Smoothing Parameter Shrinkage in Exponential Smoothing

Kandrika F. Pritularga[a,*], Ivan Svetunkov[a], Nikolaos Kourentzes[b]

[a] *Centre for Marketing Analytics and Forecasting, Department of Management Science, Lancaster University Management School, UK*
[b] *Skövde Artificial Intelligence Lab, School of Informatics, University of Skövde, Sweden.*

Abstract

Exponential smoothing is widely used in practice and has shown its efficacy and reliability in many business cases. Oftentimes, the time series in business practices are limited and short. This poses a challenge in the smoothing parameter estimation, especially under maximum likelihood framework. The model may suffer from incorrect parameters and harm the forecasting accuracy.

Motivated by the challenges in smoothing parameter estimation, we consider the use of shrinkage estimators for exponential smoothing. Regularisation can help with parameter estimation, mitigating parameter uncertainties. Building on the regularisation literature, we explore ℓ_1 and ℓ_2 regularisation. We also implement different loss functions to accommodate different shrinkage rates for each smoothing parameter. A case study of A&E admission forecasting demonstrates that regularising the smoothing parameters improve the forecast accuracy in many cases.

Keywords: forecasting, state-space model, parameter estimation, regularisation

1. Introduction

Forecasting is essential to support decisions in many organisations. For example, forecasts are shared across echelons in a supply chain to provide the information of future demand. This information is used by the members of the supply chain to make decisions, such as inventory, procurement, and production. The decisions require reliable information where the forecasts are consistent enough so that the unnecessary costs due to bullwhip effects can be avoided (Chen et al., 2000; Sadeghi, 2015). In contrast, volatile forecasts potentially lead to inducing more costs due to re-planning, schedule instability, and low service-level (Kadipasaoglu and Sridharan, 1995). Thus, supporting the management with reliable and consistent forecasts are important. Consistent forecasts also mitigate potential issues due to overfitting and unstable forecast selection (Barrow et al., 2020). We can say that the consistency in the forecasts is important to promote and ensure reliable decisions across the organisation, or the supply chain.

*Correspondance: K Pritularga, Department of Management Science, Lancaster University Management School, Lancaster, Lancashire, LA1 4YX, UK. Tel.: +44 1524 592911

Email addresses: k.pritularga@lancaster.ac.uk (Kandrika F. Pritularga), i.svetunkov@lancaster.ac.uk (Ivan Svetunkov), nikolaos@kourentzes.com (Nikolaos Kourentzes)

In order to achieve reliable and accurate forecasts, exponential smoothing (ETS) is widely used in business practices. It is robust, easy to implement, and amongst top-performing models in forecasting competitions (Fildes et al., 1998; Makridakis and Hibon, 2000; Makridakis et al., 2018). It also has been developed quite intensively (see Gardner, 2006). In ETS there are two important groups of parameters, namely smoothing parameters and initial values. The smoothing parameters control how new information affecting the states of the model while the initial values act as proxies of prior information before collected observations. The conventional methodology in ETS utilises the single source of error (SSOE) framework (Snyder, 1985). Hyndman et al. (2002, 2008) automate the methodology by employing the maximum likelihood estimation (MLE) for parameters in order to select the most appropriate model to the data. However, in business practices time series are typically short. Consequently, the estimators are prone to inefficiency, especially when the observations are fewer than the number of parameters. We discuss this issue in Section 2.

We develop an estimation procedure in ETS models which produce reliable and consistent forecasts. We regularise the smoothing parameters in ETS to control the effect of new information on the states of the ETS model. By shrinking the smoothing parameters, we essentially reduce the effect of new information on the states and accentuate the long memory processes. On the other hand, a concept of regularisation is widely applied in regression, such as LASSO and ridge regression (Tibshirani, 1996; Hoerl and Kennard, 2000). Both reduce the effect of the explanatory variables on the target variable, depending on the penalty function. Although we aim to achieve similar effects from shrinking the smoothing parameters, we argue that regularisation in ETS is conceptually different from the one in regression. Suppose that we shrink the smoothing parameters to zero, regularising the smoothing parameter means that we reduce the effect of decreasing weights on the long memory processes. This is contrary to shrinking regression parameter estimates to zero, where the shrinkage eliminates the contribution of a specific explanatory variable to the target variable. Special cases such as the original Theta method can be seen as having some parameters set to zero, resulting in deterministic states (Assimakopoulos and Nikolopoulos, 2000; Hyndman and Billah, 2003).

We aim (a) to demonstrate the implementation of regularisation in ETS, (b) to explain the mechanics of regularisation in ETS, and (c) to explain the effect of shrinkage on the parameters and on the predictive accuracy. Thus, we conducted an experiment with A&E admission data. Given a pre-determined model provided by an automatic selection via information criterion,

we find that the ETS with regularisation gains forecast accuracy improvement. Models with weighted regularisation outperform the other models in most cases. The choice of the penalty function does not have statistical differences in improving the forecast accuracy. We also find that the regularisation is able to shrink the level smoothing parameter, but it does not shrink the seasonal smoothing parameter much. We also note that the initial values do not change much due to smoothing parameter shrinkage. Lastly, small shrinkage parameters are sufficient to improve the forecast accuracy

In Section 2, we discuss the theory behind exponential smoothing with regularisation. Section 3 describes the experimental settings and discusses the findings. We apply the proposed model to the real-life time series where the data generating process is unknown in Section 4. In Section 5, we discuss and conclude our proposed model with its findings.

2. Exponential Smoothing with Parameter Shrinkage

ETS reconstructs the time series from its unobserved components, such as its level, trend, and seasonality. Hyndman et al. (2008) build a taxonomy for naming ETS models, such as ETS(ANN), where ANN means an additive error (A) with no trend and seasonality (N). Multiplicative components are denoted as M, and Ad or Md denote additive or multiplicative damped trend. Let y_t be an observed time series, at period t, assuming that the time series is constructed from different unobserved components ($\boldsymbol{x}_t$), such as trend and seasonality. The states here usually identify the data characteristics. We can write a general exponential smoothing model according to Hyndman et al. (2008),

$$y_t = \boldsymbol{w}^\top \boldsymbol{x}_{t-1} + \varepsilon_t \tag{1}$$

$$\boldsymbol{x}_t = \boldsymbol{F}\boldsymbol{x}_{t-1} + \boldsymbol{g}\varepsilon_t, \tag{2}$$

where ε_t is the error term which has zero mean and variance of σ^2. $\boldsymbol{w}$ is the measurement vector, $\boldsymbol{F}$ is the transition matrix, and $\boldsymbol{g}$ is the persistence vector, which includes smoothing parameters for all states in the model. $\boldsymbol{x}_t$ may contain a level (l_t), a trend (b_t), and a seasonality (s_t). Following the states, $\boldsymbol{g}$ also contains the level (α), the trend (β), and the seasonal smoothing parameter (γ). In a damped trend model, the dampening parameter, or ϕ, is added into the transition matrix and the measurement vector. Conventionally, we call (1) and (2) as the measurement and the transition equation, respectively. Since only a single error term affects

both equations, we call it a single source of error state-space framework (SSOE). We refer to Hyndman et al. (2008) for more details.

Suppose we produce one-step ahead forecast from ETS model, denoted as $\mu_{t+1|t}$. We can measure the one-step ahead forecast error, $e_{t+1|t} = y_{t+1} - \mu_{t+1|t}$, where $\mu_{t+1|t} = \boldsymbol{w}^\top \boldsymbol{x}_t$. In general, the parameters are unknown but we assume that $\boldsymbol{w}^\top$ and $\boldsymbol{F}$ are known, depending on the structure of the model, i.e., whether there is trend or seasonality. In a dampening trend model, ϕ is inserted into $\boldsymbol{w}^\top$ and $\boldsymbol{F}$ and we still need to estimate it. Apart from that, we need to estimate the elements of $\boldsymbol{g}$ to determine the effect of new information on the states.

The current methodology employs MLE for $\boldsymbol{\theta}$, $\boldsymbol{x}_0$, and σ^2 (Hyndman et al., 2008), where $\boldsymbol{\theta} = \{\boldsymbol{g}, \phi\} = \{\alpha, \beta, \gamma, \phi\}$. Not only such parameters, but also a variance is estimated from the MLE. This allows us producing prediction intervals. Apart from that, it also enables us to select the best approximated model via an information criterion. According to Hyndman et al. (2008), the likelihood function is shown as,

$$\mathcal{L}^*(\boldsymbol{\theta}, \boldsymbol{x}_0 | h(\cdot), \mathcal{Y}_t) = n \log \left(\sum_{t=1}^{T} \hat{e}_{t+1|t}^2 \right) + 2 \sum_{t=1}^{T} \log |r(\boldsymbol{x}_t)|,$$

where $r(\cdot)$ is a scalar function which determines whether the error is additive or multiplicative. T is the number of observations, and $\boldsymbol{x}_0$ is the vector of the initial states. $h(\cdot)$ denotes a pre-determined model structure, i.e., determined by a model selection, and $\mathcal{Y}_t$ denotes the available information at time t. Hyndman et al. (2008) show that the likelihood function when maximised is reduced to the augmented sum of squared errors criterion, shown as,

$$S(\boldsymbol{\theta}, \boldsymbol{x}_0 | h(\cdot), \mathcal{Y}_t) = \left| \prod_{t=1}^{T} r(\boldsymbol{x}_t) \right|^{2/n} \sum_{t=1}^{T} e_{t+1|t}^2. \tag{3}$$

In the case of additive errors $|r(\boldsymbol{x}_t)| = 1$, Eq 3 reduces to the sum squared of errors. Under MLE, consistent and efficient estimators can be achieved when $T \to \infty$. However, in business practices large observations are rarely obtained due to product life cycle, product discontinuity, or low-quality data management. This may harm the efficiency of the estimators, and eventually worsen the predictive accuracy. Moreover, MLE assumes that the probability model is known and correct. Since the data generating process is generally unknown it is more likely that we suffer from model misspecification and it may harm the forecast accuracy as well.

Hyndman et al. (2008) note that we can utilise (3) because it is computationally less

expensive than MLE and it avoids variance estimation instability, which sometimes occurs in MLE. However, due to limited sample sizes the estimators may reach the upper bound of the parameter spaces. This potentially results in an overfit model, where the model fits the in-sample very well but does not behave well in the out-of-sample. Following this problem, we modify the loss function via minimising the root mean squared one-step ahead in-sample forecast error (RMSE) with regularisation. The conventional loss function, which is widely applied in regression problems, is shown as,

$$\text{RMSE}(\boldsymbol{\theta}, \boldsymbol{x_0}|h(\cdot), \mathcal{Y}_t) + \lambda^* p(\boldsymbol{\theta}), \tag{4}$$

where λ^* is the shrinkage parameter and $\lambda^* \geq 0$. It is possible that the upper bound of λ^* is infinity. Hence, it could be difficult to find the optimal value of λ^*. We aim to modify (4) so that the shrinkage parameter has a finite upper bound. Thus, the penalised loss function is shown as,

$$(1-\lambda)\text{RMSE}(\boldsymbol{\theta}, \boldsymbol{x_0}|h(\cdot), \mathcal{Y}_t) + \lambda p(\boldsymbol{\theta}), \tag{5}$$

where $\lambda \in [0, 1]$, $\lambda \in \boldsymbol{\Lambda}$ and is the shrinkage parameter which controls the shrinkage rate of each smoothing parameter while $\boldsymbol{\Lambda}$ is the parameter space of λ. $p(\boldsymbol{\theta})$ is the penalty function, which depends on the type of penalisation. We can write the penalty function with ℓ_1 regularisation as, $p(\boldsymbol{\theta}) = |\alpha|+|\beta|+|\gamma|+|1-\phi|$ and for the ℓ_2 regularisation, $p(\boldsymbol{\theta}) = (\alpha)^2+(\beta)^2+(\gamma)^2+(1-\phi)^2$. In the case of ϕ, we shrink it towards 1 instead of 0 because we need to preserve the trend. Suppose that we shrink ϕ to zero, then the trend vanishes. In particular, (5) demonstrates a trade-off between model fitness and model responsiveness. When λ is close to 0, the estimator puts more weights on the fitness. On the other hand, when λ is close to 1, the estimator puts more weights on model responsiveness.

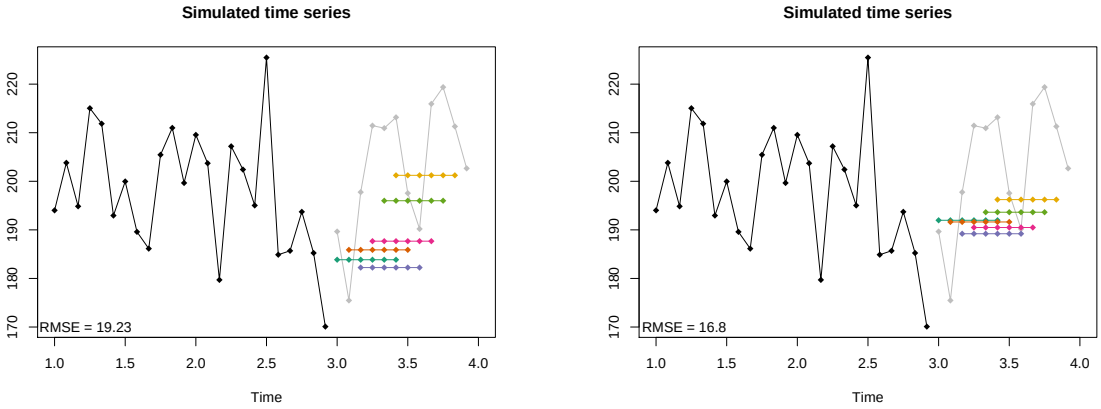
We can demonstrate a hypothetical effect of lowering the smoothing parameters on the states and consequently on the forecasts. In this simple illustration, we employ ETS(ANN). For such model, $\boldsymbol{\theta} = \{\alpha\}$ and $\boldsymbol{x}_t = \{\ell_t\}$, and the model is shown as,

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

Suppose that y_t has the data generating process of ETS(ANN), wih $\alpha = 0.4$ and $\ell_0 = 200$. It

is a monthly time series and has 36 observations. We reserve the first 24 observations as the in-sample and apply two ETS(ANN) models: one with $\hat{\alpha} = 0.35$ and the other one with $\hat{\alpha} = 0.15$ for 6 origins. So that, we produce 1-6 step ahead forecasts 6 times.

Figure 1 demonstrates the difference between two values of α. When $\hat{\alpha}$ is close to the true α, i.e., $\hat{\alpha} = 0.35$, the forecasts are volatile. However, when $\hat{\alpha}$ is reduced to a small size, e.g., 0.15, the forecasts are relatively stable than the former. We can explain this phenomenon by looking at the effect of $\hat{\alpha}$ on the new information (ε_t). A small size of $\hat{\alpha}$ reduces the effect of new information on the state significantly, i.e., $\hat{\alpha} \to 0$, then $\hat{\alpha}\varepsilon_t \to 0$. Consequently, $l_t \approx l_0$. This induces not only more stable forecasts but also more reliable forecasts. We expect that similar effects happen in the forecasts due to regularising $\boldsymbol{\theta}$.



(a) $\hat{\alpha} = 0.35$ (b) $\hat{\alpha} = 0.15$

Figure 1: Examples of two forecasts for 6 origins from different sizes of $\hat{\alpha}$

2.1. Weighted Regularisation

(5) assumes that we shrink the smoothing parameters at the same rate. However, this may not be the case. For example, we have a seasonal time series with a weak trend and we model it using ETS with trend and seasonality, or ETS(AAA), with regularisation. As we know that the time series is seasonal but does not have a strong trend, we would like to shrink the trend smoothing parameter more than the seasonal one. Thus, we need to adjust (5) in order to accommodate this. We propose to add weights for each smoothing parameter of the penalty function. This potentially mitigates the uncertainty from the structure of the model. The loss

function with weighted regularisation is shown as,

$$\{\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{x}}_0\}_{\ell_1} = \arg\min(1-\lambda)\text{RMSE}(\boldsymbol{\theta}, \boldsymbol{x_0}|h(\cdot), \mathcal{I}_t) + \lambda \sum_{j=1}^{k} \omega_j |\theta_j|, \tag{6}$$

$$\{\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{x}}_0\}_{\ell_2} = \arg\min(1-\lambda)\text{RMSE}(\boldsymbol{\theta}, \boldsymbol{x_0}|h(\cdot), \mathcal{I}_t) + \lambda \sum_{j=1}^{k} \omega_j \theta_j^2, \tag{7}$$

where ω_j is a weight of each smoothing parameter, $\omega_j \in [0, 1] \in \boldsymbol{\Omega}$, $\sum_{j=1}^{k} \omega_j = 1$, and $\boldsymbol{\Omega}$ is the parameter space of ω_j. (6) and (7) are the loss function for different penalty functions, namely ℓ_1 and ℓ_2 regularisation, respectively. A similar penalty function is used on adaptive regularisation (Zou, 2006), where each parameters in the penalty function may have different shrinkage parameters whether in groups or individually.

2.2. On choosing λ

Having discussed the effect of regularised smoothing parameters on the states, we need to find good estimates for the shrinkage parameters. Grid search is widely used for shrinkage parameter optimisation (Hoerl and Kennard, 2000; Bergstra et al., 2012). It is a relatively simple operation, but computationally expensive operation.

We propose to implement a derivative-free shrinkage parameter optimisation. We aim to minimise the cross-validated root mean squared one-step ahead out-of-sample forecast error, shown as,

$$\{\hat{\lambda}, \hat{\boldsymbol{\omega}}\} = \arg\min_{\lambda, \boldsymbol{\omega}} \frac{1}{K} \sum_{l=1}^{K} \text{RMSE}(\lambda, \boldsymbol{\omega}, \hat{\boldsymbol{\theta}}_l, \hat{\boldsymbol{x}}_{l,0}, y_{1:l}), \tag{8}$$

where K is the number of origins and $y_{1:l}$ is the in-sample time series starting from $t = 1$ to $t = l$. $\hat{\boldsymbol{\theta}}_l$ and $\hat{\boldsymbol{x}}_{l,0}$ are the estimated parameters and initial values for origin l. Essentially, we propose a two-step estimation. First, we estimate the smoothing parameters given the $\hat{\lambda}$ and $\hat{\boldsymbol{\omega}}$. Then, we estimate $\hat{\lambda}$ and $\hat{\boldsymbol{\omega}}$ minimising (8). We use a derivative-free optimisation because finding the gradient from the cross-validation is not a trivial task.

3. NHS A&E Admission Case Study

3.1. Experimental Design

In this case study, we apply our proposed model to NHS A&E admission data of a hospital in the northeast of England. The data contains number of incidents in a day for different classifications, such as age (under 3 years old, between 4-16 years old, between 17-74 years old,

and more than 75 years old), sex (male, female), and type of disposal (admitted, discharged, referred to clinics, transferred, died, referred to health care professionals, left, and others). In total, we have 135 time series. The data itself spans from January 2009 to October 2019 and on a monthly level. In order to see the effect of sample sizes, we employ two different samples sizes. First, the time series is short, which has 36 months (2009-2012). For a longer sample size, we have 108 months (2009-2018). For each sample size, we apply rolling-origin with 5 origins to produce 1 to 12-steps ahead forecast, with the same model structure.

We aim to demonstrate the effect of regularisation by comparing a model with and without it. Before applying regularisation, we need to determine the structure of the model. We use an automatic selection based on the corrected Akaike Information Criteria (AICc), provided by adam() function in smooth package Svetunkov (2021), that implements exponential smoothing. After that, we use the model as the benchmark and then apply regularisation to it. In essence, we assume that the structure of the model is defined correctly by AICc and we take care of the parameter estimation issue with regularisation. Hence, the benchmark model is a model selected from AICc without regularisation and the proposed model is the same model structure with regularisation. In this design, we have four scenarios, which are a combination of ℓ_1, ℓ_2, unweighted, and weighted regularisation. Table 1 summarises the notations for the scenarios.

Regularisation	ℓ_1	ℓ_2
Unweighted	L1:UR	L2:UR
Weighted	L1:WR	L2:WR

Table 1: A summary of scenarios, a combination between different types of regularisation. L1 and L2 denote ℓ_1 and ell_2 regularisation. UR and WR denote the unweighted and weighted regularisation.

In order to evaluate the forecasting performance of each scenario we use three measures, namely RMSE, mean absolute error (MAE), and mean error (ME). These measures are calculated as follows,

$$\text{RMSE} = \sqrt{\frac{1}{h}\sum_{i=1}^{h}(y_{t+i} - \hat{y}_{t+i|t})^2},$$

$$\text{MAE} = \frac{1}{h}\sum_{i=1}^{h}|y_{t+i} - \hat{y}_{t+i|t}|,$$

$$\text{ME} = \frac{1}{h}\sum_{i=1}^{h}(y_{t+i} - \hat{y}_{t+i|t}).$$

Since we are interested in forecast accuracy improvement, we take a percentage difference between the proposed model and the benchmark model. For example, in the case of RMSE,

$$\text{dRMSE} = 100 \times \frac{(\text{RMSE}_a - \text{RMSE}_b)}{\text{RMSE}_b}$$

where RMSE_a and RMSE_b denote the RMSE from the proposed and the benchmark model, respectively, and the benchmark model here is the model without regularisation. As for ME, we take an absolute value of each ME before taking the percentage difference to eliminate the negative sign. However, we should acknowledge that the percentage difference of absolute MEs results in the magnitude of the bias without knowing whether it is biased negatively or positively.

3.1.1. Choosing the shrinkage parameters

In implementing our approach, we use Nelder-Mead algorithm, implemented in nloptr package in R (Johnson, 2021). For the algorithm, the maximum iteration is 1000 and the stopping criterion is 1e-08. For the unweighted regularisation, we need to find λ only. The optimal λ is then applied to all smoothing parameters. On the other hand, the weighted regularisation requires four parameters to estimate, namely λ, ω_α, ω_β, ω_γ, ω_ϕ, where ω_α, ω_β, ω_γ, ω_ϕ are the weight of level, trend, seasonal smoothing parameter, and dampening parameter. We aim to initialise the optimisation with uninformative initial values. In this experiment, we use the initialisation of 0.1 for λ. As for ω, we use an equal weights. For example, for ETS(AAN) we need to estimate ω_α and ω_β hence we only need $\omega_\alpha = 0.5$ while $\omega_\phi = 1 - \omega_\alpha$.

3.2. Findings

Table 2 presents the average forecast accuracy improvement due to parameter regularisation in the percentage difference of RMSEs from 135 time series. Negative numbers show the percentage improvement from the benchmark, and vice versa. Note here that all models are ETS(ANA) for both sample sizes, i.e., 36 months and 108 months, which are selected from AICc.

We can see that across different scenarios the proposed estimation procedure outperforms the benchmark model by 2-3%, on average. We also observe that the weighted regularisation outperforms the unweighted regularisation, even though the latter shows improvements in some cases. Looking at the small sample size, the regularised models outperform the benchmark models in all cases in comparison with the benchmark. However, when we have long time series, the models with the unweighted regularisation produce less accurate forecasts than the benchmark

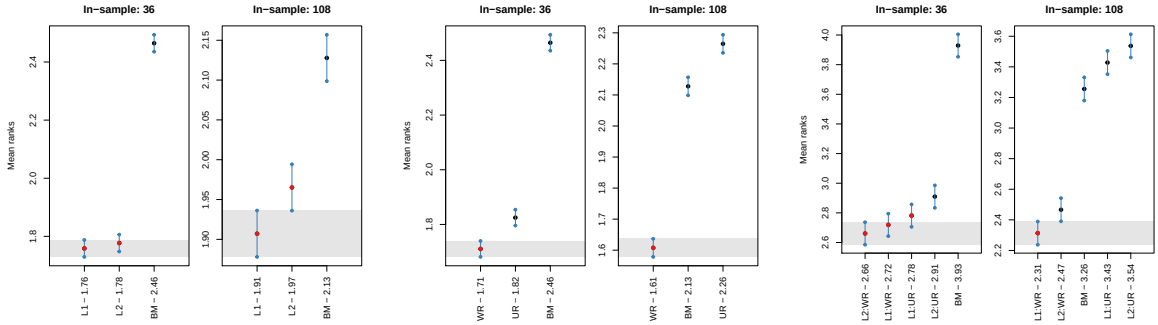
and the model with the weighted regularisation outperforms the benchmark. Nonetheless, the model with regularisation induces biases as shown in AbsME.

Panel A: accuracy measures for 1 step ahead forecast					
		L1		L2	
Accuracy	Sample	UR	WR	UR	WR
dRMSE	36	-0.06	-2.06	-1.97	**-2.16**
	108	3.78	**-3.07**	3.97	-2.49
dMAE	36	-0.06	-2.06	-1.97	**-2.16**
	108	3.78	**-3.07**	3.97	-2.49
dAbsME	36	95.31	87.66	**11.27**	23.81
	108	30.64	**22.59**	22.93	27.38
Panel B: accuracy measures for 1-12 step ahead forecast					
		L1		L2	
Accuracy	Sample	UR	WR	UR	WR
dRMSE	36	-2.19	-2.33	-2.30	**-2.35**
	108	0.68	**-1.28**	0.68	-1.31
dMAE	36	-2.58	-3.06	-2.60	**-2.99**
	108	1.16	**-1.84**	1.02	-1.17
dAbsME	36	6.55	**-1.07**	-3.16	-3.87
	108	104.26	103.63	134.83	**84.67**

Table 2: A summary of the percentage difference for different accuracy measures, loss functions, and forecast horizons. Negative bold numbers show the best performance, except for dAbsME where the bold number show the least biased forecasts.

In addition to Table 2, we conducted Nemenyi-Friedman non-parametric test to see whether some types of regularisation perform better than the others and the benchmark, statistically (see Figure 2). We use the RMSE for all forecast horizons and time series in a hierarchy. Note here that if the intervals in Figure 2 intersect, then there is no statistical difference. Panel 2a present the effect of the penalty function on the accuracy improvement. We can see that there is no significant difference between ℓ_1 and ℓ_2 regularisation and both outperform the benchmark, for the small and the large sample size. Figure 2b present the effect and the unweighted and the weighted regularisation. For both sample sizes, we can see that the weighted regularisation improves the forecast accuracy significantly. The last panel depict the effect of a combination between the type of regularisation and the penalty function. For the small sample size, L2:WR, L1:WR, and L1:UR perform similar to each other and outperforms the benchmark, statistically. However, when the sample size is larger, only L1:WR outperforms the others statistically. In

the case when the unweighted regularisation performs similar to the weighted one statistically, it is preferable to choose the unweighted one due to less expensive computation.

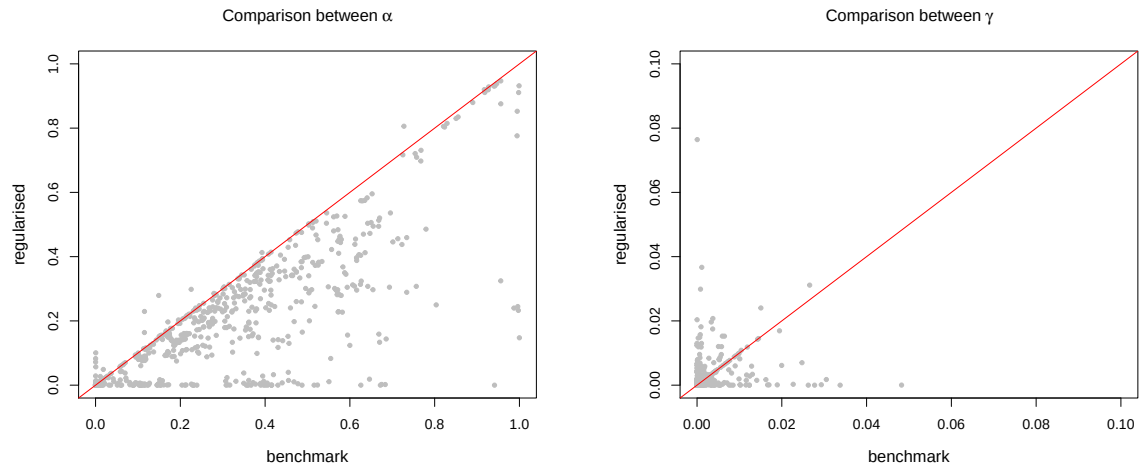


(a) Penalty function (b) Type of regularisation (c) Combination

Figure 2: A non-parametric Nemenyi-Friedman test of RMSE for different types of loss function. In Panel (a), we distinguish the effect of ℓ_1 and ℓ_2 regularisation. In Panel (b) we distinguish the effect of weighted and unweighted regularisation, for different sample sizes and forecast horizons.

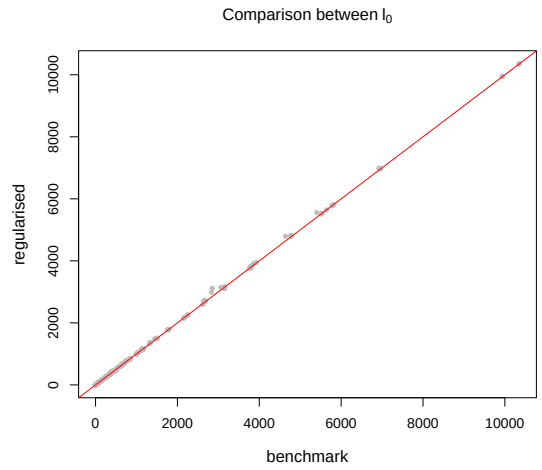
Since L1:WR performs well in both sample sizes, we focus on comparing the parameters of the benchmark and the regularised model in L1:WR. Figure 3 shows the comparisons. Panels 3a-3c are the scatterplots between the parameters of the benchmark model (x-axis) and the proposed model (y-axis). The red diagonal lines depict the equality between both parameters. Anything below the red line shows that the parameters are shrunk due to the regularisation, vice versa. Panel 3d shows the boxplots of the seasonal initial values, from the benchmark and the proposed model.

We can see from Panel 3a that the proposed estimation procedure is able to shrink the level smoothing parameter most of the time. However, looking at Panel 3b we can see that $\hat{\gamma}$ from the proposed model does not shrink as it intended to. Interestingly enough, the regularisation does not affect the estimation of the initial values. Panel 3c shows that the grey dots are mostly aligned to the red line, which means that the level initial values are not shrunk in the response of the level smoothing parameter shrinkage. From Panel 3d we do not see differences in the distribution of each seasonal initial value. These show that the regularisation works as it is intended to, especially for the level smoothing parameter, and it does not affect the estimation of the initial values.

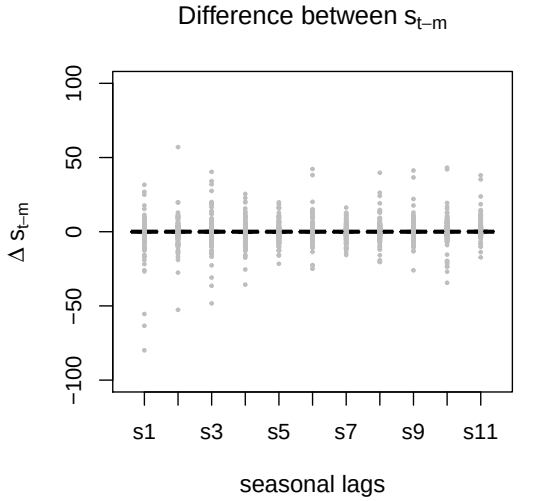


(a) $\hat{\alpha}$

(b) $\hat{\gamma}$



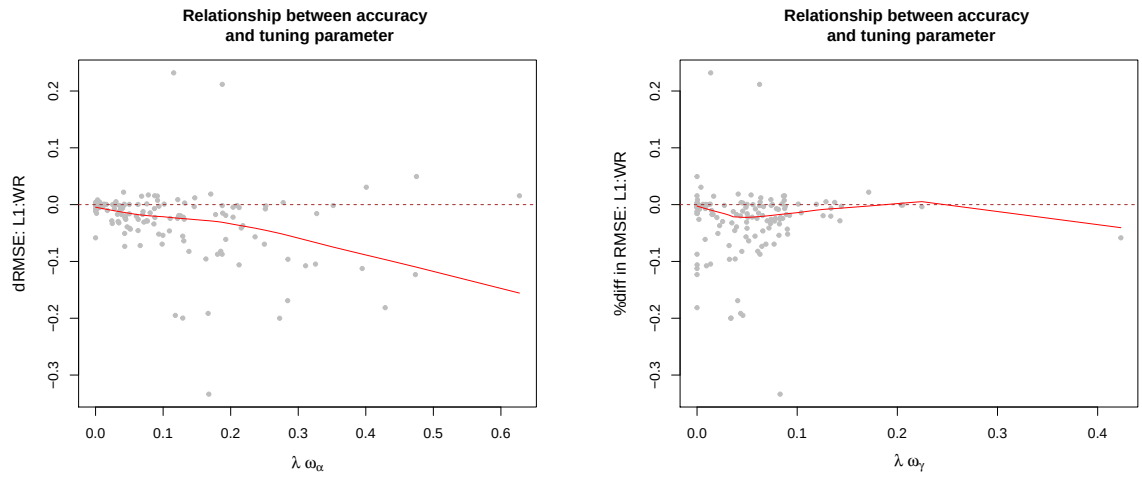
(c) $\hat{l}_0$



(d) $\Delta\hat{s}_{t-m}$

Figure 3: Parameter comparisons between the regularised and the benchmark models for L1 weighted regularisation, with small sample size. Panel (d) presents the difference between the seasonal initial values of the proposed model and the benchmark. m is the seasonal lag

Apart from the effect of shrinkage on the parameters, we are interested in the effect of the shrinkage parameters on the accuracy improvement. Figure 4 represents the scatterplots between the percentage difference of RMSE (dRMSE) for all forecast horizons and the shrinkage parameters. In Panel (a) we can see that forecast accuracy improvement is concentrated when $\lambda\omega_\alpha$ is less than 0.2. On the other hand, $\lambda\omega_\gamma$ is concentrated between 0 and 0.1 and many of them show the improvement of forecast accuracy. Nonetheless, with relatively small $\lambda\boldsymbol{\omega}$ the forecasts from the proposed model generally improve the forecast accuracy.



(a) $\lambda\omega_\alpha$ (b) $\lambda\omega_\gamma$

Figure 4: Relationships between dRMSE and shrinkage parameters. Anything lower than the red line denotes better performances than the benchmark.

4. Questions arising from the preliminary results

- Distinguishing the effect of different loss functions on the forecast accuracy. It is obvious from the current experiment that the weighted regularisation outperforms the unweighthed one. However, it is not obvious how the choice of the penalty norm affects the forecasting performance. We need to find a way to distinguish the effect of scenarios on the forecast accuracy measure.

- Discussion on the effect of $\hat{\alpha}$ with shrinkage on the other smoothing parameters. Hyndman et al. (2008, p. 46) note that the traditional parameter space of the trend and the seasonal smoothing parameter depends on the level smoothing parameter, i.e., $0 < \alpha < 1$, $0 < \beta < \alpha$, $0 < \gamma < 1 - \alpha$. Shrinking α will potentially shrink the parameter space of β, but might enlarge the parameter space of γ.

References

Assimakopoulos, V., Nikolopoulos, K., 2000. The theta model: a decomposition approach to forecasting. International journal of forecasting 16, 521–530. doi:10.1016/S0169-2070(00)00066-2.

Barrow, D., Kourentzes, N., Sandberg, R., Niklewski, J., 2020. Automatic robust estimation for exponential smoothing: perspectives from statistics and machine learning. Expert Systems with Applications 160.

Bergstra, J., Ca, J.B., Ca, Y.B., 2012. Random search for hyper-parameter optimization yoshua bengio. Journal of Machine Learning Research 13, 281–305. URL: http://scikit-learn.sourceforge.net.

Chen, F., Ryan, J.K., Simchi-Levi, D., 2000. The impact of exponential smoothing forecasts on the bullwhip effect. Naval Research Logistics 47, 269–286. doi:10.1002/(SICI)1520-6750(200006)47:4¡269::AID-NAV1¿3.0.CO;2-Q.

Fildes, R., Hibon, M., Makridakis, S., Meade, N., 1998. Generalising about univariate forecasting methods: further empirical evidence. International journal of forecasting 14, 339–358. doi:10.1016/S0169-2070(98)00009-0.

Gardner, E.S., 2006. Exponential smoothing: The state of the art—part ii. International journal of forecasting 22, 637–666. doi:10.1016/j.ijforecast.2006.03.005.

Hoerl, A.E., Kennard, R.W., 2000. Ridge regression: Biased estimation for nonorthogonal problems. Technometrics 42, 80–86. doi:10.1080/00401706.2000.10485983.

Hyndman, R.J., Billah, B., 2003. Unmasking the theta method. International journal of forecasting 19, 287–290. doi:10.1016/S0169-2070(01)00143-1.

Hyndman, R.J., Grose, S., Koehler, A.B., Snyder, R.D., 2002. A state space framework for automatic forecasting using exponential smoothing methods. International journal of forecasting 18, 439–454. doi:10.1016/S0169-2070(01)00110-8.

Hyndman, R.J., Koehler, A.B., Ord, J.K., Snyder, R.D., 2008. Forecasting with exponential smoothing : the state space approach. Springer.

Johnson, S.G., 2021. The nlopt nonlinear-optimization package URL: http://github.com/stevengj/nlopt.

Kadipasaoglu, S.N., Sridharan, V., 1995. Alternative approaches for reducing schedule instability in multistage manufacturing under demand uncertainty. Journal of operations management 13, 193–211. doi:10.1016/0272-6963(95)00023-L.

Makridakis, S., Hibon, M., 2000. The m3-competition: results, conclusions and implications. International journal of forecasting 16, 451–476. doi:10.1016/S0169-2070(00)00057-1.

Makridakis, S., Spiliotis, E., Assimakopoulos, V., 2018. The m4 competition: Results, findings, conclusion and way forward. International journal of forecasting 34, 802–808. doi:10.1016/j.ijforecast.2018.06.001.

Sadeghi, A., 2015. Providing a measure for bullwhip effect in a two-product supply chain with

exponential smoothing forecasts. International journal of production economics 169, 44–54. doi:10.1016/j.ijpe.2015.07.012.

Snyder, R.D., 1985. Recursive estimation of dynamic linear models. Journal of the Royal Statistical Society. Series B, Methodological 47, 272–276. doi:10.1111/j.2517-6161.1985.tb01355.x.

Svetunkov, I., 2021. smooth: Forecasting Using State Space Models. URL: `https://github.com/config-i1/smooth`. r package version 3.1.2.41023.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B, Methodological 58, 267–288. doi:10.1111/j.2517-6161.1996.tb02080.x.

Zou, H., 2006. The adaptive lasso and its oracle properties. Journal of the American Statistical Association 101, 1418–1429. URL: `https://www.tandfonline.com/action/journalInformation?journalCode=uasa20`, doi:10.1198/016214506000000735.