

Trace likelihood and shrinkage in estimation of the forecasting models

Ivan Svetunkov¹ Nikos Kourentzes¹

¹Lancaster Centre for Forecasting
Lancaster University

EURO 2015

Introduction

Several assumptions in model estimation:

- 1 Model is specified correctly,
- 2 No heteroscedasticity in residuals,
- 3 No autocorrelation in residuals,
- 4 Residuals are distributed normally, $\epsilon_t \sim N(0, \sigma^2)$

Violation of any of these assumptions may lead to bias in forecasts.

Introduction

Possible solution:

Change cost function to sum of multi-step ahead forecast errors (which corresponds to Total Variation):

$$TV = \sum_{j=1}^h \sigma_j^2, \quad (1)$$

where $\sigma_j^2 = \frac{1}{T} \sum_{t=1}^T \epsilon_{t+j}^2 = \frac{1}{T} \sum_{t=1}^T (y_{t+j} - \mu_{t+j|t})^2$ is conditional variance of errors for the j observations ahead and $j = 1, \dots, h$

- ① Using 1-step ahead forecast error may lead to bias even when model is specified correctly due to finite sample (Clements and Hendry [1996], Hansen [2010]).
- ② Using (1) in this situation reduces bias and leads to more robust parameters estimation (Tiao and Xu [1993], Xia and Tong [2011]).
- ③ Using (1) for trace forecasts is beneficial [Weiss and Andersen, 1984].

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

Function (1) has a very complicated rationale.

Beware of formulas!

Solution

Estimate the joint distribution of 1 to h steps ahead errors.
Introducing the covariance matrix of errors:

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

where $\sigma_{i,j} = \text{cov}(\epsilon_{t+i}, \epsilon_{t+j})$

Using (2) the general likelihood for trace forecast is:

$$L(\theta, \Sigma | \mathbf{y}) = \prod_{t=1}^T p(y_{t+1}, y_{t+2}, \dots, y_{t+h} | \theta, \Sigma), \quad (3)$$

where \mathbf{y} is the vector of all the actual values,
 θ is the vector of all the parameters of a model.

Solution

Assuming normal distribution of the residuals...

...taking logarithms...

...using estimated covariance matrix...

...and after several magic passes with trace function:

$$\ell(\theta, \Sigma | \mathbf{y}) = -\frac{T}{2} \left(h \log(2\pi e) + \log |\hat{\Sigma}| \right), \quad (4)$$

Conclusion

- Model selection can be performed using any information criteria (for example, AIC);
- Maximisation of (4) is equivalent to minimisation of the generalised variance (GV):

$$GV = \log(|\hat{\Sigma}|) \quad (5)$$

Cost functions

Minimising GV, the forecast autocorrelation may increase.

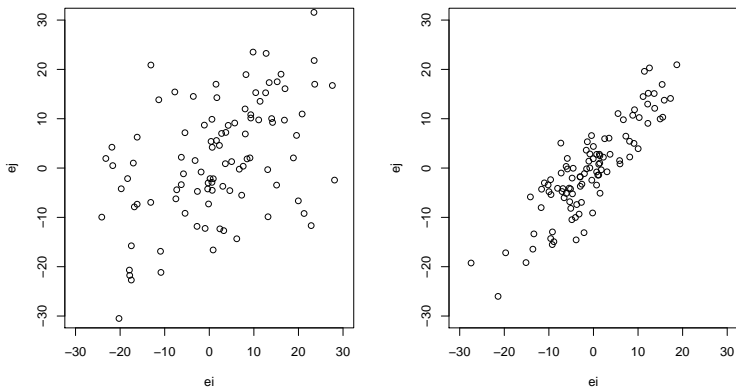


Figure: Scatter plots before and after the minimisation of GV.

Cost functions

One of the possible solutions - ignore the problem:
(assume no autocorrelation)

$$\Sigma = \begin{pmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \sigma_h^2 \end{pmatrix}, \quad (6)$$

The cost function (5) changes to:

$$TLV = \sum_{j=1}^h \log(\sigma_j^2) \quad (7)$$

This is a different cost function than (1):

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

State-space models

Due to Hyndman et al. [2008] the h-steps ahead variance for additive state-space models is:

$$\sigma_j^2 = \begin{cases} \sigma_1^2 & \text{if } j = 1 \\ \sigma_1^2 \left(1 + \sum_{i=1}^{j-1} c_i^2 \right) & \text{if } j > 1 \end{cases} \quad (8)$$

Substituting (8) in (7) leads to:

$$TLV = \sum_{j=1}^h \log \left(\sigma_1^2 \left(1 + \sum_{i=1}^{j-1} c_i^2 \right) \right) \quad (9)$$

or:

$$TLV = h \log (\sigma_1^2) + \sum_{j=1}^h \log \left(1 + \sum_{i=1}^{j-1} c_i^2 \right) \quad (10)$$

For ETS(A,N,N) $c_j = \alpha$. Substituting it in (10) leads to:

$$TLV = h \log(\sigma_1^2) + \sum_{j=1}^h \log \left(1 + \sum_{i=1}^{j-1} \alpha^2 \right) \quad (11)$$

Conclusion

- 1 Using the trace likelihood on ETS(A,N,N) leads to shrinkage of α ;
- 2 Speed of shrinkage increases with the increase of the horizon.

ETS examples

For ETS(A,A,N) $c_i = \alpha + \beta i$. Which gives:

$$TLV = h \log(\sigma_1^2) + \sum_{j=1}^h \log \left(1 + \sum_{i=1}^{j-1} (\alpha + \beta i)^2 \right) \quad (12)$$

Conclusion

- 1 Both α and β shrink;
- 2 β shrinks faster with the increase of horizon.

Remark

In ETS(A,Ad,N) ϕ slows down the shrinkage of β .

ARIMA example

ARIMA can be represented in state-space form [Hyndman et al., 2008, p.174]. For general ARIMA c_i can be:

$$c_i = (1 \quad 0 \quad \dots \quad 0) \begin{pmatrix} \eta_1 & I_{k-1} \\ \dots & \dots \\ \eta_k & 0 \end{pmatrix}^{i-1} \begin{pmatrix} \eta_1 - \theta_1 \\ \dots \\ \eta_k - \theta_k \end{pmatrix} \quad (13)$$

Conclusion

- 1 In general AR terms shrink towards zero in a non-linear manner;
- 2 MA terms shrink towards AR terms.

Regression example

Any regression model can also be represented in state-space form.

But variance of a regression depends only on exogenous variables values.

Conclusion

- 1 Coefficients of such a regression do not shrink when any multi-step ahead cost function is used.

Simulation. Setup

Data was generated using ETS(A,N,N), ETS(A,A,N), ETS(A,Ad,N).
1000 time series in each.
 $T = 25, 50, 100, 1000$.

The same models were estimated using:

- the conventional cost function - "None",
- standard h-steps ahead cost function (Total Variation) - "TV",
- full Σ matrix (Generalised Variance) - "GV",
- diagonal Σ matrix (Total Logarithmic Variation) - "TLV".

$h = 5, 10, 20, 100$

MASE was used in the forecast error estimation.

"sim.ets" and "ets2" in "TStools" for R:

<https://github.com/trnnick/TStools>

Using correct models to forecast the data

Simulation. Correct model

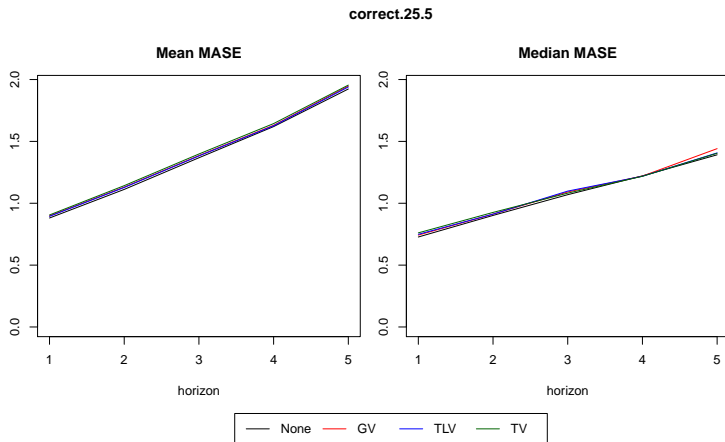


Figure: $T=25$, $h=5$

Simulation. Correct model

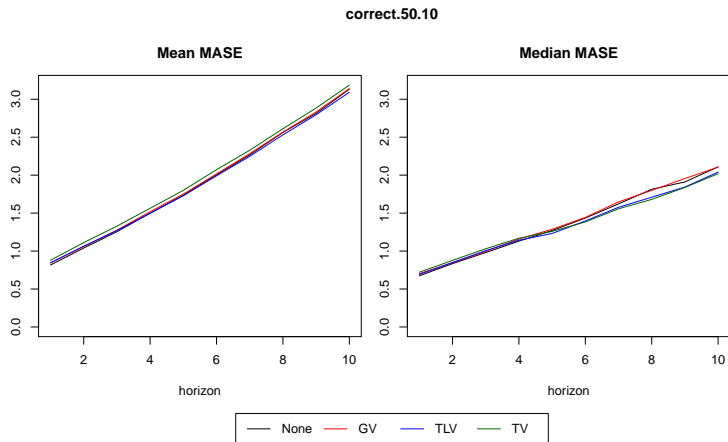


Figure: $T=50$, $h=10$

Simulation. Correct model

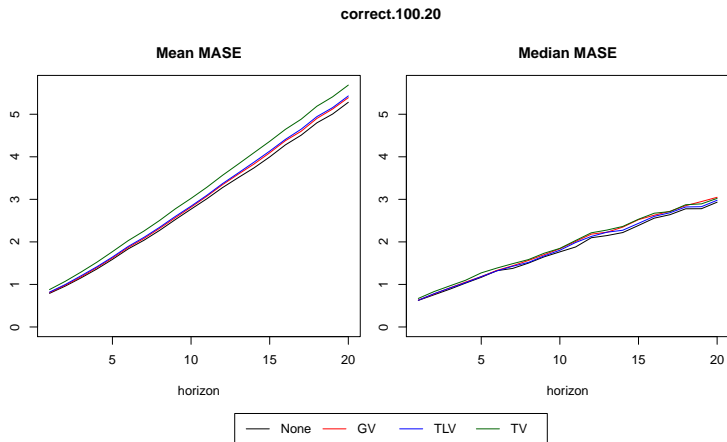


Figure: $T=100$, $h=20$

Simulation. Correct model

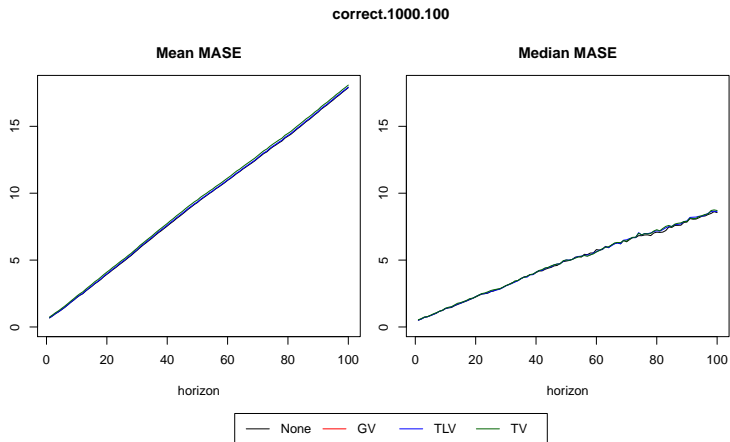


Figure: $T=1000$, $h=100$

Using wrong models

Simulation. Wrong model

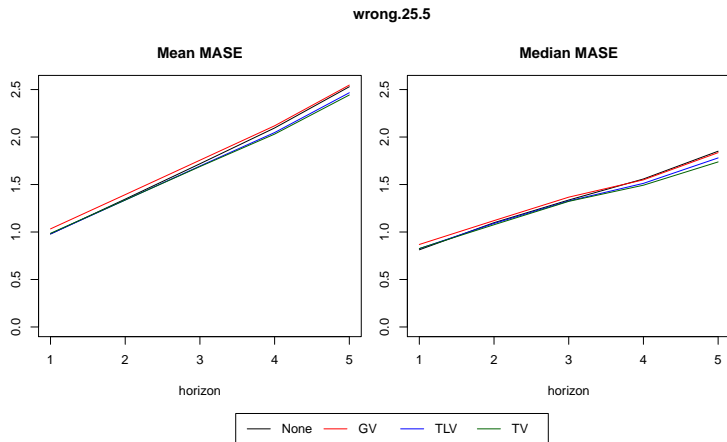


Figure: $T=25$, $h=5$

Simulation. Wrong model

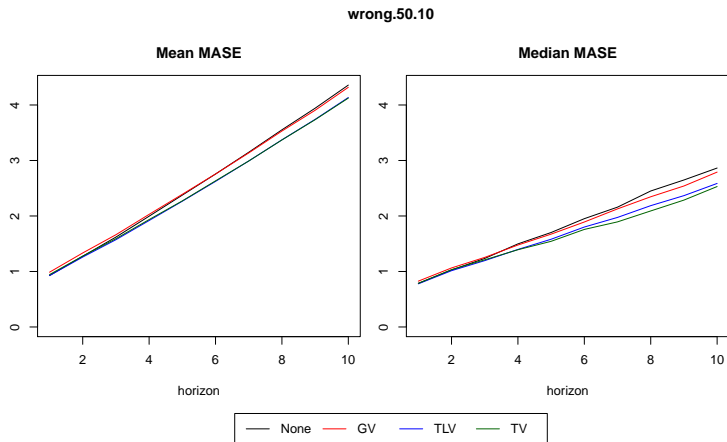


Figure: $T=50$, $h=10$

Simulation. Wrong model

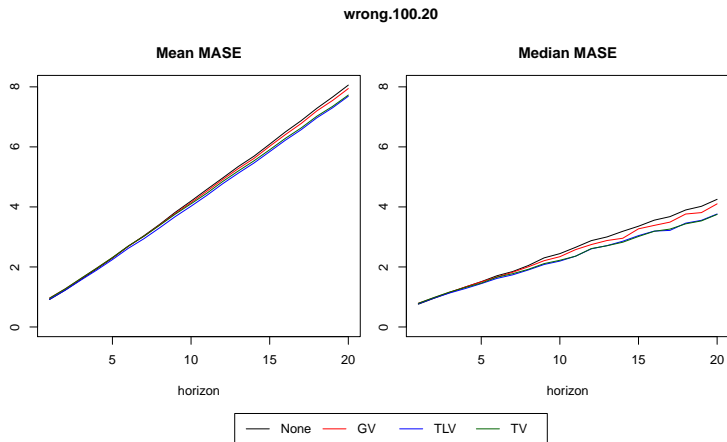


Figure: $T=100$, $h=20$

Simulation. Wrong model

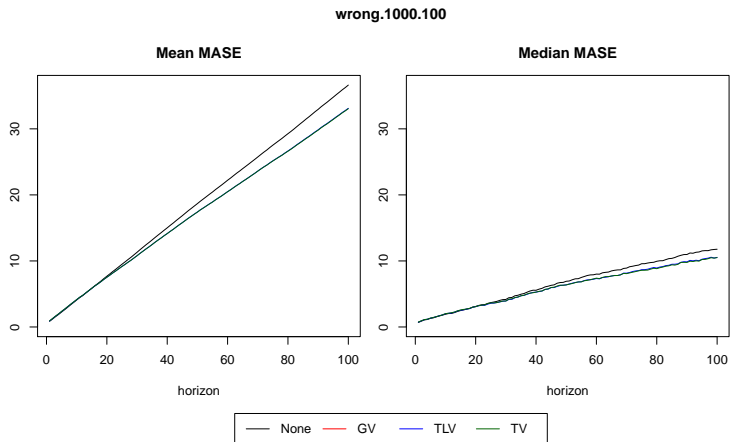


Figure: $T=1000$, $h=100$

Conclusions

- Trace likelihood gives a statistical explanation of multiple steps ahead cost function;
- Model selection can easily be done using trace likelihood;
- Maximisation of trace likelihood is equivalent to minimisation of generalised variance;
- Using diagonal Σ leads to the shrinkage of parameters;
- Shrinkage happens naturally in trace likelihood;
- Using the method leads to the increase in the accuracy in the long-term...
- ... without a significant loss in the short-term.

That's all, folks!

Thank you for your attention!

Ivan Svetunkov, Nikolaos Kourentzes

Lancaster Centre for Forecasting,
Lancaster University

i.svetunkov@lancaster.ac.uk

Several more kilograms of slides...

Using correct models

Appendix. Correct model, ETS(A,A,N)

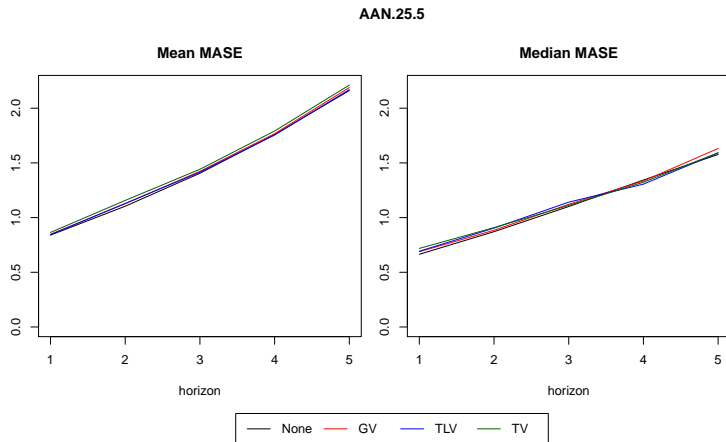


Figure: $T=25, h=5$

Appendix. Correct model, ETS(A,A,N)

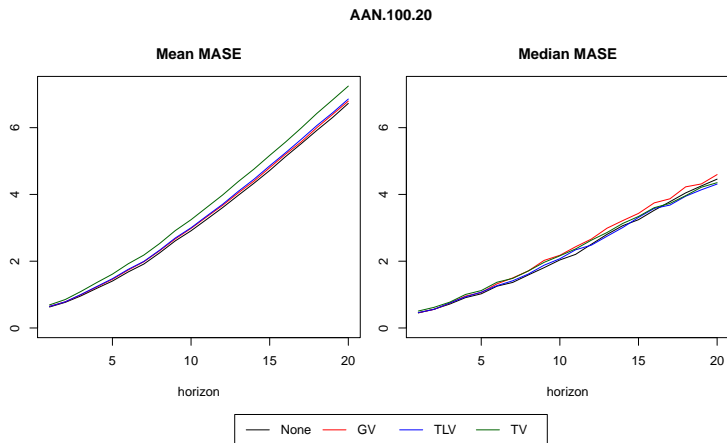


Figure: $T=100$, $h=20$

Appendix. Correct model, ETS(A,A,N)

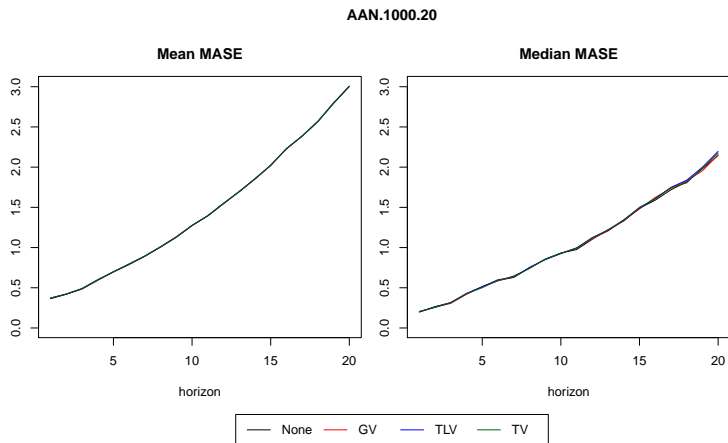


Figure: $T=1000$, $h=20$

Appendix. Correct model, ETS(A,A,N)

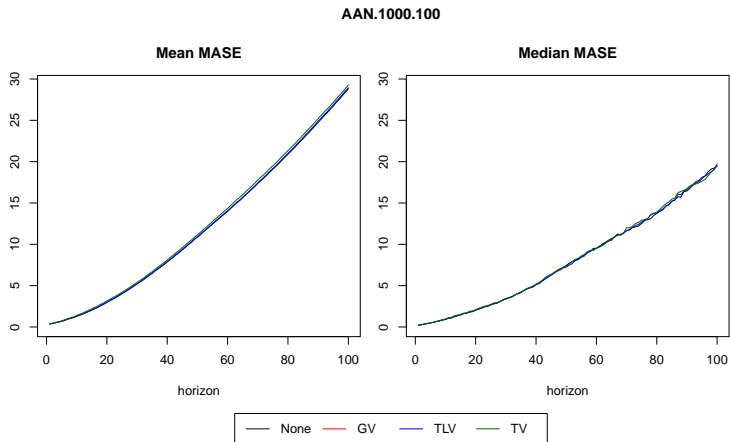


Figure: $T=1000$, $h=100$

Appendix. Correct model, ETS(A,Ad,N)

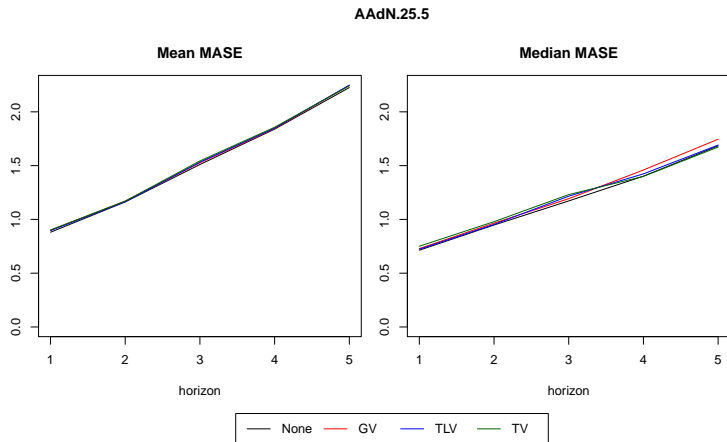


Figure: $T=25, h=5$

Appendix. Correct model, ETS(A,Ad,N)

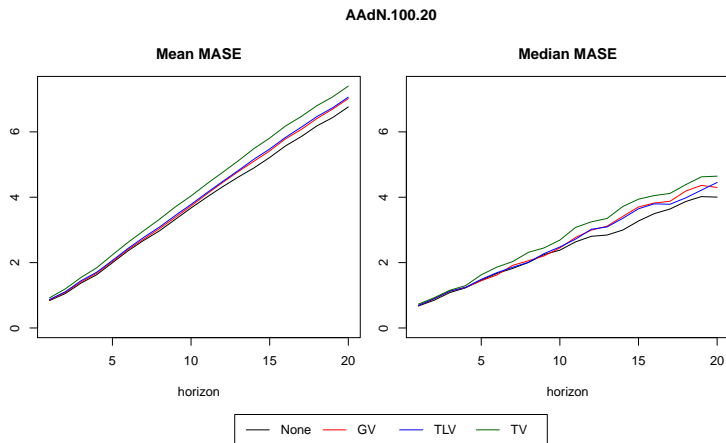


Figure: $T=100$, $h=20$

Appendix. Correct model, ETS(A,Ad,N)

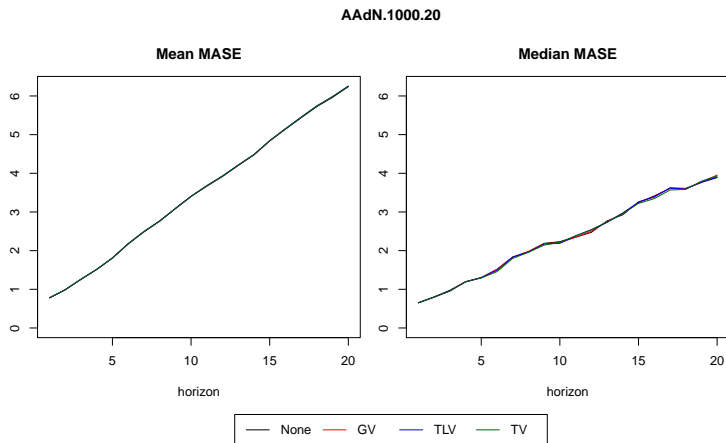


Figure: $T=1000$, $h=20$

Appendix. Correct model, ETS(A,Ad,N)

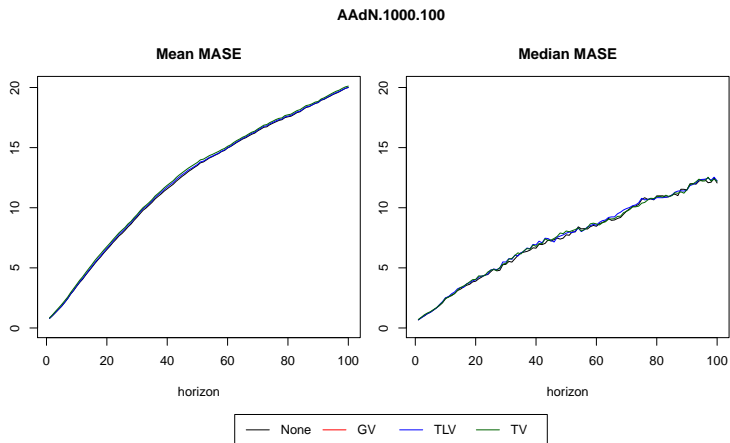


Figure: $T=1000$, $h=100$

Using wrong models

Appendix. ETS(A,A,N) on ETS(A,N,N) data

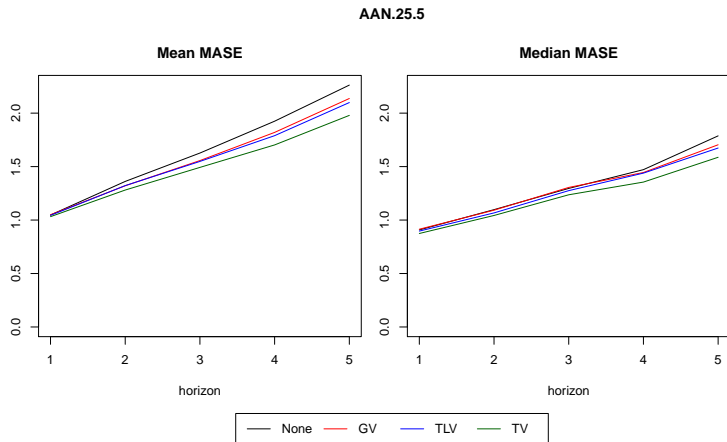


Figure: $T=25$, $h=5$

Appendix. ETS(A,A,N) on ETS(A,N,N) data

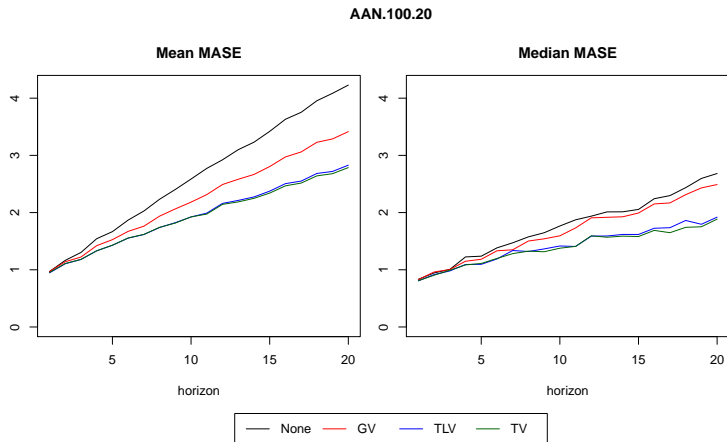


Figure: $T=100$, $h=20$

Appendix. ETS(A,A,N) on ETS(A,N,N) data

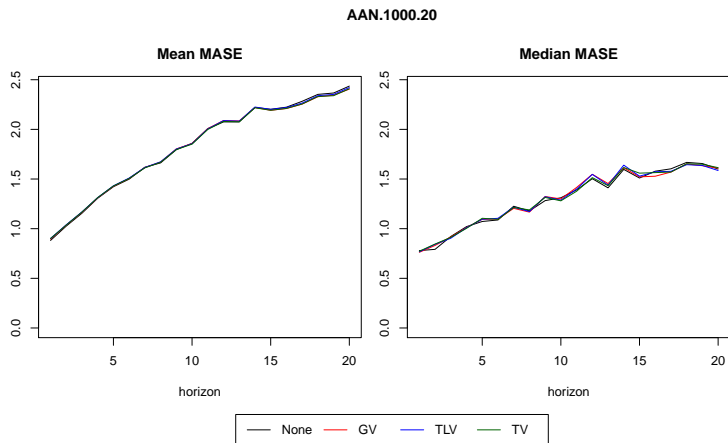


Figure: $T=1000$, $h=20$

Appendix. ETS(A,A,N) on ETS(A,N,N) data

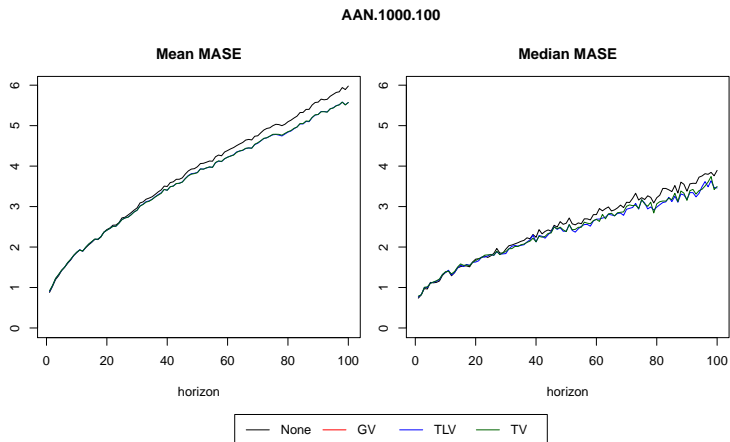


Figure: $T=1000$, $h=100$

Appendix. ETS(A,Ad,N) on ETS(A,N,N) data

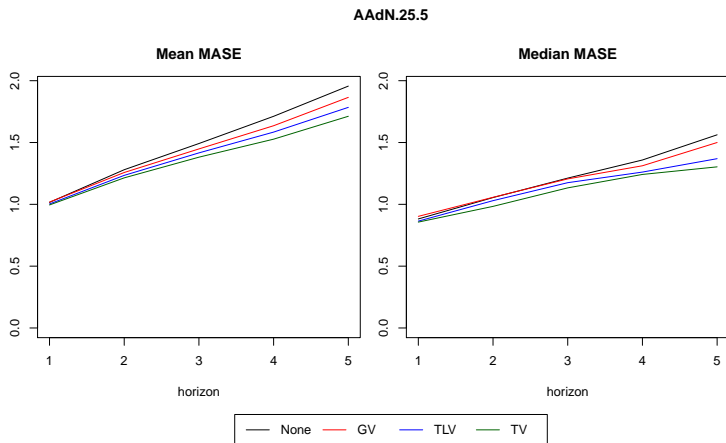


Figure: $T=25, h=5$

Appendix. ETS(A,Ad,N) on ETS(A,N,N) data

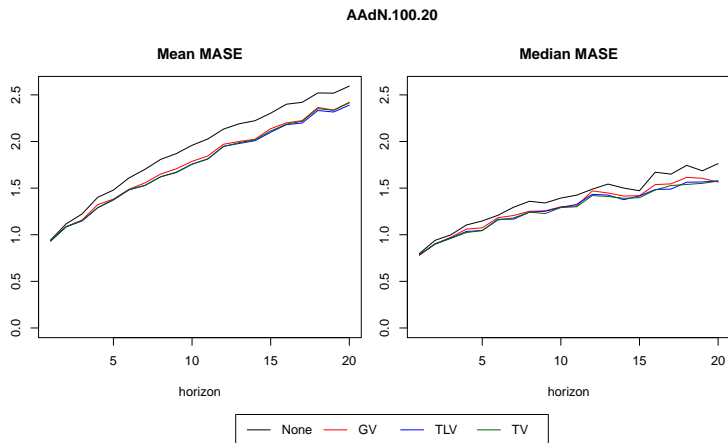


Figure: $T=100$, $h=20$

Appendix. ETS(A,Ad,N) on ETS(A,N,N) data

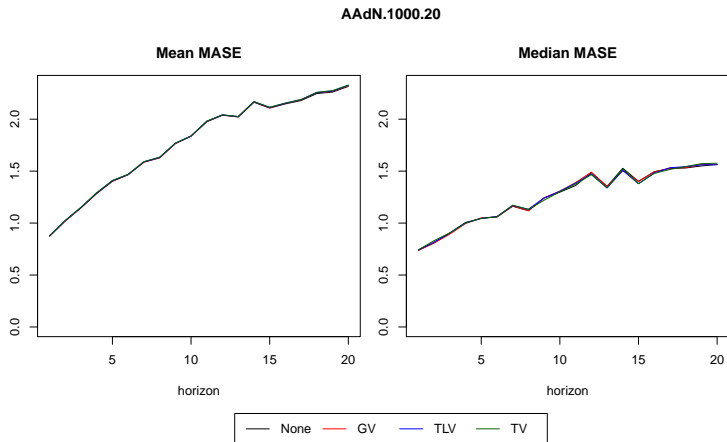


Figure: $T=1000$, $h=20$

Appendix. ETS(A,Ad,N) on ETS(A,N,N) data

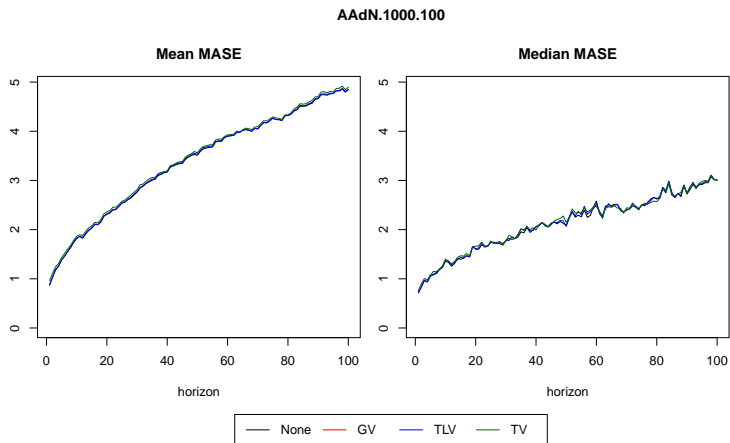


Figure: $T=1000$, $h=100$

References

- Clements, M. P., Hendry, D. F., 1996. Multi-step estimation for forecasting. *Oxford Bulletin of Economics and Statistics* 58 (4), 657–684.
- Hansen, B. E., 2010. Multi-step forecast model selection.
URL <http://www.ssc.wisc.edu/~bhansen/papers/hstep.pdf>
- Hyndman, R. J., Koehler, A. B., Ord, J. K., Snyder, R. D., 2008. *Forecasting With Exponential Smoothing: The State Space Approach*. Springer-Verlag Berlin Heidelberg.
URL <http://link.springer.com/book/10.1007%2F978-3-540-71918-2>
- Tiao, G. C., Xu, D., 1993. Robustness of maximum likelihood estimates for multi-step predictions: The exponential smoothing case. *Biometrika* 80 (3), pp. 623–641.
- Weiss, A., Andersen, A., 1984. Estimating time series models using the relevant forecast evaluation criterion. *Journal of Royal Statistical Society (A)* 147, 484–487.
- Xia, Y., Tong, H., 2011. Feature matching in time series modeling. *Statistical Science* 26 (1), 21 – 46.