Multi-step Estimators and Shrinkage Effect in Time Series Models

Ivan Svetunkov

CEBA Talk 2021, based on Svetunkov et al. (2021)

26 March 2021

Marketing Analytics and Forecasting

Lancaster University Management School

Introduction

We all have our favourite time series models (do we?)

ETS, ARIMA, GARCH, STAR, ...

We know how to estimate them.

We have our favourite loss function.

How do you estimate your model?

Introduction

Typically, it is a likelihood maximisation.

In case of Normal distribution, MLE corresponds to OLS.

OLS is equivalent to minimsation of Mean Squared Error (MSE):

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

where $e_{t+1|t} = y_{t+1} - \hat{y}_{t+1|t}$ is in-sample forecast error.

Introduction

When we are interested in h steps ahead forecast, we can use a different approach.

Produce h steps ahead forecasts in sample and minimise:

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

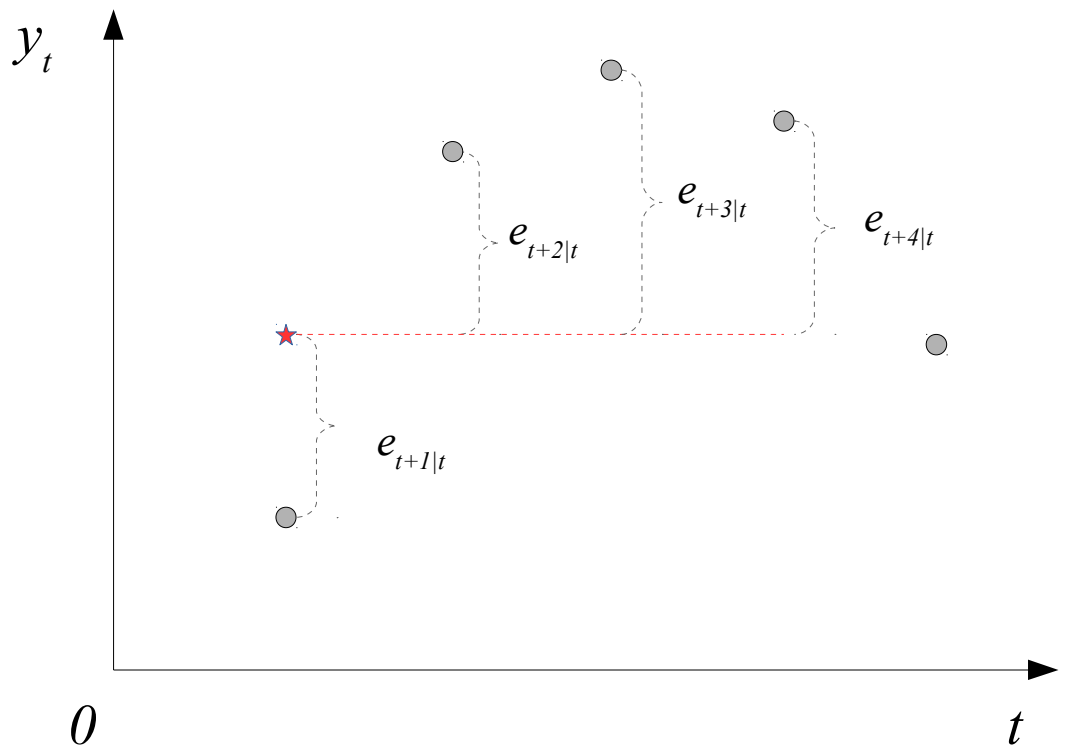
where $e_{t+h|t} = y_{t+h} - \hat{y}_{t+h|t}$.

MSE_h can be used for different h (Kang, 2003; Chevillon and Hendry, 2005; Pesaran et al., 2010)

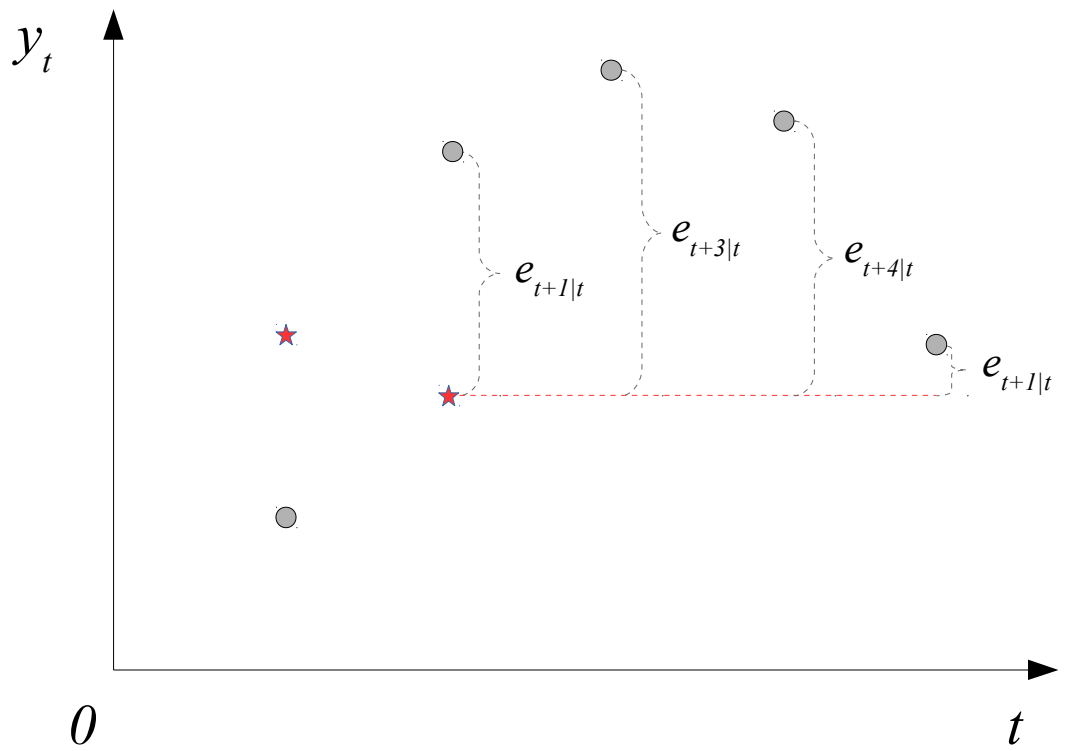
This leads to h different values of parameters for each $j = 1, \ldots, h$.

This is also called "direct multi-step estimator" (DMS, Chevillon, 2007).

Multistep losses imply producing h steps ahead forecasts from each t:



Multistep losses imply producing h steps ahead forecasts from each t:



Introduction

MSE_h produces estimates of parameters that are less efficient than with MSE.

(Tiao and Xu, 1993)

But it is more robust and increases forecast accuracy of models.

(McElroy and Wildi, 2013; Tiao and Xu, 1993; Clements and Hendry, 1996)

Although Marcellino et al. (2006) demonstrate that the forecast accuracy using MSE_h is lower than using MSE.

Introduction

Estimation process of MSE_h is complex (h models to fit).

Weiss and Andersen (1984); Xia and Tong (2011) proposed using "Trace Mean Squared Error" (TMSE):

$$\text{TMSE} = \sum_{j=1}^{h} \frac{1}{T-h} \sum_{t=1}^{T} e_{t+j|t}^2. \tag{3}$$

We only fit model once, but estimates of parameters will have some benefits from MSE_h.

Kourentzes and Trapero (2018) show that the losses in accuracy in TMSE in comparison with MSE_h are marginal.

Introduction

One more multistep estimator.

Kourentzes et al. (2019) proposed using "Mean Squared Cumulative Error" (MSCE):

$$\text{MSCE} = \frac{1}{T-h}\sum_{t=1}^{T}\left(\sum_{j=1}^{h} e_{t+j|t}\right)^2. \tag{4}$$

The motivation is to align this loss with typical inventory decisions.

We might be interested in cumulative demand over the lead time h.

Kourentzes et al. (2019) showed that it works well in context of inventory control.

Introduction

So, what's the problem?

- The results are ambiguous;
- Estimates of parameters are inefficient;
- Estimates of parameters could be unstable;
- Nobody has ever explained why multistep estimators work / don't work;

There is no comprehensible explanation of what happens with models, when multistep estimators are used.

Shrinkage effect

A shot from Rick and Morty, S3E10 “The Rickchurian Mortydate”

Shrinkage effect

MSE is proportional to variance of one-step-ahead error.

If the forecasts are unbiased then:

$$\text{MSE}_h = \hat{\sigma}_h^2,$$

$$\text{TMSE} = \sum_{j=1}^{h} \sigma_j^2,$$

$$\text{MSCE} = \sum_{j=1}^{h} \sigma_j^2 + 2 \sum_{j=2}^{h} \sum_{i=1}^{j-1} \sigma_{i,j},$$

where $\sigma_{i,j}$ is the covariance between i^{th} and j^{th} steps ahead forecast errors.

Shrinkage effect

When we use multistep estimators, we minimise respective variances of multistep errors.

So what?

I like ETS, let's see what it means for this model.

General form of pure additive ETS model (Snyder, 1985; Ord et al., 1997):

$$\begin{cases} y_t = \mathbf{w}' \mathbf{v}_{t-1} + \varepsilon_t \\ \mathbf{v}_t = \mathbf{F} \mathbf{v}_{t-1} + \mathbf{g} \varepsilon_t \end{cases}, \tag{5}$$

see Svetunkov et al. (2021) for detail.

Shrinkage effect

The h steps ahead variance of this model is:

$$\sigma_h^2 = \begin{cases} \sigma_1^2 \left(1 + \sum_{j=1}^{h-1} c_j^2 \right) & \text{when } h > 1 \\ \sigma_1^2 & \text{when } h = 1 \end{cases}, \tag{6}$$

where

$$c_j = \mathbf{w}' \mathbf{F}^{j-1} \mathbf{g}. \tag{7}$$

When we use multistep estimators, we minimise $\hat{\sigma}_1^2$ together with $\hat{c}_j$.

Shrinkage effect examples. ETS(A,N,N)

An example with local level model, ETS(A,N,N):

$$y_t = l_{t-1} + \varepsilon_t$$
$$l_t = l_{t-1} + \alpha\varepsilon_t,$$

so that $\mathbf{F} = 1$, $\mathbf{w} = 1$, $\mathbf{g} = \alpha$ and $\mathbf{v}_t = l_t$.

Shrinkage effect examples. ETS(A,N,N)

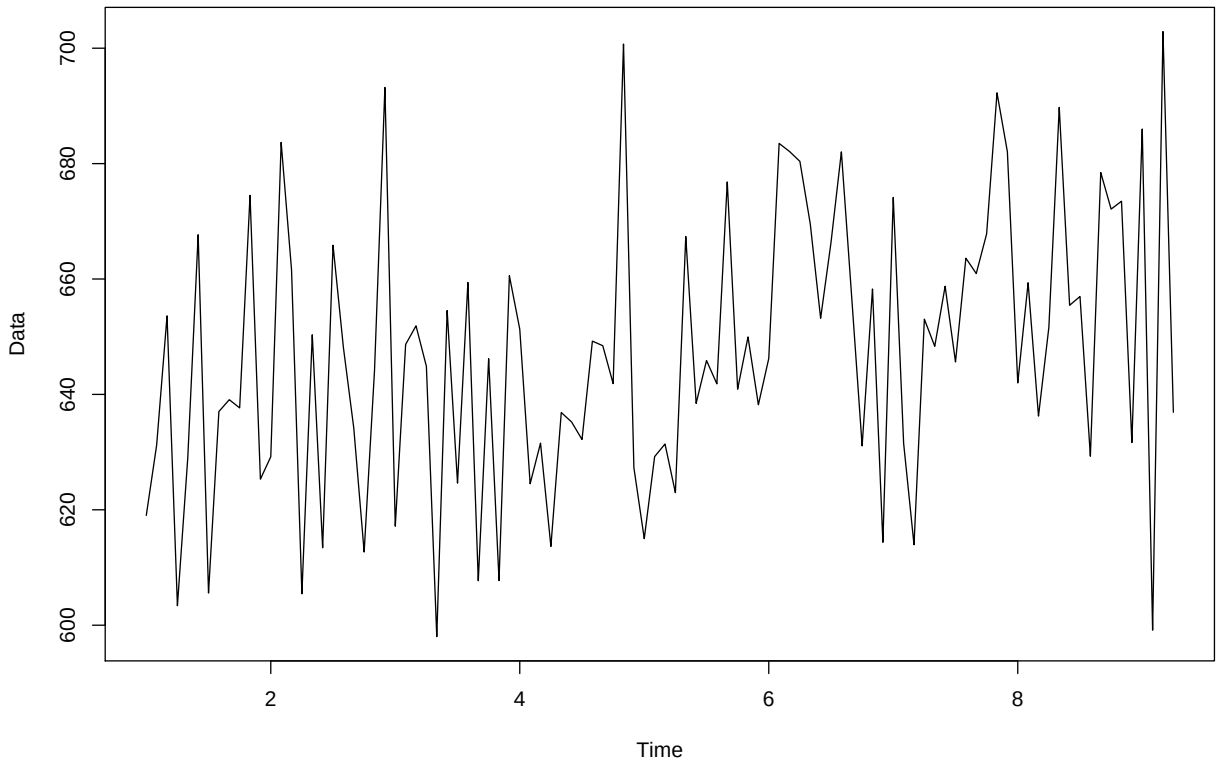


Figure: ETS(A,N,N) - local level model

Shrinkage effect examples. ETS(A,N,N)

Values for ETS(A,N,N)

$$\begin{aligned} \hat{c}_j &= \hat{\alpha} \\ \hat{\sigma}_h^2 &= \hat{\sigma}_1^2 \left(1 + (h-1)\hat{\alpha}^2\right) \end{aligned} \tag{8}$$

So for this model:

$$\text{MSE}_h = \hat{\sigma}_1^2 \left(1 + (h-1)\hat{\alpha}^2\right)$$

If we use multistep estimators, **we shrink smoothing parameter** $\hat{\alpha}$ towards zero.

ETS(A,N,N) with $\hat{\alpha} = 0$ is a global level model (deterministic).

Shrinkage effect examples. ETS(A,N,N)

In case of TMSE, the shrinkage will be weaker for the same h:

$$\text{TMSE} = \sum_{j=1}^{h} \hat{\sigma}_j^2 = \hat{\sigma}_1^2 \left(h + \sum_{j=2}^{h} \sum_{i=1}^{j-1} \hat{c}_{|j-i|}^2 \right) = \hat{\sigma}_1^2 h \left(1 + \frac{h-1}{2} \hat{\alpha}^2 \right) \tag{9}$$

In MSCE, it will be potentially stronget than in TMSE.

Shrinkage effect. Other ETS examples

Shrinkage is different for different models.

Model	$\hat{c}_j$	$\hat{\sigma}_h^2$
ETS(A,N,N)	$\hat{\alpha}$	$\hat{\sigma}_1^2\left(1+(h-1)\hat{\alpha}^2\right)$
ETS(A,A,N)	$\hat{\alpha}+\hat{\beta}j$	$\hat{\sigma}_1^2\left(1+\sum_{j=1}^{h-1}\left(\hat{\alpha}+\hat{\beta}j\right)^2\right)$
ETS(A,Ad,N)	$\hat{\alpha}+\hat{\beta}\sum_{i=1}^{j}\hat{\phi}^i$	$\hat{\sigma}_1^2\left(1+\sum_{j=1}^{h-1}\left(\hat{\alpha}+\hat{\beta}\sum_{i=1}^{j}\hat{\phi}^i\right)^2\right)$
ETS(A,N,A)	$\hat{\alpha}+\hat{\gamma}j_m$	$\hat{\sigma}_1^2\left(1+\sum_{j=1}^{h-1}(\hat{\alpha}+\hat{\gamma}j_m)^2\right)$

Table: Parameters of ETS models (Svetunkov et al., 2021). $j_m = \lfloor \frac{j-1}{m} \rfloor$, where m is seasonal periodicity.

Shrinkage effect. ETS examples

Subconclusions:

1. Multistep estimators impose shrinkage on parameters;
2. This shrinkage makes models deterministic;
3. Strength of shrinkage is proportional to forecast horizon h;
4. Shrinkage effect will differ from one model to another.

What about other models?

Shrinkage effect. ARIMA examples

ARIMA can be represented in state space form (Snyder, 1985). First expand it:

$$y_t = \sum_{j=1}^{K} \varphi_j B^j y_{t-j} + \sum_{j=1}^{K} \eta_j B^j \varepsilon_{t-j} + \beta + \varepsilon_t. \tag{10}$$

Then:

$$\mathbf{w} = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix}, \mathbf{F} = \begin{pmatrix} \varphi_1 & 1 & 0 & \dots & 0 & 1 \\ \varphi_2 & 0 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \varphi_K & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & 0 & \dots & 0 & 1 \end{pmatrix}, \mathbf{g} = \begin{pmatrix} \varphi_1 + \eta_1 \\ \varphi_2 + \eta_2 \\ \vdots \\ \varphi_K + \eta_K \\ 0 \end{pmatrix}. \tag{11}$$

Shrinkage effect. ARIMA examples

Using the same principles, we can have multiple steps ahead variance (8).

But it cannot be analysed in general... So, examples!

ARIMA(1,1,1):

$$\mathbf{w} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \mathbf{F} = \begin{pmatrix} 1+\phi_1 & 1 \\ -\phi_1 & 0 \end{pmatrix}, \mathbf{g} = \begin{pmatrix} 1+\phi_1+\theta_1 \\ -\phi_1 \end{pmatrix}.$$

Shrinkage effect. ARIMA examples

$$\hat{c}_j = 1 + \left(\hat{\phi}_1 + \hat{\theta}_1\right) \sum_{i=1}^{j} \hat{\phi}_1^{i-1}. \tag{12}$$

Substituting (12) into the formula of h steps ahead variance results in the following:

$$\hat{\sigma}_h^2 = \hat{\sigma}_1^2 \left(1 + \sum_{j=1}^{h-1} \left(1 + \left(\hat{\phi}_1 + \hat{\theta}_1\right) \sum_{i=1}^{j} \hat{\phi}_1^{i-1}\right)^2\right).$$

What does this mean?

Shrinkage effect. ARIMA examples

We deal with geometric progression, thus:

$$\left(1+\left(\hat{\phi}_1+\hat{\theta}_1\right)\sum_{i=1}^{j}\hat{\phi}_1^{i-1}\right)^2=\left(\frac{\hat{\theta}_1(1-\hat{\phi}_1^j)+1-\hat{\phi}_1^{j+1}}{1-\hat{\phi}_1}\right)^2,$$

This ratio will be minimised, when:

- $\hat{\phi}_1 \to -1$;
- $\hat{\theta}_1(1-\hat{\phi}_1^j)+1-\hat{\phi}_1^{j+1} \to 0$;

(2) implies: $\hat{\theta}_1 \to -\frac{1-\hat{\phi}_1^{j+1}}{1-\hat{\phi}_1^j}$

Together with (1), this means: $\hat{\theta}_1 \to -1$.

Shrinkage effect. ARIMA examples

In ARIMA(1,1,1) context it means that this model:

$$(1-B)(1-\hat{\phi}_1 B)y_t = (1+\hat{\theta}_1 B)e_t \tag{13}$$

will become:

$$(1-B)(1+B)y_t = (1-B)e_t,$$

or equivalently:

$$(1+B)y_t = e_t,$$

which is a non-stationary ARIMA(1,0,0).

Shrinkage effect. ARIMA examples

Other ARIMA models can be analysed in a similar way.

Subconclusions:

1. Shrinkage in ARIMA leads to degenerate models;
2. AR and MA parameters would shrink towards bounds;
3. These are typically "deterministic" models.

If you need to analyse parameters of ARIMA, be careful, when using multistep estimators.

A new estimator

Is it possible to weaken shrinkage, but still have benefits of multistep estimators?

The variance h steps ahead will be typically larger than 1 step ahead.

MSE_h will have strong shrinkage because of that.

TMSE will put emphasise on errors h steps ahead.

A possible solution – "Geometric TMSE" (GTMSE):

$$\text{GTMSE} = \sum_{j=1}^{h} \log \left(\frac{1}{T-h} \sum_{t=1}^{T-h} e_{t+j}^2 \right). \tag{14}$$

A new estimator

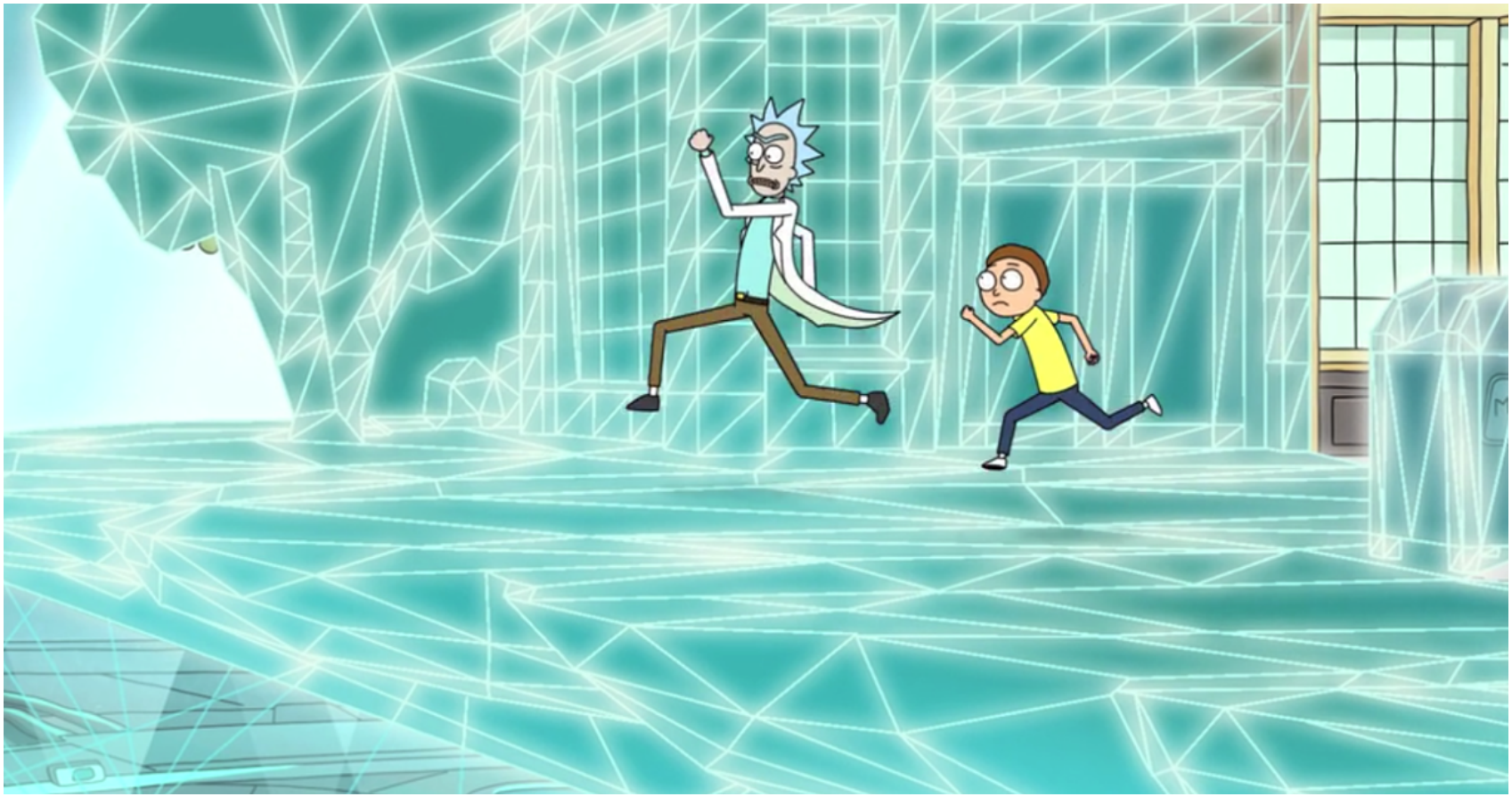
For the state space model we will have:

$$\text{GTMSE} = h \log\left(\hat{\sigma}_1^2\right) + \sum_{j=2}^{h} \log\left(1 + \sum_{i=1}^{j-1} \hat{c}_i^2\right). \qquad (15)$$

Due to logarithms, the one-step-ahead variance is balanced out with the sum of $\hat{c}_j$ elements.

There is a shrinkagein case of (14), but it is reduced in comparison with MSE_h, TMSE or MSCE.

Simulation experiment



A shot from Rick and Morty, S1E4 “M. Night Shaym-Aliens!”

Simulation experiment

Two DGPs:

1. ETS(A,N,N) with $\alpha = 0.2$;
2. ARIMA(0,1,1) with $\alpha = 0.6$;

500 time series for each

5000 observations

Get subsamples of 20, 50, 100, 200, 500, 1000 and 5000 observations.

Simulation experiment

Apply several models:

1. ETS(A,N,N);
2. ETS(A,A,N);
3. ARIMA(0,1,1);
4. ARIMA(1,1,1);

This covers:

- The model is correctly specified;
- The model is analogous to the true one (e.g., ETS(A,N,N) and ARIMA(0,1,1));
- The model is misspecified (e.g., ARIMA(1,1,1) and ARIMA(0,1,1));
- The model is wrong (e.g., ETS(A,A,N) and ARIMA(0,1,1)).

Simulation experiment

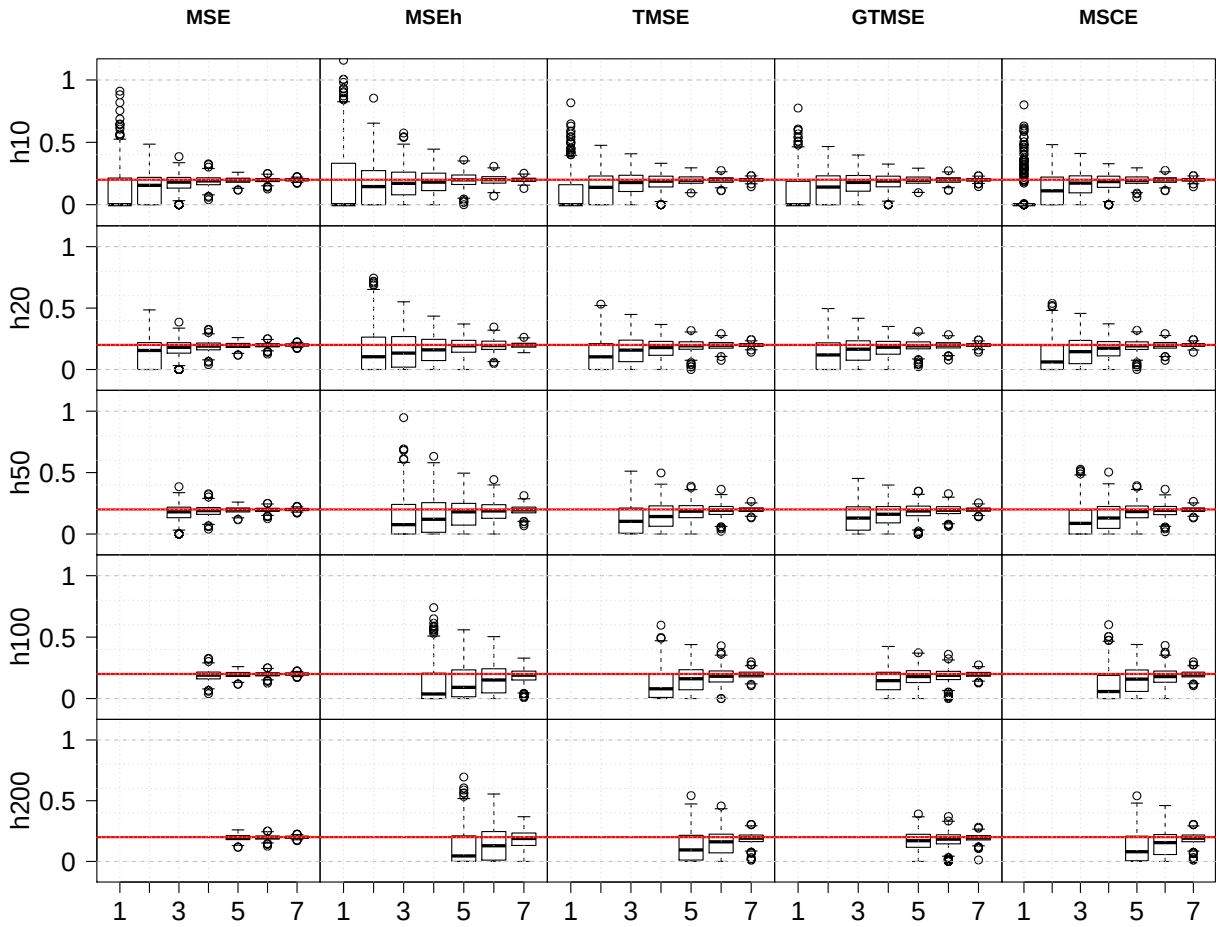
Use all the discussed estimators:

1. MSE;
2. MSE_h;
3. TMSE;
4. GTMSE;
5. MSCE.

Record estimated parameters.

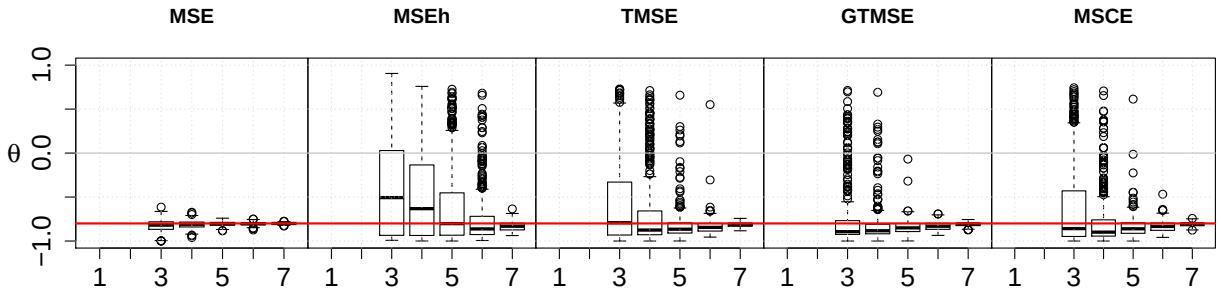
Use `es()`, `ssarima()`, `sim.es()`, `sim.ssarima()` functions from `smooth` v2.5.3 package (**?**) for R (R Core Team, 2018).

ETS(A,N,N) applied to ETS(A,N,N) data



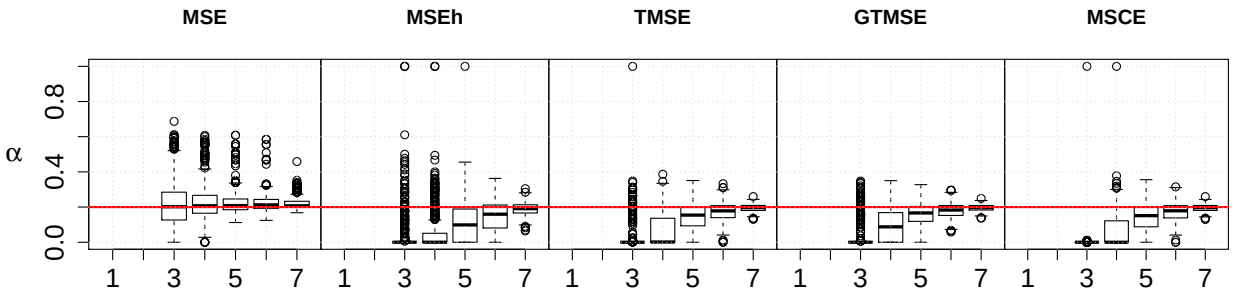
ETS(A,N,N) data, h=50

ARIMA(0,1,1)

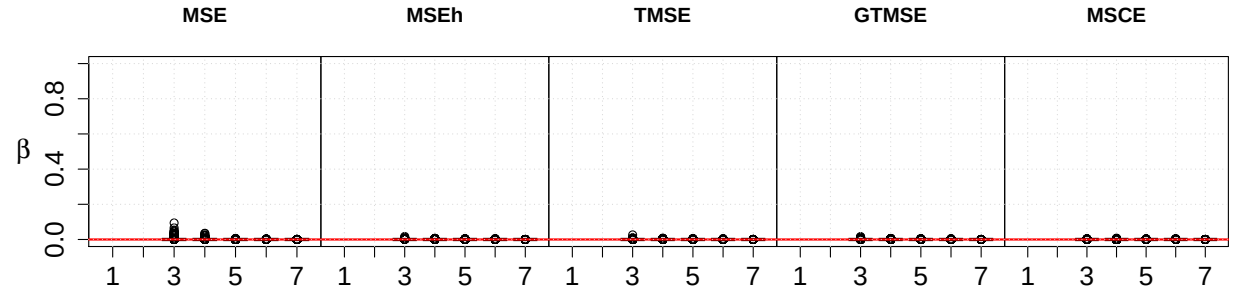


ETS(A,N,N) data, h=50

ETS(A,A,N), α

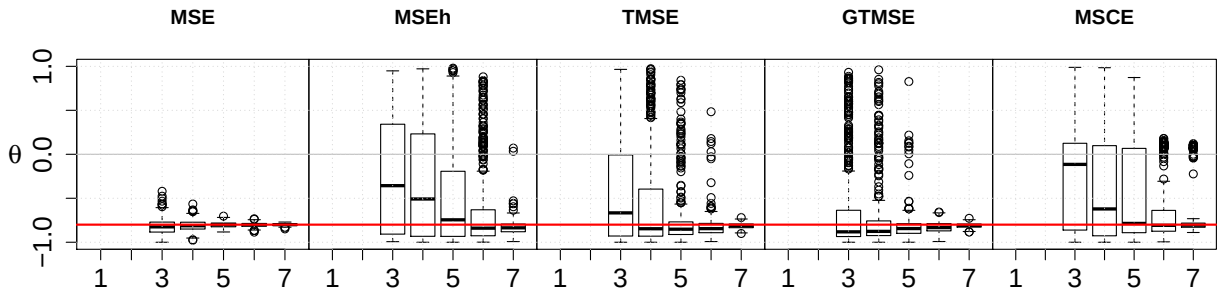


ETS(A,A,N), β

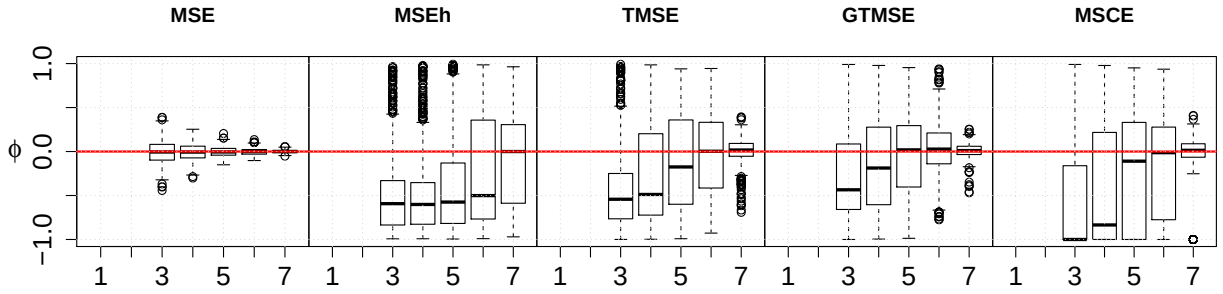


ETS(A,N,N) data, h=50

ARIMA(1,1,1), θ



ARIMA(1,1,1), ϕ



Simulation experiment results

Summarising the plots above (and those that are not shown here):

- Shrinkage is the strongest in MSCE, followed by MSE_h;
- GTMSE has the mildest shrinkage;
- It helps in case of misspecified model;
- Estimates of parameters are biased but can be more efficient than in case of MSE;
- Shrinkage is neutralise, when sample size increases.

Why bother with multistep estimators then?

An example of application

A shot from Rick and Morty, S3E9 "The ABC's of Beth"

An example of application

Hadley Centre / Climatic Research Unit data of Mean Surface Temperatures.

2016 observations, witholding 60 observations.

Apply ARIMA(1,1,2) (see Beaulieu and Killick, 2018, and discussion within) with:

1. MSE,
2. MSE_h,
3. TMSE,
4. GTMSE,
5. and MSCE.

An example of application

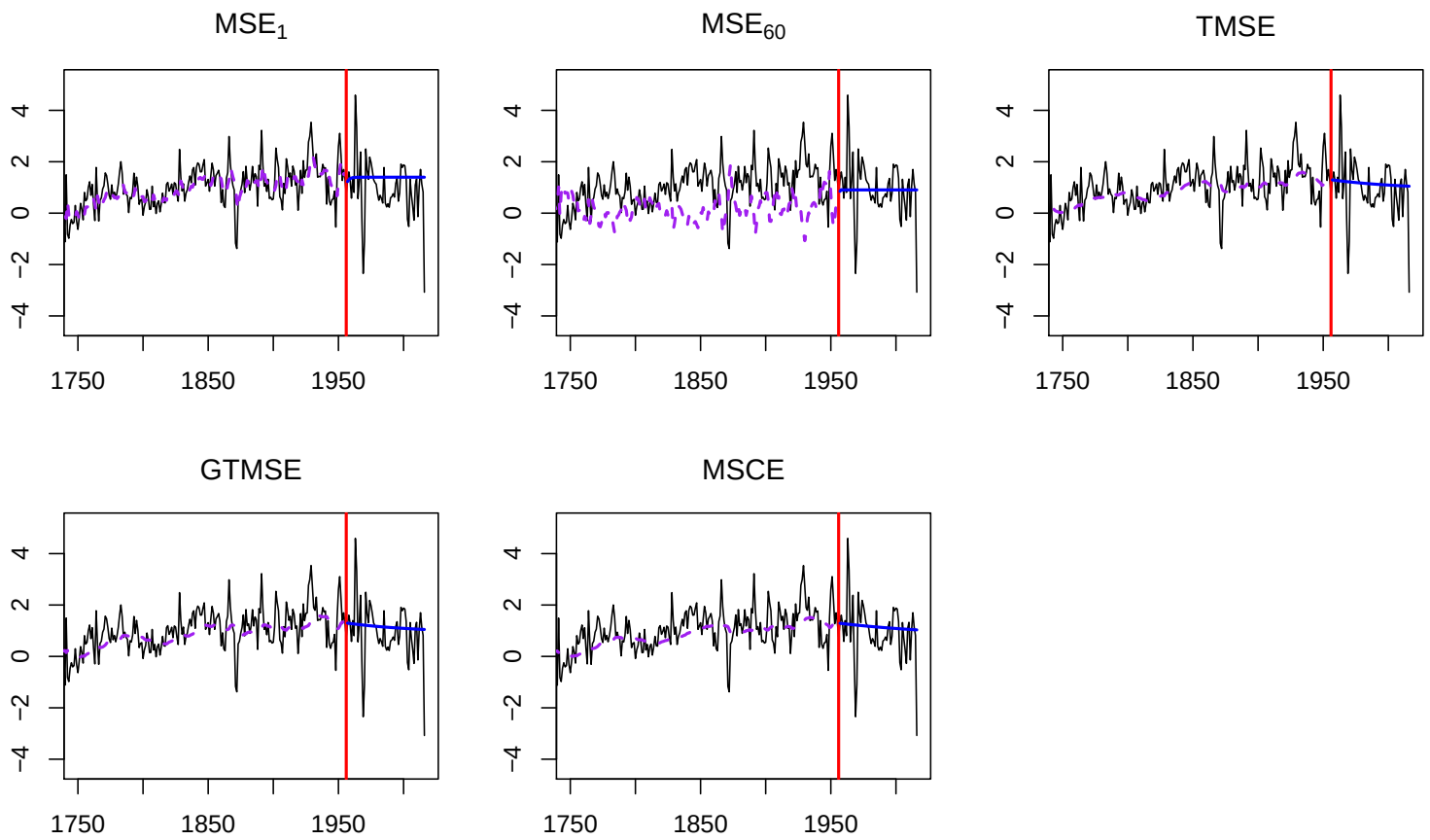
Measure performance using:

$$\text{RMSE} = \sqrt{\frac{1}{h}\sum_{j=1}^{h} e_{t+j|t}^2},$$

$$\text{MAE} = \frac{1}{h}\sum_{j=1}^{h} |e_{t+j|t}|,$$

$$\text{ME} = \frac{1}{h}\sum_{j=1}^{h} e_{t+j|t}.$$

An example of application



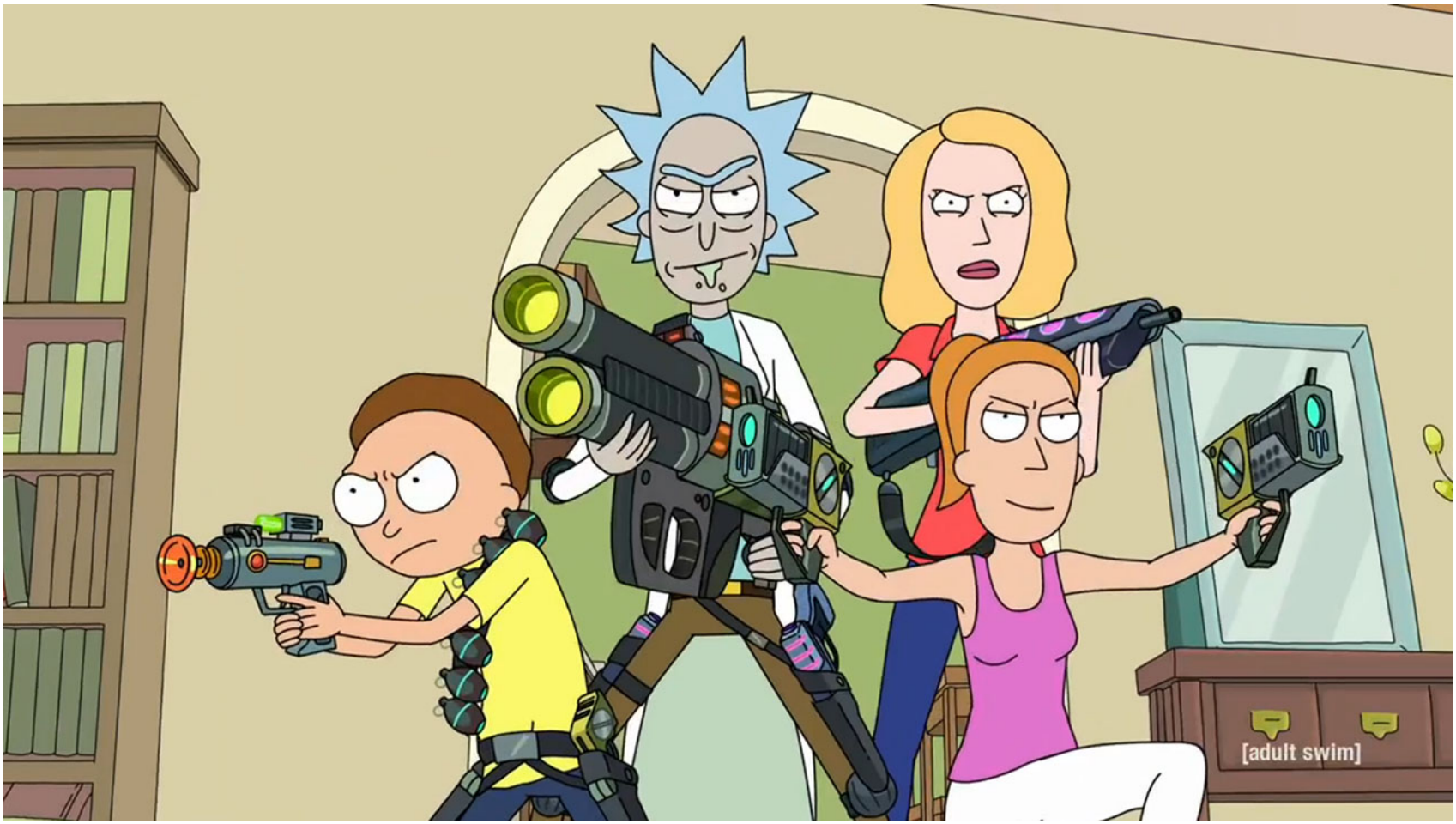
An example of application

Estimators	RMSE	MAE	ME	AR(1)	MA(1)	MA(2)
MSE_1	1.216	0.848	-26.919	0.569	-1.248	0.277
MSE_h	**1.129**	0.785	**2.342**	-0.200	-1.311	0.315
TMSE	1.131	0.768	-12.470	0.983	-1.924	0.924
GTMSE	1.130	0.768	-12.441	0.982	-1.921	0.921
MSCE	**1.129**	**0.767**	-12.277	0.989	-1.941	0.941

Table: Performance of estimators on an example of a time series.

TMSE, GTMSE and MSCE estimates of the AR(1) parameter are consistent with Hartmann et al. (2013); Cahill et al. (2015); Rahmstorf et al. (2017)

Conclusions



A shot from Rick and Morty, S2E4 “Total Rickall”

Conclusions

- Multistep estimators imply shrinkage of parameters;
- Parameters of ETS and ARIMA shrink differently;
- They both imply moving towards deterministic model;
- This gives robustness to models and help in long-term forecasting;
- ETS parameters shrink towards zero;
- The main benefits in terms of accuracy appear on big samples;

Conclusions

- If you want accuracy for a specific horizon, use MSE_h;
- You can align the estimator with a specific managerial decision;
- But! Parameters may overshrink when estimated using MSE_h and MSTFE;
- If you care for parameters, MSE is a safer choice;
- If you are unsure, GTMSE is a balanced estimator;

Thank you for your attention!

Ivan Svetunkov

i.svetunkov@lancaster.ac.uk

https://forecasting.svetunkov.ru

Twitter: @iSvetunkov

Marketing Analytics and Forecasting

Lancaster University Management School

References I

Beaulieu, C., Killick, R., 2018. Distinguishing trends and shifts from memory in climate data. Journal of Climate 31 (23), 9519–9543.

Cahill, N., Rahmstorf, S., Parnell, A. C., 2015. Change points of global temperature. Environmental Research Letters 10 (8), 084002.

Chevillon, G., sep 2007. Direct Multi-Step Estimation and Forecasting. Journal of Economic Surveys 21 (4), 746–785. URL http://doi.wiley.com/10.1111/j.1467-6419.2007.00518.x

References II

Chevillon, G., Hendry, D. F., apr 2005. Non-parametric direct multi-step estimation for forecasting economic processes. International Journal of Forecasting 21 (2), 201–218. URL http://linkinghub.elsevier.com/retrieve/pii/S016920700400069X

Clements, M. P., Hendry, D. F., may 1996. Multi-Step Estimation for Forecasting. Oxford Bulletin of Economics and Statistics 58 (4), 657–684. URL http://doi.wiley.com/10.1111/j.1468-0084.1996.mp58004005.x

References III

Hartmann, D., Klein Tank, A., Rusticucci, M., Alexander, L., Brönnimann, S., Charabi, Y., Dentener, F., Dlugokencky, E., Easterling, D., Kaplan, A., Soden, B., Thorne, P., Wild, M., Zhai, P., 2013. Observations: Atmosphere and Surface. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, book section 2, pp. 159–254.

Kang, I.-B., 2003. Multi-period forecasting using different models for different horizons: an application to U.S. economic time series data. International Journal of Forecasting 19 (3), 387–400. URL http://linkinghub.elsevier.com/retrieve/pii/S0169207002000109

References IV

Kourentzes, N., Li, D., Strauss, A. K., 2019. Unconstraining methods for revenue management systems under small demand. Journal of Revenue and Pricing Management 18 (1), 27–41.

Kourentzes, N., Trapero, J. R., 2018. On the use of multi-step cost functions for generating forecasts.

Marcellino, M., Stock, J. H., Watson, M. W., nov 2006. A comparison of direct and iterated multistep AR methods for forecasting macroeconomic time series. Journal of Econometrics 135 (1-2), 499–526.
URL http://linkinghub.elsevier.com/retrieve/pii/S030440760500165X

References V

McElroy, T., Wildi, M., jul 2013. Multi-step-ahead estimation of time series models. International Journal of Forecasting 29 (3), 378–394.
URL http://www.sciencedirect.com/science/article/pii/S0169207012001148

Ord, J. K., Koehler, A. B., Snyder, R. D., dec 1997. Estimation and Prediction for a Class of Dynamic Nonlinear Statistical Models. Journal of the American Statistical Association 92 (440), 1621–1629.
URL http://www.tandfonline.com/doi/abs/10.1080/01621459.1997.10473684

References VI

Pesaran, M. H., Pick, A., Timmermann, A., 2010. Variable Selection, Estimation and Inference for Multi-period Forecasting Problems.

R Core Team, 2018. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.
URL `https://www.R-project.org/`

Rahmstorf, S., Foster, G., Cahill, N., apr 2017. Global temperature evolution: recent trends and some pitfalls. Environmental Research Letters 12 (5), 054001.

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

References VII

Svetunkov, I., 2019. smooth: Forecasting Using State Space Models. R package version 2.5.3.
URL `https://github.com/config-i1/smooth`

Svetunkov, I., Kourentzes, N., Killick, R., 2021. Multi-step Estimators and Shrinkage Effect in Time Series Models.

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

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

References VIII

Xia, Y., Tong, H., 2011. Feature matching in time series modeling. Statistical Science 26 (1), 21–46.
URL http://www.imstat.org/sts/