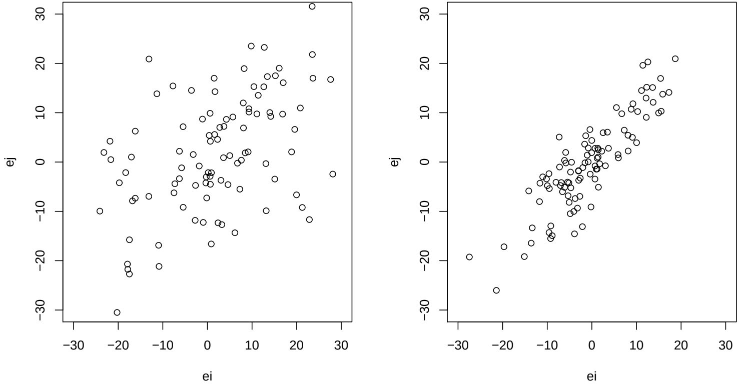

Figure: Scatter plots before and after the minimisation of GV.

Solution

One of the options – ignore the correlation matrix and minimise log variances using:

$$TLV = \sum_{j=1}^{h} \log\left(\sigma_j^2\right) \tag{11}$$

This is a different cost function than *TFMSE* (which is proportional to “Total Variation”):

$$TV = \sum_{j=1}^{h} \sigma_j^2$$

Solution

GV and TLV have advantages over TFMSE and MSE_h:

1. The variance of short-term forecast errors is not concealed by the long-term.
2. In a way GV encompasses other cost functions mentioned above.
3. GV can be used in the likelihood estimation.

State-space models

Using Hyndman et al. (2008) SSOE additive state-space model the j-steps ahead variance is:

$$\sigma_j^2 = \begin{cases} \sigma_1^2 & \text{if } j = 1 \\ \sigma_1^2 \left(1 + \sum_{i=1}^{j-1} (c_{i,j})^2\right) & \text{if } j > 1 \end{cases} \tag{12}$$

Substituting (12) in (11) leads to:

$$\log GV = h \log\left(\sigma_1^2\right) + \sum_{j=1}^{h} \log\left(1 + \sum_{i=1}^{j-1} c_{i,j}^2\right) + \log|R| \tag{13}$$

where $c_{i,j} = w' F^{j-i-1} g$

Some conclusions:

1. **This is shrinkage!**
2. All the multi steps functions impose shrinkage on parameters;
3. But this is not LASSO style shrinkage;
4. When $h \to \infty$, model becomes deterministic;
5. TLV and GV shrink more gently, because of the change of scale.

Parameters of ETS and ARIMA shrink.
Parameters of regression do not shrink.

log GV can also be represented as:

$$\log GV = h \log \sigma_1^2 + \log |A|, \tag{14}$$

where $A = \begin{pmatrix} a_{1,1} & \dots & a_{1,h} \\ \vdots & \ddots & \vdots \\ a_{1,h} & \dots & a_{h,h} \end{pmatrix}$;

$$a_{j,k} = \begin{cases} 1 + \sum_{i=1}^{j-1} c_{i,j}^2 & ,j = k \\ c_{k,j} + \sum_{i=1}^{j-1} c_{i,k} c_{i,j} & ,j \leq k \end{cases} ;\; c_{i,j} = w' F^{j-i-1} g.$$

State-space models

The previous slide means that by minimising log GV:

1. The one step ahead variance is minimised;
2. The shrinkage is imposed on parameters;
3. The shrinkage effect is weakened (because of $a_{j,k \neq j}$);
4. So the parameters will not over shrink as h increases.

M3 Data Experiment

M3 monthly data, fixed origin, $h = 18$.

ETS in "smooth" package for R (`https://github.com/config-i1/smooth`).

Model selection using the conventional AIC based on 1 step ahead error.

Several estimation methods:

1. Conventional Hyndman and Khandakar (2008),
2. MSE_h optimised only once (MSE_h),
3. *TFMSE*, aka Total Variation (TV),
4. Total Logarithmic Variation (TLV),
5. Generalised Variance (GV),

M3 Data Experiment

Table: Mean errors

Method	MPE	MAPE	SMAPE	MASE
Conventional	-11.861	23.559	14.373	2.091
MSE_h	**-7.540**	**21.566**	15.597	2.479
TV	-10.019	22.179	14.645	2.235
TLV	-9.628	21.971	**14.598**	2.234
GV	-9.750	23.357	15.508	**2.226**

M3 Data Experiment

Table: Median errors

Method	MPE	MAPE	SMAPE	MASE
Conventional	0.206	9.184	9.106	1.087
MSE_h	0.382	10.025	10.148	1.177
TV	**0.333**	9.362	9.296	1.115
TLV	0.548	**9.201**	**9.235**	**1.108**
GV	0.572	9.260	9.348	1.166

Conclusions

- Trace forecast likelihood gives a statistical rationale for some multiple steps ahead cost functions;
- Model selection can easily be done using GV;
- Maximisation of trace likelihood is equivalent to minimisation of GV;
- Any multi-steps ahead cost function implies shrinkage (towards deterministic function);
- Shrinkage happens naturally in trace forecast likelihood;
- In theory the proposed approach is wonderful...
- ...in practise it doesn't work... yet!

That's all, folks!

Thank you for your attention!

Ivan Svetunkov, Nikolaos Kourentzes

Lancaster Centre for Forecasting,
Lancaster University

i.svetunkov@lancaster.ac.uk

References

Clements, M. P., Hendry, D. F., 2008. Multi-Step Estimation for Forecasting.

Hyndman, R. J., Khandakar, Y., 2008. Automatic time series forecasting: the forecast package for R. Journal Of Statistical Software 27 (3), 1–22.

Hyndman, R. J., Koehler, A., Ord, K., Snyder, R., 2008. Forecasting with Exponential Smoothing. Springer Berlin Heidelberg.

McElroy, T., Wildi, M., 2013. Multi-step-ahead estimation of time series models. International Journal of Forecasting 29 (3), 378–394.

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.