Shrinkage Estimator for Exponential Smoothing Models

Kandrika 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 the practice and has shown its efficacy and reliability in many business applications. Yet, there are cases, for example when the estimation sample is limited, that the estimated smoothing parameters can be erroneous, often unnecessarily large. This can lead to over-reactive forecasts, and high forecast errors. Motivated by these challenges, we investigate the use of shrinkage estimators for exponential smoothing. This can help with parameter estimation and mitigating parameter uncertainty. Building on the shrinkage literature, we explore ℓ_1 and ℓ_2 shrinkage for different time series and exponential smoothing model specifications. From the simulation and the empirical study, we find that using shrinkage in exponential smoothing results in forecast accuracy improvements and better prediction intervals. In addition, using bias-variance decomposition we show the interdependence between smoothing parameters and initial values and the importance of the initial value estimation on point forecasts and prediction intervals.

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

1. Introduction

Forecasting is essential to support decisions such as inventory management, production planning, procurement, and other. To effectively support decisions, forecasts are expected to consistent and reliable. In contrast, volatile forecasts can induce more costs due to re-planning, schedule instability, and low service level (Kadipasaoglu and Sridharan, 1995). Forecasts with such characteristics can also mitigate potential issues related to overfitting and erratic forecast selection (Barrow et al., 2021).

Exponential smoothing is widely used in forecasting. It is robust, easy to implement, and amongst the top-performing methods in forecasting competitions (Fildes et al., 1998; Makridakis and Hibon, 2000; Makridakis et al., 2018), available widely in many software. It has been developed extensively over the years (Gardner, 2006). Hyndman et al. (2002) introduced a state-space model formulation for exponential smoothing, to handle time series with different trend

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

Email address: k.pritularga@lancaster.ac.uk (Kandrika Pritularga) and seasonal components. Hyndman et al. (2008b) expanded the model taxonomy to include a variety of different Error, Trend, Seasonal components, leading to the acronym ETS that is commonly used for the exponential smoothing family of models. It has two important groups of parameters, namely the smoothing parameters and and the initial values. The smoothing parameters control how new information impacts the forecasts of the model and the initial values act as proxies for the information prior to the collected observations. In a trend model, we can include a parameter which dampens the otherwise linear trend. The conventional methodology in ETS utilises the single source of error (SSOE) framework (Snyder, 1985). Hyndman et al. (2002, 2008b) automated the methodology by employing the maximum likelihood estimation (MLE) and information criteria in order to select the most appropriate model to the data. This approach produces consistent and efficient estimates of parameters asymptotically, in the statistical sense. However, in practice time series are typically short, which can affect the quality of the estimation. As a result, ETS potentially suffers from overfitting and might produce inaccurate forecasts (Barrow et al., 2021).

In addition, the literature highlights the benefits of lower smoothing parameters in forecasting. Johnston and Boylan (1994) argue that a lower smoothing parameter in simple (level) exponential smoothing can reduce forecast errors. Special cases, such as the original Theta method, can be seen as having a parameter set to zero, resulting in a deterministic state (Assimakopoulos and Nikolopoulos, 2000; Hyndman and Billah, 2003). The Theta method has shown good performance in both the M3 and M4 competitions, and the authors argue that the Theta method is able to capture long-term trends, modelled as deterministic. We agree that this can be beneficial, and that more consideration needs to be given to the potential benefits of low smoothing parameters, but this should be done in an non-arbitrary manner.

We propose to control the smoothing parameters by introducing shrinkage in ETS. The concept of shrinkage is widely used in the regression context, with LASSO and ridge being the two most popular (Tibshirani, 1996; Hoerl and Kennard, 2000). Both reduce the effect of explanatory variables on the target variable by shrinking their respective coefficients towards zero. Forecasting with shrinkage regression has been shown to produce accurate forecasts (Sagaert et al., 2019). Here, we develop a shrinkage estimation procedure for ETS, to obtain reliable and consistent forecasts. We shrink the smoothing parameters in ETS to control the effect of new information on the states of the model. However, the effect of shrinkage on ETS is different from the one in regression models. Shrinkage can have two main effects for regression. First, by resulting in smaller coefficients, models become more resistant to estimation issues due to sampling uncertainty. Second, shrinking parameters to zero, as with LASSO regression, variables can be eliminated from the model. Irrespective whether LASSO, ridge, or other variants of shrinkage are used, there are additional modelling benefits, such as being able to estimate models when the number of explanatory variables exceeds the sample size and being able to obtain estimates for highly multicollinear systems (Hastie et al., 2015). The proposed shrinkage in ETS matches the first operation of shrinkage, but does not eliminate states from the model. By shrinking the smoothing parameters, we reduce the effect of new information on the model states. The literature has investigated the positive effect of shrinking the value of model parameters to improve forecasting performance. One such application is various approaches that shrink the size of the seasonal estimates, demonstrating positive effects on accuracy (Bunn and Vassilopoulos, 1999; Miller and Williams, 2003, 2004; Kourentzes et al., 2014). A motivation for using shrinkage for seasonal estimates is the relatively large number of parameters needed to estimate a seasonal profile in relation to the available seasons in the fitting sample. A similar application can be seen in using shrinkage estimators for large variance-covariance matrices (Daniels and Kass, 2001; Schäfer and Strimmer, 2005) with applications, for instance, in hierarchical forecasting (Wickramasuriya et al., 2019). Barrow et al. (2021) demonstrate that these benefits are relevant even in the case of ETS with only a few parameters to estimate.

Recognising the estimation issues originating from limited samples, other approaches have investigated improving parameter estimates by regularising parameters towards a pooled estimate across time series, or an informative prior, for instance by using empirical Bayes (Greis and Gilstein, 1991). Although the prior can be geared towards shrinking parameters to zero, these approaches often have different targets, and relate to pooled estimation methods (e.g., Trapero et al., 2015, use pooled estimation to obtain more reliable estimates for promotional effects) and more recently global learning methods (Makridakis et al., 2021). These approaches go beyond the univariate case, and do not directly relate to the proposed univariate ETS shrinkage, yet they demonstrate the longstanding interest in the literature to investigate parameter shrinkage estimation improvements via shrinkage and regularisation.

The shrinkage estimators investigated here have parallels with other recent contributions in the literature for mitigating parameter estimation issues for ETS (e.g., Barrow et al., 2021, use M-estimators and boosting), or limiting the pool of exponential smoothing models (Kourentzes et al., 2019; Meira et al., 2021), all aiming to reduce the inconsistency of forecasts. Our approach builds on the well-researched shrinkage estimators and does not require the introduction of ad-hoc heuristics, while offering a data-driven way to identify how much to diverge from conventional estimators.

We (a) propose an implementation of shrinkage estimates for ETS; (b) explain the mechanism of shrinkage in ETS; and (c) evidence the effect of shrinkage on predictive accuracy. Using simulated and real data, we show that our estimation procedure works well in the ETS framework and leads to a reduction in bias for longer horizons, more accurate point forecasts, and more precise prediction intervals. In Section 2, we present the proposed shrinkage for exponential smoothing. Section 3 describes the experimental setup used to validate the efficacy of shrinkage for ETS and the findings. We use the proposed estimator to a real-life application in Section 4 and conclude in Section 5.

2. Exponential Smoothing with Parameter Shrinkage

ETS reconstructs the time series from its unobserved states: level, trend, and seasonality. Hyndman et al. (2008b) build a taxonomy based on this, providing a naming scheme for the different model forms, such as ETS(A,N,N), which means a model with an additive error (A) no trend (N) no seasonality (N). Multiplicative states are denoted by M, and Ad and Md denote additive and multiplicative damped trends respectively. Let y_t be an observed time series, at period t. Assuming that the time series is constructed from different unobserved states ($\boldsymbol{x}_t$), we can write a general exponential smoothing model according Hyndman et al. (2008b):

$$y_t = w(\boldsymbol{x}_{t-1}) + r(\boldsymbol{x}_{t-1})\varepsilon_t, \tag{1}$$

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

where Eq. (1) and Eq. (2) are the measurement and the transition equations. the error term (ε_t) has zero mean and variance of σ^2. Oftentimes, ε_t is assumed to be Gaussian, but can be relaxed for non-negative time series. In that case, the log-normal distribution can be an alternative to the normal distribution (Svetunkov, 2022). Let us consider two simple example from additive and multiplicative models, the ETS(A,N,N) and the ETS(M,N,N). For ETS(A,N,N) $w(\boldsymbol{x}_{t-1}) = f(\boldsymbol{x}_{t-1}) = l_{t-1}$, $r(\boldsymbol{x}_{t-1}) = 1$, and $g(\boldsymbol{x}_{t-1}) = \boldsymbol{g} = \alpha$, where l_t is the level of the time series at time t. As a result $\varepsilon_t = y_t - l_{t-1}$. On the other hand, for ETS(M,N,N), $w(\boldsymbol{x}_{t-1}) = f(\boldsymbol{x}_{t-1}) = r(\boldsymbol{x}_{t-1}) = l_{t-1}$, and $g(\boldsymbol{x}_{t-1}) = \alpha l_{t-1}$. The multiplicative error is a relative error, i.e., $\varepsilon_t = \frac{y_t - l_{t-1}}{l_{t-1}}$.

In the general case, $\boldsymbol{g}$ contains all level (α), trend (β), and seasonal (γ) smoothing parameters. While there are different types of parameter restrictions (Hyndman et al., 2008a), the most popular are: $0 < \alpha < 1$, $0 < \alpha < \beta$, and $0 < \gamma < 1 - \alpha$. We use these parameter restrictions here.

Gardner and McKenzie (1985) proposed the ϕ parameter in the ETS models to dampen the linear trend and we call it the dampening parameter. In a damped trend model ϕ is added into the transition matrix and the measurement vector. The parameter space is $0 < \phi \leq 1$ (Hyndman et al., 2008b, p. 157). In practice, ϕ is typically close to 1, since the trend still persists, but is slightly diminished.

The modelling employs MLE of the smoothing parameters, the dampening parameter, and the vector of initial values ($\boldsymbol{x}_0$), and σ^2. The likelihood is often simplified to get the concentrated likelihood function (see, for example, Hyndman et al., 2008b, p. 69)

$$\ell^*(\boldsymbol{\theta}, \boldsymbol{x}_0 | \mathcal{Y}_t) = -\frac{T}{2} \log \left(2\pi e \frac{1}{T} \sum_{t=1}^{T} {\varepsilon_t}^2 \right) - \sum_{t=1}^{T} \log |r(\boldsymbol{x}_{t-1})|, \tag{3}$$

where $\boldsymbol{\theta} = \{\boldsymbol{g}, \phi\} = \{\alpha, \beta, \gamma, \phi\}$, T is the number of observations, and $\mathcal{Y}_t$ denotes the available information up until the time t. ε_t is the error term, corresponds to the forecast error in the case a model matching the data generating process. Throughout the paper 'smoothing parameters' refers to $\boldsymbol{\theta}$, unless stated otherwise. Under MLE, consistent and efficient estimators, in the statistical sense, can be achieved when $T \to \infty$. However, in practice, large sample sizes, for which the asymptotic behaviour becomes relevant, are rarely obtained due to product life cycle, product discontinuity, or low-quality data management (Ord et al., 2017). This may harm the efficiency of the estimates, and eventually worsen the predictive accuracy. In the case of misspecified models due to unknown data generating processes, these problems can be exaggerated more.

When we maximise Eq. (3), it is equivalent to minimising the augmented mean squared error. In case of additive error models it reduces to Mean Squared Error, which is often used in estimation of statistical models in a variety of contexts:

$$\mathrm{MSE}(\boldsymbol{\theta}, \boldsymbol{x_0} | \mathcal{Y}_t) = \frac{1}{T} \sum_{t=1}^{T} \varepsilon_t^2. \tag{4}$$

The MSE-based estimation is also known to produce consistent and efficient estimates of parameters, but suffers from the issues similar to MLE on small samples. To counter the effect of small-sample inefficiency, we modify Eq. (4), introducing a shrinkage component for the smoothing parameters. The conventional loss function, which is widely applied in regression problems (Tibshirani, 1996; Hoerl and Kennard, 2000), is shown as,

$$\text{MSE}(\boldsymbol{\theta}, \boldsymbol{x_0}|\mathcal{Y}_t) + \lambda^* p(\boldsymbol{\theta})\mathbf{1}, \tag{5}$$

where $\lambda^* \geq 0$ is the shrinkage hyper-parameter. The λ^* is estimated separately from the smoothing parameters (see Section 2.3) and hence treated separately. Given that λ^* has no upper bound, it can be difficult to find its optimal value. We modify Eq. (5) so that the shrinkage hyper-parameter has a finite upper bound:

$$(1 - \lambda)\text{MSE}(\boldsymbol{\theta}, \boldsymbol{x_0}|\mathcal{Y}_t) + \lambda p(\boldsymbol{\theta})\mathbf{1}, \tag{6}$$

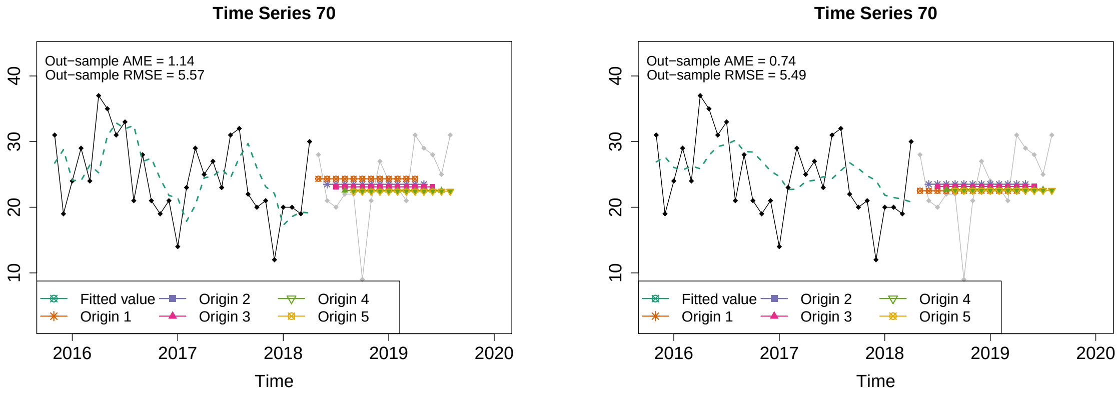
where $p(\boldsymbol{\theta}) = \begin{bmatrix} p(\alpha) & p(\beta) & p(\gamma) & p(\phi) \end{bmatrix}$. $p(\cdot)$ is characterised by the selected norm, and $\mathbf{1}$ is a vector of ones. $\lambda \in [0, 1]$ and controls the shrinkage rate of the smoothing parameters. With the ℓ_1 norm, $p(\boldsymbol{\theta})\mathbf{1} = |\alpha|+|\beta|+|\gamma|+|1-\phi|$, and with the ℓ_2 norm, $p(\boldsymbol{\theta})\mathbf{1} = (\alpha)^2+(\beta)^2+(\gamma)^2+(1-\phi)^2$. Both norms shrink $\boldsymbol{g}$ to zero, but the ℓ_1 loss can result in sparser solutions, i.e., with more parameters equal to zero (Hastie et al., 2015). Note that in the case of ϕ we regularise it to 1, simplifying the damped trend model to the linear trend one.

By shrinking the parameters, we do not eliminate the states, but we reduce the degree of stochasticity of the states. We transition from a stochastic to a more deterministic model. Suppose that we have a trended model, e.g., ETS(A,A,N). If we shrink β to zero, we will get a deterministic trend. This is so as the additive trend state, $b_t = b_{t-1} + \beta\varepsilon_t$, is no longer updated by ε_t. In a seasonal model, e.g., ETS(A,N,A), shrinking γ to zero means that the seasonality becomes deterministic, i.e., remaining identical across periods. As $\boldsymbol{\theta} \to \mathbf{0}$ then $\boldsymbol{x}_t \to \boldsymbol{x}_0$. The structure of both models are still intact, but the influence of the stochastic ε_t changes. Consequently, shrinking the smoothing parameters has implications on the initial values. As the corresponding states become more deterministic, it increases the importance of the estimation of the initial values. For example, for ETS(A,A,N) with $\beta = 0$, the model resembles the idea of the Theta method (Assimakopoulos and Nikolopoulos, 2000; Hyndman et al., 2002), where the trend smoothing parameter shrinkage results in a drift trend. The method reinforced the deterministic trend, and empirically performed well in M3 and M4 Competition (Makridakis and

Hibon, 2000; Makridakis et al., 2018). However, instead of forcing the parameters to exactly zero arbitrarily (and to 1 for ϕ), we shrink the smoothing parameters conditional on the data itself. The shrinkage is data-dependent and when it is not needed λ can become close to zero, leading to a minimisation similar to the minimisation of MSE. We do not shrink the initial values and therefore the states (level, trend, seasonality) are not eliminated. As the states become more deterministic, the reliance on the initial values increases, and accordingly affects their estimation. Moreover, the model does not simplify in terms of modelled time series components, but reduces in terms of parameters. This differs from standard regression shrinkage estimators, where as coefficients are reduced to zero, inputs are eliminated from the now simpler regression.

We can interpret Eq. (6) as a trade-off between model fitness and model inertia. When λ is close to 0, the estimator puts more weight on the fitness. On the other hand, when λ is close to 1, the estimator puts more weight on the model inertia. The model inertia is defined as a situation where the updated information does not affect the forecasts. Note here that Eq. (6) is applicable to both additive and multiplicative models because they are seen to be identical via their recursive relationship (Hyndman et al., 2008b, p. 55). As a result, their point forecasts are identical but they differ in the prediction intervals. Thus, the shrinkage estimation focuses on controlling how the model behaves and can be implemented in any types of model structure.

We demonstrate the effect of the smoothing parameter shrinkage on the states and the forecasts. In the example, we use a time series from the empirical evaluation (Section 4) with the sample size of 30. We identify its model structure to be ETS(M,N,N), and estimate the smoothing parameters with (a) MLE and (b) shrinkage. In terms of the forecasting task, we produce 1-12 step ahead forecasts from 5 origins.



(a) MLE, $\hat{\alpha}_{\text{MLE}} = 0.47$

(b) Shrinkage, $\hat{\alpha}_{\text{REG}} = 0.19$

Figure 1: Examples of different estimation approaches for 5 origins. AME is the absolute mean error and both error measures calculates the accuracy of 1-12 step ahead forecasts.

Figure 1 demonstrates the effect of the smoothing parameter shrinkage on the single-state in the fitted value in-sample and the out-of-sample forecasts. The shrinkage reduces α by more than a half. By doing so we reduce the effect of updating information ($\hat{\alpha}\varepsilon_t$) on the state and produce a smoother fit. On the other hand, with a higher α, the state is updated more and results in a more volatile in-sample fit. Consequently, the forecasts from the shrunk model become more stable and consistent than the ones from the MLE. Apart from that, the shrunk model produces more accurate and less biased forecasts than MLE does.

2.1. Weighted Shrinkage

Eq. (6) assumes that we shrink all smoothing parameters at the same rate. However, this may not be reasonable. Different shrinkage rates would imply that we expect each state to have a different level of stochasticity. By having different shrinkage rates, for instance, we can have a model with a deterministic seasonal pattern and a stochastic trend. We adjust Eq. (6) to accommodate different levels of stochasticity for each state, by adding weights for each parameter in the penalty function. The loss function with weighted shrinkage is shown as,

$$\{\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{x}}_0\} = \underset{\theta, \boldsymbol{x}_0}{\arg\min} \left((1-\lambda)\mathrm{MSE}(\boldsymbol{\theta}, \boldsymbol{x_0}|\mathcal{Y}_t) + \lambda p(\boldsymbol{\theta})\boldsymbol{\omega} \right), \tag{7}$$

where $\boldsymbol{\omega}$ is a vector of weights, with $\omega_i \in [0,1]$, $\sum_{i=1}^{k} \omega_i = 1$, and k is the number of states. Suppose that λ is 0.2 for ETS(A,A,N). When ω_α and ω_β are 0.5, both states are shrunk at the same rate. However, when ω_α and ω_β are 0.1 and 0.9 respectively, the trend has a tendency to be more deterministic than the level because the trend is shrunk to zero at a faster rate than the level. A similar penalty function is used on the adaptive shrinkage by Zou (2006), where each parameter in the penalty function may have different shrinkage hyperparameters in groups or individually.

2.2. Prediction Intervals

Decision makers may often require the predictive distribution of forecasts as they need to take into account the future uncertainty. Quantile forecasts are relevant in organisations, such as in retail, warehouse staffing decisions, and dynamic pricing optimisation (Sillanpää and Liesiö, 2018; Sanders and Graman, 2009; Chen and Chen, 2018). In this section, we dissect the construction of theoretical forecast variance and discuss the effect of shrinkage on each of its elements.

Producing prediction intervals requires multi-step ahead forecast variances. Literature provides many approaches to produce the forecast variance and in this paper, we discuss two of them, namely the theoretical and the empirical prediction intervals. For additive models, we can construct the theoretical prediction intervals analytically, where they are affected by the forecast horizon, the parameters, and the in-sample residuals. The analytical variance is tractable and the in-sample residuals are independent and identically distributed (i.i.d.) (Hyndman et al., 2008b). In other words, we have to assume that the residuals are homoscedastic and are not autocorrelated. For the other classes of models such as multiplicative and mixed models, we can use approximate or simulation-based prediction intervals.

To calculate the theoretical prediction interval, we require to estimate the forecast variance at the forecast horizon h. According to Hyndman et al. (2008b), for linear homoscedastic models, the forecast variance is:

$$\mathrm{V}(y_{t+h|t}) = \begin{cases} \sigma^2, & \text{if } h = 1 \\ \sigma^2 \left[1 + \sum_{j=1}^{h-1} c_j^2\right], & \text{if } h \geq 2 \end{cases}, \tag{8}$$

where c_j depends on the forecast horizon and smoothing parameters (Hyndman et al., 2008b, p. 82). For example, ETS(A,N,N), $\sum_{j=1}^{h-1} c_j^2 = \alpha^2(h-1)$. In general, there are three factors that affect the forecast variance, namely $\boldsymbol{g}$, h, and σ^2 (and ϕ). In most cases, σ^2 is unknown and is estimated from the in-sample residuals. Due to the recursive nature of the model, the residuals depend on the estimates of $\boldsymbol{g}$ and $\boldsymbol{x}_0$ (and ϕ). We utilise the bias-variance decomposition of the sum squared in-sample residuals to explain how each of these affects the residuals. The derivation assumes a non-damped trend additive time series, which leads to a fixed $\boldsymbol{F}$. Otherwise, the conditional variance of the states does not have a closed-form due to the multiplication of the transition matrix (which contains the dampening parameter) by the states.

$$\mathrm{E}\left(y_t - \hat{y}_{t|t-1}|\mathcal{Y}_t\right)^2 = \underbrace{\mathrm{E}\left(\boldsymbol{\mu}_t - \mathrm{E}\left(\hat{y}_{t|t-1}\right)|\mathcal{Y}_t\right)^2}_{\text{model bias}} + \underbrace{\mathrm{E}\left(\mathrm{E}\left(\hat{y}_{t|t-1}\right) - \hat{y}_{t|t-1}|\mathcal{Y}_t\right)^2}_{\text{model variance}} + \underbrace{\sigma^2}_{\substack{\text{irreducible}\\ \text{error}}}, \tag{9}$$

where $y_t = \boldsymbol{\mu}_t + \varepsilon_t$ and $\boldsymbol{\mu}_t$ is the structure of the time series. $\hat{y}_{t|t-1}$ is the fitted value of y_t conditional on the previous information and its recursive form is shown as,

$$\hat{y}_{t|t-1} = \boldsymbol{w}^\top \boldsymbol{F}^{t-1}\hat{\boldsymbol{x}}_0 + \sum_{i=1}^{t-1} \boldsymbol{w}^\top \boldsymbol{F}^{t-i-1}\hat{\boldsymbol{g}} e_{i|i-1}, \tag{10}$$

where $e_{i|i-1}$ is the residual at time i, which is conditional on the previous information. There are two conditions that determine the model bias, namely a correctly-specified model and a mis-specified model. In the former case, $\mathrm{E}(\hat{y}_{t|t-1}|\mathcal{Y}_{x_0}) = \boldsymbol{\mu}_t = \boldsymbol{w}^\top \boldsymbol{F}^{t-1}\boldsymbol{x}_0$, where $\mathcal{Y}_{x_0}$ is the information related to $\boldsymbol{x}_0$. Since $\mathrm{E}(\hat{\boldsymbol{x}}_0|\mathcal{Y}_{x_0}) = \boldsymbol{x}_0$, thus the model bias vanishes. However, when the model is incorrectly specified, $\mathrm{E}(\hat{\boldsymbol{x}}_0|\mathcal{Y}_{x_0}) \neq \boldsymbol{x}_0$, hence the model bias should be non-zero.

To dissect the model variance, we insert Eq. (10) into the model variance in Eq. (9) and the decomposition is formulated as:

$$\begin{aligned}\mathrm{E}\left(\mathrm{E}\left(\hat{y}_{t|t-1}\right) - \hat{y}_{t|t-1}|\mathcal{Y}_t\right)^2 = \mathrm{E}\left(\boldsymbol{w}^\top \boldsymbol{F}^{t-1}(\boldsymbol{x}_0 - \hat{\boldsymbol{x}}_0)|\mathcal{Y}_t\right)^2 + \mathrm{E}\left(\sum_{i=1}^{t-1} \boldsymbol{w}^\top \boldsymbol{F}^{t-i-1}\hat{\boldsymbol{g}}e_{i|i-1}|\mathcal{Y}_t\right)^2 \quad (11)\\ - 2\mathrm{cov}\left(\boldsymbol{w}^\top \boldsymbol{F}^{t-1}(\boldsymbol{x}_0 - \hat{\boldsymbol{x}}_0), \sum_{i=1}^{t-1} \boldsymbol{w}^\top \boldsymbol{F}^{t-i-1}\hat{\boldsymbol{g}}e_{i|i-1}|\mathcal{Y}_t\right).\end{aligned}$$

Eq. (11) shows that the model variance is affected by the variance of $\hat{\boldsymbol{g}}$ and $\hat{\boldsymbol{x}}_0$, and the covariance between them. This shows that it is important to take care of not only the smoothing parameters but also the estimation of initial values. If the model is assumed to be unbiased, the sum squared of the in-sample residuals contains σ^2 and the model variance, where the model variance depends on the estimation of the smoothing parameters and the initial values. Linking back to Eq (8), we can say that the forecast variance is affected not only by $\hat{\boldsymbol{\theta}}$ and h, but also by $\hat{\boldsymbol{x}}_0$, as a result of the unknown σ^2.

Our forecast variance analysis is under the assumption of independent and identically distributed residuals. However, this assumption can be violated because the data generating process is unknown. As a result, the residuals may be correlated over time and/ or have time-varying variance. The literature suggests using empirical prediction intervals (Chatfield, 2001). We estimate the multi-step ahead forecast error and construct the requested quantiles for the empirical distribution. The empirical approach is robust and usually outperforms the theoretical one, when the normality assumption does not hold (Lee and Scholtes, 2014; Trapero et al., 2019).

2.3. On choosing hyper-parameters

We need to develop an approach for obtaining the shrinkage hyper-parameters. Grid search is widely used for this purpose (Hoerl and Kennard, 2000; Bergstra et al., 2012). It is relatively simple but computationally expensive. Instead, we propose to implement a derivative-free shrinkage hyper-parameter optimisation. We aim to minimise the mean squared one-step ahead holdout forecast error, shown as,

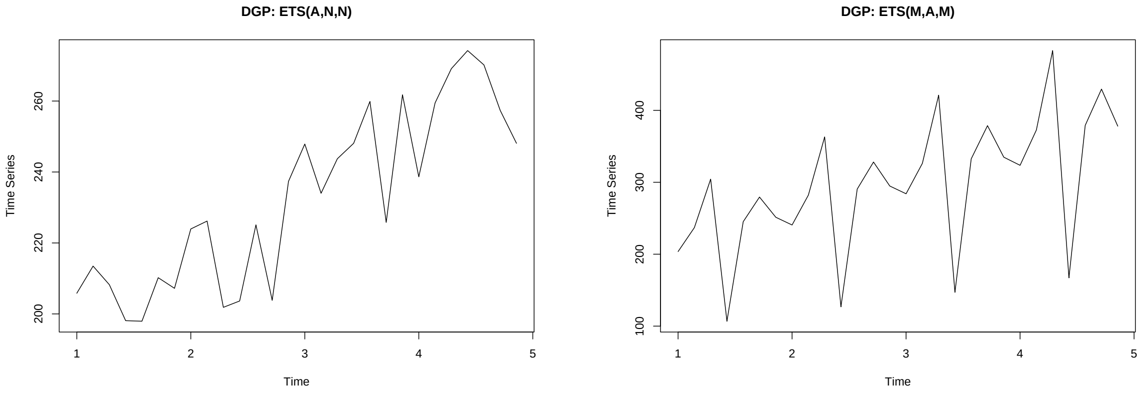
$$\{\hat{\lambda}, \hat{\boldsymbol{\omega}}\} = \arg\min_{\lambda, \boldsymbol{\omega}} \frac{1}{K} \sum_{i=1}^{K} \text{MSE}(\lambda, \boldsymbol{\omega}, \hat{\boldsymbol{\theta}}_i, \hat{\boldsymbol{x}}_{i,0}, y_{1:i}), \tag{12}$$

where K is the number of forecast origins and $y_{1:i}$ is the in-sample time series starting from $t = 1$ to $t = i$. $\hat{\boldsymbol{\theta}}_i$ and $\hat{\boldsymbol{x}}_{i,0}$ are the estimated parameters and initial values for origin l. First, we estimate $\hat{\boldsymbol{\theta}}_i$ and $\hat{\boldsymbol{x}}_{i,0}$ given $\hat{\lambda}$ and $\hat{\boldsymbol{\omega}}$. Then, we change $\hat{\lambda}$ and $\hat{\boldsymbol{\omega}}$, minimising Eq. (12). We use the Nelder-Mead algorithm. The algorithm terminates when it meets a stopping criteria, i.e., the parameter tolerance or the maximum number of iterations. We use an uninformative initialisation, with 0.1 for λ and equal weights for each smoothing parameter.

3. Simulation Study

We perform four simulations to demonstrate the efficacy of the proposed estimation approach in terms of its point forecasts and prediction interval accuracy. The four simulation experiments investigate alternative shrinkage approaches: (a) the ℓ_1 unweighted shrinkage (ℓ_1-US), (b) the ℓ_1 weighted shrinkage (ℓ_1-WS), (c) the ℓ_2 unweighted shrinkage (ℓ_2-US), and (d) the ℓ_2 weighted shrinkage (ℓ_2-WS). US and WS denote the unweighted and the weighted shrinkage estimators.

3.1. Experimental Design



(a) DGP: ETS(ANN)

(b) DGP: ETS(MAM)

Figure 2: Examples of the simulated time series

In this simulation study, we generate data using seven different Data Generating Processes (DGP), where six of them are additive models and one of them is a mixed model with multiplicative errors and seasonality. The parameters of these models are summarised in Table 1.

Values in Table 1 reflect generated time series that have a moderate amount of stochasticity, being neither too smooth nor volatile. As for the parameters for ETS(M,A,M), they guarantee a non-explosive behaviour. Figure 2 presents examples of simulated time series for different data generating processes. The error for the additive time series follows a normal distribution with zero mean and unit standard deviation. The multiplicative error $(1 + \varepsilon_t)$ follows a normal distribution with a mean of 1 and a standard deviation of 0.0075 to avoid explosive time series. We generate daily time series with the sample sizes of 28 and 420 to explore shrinkage on short (a month of daily time series) and relatively long time series (15 months of daily time series), influenced by the typical business forecasting context. For the former case, we expect that the ETS with shrinkage performs better than MLE and for the latter case we expect MLE to be more competitive, with the longer sample aiding the MLE estimators.

DGP	α	β	γ	ϕ	l_0	b_0
ETS(A,N,N)	0.400				200	0.500
ETS(A,A,N)	0.400	0.300			200	0.500
ETS(A,N,A)	0.400		0.100		200	0.500
ETS(A,A,A)	0.400	0.300	0.100		200	0.500
ETS(A,Ad,A)	0.400	0.300	0.100	0.940	200	0.500
ETS(A,Ad,N)	0.400	0.300		0.940	200	0.500
ETS(M,A,M)	0.100	0.075	0.050		200	5.000

Table 1: A summary of data generating processes. The seasonal initial values are generated randomly.

We apply seven models to each generated series, namely the ETS(A,N,N), ETS(A,A,N), ETS(A,N,A), ETS(A,A,A), ETS(A,Ad,A), ETS(A,Ad,N), and ETS(M,A,M). Thus, we have 49 combinations of DGPs and models. We classify them according to model misspecification, resulting in five groups: correctly specified models (CR), over-specified models (OV), under-specified models (UN), incorrect-state models (IS), and incorrect-error-seasonality models (IE). We include seasonal processes because many time series may be seasonal in practice.

We produce 1-7 step-ahead point forecasts, matching a complete week of daily data, and construct prediction intervals of 80%, 85%, 90%, 95%, and 99% confidence levels, theoretically and empirically. We repeat the simulation experiment for 500 times, which was sufficient for the summary statistics to converge. We use the sim.es() function to generate data and the adam() function for model fitting. Both are available in the smooth package for R (Svetunkov, 2021; R Core Team, 2021). As for the hyper-parameter estimation, we use the nloptr package for R (Johnson, 2021).

We use two error measures to evaluate point forecasts: RMSE and AME. The former

Model		ANN	AAN	AAdN	ANA	AAA	AAdA	MAM
DGP	ANN	CR	OV	OV	OV	OV	OV	IE
	AAN	UN	CR	OV	IS	OV	OV	IE
	AAdN	UN	UN	CR	IS	IS	OV	IE
	ANA	UN	IS	IS	CR	OV	OV	IE
	AAA	UN	UN	IS	UN	CR	OV	IE
	AAdA	UN	UN	UN	UN	UN	CR	IE
	MAM	IE	IE	IE	IE	IE	IE	CR

Table 2: A classification based on pairs of DGPs and applied models.

measures the accuracy while the latter measures the size of the bias:

$$\text{RMSE} = \sqrt{\frac{1}{h}\sum_{k=1}^{h}\left(y_{t+k} - \hat{y}_{t+k|t+k-1}\right)^2}, \quad \text{AME} = \left|\frac{1}{h}\sum_{k=1}^{h} y_{t+k} - \hat{y}_{t+k|t+k-1}\right|.$$

We use a percentage difference between the measures to quantify the improvement or deterioration in accuracy of shrinkage models in comparison with MLE ones.

$$\text{dRMSE} = \frac{(\text{RMSE}_{\text{MLE}} - \text{RMSE}_{\text{SHR}})}{\text{RMSE}_{\text{MLE}}}$$

where SHR denotes the shrinkage. A similar formula is also applied to AME.

For the prediction interval, we use the scaled Mean Interval Score or sMIS (Koenker and Bassett, 1978) and the scaled Pinball Score (Gneiting and Raftery, 2007) that not only take the width of the interval into account, but also how well the interval captures the uncertainty. Narrow intervals do not necessarily indicate precise intervals, and may indicate model misspecification (Chatfield, 2001). We scale both measures with the mean of the absolute in-sample value:

$$\text{sMIS} = \frac{\frac{1}{h}\sum_{k=1}^{h}\big((ub_{t+k} - lb_{t+k}) + \frac{2}{\tau}(lb_{t+k} - y_{t+k}1(y_{t+k} < lb_{t+k}) + \frac{2}{\tau}(y_{t+k} - ub_{t+k}1(y_{t+k} > ub_{t+k}))\big)}{|\bar{y}|}$$

$$\text{sPinball} = \frac{(1-\tau)\sum_{y_{t+j}<q_{t+k},k=1,..,h}|y_{t+k} - q_{t+k}| + \tau\sum_{y_{t+k}>q_{t+k},k=1,..,h}|y_{t+k} - q_{t+k}|}{|\bar{y}|}$$

where $|\bar{y}|$ is the mean of the absolute in-sample, τ is the level of confidence, ub_t and lb_t are the upper and lower bounds, and q is the value of a specific quantile of the distribution.

Penalty	Type	CR	OV	UN	IS	IE	Overall
ℓ_1	US	2.22	1.36	4.02	1.44	2.36	2.28
	WS	7.06	0.94	13.18	**3.05**	**6.78**	6.20
ℓ_2	US	3.00	1.12	3.61	1.49	2.57	2.36
	WS	**13.17**	**2.25**	**15.14**	2.34	6.54	**7.89**

Table 3: A summary of aggregate performances in dRMSE for ℓ_1 and ℓ_2 shrinkage for different model specifications, for the small sample size of 28 and 1-7 steps-ahead forecasts. Positive values indicate percentage gain over the performance of the MLE

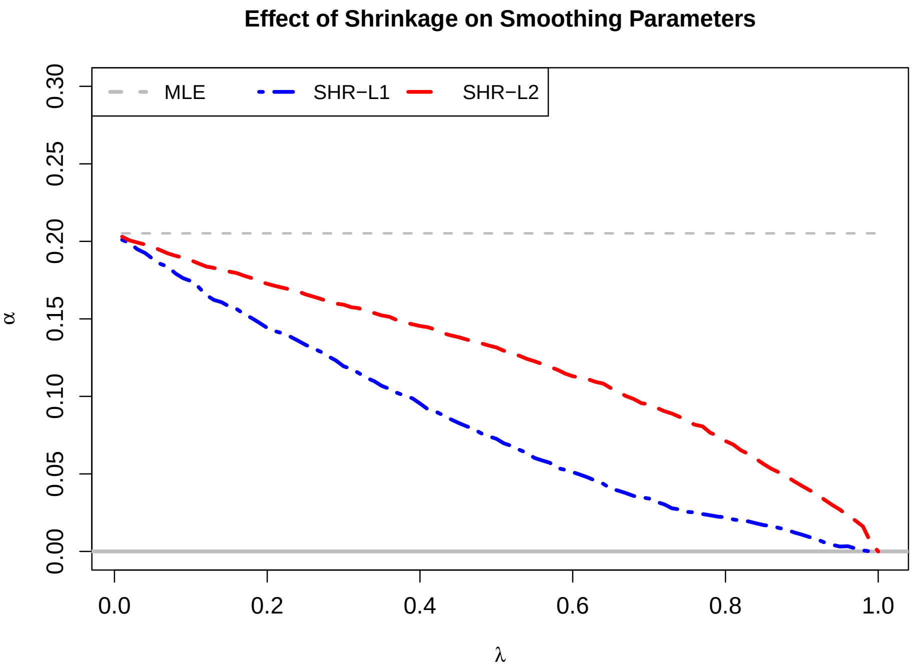


Figure 3: Effects of ℓ_1 and ℓ_2 shrinkage on smoothing parameters; the DGP is ETS(A,N,N), and the fitted model is ETS(A,N,N).

3.2. Findings

3.2.1. Accuracy

Table 3 presents the aggregate performances of ETS models with shrinkage in the four simulation settings. Positive numbers show percentage improvement in accuracy from the MLE and otherwise. Bold numbers denote the best performing result for each column. We can see that ℓ_2-WS outperforms the other approaches, improving the forecast accuracy overall by 7.9% and ℓ_1-US performs the worst among all, although still outperforming MLE. Looking at the model specification, for most cases, ℓ_2-WS outperforms the others, except for the IS and IE scenarios. Regardless of the type and the penalty function, shrinking the smoothing parameters improves the forecasting accuracy, but ℓ_2-WS performs the best. Figure 3 shows the values of the smoothing parameter α for a case where the data generating process is ETS(A,N,N) and the correct model is fitted. For $\lambda = 0$ all MLE, ℓ_1, and ℓ_2 give the same α, as there is no shrinkage. As λ increases ℓ_1 imposes a higher penalty, leading a smaller α. For the extreme case of $\lambda = 1$, both ℓ_1 and ℓ_2 result in $\alpha = 0$. When a model has more smoothing parameters than one, restrictions imposed by ETS can affect how β and γ shrink, with WS providing additional flexibility. For example, since $0 < \gamma < 1 - \alpha$, as α shrinks the possible values for γ changes: a slower shrinkage of α provides more options for γ. Nonetheless, as the optimal λ for ℓ_1 and ℓ_2 can differ, same smoothing parameters are possible for both penalties. Therefore, while in regression models ℓ_1 and ℓ_2 behave differently in terms of model sparsity, for ETS this is not the case. As the smoothing parameters become equal to zero, the respective states become deterministic, but they are not removed from the model. Instead, the two penalties have mainly implications for how the optimizer searches the parameter space. Hereafter, we present the only results for ℓ_2 for brevity since we empirically find it to perform marginally better. The findings we present are applicable to ℓ_1, but with smaller gains over the MLE.

Model	CR	OV	UN	IS	IE	Overall
			$h = 1$			
Measure			dRMSE			
US	8.332	4.151	10.745	**6.531**	9.983	7.948
WS	**26.704**	**6.649**	**27.655**	6.315	**14.878**	**16.440**
Overall	17.518	5.400	19.200	6.423	12.431	12.194
Measure			dAME			
US	**-1.994**	0.493	**0.052**	2.447	-0.901	**0.019**
WS	-15.660	**1.170**	-8.986	**9.256**	**5.166**	-1.811
Overall	-8.827	0.832	-4.467	5.851	2.132	-0.896
			$h = 1 - 7$			
Measure			dRMSE			
US	3.000	1.119	3.612	1.487	2.572	2.358
WS	**13.166**	**2.248**	**15.138**	**2.338**	**6.538**	**7.886**
Overall	8.083	1.684	9.375	1.912	4.555	5.122
Measure			dAME			
US	1.402	0.657	2.754	3.228	2.059	2.020
WS	**4.786**	**2.227**	**13.644**	**7.647**	**7.385**	**7.138**
Overall	3.094	1.442	8.199	5.438	4.722	4.579

Table 4: A summary of performances in dRMSE and dAME for ℓ_2 shrinkage for different model specification and the sample size of 28. Positive values indicate percentage gain over the performance of the MLE.

Building on the understanding of the effects of ℓ_1 and ℓ_2 penalties on the aggressiveness of the shrinkage, we can interpret the marginally better performance of ℓ_1 in the IS and IE cases. In Section 3.2.2 we show that the shrinkage of the smoothing parameters indirectly influences the estimation of the initial values. The ℓ_1 penalty forces the initial values of superfluous states to become smaller faster (see Figure 8), lessening their overall effect. Similarly, when appropriate states are missing (e.g., seasonality) the ℓ_1 penalty, being more aggressive, can keep the smoothing parameters low and therefore guard the existing states from rapidly updating and overfitting. As the proposed shrinkage approach does not focus on model selection, any influence on the initial values is indirect, and we argue that ℓ_2 in general performs the best. In addition, we conducted additional experiments to see if other DGPs and applied models would behave differently with the proposed estimators. Our investigation shows that the results hold for a wide variety of ETS models, including multiplicative ones.

Table 4 presents the summary of the forecasting performance for dRMSE and dAME for the forecast horizons of 1 and 1-7, and sample size of 28. The table is structured similarly to Table 2. In terms of dRMSE for $h = 1$, we see a significant forecast improvement over MLE by 12%. Analysing by the type of shrinkage, the weighted one outperforms the unweighted one across all model specifications, improving performance by 16% on average. Looking at the model specification, the biggest improvement occurs when the model is under-specified, followed by the correctly-specified model, with improvements of 27% and 26%, respectively. If we have an incorrect error and seasonality model, we gain a 14% performance. The least improvements occur in the case of the over-specified and the incorrect-state models, yielding a 6% performance. Both weighted and unweighted shrinkage improve the forecast accuracy. For the longer horizon, the improvement is at 5% overall. As the errors for both MLE and shrinkage increase for longer horizons, the relative improvement decreases, since it is more difficult to forecast in the long run.

In terms of dAME, for h = 1, we can see that the forecasts become more biased than in the case of the MLE. On the other hand, the incorrect models, i.e., the incorrect state and the incorrect error and seasonality, gain the most improvement. We observe an improvement for the longer forecast horizon. Overall, the forecasts are 5% less biased than MLE with the under-specified and the over-specified models gaining the most and the least improvement, respectively.

Figure 4 shows the effect of sample sizes on dRMSE in the five scenarios. We observe higher accuracy improvement on smaller samples. When we have more observations the benchmark MLE performance improves. Similarly to the application of shrinkage in regression, we see that the proposed shrinkage estimator for ETS mitigates the sampling uncertainty that limits the efficiency of MLE for smaller samples.

Overall, for shrinkage for ETS is shown to be beneficial for accuracy. We gain more improvement when we use weighted shrinkage. In terms of forecast bias, shrinkage benefits for

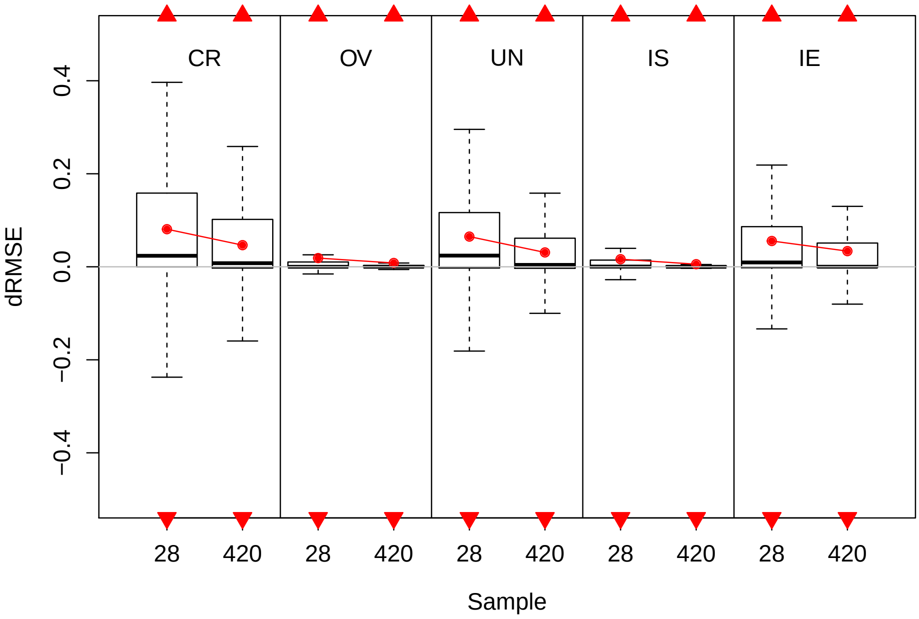


Figure 4: Distributions of dRMSE across different model specifications and sample sizes. The dotted lines denote the mean of the distributions and the arrows show some parts of the distributions outside of the plotting area.

longer horizons.

3.2.2. Prediction Intervals

In this subsection, we demonstrate the performance of prediction intervals of the shrinkage and the MLE models, and explain why the theoretical prediction intervals from the shrinkage outperform the MLE ones, and lastly, we also demonstrate how well the theoretical forecast variance approximates the empirical conditional forecast variance.

Instead of presenting all combinations as in Table 2, we consider several special cases to present the performance of the ℓ_2-WS shrinkage as the models are linear homoscedastic and their prediction intervals are tractable analytically. The special cases are summarised in Table 5.

Specification	CR	OV	UN	IS	IE
DGP	ETS(A,N,N)	ETS(A,N,N)	ETS(A,N,N)	ETS(A,A,N)	ETS(M,A,M)
Model	ETS(A,N,N)	ETS(A,A,N)	ETS(A,N,A)	ETS(A,N,A)	ETS(A,A,A)

Table 5: Several combinations from five levels of model specification

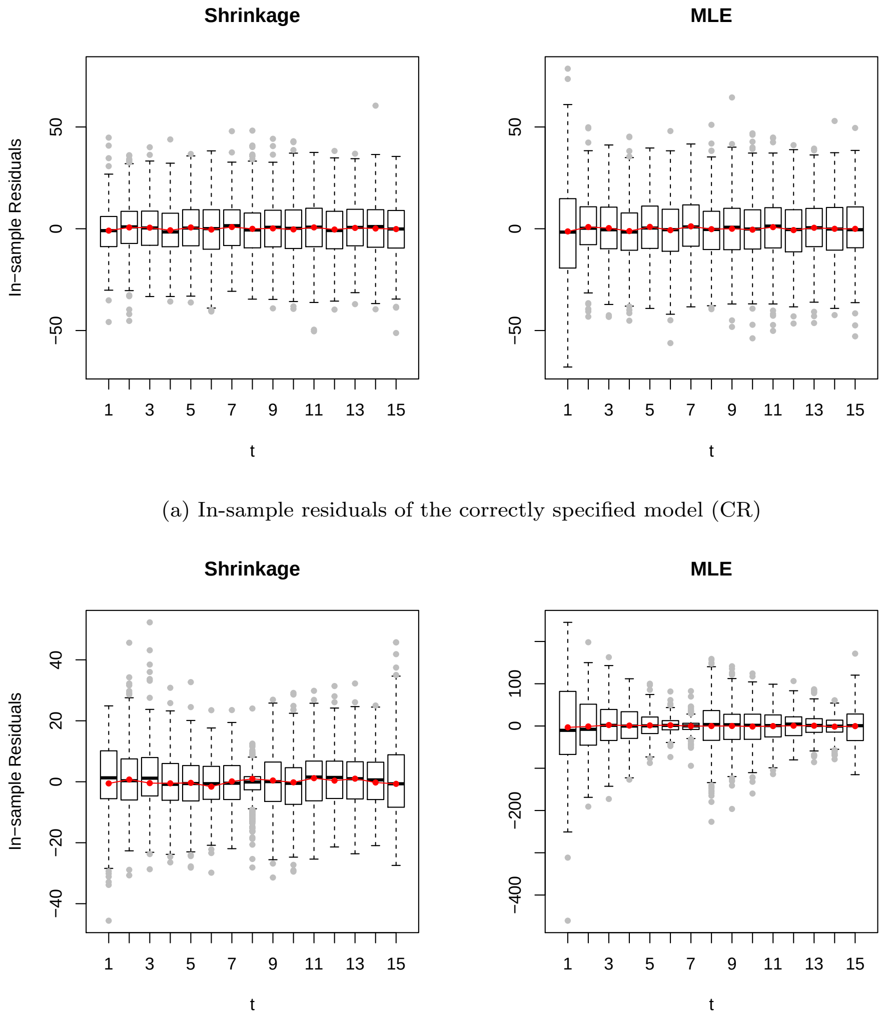
Table 6 demonstrates the performance of the theoretical and the empirical prediction intervals of the shrinkage and the MLE models, for the small sample size and for both measures. For the MLE models, the empirical prediction intervals perform better than the theoretical ones, as the residual assumptions of the MLE models do not hold. This aligns well with the literature that the empirical ones are a robust approach in producing prediction intervals. On the other hand, for the shrinkage models the theoretical prediction intervals perform better than the empirical intervals. In order to explain this finding, we discuss (a) the assumptions of the residuals from the shrinkage models, (b) the distribution of the standard error between the shrinkage and the MLE models, and (c) the distribution of the initial values.

Measure	sMIS				sPinball			
Approach	Shrinkage		MLE		Shrinkage		MLE	
	THE	EMP	THE	EMP	THE	EMP	THE	EMP
80	**1.22**	1.37	2.20	1.76	**0.70**	0.69	1.52	1.10
85	**1.33**	1.55	2.44	1.91	**0.57**	**0.57**	1.27	0.89
90	**1.49**	1.86	2.77	2.14	**0.43**	0.44	0.96	0.65
95	**1.81**	2.75	3.28	2.65	**0.26**	0.29	0.57	0.37
99	**2.91**	9.77	4.34	6.09	**0.07**	0.17	0.15	0.12
Average	**1.75**	3.46	3.01	2.91	**0.41**	0.43	0.90	0.63

Table 6: Overall performance of the prediction interval across scenarios, for small sample size and $h = 1 - 7$. THE and EMP are the theoretical and the empirical prediction intervals. Bold numbers denote the best performing prediction intervals.

We conducted several diagnostics of the residuals, namely checking the normality of the residuals with the Shapiro-Wilk test, especially for the sample sample size, visualising the distribution of the residuals over time, and the summary of autocorrelation testing of the residuals using the partial autocorrelation function (PACF). For the normality test, we recorded the number of instances that the residuals are normally distributed, i.e., when the null hypothesis of normality is not rejected, with a critical level of 5%. We found that for all model specifications the shrinkage models produced normally distributed residuals at 86% of all instances. On the other hand, the MLE models produced normally distributed residuals, at 71% of all instances. Therefore, the proposed estimation results in normally distributed residuals more frequently, compared to MLE, satisfying the assumptions for the theoretical prediction intervals more often, which is reflected in the more accurate prediction intervals reported in Table 6.

In terms of how consistent the variance of the residuals was over time, we examined the boxplot of the residuals over time in the cases of CR and IE. Figures 5a and 5b show that the residuals of the shrinkage models are less erratic than the residuals of the MLE ones. We also see a significant difference between the first residual boxplot of the models, where the shrinkage model exhibits a narrower distribution than that of the MLE model. Overall, the residuals of the shrinkage models seem to have a more stable variance over time.



(a) In-sample residuals of the correctly specified model (CR)

(b) In-sample residuals of the incorrect error and seasonality model (IE)

Figure 5: Residual diagnostics. The red dots denote the mean of each distribution and the grey dots denote the outliers of the distribution.

Across all model specifications, we observe that the shrinkage models produce fewer autocorrelated residuals than the MLE models do, as shown in Figure 6. We argue that the residuals of the shrinkage models meet the assumptions better than the MLE models. Thus, it is sensible to use the theoretical prediction intervals when we have linear homoscedastic shrinkage models.

Next, we discuss the standard error of ETS with Shrinkage and MLE. Figure 7 presents the distribution of the standard error for each model specification. In general, we observe that the shrinkage models produce lower standard errors than the MLE ones. The shrinkage models have a better fit than the MLE, especially when the model structure is wrong. Shrinkage not only improves the out-of-sample performance but also the in-sample performance. Looking at OV where some states are redundant, the shrinkage is able to reduce the standard error substantially. Linking to Eq. (9), we argue that the model bias should be non-zero, but the

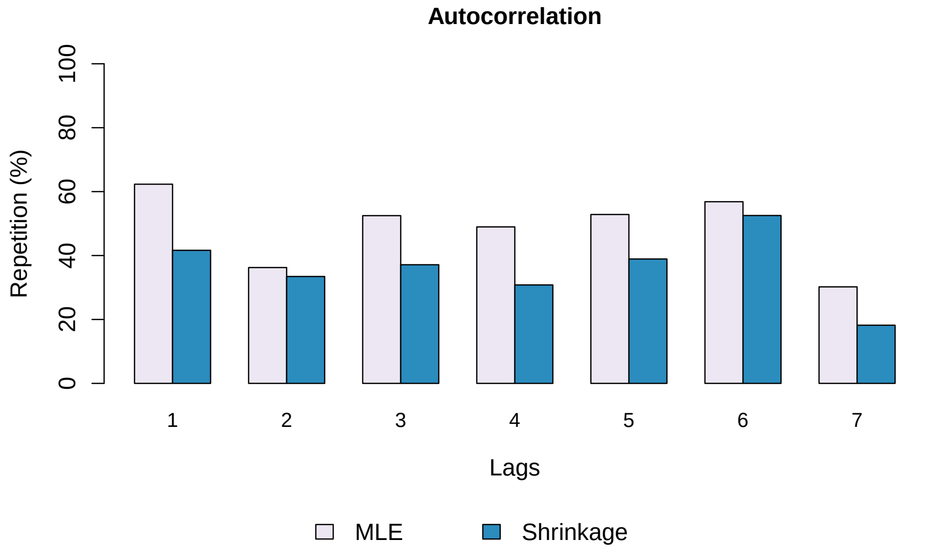


Figure 6: Number of instances that contains autocorrelated residuals

model variance decreases due to the smaller smoothing parameters and a stronger covariance between the smoothing parameters and the initial values. On the other hand, when the model bias is close to zero, i.e., in the case of CR, the standard error reduces moderately, where the effect of shrinkage in CR is not as apparent as in OV or IE.

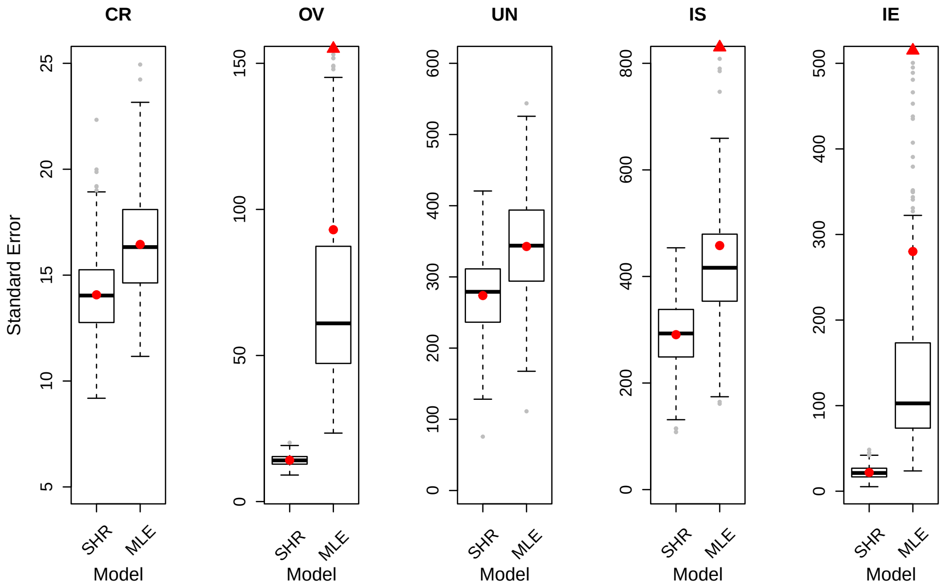


Figure 7: Distributions of standard error between the MLE and the shrinkage models. The red arrows denote that some parts of the boxplot are not plotted and the red dots denote the means of each distribution.

Next we turn our attention to the initial values. Figure 8 presents the standardised level initial values across origins and model specifications. On average the initial values are estimated well, however, there are substantial differences in the spread of the distributions. The variability

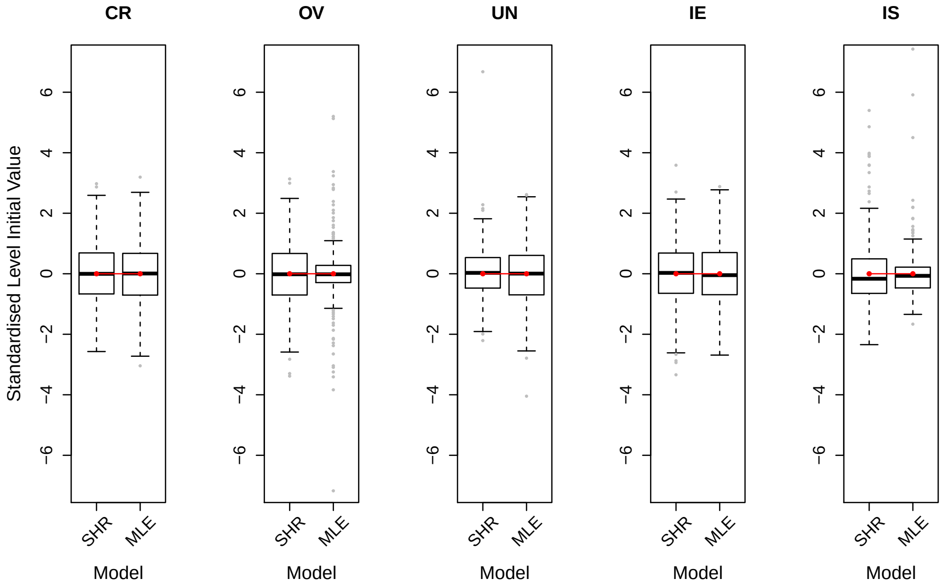
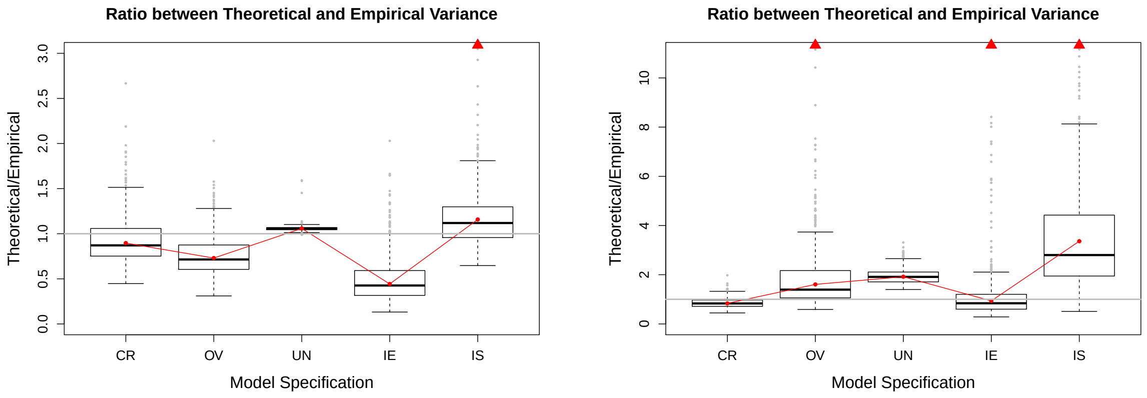


Figure 8: Distributions of the standardised initial values from different model specifications. The red dots denote the mean of each distribution and the grey dots denote the outliers of the distribution.

of the initial value from the MLE model is substantially larger than the one from the shrinkage model. Looking at the OV case, we can see that the MLE produces extreme initial values for many instances. The estimation of the initial values is affected by the shrinkage of the smoothing parameters. We show that this often lends to better estimated initial values and result in lower standard errors and consequently better performing prediction intervals.

Lastly, we investigate whether the theoretical conditional forecast variance can be a good approximation to the empirical conditional forecast variance for shrinkage models. We expect that these two to provide similar results, i.e., the theoretical does not overestimate or underestimate the observed variance. The theoretical variance is computationally more efficient and useful when there is only limited sample. For this analysis, we include more forecast origins (from 5 to 30) to aid approximating the empirical variance. Hence, we also increase the sample size from 28 to 63. We calculate the variance of the 7-step ahead forecast using its RMSE from all 30 origins to estimate the conditional variance of the point forecast in the out-of-sample, giving the empirical conditional variance. Figure 9 presents the ratio between the theoretical and the empirical conditional variance. For the shrinkage models, on average the theoretical variance matches the empirical conditional variance. For severely mis-specified models, the theoretical variance underestimates the empirical conditional variance. As for the MLE models, the theoretical variance significantly overestimates the empirical one, for the OV, UN, and IS cases. This shows that in general, with shrinkage we can approximate the empirical conditional variance with the theoretical variance. With this finding, we argue that we should use the theoretical prediction intervals when we shrink the smoothing parameters.



(a) Shrinkage

(b) Benchmark

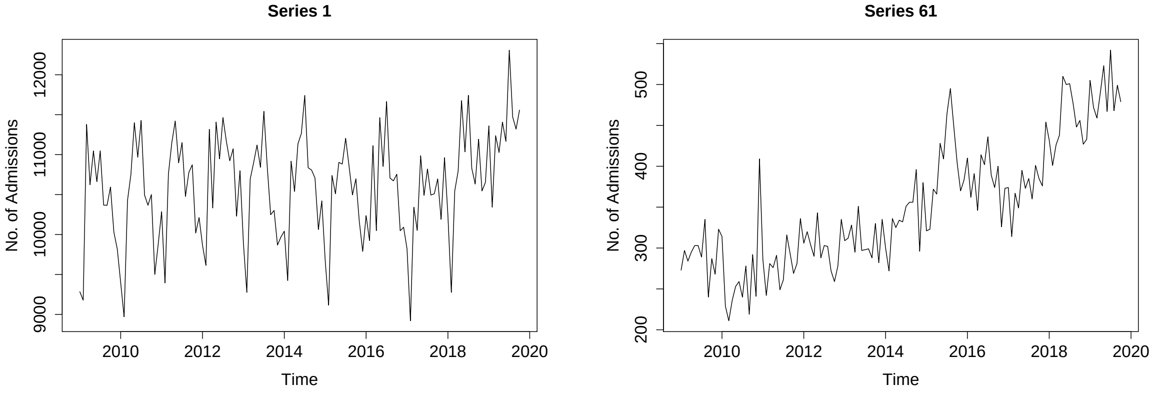
Figure 9: Ratios between the theoretical and the empirical conditional variance of the forecasts, for different levels of model specification. The red dots denote the mean of each distribution, the grey dots denote the outliers of the distribution, and the red arrows denote that some parts of the boxplot are not plotted

Our discussion from the simulation study leans towards the usage of the theoretical prediction interval because (i) the residual assumptions are met more frequently than the MLE; therefore (ii) the theoretical prediction interval is better than the empirical prediction interval for most cases; and (iii) the theoretical variance of the shrinkage is more appropriate than that of the MLE.

4. Application of ETS Shrinkage for the UK NHS

4.1. Experimental Design

In this case study, we apply the proposed estimation procedure to the A&E admissions of a hospital in the northeast of England. The data contains the number of incidents in a day, which is classified by 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 daily time series, but we use only 86 of them because we exclude all time series with zero values. The data spans from January 2009 to October 2019 and we aggregate the time series to the monthly level. To investigate the effect of sample size, we consider two different sizes, namely the case of short time series with 36 observations (November 2015 - April 2018) and the case of long time series with 108 observations (January 2009 - April 2018). The chosen dataset includes types of time series beyond the ones used in the simulations, exhibiting additional complexities due to the nature of the collected data. This enables us to evaluate shrinkage further, with more diverse and complex time series. Figure 10 depicts the two example time series from our data. We observe that some time series are trended and have seasonality. For each sample size, we apply rolling-origin with 5 origins to produce 1 to 12-steps ahead forecasts, with the same model structure. We use the accuracy measures introduced in Section 3. However, if the model is multiplicative or mixed, we produce a simulated-based prediction interval, instead of the analytical ones, because the latter might not be available for such models.



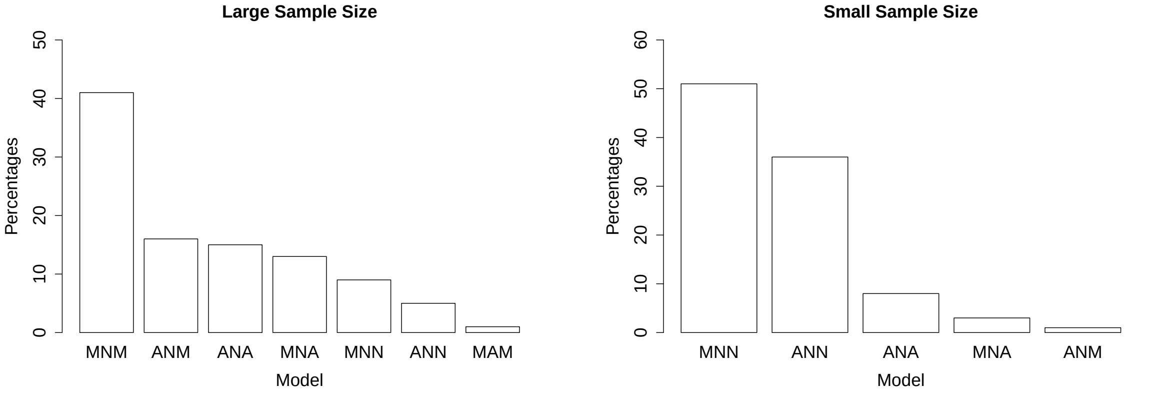
(a) Time series 1

(b) Time series 61

Figure 10: Examples of the time series from the A&E NHS dataset

We demonstrate the effect of shrinkage by comparing models with and without it. Before applying shrinkage, we determine the structure of the model using automatic selection based on the corrected Akaike Information Criteria (AICc). Then, we use the MLE and the shrinkage to estimate the smoothing parameters and the initial values.

4.2. Findings



(a) 2009-2018

(b) 2015-2018

Figure 11: Numbers of model structures for both sample sizes, in percentages

Prior to analysing the performance of the proposed estimation approach, we present the identified models. We identified a mixed pool of models, such as additive, multiplicative, and mixed models. The distribution of model types is shown in Figure 11. For both sample sizes, we can see that multiplicative models dominate. When the sample size is larger, there are more seasonal models.

Sample	2009-2018				2015-2018			
Measure	dRMSE		dAME		dRMSE		dAME	
Type	US	WS	US	WS	US	WS	US	WS
$h = 1$	10.052	**10.344**	-6.210	**-5.264**	8.741	**9.639**	**-1.030**	-3.247
$h = 1 - 6$	**5.655**	5.634	**0.442**	0.363	**4.531**	3.564	**6.619**	4.367
$h = 1 - 12$	**5.639**	5.613	**9.056**	9.018	**4.404**	3.483	**7.762**	5.504

Table 7: A summary of dRMSE and dAME for all time series with five origins. US and WS are the unweighted and the weighted shrinkage, and MLE is the benchmark model. A bold number demonstrates the lowest among the others.

Table 7 present the summary of performances of approaches in terms of dRMSE and dAME for different forecast horizons. For dRMSE, we see that the shrinkage improves the forecast accuracy across all forecast horizons. The performance differences between US and WS become marginal when the sample size is large. However, when the sample size is small, the performance differences become more pronounced. Overall WS performs well when $h = 1$ and US performs well for longer forecast horizons. For dAME, our empirical findings align well with our simulation findings: the forecasts are biased for short horizons, but they become less biased for longer ones. The improvement is substantial for $h = 1 - 12$, namely 9% and 7% for the large and the small sample size, respectively. Regardless of the type of shrinkage, on average, it improves forecast accuracy and forecast bias.

In addition to the forecast accuracy, we also measure the performance of the prediction intervals, presented in Table 8. The unweighted shrinkage produces more precise prediction intervals than the rest, even though the difference with the weighted are marginal. Also note that the theoretical prediction intervals from the shrinkage models perform better than the theoretical and the empirical intervals from the MLE, which aligns with our findings in Section 3. A potential issue might arise from ETS(A,N,M), where the forecast variance might be infinite (Hyndman et al., 2008b, p. 257). We find that in our case study the forecast variances are finite and Table 8 demonstrates that the prediction interval of the models estimated using the proposed shrinkage approach are adequate.

		US		WS		MLE	
Measure	Interval	THE	EMP	THE	EMP	THE	EMP
			2015-2018				
sMIS	80	**0.652**	0.704	**0.652**	0.704	0.709	0.768
sMIS	85	**0.702**	0.773	**0.702**	0.773	0.769	0.841
sMIS	90	**0.776**	0.876	**0.776**	0.876	0.856	0.946
sMIS	95	**0.910**	1.089	0.911	1.090	1.011	1.169
sMIS	99	**1.306**	2.466	1.313	2.472	1.440	2.637
sPinball	80	**0.705**	0.735	0.705	0.735	0.830	0.834
sPinball	85	**0.589**	0.613	**0.589**	0.613	0.697	0.688
sPinball	90	**0.449**	**0.470**	0.449	0.470	0.535	0.516
sPinball	95	**0.271**	0.290	0.272	0.290	0.327	0.310
sPinball	99	**0.078**	0.112	0.079	0.112	0.095	0.106

Table 8: Performances of the prediction interval, confidence levels, type of shrinkage, and prediction interval calculation for 1-12 step-ahead forecasts and small sample size. THE and EMP are the theoretical and the empirical prediction intervals. Bold numbers denote the best performing prediction intervals.

5. Conclusion

To reach reliable decisions reliable forecasts are needed. ETS has been a widely used forecasting model, providing reliable and robust forecasts. The current methodology utilises the maximum likelihood to estimate the smoothing parameters, the initial values, and the dampening parameter. However, with limited sample, the estimation of the smoothing parameters becomes unreliable. On the other hand, there is evidence that low smoothing parameters may reduce the forecast error. We propose to estimate the smoothing and the dampening parameters using shrinkage. However, the shrinkage in ETS differs from regression, as it does not eliminate model states, but instead reinforces deterministic patterns.

Using simulation and a real case study, we find the proposed shrinkage produces more accurate and less biased forecasts. We also find that its theoretical prediction intervals perform best, compared to empirical prediction intervals. Using the bias-variance decomposition we demonstrate that the smoothing parameter shrinkage and the dependent initial value estimation affect the in-sample standard error, and result in improved prediction intervals. Thus, we emphasise the importance of the initial value estimation, especially when the shrinkage approach is used. However, we do not shrink the initial values. The effect of alternative estimators for the initial values remains interesting for future research, as it has the potential to enable shrinkage to model select as well, when an ℓ_1 penalty is used. By setting to zero both the smoothing parameter and the initial value of a state of the ETS model, it could simplify to a smaller model. To achieve this, future work should resolve issues with the scaling of the various parameters to be shrunk efficiently.

Although we demonstrate the benefits of shrinkage on real time series that can contain various artefacts, our evaluation is not exhaustive. For instance, we do not investigate the effectiveness of shrinkage in the presence of structural breaks, as these may introduce the need for specific treatments (e.g., adding indicator variables), and lead to modelling questions that are beyond the focus of this work.

Moreover, the shrinkage estimators investigated here have parallels with other recent contributions in the literature for mitigating parameter estimation issues, for instance, by limiting the pool of exponential smoothing models (Kourentzes et al., 2019; Meira et al., 2021), aiming to reduce the instability of forecasts. Future research should investigate these approaches for complementarities.

6. Acknowledgement

We would like to thank Juan Ramon Trapero for his constructive feedback. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. We declare no conflicts of interest.

References

Assimakopoulos, V., Nikolopoulos, K., 2000. The theta model: A decomposition approach to forecasting. International Journal of Forecasting 16, 521–530.

Barrow, D., Kourentzes, N., Sandberg, R., Niklewski, J., 2021. Automatic robust estimation for exponential smoothing: Perspectives from statistics and machine learning. Expert Systems with Applications 160.

Bergstra, J., Bergstra, J., Bengio, Y., 2012. Random search for hyper-parameter optimization. Journal of Machine Learning Research 13, 281–305.

Bunn, D.W., Vassilopoulos, A.I., 1999. Comparison of seasonal estimation methods in multi-item short-term forecasting. International Journal of Forecasting 15, 431–443.

Chatfield, C., 2001. Time Series Forecasting. Chapman & Hall/CRC, Boca Raton.

Chen, M., Chen, Z.L., 2018. Robust dynamic pricing with two substitutable products. Manufacturing & Service Operations Management 20, 249–268.

Daniels, M.J., Kass, R.E., 2001. Shrinkage estimators for covariance matrices. Biometrics 57, 1173–1184.

Fildes, R., Hibon, M., Makridakis, S., Meade, N., 1998. Generalising about univariate forecasting methods: Further empirical evidence. International Journal of Forecasting 14, 339–358.

Gardner, E.S., 2006. Exponential smoothing: The state of the art — part II. International Journal of Forecasting 22, 637–666.

Gardner, E.S., McKenzie, E., 1985. Forecasting trends in time series. Management Science 31, 1237–1246.

Gneiting, T., Raftery, A.E., 2007. Strictly proper scoring rules, prediction, and estimation. Journal of the American Statistical Association 102, 359–378.

Greis, N.P., Gilstein, C.Z., 1991. Empirical bayes methods for telecommunications forecasting. International Journal of Forecasting 7, 183.

Hastie, T., Tibshirani, R., Wainwright, M., 2015. Statistical Learning with Sparsity the Lasso and Generalizations. Taylor & Francis Group.

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

Hyndman, R.J., Akram, M., Archibald, B.C., 2008a. The admissible parameter space for exponential smoothing models. Annals of the Institute of Statistical Mathematics 60, 407–426.

Hyndman, R.J., Billah, B., 2003. Unmasking the theta method. International Journal of Forecasting 19, 287–290.

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.

Hyndman, R.J., Koehler, A.B., Ord, J.K., Snyder, R.D., 2008b. Forecasting with Exponential Smoothing: The State Space Approach. Springer.

Johnson, S.G., 2021. The NLopt Nonlinear-optimization package. URL: `http://ab-initio.mit.edu/nlopt`.

Johnston, F.R., Boylan, J.E., 1994. How far ahead can an ewma model be extrapolated? The Journal of the Operational Research Society 45, 710–713.

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.

Koenker, R., Bassett, G., 1978. Regression quantiles. Econometrica 46, 33–50.

Kourentzes, N., Barrow, D., Petropoulos, F., 2019. Another look at forecast selection and combination: Evidence from forecast pooling. International Journal of Production Economics 209, 226–235.

Kourentzes, N., Petropoulos, F., Trapero, J.R., 2014. Improving forecasting by estimating time series structural components across multiple frequencies. International Journal of Forecasting 30, 291–302.

Lee, Y.S., Scholtes, S., 2014. Empirical prediction intervals revisited. International Journal of Forecasting 30, 217–234.

Makridakis, S., Hibon, M., 2000. The M3 competition: Results, conclusions and implications. International Journal of Forecasting 16, 451–476.

Makridakis, S., Spiliotis, E., Assimakopoulos, V., 2018. The M4 competition: Results, findings, conclusion and way forward. International Journal of Forecasting 34, 802–808.

Makridakis, S., Spiliotis, E., Assimakopoulos, V., 2021. Predicting/hypothesizing the findings of the M5 competition. International Journal of Forecasting .

Meira, E., Luiz, F., Oliveira, C., Jeon, J., 2021. Treating and pruning: New approaches to forecasting model selection and combination using prediction intervals. International Journal of Forecasting 37, 547–568.

Miller, D.M., Williams, D., 2003. Shrinkage estimators of time series seasonal factors and their effect on forecasting accuracy. International Journal of Forecasting 19, 669–684.

Miller, D.M., Williams, D., 2004. Damping seasonal factors: Shrinkage estimators for the X-12-ARIMA program. International Journal of Forecasting 20, 529–549.

Ord, J.K., Fildes, R., Kourentzes, N., 2017. Principles of business forecasting. Second edition. ed., Wessex Press Inc., New York.

R Core Team, 2021. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria. URL: `https://www.R-project.org/`.

Sagaert, Y.R., Kourentzes, N., Vuyst, S.D., Aghezzaf, E.H., Desmet, B., 2019. Incorporating macroeconomic leading indicators in tactical capacity planning. International Journal of Production Economics 209, 12–19.

Sanders, N.R., Graman, G.A., 2009. Quantifying costs of forecast errors: A case study of the warehouse environment. Omega 37, 116–125.

Schäfer, J., Strimmer, K., 2005. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. Statistical Applications in Genetics and Molecular Biology 4, 1–30.

Sillanpää, V., Liesiö, J., 2018. Forecasting replenishment orders in retail: Value of modelling low and intermittent consumer demand with distributions. International Journal of Production Research 56, 4168–4185.

Snyder, R.D., 1985. Recursive estimation of dynamic linear models. Journal of the Royal Statistical Society. Series B, Methodological 47, 272–276.

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

Svetunkov, I., 2022. Forecasting and analytics with ADAM. OpenForecast. URL: `https://openforecast.org/adam/`. (Version: 18th February 2022).

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B, Methodological 58, 267–288.

Trapero, J.R., Cardós, M., Kourentzes, N., 2019. Empirical safety stock estimation based on kernel and GARCH models. Omega 84, 199–211.

Trapero, J.R., Kourentzes, N., Fildes, R., 2015. On the identification of sales forecasting models in the presence of promotions. Journal of the Operational Research Society 66, 299–307.

Wickramasuriya, S.L., Athanasopoulos, G., Hyndman, R.J., 2019. Optimal forecast reconciliation for hierarchical and grouped time series through trace minimization. Journal of the American Statistical Association .

Zou, H., 2006. The adaptive lasso and its oracle properties. Journal of the American Statistical Association 101, 1418–1429.