Mitigating parameter uncertainty in business forecasting

Kandrika Fadhlan Pritularga

Submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy at the Department of Management Science, Lancaster University

December, 2022

Abstract

Organisations make multiple decisions, and each layer requires different types of information. The main task of forecasting in these different situations is to support each decision with relevant future information in point and interval forecasts. The consequence is that we have multiple and correlated time series. To produce point and interval forecasts for multiple decisions, we have three modeling options, namely:

- modeling each time series with a set of univariate Exponential Smoothing methods,
- modeling all time series with a Vector Exponential Smoothing model, and
- utilising state-of-the-art forecast reconciliation.

Each option has the same idea: to approximate the 'true' data-generating process. Consequently, we have uncertainties around each modeling option, namely (a) model structure, (b) parameter, and (c) sampling uncertainty. The literature mainly focuses on mitigating the model structure uncertainty, which is believed to harm forecast accuracy significantly. On the other hand, this thesis mitigates the parameter uncertainty in each modeling case (Exponential Smoothing, Vector Exponential Smoothing, and Forecast Reconciliation). We propose parameter shrinkage in each modeling option. Specifically, we propose a shrinkage estimator for the univariate and the multivariate exponential smoothing. We also suggest forcing some covariances to zero to mitigate the covariance matrix estimation uncertainty in the forecast reconciliation.

Our study relies on theoretical investigations, simulation, and empirical studies. The theoretical analysis provides solid and rational arguments to mitigate the parameter uncertainty. We complement it with empirical findings, where the difference between the simulation and the empirical study is how much we can control the experimental designs. We also ensure that each design follows sound principles of forecasting evaluation.

Our findings show that the shrinkage estimator improves forecast accuracy. However, the results are mixed for the Vector Exponential Smoothing. We also find that forcing some covariances in the covariance matrix approximation improves both the forecast accuracy and the variability of the forecasting performance. By understanding the parameter uncertainty, we find important correlations between parameters that may affect forecast accuracy. We also propose the concept of stochastic coherency to encapsulate the overlooked uncertainties in forecast reconciliation.

Our thesis emphasises the importance of revisiting uncertainty in business forecasting. We decipher it via the bias-variance decomposition and understand how the interdependence between parameters affects our understanding of the uncertainty. It is not only essential to address each uncertainty individually but also to address all uncertainties comprehensively. In particular, we propose different types of parameter shrinkage. The implementation depends on whether we have sufficient information to estimate parameters in the model. In the univariate case, the parameters' estimates tend to be inefficient when the sample size is limited. In the multivariate case, either the shrinkage estimator or forcing some parameters to zero by design is also a potential solution to the problem. These forms of shrinkage avoid overfitting and potentially improve foecast accuracy.

Concerning decision makers, our understanding of uncertainty highlights the importance of reliability in forecasting, i.e., unmitigated parameter uncertainty results in unreliable forecasting performance. This reliability is essential to gain the decision-maker's trust in our forecasts. It is a new business forecasting concept and is open to investigation.

Contents

List of Tables

List of Figures

Acknowledgement

Words cannot express my gratitude to my supervisors, Dr. Ivan Svetunkov and Prof. Nikolaos Kourentzes, for their invaluable time, effort, and patience in supervising me. Your guidance and mentorship are what education should be. I would like to thank all members of the Centre of Marketing Analytics and Forecasting for a lovely and warm research community. This endeavour would not have been possible without the generous support from the Department of Management Science, Lancaster University, who funded my doctoral training.

I am also grateful to my friends, especially Anna Sroginis, for her invaluable academic and mental health supports. I would also thank to Margono Robertus Robert and Cesty Nur Tribuana for keeping me on the ground to enjoy my life a little longer.

I would be remiss in not mentioning my strong and independent mother for her moral and academic supports. Lastly, I present this thesis for my beloved late sister, who always supported my academic journey until the end of her life. This is for you, Mbak Icha :)

Declaration

This thesis is my own work and it has not been submitted in support of an application for another higher degree or qualification elsewhere.

Kandrika Fadhlan Pritularga

Certainty is inseparable from emptiness:

there is no synthetic a priori.

Hans Reichenbach, 1968

Chapter 1

Introduction

1.1 Motivation and Background

Organisations make decisions to achieve their goals. We can classify these decisions according to the organisational hierarchy or functions, product categories, market segments, etc. (Hyndman et al., 2011; Athanasopoulos et al., 2017; Kourentzes and Athanasopoulos, 2019; Oliveira and Ramos, 2019; Babai et al., 2022). Each of these decisions requires different information. For example, the strategic level requires macro-level information (national sales, macroeconomic indicators). In contrast, the tactical level requires business unit or store-related information (financial performance of each business unit or branch). The main task of forecasting in these different situations is to support each decision with relevant future information, especially in projecting the most likely future (point forecasts) and reducing future uncertainty (interval forecasts) amid uncertainties in the market (Ord et al., 2017).

As supporting multiple decisions in any organisation, forecasters need to deal with many time series and produce multiple point and interval forecasts for each decision. A straightforward approach is to model each time series independently using a univariate statistical time series model. Among them, Exponential Smoothing is one of the most popular methods because it is easy to implement, intuitive, robust, and performs well in practice (Makridakis and Hibon, 2000; Makridakis et al., 2018, 2021). Exponential Smoothing was initially developed by Brown (1956) in the form of a Simple Exponential Smoothing, while Holt (2004) extended it to include trend and seasonality (Winters, 1960). Snyder (1985) proposed a Single Source of Error (SSOE) state-space model, underlying Simple Exponential Smoothing. Ord et al. (1997), Hyndman et al. (2002) and Hyndman et al. (2008b) improved the methodology by developing a new taxonomy that they called 'ETS' or Error-Trend-Seasonality. They used the maximum likelihood to estimate parameters in Exponential Smoothing models. By doing so, they could automate the modelling process by estimating the model parameters and selecting the best-approximating model via minimising an information criterion.

In practice, correlated time series are common (Ma et al., 2016). Implementing ETS in a set of correlated time series is computationally inexpensive, at the cost of losing the interdependence information. A potential approach to incorporate interdependence is to employ a multivariate forecasting model. Such models capture two types of interdependence: (a) functional and (b) covariance, where the former demonstrates the functional relationship between time series and the latter shows the correlations between errors. A multivariate extension of Exponential Smoothing was initially proposed by Jones (1966). Duncan and Horn (1972) developed it further with a Bayesian approach. de Silva et al. (2010) propose a vector exponential smoothing (VES) within the SSOE framework as the extension of the ETS methodology. Like any multivariate model, it is likely to suffer from 'the curse of dimensionality', where the number of parameters may exceed the number of observations. Consequently, the model may be estimable but suffer from overfitting as the number of observations might not be sufficient to obtain efficient estimates of parameters in a statistical sense. Svetunkov et al. (2022a) attempted to solve the dimensionality problems by imposing commonality in parameters.

As incorporating interdependence into a forecasting model is not straightforward, we can split the multivariate estimation into a two-step estimation via forecast reconciliation (Athanasopoulos et al., 2009; Hyndman et al., 2011, 2016; Wickramasuriya et al., 2019; Panagiotelis et al., 2022). First, we generate independent unbiased base forecasts from univariate time series models. Second, we combine forecasts using a reconciliation weights matrix that adheres to the linear constraint. This reconciliation lends information from the linear aggregation to improve the forecast accuracy, and it is estimated via an ill-posed least square approach. The weights distribute the 'left-over' information from the univariate models and add 'interdependence' information to the final reconciled forecasts.

Modelling a multivariate time series is not straightforward. We have three approaches to choose from, and there is always a trade-off between these approaches. For example, employing univariate models is easy, but we lose important information. On the other hand, employing a multivariate model captures available information, but it is difficult to estimate and prone to significant estimation errors. We certainly can use forecast reconciliation, but the performance depends on the covariance matrix approximation, and the multivariate connection remains limited, e.g., no lags. Regardless of each approach's advantages and disadvantages, the fundamental idea is to approximate the unknown 'true' data-generating process (DGP). As in practice, the 'true' DGP is unknown; we need to estimate the model structure and parameters, conditional on the data at hand. It significantly affects how we approximate the DGP, namely handling uncertainties.

Chatfield (2000) discussed three fundamental sources of uncertainties in statistical models: (a) model structure, (b) parameter estimation, and (c) sampling one. He argues that the model structure has the most severe effect on forecast accuracy and the others do not. This argument implicitly assumes that the sample size is sufficient to obtain efficient estimates of parameters. In contrast, in business, sample sizes are often limited due to product life cycles or poor data management (Ord et al., 2017). The parameters may be less efficient. It may harm forecast accuracy and make forecasts unreliable, i.e., inconsistent across forecast origins, and eventually make forecasts untrustworthy to the managers (Spavound and Kourentzes, 2022).

1.2 Research Question and Methodology

We discuss three modelling approaches in business forecasting and potential issues due to uncertainties in model estimation. This thesis investigates the effect of parameter uncertainty on forecast performance and proposes parameter shrinkage approaches to mitigate this issue. We address the following overarching research question,

> **Research Question** *How do we mitigate the effect of parameter uncertainty statistical models/ methods in forecasting performance?*

In particular, we investigate this with three modelling approaches: ETS, VES, and Forecast Reconciliation, each corresponding to a chapter in this thesis. Each looks at a different case of parameter uncertainty. We validate our investigation through different approaches, namely theoretical investigations, simulations, and empirical studies (Naylor and Finger, 1967; Kleindorfer et al., 1998; Fildes and Kourentzes, 2011).

We theoretically investigate the overall forecast uncertainty by decomposing it into several components, e.g., uncertainty related to model structure, parameter, and randomness. From this, we can trace the contribution of each component to the overall uncertainty. We also use the bias-variance trade-off (Hastie et al., 2015) to find any potential interaction between parameters that affect the uncertainty and formalise it. These theoretical insights provide a sound basis to propose several parameter shrinkage approaches. We modify the loss function of ETS and VES and propose new covariance matrix approximations relevant to the forecast reconciliation problem.

We substantiate and complement our theoretical insights with simulated and real data experiments. The main difference between both experiments is whether we know and can control the DGP. For the former, first, we set up the DGP. We then implement different models with different specifications, with and without the parameter shrinkage treatment. For example, in Chapter 4, we construct simple autoregressive time series and purposefully model them with an autoregressive model and ETS. That way, we can see the effect of different degrees of uncertainty on forecasting performance. However, we argue that simulated experiments are insufficient as the DGP remains fully controlled and relatively simplistic. We provide real data experiments where we cannot control or know the 'true' DGP. We use an Accident & Emergency Admission dataset for Chapter 2 and Chapter 4. In this experiment, we assume that the model structure is approximated adequately, to isolate the model structure uncertainty. We observe forecast accuracy improvement with parameter shrinkage when we do not know the DGP.

We also ensure that each design follows sound principles of forecasting evaluation (Tashman, 2000). We split the time series in each experiment into a training and a test set. We use the former to estimate the model and the latter to assess the point and interval forecast accuracy for short and longer forecast horizons. The evaluation was a rolling origin scheme. We compare any findings with benchmark models. Our benchmarks are state-of-the-art implementations and, therefore, without any of the proposed shrinkage estimators. For example, in Chapter 2, we use ETS without any parameter shrinkage. We assess the forecasting performance by investigating the forecast error comprehensively. We look at the mean, the mean squared of the forecast error, and the distribution of the mean squared error via boxplots. We attempt to observe the variability of the forecast error variance as a proxy for the reliability of the forecasts (Spavound and Kourentzes, 2022). It allows us to evaluate any proposed methods comprehensively.

1.3 Contributions

Chapter 2 discusses a shrinkage estimator for ETS, where we shrink the smoothing parameters toward zero. By doing that, we can control the stochasticity of the model states, i.e., becoming more or less stochastic. This approach results in not only more accurate point forecasts but also more accurate prediction intervals. We also find a correlation between smoothing parameters and initial values. Lowering the smoothing parameters makes the states 'remember' the past information more, resulting in a more efficient initial value estimation. The estimated initial values become more efficient, which leads to a decrease in the in-sample standard error and improves prediction interval accuracy. It is in contrast to Makridakis and Winkler (1989)'s finding that the initial values do not affect forecasting performance. This chapter is published at *the International Journal of Forecasting* (Pritularga et al., 2022).

Chapter 3 extends the proposed shrinkage estimator for ETS to VES. In this case, we shrink the persistence matrix containing the time series' smoothing parameters. The idea is similar to Chapter 2, where we control the state stochasticity. At the same time, we aim to mitigate the curse of dimensionality issues in VES, which bridges the gap between a highly restrictive VETS by Svetunkov et al. (2022a) and an unrestricted one by de Silva et al. (2010). However, we find that this is not straightforward. The shrinkage can estimate shrink smoothing parameters in the persistence matrix to zero, but this does not necessarily translate to forecast accuracy improvement. We observe a compensating effect between the persistence matrix and the covariance matrix, i.e., the variances and the covariance compensate for the lowered smoothing parameters. As a result, the covariance matrix estimation error becomes unnecessarily large in some instances. We propose to develop an extended shrinkage estimator that includes the persistence matrix and the covariance matrix, similar to Wilms and Croux (2018) that developed the shrinkage estimator for the vector autoregressive model. Our findings provide a strong argument as to why we need to include the covariance matrix in the shrinkage estimator.

Chapter 4 explores the effect of parameter uncertainty on forecast reconciliation and proposes the concept of 'stochastic coherency' in forecast reconciliation. We find that forecast reconciliation improves the point forecast accuracy, on average, as suggested by Wickramasuriya et al. (2019), and validated in other works, e.g., Oliveira and Ramos (2019). However, we also find variability in the error measures. The variability originates from the reconciliation weights matrix uncertainty, and the weights uncertainty originates from the covariance matrix estimation error and the forecasting model uncertainty. We see an amplification effect from the forecasting model to the final reconciled forecasts. Stochastic coherency emphasises that the notion of coherency is subject to uncertainty, and forecasters should be aware of that. Practically speaking, we need slack to accommodate the discrepancy between levels of aggregation due to measurement uncertainty. The concept accommodates overlooked uncertainties in forecast reconciliation. In light of such uncertainty, we also propose several alternatives to the covariance matrix approximations, i.e., to force some covariances to zero to mitigate the covariance matrix estimation error. This chapter is published *at the International Journal of Production Economics* (Pritularga et al., 2021).

We should handle uncertainties in any statistical method/ model comprehensively. A well-known approach that is believed to mitigate overall uncertainty is the linear combination of forecasts. It could be in the form of forecast combination (Bates and Granger, 1969; Clemen, 1989; Kourentzes et al., 2014; Barrow and Kourentzes, 2016), the Multi Aggregation Prediction Algorithm (Kourentzes et al., 2014, 2017), or forecast reconciliation (Athanasopoulos et al., 2009; Hyndman et al., 2011; Athanasopoulos et al., 2017; Jeon et al., 2019; Wickramasuriya et al., 2019; Kourentzes and Athanasopoulos, 2019; di Fonzo and Girolimetto, 2021; Panagiotelis et al., 2022). It is a post-processing approach, where first, we estimate the forecasting models and use some linear combination approaches to combine the forecasts. It is useful, but if the combination is not done properly, it can harm the forecast accuracy (Smith and Wallis, 2009; Claeskens et al., 2016). In light of that, some studies attempt to mitigate the uncertainty during the modelling process. For example, Burnham and Anderson (2002) and Hyndman et al. (2008b) select the best-approximated model by minimising an information criterion and therefore mitigate the model structure uncertainty prior to any combination. Similarly, one can eliminate forecasts from the combination pool (Kourentzes et al., 2019). However, the parameter uncertainty is not explicitly targeted. Our contribution is to fill this gap, where we mitigate the parameter uncertainty in the modelling process. We propose parameter shrinkage approaches and provide strong evidence that a shrinkage estimator can be a sensible solution beyond its standard regression use in the univariate (see Chapter 2) and multivariate context (see Chapter 3 and 4).

1.4 Structure of the Thesis

The remainder of this thesis is structured as follows. The next chapter presents the shrinkage estimator for ETS, and Chapter 3 extends its implementation to VES models. In Chapter 4, we propose the notion of stochastic coherency to accommodate overlooked uncertainties in hierarchical forecasting and design alternatives to the covariance matrix estimation approximations. We discuss and conclude our thesis in Chapter 5.

Chapter 2

Shrinkage Estimator for Exponential Smoothing Models

In this chapter we propose a new estimator for exponential smoothing models to mitigate the parameter uncertainty. This results in more reliable point and interval forecasts. All materials in this chapter are based on an article published at the International Journal of Forecasting (Pritularga et al., 2022).

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 in-

terdependence between smoothing parameters and initial values and the importance of the initial value estimation on point forecasts and prediction intervals.

2.1 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 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 2.2, we present the proposed shrinkage for exponential smoothing. Section 2.3 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 2.4 and conclude in Section 2.5.

2.2 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):

$$y_t = w(\boldsymbol{x}_{t-1}) + r(\boldsymbol{x}_{t-1})\varepsilon_t, \tag{2.1}$$

$$\boldsymbol{x}_t = f(\boldsymbol{x}_{t-1}) + g(\boldsymbol{x}_{t-1})\varepsilon_t, \tag{2.2}$$

where Eq. (2.1) and Eq. (2.2) 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, 2022a). 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. Those parameters update new information coming into the model states. 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)

$$\ell^*(\boldsymbol{\theta}, \boldsymbol{x}_0|\mathcal{I}_t) = -\frac{T}{2}\log\left(2\pi e\frac{1}{T}\sum_{t=1}^{T}{\varepsilon_t}^2\right) - \sum_{t=1}^{T}\log|r(\boldsymbol{x}_{t-1})|, \tag{2.3}$$

where $\boldsymbol{\theta} = \{\boldsymbol{g}, \phi\} = \{\alpha, \beta, \gamma, \phi\}$, T is the number of observations, and $\mathcal{I}_t$ 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. (2.3), 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:

$$\text{MSE}(\boldsymbol{\theta}, \boldsymbol{x_0}|\mathcal{I}_t) = \frac{1}{T}\sum_{t=1}^{T}\varepsilon_t^2. \tag{2.4}$$

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. (2.4), 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,

$$\text{MSE}(\boldsymbol{\theta}, \boldsymbol{x_0}|\mathcal{I}_t) + \lambda^* p(\boldsymbol{\theta})\mathbf{1}, \tag{2.5}$$

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. (2.5) so that the shrinkage hyper-parameter has a finite upper bound:

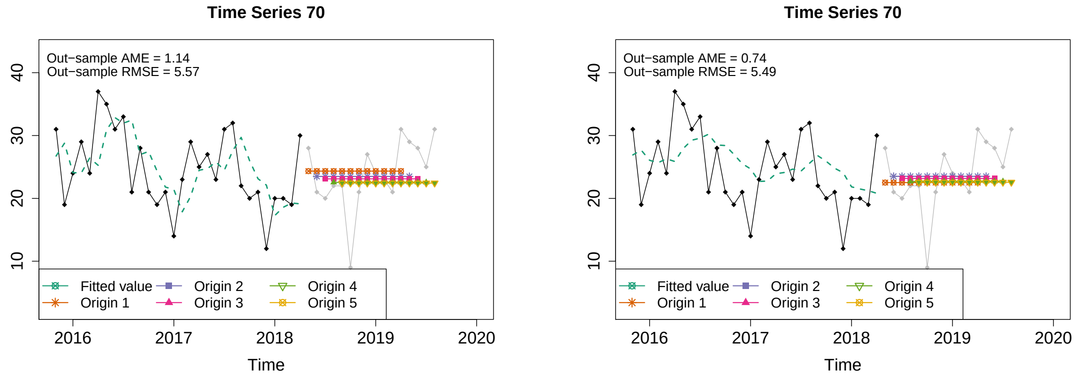
$$(1-\lambda)\text{MSE}(\boldsymbol{\theta}, \boldsymbol{x_0}|\mathcal{I}_t) + \lambda p(\boldsymbol{\theta})\mathbf{1}, \tag{2.6}$$

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. (2.6) 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. (2.6) 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 2.4) 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 2.1: Examples of different estimation approaches for 5 origins.

Figure 2.1 demonstrates the effect of the smoothing parameter shrinkage on the single-state in the fitted value in-sample and the out-of-sample forecasts. RMSE is the root mean squared error and AME is the absolute mean error. Both error measures calculates the accuracy of 1-12 step ahead 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.

2.2.1 Weighted Shrinkage

Eq. (2.6) 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. (2.6) 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,

$$\{\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{x}}_0\} = \underset{\theta, \boldsymbol{x}_0}{\arg\min} \left((1-\lambda)\mathrm{MSE}(\boldsymbol{\theta}, \boldsymbol{x_0}|\mathcal{I}_t) + \lambda p(\boldsymbol{\theta})\boldsymbol{\omega}\right), \tag{2.7}$$

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.

2.2.2 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:

$$\mathrm{V}(y_{t+h|t}) = \begin{cases} \sigma^2, & \text{if } h = 1 \\ \sigma^2 \left[1 + \sum_{j=1}^{h-1} c_j^2\right], & \text{if } h \geq 2 \end{cases}, \tag{2.8}$$

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 transition matrix or $\boldsymbol{F}$. This matrix shows the relationship between the current and the previous states. 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.

$$\mathrm{E}\left(y_t - \hat{y}_{t|t-1}|\mathcal{I}_t\right)^2 = \underbrace{\mathrm{E}\left(\boldsymbol{\mu}_t - \mathrm{E}\left(\hat{y}_{t|t-1}\right)|\mathcal{I}_t\right)^2}_{\text{model bias}} + \underbrace{\mathrm{E}\left(\mathrm{E}\left(\hat{y}_{t|t-1}\right) - \hat{y}_{t|t-1}|\mathcal{I}_t\right)^2}_{\text{model variance}} + \underbrace{\sigma^2}_{\substack{\text{irreducible} \\ \text{variance}}}, \tag{2.9}$$

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,

$$\hat{y}_{t|t-1} = \boldsymbol{w}^\top \boldsymbol{F}^{t-1}\hat{\boldsymbol{x}}_0 + \sum_{i=1}^{t-1} \boldsymbol{w}^\top \boldsymbol{F}^{t-i-1}\hat{\boldsymbol{g}} e_{i|i-1}, \tag{2.10}$$

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, $\mathrm{E}(\hat{y}_{t|t-1}|\mathcal{I}_{x_0}) = \boldsymbol{\mu}_t = \boldsymbol{w}^\top \boldsymbol{F}^{t-1}\boldsymbol{x}_0$, where $\mathcal{I}_{x_0}$ is the information related to $\boldsymbol{x}_0$. Since $\mathrm{E}(\hat{\boldsymbol{x}}_0|\mathcal{I}_{x_0}) = \boldsymbol{x}_0$, thus the model bias vanishes. However, when the model is incorrectly specified, $\mathrm{E}(\hat{\boldsymbol{x}}_0|\mathcal{I}_{x_0}) \neq \boldsymbol{x}_0$, hence the model bias should be non-zero.

To dissect the model variance, we insert Eq. (2.10) into the model variance in Eq. (2.9) and the decomposition is formulated as:

$$\begin{aligned}\mathrm{E}\left(\mathrm{E}\left(\hat{y}_{t|t-1}\right) - \hat{y}_{t|t-1}|\mathcal{I}_t\right)^2 = \mathrm{E}\left(\boldsymbol{w}^\top \boldsymbol{F}^{t-1}(\boldsymbol{x}_0 - \hat{\boldsymbol{x}}_0)|\mathcal{I}_t\right)^2 + \mathrm{E}\left(\sum_{i=1}^{t-1} \boldsymbol{w}^\top \boldsymbol{F}^{t-i-1}\hat{\boldsymbol{g}} e_{i|i-1}|\mathcal{I}_t\right)^2 \\ - 2\mathrm{cov}\left(\boldsymbol{w}^\top \boldsymbol{F}^{t-1}(\boldsymbol{x}_0 - \hat{\boldsymbol{x}}_0), \sum_{i=1}^{t-1} \boldsymbol{w}^\top \boldsymbol{F}^{t-i-1}\hat{\boldsymbol{g}} e_{i|i-1}|\mathcal{I}_t\right).\end{aligned} \tag{2.11}$$

Eq. (2.11) 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 (2.8), 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, 2000). 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).

2.2.3 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,

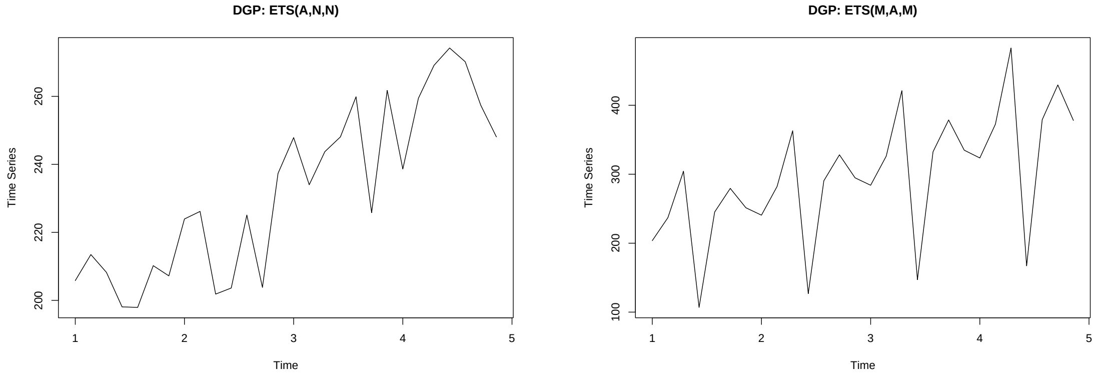
$$\{\hat{\lambda}, \hat{\boldsymbol{\omega}}\} = \arg\min_{\lambda, \boldsymbol{\omega}} \frac{1}{K} \sum_{i=1}^{K} \text{MSE}(\lambda, \boldsymbol{\omega}, \hat{\boldsymbol{\theta}}_i, \hat{\boldsymbol{x}}_{i,0}, y_{1:i}), \tag{2.12}$$

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. (2.12). 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.

2.3 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 one (ℓ_1-US), (b) the ℓ_1 weighted one (ℓ_1-WS), (c) the ℓ_2 unweighted one (ℓ_2-US), and (d) the ℓ_2 weighted one (ℓ_2-WS). US and WS denote the unweighted and the weighted shrinkage estimators.

2.3.1 Experimental Design



(a) DGP: ETS(ANN)

(b) DGP: ETS(MAM)

Figure 2.2: 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 2.1. Values in Table 2.1 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. In addition, the seasonal initial values are generated randomly. Figure 2.2 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 red (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 2.1: A summary of data generating processes.

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, 2022). As for the hyper-parameter estimation, we use the nloptr package for R (Johnson, 2022).

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

Model		ANN	AAN	AAdN	ANA	AAA	AAdA	MAM
	ANN	CR	OV	OV	OV	OV	OV	IE
	AAN	UN	CR	OV	IS	OV	OV	IE
	AAdN	UN	UN	CR	IS	IS	OV	IE
DGP	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 2.2: A classification based on pairs of DGPs and applied models.

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

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

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

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

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, 2000). We scale both measures with the mean of the absolute in-sample value:

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

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

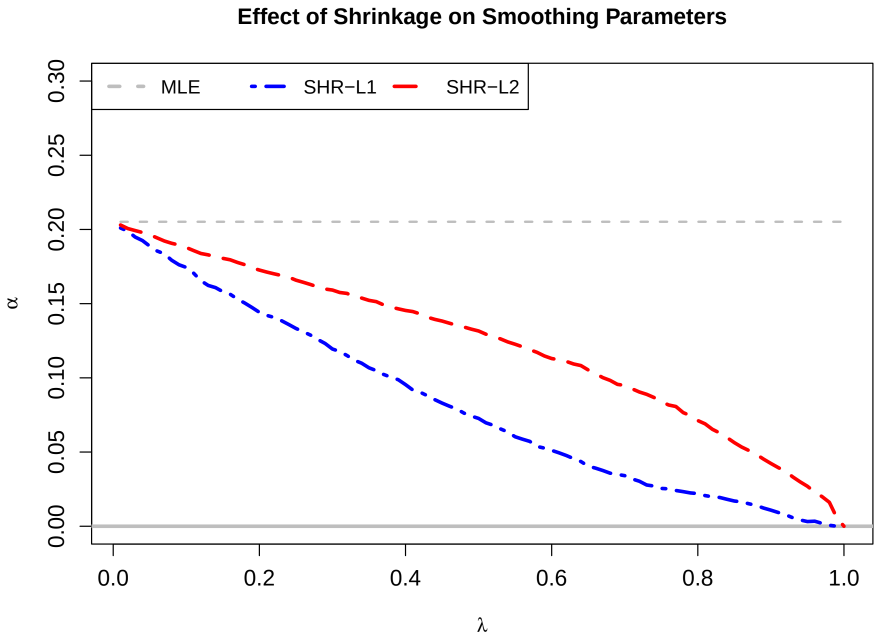


Figure 2.3: Effects of ℓ_1 and ℓ_2 shrinkage on smoothing parameters.

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.

$Ep =$ (Required Personality Level of Each Resource - Assigned Personality Level of each Resources) × (da_{i})

2.3.2 Findings

Accuracy

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 2.3: A summary of aggregate performances in dRMSE.

Table 2.3 presents the aggregate performances of ETS models with shrinkage in the four simulation settings, 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 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 2.3 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.

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

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 2.4: A summary of performances in dRMSE and dAME.

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 2.8), 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 2.4 presents the summary of the forecasting performance for dRMSE and dAME for the forecast horizons of 1 and 1-7, sample size of 28, and ℓ_2 shrinkage. Positive values indicate percentage gain over the performance of the MLE. The table is structured similarly to Table 2.2. 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 2.4 shows the effect of sample sizes on dRMSE in the five scenarios. The dotted lines denote the mean of the distributions and the arrows show some parts of the distributions outside of the plotting area. 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 longer horizons.

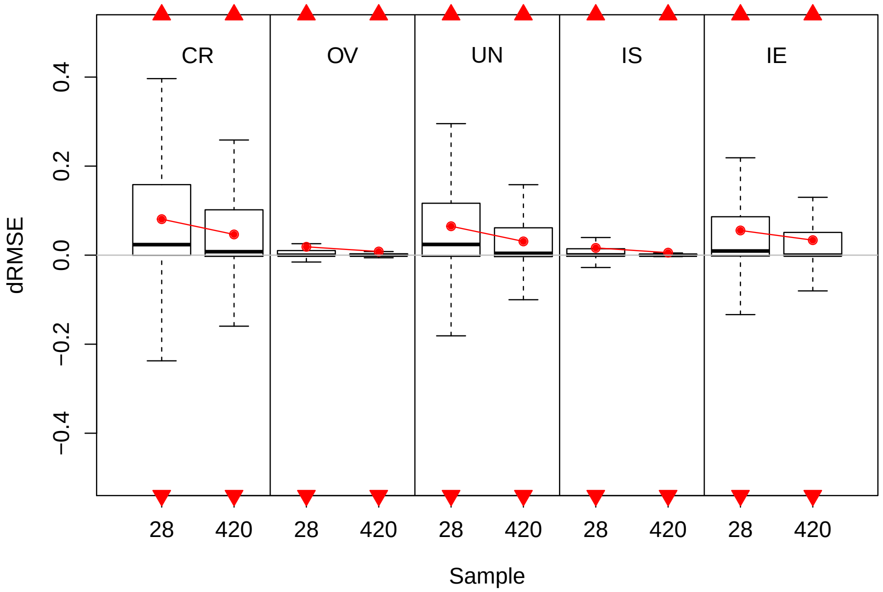


Figure 2.4: Distributions of dRMSE across different model specifications and sample sizes.

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 2.2, 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 2.5.

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 2.5: Several combinations from five levels of model specification

Table 2.6 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

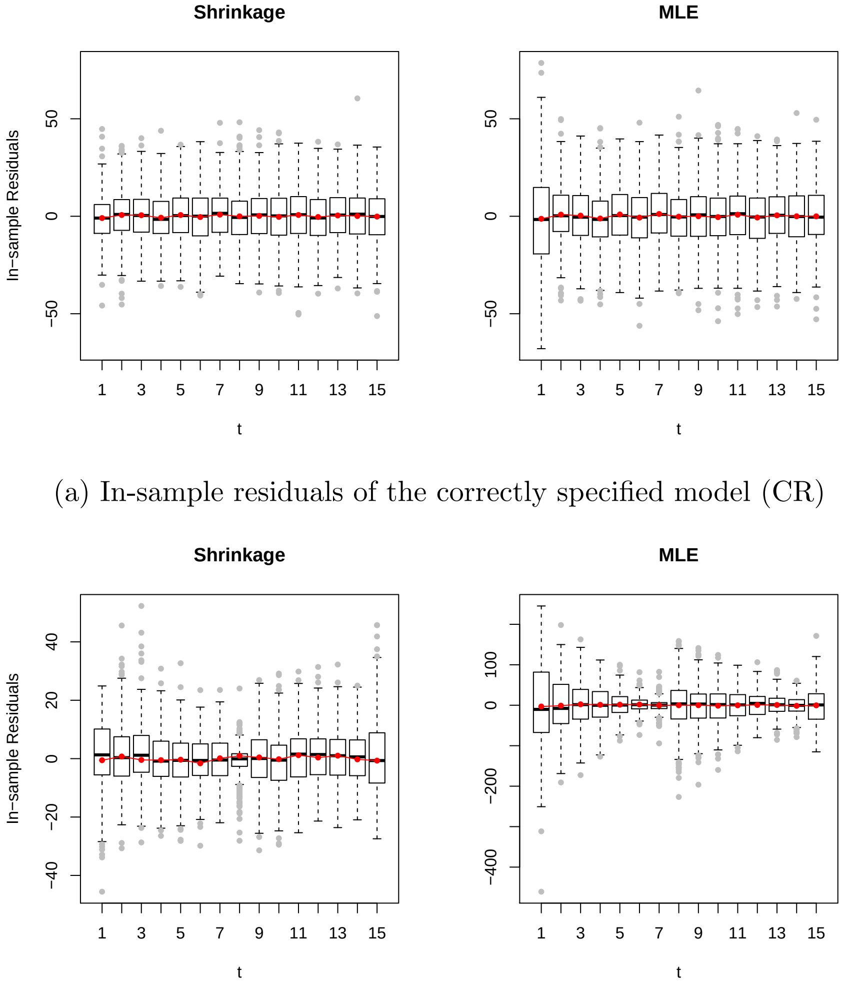
Table 2.6: Overall performance of the prediction interval across scenarios, for small sample size and $h = 1 - 7$.

We conducted several diagnostics of the residuals, namely checking the normality of the residuals with the Shapiro-Wilk test, especially for the small 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 2.6. THE and EMP are the theoretical and the empirical prediction intervals. Bold numbers denote the best performing prediction intervals.

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 2.5a and 2.5b show that the residuals of the shrinkage models are less erratic than the residuals of the MLE ones, where the red dots denote the mean of each distribution and the grey dots denote the outliers of the distribution. 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.

Across all model specifications, we observe that the shrinkage models produce fewer autocorrelated residuals than the MLE models do, as shown in Figure 2.6. 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 2.7 presents the distribution of the standard error for each model specification. The red arrows denote that some parts of the boxplot are not plotted and the red dots denote the means of each distribution. 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. (2.9), we argue that the model bias



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

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

Figure 2.5: Residual diagnostics.

should be non-zero, but the 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.

Next we turn our attention to the initial values. Figure 2.8 presents the standardised level initial values across origins and model specifications. The red dots denote the mean of each distribution and the grey dots denote the outliers of the distribution.

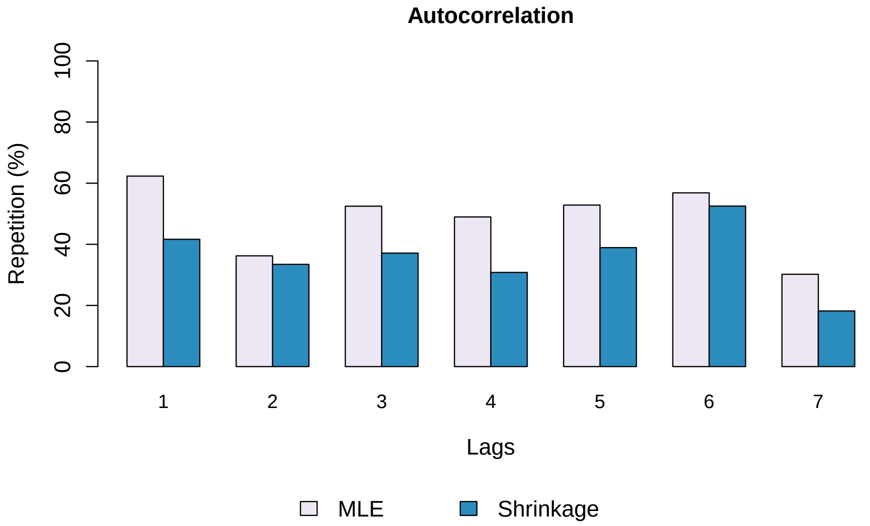


Figure 2.6: Number of instances that contains autocorrelated residuals

On average the initial values are estimated well, however, there are substantial differences in the spread of the distributions. The variability 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

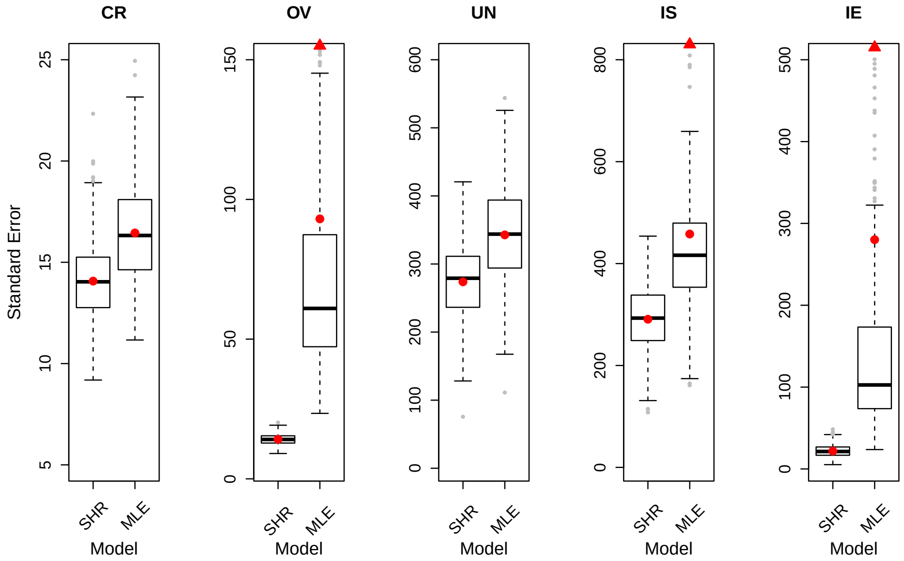


Figure 2.7: Distributions of standard error between the MLE and the shrinkage models.

the conditional variance of the point forecast in the out-of-sample, giving the empirical conditional variance. Figure 2.9 presents the ratio between the theoretical and the empirical conditional variance. 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. For the shrinkage models, on average the theoretical variance matches the empirical conditional variance. For severely misspecified 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.

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

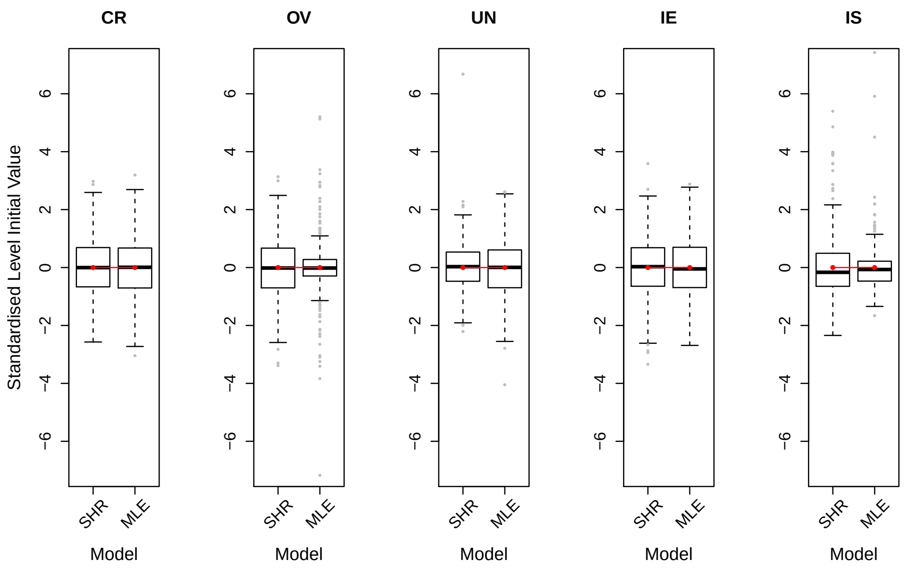


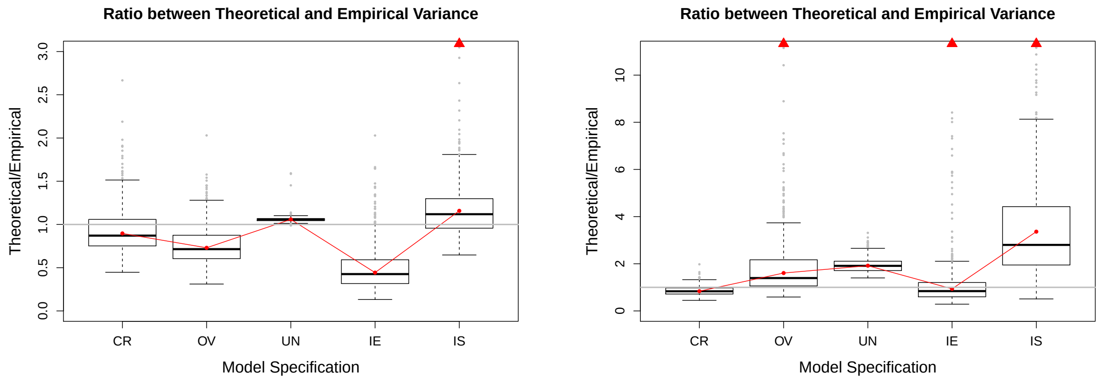
Figure 2.8: Distributions of the standardised initial values from different model specifications.

shrinkage is more appropriate than that of the MLE.

2.4 Application of ETS Shrinkage for the UK NHS

2.4.1 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. Some time series still retain zero values after temporally aggregating them due to its intermittent characteristics and ETS is not appropriate for such time series (Boylan and Syntetos, 2021). The data spans from January 2009 to October 2019 and we aggregate the time



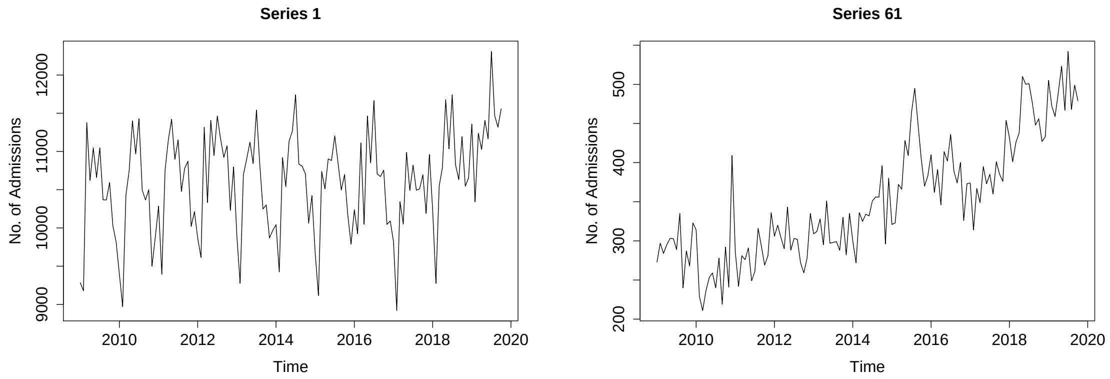
(a) Shrinkage

(b) Benchmark

Figure 2.9: Ratios between the theoretical and the empirical conditional variance of the forecasts, for different levels of model specification.

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 2.10 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 2.3. 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.

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.

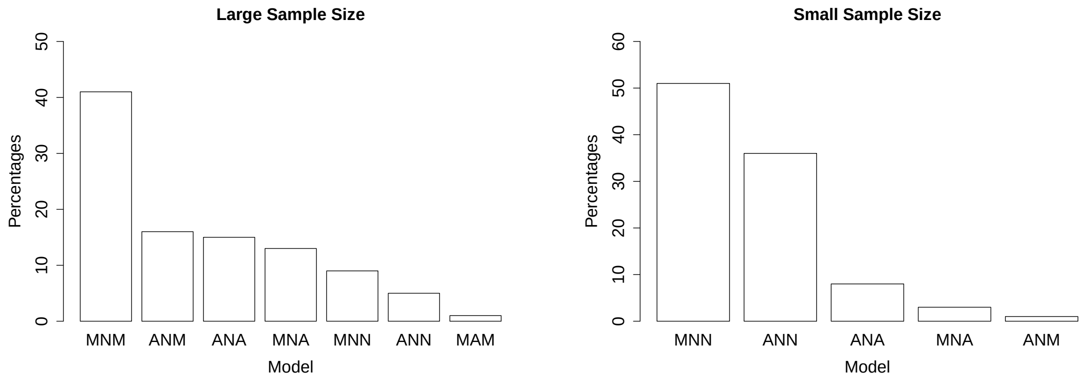


(a) Time series 1

(b) Time series 61

Figure 2.10: Examples of the time series from the A&E NHS dataset

2.4.2 Findings



(a) 2009-2018

(b) 2015-2018

Figure 2.11: 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 2.11. For both sample sizes, we can see that multiplicative models dominate. When the sample size is larger, there are more seasonal models.

Table 2.7 present the summary of performances of approaches in terms of dRMSE and dAME for different forecast horizons. US and WS are the unweighted and the weighted shrinkage, and MLE is the benchmark model. A bold number demonstrates

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 2.7: A summary of dRMSE and dAME for all time series with five origins.

the lowest among the others. 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.

		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 2.8: Performances of the prediction interval.

In addition to the forecast accuracy, we also measure the performance of the prediction intervals, presented in Table 2.8. It contains the prediction intervals for different confidence levels types of shrinkage, forecast horizons, and the small sample size only. THE and EMP are the theoretical and the empirical prediction intervals and bold numbers denote the best performing prediction intervals. 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 2.3. 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 2.8 demonstrates that the prediction interval of the models estimated using the proposed shrinkage approach are adequate.

2.5 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.

Chapter 3

Shrinkage Estimators for Vector Exponential Smoothing

In this chapter, we extend the understanding of a shrinkage estimator from ETS to Vector Exponential Smoothing or VES. We present the results and provide some discussion as to why this approach does not work as we expect. In the end, we discuss possible improvements for the study.

Abstract

Vector Exponential Smoothing (VES) is a multivariate extension of Exponential Smoothing. Like any other multivariate models, it suffers from overparameterisation and overfitting when the sample size is limited and these affect the forecasting performance of the model. We postulate that shrinkage solves these issues and highlight that forcing some parameters to zero arbitrarily is an ad-hoc form of shrinkage. In addition, we also propose to shrink smoothing parameters in VES by developing a shrinkage estimator for VES models. We conducted simulation experiments to test the efficacy of the proposed estimator. The findings show that univariate Exponential Smoothing models outperform the other multivariate models in a multivariate setting. The multivariate models do not perform well because of a compensating effect between the smoothing parameters and the covariance matrix. We suggest exploring

the properties and the behaviour of the loss function of Exponential Smoothing and VES and implementing shrinkage not only on the smoothing parameters but also on the covariance matrix. An improved experimental design should also be investigated, as the results from the VES with shrinkage do not converge to univariate models.

3.1 Introduction

Exponential smoothing is a popular forecasting method, which is widely used in business. It is easy to understand and implement. Its forecasts are fairly accurate for a relatively simple model, as recorded in several forecasting competitions (Makridakis and Hibon, 2000; Makridakis et al., 2018, 2021). Brown (1959) proposed it initially and its development has been growing since (Gardner, 2006). Hyndman et al. (2002) and Hyndman et al. (2008a) automated the methodology and developed a taxonomy of Error-Trend-Seasonal models (ETS) under a Single Source of Error (SSOE) framework. ETS is constructed from unobserved states, which consist of a level, a trend, and a seasonality. For each state, there are two important groups of parameters: smoothing parameters and initial values. Smoothing parameters update new information on the states and initial values are proxies of information prior to data collection. These parameters are generally unknown and we need to estimate them. Hyndman et al. (2002) and Hyndman et al. (2008a) proposed using Maximum Likelihood to estimate them and to select the best approximating ETS model to a target series via the minimisation of an information criterion. This approach works well on larger samples of data, as Maximum Likelihood guarantees consistent and efficient estimated parameters asymptotically. However, this might not be the case in practice, where sample sizes can be small due to, for example, short product life cycles and poor data management. Due to the small sample size, the parameters tend to be higher, which could potentially deteriorate the forecasting performance of ETS due to overfitting. Barrow et al. (2021) and Pritularga et al. (2022) proposed different estimators to deal with the overfitting. Specifically, Pritularga et al. (2022) suggested shrinking the smoothing parameters and found that it controls the stochasticity of states of ETS that leads to more accurate and reliable forecasts.

In practice, we can have a set of connected time series. For example, two products can have complimentary or substitution effects in the retail industry. This also affects the forecasting models. If we model the series via a set of univariate models, e.g., ETS, we cannot model their interdependence and suffer from mis-specification issues. An alternative approach is to employ a multivariate forecasting model, e.g., a vector autoregressive model (VAR). An application of VAR in retail forecasting is well documented by Wilms and Croux (2018) and M5 Competition has demonstrated the importance of 'cross-learning' between time series to produce more accurate forecasts (Makridakis et al., 2021). An alternative to VAR models is a multivariate exponential smoothing model. It was initially introduced by Jones (1966) under the Multiple Source of Error (MSOE) framework, and Duncan et al. (1993) implemented a Bayesian estimation. de Silva et al. (2010) developed the Vector Innovations Structural Time Series (VISTS) that utilises the SSOE framework. The same framework is called Vector Exponential Smoothing (VES) by Hyndman et al. (2008a) and Athanasopoulos and de Silva (2012) extended it by adding seasonality. Throughout the paper we refer to the SSOE multivariate exponential smoothing as 'VES' unless stated otherwise.

VES is a multivariate extension of ETS, and it is constructed from unobserved states, i.e., level, trend, and seasonality. VES can accommodate interdependence between time series in each state. Interdependence in VES can be observed from two parameters: the persistence matrix and the covariance matrix. The persistence matrix contains smoothing parameters of and between time series. The matrix controls both the effect of new information on its own time series and the effect of new cross-series information. For example, the new information from time series 1 has an effect on time series 2 and vice versa, which we call it the cross-smoothing effect. This is basically the extension of smoothing parameters in a matrix form. On the other hand, the covariance matrix is a statistical property of the time series' errors. It shows contemporaneous correlations between errors in each time series. Similar to ETS, VES has two other groups of parameters: initial values and potentially dampening parameters. We omit dampening parameters in this paper to limit additional complications. In addition, VES models have a close connection with Vector Autoregressive Integrated Moving Average (VARIMA) models (de Silva et al., 2010). This can help us to understand parameters in VES better from VARIMA perspectives.

The state-of-the-art VES methodology employs Maximum Likelihood estimation (MLE) of parameters. The MLE works well when the sample size is large enough to produce unbiased and efficient estimates, i.e., the parameters are well estimated asymptotically. However, the estimation of any vector model is not straightforward (James and Stein, 1961; Basu and Michailidis, 2015; Nicholson et al., 2017; Wilms and Croux, 2018; Wilms et al., 2021). The model may suffer from overparameterisation. The number of parameters may increase exponentially as the number of time series increases the number of time series. Given limited sample sizes and overparameterisation, the resulting parameters tend to be biased and/ or inefficient. This may lead to overfitting, where the model attempts to fit the dataset but fails to differentiate between randomness and the underlying structure. This can harm forecast accuracy.

In light of the dimensionality issue, several studies implement parameter shrinkage in VAR models (Basu and Michailidis, 2015; Nicholson et al., 2017; Wilms and Croux, 2018), which makes the model easier to identify and interpret by reducing the number of lags and the cross-effect between time series. Duncan et al. (1993) introduced common parameters in a multivariate exponential smoothing model with the Bayesian approach, but they argue that it may be too restrictive. This model does not acknowledge slight differences between the time series. A new development of VES is proposed by Svetunkov et al. (2022a). They propose a new taxonomy called 'VETS-PIC', where common restrictions are implemented to smoothing (P)arameters, (I)nitial values, and (C)omponents or states. By introducing commonality, the number of parameters reduces significantly and the model becomes easier to identify and estimate. Nonetheless, both studies put strong restrictions on the model. Furthermore, the commonality may not improve the accuracy when the time series are heterogeneous.

The literature has shown two extremes in modelling VES. We can model it without any restrictions. This offers flexibility and fully uses information, but it might suffer from overfitting. On the other hand, we can impose strong restrictions on the model, but this does not accommodate potentially different smoothing of each time series, even though the model is easier to estimate. We argue that we need a bridge between the two extremes, where it can solve the dimensionality issue and allow for some variations in the model. A possible solution is to implement parameter shrinkage. Shrinkage has been known in dealing with high-dimensionality issues Hoerl and Kennard (2000); Tibshirani (1996); Sagaert et al. (2018); Wilms and Croux (2018); Wilms et al. (2021). It shrinks the parameters to zero, conditional on the data, and the model is easier to identify and estimate. Specifically, (Wilms et al., 2021) shrink the autoregressive and the moving average coefficients of a Vector Autoregressive Moving Average (VARMA) to zero, meaning that they reduce the number of lags to make the model more identifiable. On the other hand, parameter shrinkage in VES focuses on reducing the cross-effect between states only, as the model structure has identified the number of lags. Pritularga et al. (2022) show that shrinking the smoothing parameters reduces the effect of new information on each state and eventually reduces the stochasticity of each state. In light of the benefits of shrinkage, we propose to shrink smoothing parameters in VES to deal with the high-dimensionality issue and control the stochasticity of each state. As the smoothing parameters are collected in the persistence matrix, we focus on that matrix and propose several penalty functions to make each state less stochastic (near deterministic).

Section 3.2 discusses VES in general, followed by a detailed discussion of a simple VES model in Section 3.3. We propose the shrinkage implementation in Section 3.5. Section 3.6 describes the simulation setting and demonstrates the findings. Lastly, Section 3.7 discusses and concludes the findings.

3.2 Vector Exponential Smoothing

Similar to ETS, VES reconstructs a set of time series from their vectors of unobserved states, e.g., level ($\boldsymbol{l}$), trend ($\boldsymbol{b}$), and seasonality ($\boldsymbol{s}$). The taxonomy of VES is similar to ETS, provided by Hyndman et al. (2008a). For example, VES(A, A, N) consists of additive errors (A), trends (A), and no seasonality (N). When states have a multiplicative relationship, they will be denoted as M. A multiplicative VES would be a multiplication between states and time series, which it deems complicated. A way to mitigate this issue is to take a logarithm of the model and treat it as an additive model (Svetunkov et al., 2022a). Lastly, additive and multiplicative damped trends are denoted as Ad and Md. This paper focuses on additive models only.

Let $\boldsymbol{y}_t$ be a N-vector observed time series at time t, where N is the number of time series and $t = 1, ..., T$. $\boldsymbol{y}_t$ is constructed from unobserved states ($\boldsymbol{x}_t$), where $\boldsymbol{x}_t$ contains K states. For example, $\boldsymbol{x}_t$ contains three states and can be presented as $\boldsymbol{x}_t = \{\boldsymbol{l}_t, \boldsymbol{b}_t, \boldsymbol{s}_t\}$, where each state has the same dimensionality as $\boldsymbol{y}_t$. The linear function between $\boldsymbol{y}_t$ and $\boldsymbol{x}_t$ or the measurement equation is formulated as de Silva et al. (2010); Hyndman et al. (2008a),

$$\boldsymbol{y}_t = \boldsymbol{W}\boldsymbol{x}_{t-1} + \boldsymbol{\varepsilon}_t, \tag{3.1}$$

where $\boldsymbol{W}$ is a $N \times NK$ measurement matrix and $\boldsymbol{\varepsilon}_t$ is N-vector of errors that have zero means and the $N \times N$ covariance matrix of $\mathbf{\Sigma}$. $\boldsymbol{x}_t$ has an additive dynamic relationship with the past, which is modelled via the transition equation (de Silva et al., 2010; Hyndman et al., 2008a)

$$\boldsymbol{x}_t = \boldsymbol{F}\boldsymbol{x}_{t-1} + \boldsymbol{P}\boldsymbol{\varepsilon}_t, \tag{3.2}$$

where $\boldsymbol{\varepsilon}_t$ has zero means and a covariance matrix of $\mathbf{\Sigma}$, $\boldsymbol{F}$ is $NK \times NK$ transition matrix, and $\boldsymbol{P}$ is a $NK \times N$ persistence matrix that contains a collection of smoothing parameters. Eq. (3.1) and Eq. (3.2) are the parts of an additive VES model.

We can transform the model structure of VES, i.e., Eq. (3.1) and Eq. (3.2), to become a Vector AutoRegressive Integrated Moving Average model or VARIMA(p,d,q), where p and q denote the lag of autoregressive and moving average component and d denotes the level of integration. For example, according to de Silva et al. (2010, p. 358), VES(A,N,N) and VES(A,A,N) are equivalent to VARIMA(0,1,1) and VARIMA(0,2,2). This equivalence significantly explains the relationship between parameters in $\boldsymbol{P}$ and $\mathbf{\Sigma}$. The details of the equivalence are available in de Silva et al. (2010), and we discuss the effect of this relationship on the forecasts in Section 3.3.2.

3.2.1 Estimation

In general, parameters in $\boldsymbol{W}$, $\boldsymbol{F}$, $\boldsymbol{P}$, and $\mathbf{\Sigma}$ are unknown and we need to estimate them. Literature utilises the maximum likelihood approach to estimate these parameters (Hyndman et al., 2008a; de Silva et al., 2010; Svetunkov et al., 2022a). According to Snyder et al. (2017), the likelihood function of a multivariate normal distribution is

$$\ln \boldsymbol{\mathcal{L}}(\boldsymbol{P}, \boldsymbol{x}_0, \mathbf{\Sigma} | \boldsymbol{\mathcal{I}}_T) = -\frac{TN}{2}(\ln 2\pi) - \frac{N}{2}\ln|\mathbf{\Sigma}| - \frac{1}{2}\sum_{t=1}^{T} \boldsymbol{\varepsilon}_t^{\top} \mathbf{\Sigma}^{-1} \boldsymbol{\varepsilon}_t,$$

where $\boldsymbol{\mathcal{I}}_t$ is the collection of time series up to time t. $\boldsymbol{\varepsilon}_t$ refers to additive errors, as presented in Eq. (3.1).

It is often simplified to get the concentrated likelihood,

$$\ln \boldsymbol{\mathcal{L}}(\boldsymbol{P}, \boldsymbol{x}_0, |\hat{\boldsymbol{\Sigma}}, \boldsymbol{\mathcal{I}}_T) = -\frac{TN}{2}(\ln 2\pi) - \frac{N}{2}\ln|\hat{\boldsymbol{\Sigma}}| - \frac{TN}{2}, \tag{3.3}$$

where $\hat{\boldsymbol{\Sigma}} = \frac{1}{T}\hat{\boldsymbol{e}}_{t|t-1}^{\top}\hat{\boldsymbol{e}}_{t|t-1}$. $\hat{\boldsymbol{e}}_{t|t-1}$ is the one-step ahead forecast error and this contains the information about the smoothing parameter and the initial value estimation.

Snyder et al. (2017) postulate that maximising Eq. (3.3) is equivalent to minimising the determinant of $\hat{\boldsymbol{\Sigma}}$, which is called the generalised variance. In the case when the covariances of $\boldsymbol{\Sigma}$ are zero, maximising $|\boldsymbol{\Sigma}|$ is implicitly the same as minimising the sum of variances in $\boldsymbol{\Sigma}$. Hence, we have a trace of variances in the covariance matrix as the loss function. It is defined as,

$$\mathrm{Tr}\left(\mathrm{MSE}_N(\boldsymbol{P}, \boldsymbol{x}_0|\boldsymbol{\mathcal{I}}_T)\right) = \frac{1}{TN}\sum_{i=1}^{N}\sum_{t=1}^{T}\varepsilon_{i,t}^2. \tag{3.4}$$

where MSE_N is a diagonal matrix of MSEs for each time series and $\mathrm{Tr}(\cdot)$ is a trace of a matrix.

Regardless of the estimation, we need to ensure that the model is stable. This means that the past information becomes less relevant to the current information, i.e., states indicate a monotonic behaviour. In such a case, the eigenvalues of the discount matrix should lie *in* a unit circle. The discount matrix is defined as

$$\boldsymbol{D} = \boldsymbol{F} - \boldsymbol{P}\boldsymbol{W}.$$

A relaxation of the stability restrictions allows the eigenvalues to lie on the unit circle, where we can use recent information to predict the future. This model does not need to be stable but it is forecastable.

This stability and/ or the forecastability restriction has several implications for the model. First, the smoothing parameters can be positive or negative, as defined in the admission bound (Hyndman et al., 2008a). The traditional parameter space, i.e., between 0 and 1, no longer holds. Second, there is a possibility that the model is deterministic because we allow the eigenvalues to be on the unit circle. For example, for a deterministic VES(A,N,N), $\boldsymbol{D} = \boldsymbol{I}$ as $\boldsymbol{P} = \boldsymbol{0}$ and the eigenvalues of $\boldsymbol{D}$ are 1. Note that the eigenvalues of $\boldsymbol{D}$ for a deterministic model are equal to 1. On the other hand, a model with the unitary eigenvalues of $\boldsymbol{D}$ does not necessarily become a deterministic model; otherwise it could become an unstable but forecastable model.

3.2.2 Prediction

After estimating the parameters, we can produce forecasts from the model. The one-step ahead point forecasts are denoted as,

$$\begin{aligned} \hat{\boldsymbol{y}}_{t+1|t} &= \boldsymbol{W}\hat{\boldsymbol{x}}_t, \\ \hat{\boldsymbol{x}}_t &= \boldsymbol{F}\hat{\boldsymbol{x}}_{t-1} + \hat{\boldsymbol{P}}\hat{\boldsymbol{e}}_{t|t-1}, \end{aligned}$$

where $\hat{\boldsymbol{P}}$ is the estimated $\boldsymbol{P}$, $\hat{\boldsymbol{x}}_t$ is the estimated state. Note here that as we can structure the model based on which states are incorporated, $\boldsymbol{W}$ linearly structures the $\boldsymbol{y}_t$ and it is treated to be known. Unlike Hyndman et al. (2008b, p. 292), we attempt to incorporate the effects of the in-sample estimation on the forecasts, meaning that $\boldsymbol{x}_t$ is estimated. Thus, the recursive relationship of the forecasts can be formulated as

$$\hat{\boldsymbol{y}}_{t+1|t} = \boldsymbol{W}\hat{\boldsymbol{D}}^t\hat{\boldsymbol{x}}_0 + \sum_{r=0}^{t} \boldsymbol{W}\hat{\boldsymbol{D}}^r\hat{\boldsymbol{P}}\boldsymbol{y}_{t-r}, \tag{3.5}$$

where $\hat{\boldsymbol{D}} = \boldsymbol{F} - \hat{\boldsymbol{P}}\boldsymbol{W}$, and $\hat{\boldsymbol{x}}_0$ is the estimated initial value, and the one-step ahead forecast variance-covariance is

$$\begin{aligned}\mathrm{V}(\hat{\boldsymbol{y}}_{t+1}|\hat{\boldsymbol{P}}, \hat{\boldsymbol{x}}_0, \hat{\boldsymbol{\Sigma}}, \boldsymbol{\mathcal{I}}_T) &= \boldsymbol{F}^{t+1}\hat{\boldsymbol{x}}_0\hat{\boldsymbol{x}}_0^\top(\boldsymbol{F}^{t+1})^\top + \sum_{r=0}^{t+1} \boldsymbol{W}^\top \boldsymbol{F}^r \hat{\boldsymbol{P}}\hat{\boldsymbol{\Sigma}}_{t+1-r}\hat{\boldsymbol{P}}^\top(\boldsymbol{F}^r)^\top \boldsymbol{W} \\ &= \underbrace{\boldsymbol{F}^{t+1}\hat{\boldsymbol{x}}_0\hat{\boldsymbol{x}}_0^\top(\boldsymbol{F}^{t+1})^\top}_{\text{Initial Value Effect}} + \underbrace{\boldsymbol{W}^\top \hat{\boldsymbol{P}}\hat{\boldsymbol{\Sigma}}\hat{\boldsymbol{P}}^\top \boldsymbol{W} \sum_{r=0}^{t+1} \boldsymbol{F}^r(\boldsymbol{F}^r)^\top}_{\text{Covariance and Persistence Matrix Effect}} \end{aligned} \tag{3.6}$$

where we assume that the covariance matrix is time-invariant, i.e., $\hat{\boldsymbol{\Sigma}}_{t+h-1} = \hat{\boldsymbol{\Sigma}}_{t+h-2} = ... = \hat{\boldsymbol{\Sigma}}_t = \hat{\boldsymbol{\Sigma}}_{t-1} = ... = \hat{\boldsymbol{\Sigma}}_0 = \hat{\boldsymbol{\Sigma}}$, for the forecast horizon of h. We can see that the uncertainty in the in-sample propagates to the forecast variance-covariance through the variance of $\hat{\boldsymbol{x}}_0$ and the multiplication between $\hat{\boldsymbol{P}}$ and $\hat{\boldsymbol{\Sigma}}$. This shows that we need to be cautious about estimating $\boldsymbol{P}$ and $\boldsymbol{\Sigma}$.

Next, we illustrate a simple bivariate VES(A,N,N) to demonstrate: (a) the relationship between parameters in $\boldsymbol{P}$ via expanding the VARIMA(0,1,1); (b) the relationship between $\boldsymbol{P}$ and $\boldsymbol{\Sigma}$; (c) the effects of forcing some parameters in $\boldsymbol{P}$ and $\boldsymbol{\Sigma}$ to zero on the prediction distribution; lastly (d) an equivalent situation between VES and ETS.

3.3 Special Cases of VES(A,N,N)

The bivariate VES(A, N, N) is used to explore parameters in VES. It has a single state, level only, and two time series, thus $K = 1$ and $N = 2$. The measurement and the transition equation are,

$$\begin{aligned}\boldsymbol{y}_t &= \boldsymbol{l}_{t-1} + \boldsymbol{\varepsilon}_t \\ \boldsymbol{l}_t &= \boldsymbol{l}_{t-1} + \boldsymbol{P}\boldsymbol{\varepsilon}_t,\end{aligned}$$

where $\boldsymbol{y}_t = \begin{bmatrix} y_{1t} & y_{2t} \end{bmatrix}^\top$, $\boldsymbol{l}_t = \begin{bmatrix} l_{1t} & l_{2t} \end{bmatrix}^\top$, $\boldsymbol{\varepsilon}_t = \begin{bmatrix} \varepsilon_{1t} & \varepsilon_{2t} \end{bmatrix}^\top$, $\boldsymbol{W} = \boldsymbol{F} = \boldsymbol{I}_2$, and $\boldsymbol{I}_2$ is a 2×2 identity matrix, and $\boldsymbol{\varepsilon}_t\ MVN(\mathbf{0}, \boldsymbol{\Sigma})$. $\boldsymbol{P}$ and $\boldsymbol{\Sigma}$ are,

$$\boldsymbol{P} = \begin{bmatrix} \alpha_{11} & \alpha_{12} \\ \alpha_{21} & \alpha_{22} \end{bmatrix}, \quad \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{11}^2 & \sigma_{12} \\ \sigma_{21} & \sigma_{22}^2 \end{bmatrix}.$$

Parameters α_{11} and α_{22} model the smoothing on each series individually, while α_{12} and α_{21} model the cross-smoothing.

3.3.1 VARIMA equivalence

de Silva et al. (2010) has shown an equivalence between VES and VARIMA, where in this case, VES(A,N,N) is equivalent to VARIMA(0,1,1). However, the literature does not explicitly explain the interaction between parameters in $\boldsymbol{P}$ via VARIMA. We transform VES into VARIMA via linear algebra and expand the model equations to uncover the interaction between α_{ii} and α_{ij}, for all $i, j \in N$, to understand the complexity of modelling VES. The linear transformation is:

$$\begin{aligned} (\boldsymbol{I}_2 - \boldsymbol{I}_2 L)\boldsymbol{y}_t &= \boldsymbol{\varepsilon}_t - (\boldsymbol{I}_2 - \boldsymbol{P})\boldsymbol{\varepsilon}_{t-1}, \\ &= \boldsymbol{\varepsilon}_t - \boldsymbol{Z}\boldsymbol{\varepsilon}_{t-1} \end{aligned} \tag{3.7}$$

where $\boldsymbol{Z} = \boldsymbol{I}_2 - \boldsymbol{P}$ and it is the coefficient matrix of the moving average component. If we expand it into a system of equations, these are:

$$\begin{aligned} y_{1t} - y_{1t-1} &= \varepsilon_{1t} - (1 - \alpha_{11})\varepsilon_{1t-1} + \alpha_{12}\varepsilon_{2t-1}, \\ y_{2t} - y_{2t-1} &= \varepsilon_{2t} - (1 - \alpha_{22})\varepsilon_{2t-1} + \alpha_{21}\varepsilon_{1t-1}. \end{aligned} \tag{3.8}$$

We focus on the first equation in Eq. (3.8) only. We can transform Eq. (3.8) with a backshift operator L and we define $z_{11} = 1 - \alpha_{11}$ and $z_{12} = -\alpha_{12}$,

$$\begin{aligned} (1 - z_{11}L)^{-1}(1 - L)y_{1t} &= \varepsilon_{1t} + (1 - z_{11}L)^{-1} z_{12} L \varepsilon_{2t} \\ \left(1 + z_{11}L + z_{11}^2 L^2 + ...\right)(1 - L)y_{1t} &= \varepsilon_{1t} + \left(1 + z_{11}L + z_{11}^2 L^2 + ...\right) z_{12} L \varepsilon_{2t}, \end{aligned} \tag{3.9}$$

where $(1 - z_{11}L)^{-1} = (1 + z_{11}L + z_{11}^2 L^2 + ...)$. Eq. (3.9) shows an ARIMA process with the additional ε_{2t}, where z_{11} affects the autoregressive component and $z_{11}^t z_{12}$, for $t = 1, ..., T$, determines the effects of ε_{2t} on y_{1t}. This shows a complicated interaction between the autoregressive and the moving average component, as well as, between z_{11} and z_{12}. The effect of ε_{2t} may persist or even overwhelm the effect of ε_{1t} on y_{1t}. Therefore, we need to be cautious when dealing with the off-diagonal parameters of $\boldsymbol{P}$, where omitting z_{12} may simplify the model but lose information about ε_{2t}.

3.3.2 Forecast Variance-Covariance

We have discussed the interaction between z_{11} and z_{12} on Eq. (3.9), and now we explore the interaction between parameters in $\boldsymbol{P}$ and $\boldsymbol{\Sigma}$ from the forecast variance-covariance. Since in this example we use VES(A,N,N), its forecast variance-covariance is,

$$\mathrm{V}(\hat{\boldsymbol{y}}_{t+1} | \hat{\boldsymbol{P}}, \hat{\boldsymbol{l}}_0, \hat{\boldsymbol{\Sigma}}, \boldsymbol{\mathcal{I}}_T) = \mathrm{V}(\hat{\boldsymbol{l}}_0) + \hat{\boldsymbol{P}} \hat{\boldsymbol{\Sigma}} \hat{\boldsymbol{P}}^\top \sum_{r=0}^{t+1} \boldsymbol{I}^r (\boldsymbol{I}^r)^\top, \tag{3.10}$$

where the one-step ahead forecast variance-covariance consists of (a) the variance of the estimated initial value and (b) the multiplication between $\hat{\boldsymbol{P}}$ and $\hat{\boldsymbol{\Sigma}}$. The forecast variance-covariance is affected by the estimation in the in-sample. Now, we break

$\hat{\boldsymbol{P}}\hat{\boldsymbol{\Sigma}}\hat{\boldsymbol{P}}$ and the matrix becomes

$$\hat{\boldsymbol{P}}\hat{\boldsymbol{\Sigma}}\hat{\boldsymbol{P}}^\top = \begin{bmatrix} \hat{\alpha}_{11}^2(\hat{\alpha}_{11}\hat{\sigma}_{11}^2 + \hat{\alpha}_{12}\hat{\sigma}_{21}) + \hat{\alpha}_{12}^2(\hat{\alpha}_{11}\hat{\sigma}_{12} + \hat{\alpha}_{12}\hat{\sigma}_{22}^2) & \hat{\alpha}_{21}^2(\hat{\alpha}_{11}\hat{\sigma}_{11}^2 + \hat{\alpha}_{12}\hat{\sigma}_{21}) + \hat{\alpha}_{22}^2(\hat{\alpha}_{11}\hat{\sigma}_{12} + \hat{\alpha}_{12}\hat{\sigma}_{22}^2) \\ \hat{\alpha}_{11}^2(\hat{\alpha}_{21}\hat{\sigma}_{11}^2 + \hat{\alpha}_{22}\hat{\sigma}_{21}) + \hat{\alpha}_{12}^2(\hat{\alpha}_{21}\hat{\sigma}_{12} + \hat{\alpha}_{12}\hat{\sigma}_{22}^2) & \hat{\alpha}_{21}^2(\hat{\alpha}_{21}\hat{\sigma}_{11}^2 + \hat{\alpha}_{22}\hat{\sigma}_{21}) + \hat{\alpha}_{22}^2(\hat{\alpha}_{21}\hat{\sigma}_{12} + \hat{\alpha}_{12}\hat{\sigma}_{22}^2) \end{bmatrix}. \tag{3.11}$$

From Eq. (3.11), we can see that the forecast variance-covariance is constructed from several multiplications between σ_{ii}^2, σ_{ij}, α_{ii}, and α_{ij}. As we need to estimate $\boldsymbol{P}$ and $\boldsymbol{\Sigma}$, this affects the construction of Eq. (3.10) and, in turn, the accuracy of the prediction intervals. This shows that in dealing with parameter uncertainty, we need to deal with $\boldsymbol{P}$ and $\boldsymbol{\Sigma}$ in some way or another.

3.3.3 Estimation in VES

In our discussion, capturing the interconnections between time series requires a full estimation of $\boldsymbol{P}$ and $\boldsymbol{\Sigma}$, thus both parameters are of our interests, i.e., as to estimate them conventionally or with shrinkage. Our previous discussion demonstrates the relationships between parameters in $\boldsymbol{P}$ and parameters between $\boldsymbol{P}$ and $\boldsymbol{\Sigma}$. When the sample size is sufficient enough to produce consistent and efficient estimates of parameters, we can estimate these with no adverse impact on the forecast accuracy. However, as discussed in Section 1, the sample size is often limited for various reasons. Thus, the estimation of VES might result inefficient parameters, either in $\boldsymbol{P}$ or/and $\boldsymbol{\Sigma}$. This can make the interaction between parameters unhelpful, exacerbating uncertainties and propagating the uncertainty to forecast accuracy. We argue that parameter shrinkage can mitigate this issue.

Shrinkage has a wide spectrum. If we look at VAR models, a structural VAR (Lütkepohl, 2005) can be seen as a specific form of shrinkage, where some coefficients are forced to be zero so that the model is identifiable and interpretable. In the hierarchical forecasting literature, if seen as a multivariate modelling problem, some studies propose a covariance matrix with zero covariances as an approximation to the $h-$

step ahead forecast error covariance matrix (Hyndman et al., 2016; Athanasopoulos et al., 2017; Wickramasuriya et al., 2019; Nystrup et al., 2020). Pennings and van Dalen (2017) proposed a multivariate exponential smoothing model with MSOE to model the hierarchical time series, where the off-diagonal parameters in $\boldsymbol{P}$ are zero. In the VES literature, a commonality restriction proposed by Duncan et al. (1993) and Svetunkov et al. (2022a) is also another form of shrinkage. Instead of shrinking the parameters to zero, the parameters are regularised to a common value. de Silva et al. (2010) and Svetunkov et al. (2022a) lean toward using a diagonal covariance matrix instead of a fully estimated covariance matrix. These studies show the efficacy of restricting some parameters to zero in forecasting problems. On the other hand, we typically understand shrinkage as data-dependent, meaning that the amount of parameter shrinkage is conditional to the data at hand. Such shrinkage is often proposed to shrink the covariance matrix (Daniels and Kass, 2001; Schäfer and Strimmer, 2005). In the VAR literature, several autoregressive coefficient shrinkage approaches have been proposed (Basu and Michailidis, 2015; Nicholson et al., 2017; Wilms and Croux, 2018; Wilms et al., 2021).

There are two streams in implementing shrinkage: (a) fixed shrinkage, where some parameters are forced to be zero and (b) data-dependent shrinkage. We explore the effects of the fixed shrinkage in VES and propose a shrinkage estimator with several options for the penalty function, investigating both streams.

3.4 The Effects of Fixed Shrinkage on VES

Having discussed the fixed shrinkage on multivariate forecasting models, we illustrate the effects of the fixed shrinkage on VES using VES(A,N,N), specifically Eq. (3.9) and Eq. (3.11). We set α_{ij} and σ_{ij} to be zero. When α_{ij} is zero, then Eq. (3.9) becomes

$$\left(1 + z_{11}L + z_{11}^2 L^2 + ...\right)(1 - L)y_{1t} = \varepsilon_{1t},$$

where it is a much simpler ARIMA process, and the effect of ε_{2t} on y_{1t} disappears. As long as the effect of z_{11} on y_{1t} diminishes over time, then it is easier to estimate this model.

With regards to Eq. (3.11), if we set α_{ij} and σ_{ij} to zero, the matrix becomes,

$$\hat{\boldsymbol{P}}\hat{\boldsymbol{\Sigma}}\hat{\boldsymbol{P}}^\top = \begin{bmatrix} \hat{\sigma}_{11}^2 \hat{\alpha}_{11}^2 & 0 \\ 0 & \hat{\sigma}_{22}^2 \hat{\alpha}_{22}^2 \end{bmatrix}.$$

We can see the forecast variance-covariance becomes less complicated. Thus, from the model structure, forcing some parameters to zero makes the model easier to identify and estimate. It also affects the construction of the forecast variance-covariance, potentially improving the forecast accuracy.

3.4.1 Equivalence condition between VES and ETS

Despite the benefits of the fixed shrinkage in VES, this shrinkage highlights a situation where the model structure of VES is equivalent to a set of ETS, mathematically. Using the same example of VES(A,N,N), the matrix form of VES is shown as,

$$\begin{bmatrix} y_{1t} \\ y_{2t} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} l_{1t-1} \\ l_{2t-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{bmatrix}, \tag{3.12}$$

$$\begin{bmatrix} l_{1t} \\ l_{2t} \end{bmatrix} = \begin{bmatrix} l_{1t-1} \\ l_{2t-1} \end{bmatrix} + \begin{bmatrix} \alpha_{11} & 0 \\ 0 & \alpha_{22} \end{bmatrix} \begin{bmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{bmatrix}. \tag{3.13}$$

Both Eq. (3.12) and Eq. (3.13) show that the bivariate VES(A,N,N) are the same as two ETS(A,N,N) models.

In terms of the estimation of both models, the equivalence in the parameters can be achieved when the covariances in the covariance matrix is omitted, meaning that the covariances are zero. In that case, the loss function of the VES is shown as,

$$\{\hat{\alpha}_{11}, \hat{\alpha}_{22}, \hat{x}_{01}, \hat{x}_{02}\}^{\text{VES}} = \min \ \left(\text{MSE}(\alpha_{11}, x_{01}|\mathcal{I}_{1t}) + \text{MSE}(\alpha_{22}, x_{02}|\mathcal{I}_{2t})\right), \qquad (3.14)$$

where $\{\hat{\alpha}_{11}, \hat{\alpha}_{22}, \hat{x}_{01}, \hat{x}_{02}\}^{\text{VES}} = \hat{\theta}^{\text{VES}}$, and $\mathcal{I}_{it}$ is the time series up to time t for i time series, for i in N. Given the same $\boldsymbol{\Sigma}$, the sum of ETS loss functions is the same as the VES loss function, as shown,

$$\min \ \left(\text{MSE}(\alpha_{11}, x_{01}|\mathcal{I}_{1t}) + \text{MSE}(\alpha_{22}, x_{02}|\mathcal{I}_{2t})\right) = \min \text{MSE}(\alpha_{1}, x_{01}|\mathcal{I}_{1t}) + \min \text{MSE}(\alpha_{2}, x_{02}|\mathcal{I}_{2t}).$$

If this condition is achieved, the parameters of VES are equal to those of ETS, or $\hat{\theta}^{\text{VES}} = \hat{\theta}^{\text{ETS}}$. Mathematically, there are two conditions that promote equivalence between VES and ETS, namely (a) the same model structure and (b) the same loss function.

In practice, we minimise the loss function with a numerical optimisation routine. The sum of ETS losses and the VES loss function should result in the same parameters within a tolerance margin. However, as the optimiser in the VES case has to solve a problem with much bigger dimensionality, the algorithmic aspects of the numerical optimisation become relevant and result in an additional source for the deviation between the VES and ETS parameters, even under the restrictions discussed above.

3.5 Shrinkage Estimator and Penalty Function

We see that theoretically, shrinkage in $\boldsymbol{P}$ to zero simplifies the interaction between parameters and avoids the estimation error propagation. We propose a shrinkage estimator that shrinks $\boldsymbol{P}$ so that we can mitigate the parameter uncertainty and control the state stochasticity without losing the structural interdependence between time series. Thus, we modify the loss function to

$$(1-\lambda)\boldsymbol{\mathcal{C}} + \lambda f(\cdot), \tag{3.15}$$

where λ is a scalar shrinkage hyper-parameter that regularises the parameters uniformly across time series and state, $0 \leq \lambda \leq 1$, $\boldsymbol{\mathcal{C}}$ is a multivariate loss function, where it can be Eq. (3.3) or Eq. (3.4), and $f(\cdot)$ is a penalty function that we discuss in this section. We can interpret Eq. (3.15) as a trade-off between model fit and model inertia. If λ is 0, the effect of the penalty is zero. On the other hand, if λ is 1, the estimator leans towards the model inertia, meaning that the new information does not affect the states and the model becomes anchored to the initial values, hence more deterministic (Pritularga et al., 2022).

We focus on $\boldsymbol{D}$, where it collects much information about the model, as stated in Eq. (3.5). This motivates us to propose different penalty functions, namely shrinking $\boldsymbol{P}$, $\boldsymbol{PW}$ and the eigenvalues of $\boldsymbol{D}$ to 1, as all of them potentially lead to a less stochastic model. As for the norm function, we use ℓ_2 norm as Pritularga et al. (2022) has shown that the ℓ_2 norm outperforms the ℓ_1 one and the differences are marginal, in the context of univariate ETS.

3.5.1 Shrinking $\boldsymbol{P}$

The penalty function is formulated in this case as:

$$f(\boldsymbol{P}) = \sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{k=1}^{K} p_{ijk}^2,$$

where $i, j = \{1, .., N\}$, $k = \{1, ..., K\}$, N is the number of time series and k is the number of states. Suppose that $\lambda \to 1$, $p_{ijk} \to 0$, and $\boldsymbol{P} \to \boldsymbol{0}$, where $\boldsymbol{0}$ is a null matrix. Consequently, because $\boldsymbol{D} \approx \boldsymbol{F}$, Eq. (3.5) becomes,

$$\hat{\boldsymbol{y}}_{t+1|t} = \boldsymbol{W}\boldsymbol{F}^{t+1}\hat{\boldsymbol{x}}_0.$$

The new information effect vanishes as $\boldsymbol{P} \to \boldsymbol{0}$. Thus, shrinking $\boldsymbol{P}$ means that the initial values become more prominent in constructing the forecasts. As shown by (Pritularga et al., 2022) the estimated initial values reflect this.

3.5.2 Shrinking $\boldsymbol{PW}$

A similar approach to Section 3.1 is to shrink parameters in $\boldsymbol{PW}$ to zero so that $\boldsymbol{D} \to \boldsymbol{F}$. However, since we multiply $\boldsymbol{P}$ and $\boldsymbol{W}$, it is the same as multiplying the sum of p_{ij} by the number of states. For example, $\sum_{i,j=1}^{N}\sum_{k=1}^{q} p_{ijk}w_{ijk} = 2\sum_{i,j=1}^{N}\sum_{k=1}^{q} p_{ijk}$, where w_{ijk} is each value in $\boldsymbol{W}$ for all i, j, and k. As a result, this penalises the loss function heavier and faster than shrinkage in 3.5.1. Thus, the penalty function is defined as

$$f(\boldsymbol{PW}) = \sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{k=1}^{K}(p_{ijk}w_{ijk})^2.$$

3.5.3 Shrinking eigenvalues to 1

The next approach is to shrink the eigenvalues of $\boldsymbol{D}$ to 1. $\boldsymbol{D}$ contains 'information' of the models, especially in its eigenvalues, denoted as ξ. The intuition follows from Hyndman et al. (2008a), where the authors decompose a discount matrix for a univariate ETS to explain the forecastability condition. In the same manner, we use the same decomposition in the multivariate setting. We decompose $\boldsymbol{D}$ so that $\boldsymbol{D} = \boldsymbol{U}\boldsymbol{H}\boldsymbol{U}^{-1}$, where $\boldsymbol{U}$ is a eigenvector and $\boldsymbol{H}$ is a diagonal matrix that contains the eigenvalues of $\boldsymbol{D}$. This decomposition affects the recursive relationship in the time series, shown as:

$$\begin{aligned}\boldsymbol{y}_t &= \boldsymbol{W}\left(\boldsymbol{U}\boldsymbol{H}\boldsymbol{U}^{-1}\right)^{t-1}\boldsymbol{x}_0 + \sum_{k=1}^{t-1}\boldsymbol{W}\left(\boldsymbol{U}\boldsymbol{H}\boldsymbol{U}^{-1}\right)^{k-1}\boldsymbol{P}\boldsymbol{y}_{t-k} + \boldsymbol{\varepsilon}_t,\\ &= \boldsymbol{W}\boldsymbol{U}\boldsymbol{H}^{t-1}\boldsymbol{U}^{-1}\boldsymbol{x}_0 + \sum_{k=1}^{t-1}\boldsymbol{W}\boldsymbol{U}\boldsymbol{H}^{k-1}\boldsymbol{U}^{-1}\boldsymbol{P}\boldsymbol{y}_{t-k} + \boldsymbol{\varepsilon}_t. \qquad (3.16)\end{aligned}$$

First, we can see that $\boldsymbol{U}$ does not give much information and $\boldsymbol{H}$ determines the recursive relationship of $\boldsymbol{y}_t$. It is also clear that the effect of $\boldsymbol{H}$ vanishes over time. Suppose that $\boldsymbol{H} \to \boldsymbol{I}$ when $\xi \to 1$, then the recursive relation can be written as,

$$\boldsymbol{y}_t = \boldsymbol{W}\boldsymbol{x}_0 + \sum_{k=1}^{t-1}\boldsymbol{W}\boldsymbol{P}^{\dagger\dagger}\boldsymbol{y}_{t-k} + \boldsymbol{\varepsilon}_t,$$

where $\boldsymbol{P}^{\dagger\dagger}$ is the persistence matrix when $\boldsymbol{H} \to \boldsymbol{I}$. In contrast, if $\boldsymbol{H} \to \boldsymbol{0}$, then $\boldsymbol{y}_t \to \varepsilon_t$, making model behave like Random Walk. Given this rationale we propose to shrinking ξ_i to 1, where $i = \{1, ..., N\}$. Thus, the penalty in this case is defined as,

$$f(\boldsymbol{H}) = \sum_{i=1}^{N}\sum_{k=1}^{K}(1 - \xi_{ik})^2.$$

3.5.4 Hyper-parameter Optimisation

We employ a derivative-free numerical optimisation to find a scalar λ, similar to Pritularga et al. (2022). This implies that the shrinkage rate for all states and time series is the same. We aim to minimise the mean squared one-step ahead holdout forecast error, over Q forecast origins and N time series in the multivariate system. Asymptotically, the one-step ahead holdout forecast error has the same properties as the in-sample residuals (Chatfield, 2000). The objective function is

$$\hat{\lambda} = \arg\min_{\lambda} \frac{1}{Q} \sum_{q=1}^{Q} \mathrm{MSE}_{q,N}(\lambda, \hat{\boldsymbol{P}}_q, \hat{\boldsymbol{x}}_{q,0}, \boldsymbol{y}_{1:q}),$$

where $\mathrm{MSE}_{o,N}$ is the mean squared of the one-step ahead holdout forecast error at the forecast origin of q and $q = \{1, ..., Q\}$. We use the Nelder-Mead optimisation routine (Nelder and Mead, 1965) to find the optimal λ.

3.6 Simulation Study

In this section, we describe our experimental design to demonstrate the efficacy of the proposed estimators, in comparison with the other models. We also discuss the findings from our experiments.

3.6.1 Experimental Design

$\boldsymbol{P}/\boldsymbol{\Sigma}$	Uncorrelated	Correlated
Independent	DGP(IDP,UNCORREL)	DGP(IDP,CORREL)
Dependent	DGP(DEP,UNCORREL)	DGP(DEP,CORREL)

Table 3.1: A combination of $\boldsymbol{P}$ and $\boldsymbol{\Sigma}$ in DGP

We design four simulation experiments based on types of interdependence in $\boldsymbol{P}$ and $\boldsymbol{\Sigma}$. We aim to demonstrate the efficacy of the proposed estimation approaches, i.e., the uncorrelated and correlated innovations (named UNCORREL, CORREL), and the independent and dependent persistence matrix (named IDP, DEP). Thus, we have UNCORREL-IDP, UNCORREL-DEP, CORR-IDP, and CORREL-DEP, summarised in Table 3.1. Of each Data Generating Process (DGP), we have four combinations that represent different levels of model specification, shown in Table 3.2.

Specification	CR	OV	UN	IE
DGP	A,N,N	A,N,N	A,N,A	A,N,A
Model	A,N,N	A,A,N	A,N,N	A,A,N

Table 3.2: Model specification

Table 3.2 lists four groups of model specification, namely correctly-specified model (CR), overly-specified model (OV), under-specified model (UN), and incorrect model specification (IS). In the last case, the model incorrectly includes the trend and excludes the seasonality, where there should have had a seasonality. We attempt to distinguish different types of model misspecification, i.e., OV, UN, and IS.

Of each combination, we have 12 models with the loss function of the maximum likelihood (LIKE) and the trace (TR), referring to Eq. (3.3) and Eq. (3.4) respectively. This leads to a univariate ETS (ETS), a univariate ETS with shrinkage (ETS-SHR), VES with independent $\boldsymbol{P}$ (LIKE-IDP and TR-IDP), VES with dependent $\boldsymbol{P}$ (LIKE-DEP and TR-DEP), VES with dependent $\boldsymbol{P}$ and shrinkage in $\boldsymbol{P}$ (LIKE-P and TR-P), VES with dependent $\boldsymbol{P}$ and shrinkage in $\boldsymbol{PW}$ (LIKE-PW and TR-PW), and VES with dependent $\boldsymbol{P}$ and shrinkage in the eigenvalues (LIKE-1L and TR-1L). Table 3.3 summarises the combinations of the models, the loss function, $\boldsymbol{P}$, the penalty functions, and the model naming. The standard models in the literature are ETS, LIKE-IDP, TR-IDP, LIKE-DEP, TR-DEP, which we consider as benchmarks.

We generate monthly time series with a frequency of 12. The details of the parameters are summarised in Table 3.4. Values in Table 3.4 ensure the stability

		Name	ETS	ETS-SHR	LIKE-DEP	LIKE-IDP	LIKE-P	LIKE-PW	LIKE-1L	TR-DEP	TR-IDP	TR-P	TR-PW	TR-1L
	Struc	ETS	✓	✓										
		VES			✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	Loss	MSE	✓											
		MSE-SHR		✓										
		Likelihood			✓	✓	✓	✓	✓					
		Trace								✓	✓	✓	✓	✓
	Full	Dependent (DEP)			✓					✓				
Shrinkage	Fix	Independent (IDP)				✓					✓			
	Data	$\boldsymbol{P}$					✓					✓		
		$\boldsymbol{PW}$						✓					✓	
		$\sum(1-\xi_i)^2$							✓					✓

Table 3.3: A combination of treatments in the models.

of the processes, where U(·) and MVN(·) denote the uniform and the multivariate normal distribution. Note that when the process is A, N, N, then $\gamma_{ij} = 0$. As for the covariance matrix, the setup is summarised in Table 3.5 and UNCORR and CORR denote the uncorrelated and correlated covariance matrices. Seasonal initial values are randomly set and we ensure that they have similar similar seasonal patterns and are normalised. We employ 5 time series in the multivariate system. We also ensure that the number of degrees of freedom is sufficient to estimate the model.

We produce 1 and 1-12 steps ahead point forecasts, and generate theoretical prediction intervals of 80%, 85%, 90%, 95%, and 99% confidence levels. The theoretical prediction intervals implicitly assume that the errors follow a multivariate normal distribution with zero covariances. To generalise our findings, we repeat the simulation 500 times, which was sufficient for the summary statistics to converge. We use the sim.ves() function to generate the time series. For univariate models, we use the adam() function from smooth package (Svetunkov, 2022b) while for multivariate mod-

Smoothing parameters	
α_{ii}	$\mathrm{U}[0.3, 0.6]$, $i \in N$
α_{ij}	$(0.9 - \alpha_{ii})/n$, $i, j \in N$, $i \neq j$
γ_{ii}	$\mathrm{U}[0.2, 1 - \alpha_{ii}]$, $i \in N$
γ_{ij}	$(0.6 - \gamma_{ii})/N$, $i, j \in N$, $i \neq j$
Parameter	Values
Seasonal cycle	12
Initial value (l_{i0})	$\mathrm{U}[100, 1000]$, $i \in N$
Number of observations	$T = 36$ months
Holdout size	12 months
Forecast horizon	1-12 months
Group sizes	5
Noise	$\boldsymbol{\varepsilon}_t \sim \mathrm{MVN}(\mathbf{0}, \boldsymbol{\Sigma})$
Loss function	Trace (TR) and Likelihood (LIKE)

Table 3.4: A simulation setup for the data generating processes.

UNCORR	$\sigma_{i^{\ddagger}i^{\ddagger}}$	10, $i^{\ddagger} = \{1, 2, ..., \lfloor \frac{N}{2} \rfloor\}$, $i^{\ddagger} = j^{\ddagger}$
	$\sigma_{i^{\ddagger\ddagger}i^{\ddagger\ddagger}}$	7, $i^{\ddagger\ddagger} = \{\lceil \frac{N}{2} \rceil, ..., N\}$, $i^{\ddagger\ddagger} = j^{\ddagger\ddagger}$
	σ_{ij}	0, $i \in \{i^{\ddagger}, i^{\ddagger\ddagger}\}$, $j \in \{j^{\ddagger}, j^{\ddagger\ddagger}\}$
CORR	$\sigma_{i^{\ddagger}i^{\ddagger}}$	10, $i^{\ddagger} = \{1, 2, ..., \lfloor \frac{N}{2} \rfloor\}$, $i^{\ddagger} = j^{\ddagger}$
	$\sigma_{i^{\ddagger\ddagger}i^{\ddagger\dagger}}$	7, $i^{\ddagger\ddagger} = \{\lceil \frac{N}{2} \rceil, ..., N\}$, $i^{\ddagger\ddagger} = j^{\ddagger\ddagger}$
	σ_{ij}	6, $i \in \{i^{\ddagger}, i^{\ddagger\ddagger}\}$, $j \in \{j^{\ddagger}, j^{\ddagger\ddagger}\}$

Table 3.5: A simulation setup for $\boldsymbol{\Sigma}$.

els we use the ves() function from legion package (Svetunkov and Pritularga, 2022) for R (R Core Team, 2022).

We assess the performance of the proposed estimators with the Mean Squared Error (MSE) to measure the accuracy of point forecasts and the Absolute Mean Error to measure the magnitude of the bias:

$$\mathrm{MSE} = \tfrac{1}{Nh} \textstyle\sum_{i=1}^{N} \sum_{k=1}^{h} \left(y_{i,t+k} - \hat{y}_{i,t+k|t+k-1} \right)^2 \quad \mathrm{AME} = \tfrac{1}{N} \textstyle\sum_{i=1}^{N} \left| \tfrac{1}{h} \sum_{k=1}^{h} \left(y_{i,t+k} - \hat{y}_{i,t+k|t+k-1} \right) \right|,$$

and as for the prediction interval, we use the Mean Interval Score (MIS),

$$\mathrm{MIS} = \tfrac{1}{h} \textstyle\sum_{k=1}^{h} \left((ub_{t+k} - lb_{t+k}) + \tfrac{2}{\tau}(lb_{t+k} - y_{t+k}\mathbf{1}(y_{t+k} < lb_{t+k}) + \tfrac{2}{\tau}(y_{t+k} - ub_{t+k}\mathbf{1}(y_{t+k} > ub_{t+k})) \right),$$

where τ is the level of confidence, ub_t and lb_t are the upper and lower bounds, and $\mathbf{1}(\cdot)$ is an indicator function that returns 1 if the condition inside it is met, and otherwise.

We choose MIS over the other measures such as Pinball and Coverage because MIS can handle small sample sizes and covers both sides of the confidence level (Gneiting and Raftery, 2007).

We repeat this experiment 500 times. It is important to make error measures comparable and we use a percentage difference between the benchmark and the model of interest. The formula is shown as,

$$\text{dMSE} = \frac{\text{MSE}_\text{B} - \text{MSE}_\text{A}}{\text{MSE}_\text{B}},$$

where B is the benchmark, and A is the model of interest. Similar formulas apply to AME and MIS.

We also measure how well the smoothing parameters in $\boldsymbol{P}$ and the variances and covariances in $\boldsymbol{\Sigma}$ are. We measure dSmooth, a percentage difference of the true and the estimated smoothing parameters in $\boldsymbol{P}$, which is calculated as,

$$\text{dSmooth} = \begin{cases} \frac{1}{N^2K}\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{k=1}^{K}\frac{p_{ijk}-\hat{p}_{ijk}}{p_{ijk}}, & i,j \in N \\ \frac{1}{NK}\sum_{i=j=1}^{N}\sum_{k=1}^{K}\frac{p_{ijk}-\hat{p}_{ijk}}{p_{ijk}}, & i=j \in N \\ \frac{1}{N^{\dagger 2}K}\sum_{i=1,i\neq j}^{N^\dagger}\sum_{j=1,i\neq j}^{N^\dagger}\sum_{k=1}^{K}\frac{p_{ijk}-\hat{p}_{ijk}}{p_{ijk}}, & i \neq j \in N^\dagger. \end{cases}$$

where when $i,j \in N$ dSmooth includes all smoothing parameters in $\boldsymbol{P}$. When $i = j \in N$, it calculates the percentage difference of the diagonals of $\boldsymbol{P}$. When $i \neq j \in N^\dagger$, it calculates the percentage difference of the off-diagonals of $\boldsymbol{P}$, for $k \in K$ and $N^\dagger = \frac{1}{2}N(N+1)$. We also measure the percentage difference between the true and the estimated variances and covariances, as,

$$\text{dVar} = \frac{1}{N}\sum_{i=j=1}^{N}\frac{\sigma_{ii}^2 - \hat{\sigma}_{ii}^2}{\sigma_{ii}^2}, \quad \text{and,} \quad \text{dCov} = \frac{1}{N^\dagger}\sum_{i,j=1,i\neq j}^{N^\dagger}\frac{\sigma_{ij}^2 - \hat{\sigma}_{ij}^2}{\sigma_{ij}^2}.$$

3.6.2 Findings

Overall performance: univariate v.s. multivariate models

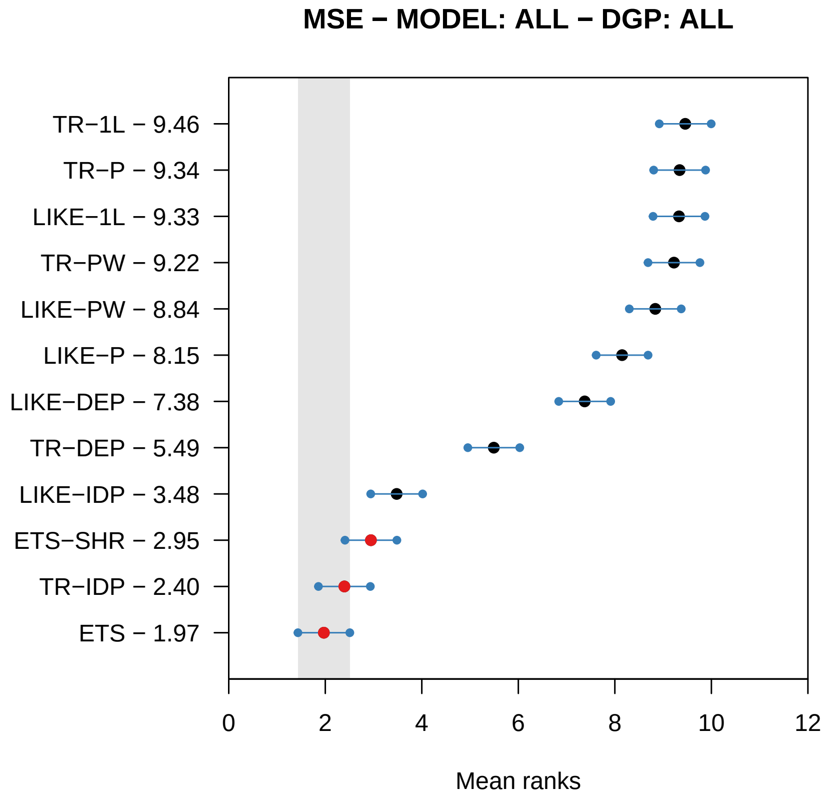


Figure 3.1: Overall performance of MSE ranks with an MCB test

Figure 3.1 presents the overall point forecast accuracy performance of 12 models across the data generating processes, model specifications, and forecast horizons, presented in a format of MCB test (Koning et al., 2005; Demšar, 2006). All models that intersect with the grey shaded area show no evidence of significant differences at 5% level, according to the Nemenyi test (Koning et al., 2005). We can see that ETS, TR-IDP, and ETS-SHR outperform the multivariate models. LIKE-IDP performs worse than the first three models but it outperforms the other LIKE models. Then, TR-DEP and LIKE-DEP follow, and the rest are the proposed shrinkage approaches. In this instance, forecasts from the proposed estimator are less accurate than the ones from ETS, TR-DEP, or LIKE-DEP. This shows the outperforming models with fewer estimated parameters in light of correlated time series and limited sample sizes. We separate the results in Figure 3.1 according to the model specifications and we present the first six models for each of them in Table 3.6. For the latter, we perform the non-parametric Nemenyi test for the MSE of each model specification, where bold model names denote the models with no evidence of statistical differences at 5% significance or having similar performance statistically. We can see that either ETS or VES with a diagonal $\boldsymbol{P}$ consistently outperform the other models. In the case of the correct specified models, TR-IDP and ETS perform similarly statistically. This shows that given the same model structure, different estimation approaches, i.e., simultaneously or independently, lead to different results. Further exploration is needed to understand why this happens. We also notice that ETS-SHR performs worse than ETS, which contradicts to the results in Pritularga et al. (2022). This happens because we apply ETS-SHR in a multivariate system, whereas Pritularga et al. (2022) apply ETS-SHR in the uncorrelated time series. Nonetheless, ETS-SHR performs well when the model is misspecified.

Rank	1	2	3	4	5	6
CR	**TR-IDP**	**ETS**	LIKE-IDP	TR-DEP	LIKE-DEP	ETS-SHR
OV	**ETS**	**ETS-SHR**	**TR-IDP**	LIKE-IDP	TR-DEP	TR-1L
UN	**ETS**	**TR-IDP**	**ETS-SHR**	LIKE-IDP	TR-DEP	LIKE-1L
IS	**ETS-SHR**	**LIKE-IDP**	**ETS**	**TR-IDP**	TR-PW	TR-DEP

Table 3.6: The first six performing models according to the MSE ranks.

Our findings are seemingly counterintuitive because univariate models outperform a multivariate model in a multivariate system, whereas the multivariate models are expected to perform well due to the DGP being indeed a multivariate one. In order to understand our finding, we demonstrate the distribution of the percentage difference of all smoothing parameters in $\boldsymbol{P}$, dSmooth, for each model for the case of the correctly-specified model, with DGP(DEP, CORREL). The result is shown in Figure 3.2, where the closer the value is to zero, the more accurate the smoothing parameter estimation is. A positive number shows shrinkage in smoothing parameters; otherwise. Specifically, ETS, ETS-SHR, TR-IDP, and LIKE-IDP contain the smoothing parameters of their time series, while the others contain all smoothing parameters in $\boldsymbol{P}$. Red dotted lines are the mean of each distribution, and red arrows denote some parts of the distribution that are not plotted.

We can see that the estimated smoothing parameters for ETS, ETS-SHR, TR-IDP, and LIKE-IDP have tight distributions, although the smoothing parameters are generally biased. On the other hand, for LIKE, the smoothing parameters are estimated well on average, but for some instances, they have large estimation errors. As for TR, most of them are biased but more efficient than LIKE. Regardless of the unbiasedness of the smoothing parameters, they tend to be inefficient in multivariate models, meaning that they have large estimation errors. This potentially leads to underperforming multivariate models on average. As LIKE has more estimated parameters, it depicts the true DGP, where the errors are correlated. However, incorporating too many parameters can be challenging, even though this depicts the reality we attempt to model. Note that here we use only 5 time series, and the estimation challenge is already very evident. Thus, there should be a balance between depicting reality fully and being a useful model to predict the future.

Prediction intervals

Figure 3.3 demonstrates the MCB test for MIS across forecast horizons, model specification, and DGPs. Similar to the point forecast accuracy, the top-performing models for MIS are ETS, TR-IDP, ETS-SHR, and LIKE-IDP. The next top-performing models with a similar rank are TR-DEP, TR-PW, and TR-P, and the rest are models estimated via likelihood. This shows that estimating covariances can be tricky and harm forecast accuracy. Table 3.7 presents dMIS between DEP as the benchmark and other multivariate models for DGP(DEP, CORREL). Positive numbers show that the

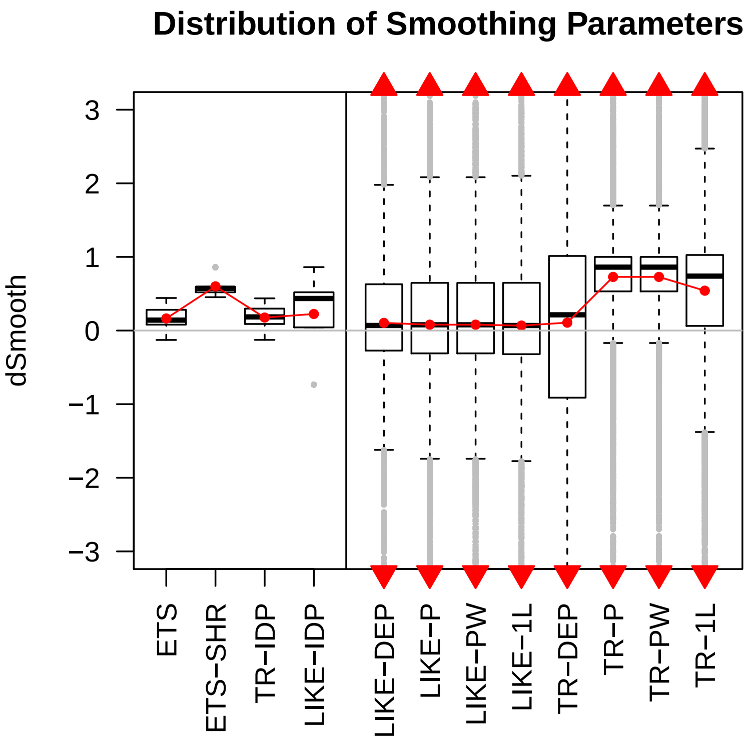


Figure 3.2: Distributions of smoothing parameters of each model for the correct specified model.

proposed approaches are better than DEP, and bold numbers show the best among the four models. LIKE-IDP and TR-IDP are still the best multivariate models compared to the others. Importantly, TR models perform better than LIKE models.

Loss	LIKE				TR			
Spec.	VES.IDP	VES.P	VES.PW	VES.1L	VES.IDP	VES.P	VES.PW	VES.1L
CR	**50.07**	-96.57	-96.57	-399.55	**40.81**	-15.50	-15.50	-125.75
OV	**78.69**	-250.40	-290.10	-150.29	**98.32**	3.20	-2.17	-2.96
UN	**12.78**	4.85	4.85	-19.94	**75.72**	70.31	70.31	70.44
IE	**90.34**	23.70	25.22	-57.63	**98.57**	3.54	0.99	0.14

Table 3.7: dMIS across confidence levels for the correct specified model only.

We connect Figure 3.3 and Table 3.7 with Eq. (3.11) to discuss estimated variances, covariances, and smoothing parameters. The first three models, namely ETS, TR-IDP, and ETS-SHR, have independent model structures and zero covariances. When the forecast variance-covariance relies on $\boldsymbol{P\Sigma P}^\top$, then the multiplication becomes less complicated. This can mitigate any potential issues of parameter estima-

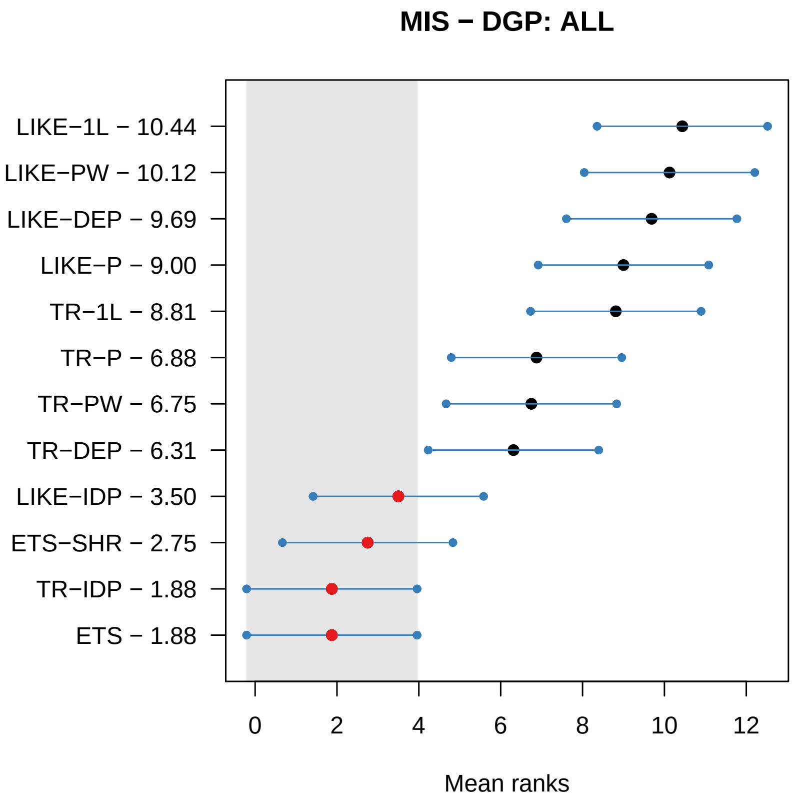
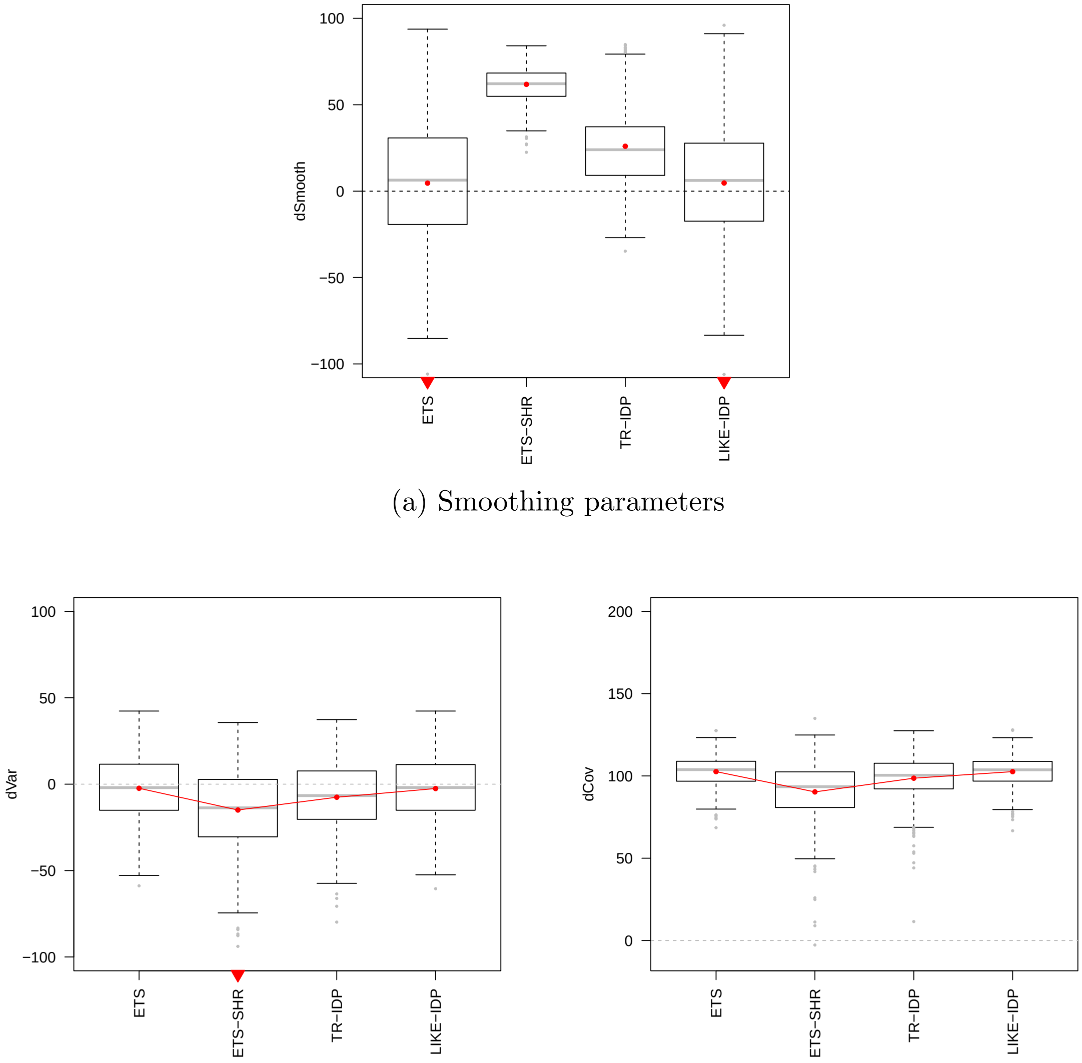


Figure 3.3: An MCB test for MIS across forecast horizons, model specifications, and DGPs

tion uncertainty. The next group consists of TR-DEP, TR-PW, and TR-P, where the covariances in $\boldsymbol{\Sigma}$ are omitted, but the smoothing parameters in $\boldsymbol{P}$ remain. This also shows that forcing some covariances to zero may reduce the estimation errors, thus providing accurate prediction intervals. This group certainly outperforms the LIKE models. Thus, we also see a clear benefit of the fixed shrinkage in $\boldsymbol{\Sigma}$.

Parameter estimation accuracy



(a) Smoothing parameters

(b) Residual variances

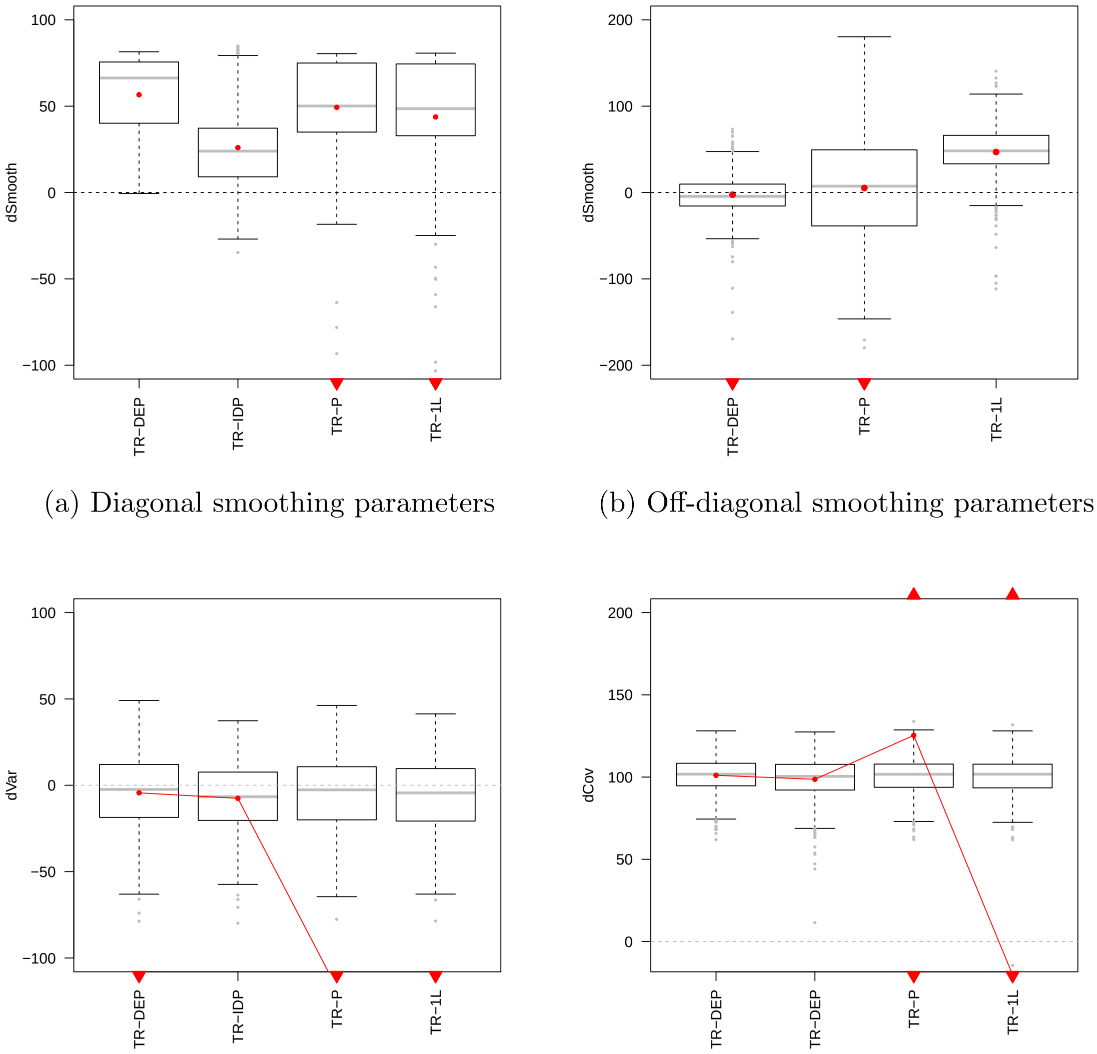
(c) Residual covariances

Figure 3.4: Percentage differences between the true and estimated smoothing parameters, variances, and covariances, for the correct specified model with independent model structures such as ETS, ETS-SHR, TR-IDP, and LIKE-IDP.

We observe that omitting the off-diagonals of $\boldsymbol{P}$ and $\boldsymbol{\Sigma}$ results in more accurate prediction intervals. It is worth inspecting the estimated diagonals and the off-diagonals of $\boldsymbol{P}$, and the estimated residual variances and covariances in $\boldsymbol{\Sigma}$, shown in Figure 3.4. A positive number shows that the estimated smoothing parameter is lower than the true one. A more than 100% dSmooth represents a change of direction in the parameters, for example the estimated one is negative whereas the true one is positive. Negative dVar and dCov show that the estimated ones are larger than the true ones. The dots represent the average of each distribution, and the red arrows show that some parts of the distribution are not plotted. This description applies to Figure 3.5 and 3.6.

We discuss the dSmooth of the top-performing models, where the smoothing parameters are of each time series. ETS-SHR produces lower smoothing parameters due to shrinkage than ETS, TR-IDP, and LIKE-IDP. Moreover, we observe some shrinkage in TR-IDP, whereas ETS and LIKE-IDP estimate unbiased smoothing parameters on average. Looking at the estimation of $\boldsymbol{\Sigma}$, the shrinkage in ETS-SHR results in increased variances and covariances. This may explain why ETS-SHR does not perform better than ETS, as the estimated residual variances and covariances compensate for the smoothing parameter shrinkage. On the other hand, on average, ETS and LIKE-IDP result in zero covariances because the smoothing parameters are estimated well. Hence, there is a possible connection between smoothing parameters, variances, and covariances, where all parameters in $\boldsymbol{P}$ and $\boldsymbol{\Sigma}$ have to be estimated optimally to avoid any compensating effects.

Next we discuss the parameter accuracy for TR only, shown in Figure 3.5. We can see that the diagonals of $\boldsymbol{P}$ have a shrinkage effect, even though TR-IDP has the smallest positive dSmooth. Still, the estimated diagonals of $\boldsymbol{P}$ in TR-P and TR-1L are unnecessarily large for some instances. Note that we do not include TR-PW because TR-P and TR-PW, in the case of the correct specification, have the same results, i.e. $\boldsymbol{PW}$ and $\boldsymbol{P}$ for VES(A,N,N) are the same. Only TR-1L can significantly shrink the off-diagonal of $\boldsymbol{P}$. On the other hand, TR-DEP and TR-P estimate the off-diagonals well with a higher variability for the latter. Regarding the variance and the covariance estimation, TR-DEP, TR-IDP, TR-P, and TR-1L have similar interquartile



(a) Diagonal smoothing parameters

(b) Off-diagonal smoothing parameters

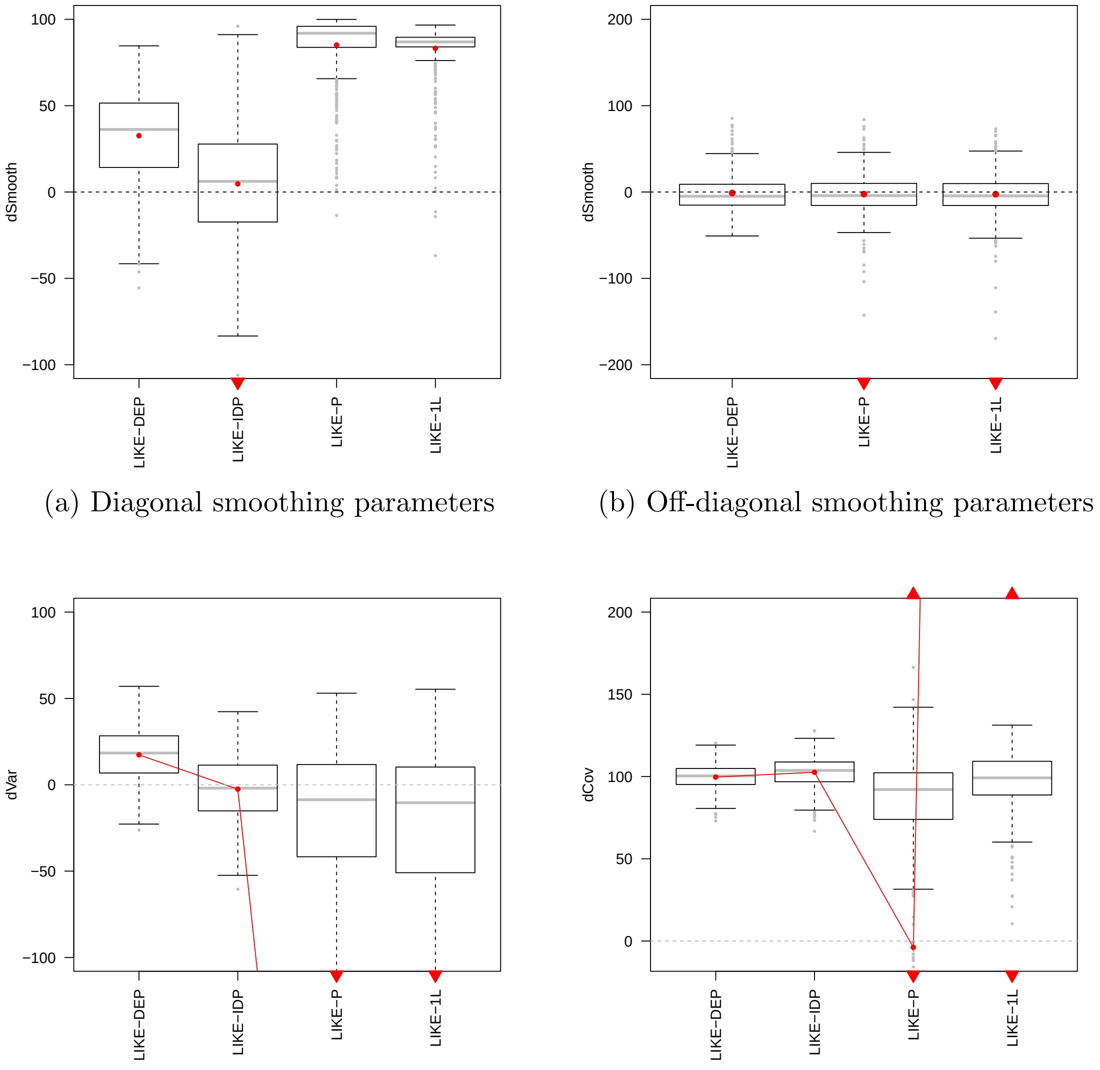
(c) Residual variances

(d) Residual covariances

Figure 3.5: Percentage differences between the true and estimated smoothing parameters, variances, and covariances, for the correct specified models with a dependent model structure and a trace loss function.

ranges. In some instances, the estimated variances and covariances of TR-P and TR-1L explode. We have evidence of a relatively complicated connection between smoothing parameters, variances, and covariances. This finding also highlights the potential for misbehaving covariance matrix when shrinkage in $\boldsymbol{P}$ is implemented, especially when TR is used in DGP(DEP, CORREL).

We now focus on models with the likelihood (LIKE), where the covariances are



(a) Diagonal smoothing parameters

(b) Off-diagonal smoothing parameters

(c) Residual variances

(d) Residual covariances

Figure 3.6: Percentage differences between the true and estimated smoothing parameters, variances, and covariances, for the correct specified models with a dependent model structure and a maximum likelihood loss function.

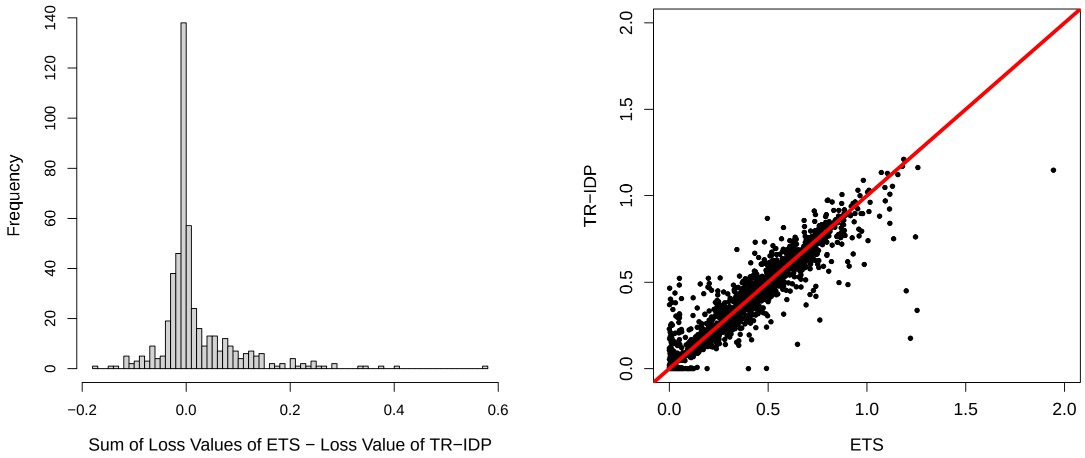
included in the loss function. On average, LIKE-IDP can estimate the diagonals of $\boldsymbol{P}$ well with quite significant variability, while there is a slight shrinkage effect in LIKE-DEP. However, LIKE-P and LIKE-1L are able to shrink the diagonals of $\boldsymbol{P}$ significantly, about 80%, with some outliers. On the other hand, on average, LIKE-DEP, LIKE-P and LIKE-1L can estimate the off-diagonal of $\boldsymbol{P}$ well.

In contrast to the smoothing parameter estimation, the distribution of dVar and dCov reveals some interesting patterns. LIKE-DEP results in smaller estimated variances than the true ones and are able to estimate the covariance to zero. This shows a possibility of an overfit model because the variance is unnecessarily smaller than the true one. For LIKE-IDP, the variances are estimated well on average with some variability. However, LIKE-P and LIKE-1L result in unnecessarily large estimated variances with relatively high variability. As for the dCov, we can see that LIKE-P and LIKE-1L have decent interquartile, but the averages are highly affected by outliers. This means that some of the estimated covariances are unnecessarily high in either direction, i.e., the values of the covariances can be large, either in a positive or a negative direction. Thus, in the DGP(DEP, CORREL) setting, it is important to include $\boldsymbol{\Sigma}$ in the shrinkage loss function to avoid large estimation errors either in $\boldsymbol{P}$ or $\boldsymbol{\Sigma}$.

A comparison between ETS and TR-IDP

Having discussed the overall performance, we are interested in discussing the performance of ETS and TR-IDP. Both models are a specific case of the fixed shrinkage, where both have the same structure but differ in their estimation. ETS estimates the parameters of each model independently, whereas TR-IDP estimates the parameters simultaneously. In discussing these findings, we use DGP(IDP, UNCORREL) specifically to avoid additional uncertainties. Figure 3.7 shows evidence that the two models have different estimated parameters. A red line on Panel (b) denotes an equivalence between parameters. Figure 3.7a demonstrates the percentage difference between the sum of ETS's loss function and the TR-IDP's loss function. The figure shows discrepancies between the loss values of ETS and TR-IDP. Ideally, they would have had the same loss value, assuming that the optimisation routine worked the same.

In addition, we show a comparison between the smoothing parameters of both models. Figure 3.7b shows that the smoothing parameters of ETS have higher vari-



(a) The difference between TR and the sum of MSEs

(b) Smoothing parameters of ETS and TR-IDP

Figure 3.7: A comparison between ETS and TR-IDP, in their loss values and smoothing parameters.

ability than those of TR-IDP. Since each time series is modelled independently, there is a chance that the smoothing parameters of ETS are more freely estimated than those of TR-IDP. TR-IDP estimates the smoothing parameters in a bundle, which means that they are estimated together in the optimisation routine. The parameters do not have equal freedom to find the optimal values. These results hold when we used the same optimisers for the two models with BOBYQA and Nelder-Mead (Nelder and Mead, 1965; Johnson, 2022). This potentially affects the parameter estimation error and impacts forecast accuracy. Admittedly, optimisation quality has been largely ignored in the forecasting literature. Our results show that, for the multivariate models, it can have a very substantive impact.

3.7 Discussion and Conclusion

We propose a shrinkage approach with several penalty functions to mitigate smoothing parameter estimation uncertainty in VES. Our findings demonstrate that the proposed approaches are able to shrink the smoothing parameters effectively; however, they do not produce accurate point and interval forecasts as expected. Another finding, rather counterintuitive, in our experiment, is that the univariate models outperform the multivariate models in the multivariate setting. Here, we discuss a possible explanation as to why our approaches do not result in forecast accuracy improvement and why the univariate models outperform the multivariate ones.

We argue that our approach performs as intended, i.e., shrinking the smoothing parameters. In principle, this should lead to better performance, as shown by (Pritularga et al., 2022) for the univariate case. However, this increases the estimation error of $\boldsymbol{\Sigma}$, and we observe a 'compensating effect' between $\boldsymbol{P}$ and $\boldsymbol{\Sigma}$. For example, as the estimated off-diagonals of $\boldsymbol{P}$ are shrunk, the covariances in $\boldsymbol{\Sigma}$ increase for some instances. This indicates a potential identification issue in estimating both $\boldsymbol{P}$ and $\boldsymbol{\Sigma}$. If this identification issue persists, a model restriction such as a fixed shrinkage in $\boldsymbol{P}$ and $\boldsymbol{\Sigma}$ is a potential solution, which leads to the more accurate point and interval forecasts, as the models are able to estimate the smoothing parameters efficiently in the statistical sense.

There are several choices in modelling a multivariate system with a limited sample size. First, we can employ the fixed shrinkage in a multivariate model. For example, TR-IDP estimates the smoothing parameters well and can produce accurate point and interval forecasts. The other option is to employ a set of ETS or ETS-SHR. These align well with the previous studies in VES, where implementing TR is sufficient to produce accurate forecasts (de Silva et al., 2010; Svetunkov et al., 2022a). However, implementing ETS-SHR in a multivariate correlated time series needs to be done with care as the smoothing parameter shrinkage can also increase the estimated residual covariances.

One may argue that we can incorporate the interdependence in $\boldsymbol{P}$ and the fixed shrinkage in $\boldsymbol{\Sigma}$, i.e. TR-DEP. We see that there is an unmitigated shrinkage in the diagonals of $\boldsymbol{P}$, but at the same time, some estimated variances are too large. The diagonal smoothing parameter shrinkage potentially increases the estimated variances. Thus, this is not an option either.

The previous examples show that we should shrink $\boldsymbol{P}$ and $\boldsymbol{\Sigma}$ simultaneously to avoid any compensating effect between both parameters. Hence, it is important to shrink or regularise $\boldsymbol{\Sigma}$ in the multivariate framework. Similar studies have been proposed by Wilms and Croux (2018); Wilms et al. (2021) to shrink the autoregressive coefficient and $\boldsymbol{\Sigma}$ in a VAR model. Our study complements Wilms and Croux (2018); Wilms et al. (2021). We explain why we not only need to parameterise the coefficients but also $\boldsymbol{\Sigma}$ in a multivariate time series model. We also highlight potential estimation issues when only shrinkage is implemented in one of them.

An alternative to incorporating cross-learning information into the forecasts is done via forecast reconciliation. Forecast reconciliation combines forecasts according to some linear restrictions (Athanasopoulos et al., 2009; Hyndman et al., 2011, 2016; Wickramasuriya et al., 2019; Athanasopoulos et al., 2020; Panagiotelis et al., 2021) via an ill-posed least square estimation. Forecast reconciliation is a multivariate problem with a two-step estimation, i.e., (1) producing independent forecasts via univariate models and (2) estimating a reconciliation weight matrix to linearly combine the forecasts, where the matrix incorporates the cross-learning information from the residuals of the univariate models. We mitigate the model uncertainty issues by utilising the 'left-over' information captured by the residuals to combine the forecasts. Nonetheless, in a multivariate system with limited sample sizes, we have three alternatives: (a) implementing ETS, ETS-SHR, or TR-IDP models, or (b) implementing LIKE-DEP with shrinkage in $\boldsymbol{P}$ and $\boldsymbol{\Sigma}$, or (c) implementing (a) and then employing forecast reconciliation to incorporate the cross-learning information.

Our study focuses on estimating univariate and multivariate models in a multivariate system with limited sample sizes. To some extent, it is also important to provide evidence when the sample size is large enough to observe how they behave asymptotically. This will give us a bigger picture and a fair treatment of LIKE-DEP or LIKE-1L regarding how we handle a multivariate model with a relatively large sample size.

It is also worth discussing the difference in performance between ETS and TR-IDP. Both have the same model structure but have different estimation approaches. We expect both to perform similarly, but our findings show that, in some cases, they differ. We suspect this is because of how the optimisation operates differently on ETS and TR-IDP. TR-IDP has more parameters to estimate than each univariate ETS. In the case of the Nelder-Mean algorithm, as the number of parameters increases, the progress in finding the local optima becomes smaller for each iteration. Eventually, the search stops (Han and Neumann, 2006; Conn et al., 2009). We can confirm this argument from our results. In ETS, the parameters have more freedom to reach the local minimum, and the estimated parameters exhibit larger movement from the initial values.

On the other hand, the estimated parameters of TR-IDP are less diverse than the ETS ones due to the little progress for each iteration. We argue for more explorations of the optimisation algorithm implications on multivariate models. It is important to investigate the properties and behaviour of the loss function in a univariate and a multivariate state-space model in more detail. A similar study has been done by Farnum (1992), where they studied the loss function of a Simple Exponential Smoothing when the smoothing parameter is close to 0 and 1. Understanding the properties and the behaviour of the loss function in depth enables us to devise an optimisation algorithm suitable for the specific problem. This may lead to better parameter estimation and eventually improve forecasting performance.

Second, it is also possible to investigate the stability and forecastability conditions in VES. These conditions allow negative smoothing parameters. A negative smoothing parameter results in a harmonic pattern, whereas states ETS itself, by nature, have a monotonic pattern if the smoothing parameters are between 0 and 1. Tsay (2014) notes that the stability condition ensures the eigenvalues decay exponentially, but not in individual elements in $\boldsymbol{Z}$ (see Eq. (3.7)). Whether negative smoothing parameters impact the forecasting performance needs to be explored further. We argue that we may need to introduce additional conditions or restrictions on top of forecastibility for either ETS or VES. For example, one could force the smoothing parameters to be positive so that both the eigenvalues and the elements inside the discount matrix decay exponentially. This potentially affects the smoothing parameters' parameter spaces, which might be useful when we have limited sample sizes. This, along with other alternatives, needs to be investigated further.

In conclusion, a shrinkage implementation in VES is not as straightforward as the one in ETS. There are interactions between smoothing parameters in $\boldsymbol{P}$ and $\boldsymbol{\Sigma}$ and between them, which lead to a compensating effect between $\boldsymbol{P}$ and $\boldsymbol{\Sigma}$. Due to the relationships between parameters in VES, shrinkage should be implemented in both $\boldsymbol{P}$ and $\boldsymbol{\Sigma}$. Alternatively, we can employ a set of ETS models or VES with a diagonal $\boldsymbol{P}$ and $\boldsymbol{\Sigma}$ to mitigate the parameter uncertainty. Then if cross-learning is needed, forecast reconciliation can be used. Apart from that, the choice between ETS and TR-IDP is still an open question. The numerical optimisation for the multivariate models needs more exploration.

Chapter 4

Stochastic Coherency in Forecast Reconciliation

The previous chapters discuss the efficacy of parameter shrinkage for ETS and VES. We show that we need to be careful in employing a multivariate model in a multivariate time series. Alternatively, one can use the forecast reconciliation, as several studies demonstrate its efficacy.

In this chapter we analyse the forecast reconciliation in detail, especially its sources of uncertainties. We propose a concept of 'stochastic coherency' to accommodate the overlooked uncertainties. We also propose several covariance matrix approximations to mitigate this issue. All the materials in this chapter are based on an article published in the International Journal of Production Economics (Pritularga et al., 2021).

Abstract

Hierarchical forecasting has been receiving increasing attention in the literature. The notion of coherency is central to this, which implies that the hierarchical time series follows some linear aggregation constraints. This notion, however, does not take several modelling uncertainties into account. We propose to redefine coherency as stochastic. This allows to accommodate overlooked uncertainties in forecast recon-

ciliation. We show analytically that there are two potential sources of uncertainty in forecast reconciliation. We use simulated data to demonstrate how these uncertainties propagate to the covariance matrix estimation, introducing uncertainty in the reconciliation weights matrix. This then increases the uncertainty of the reconciled forecasts. We apply our understanding to modelling accident and emergency admissions in a UK hospital. Our analysis confirms the insights from stochastic coherency in forecast reconciliation. It shows that we gain accuracy improvement from forecast reconciliation, on average, at the cost of the variability of the forecast error distribution. Users may opt to prefer less volatile error distributions to assist decision making.

4.1 Introduction

Forecasting is an essential activity for decision making in organisations. Often forecasts and supported decisions are organised in hierarchies. These hierarchies can be constructed from market segments, products, or other demarcations (Athanasopoulos et al., 2009). Beyond cross-sectional hierarchies, there are temporal hierarchies where different functions in an organisation require forecasts at different sampling frequencies and planning horizons (Athanasopoulos et al., 2017). Combining both is also possible, which aims to provide a coherent view of the future across both dimensions (Kourentzes and Athanasopoulos, 2019).

All the hierarchical forecasting methods are based on the property of coherency (Wickramasuriya et al., 2019; Jeon et al., 2019; Taieb et al., 2020; Athanasopoulos et al., 2020). It implies that the lower level forecasts add up to the forecasts of the higher levels. For example, sales of individual products in a hierarchy sum up to product category sales at higher levels, in observations and forecasts. When forecasts are produced independently, they are typically not coherent, and this has been one of the motivations for developing hierarchical forecasting methods. This enables aligned planning and actions throughout organisations and stake-holders (Kourentzes and Athanasopoulos, 2019), which is the main motivation for hierarchical forecasting in an organisational context. The concept of coherency has been central in temporal dissagregation, establishing a link between high and low frequency time series. For example, Chow and Lin (1971) uses a highly restrictive generalised linear regression model to this purpose.

In the past, studies attempted to tackle this problem by employing a bottom-up or top-down approach (Fliedner, 2001). The main issue with these methods is that they ignore information either at higher levels or lower levels (Athanasopoulos et al., 2009; Ord et al., 2017), thus leading to less accurate forecasts. Furthermore, implicitly we accept increased modelling risk, as all forecasts in the hierarchy are based on a single (top-down) or a few (bottom-up) forecasting models, which may be misspecified. This misspecification can have adverse effects on the uncertainty of the forecasts across the hierarchy, resulting in increased costs of any supported decisions, such as unmet demand due to poor forecasts.

Nowadays, hierarchical forecasting is seen as a reconciliation problem, where forecasts are generated at all levels and then are reconciled to a common view of the future (Hyndman et al., 2011, 2016; Wickramasuriya et al., 2019). Several studies have shown significant improvements in forecast accuracy in different contexts, when hierarchical reconciliation techniques are used (Yang et al., 2016; Oliveira and Ramos, 2019; Kourentzes and Athanasopoulos, 2019, 2021). In brief, forecast reconciliation is achieved by linearly combining all forecasts from the hierarchy to a set of adjusted bottom-level forecasts, which by construction make use of all available information, and then aggregating these to reconciled forecasts for the complete hierarchy. Note that since we no longer rely on forecasts at any specific level, we mitigate uncertainties stemming from the specification of the forecasting methods that both top-down and bottom-up methods suffer from. Apart from providing coherent forecasts, the reconciliation methods also often improve upon the accuracy of the base independent forecasts in cross-sectional, temporal, and cross-temporal hierarchies (Hyndman et al., 2011; Athanasopoulos et al., 2017; Wickramasuriya et al., 2019; Kourentzes and Athanasopoulos, 2019; Kourentzes et al., 2021).

Nonetheless, in the literature there is empirical evidence that hierarchical forecasting does not universally result in reduced forecast uncertainty and better forecast accuracy. In temporal hierarchies, Athanasopoulos et al. (2017) demonstrate that model selection uncertainty affects the efficacy of forecast reconciliation. Base forecasts from well-specified forecasting models gain little benefit from forecast reconciliation, whereas forecasts from mis-specified models are improved significantly. Furthermore, Kourentzes and Athanasopoulos (2019) find that combining cross-sectional and temporal hierarchies offers 'small yet significant' improvement upon the accuracy of the base forecasts, as the first dimension, the temporal, already mitigates much of the uncertainty in base forecasts. However, in order to reconcile forecasts, a reconciliation weights matrix is needed, and defining this matrix in the cross-temporal case can be challenging (see also di Fonzo and Girolimetto, 2021). In cross-temporal hierarchies, Kourentzes and Athanasopoulos (2019) average across multiple estimates of the reconciliation weights matrix to avoid unnecessary estimation uncertainty, while retaining coherency. Results in Panagiotelis et al. (2021) show substantially different forecast error variances depending on how the reconciliation weights matrix is calculated. Empirical results from the retail and tourism sectors further demonstrate this variability of performance that appears to depend on the calculation of the reconciliation weights (Wickramasuriya et al., 2019; Oliveira and Ramos, 2019). We argue that there are inherent uncertainties in forecast reconciliation that have not been explored in the literature, which we investigate here.

Thus, it appears that recent studies have overlooked the effect of uncertainties in forecast reconciliation. Panagiotelis et al. (2022, Theorem 3.1) demonstrate that the only source of uncertainty is originating from the base forecasts, and the reconciliation weights matrix is assumed to have no uncertainty and effectively treated as known. Nevertheless, since we estimate the reconciliation weights, we face uncertainty in their estimation. Furthermore, as there are different approximations for the covariance matrix (Hyndman et al., 2011, 2016; Athanasopoulos et al., 2017; Wickramasuriya et al., 2019; Nystrup et al., 2020), this leads to a selection question. Thus, the conventional reconciled forecast variance is potentially underestimated. Note that we consider parameter estimation and forecasting method selection uncertainties as two aspects of the same modelling issue.

Apart from that, there is another complication with hierarchical time series. There is a gap between how the hierarchical time series are collected in practice and how we use the data for forecasting. Suppose that we see the original information coming from the bottom-level of the hierarchy. For example, in macroeconomic variables the data is collected either by surveys, estimates, or a combination of them. As the data are collected for different nodes or levels of the hierarchy, the bottom level does not always add up to the higher levels of the hierarchy. Statistics bureaux use the account ‘statistical discrepancy’ to fill the gap. Athanasopoulos et al. (2020) treat the discrepancy as another time series in forecast reconciliation. This affects how we perceive coherency in hierarchical time series, both in the observational and population levels, as well as how we understand modelling uncertainty in forecast reconciliation.

In order to address all these issues, we propose the notion of “stochastic coherency”. Stochastic coherency is easy to understand if we use the geometric interpretation of forecast reconciliation (Panagiotelis et al., 2021). Incoherent base forecasts (the initial forecasts for each node of the hierarchy) are projected to a coherent subspace. Conventionally, this projection has no uncertainty, while with stochastic coherency, the projection becomes stochastic. Equivalently, if we see forecast reconciliation from a forecast combination interpretation (e.g., Kourentzes and Athanasopoulos, 2019), the combination weights are stochastic.

Suppose we have a set of forecasts from a sample of time series and a hierarchy is given. When we collect additional samples and re-estimate the reconciliation weights matrix, that is bound to change due to the estimation of covariance matrix approximations. However, as long as the estimated reconciliation weights matrix meets the coherency constraint (Wickramasuriya et al., 2019), the forecasts are coherent, but they will change as the weight matrix changes. The key here is to acknowledge that the uncertainty of coherent forecasts comes from the incoherent base forecasts and propagates to the estimation of the reconciliation weights matrix. This increases the uncertainty of coherent forecasts.

Another differentiating characteristic of stochastic coherency is how the error terms in the data generating process are treated. We realise that the error term in the hierarchical time series itself may contain not only the innovations but also potential errors coming from data collection, such as sampling and measurement errors. On top of that, modelling uncertainty is introduced when we produce forecasts. Hence, it allows us to decompose the variance of coherent forecasts. As we show later in the paper, this has important implications for the construction of the estimated covariance matrix and coherent forecasts.

Stochastic coherency affects not only point forecasts but also probabilistic forecasts. In order to understand the effect of stochastic coherency on probabilistic coherent forecasts better, we refer to its definition by Taieb and Koo (2019) and Panagiotelis et al. (2022). The former defines the probabilistic coherent forecasts as convolutions of linear constraints, while the latter defines them in a more flexible manner as to extend to non-linear constraints (Panagiotelis et al., 2022, p. 8). However, both definitions are rooted from the idea of forecast reconciliation where the base point forecasts are projected onto the coherent space by the reconciliation weights matrix, and they assume that the weights matrix is known. Our stochastic coherency highlights the uncertainty of these weights, or the projection, and it will affect both the point and the probabilistic forecasts. This will increase the coherent probabilistic forecasts uncertainty. We argue that these are important characteristics in the application of hierarchical forecasting in organisations, where understanding and controlling the sources of uncertainty is important for mitigating risks associated with decision making, beyond any accuracy improvements.

We explore stochastic coherency in detail in Section 4.2 and 4.3. We show when forecast reconciliation becomes beneficial, and we explore uncertainties in forecast reconciliation further, attributing them to their sources. We find that the more complete the covariance matrix approximation is, the better the resulting point forecast accuracy can be but at the cost of the increased variance of the reconciled forecast errors. In Section 4.4, we conduct a simulation experiment to validate our understanding. In Section 4.5 we apply this to modelling accident and emergency admissions at a UK hospital, demonstrating the effect of stochastic coherency on a real complex problem. Based on these findings, we discuss and conclude our work in Section 4.6.

4.2 Classical and Stochastic Coherency

The notion of coherency in hierarchical forecasting has been proposed and elaborated by a series of hierarchical forecasting works (Athanasopoulos et al., 2009; Hyndman et al., 2011, 2016; Wickramasuriya et al., 2019; Panagiotelis et al., 2021; Athanasopoulos et al., 2020). The literature defines forecasts as coherent forecasts if they adhere to a linear constraint, e.g. they add up according to the hierarchy, often simplified as the bottom level forecasts aggregating to the higher level forecasts. In a similar manner, Taieb et al. (2020) define mean coherent forecasts when the errors between aggregated bottom-level forecasts and the independent forecasts at the upper-level are zero.

Let us explore the hierarchical approach in detail. Suppose that $\boldsymbol{y}_t$ is a $N \times 1$ vector of hierarchical time series across the hierarchy, at period t, where $\boldsymbol{y}_t$ is constructed from $\boldsymbol{b}_t$, a $m \times 1$ vector of the bottom-level time series, and a summation matrix, $\boldsymbol{S}$. In this case, $\boldsymbol{S}$ maps the bottom-level onto the upper-level of the hierarchy. The coherent hierarchical time series is denoted as,

$$\boldsymbol{y}_t = \boldsymbol{S}\boldsymbol{b}_t. \tag{4.1}$$

We argue that Eq (4.1) is not general. Let us consider how time series data is collected in different organisations. In any retailer which records demand of every stock keeping unit at the bottom-level in real-time, they can update new information in the middle and the top of hierarchy at time t, across the hierarchy, instantaneously. This means that the hierarchical time series are coherent, even as new information becomes available. On the other extreme, we may need to estimate the data, even though by nature it is a part of a hierarchy, for example the gross domestic product (GDP). For instance, the Office of National Statistics United Kingdom measures national accounts through surveys, forecasts, and estimates from models, which are subject to errors (Office for National Statistics, 2011). Once an account is measured, they need to reconcile the number from different methods and sources. Hence, in the case of GDP, the values in the hierarchy from aggregating the bottom-level data and collecting data from each level will be different. To accommodate the potential gap, the statistical bureaux create an account called statistical discrepancy (Australian Bureau of Statistics, 2015, p. 471). This discrepancy captures any potential error coming from measurement and sampling errors. Athanasopoulos et al. (2020) treat the discrepancy as a time series. It is easy to identify scenarios where such measurement issues violate the classical coherency, from individual companies, to national statistics. Therefore, due to the measurement errors, we redefine hierarchical time series $\boldsymbol{y}_t$ as,

$$\boldsymbol{y}_t = \boldsymbol{S}\boldsymbol{b}_t + \boldsymbol{\delta}_t,$$

where $\boldsymbol{\delta}_t$ is the statistical discrepancy at time t. By nature, $\boldsymbol{\delta}_t$ is zero when data collection is done perfectly and able to measure the variables of interest accurately.

First, we discuss the time series in population. Suppose that we know the true data generating process of $\boldsymbol{b}_t$, which has an additive state-space structure. We use this framework illustratively and we are not restricted to it. Nonetheless, the state-space modelling framework is very flexible and encompasses many popular forecasting model families. Let:

$$\boldsymbol{b}_t = \boldsymbol{\mu}_{b,t} + \boldsymbol{\eta}_{b,t}, \tag{4.2}$$

where $\boldsymbol{\mu}_{b,t}$ denotes the structure of the time series and $\boldsymbol{\eta}_{b,t}$ is the innovation term at period t, which for simplicity follows a multivariate normal distribution with zero mean and has a covariance matrix of $\boldsymbol{\Sigma}_b$. We aggregate $\boldsymbol{b}_t$ by multiplying with $\boldsymbol{S}$ and from Eq (4.2), and we get $\boldsymbol{y}_t$ as a vector time series,

$$\begin{aligned}\boldsymbol{y}_t &= \boldsymbol{S}\boldsymbol{b}_t + \boldsymbol{\delta}_t \\ &= \boldsymbol{S}\boldsymbol{\mu}_{b,t} + \boldsymbol{S}\boldsymbol{\eta}_{b,t} + \boldsymbol{\delta}_t \\ &= \boldsymbol{\mu}_t + \boldsymbol{\varepsilon}_t \end{aligned} \tag{4.3}$$

where $\boldsymbol{\varepsilon}_t$ is the total residual of the process, which consists of the aggregated innovations and the statistical discrepancy, denoted as $\boldsymbol{\varepsilon}_t = \boldsymbol{S}\boldsymbol{\eta}_{b,t} + \boldsymbol{\delta}_t$, and $\boldsymbol{\mu}_t = \boldsymbol{S}\boldsymbol{\mu}_{b,t}$. In this case, we assume that $\mathrm{E}(\boldsymbol{\delta}_t|\boldsymbol{\mathcal{I}}_t) = \mathbf{0}$, and from definition $\mathrm{E}(\boldsymbol{\eta}_{b,t}|\boldsymbol{\mathcal{I}}_t) = \mathbf{0}$, thus

$\mathrm{E}(\boldsymbol{S}\boldsymbol{\eta}_{b,t}|\boldsymbol{\mathcal{I}}_t) = \mathbf{0}$. In expectation, Eq (4.3) becomes,

$$\begin{aligned}\mathrm{E}(\boldsymbol{y}_t|\boldsymbol{\mathcal{I}}_t) &= \mathrm{E}(\boldsymbol{S}\boldsymbol{b}_t + \boldsymbol{\delta}_t|\boldsymbol{\mathcal{I}}_t)\\ &= \mathrm{E}(\boldsymbol{S}\boldsymbol{\mu}_{b,t}|\boldsymbol{\mathcal{I}}_t) + \mathrm{E}(\boldsymbol{S}\boldsymbol{\eta}_{b,t}|\boldsymbol{\mathcal{I}}_t) + \mathrm{E}(\boldsymbol{\delta}_t|\boldsymbol{\mathcal{I}}_t)\\ &= \mathrm{E}(\boldsymbol{\mu}_t|\boldsymbol{\mathcal{I}}_t)\end{aligned}$$

where $\boldsymbol{\mathcal{I}}_t$ is the available information at t. We can also infer that $\boldsymbol{\mu}_t = \boldsymbol{S}\boldsymbol{\mu}_{b,t}$ and this also holds at period $t+h$. This shows that the time series is coherent in expectations, meaning that the linear hierarchical structure, $\boldsymbol{S}$, guarantees coherency in the structures of the time series, but does not necessarily guarantee coherency in the residuals.

In observations, we exploit $\boldsymbol{\mathcal{I}}_t$ by differentiating between the type of the information, namely $\boldsymbol{\Theta}$ as a set of forecasting models in the hierarchy, and $\boldsymbol{Y}_t$ as the available hierarchical time series, where $\boldsymbol{Y}_t = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_t\}$. Note that $\boldsymbol{\Theta}$ is not restricted to a single family of forecasting models and can be different forecasting models or methods for each series across the hierarchy.

Using forecasting models $\boldsymbol{\Theta}$, we produce h-step ahead base forecasts. The forecasts, typically, adhere to the classical coherency, but are inaccurate. Following the hierarchical forecasting literature, we can reconcile base forecast as:

$$\tilde{\boldsymbol{y}}_{t+h|t} = \boldsymbol{S}\boldsymbol{G}\hat{\boldsymbol{y}}_{t+h|t}, \tag{4.4}$$

where $\tilde{\boldsymbol{y}}_{t+h|t}$ is h-step ahead reconciled forecast, and $\boldsymbol{G}$ is a reconciliation weights matrix, which combines all forecasts across the hierarchy to create adjusted bottom-level forecasts. As $\boldsymbol{S}$ and $\hat{\boldsymbol{y}}_{t+h|t}$ are available prior to the reconciliation, $\boldsymbol{G}$ is estimated. Wickramasuriya et al. (2019) propose the MinT Reconciliation to obtain $\boldsymbol{G}$, by minimising the trace of covariance matrix of the reconciled forecast error

$(\tilde{\boldsymbol{e}}_{t+h|t} = \boldsymbol{y}_{t+h|t} - \tilde{\boldsymbol{y}}_{t+h|t})$, instead of reconciliation error $(\boldsymbol{\epsilon}_{t+h|t} = \hat{\boldsymbol{y}}_{t+h|t} - \tilde{\boldsymbol{y}}_{t+h|t})$:

$$\min \operatorname{Tr}\left(\boldsymbol{S}\boldsymbol{G}\boldsymbol{W}_{t+h|t}\boldsymbol{G}^{\top}\boldsymbol{S}^{\top}\right)$$

subject to $\boldsymbol{S}\boldsymbol{G}\boldsymbol{S} = \boldsymbol{S}$, or alternatively $\boldsymbol{G}\boldsymbol{S} = \boldsymbol{I}$, where $\boldsymbol{W}_{t+h|t} = \mathrm{E}(\hat{\boldsymbol{e}}_{t+h|t}\hat{\boldsymbol{e}}_{t+h|t}^{\top}|\boldsymbol{\mathcal{I}}_t)$ and $\hat{\boldsymbol{e}}_{t+h|t}$ is the h-step ahead base forecast error, $\boldsymbol{y}_{t+h} - \hat{\boldsymbol{y}}_{t+h|t}$. They show that forecasts are unbiasedly coherent when the unbiasedness constraint, or $\boldsymbol{S}\boldsymbol{G}\boldsymbol{S} = \boldsymbol{S}$, holds and also implies that $\boldsymbol{S}\boldsymbol{G}$ is a projection matrix. Thus, the optimal reconciliation weights matrix is formulated as:

$$\boldsymbol{G} = (\boldsymbol{S}^{\top}\boldsymbol{W}_{t+h|t}^{-1}\boldsymbol{S})^{\top}\boldsymbol{S}^{\top}\boldsymbol{W}_{t+h|t}^{-1}. \tag{4.5}$$

Eq (4.5) shows that $\boldsymbol{G}$ is valid under a set of forecasting models $\boldsymbol{\Theta}$ and depends on the expected value of the h-step ahead base forecast error covariance matrix, which contains the uncertainties from the corresponding forecasting models. In a limited sample, $\hat{\boldsymbol{W}}_{t+h|t}$, is constructed from the estimated parameters of the forecasting models $\boldsymbol{\Theta}$ and the one-step ahead base forecast error covariance matrix, $\hat{\boldsymbol{W}}_{t+1|t}$. Being estimated, $\hat{\boldsymbol{W}}_{t+h|t}$ is uncertain due to modelling uncertainty, and this influences the uncertainty of $\boldsymbol{G}$. In the observational level where the sample size is limited, we denote the estimated reconciliation weights matrix as $\hat{\boldsymbol{G}}$, thus $\boldsymbol{S}\hat{\boldsymbol{G}}\boldsymbol{S} = \boldsymbol{S}$ is subject to uncertainty and the coherency constraint depends on how we utilise the available information, given a limited sample. To avoid confusion we clarify the notion here: $\boldsymbol{G}$ refers to the weights matrix of the conventional coherency from the literature. Here, we use $\hat{\boldsymbol{G}}$ to highlight that $\hat{\boldsymbol{G}}$ is estimated. In our stochastic coherency framework $\boldsymbol{G}$ and $\hat{\boldsymbol{G}}$ are coincide. We also introduce $\boldsymbol{\Gamma}$ that is the weights matrix in population.

The expectation of the reconciled forecasts conditional to $\boldsymbol{\mathcal{I}}_t$ is:

$$\begin{aligned}
\mathrm{E}(\tilde{\boldsymbol{y}}_{t+h|t}|\boldsymbol{\mathcal{I}}_t) &= \mathrm{E}(\boldsymbol{S}\hat{\boldsymbol{G}}\hat{\boldsymbol{y}}_{t+h|t}|\boldsymbol{\mathcal{I}}_t) \\
&= \boldsymbol{S}\boldsymbol{\Gamma}\boldsymbol{\mu}_{t+h|t} \\
&= \boldsymbol{S}\boldsymbol{\Gamma}\boldsymbol{S}\boldsymbol{\mu}_{b,t+h|t},
\end{aligned}$$

where $\mathrm{E}(\hat{\boldsymbol{y}}_{t+h|t}|\boldsymbol{\mathcal{I}}_t) = \boldsymbol{\mu}_{t+h|t} = \boldsymbol{S}\boldsymbol{\mu}_{b,t+h|t}$, and $\boldsymbol{\Gamma} = \mathrm{E}(\hat{\boldsymbol{G}}|\boldsymbol{\mathcal{I}}_t)$. Coherency needs the unbiasedness property to ensure that the forecasts are coherent via distributing information across the hierarchy through linear combination. The multiplication between $\boldsymbol{S}$ and $\boldsymbol{\Gamma}$ results in a projection which maintains coherency with regard to any overlooked errors from forecasting models, such as the estimation errors or any statistical discrepancies.

As $\boldsymbol{S}\boldsymbol{\Gamma}$ and $\boldsymbol{S}\hat{\boldsymbol{G}}$ are both projection matrices, this property should be maintained. For example, we maintain the projection matrix to be idempotent. This should hold in both the population and the estimation level, $\boldsymbol{S}\boldsymbol{\Gamma}\boldsymbol{S}\boldsymbol{\Gamma} = \boldsymbol{S}\boldsymbol{\Gamma}$ and $\boldsymbol{S}\hat{\boldsymbol{G}}\boldsymbol{S}\hat{\boldsymbol{G}} = \boldsymbol{S}\hat{\boldsymbol{G}}$. The linear projection, which maintains unbiasedness and coherency, basically ensures the projected forecasts lie on the coherent subspace (Panagiotelis et al., 2021). The issue now is how uncertain the estimated projection matrix is. In the case of $\boldsymbol{S}\boldsymbol{\Gamma}$, it projects to $\boldsymbol{\mu}_{t+h|t}$, whereas $\boldsymbol{S}\hat{\boldsymbol{G}}$ may project the forecasts a bit further from $\boldsymbol{\mu}_{t+h|t}$. Thus, the uncertainty in the projection highlights the importance of modelling uncertainty, since the former originates from the latter. Therefore, we redefine coherency by treating the coherent projection matrix, $\boldsymbol{S}\boldsymbol{\Gamma}\boldsymbol{S} = \boldsymbol{S}$ to mitigate overlooked errors from the forecasting models in forecast reconciliation. We call it **stochastic coherency**.

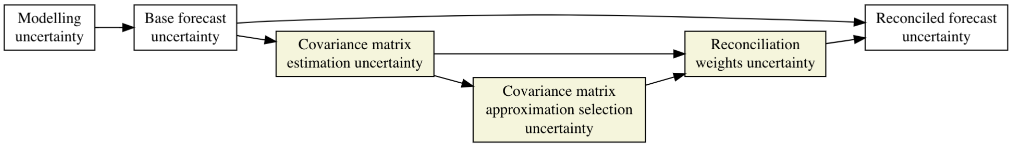

Modelling uncertainty
Base forecast uncertainty
Covariance matrix estimation uncertainty
Covariance matrix approximation selection uncertainty
Reconciliation weights uncertainty
Reconciled forecast uncertainty

Figure 4.1: Forecast reconciliation uncertainty framework. Boxes in beige depict our novel understanding in forecast reconciliation uncertainty framework.

Figure 4.1 summarises the view of the uncertainties in forecast reconciliation we gain from stochastic coherency. Modelling uncertainty leads to the uncertainty in the base forecasts, as in conventional forecasting (Chatfield, 1995). This contributes to the uncertainty of reconciled forecast, which is well understood in the hierarchical forecasting literature (for example, Athanasopoulos et al., 2017). With stochastic coherency we demonstrated that there are additional sources of uncertainty, that can help explain the observations in the literature (Panagiotelis et al., 2021). There is uncertainty in the covariance matrix approximation, which is naturally connected to the uncertainty of the base forecasts. This additional uncertainty is both due the estimation and selection of an appropriate covariance matrix approximation method. Both contribute to the uncertainty of the reconciliation weights, which adds to total uncertainty of the reconciled forecasts.

As modelling uncertainty plays an important role in forecast reconciliation, i.e. how we exploit $\boldsymbol{\mathcal{I}}_t$, we discuss the effect of model specification on the reconciliation. We illustrate the effects by focusing on the case when the forecasts are unbiased. Then, we move to two special cases, namely reconciling biased forecasts and reconciling forecasts from perfectly specified forecasting models. We demonstrate how modelling uncertainty, as in the structure of the models and the parameter estimation, affects forecast reconciliation.

4.3 Reconciling Unbiased Forecasts with Stochastic Coherency

In this scenario, we consider well-specified forecasting models. The forecasting models are able to capture the structure of the data generating process well, but suffer from parameter estimation uncertainty. Given the limited sample size in $\boldsymbol{Y}_t$, we produce h-step ahead base forecasts, $\hat{\boldsymbol{y}}_{t+h|t}$, and we expect that $\mathrm{E}(\hat{\boldsymbol{y}}_{t+h|t}|\boldsymbol{\mathcal{I}}_t) = \boldsymbol{\mu}_{t+h|t}$. Thus, the uncertainty due to parameter estimation is $\hat{\boldsymbol{y}}_{t+h|t} - \mathrm{E}(\hat{\boldsymbol{y}}_{t+h|t}) = \boldsymbol{v}_{t+h|t}$, where $\mathrm{E}(\boldsymbol{v}_{t+h|t}|\boldsymbol{\mathcal{I}}_t) = \boldsymbol{0}$. Note that the irreducible forecast error at period $t+h$ is defined as $\boldsymbol{\zeta}_{t+h|t} = \boldsymbol{y}_{t+h} - \boldsymbol{\mu}_{t+h|t}$. This differs from $\boldsymbol{\varepsilon}_{t+h}$ as the latter is unconditional.

As the base forecasts are $\hat{\boldsymbol{y}}_{t+h|t} = \boldsymbol{\mu}_{t+h|t} + \boldsymbol{v}_{t+h|t}$, the base forecast errors become:

$$\begin{aligned}\hat{\boldsymbol{e}}_{t+h|t} &= \boldsymbol{y}_{t+h} - \hat{\boldsymbol{y}}_{t+h|t} \\ &= \boldsymbol{\mu}_{t+h|t} + \boldsymbol{\zeta}_{t+h|t} - \boldsymbol{\mu}_{t+h|t} - \boldsymbol{v}_{t+h|t} \\ &= \boldsymbol{\zeta}_{t+h|t} - \boldsymbol{v}_{t+h|t}. \qquad (4.6)\end{aligned}$$

From Eq (4.6), we can see that the base forecast errors consist of the irreducible error and the error due to parameter estimation. The latter is affected by the sample sizes.

We aim to reconcile the base forecasts with regard to the hierarchical structure, using $\tilde{\boldsymbol{y}}_{t+h|t} = \boldsymbol{S}\hat{\boldsymbol{G}}\hat{\boldsymbol{y}}_{t+h|t}$. To estimate $\hat{\boldsymbol{G}}$, we need to estimate the h-step ahead base forecast error covariance matrix,

$$\hat{\boldsymbol{W}}_{t+h|t} = \boldsymbol{Z}_{t+h|t} + \boldsymbol{V}_{t+h|t} + \boldsymbol{C}_{t+h|t},$$

where $\boldsymbol{Z}_{t+h|t}$ is the covariance matrix of $\boldsymbol{\zeta}_{t+h|t}$ and $\boldsymbol{V}_{t+h|t}$ is the covariance matrix of $\boldsymbol{v}_{t+h|t}$, where $\mathrm{E}(\boldsymbol{Z}_{t+h|t}) = \boldsymbol{\Sigma}$ and $\mathrm{E}(\boldsymbol{V}_{t+h|t}) = \boldsymbol{V}$, and $\boldsymbol{C}_{t+h|t}$ is the covariance matrix between $\boldsymbol{\zeta}_{t+h|t}$ and $\boldsymbol{v}_{t+h|t}$. Therefore,

$$\hat{\boldsymbol{G}} = (\boldsymbol{S}^\top(\hat{\boldsymbol{W}}_{t+h|t})^{-1}\boldsymbol{S})^\top\boldsymbol{S}^\top(\hat{\boldsymbol{W}}_{t+h|t})^{-1}. \tag{4.7}$$

Looking at Eq (4.7), $\hat{\boldsymbol{G}}$ is uncertain, because the variances and the covariances of $\hat{\boldsymbol{W}}_{t+h|t}$ depend on the parameter estimation uncertainty, given a limited sample size. We can see that the uncertainty in forecasting models is transferred to $\hat{\boldsymbol{G}}$, which will affect the reconciled forecast errors. The reconciliation weights matrix may not be able to improve the base forecast accuracy due to this uncertainty.

From Eq (4.7), we can produce the reconciled forecasts, $\tilde{\boldsymbol{y}}_{t+h|t} = \boldsymbol{S}\hat{\boldsymbol{G}}\hat{\boldsymbol{y}}_{t+h|t}$, and decompose the reconciled forecast error:

$$\begin{aligned}
\tilde{\boldsymbol{e}}_{t+h|t} &= \boldsymbol{y}_{t+h} - \tilde{\boldsymbol{y}}_{t+h|t} \\
&= \boldsymbol{y}_{t+h} - \boldsymbol{S\Gamma}\boldsymbol{\mu}_{t+h|t} + \boldsymbol{S\Gamma}\boldsymbol{\mu}_{t+h|t} - \boldsymbol{S}\hat{\boldsymbol{G}}\hat{\boldsymbol{y}}_{t+h|t} \\
&= \boldsymbol{y}_{t+h} - \boldsymbol{S\Gamma}\boldsymbol{\mu}_{t+h|t} + \boldsymbol{S\Gamma}\boldsymbol{\mu}_{t+h|t} - \boldsymbol{S}\hat{\boldsymbol{G}}(\boldsymbol{\mu}_{t+h|t} + \boldsymbol{v}_{t+h|t}) \\
&= \underbrace{\boldsymbol{\zeta}_{t+h|t}}_{\text{irreducible error}} + \underbrace{(\boldsymbol{S\Gamma} - \boldsymbol{S}\hat{\boldsymbol{G}})\boldsymbol{\mu}_{t+h|t}}_{\substack{\text{reconciliation matrix} \\ \text{estimation error}}} + \underbrace{(-\boldsymbol{S}\hat{\boldsymbol{G}}\boldsymbol{v}_{t+h|t})}_{\substack{\text{reconciled} \\ \text{estimation error}}}
\end{aligned} \tag{4.8}$$

where $\boldsymbol{\zeta}_{t+h|t} = \boldsymbol{y}_{t+h} - \boldsymbol{\mu}_{t+h|t}$, and $\boldsymbol{S\Gamma}\boldsymbol{\mu}_{t+h|t} = \boldsymbol{\mu}_{t+h|t}$ as $\boldsymbol{S\Gamma S} = \boldsymbol{S}$ and $\boldsymbol{\mu}_{t+h|t} = \boldsymbol{S}\boldsymbol{\mu}_{b,t+h|t}$. Eq (4.8) shows that the reconciled forecast error consists of the irreducible error, the reconciliation matrix estimation error, and the reconciled estimation error, which will affect the uncertainty of reconciled forecast error variance.

Looking at the relations between different forecast errors in forecast reconciliation, Panagiotelis et al. (2022) and Panagiotelis et al. (2021) use generalised Pythagoras theorem to establish their relationships. We argue that it needs a relaxation to accommodate the uncertainty by using triangular inequality, where the relationship is shown as,

$$\begin{aligned} ||\boldsymbol{y}_{t+h} - \hat{\boldsymbol{y}}_{t+h|t}||^2 &\leq ||\boldsymbol{y}_{t+h} - \tilde{\boldsymbol{y}}_{t+h|t}||^2 + ||\hat{\boldsymbol{y}}_{t+h|t} - \tilde{\boldsymbol{y}}_{t+h|t}||^2, \qquad (4.9) \\ \text{SSE}_{base} &\leq \text{SSE}_{recon} + \text{SSE}_{\boldsymbol{\epsilon}}, \end{aligned}$$

where $\text{SSE}_{\boldsymbol{\epsilon}}$ is the sum squared reconciliation error and $\text{SSE}_{\boldsymbol{\epsilon}} = ||\hat{\boldsymbol{y}}_{t+h|t} - \tilde{\boldsymbol{y}}_{t+h|t}||^2 \geq 0$. If the left hand side of Eq (4.9) is equal to the right hand side, then $||\boldsymbol{y}_{t+h} - \hat{\boldsymbol{y}}_{t+h|t}||^2 \geq ||\boldsymbol{y}_{t+h} - \tilde{\boldsymbol{y}}_{t+h|t}||^2$. However, Eq (4.9) demonstrates that SSE_{recon} may exceed SSE_{base} as a result of the overall uncertainty in forecast reconciliation. Given the case of unbiased forecasts, we discuss two special cases when the forecasting models are mis-specified and perfectly specified.

4.3.1 Special Case I: Mis-specified Forecasting Models

Due to unknown data generating processes, it is possible to obtain mis-specified models, denoted by †, i.e. adding a redundant variable, wrong transformation, or omitted variables. In the case of mis-specified forecasting models, we produce biased h-step ahead base forecasts, where $\hat{\boldsymbol{y}}^{\dagger}_{t+h|t} = \hat{\boldsymbol{y}}_{t+h|t} + \boldsymbol{o}^{\dagger}_{t+h|t}$ and $\text{E}(\boldsymbol{o}^{\dagger}_{t+h|t}|\boldsymbol{\mathcal{I}}_t) = \boldsymbol{o}^{\dagger}$, which may be nonzero. The base forecast error is shown as,

$$\begin{aligned} \hat{\boldsymbol{e}}^{\dagger}_{t+h|t} &= \boldsymbol{y}_{t+h} - \hat{\boldsymbol{y}}^{\dagger}_{t+h|t} \\ &= \boldsymbol{\mu}_{t+h|t} + \boldsymbol{\zeta}_{t+h|t} - \boldsymbol{\mu}_{t+h|t} - \boldsymbol{v}_{t+h|t} - \boldsymbol{o}^{\dagger}_{t+h|t} \\ &= \boldsymbol{\zeta}_{t+h|t} - \boldsymbol{v}_{t+h|t} - \boldsymbol{o}^{\dagger}_{t+h|t}. \qquad (4.10) \end{aligned}$$

Consequently, the h-step ahead biased base forecast error covariance matrix can be constructed as:

$$\hat{\boldsymbol{W}}^{\dagger}_{t+h|t} = \boldsymbol{Z}_{t+h|t} + \boldsymbol{V}_{t+h|t} + \boldsymbol{O}^{\dagger}_{t+h|t} + \boldsymbol{C}^{\dagger}_{t+h|t},$$

where $\boldsymbol{O}^{\dagger}_{t+h|t}$ is the estimated covariance matrix of $\boldsymbol{o}^{\dagger}_{t+h|t}$ and $\mathrm{E}(\boldsymbol{O}^{\dagger}_{t+h|t}) = \boldsymbol{O}^{\dagger}$. In this case, $\boldsymbol{C}^{\dagger}_{t+h|t}$ collects all covariances between $\boldsymbol{\zeta}_{t+h|t}$, $\boldsymbol{v}_{t+h|t}$, and $\boldsymbol{o}^{\dagger}_{t+h|t}$. Hence, we can calculate $\hat{\boldsymbol{G}}$ in the case of mis-specified models, or $\hat{\boldsymbol{G}}^{\dagger}$, such as,

$$\hat{\boldsymbol{G}}^{\dagger} = (\boldsymbol{S}^{\top}(\hat{\boldsymbol{W}}^{\dagger}_{t+h|t})^{-1}\boldsymbol{S})^{\top}\boldsymbol{S}^{\top}(\hat{\boldsymbol{W}}^{\dagger}_{t+h|t})^{-1}. \tag{4.11}$$

Note that Eq (4.5) is obtained from the assumption of unbiased base forecasts. However, we aim to show that we are still able to reconcile the forecasts, even if the forecasts are biased, but it will come at a cost of more variability due to an additional element in the modelling uncertainty.

Using Eq (4.11), we construct the reconciled forecasts as $\tilde{\boldsymbol{y}}^{\dagger}_{t+h|t} = \boldsymbol{S}\hat{\boldsymbol{G}}^{\dagger}\hat{\boldsymbol{y}}^{\dagger}_{t+h|t}$ and the reconciled forecast error is shown as,

$$\begin{aligned}
\tilde{\boldsymbol{e}}^{\dagger}_{t+h|t} &= \boldsymbol{y}_{t+h} - \tilde{\boldsymbol{y}}^{\dagger}_{t+h|t}\\
&= \boldsymbol{y}_{t+h} - \boldsymbol{S\Gamma}(\boldsymbol{\mu}_{t+h|t} + \boldsymbol{o}^{\dagger}_{t+h|t}) + \boldsymbol{S\Gamma}(\boldsymbol{\mu}_{t+h|t} + \boldsymbol{o}^{\dagger}_{t+h|t}) - \boldsymbol{S}\hat{\boldsymbol{G}}^{\dagger}\hat{\boldsymbol{y}}^{\dagger}_{t+h|t}\\
&= \boldsymbol{y}_{t+h} - \boldsymbol{S\Gamma\mu}_{t+h|t} + \boldsymbol{S\Gamma\mu}_{t+h|t} - \boldsymbol{S}\hat{\boldsymbol{G}}^{\dagger}(\boldsymbol{\mu}_{t+h|t} + \boldsymbol{v}_{t+h|t} + \boldsymbol{o}^{\dagger}_{t+h|t})\\
&= \underbrace{\boldsymbol{\zeta}_{t+h|t}}_{\substack{\text{irreducible}\\ \text{error}}} + \underbrace{(\boldsymbol{S\Gamma} - \boldsymbol{S}\hat{\boldsymbol{G}})\boldsymbol{\mu}_{t+h|t}}_{\substack{\text{reconciliation matrix}\\ \text{estimation error}}} + \underbrace{(-\boldsymbol{S}\hat{\boldsymbol{G}}^{\dagger}\boldsymbol{v}_{t+h|t})}_{\substack{\text{reconciled}\\ \text{estimation error}}} + \underbrace{(-\boldsymbol{S}\hat{\boldsymbol{G}}^{\dagger}\boldsymbol{o}^{\dagger}_{t+h|t})}_{\substack{\text{reconciled}\\ \text{bias error}}},
\end{aligned} \tag{4.12}$$

where $\boldsymbol{\zeta}_{t+h|t} = \boldsymbol{y}_{t+h} - \boldsymbol{\mu}_{t+h|t}$, and $\boldsymbol{S\Gamma o}^{\dagger}_{t+h|t}$ cancels out and similar to Eq (4.8) $\boldsymbol{S\Gamma\mu}_{t+h|t} = \boldsymbol{\mu}_{t+h|t}$ as $\boldsymbol{S\Gamma S} = \boldsymbol{S}$ and $\boldsymbol{\mu}_{t+h|t} = \boldsymbol{S\mu}_{b,t+h|t}$. Hence, Eq (4.12) shows that the reconciled forecast error from biased unreconciled forecast consists of the irreducible error, the reconciliation matrix estimation error, the reconciled estimation error, and the reconciled bias error. This additional error affects the uncertainty of the sum squared reconciled forecast error.

4.3.2 Special Case II: Perfectly-Specified Models

Suppose we were able to produce forecasts from perfectly-specified forecasting models, where the parameters and the data generating process are known. The h-step ahead base forecasts will match with the structure of the hierarchical time series in expectations and in the observational level, shown as $\hat{\boldsymbol{y}}_{t+h|t} = \boldsymbol{\mu}_{t+h|t}$. Hence, the h-step ahead base forecast error is the irreducible error, shown as $\boldsymbol{y}_{t+h} - \hat{\boldsymbol{y}}_{t+h|t} = \boldsymbol{y}_{t+h} - \boldsymbol{\mu}_{t+h|t} = \boldsymbol{\zeta}_{t+h|t}$.

Suppose we aim to reconcile the base forecasts, the reconciled forecasts are shown as,

$$\tilde{\boldsymbol{y}}_{t+h|t} = \boldsymbol{S}\hat{\boldsymbol{G}}\boldsymbol{\mu}_{t+h|t} = \boldsymbol{S}\hat{\boldsymbol{G}}\boldsymbol{S}\boldsymbol{\mu}_{b,t+h|t} = \boldsymbol{\mu}_{t+h|t} \tag{4.13}$$

where $\boldsymbol{\mu}_{t+h|t} = \boldsymbol{S}\boldsymbol{\mu}_{b,t+h|t}$ and $\boldsymbol{S}\boldsymbol{\Gamma}\boldsymbol{S} = \boldsymbol{S}$. In this case, if the structure and the parameters are known, $\boldsymbol{S}\hat{\boldsymbol{G}} = \boldsymbol{S}\boldsymbol{\Gamma}$, and $\hat{\boldsymbol{G}}$ becomes irrelevant because the forecasts are coherent already. Following Eq (4.9), since the forecast errors between both forecasts are the same, then $||\hat{\boldsymbol{y}}_{t+h|t} - \tilde{\boldsymbol{y}}_{t+h|t}||^2 = 0$. Consequently, $||\boldsymbol{y}_{t+h} - \hat{\boldsymbol{y}}_{t+h|t}||^2 = ||\boldsymbol{y}_{t+h} - \tilde{\boldsymbol{y}}_{t+h|t}||^2$, as $\tilde{\boldsymbol{y}}_{t+h|t} = \hat{\boldsymbol{y}}_{t+h|t} = \boldsymbol{\mu}_{t+h|t}$. In perfetly specified models, MinT Reconciliation does not improve or worsen the forecast accuracy as the models are able to produce coherent structures of the time series, $\boldsymbol{\mu}_{t+h|t}$. This is in agreement with Athanasopoulos et al. (2017).

4.3.3 Uncertainty in $\boldsymbol{G}$

The previous discussion shows that the accuracy improvement due to forecast reconciliation depends on the quality of the estimated projection, $\boldsymbol{S}\hat{\boldsymbol{G}}$. Since $\hat{\boldsymbol{G}}$ is a function of $\hat{\boldsymbol{W}}_{t+h|t}$ and $\hat{\boldsymbol{W}}_{t+h|t}$ depends on the model specification, $\hat{\boldsymbol{G}}$ is stochastic. With regard to Eq (4.4), we can say that the reconciled forecasts are the result of a linear combination of all base forecasts, in which the weights are stochastic.

In order to deal with uncertain weights in $\hat{\boldsymbol{G}}$, we draw on the arguments from linear forecast combination literature by Smith and Wallis (2009) and Claeskens et al. (2016). As $\hat{\boldsymbol{G}}$ contains the estimated weights, Smith and Wallis (2009) and Claeskens et al. (2016) note that estimated combination increases the variance of the combined forecasts. Furthermore, in forecast pooling, for any forecast added in the combination to be beneficial there are conditions on the forecast variance (Kourentzes et al., 2019). In order to manage the uncertainty in forecast reconciliation, it could be possible that not all parts of $\boldsymbol{S}$ are equally informative, i.e. these may increase the uncertainty of $\hat{\boldsymbol{G}}$. This may explain the marginal improvements observed with cross-temporal hierarchies, but more importantly it suggests that $\hat{\boldsymbol{G}}$ could be restricted further.

A restricted $\hat{\boldsymbol{G}}$ can be achieved by controlling the information which enters the forecast reconciliation via $\hat{\boldsymbol{W}}_{t+h|t}$. Suppose that $\hat{\boldsymbol{W}}_{t+h|t}$ is assumed to be a fixed covariance matrix, e.g. an identity matrix, then the weights in $\hat{\boldsymbol{G}}$ are fixed and constructed from $\boldsymbol{S}$ only. Alternatively, we can include the sample variances and the covariances of the base forecast errors, but the level of the randomness on the weights in $\hat{\boldsymbol{G}}$ are subject to the uncertainty from the forecasting models. A balance between these is to use the sample variances and manage the off-diagonal elements, for example by shrinking the covariances or restricting them to zero. This may enable us to balance the trade-off between more information and reducing uncertainty of the weights in $\hat{\boldsymbol{G}}$. We note here that stochastic coherency is a general concept which can be applied to any covariance matrix in forecast reconciliation. Fixed weights in $\hat{\boldsymbol{G}}$ due to the identity covariance matrix, or OLS reconciliation, is seen as a means to limit the uncertainty of $\hat{\boldsymbol{G}}$ to zero. This way, we can restrict the uncertainty propagation from the forecasting models to the reconciled forecast uncertainty.

Stochastic coherency acknowledges two potential sources of uncertainty in forecast reconciliation, originating from the modelling or the collection of data. We demonstrate that the main source of uncertainty originates from the forecasting models. The uncertainty in the forecasting model propagates to the estimation of the reconciliation weights matrix via the covariance matrix of the h-step ahead base forecast error and this affects the uncertainty of the reconciled forecasts.

The difference between stochastic and deterministic coherency is not in the reconciled point forecasts, but rather in the variability of the error distribution. Our stochastic interpretation demonstrates that there is an increased uncertainty in the error distribution. In the following section we use simulated and real data to show that our theoretical discussion of stochastic coherency is observable in practice using the widely used MinT Reconciliation and can help explain results in the literature.

4.4 Simulation Study

4.4.1 Experimental Design

In this section we perform two simulations: first, with a small hierarchy, controlling for the model uncertainty, so as to validate the theoretical discussion above; second, with a large hierarchy, as to see the effect of the hierarchy size.

We specify the data generating process of each bottom-level time series as an AR(1) process for the small hierarchy:

$$b_{q,t} = 0.4 y_{q,t-1} + \varepsilon_{q,t},$$

where q is an index from 1 to 4, denoting the bottom level time series in the hierarchy. The innovation term $\varepsilon_{q,t} = \{\varepsilon_{1,t}, \varepsilon_{2,t}, \varepsilon_{3,t}, \varepsilon_{4,t}\}$ and $\boldsymbol{\varepsilon}_{b,t} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\varepsilon_b})$, where $\boldsymbol{\varepsilon}_{b,t} =$

$$\begin{bmatrix} \varepsilon_{1,t} & \varepsilon_{2,t} & \varepsilon_{3,t} & \varepsilon_{4,t} \end{bmatrix}^\top,$$

$$\boldsymbol{\Sigma}_{\varepsilon_b} = \begin{bmatrix} 3 & 2 & 1 & 1 \\ 2 & 3 & 1 & 1 \\ 1 & 1 & 3 & 2 \\ 1 & 1 & 2 & 3 \end{bmatrix}, \text{ and } \quad \boldsymbol{S} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ \multicolumn{4}{c}{\boldsymbol{I}_4} \end{bmatrix}.$$

The top and middle level series result from the aggregation of the bottom-level, as presented by $\boldsymbol{S}$, where $\boldsymbol{y}_t = \boldsymbol{S}\boldsymbol{b}_t$ and $\boldsymbol{b}_t = \{b_{q,t}\}$. For example, $y_{Top,t} = 0.4(b_{1,t-1} + b_{2,t-1} + b_{3,t-1} + b_{4,t-1}) + \varepsilon_{1,t} + \varepsilon_{2,t} + \varepsilon_{3,t} + \varepsilon_{4,t}$. We simulate this setting with sample sizes of 24, 120, and 240 and a burn-in period of 200, to eliminate any initialisation issues.

For the large hierarchy we use 50 bottom-level series, two levels in the middle-level and a top-level time series. All bottom-level time series are generated from ARIMA with $\boldsymbol{\varepsilon}_{b,50} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\varepsilon_{b,50}})$ and $\boldsymbol{\Sigma}_{\varepsilon_{b,50}}$ is generated randomly at each iteration of the simulation. In both simulations, we assume that $\boldsymbol{\delta}_t = \boldsymbol{0}$. For ARIMA we allow randomness in the data generating process, i.e., the AR and MA orders are sampled from 0 to 3 and the integration is from 0 to 1. We simulate the same sample sizes as for the small hierarchy with the same burn-in setting.

Forecasting Models

For the small hierarchy, we generate individual base forecasts using different model specification settings, summarised in Table 4.1. The first option, referred to as *DGP*, assumes that we know the process fully. The second option assumes the model structure is known, but the model is subject to parameter uncertainty. We call this *AR(1)*. The third option employs ARIMA with automatic model selection, named as *AutoARIMA*, which has potentially reduced model uncertainty and parameter uncertainty, as the data generating process is encompassed. The fourth option uses exponential smoothing and represents a mis-specified model by using *ETS(AAN)* that is equivalent to ARIMA(0,2,2), introducing superfluous terms. For the large hierarchy, we use ARIMA and exponential smoothing with automatic selection. These match the latter two options in Table 1. We produce 1- to 6-step ahead base forecasts for both hierarchies. For each combination of sample sizes, forecast horizon, and model specification scenario, we repeat the simulations 1000 times. ARIMA and ETS models are implemented using the auto.arima() in the forecast package (Hyndman and Khandakar, 2008) and the es() in the smooth package (Svetunkov, 2022b) for R (R Core Team, 2022), and we rely on Akaike Information Criterion for selecting the appropriate model form.

Models	DGP	AR(1)	AutoARIMA	ETS
Specification	Known	Known	Approximated	Wrong
Parameter	Known	Estimated	Estimated	Estimated

Table 4.1: Model specification for each scenario in the experimental design

Forecast Reconciliation

We reconcile the base forecasts using the MinT Reconciliation methodology. We use several approximation methods for $\hat{\boldsymbol{W}}_{t+h|t}$ from the literature, summarised in Table 4.2. Hyndman et al. (2011) use a diagonal covariance matrix with equivariant variances, they call this method *OLS*. Athanasopoulos et al. (2017) propose Structural Scaling (*SCL*), where they set equal variances to the bottom-level, and then calculate the covariance matrix as $\boldsymbol{S}\boldsymbol{\sigma}_b$. In this case, $\boldsymbol{\sigma}_b = c\boldsymbol{I}_m$, where c is a scalar and m is the number of the bottom-level time series. Hyndman et al. (2016) propose *WLS*, which uses a diagonal covariance matrix allowing for heterogeneity. Wickramasuriya et al. (2019) propose MinT-Sample, a fully unrestricted estimated covariance matrix of one-step ahead in-sample base forecast errors. This method is denoted here as *EMP*. However, as it is difficult to estimate the off-diagonals, they implement shrinkage on the off-diagonals towards zero by Schäfer and Strimmer (2005), called MinT-Shrink method. This is denoted *SHR* in Table 4.2.

Estimation	OLS	SS	WLS	SHR	EMP
Approximation	$c\boldsymbol{I}$	$\boldsymbol{S\sigma}_b$	$\hat{\boldsymbol{W}}_{d,t+1\mid t}$	$\hat{\boldsymbol{W}}^{\text{SHR}}_{t+1\mid t}$	$\hat{\boldsymbol{W}}_{t+1\mid t}$
$\hat{\boldsymbol{G}}$	$\hat{\boldsymbol{G}}_{\text{OLS}}$	$\hat{\boldsymbol{G}}_{\text{SCL}}$	$\hat{\boldsymbol{G}}_{\text{WLS}}$	$\hat{\boldsymbol{G}}_{\text{SHR}}$	$\hat{\boldsymbol{G}}_{\text{EMP}}$

Table 4.2: Different approximations of $\boldsymbol{W}_{t+h\mid t}$

Apart from the established covariance matrix approximations, we explore three alternative covariance matrices, motivated by our theoretical discussion. Our motivation is to either construct them from the bottom-level or by ignoring some of the off-diagonals in order to mitigate the uncertainty, instead of estimating the whole covariance matrix. A similar study was done by Nystrup et al. (2020) who exploited autocorrelations between time series in temporal hierarchies. Furthermore, we provide a covariance matrix approximation continuum. Figure 4.2 illustrates the covariance matrix approximations. It consists of four bottom-level, two middle-level, and a top-level time series. S on the upper-level of pShrink is constructed, then shrunk. All covariance matrix approximations are positive definite, except bSHR and EMP. bSHR is a positive semi-definite covariance matrix and the positive definiteness of EMP here depends on the sample (Hyndman et al., 2011). In the first method we collect a vector of the bottom-level variances from one-step ahead in-sample forecast errors, $\hat{\boldsymbol{\sigma}}_b$, and construct the covariance matrix with the variances of $\boldsymbol{S}\hat{\boldsymbol{\sigma}}_b$. We call it *cWLS*. Second, we estimate the bottom level covariance matrix and construct it according to the hierarchy. We force a block diagonal structure, making other elements zero and shrinking the remaining, named *bShrink*. Third, we estimate MinT-Shrink and retain the bottom-level covariance matrix. Then we aggregate it to the hierarchy and force covariances between the bottom-level and the upper-levels to be zero. We call this *pShrink*. We sacrifice information utilisation on bShrink and pShrink by forcing hierarchically block diagonals to mitigate the variability of the forecast error.

Figure 4.3 depicts the covariance matrix approximation continuum, where square points denote the alternative covariance matrix approximations. It represents the utilisation of information with regards to the forecast error variability. On the left side of the continuum, OLS provides the least information as $\hat{\boldsymbol{G}}$ is constructed from the hierarchy only. However, it produces the least variable forecast error. Beside OLS, there are SCL and WLS. They provide more information than OLS by allowing heteroscedasticity. Consequently, they produce more variable forecast error than OLS.

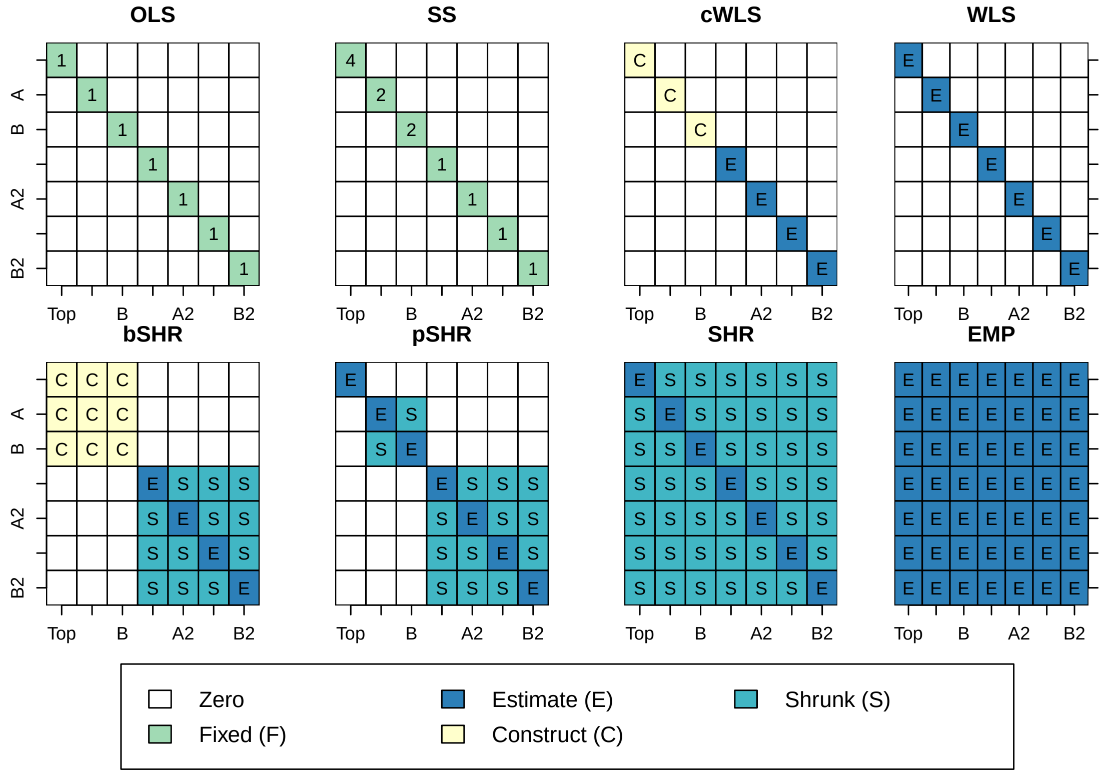


Figure 4.2: Illustration of covariance matrix approximations for a hierarchy of seven time series.

On the right hand side of the continuum, EMP provides full information as we estimate the unrestricted covariance matrix. Consequently, EMP will produce the most variable forecast error. Next to EMP, SHR provides full information with some restrictions, thus produces less variable forecast error than EMP.

Our alternative covariance matrices fill the gap between WLS and SHR. We sac-

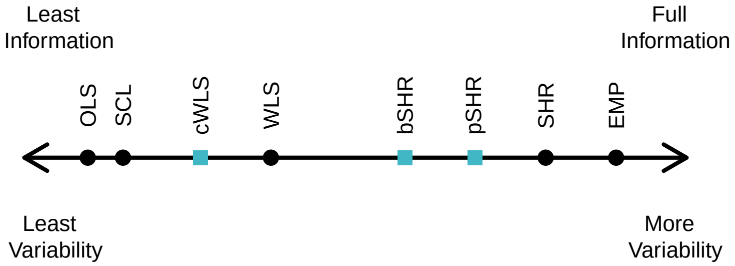


Figure 4.3: Covariance matrix approximation continuum againts information used and forecast error variability for the whole hierarchy.

rifice some correlations to manage the variability. We retain the correlations between the parent nodes and the children nodes, but we dismiss the correlations between the parent and the children from different parent nodes, and vice versa. By constructing the covariance matrix from the bottom-level information, it is expected to produce less variable forecast errors, but have a similar performance with SHR in terms of forecast accuracy.

Error Metrics

We consider two different measures in hierarchical forecasting: (a) a measure which aligns to the objective function (Wickramasuriya et al., 2019; Panagiotelis et al., 2022, 2021); (b) a measure which is more relevant to decision makers (Kourentzes et al., 2019; Athanasopoulos and Kourentzes, 2022). The former deals with measuring the average accuracy of base and reconciled forecasts across the complete hierarchy. The latter measures performance of individual time series and then summarises them across the complete hierarchy. A relevant discussion about the evaluation of hierarchical forecasts is given by Athanasopoulos and Kourentzes (2022).

We focus on the mean squared error (MSE) for each time series i, as

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

where J is the simulation run. Then, we measure the performances across the hierarchy from the loss function perspective, using Relative Total Squared Error:

$$\text{RelTotSE}_h = \frac{\sum_{i=1}^{N} MSE_{ih,recon}}{\sum_{i=1}^{N} MSE_{ih,base}},$$

where N is the number of time series in the hierarchy. Essentially, RelTotSE_h measures the relative accuracy between SSE_{recon} and SSE_{base}. From the decision-focused perspective, we use Average Relative MSE, inspired by Davydenko and Fildes (2013):

$$\text{AvgRelMSE}_h = \left(\prod_{i=1}^{N}\frac{MSE_{ih,recon}}{MSE_{ih,base}}\right)^{\frac{1}{N}}.$$

4.4.2 Findings: Small Hierarchy

Figure 4.4 and 4.5 present the distributions of RelTotSE and AvgRelMSE for different forecasting models and covariance matrix approximations for the sample size of 24. Each pair of subplots corresponds to a modelling case from Table 4.1, where the first subplot provides $t+1$ error distributions, while the second provides the average across $t+1$ to $t+6$. The red lines indicate the geometric mean. Furthermore, in all plots, grey dots denote the outliers and any part of each distribution that is not plotted is denoted by a red arrow at the top and the bottom. Note that the covariance matrix approximations are ordered by its completeness of the information, i.e. from an identity matrix to utilising variances and covariances fully to estimate the reconciliation weights matrix.

From Figure 4.4, for RelTotSE, we can see that when we have perfectly-specified

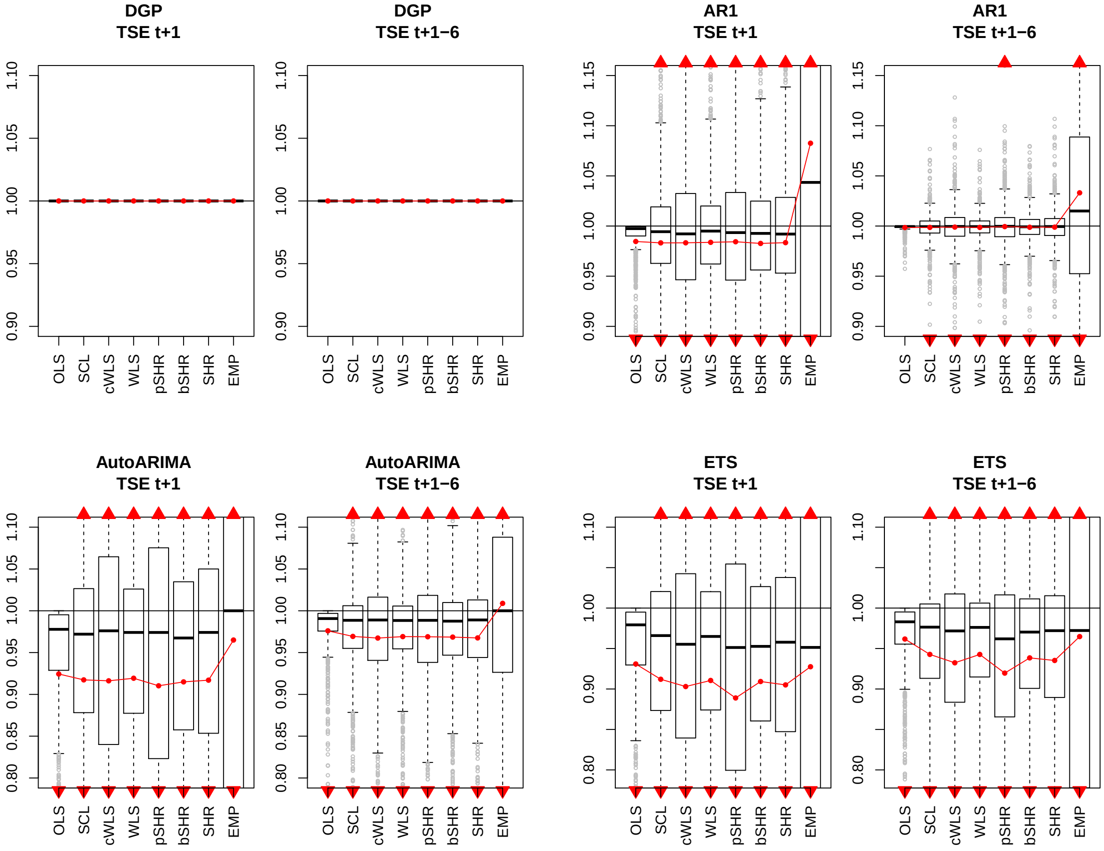


Figure 4.4: Distributions of RelTotSE (TSE) for the small hierarchy and the sample size of 24.

forecasting models, there is no gain from reconciliation. That is because the models are able to produce coherent forecasts. As the base forecasts are coherent already, the reconciled forecasts are the same as the base forecasts.

However, once we introduce modelling uncertainty, we gain some benefit from forecast reconciliation. Imposing parameter uncertainty only, i.e. employing estimated AR(1), induces relatively small gains from reconciliation. This shows that as the modelling uncertainty increases reconciliation provides gains but again at an increased variability, meaning that the mean of the relative errors decreases, but the variance of error measure distributions increases. The relative accuracy gain is more noticeable when we use AutoARIMA compared to AR(1), at the cost of higher variability of RelTotSE. Using ETS we benefit the most from forecast reconciliation, but also at the cost of the highest error variability among other modelling options. These gains are less pronounced when the multi-step base forecast errors are introduced, even though the variances are non-zero.

Looking at the covariance matrix approximations, the results verify Theorem 3.1 by Panagiotelis et al. (2021) that OLS reconciliation improves or matches the accuracy of base forecasts regardless the model specification. However, when we approximate the covariance matrix, it is possible to get less accurate reconciled forecasts on some observations, as uncertainties are introduced. We can see that some distributions go well beyond the accuracy of base forecasts. As expected the simpler the approximation of the covariance matrix is, the less the variability is, and vice versa. Our proposed covariance matrix approximations, e.g. cWLS, pSHR, and bSHR, are able to reduce the variability of RelTotSE yet provide relative accuracy, on average, similar to WLS and SHR.

In Figure 4.5 the model specification does not affect the accuracy improvement much, but affect the variability of the relative measure, for AvgRelMSE. The descriptions of any symbol are the same as the ones in Figure 4.4. We can see from the figure that the variability increases as the forecasting models become increasingly mis-specified.

Here, the effect of the covariance matrix approximations differ from RelTotSE. For AvgRelMSE there is no clear increase in error variability as more complete covariance matrix approximations are used. However, the simplest covariance matrix approximation results in very variable performance. This can be explained by considering that another role of the covariance matrix in forecast reconciliation is to scale the reconciled forecast errors. At more aggregate levels of the hierarchy the scale of errors increases. Conversely SHR is able to scale the forecast errors better than OLS. The same is true for the other approximations. This argument aligns to the discussion on temporal hierarchies where SCL performs well (Athanasopoulos et al., 2017).

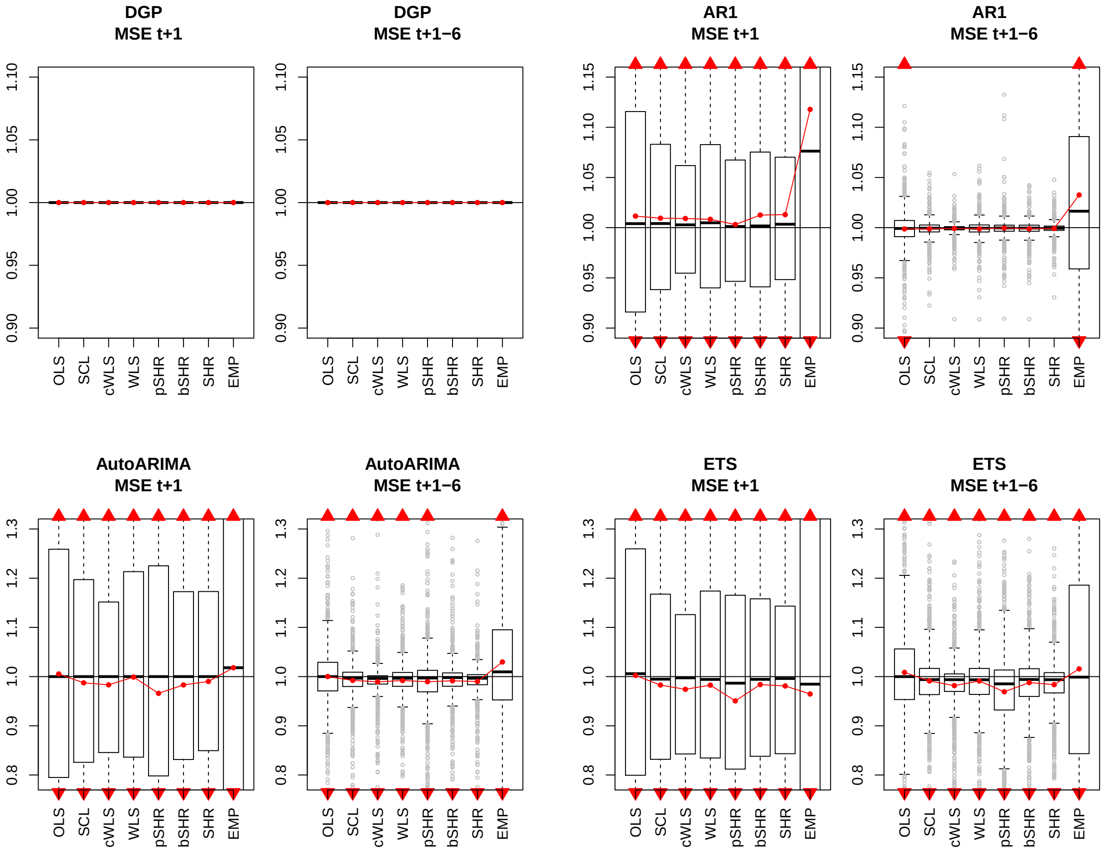


Figure 4.5: Distributions of AvgRelMSE (MSE) for the small hierarchy and the sample size of 24.

Figure 4.6 presents the effect of sample sizes using mis-specified models with RelTotSE. For RelTotSE, the benefits of forecast reconciliation reduce as the sample sizes increase, together with the decrease of the variability. As the estimation of the parameters improves, the uncertainty reduces, and therefore this result is expected. We can see this effect by looking at the lower error bars. Nevertheless, we still observe some variability in longer sample sizes.

Regardless of what error metrics is used, we observe variability in the performance of forecast reconciliation. For example, we observe a trade-off between accuracy and variability of RelTotSE, i.e. the more complete covariance matrix is, the more accurate the reconciled forecasts are. This, however, comes at a cost, which is introducing more error variability.

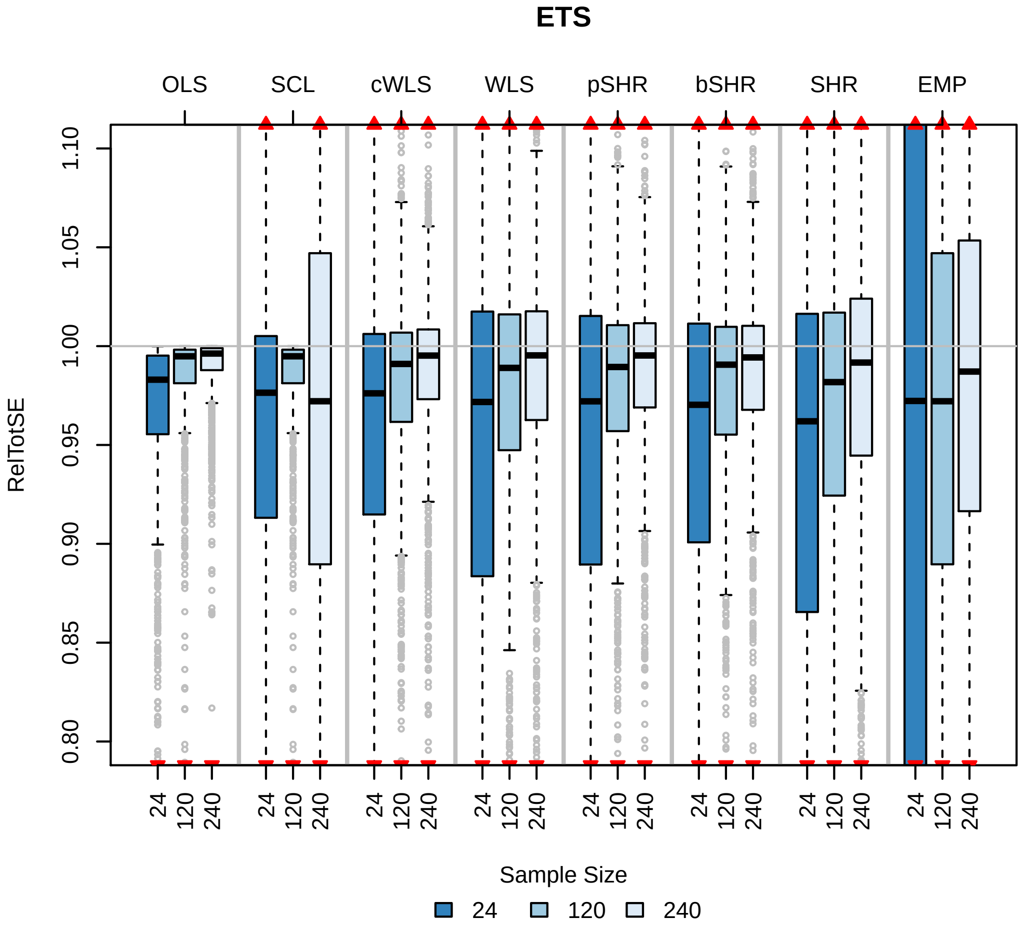


Figure 4.6: ETS, h=6, different sample sizes, RelTotSE.

4.4.3 Findings: Large Hierarchy

Next we discuss the findings from the large hierarchy. Table 4.3 presents a comparison between the small and the large hierarchy, with RelTotSE and AvgRelMSE for one-step ahead forecast and different sample sizes. We present the geometric mean and the logarithm of geometric standard deviation of the relative error distribution from ETS only. A negative (positive) number denotes an improvement (deterioration) on the error measure. The bold highlights the most accurate reconciliation approach, for the geometric mean, and the least volatile reconciliation approach, for the standard deviation. All numbers in the geometric mean are in percentages.

Considering the geometric mean of RelTotSE and AvgRelMSE, SHR outperforms

Statistics	Geometric Mean (%)				Geometric St. Dev. (log)			
Hierarchy	Small		Large		Small		Large	
Sample	24	240	24	240	24	240	24	240
				RelTotSE				
OLS	-3.8	-1.2	-7.4	-3.4	**6.1**	**2.7**	**10.1**	**4.0**
SCL	-5.7	-1.4	-18.6	-13.5	12.7	4.9	22.9	15.7
CWLS	-5.7	-1.4	-17.4	-13.0	12.8	4.9	21.4	15.3
WLS	-6.8	-1.4	-18.4	-15.3	17.8	7.0	25.4	19.8
pSHR	-6.5	-1.2	-18.4	-14.6	16.5	5.4	25.4	18.2
bSHR	-6.2	-1.4	**-21.8**	-16.8	14.7	5.4	29.8	23.3
SHR	**-8.0**	**-2.6**	-18.3	**-18.4**	20.9	11.8	25.5	29.4
EMP	-3.5	-2.3	293.2	-11.1	44.7	17.5	119.2	51.0
				AvgRelMSE				
OLS	0.9	-0.6	12.6	10.8	13.1	6.1	33.5	29.9
SCL	-0.9	-0.7	2.3	2.4	7.6	3.1	25.8	23.6
cWLS	-0.9	-0.7	-0.4	-0.4	7.8	3.1	19.4	17.9
WLS	-1.8	-0.7	-1.0	-1.5	**6.8**	**2.3**	**18.6**	16.9
pSHR	-1.6	-0.4	-1.0	-1.2	6.9	3.5	**18.6**	17.1
bSHR	-1.2	-0.7	**-1.5**	-1.4	7.6	3.0	19.4	17.8
SHR	**-3.1**	**-1.9**	-0.9	**-2.6**	11.5	8.0	19.3	**15.9**
EMP	1.6	-1.7	1730.7	3.3	34.8	13.1	156.1	26.7

Table 4.3: A comparison between the small and the large hierarchy with RelTotSE and AvgRelMSE for one-step ahead forecast.

the other alternatives, apart from the case of small sample size for the large hierarchy, where bSHR is the best. We find that pSHR ia also competitive. As expected EMP is very sensitive to estimation uncertainty. The increased size of the hierarchy substantially reduces its performance, while increasing the sample size helps.

In terms of the standard deviation of RelTotSE, we observe similar findings to the small hierarchy, where OLS is the least variable, while EMP and SHR are the most volatile. Overall, as the completeness of the covariance matrix increases, so does the variance of the errors. The larger size of the hierarchy increases it further, requiring more terms to be estimated, while sample size helps. On the other hand, for AvgRelMSE, the least variable methods are between WLS, pSHR, and SHR, which have more complete information than OLS does.

The differences between RelTotSE and AvgRelMSE can be largely explained by the changing scale across the levels of the hierarchy. Improvements in the top-level dominates RelTotSE, which is scale dependent. On the other hand, the scale independent AvgRelMSE balances the gains across all levels, and therefore differences are less pronounced. It is important to consider both views. The RelTotSE matches the operation performed by MinT Reconciliation and directly demonstrates the effects of uncertainty highlighted by the stochastic coherency. Furthermore, although on different scales, both RelTotSE and AvgRelMSE indicate that hierarchical forecasting is beneficial in terms of accuracy. As evidence, as the complexity of the covariance matrix approximation increases, so does the variance of the errors, in agreement with our theoretical discussion.

4.5 Forecasting A&E hospital admissions

We apply our understanding of stochastic coherency to Accident and Emergency (A&E) admission data in a hospital in the United Kingdom. Hospitals in the UK, as is the case globally, face increased pressure due to the global pandemic, requiring many resources. This has often caused disruptions in their normal operations, such as scheduled surgeries, but also in the operations of their A&E departments. To this end, it is important to have reliable forecasts of demand, across the different groups of interest, so that the hospital can allocate resources best. In normal conditions, A&E forecasting is important in the United Kingdom due to the worrying mismatch between the hospital service quality and financial efficiency (Limb, 2014). Forecasts can be useful for multiple decisions, such as staff scheduling, procurement of drugs and other medical supplies, bed utilisation, etc.

The time series consists of 64 bottom-level time series, which are structured according to age (under 3 years old, between 4-16 years old, between 17-74 years old, and more than 75 years old), gender (male; female), and disposal type (admitted, discharged, referred to clinics, transferred, died, referred to health care professionals, left, and others). Figure 4.8 provides a plot of representative time series from the A&E

hospital admissions dataset, where we observe that the time series exhibit seasonal patterns, local trend, and outliers. There are multiple ways to aggregate from the bottom-level time series to the total number of admissions, for example aggregating by gender, age, or type first, and then across one of the remaining two characteristics, and so on. This results in a grouped hierarchy of 135 time series with eight distinct groups/levels. A map of the hierarchy is presented in Figure 4.7. Labels on the left indicate the nature of time series at each level, while on the right provide the number of time series at that level. The lines indicate how the time series are aggregated between the different levels. Note that some time series in the bottom-level are sparse and these pose challenges in the modelling.

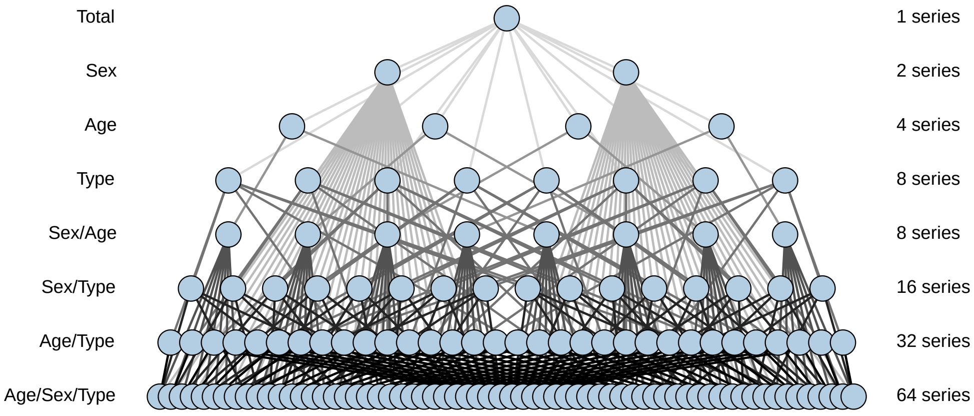
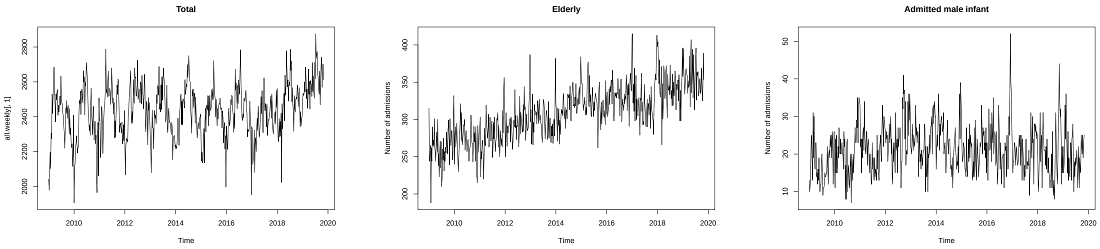


Figure 4.7: Map of the A&E admission hierarchy.

We have been provided weekly data from January 2009 to October 2019. We produce from 1- to 4-step ahead base forecasts with two sets of in-sample data. The longer set uses 536 weeks, while a much shorter set has only 100 weeks. The second introduces additional modelling uncertainty as the number of time series is larger than the number of observations making the approximation of the covariance matrix challenging. This helps us validate our findings from stochastic coherency on real complex hierarchical time series. For both cases we use the same test set of 29 weeks, allowing for 25 rolling origin forecasts.



(a) Total (b) Elderly (c) Admitted male infant

Figure 4.8: Representative time series of A&E hospital admissions dataset for total, elderly, and admitted male infant group.

Afilal et al. (2016) point that the A&E admission data can be structured in a hierarchy according to the patients' characteristics, which may be correlated. Athanasopoulos et al. (2017) use ARIMA to model UK A&E admission data, but at a country level. Forecasting models can also incorporate exogenous variables, such as special events, holidays, and temperatures, to improve forecast accuracy using regression models, ARIMA, or ETS (Kam et al., 2010; Xu et al., 2016; Rostami-Tabar and Ziel, 2022).

We use ARIMA and ETS with automatic model selection, as setup for the large simulation above. We acknowledge that these can be prone to model misspecification problems. First, we omit important information for A&E forecasting, such as special events. Second, we do not treat differently the any sparse time series at the bottom level of the hierarchy. This potential misspecification is of interest, to explore how the reconciliation approaches impact the forecasts. Thirdly, the automatic ARIMA function on forecast package does not capture seasonality while some time series are seasonal. We reconcile base forecasts using all covariance matrix approximations, as in Figure 4.2 and evaluate the forecasts using RelTotSE and AvgRelMSE.

Table 4.4 presents a comparison between ARIMA and ETS models for the short and long samples (100 and 536), for all covariance matrix approximations and over 1-4 step ahead forecasts. A negative (positive) number denotes an improvement (deterioration) of the error measure, on average. Numbers in bold highlight the best

Statisics	Geometric Mean (%)				Geometric St. Dev. (log)			
Model	ETS		ARIMA		ETS		ARIMA	
Sample	Short	Long	Short	Long	Short	Long	Short	Long
				RelTotSE				
OLS	-0.90	-0.70	-1.40	-3.10	**0.50**	**0.60**	**0.70**	**2.30**
SCL	-0.90	-2.70	-2.20	-7.10	3.70	3.50	7.20	5.90
cWLS	-0.70	-2.60	**-3.10**	-7.40	3.10	2.90	6.90	5.00
WLS	-0.90	-3.10	-2.90	-7.90	4.30	3.70	8.00	6.20
pSHR	**-1.40**	-3.30	-2.90	-5.60	4.90	3.30	8.40	4.50
bSHR	-1.10	-3.10	-1.30	-6.20	5.90	4.80	9.20	6.30
SHR	0.80	-6.60	0.30	**-11.30**	5.80	8.00	9.90	9.20
EMP	55.20	**-6.90**	64.50	-2.60	43.50	41.40	39.90	32.60
				AvgRelMSE				
OLS	0.30	1.40	1.20	1.30	3.40	3.50	4.40	7.70
SCL	-2.20	-1.60	-1.50	-1.50	2.70	**3.00**	**3.70**	4.70
cWLS	-2.00	-1.50	-2.30	-2.90	**2.60**	3.40	**3.70**	**4.80**
WLS	-2.70	-2.00	-2.40	-3.10	3.40	3.30	4.20	4.60
pSHR	**-3.50**	-2.30	**-2.70**	-1.60	3.80	3.80	4.50	6.30
bSHR	-3.20	-2.30	-1.70	-1.60	4.00	3.60	4.60	5.70
SHR	-1.30	**-5.00**	-1.20	**-7.00**	4.00	5.80	5.50	7.50
EMP	65.40	2.20	51.40	-0.20	35.30	31.70	27.80	27.10

Table 4.4: A comparison between AutoARIMA and ETS models.

performing results. Similar to Table 4.3, we present summary statistics of the error distribution, with the geometric mean and the logarithm of the geometric standard deviation. The results are ordered in terms of completeness of the covariance matrix approximation.

We note that across all results, the more complete covariance matrices, such as the pSHR, and SHR, offer good forecast accuracy. First, we focus on the cases of the short in-sample set. We find that pSHR performs overall best. For RelTotSE and ARIMA the cWLS is best but closely followed by pSHR. We note that the more complete approximations (SHR and EMP) perform poorly in terms of RelTotSE, while for AvgRelMSE the SHR improves upon the base forecasts, but still performs worse than all simpler approximations.

The results for the large sample are contrasting. The SHR performs best. In the case of RelTotSE and ETS we observe that EMP outperforms all alternatives, although closely followed by SHR. The long sample size allows for reliable estimation. We note that the less complete covariance approximations, although perform worse, all improve upon the base forecasts.

Looking at the standard deviation of the forecast errors, OLS provides the most stable relative accuracy for RelTotSE, and SCL together with cWLS for the AvgRelMSE. Overall, simpler covariance approximations exhibit a low standard deviation of the forecast errors. More complete ones, such as the SHR and EMP, exhibit increased deviations, even for the long in-sample set. The proposed covariance matrix approximation, namely cWLS, pSHR, and bSHR, they enable us to compromise between the accuracy gain and the variability of the error distribution.

The results in Table 4.4 are relative to the base forecasts and do not permit a direct comparison between ETS and ARIMA, as this is not the aim of the evaluation. If we compare the two, we find that the reconciled forecasts from ARIMA outperform the ones from ETS for small sample sizes, and vice versa.

Therefore, we argue that with complex data generating processes, observed in real data, we again find variability in the performance of forecast reconciliation, especially when we need to estimate the covariance matrix approximation, instead of relying on fixed values. This emphasises the importance of stochastic coherency to be considered in the application of hierarchical forecasting. For the particular case of A&E hospital admissions, we find that the pSHR covariance approximation that was developed with the understanding we gained from stochastic coherency resulted overall in good forecast accuracy, and stability. The performances of cWLS and bSHR were similar. As a group these performed well against approximations from the literature that were either too restrictive or they did not consider the additional uncertainty arising from stochastic coherency.

4.6 Conclusions

Stochastic coherency shifts our paradigm from deterministic to stochastic forecast reconciliation. We have to deal with the uncertainties in estimating the reconciliation weights matrix, originating modelling uncertainty due to limited sample size. This directly affects the performance of the forecast error, either on average or its variability, due to the approximation of the covariance matrix.

Our findings show that there are two sources of uncertainty in forecast reconciliation originating from modelling, namely the base forecast uncertainty and the reconciliation weight uncertainty. It becomes obvious that the base forecast uncertainty is carried forward to the reconciled forecast uncertainty. Model and parameter uncertainty contaminate the covariance matrix approximation and introduce the second source of error. Naturally, the sample size affects modelling uncertainty. Moreover, a larger hierarchy produces more uncertain reconciled forecasts, as there are more terms to estimate. These become evident with stochastic coherency and the results from both simulated and real data corroborate with this understanding.

Due to these uncertainties, we cannot say that forecast reconciliation improves the accuracy consistently all the time. Our findings show that the reconciled forecast accuracy can be worse than the base forecast accuracy in some cases, even if on average it ranks better.

In relation to different model specifications, there are some conditions when the degree of specification affects the efficacy of forecast reconciliation. As stated previously, if the forecasting models capture the bottom level data adequately, stochastic coherency indicates that the bottom-up approach is sufficient, and reconciliation will not add value. For instance, our simulation demonstrates that by having perfect information, the models can estimate coherent mean and errors, hence forecast reconciliation does not change anything. When the modelling uncertainty is limited, we obtain limited gains from forecast reconciliation. However, when we have misspecified models, forecast reconciliation becomes useful, which matches typical cases in reality.

The benefit of forecast reconciliation appears when there are modelling uncertainties. The MinT Reconciliation reduces the forecast error by redistributing the modelling uncertainty, which contains the uncertainty of parameter estimation as well as the unobservable statistical discrepancy, across the hierarchy. As long as the data generating process is unknown and the forecasts are produced from individual forecasting models, the MinT Reconciliation can help. Here we did not explore the effect of statistical discrepancy and it should be explored further in future work.

We also observe a significant accuracy improvement when forecast error covariances are incorporated into the estimated reconciliation weights matrix. Even though it improves the forecast accuracy generally, it comes at the cost of increased variability of the error measure. One of the solutions to deal with the variability is to obtain a good quality reconciliation weights matrix, which reduces the effect of modelling uncertainty. We can obtain this by managing variances and covariances on the estimated h-step ahead covariance matrix and this determines the quality of the combination weights estimation in the estimated reconciliation weights matrix. A weaker argument for this is given by Kourentzes and Athanasopoulos (2019).

We can estimate a useful reconciliation weights matrix from approximating the base forecast error covariance matrix. Simple and fixed approximations of the covariance matrix, namely OLS and SCL, are immune to modelling uncertainty and the fixed estimation of the reconciliation weights matrix is able to limit the variability of the error measure. On the other extreme, the estimation of EMP and SHR relies heavily on the base forecast errors, and is prone to modelling uncertainty. Consequently, the reconciliation weights matrix becomes uncertain. SHR relying on shrinkage remains widely useful, while EMP is useful only for a very large estimation sample size.

Managing the off-diagonals in the covariance matrix construction enables to balance the accuracy gain and the variability of the forecast error. We argue that using bSHR and pSHR are potential solutions, because they introduce restrictions, yet maintain structurally important information. Our findings also show that bSHR and pSHR results in a similar accuracy gain, but less variable to SHR, while being competitive to the simpler WLS and cWLS. Naturally, this is important in applications of hierarchical forecasting, where both aspects of accuracy and reliability over time are important. We find strong evidence of this when we model accident and emergency admissions for the UK hospital of our case study, where the covariance matrices developed with our understanding of stochastic coherency performed very competitively, offering a good balance between accurate and stable forecasts. We argue that these can aid decision making. Naturally, less variable forecasts are beneficial widely for operations. For example, in a production setting less erratic forecasts result in more resilient plans and lower costs (Sagaert et al., 2018). Similar examples can be drawn from inventory management, where maximum accuracy forecasts do not necessarily result in the best inventory performance (Kourentzes et al., 2020).

Our discussion extends to probabilistic hierarchical forecasting. The literature does not take into account modelling uncertainty (Jeon et al., 2019; Taieb et al., 2020). The density of the reconciled forecasts is also affected by the reconciliation weights matrix and so is their performance. Future research on this area will help highlighting the exact influence of modelling uncertainties on probabilistic hierarchical forecasting.

In conclusion, we introduce stochastic coherency to overcome a limitation in the definition of classical coherency in forecast reconciliation and hierarchical forecasting. Using the concept of stochastic coherency, we give more attention to the error term from the data generating process. We are able to demonstrate that stochastic coherency is relevant to forecast reconciliation via simulations and a case study of A&E admissions in a hospital. It allows us to explain observations from the literature, where well performing approximations for the covariance matrix introduce variability in the error distribution, and provides a framework to consider the setup of hierarchical forecasting in applications.

Chapter 5

Discussion and Conclusion

This thesis is built on understanding the sources of uncertainty in business forecasting, as classified by Chatfield (2000), regardless of the choice of modelling approaches (ETS, VES, or Forecast Reconciliation). In particular, we argue that not only model structure but parameter estimation uncertainty has significant effects on forecasting performance. Although this is not a new finding, we provide specific details on the effect and remedies. These unmitigated uncertainties affect the variability of the point forecasts and the forecasting performance. We address these problems using parameter shrinkage. Collating from the previous chapters, we discuss three important ideas: (a) understanding uncertainty in business forecasting, (b) estimation and shrinkage, and (c) reliability in business forecasting. At the end, we discuss the practical implications, limitations of this thesis and further research avenues.

5.1 Uncertainty in Business Forecasting

Chatfield (2000) classifies three fundamental sources of uncertainty in any statistical model: (a) model structure, (b) parameter estimation, and (c) sampling. We try to understand this classification from the bias-variance decomposition. We rewrite Eq.

2.9 in this chapter to show the decomposition, shown as

$$\mathrm{E}\left(y_t - \hat{y}_{t|t-1}|\mathcal{I}_t\right)^2 = \underbrace{\sigma^2}_{\substack{\text{irreducible}\\ \text{variance}}} + \underbrace{\mathrm{E}\left(\boldsymbol{\mu}_t - \mathrm{E}\left(\hat{y}_{t|t-1}\right)|\mathcal{I}_t\right)^2}_{\text{model bias}} + \underbrace{\mathrm{E}\left(\mathrm{E}\left(\hat{y}_{t|t-1}\right) - \hat{y}_{t|t-1}|\mathcal{I}_t\right)^2}_{\text{model variance}}. \quad (5.1)$$

Eq. 5.1 demonstrates that the total variance has three components, namely the irreducible variance, the model bias, and the model variance. The irreducible error cannot be mitigated further; meanwhile, the other two components can be reduced by having an appropriate statistical model. In this thesis, we assume that we can identify the model structure and isolate the model structure uncertainty. In other words, the total variance is conditional to the 'known' or well-identified model structure. We can argue that the parameter estimation and the sampling uncertainty are captured in the model variance, and the literature overlooks those. Table 5.1 combines our understanding of Chatfield's uncertainty classification and the bias-variance decomposition. Chatfield (2000) argued that the model structure uncertainty is the most problematic source. Burnham and Anderson (2002) and Hyndman et al. (2008b) take this idea further by selecting the best-approximating model via minimising an information criterion to identify the best model structure.

Uncertainty	Model Bias	Model Variance
Model structure	✓	
Model parameter		✓
Sampling		✓

Table 5.1: The uncertainty classification and the bias-variance decomposition

We also need to mitigate the uncertainties in the model variance and this thesis provides the evidence. If not, they might impact forecasting performance. For example, Chapter 4 demonstrates the effects of unmitigated parameter uncertainty on the reliability of the forecasting performance of hierarchical methods over forecast origins. Chapter 2 and 3 attempt to mitigate this issue by implementing a shrinkage estimator, either in a univariate or a multivariate model. The results are mixed in the sense that the implementation results in more accurate point and interval forecasts for the univariate model but inaccurate ones for the multivariate model. In addition, exploring the model variance enables us to understand the interaction between parameters in models. For example, in Chapter 2, we demonstrate a covariance between smoothing parameters and initial value estimation, which affects the prediction interval performance. In Chapter 3, we show a compensating effect between the covariance matrix and the persistence matrix.

We also demonstrate an intriguing finding in Chapter 3. Given the same model structure, estimating parameters independently and simultaneously with a numerical optimisation routine results in different estimates. It indicates that the parameter search between univariate and multivariate models is different, which is worth exploring. We rely on the numerical optimisation routine to estimate ETS and VES parameters instead of finding the solution analytically. Thus, we propose to incorporate 'optimisation uncertainty' in the classification.

This thesis provides evidence that we need to revise the uncertainty classification and change our approach to dealing with these uncertainties. We shall mitigate each uncertainty individually and show that parameter shrinkage can be a sensible solution and that it produces accurate forecasts.

5.2 Estimation and Shrinkage

The amount of information we know about the model affects how we estimate the model. Table 5.2 describes four modeling scenarios, where each depends on how much we know about the true model structure, parameters, and sample size. We have similar experimental designs in each chapter to control the source of uncertainty, where we adapted the scenario from Athanasopoulos et al. (2017).

Scenario	1	2	3	4
True model	Known	Known	Unknown	Unknown
True parameter	Known	Unknown	Unknown	Unknown
Sample size	Sufficient	Sufficient	Sufficient	Limited

Table 5.2: Scenarios in which we know the true model and have sufficient sample sizes.

Scenario 1 would not happen in reality, where we know the model perfectly, and the sample size becomes irrelevant anymore. We only have the irreducible uncertainty. Scenario 2 and 3 are suitable for the maximum likelihood estimation, as we have sufficient sample sizes to estimate parameters consistently and efficiently. However, Scenario 3 has more uncertainties than Scenario 2 because we have the model structure uncertainty. Examples of Scenario 2 are apparent in any statistical inference book, especially in estimating parameters given a known probability distribution (Casella and Berger, 2002). Meanwhile, the methodology proposed by Hyndman et al. (2008b) resembles Scenario 3, where they select the model structure via the information criterion and estimate the parameters with the maximum likelihood.

Scenario 4 represents reality, especially in business, where we have limited relevant information and know nothing about the true model and parameters. In this case, the model may suffer from overfitting, where the model tries to fit the data perfectly. For example, smoothing parameters in ETS tend to be higher than the optimal ones. This phenomenon has been observed by Barrow et al. (2021). If we stick with the uncertainty classification, the sampling uncertainty potentially results in parameter uncertainty because we need more information to estimate the parameters efficiently. As a result, these uncertainties become a joint uncertainty that cannot be mitigated independently.

This thesis focuses on the parameter estimation in Scenario 4, where we do not have enough sample sizes to conform to the maximum likelihood properties. In such cases, the estimates of parameters may not be efficient and affect forecasting performance. As discussed in Chapter 1, the linear combination approaches in business forecasting can mitigate the overall uncertainty, improving the forecast accuracy on average. However, if the combination is not done carefully, as seen in Chapter 4, this might amplify the final forecast uncertainty. We need some mitigation to handle the model structure and the parameter uncertainty in the first place. As for the former, selecting the best-approximated model via minimising an information criterion is promising.

We propose implementing parameter shrinkage approaches to overcome the parameter uncertainty. Shrinkage has a wide spectrum, namely the fixed and data-dependent shrinkage. The fixed shrinkage is when we force some parameters to zero a priori. There are many reasons to force them to zero. First, one may have a prior subjective belief that some parameters have to be zero. It is closely related to the modellers' judgement. We can extend this to the Bayesian approach where the parameters might have a prior distribution with a zero mean (Bhattacharya et al., 2015; Kastner and Huber, 2020). Second, model identification may be the main reason for the fixed shrinkage. For example, some parameters in VAR models are set to zero to mitigate issues from simultaneous equations, resulting in structural VAR models (Lütkepohl, 2005). Third, the fixed shrinkage can help avoid significant estimation errors. In a covariance matrix, we often set the covariances to be zero as it is difficult to estimate a full covariance matrix. Oftentimes, it needs a shrinkage estimator (Daniels and Kass, 2001; Schäfer and Strimmer, 2005). We provide evidence that the fixed shrinkage mitigates the parameter uncertainty. In Chapter 3, we see that VES models perform well when the cross-smoothing parameters and the covariances are shrunk to zero arbitrarily. In Chapter 4, forcing some covariances to zero improves the forecast accuracy of the forecast reconciliation on average and results in a less variable forecast error distribution.

Instead of shrinking parameters to zero heuristically, we can control the amount of parameter shrinkage depending on the data at hand. We call this as the data-dependent shrinkage. It is often used in regression models, for example, ridge, lasso, and elastic net regression (Hoerl and Kennard, 2000; Tibshirani and Taylor, 2011; Zhou, 2007; Hastie et al., 2009, 2015). We implement the data-dependent shrinkage on different time series models by modifying the loss function to include a penalty function. In the new loss function, a hyper-parameter controls the amount of the penalisation. We have evidence in Chapter 2 that shrinking smoothing parameters yields forecast accuracy improvement. Linking back to Table 5.2, we can move from Scenario 3 to 4 using the shrinkage estimator. Suppose we have a limited sample size; the estimator will typically result in some amount of shrinkage in the parameters. When the sample size increases, we can still use the shrinkage estimator. We anticipate that the shrinkage would be limited, and the shrinkage hyper-parameter would be close to zero. In other words, the estimation will be very similar to the maximum likelihood or the one proposed by (Hyndman et al., 2008b).

We have sufficient evidence to postulate that it is important to implement parameter shrinkage estimation in any model/ method. It is a flexible implementation, as we can use it when we have a limited or sufficient sample. It can be a potential safety net to mitigate issues from overfitting. In terms of mitigation, we have two options depending on the situation. Suppose we have a large model and a limited sample size; then implementing the fixed shrinkage can be our 'best bet' solution. On the other hand, the data-dependent shrinkage estimator is a sensible solution as we can transition from it to the maximum likelihood, conditional on the sample size.

5.3 Reliability in Business Forecasting

Apart from communicating our forecasts to the decision-makers, we need to make our forecasts trustworthy enough for them (Gönül et al., 2012). Spavound and Kourentzes (2022) argue that reliability in forecasting is an important feature to gaining decision-makers' trust. Reliability in forecasting can mean two things. First, a forecasting model/ method is reliable when it performs well consistently across forecast origins. Second, reliable forecasts suggest that the forecasts are concentrated around its global model. For example, for a local-level model, the reliable forecasts will be close to its mean. We suspect that this reliability may result in cost efficiency due to avoiding erratic scheduling and low service-level (Kadipasaoglu and Sridharan, 1995) and the effectiveness of the forecasting task to support decisions.

This thesis provides two examples of reliability in forecasting. Chapter 2 shows that shrinking smoothing parameters will result in forecasts close to its global model. This implementation leads to more accurate point and interval forecasts. Chapter 4 shows that there is a trade-off between the average performance and the variability in performance, i.e. when we use available information to estimate the covariance matrix, we may get the most accurate forecasts on average at the cost of increased variability. On the other hand, forcing the covariances to zero may lead to moderate forecast accuracy with manageable forecast errors. It has an important practical implication. We could handle thousands of time series in forecast reconciliation, which may consume many resources. Ensuring that the approach is reliable would be a safety net for decision makers that they get accuracy improvements when dedicating a considerable cost of resources.

Reliability is a relatively new concept in business forecasting and is open to discussion. First, reliability can be a sensible candidate for defining 'good' forecasts, as some decision-makers may emphasise reliability over accuracy. Second, there is a possibility to measure reliability, either in the forecasting performance or the forecasts. Third, we need to understand whether this concept will lead to better decisions. For example, in inventory management, reliable forecasting might reduce inventory costs. Overall, reliability in business forecasting may become an interesting research avenue, especially in a more uncertain and pessimistic society about the future (Tutton, 2022).

5.4 Practical Implications

In this section, we discuss the possible implications of this thesis for practice, which are relevant to managers and researchers. This thesis provides evidence that a multivariate model for a multivariate time series can be tricky. The model requires a large sample size to estimate all parameters efficiently. Given the limited resources available to a manager, it would be sensible to implement univariate time series models or the forecast reconciliation with a simple covariance matrix estimation approximation (Athanasopoulos et al., 2017; Pritularga et al., 2021). Utilising all information may not be necessary to obtain accurate forecasts. Instead, using relevant information, i.e., sacrificing cross-time series information, suffices to obtain accurate forecasts in light of potential large estimation errors.

Introducing a hyperparameter to a forecasting model offers greater control over estimating the model parameters. An additional ability to tune the model is a way to remedy the effect of ‘algorithm aversion’ (Dietvorst et al., 2015, 2018). For example, the modeller/ decision maker can decide which parameters are set to zero and this can make the model more trustworthy. They can also adjust the hyper-parameter optimisation according to their utility functions. However, the application has to be done carefully as Sroginis (2021) shows that the forecasting performance of judgmentally tuning the model can be low. Linking to the decision maker’s utility function, Athanasopoulos and Kourentzes (2022) provide a series of practical recommendations to accommodate the decision maker's utility function in the hierarchical forecasting problem.

5.5 Limitations

Due to its underlying assumptions, this thesis has some limitations relating to the model structure, the data-generating process, the loss function in a multivariate setting, and the hyper-parameter estimation.

Our methodology assumes that the model structure is known to observe the effect of parameter uncertainty. Specifically, in our case study, the model structure is selected via minimising an information criterion as suggested by Burnham and Anderson (2002) and Hyndman et al. (2008b), and it is sufficient for our purpose. Note that our case differs from shrinkage in regression, as the latter mitigates the parameter and the model structure uncertainty simultaneously. Ours only mitigates the parameter uncertainty and we do not perform any model selection in our implementation.

Our findings are tied to our design of the DGP. We need evidence of whether our parameter shrinkage performs well in different types of DGP. Our DGP also has a 'moderate' sampling frequency, which is neither high nor intermittent. We speculate that the connection between smoothing parameters and initial values will impact both cases, albeit differently. We also suspect that our implementation may be sensitive to a structural break, where the location of the time series shifts. Lower smoothing parameters mean that the ETS remembers the past more, and the structural break effect carries over to the forecasts. It means that the model may need to be more adaptive to handle it.

Third, we need to rethink our multivariate shrinkage loss function. Currently, we only include the persistence matrix in the penalty function. We may need to include the covariance matrix because we observe a compensating effect between both parameters. More philosophically, a multivariate model implicitly assumes that the whole system has the same loss function. Suppose we have sales data for different stores; employing a multivariate model means that each store has the same loss function. Athanasopoulos and Kourentzes (2022) and Babai et al. (2022) have pointed this problem in hierarchical forecasting.

Lastly, our shrinkage approach in ETS is computationally more expensive than the maximum likelihood estimation because we have an additional hyper-parameter. It might not be appealing for forecasters as business often has thousands of time series to forecast with the potential to investigate more efficient routines. Currently, we rely on derivative-free optimisation algorithms available on R (Johnson, 2022), and we argue for the importance of examining the optimisation problem in forecasting, specifically.

5.6 Future Research

This thesis provides various options for future research in business forecasting, especially in methodology development and applications, such as inventory management and retail forecasting. First, we can extend our shrinkage estimator to conduct model selection by shrinking the initial values. The current model selection in ETS has a binary approach, i.e., whether a state exists or does not. However, we can extend our modelling to those abrupt jumps between models (Svetunkov et al., 2022b). Instead, we can shrink states to zero and effectively reduce the effect of the corresponding states on the actual time series. We can start from a general ETS model with possible states and let the shrinkage estimator select the model. The same idea can be applied in VES. The current VES requires homogeneous time series, e.g., all time series have seasonality. By implementing model selection with shrinkage, we can model VES with heterogeneous time series.

Secondly, we can extend the shrinkage implementation with Bayesian estimators. The subjective belief that some parameters are zero can be translated to a prior distribution of some parameters (Bhattacharya et al., 2015). From Chapter 2, we can see a correlation between the smoothing parameters' value and the initial values' variance. It shows that each parameter's statistical properties affect the model's estimation and performance. A study by Andrawis and Atiya (2009) implements Bayesian estimation, showing promise, and we can generalise it to develop a new estimation and selection methodology in SSOE state-space models.

Third, we need to investigate the loss function in a multivariate model. The current loss function assumes that the whole system has a single loss function. On the other hand, each decision related to each time series in the multivariate system might have different objectives. For example, a procurement manager has a different objective function than a marketing manager. It relates to a question about connecting the decision-maker's loss function to the model's one. Trapero et al. (2019), Kourentzes et al. (2020) and Liu et al. (2022) attempt to combine the optimisation and the forecasting problem in the inventory setting. This issue becomes more prominent in hierarchical forecasting because each node might have a different, possibly conflicting, objective function. For example, decision-makers at the strategic level have different goals from ones at the operational level. Athanasopoulos and Kourentzes (2022) and Babai et al. (2022) raise this issue, and the former proposes a Pareto chart to incorporate the multi-objective decisions in hierarchical forecasting.

Regarding hierarchical forecasting, the literature often treats the hierarchy as a part of the problem setup, and it is known. However, we can argue that the hierarchy is prone to uncertainty over time. For example, a product category hierarchy can be uncertain because a product might be discontinued or a new product might be introduced. There are also some problems defining the granularity of time series in a temporal hierarchy. These counterexamples highlight a possibility of a stochastic hierarchy in terms of structure and not estimates. It might be challenging to say whether a hierarchy represents the truth of DGP. Still, we acknowledge that it is valuable to accumulate more information outside the time series of interest.

Lastly, the concept of reliability is emerging in business forecasting, and there are many aspects that we need to explore. We need to define the concept of reliability clearly. One way would be to explore the statistical properties of the forecast error distribution. As the reliability is quantified, we can measure its effects on the decisions, e.g., the impact of reliability on inventory costs.

Bibliography

Afilal, M., Yalaoui, F., Dugardin, F., Amodeo, L., Laplanche, D., Blua, P., 2016. Forecasting the Emergency Department Patients Flow. Journal of Medical Systems 40 (7).

Andrawis, R. R., Atiya, A. F., 2009. A new bayesian formulation for holt's exponential smoothing. Journal of Forecasting 28, 218–234.

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

Athanasopoulos, G., Ahmed, R. A., Hyndman, R. J., 2009. Hierarchical forecasts for australian domestic tourism. International Journal of Forecasting 25, 146–166.

Athanasopoulos, G., de Silva, A., 2012. Multivariate exponential smoothing for forecasting tourist arrivals. Journal of Travel Research 51, 640–652.

Athanasopoulos, G., Gamakumara, P., Panagiotelis, A., Hyndman, R. J., Affan, M., 2020. Hierarchical Forecasting. In: Baltagi, B., Hong, Y., Koop, G., Krämer, W., Matyas, L. (Eds.), Advanced Studies in Theoretical and Applied Econometrics. Vol. 52. Springer, pp. 689–719.

Athanasopoulos, G., Hyndman, R. J., Kourentzes, N., Petropoulos, F., 2017. Forecasting with temporal hierarchies. European Journal of Operational Research 262 (1), 60–74.

Athanasopoulos, G., Kourentzes, N., 2022. On the evaluation of hierarchical forecasts, in press.

Australian Bureau of Statistics, 2015. Australian System of National Accounts: Concepts, Sources and Methods. Tech. rep., Australian Bureau of Statistics.

Babai, M. Z., Boylan, J. E., Rostami-Tabar, B., 2022. Demand forecasting in supply chains: a review of aggregation and hierarchical approaches. International Journal of Production Research 60, 324–348.

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.

Barrow, D. K., Kourentzes, N., 2016. Distributions of forecasting errors of forecast combinations: Implications for inventory management. International Journal of Production Economics 177, 24–33.

Basu, S., Michailidis, G., 2015. Regularized estimation in sparse high-dimensional time series models. The Annals of Statistics 43, 1535–1567.

Bates, J. M., Granger, C. W. J., 1969. The combination of forecasts. Operational Research Society 20, 451–468.

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

Bhattacharya, A., Pati, D., Pillai, N. S., Dunson, D. B., 10 2015. Dirichlet–laplace priors for optimal shrinkage. Journal of the American Statistical Association 110, 1479–1490.

Boylan, J. E., Syntetos, A., 2021. Intermittent demand forecasting: context, methods and applications. John Wiley & Sons, Inc., Hoboken, New Jersey.

Brown, R. G., 1956. Exponential smoothing for predicting demand. Arthur D. Little. Inc.

Brown, R. G., 1959. Statistical forecasting for inventory control. McGraw-Hill, New York.

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.

Burnham, K. P., Anderson, D. R., 2002. Model Selection and Multi-Model Inference: A Practical Information-Theoretic Approach, 2nd Edition. Springer, New York.

Casella, G., Berger, R. L., 2002. Statistical inference, 2nd Edition. Duxbury/Thomson Learning.

Chatfield, C., 1995. Model Uncertainty, Data Mining and Statistical Inference. Journal of the Royal Statistical Society. Series A (Statistics in Society) 158 (3), 419–466.

Chatfield, C., 2000. Time-Series Forecasting, 1st Edition. Chapman and Hall/CRC.

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

Chow, G. C., Lin, A. L., 1971. Best linear unbiased interpolation, distribution, and extrapolation of time series by related series. The Review of Economics and Statistics 53 (4), 372–375.

Claeskens, G., Magnus, J. R., Vasnev, A. L., Wang, W., 2016. The forecast combination puzzle: A simple theoretical explanation. International Journal of Forecasting 32, 754–762.

Clemen, R. T., 1989. Combining forecasts: A review and annotated bibliography. International Journal of Forecasting 5 (4), 559–583.

Conn, A. R., Scheinberg, K., Vicente, L. N., 2009. Introduction to Derivative-Free Optimization. The Society for Industrial and Applied Mathematical Programming Society.

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

Davydenko, A., Fildes, R., 2013. Measuring forecasting accuracy: The case of judgmental adjustments to SKU-level demand forecasts. International Journal of Forecasting 29, 510–522.

de Silva, A., Hyndman, R. J., Snyder, R., 2010. The vector innovations structural time series framework: a simple approach to multivariate forecasting. Statistical Modelling 10, 353–374.

Demšar, J., 2006. Statistical comparisons of classifiers over multiple data sets. Journal of Machine Learning Research 7, 1–30.

di Fonzo, T., Girolimetto, D., 2021. Cross-temporal forecast reconciliation: Optimal combination method and heuristic alternatives. International Journal of Forecasting 39 (1), 39–57.

Dietvorst, B. J., Simmons, J. P., Massey, C., 2015. People erroneously avoid algorithms after seeing them err. Journal of Experimental Psychology 144, 114–126.

Dietvorst, B. J., Simmons, J. P., Massey, C., 3 2018. Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. Management Science 64, 1155–1170.

Duncan, D. B., Horn, S. D., 1972. Linear Dynamic Recursive Estimation from the Viewpoint of Regresion Analysis. Journal of the American Statistical Association 67 (340), 815–821.

Duncan, G., Gorr, W., Szczypula, J., 1993. Bayesian forecasting for seemingly unrelated time series: Application to local government revenue forecasting. Management Science 39, 275–293.

Farnum, N. R., 1992. Exponential smoothing: Behavior of the ex-post sum of squares near 0 and 1. Journal of Forecasting 11, 47–56.

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

Fildes, R., Kourentzes, N., 2011. Validation and forecasting accuracy in models of climate change. International Journal of Forecasting 27, 968–995.

Fliedner, G., 2001. Hierarchical forecasting: Issues and use guidelines. Industrial Management & Data Systems 101 (1), 5–12.

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., 3 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.

Gönül, M. S., Önkal, D., Goodwin, P., 9 2012. Why should i trust your forecasts? Foresight, 5–9.

Han, L., Neumann, M., 2006. Effect of dimensionality on the Nelder-Mead simplex method. Optimization Methods and Software 21, 1–16.

Hastie, T., Tibshirani, R., Friedman, J. H., 2009. The Elements of Statistical Learning: Data mining, Inference, and Prediction, 2nd Edition. Springer.

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.

Holt, C. C., 1 2004. Forecasting seasonals and trends by exponentially weighted moving averages. International Journal of Forecasting 20, 5–10.

Hyndman, R. J., Ahmed, R. A., Athanasopoulos, G., Shang, H. L., 2011. Optimal combination forecasts for hierarchical time series. Computational Statistics and Data Analysis 55, 2579–2589.

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., Khandakar, Y., 2008. Automatic time series forecasting: the forecast package for R. Journal of Statistical Software 26 (3), 1–22.

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

Hyndman, R. J., Lee, A. J., Wang, E., 2016. Fast computation of reconciled forecasts for hierarchical and grouped time series. Computational Statistics and Data Analysis 97, 16–32.

James, W., Stein, C., 1961. Estimation with quadratic loss. Berkeley Symposium on Mathematical Statistics and Probability, 361–379.

Jeon, J., Panagiotelis, A., Petropoulos, F., 2019. Probabilistic forecast reconciliation with applications to wind power and electric load. European Journal of Operational Research 279, 364–379.

Johnson, S. G., 2022. 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.

Jones, R. H., 1966. Exponential smoothing for multivariate time series. Journal of the Royal Statistical Society. Series B (Methodological) 28, 241–251.

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.

Kam, H. J., Sung, J. O., Park, R. W., 2010. Prediction of daily ED patient numbers for a regional emergency medical center using time series analysis. Healthcare Information Research 16 (3), 158–165.

Kastner, G., Huber, F., 11 2020. Sparse bayesian vector autoregressions in huge dimensions. Journal of Forecasting 39, 1142–1165.

Kleindorfer, G. B., O'neill, L., Ganeshan, R., 1998. Validation in simulation: Various positions in the philosophy of science. Management Science 44, 1087–1099.

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

Koning, A. J., Franses, P. H., Hibon, M. L., Stekler, H. O., 2005. The M3 competition: Statistical tests of the results. International Journal of Forecasting 21, 397–409.

Kourentzes, N., Athanasopoulos, G., 2019. Cross-temporal coherent forecasts for Australian tourism. Annals of Tourism Research 75, 393–409.

Kourentzes, N., Athanasopoulos, G., 2021. Elucidate structure in intermittent demand series. European Journal of Operational Research 288, 141–152.

Kourentzes, N., Barrow, D., Petropoulos, F., 3 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.

Kourentzes, N., Rostami-Tabar, B., Barrow, D. K., 2017. Demand forecasting by temporal aggregation: Using optimal or multiple aggregation levels? Journal of Business Research 78 (October 2016), 1–9.

Kourentzes, N., Saayman, A., Jean-Pierre, P., Provenzano, D., Sahli, M., Seetaram, N., Volo, S., 2021. Visitor arrivals forecasts amid covid-19: A perspective from the africa team. Annals of Tourism Research 88, 103197.

Kourentzes, N., Trapero, J. R., Barrow, D. K., 2020. Optimising forecasting models for inventory planning. International Journal of Production Economics 225, 107597.

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

Limb, M., 2014. Hospitals forecast a rise in emergency admissions, while commissioners forecast a fall. BMJ (Clinical research ed.) 349, 3–4.

Liu, C., Letchford, A. N., Svetunkov, I., 2022. Newsvendor problems: An integrated method for estimation and optimisation. European Journal of Operational Research 300, 590–601.

Lütkepohl, H., 2005. New Introduction to Multiple Time Series Analysis. Springer.

Ma, S., Fildes, R., Huang, T., 2 2016. Demand forecasting with high dimensional data: The case of sku retail sales forecasting with intra- and inter-category promotional information. European Journal of Operational Research 249, 245–257.

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.

Makridakis, S., Winkler, R. L., 1989. Sampling distributions of post-sample forecasting errors. Source: Journal of the Royal Statistical Society. Series C, Applied Statistics 38, 331–342.

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.

Naylor, T. H., Finger, J. M., 1967. Verification of computer simulation models. Management Science 14, 92–101.

Nelder, J. A., Mead, R., 01 1965. A Simplex Method for Function Minimization. The Computer Journal 7 (4), 308–313.

Nicholson, W. B., Matteson, D. S., Bien, J., 2017. VARX-L: Structured regularization for large vector autoregressions with exogenous variables. International Journal of Forecasting 33 (3), 627–651.

Nystrup, P., Lindström, E., Pinson, P., Madsen, H., 2020. Temporal hierarchies with autocorrelation for load forecasting. European Journal of Operational Research 280, 876–888.

Office for National Statistics, 2011. Labour market quality and methodology information. Tech. Rep. October, Office for National Statistics.

Oliveira, J. M., Ramos, P., 2019. Assessing the performance of hierarchical forecasting methods on the retail sector. Entropy 21 (436), 1–22.

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

Ord, J. K., Koehler, A. B., Snyder, R. D., 1997. Estimation and prediction for a class of dynamic nonlinear statistical models. Journal of the American Statistical Association 92, 1621–1629.

Panagiotelis, A., Athanasopoulos, G., Gamakumara, P., Hyndman, R. J., 2021. Forecast reconciliation: A geometric view with new insights on bias correction. International Journal of Forecasting 37 (1), 343–359.

Panagiotelis, A., Gamakumara, P., Athanasopoulos, G., Hyndman, R. J., 2022. Probabilistic forecast reconciliation: Properties, evaluation and score optimisation. European Journal of Operational Research.

Pennings, C. L., van Dalen, J., 2017. Integrated hierarchical forecasting. European Journal of Operational Research 263, 412–418.

Pritularga, K., Svetunkov, I., Kourentzes, N., 2021. Stochastic coherency in forecast reconciliation. International Journal of Production Economics 240 (108221).

Pritularga, K. F., Svetunkov, I., Kourentzes, N., 2022. Shrinkage estimator for exponential smoothing models, in press.

R Core Team, 2022. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.
URL https://www.R-project.org/

Rostami-Tabar, B., Ziel, F., 7 2022. Anticipating special events in emergency department forecasting. International Journal of Forecasting 38, 1197–1213.

Sagaert, Y. R., Aghezzaf, E. H., Kourentzes, N., Desmet, B., 2018. Tactical sales forecasting using a very large set of macroeconomic indicators. European Journal of Operational Research 264, 558–569.

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.

Smith, J., Wallis, K. F., 2009. A simple explanation of the forecast combination puzzle. Oxford Bulletin of Economics and Statistics 71 (3), 331–355.

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

Snyder, R. D., Ord, J. K., Koehler, A. B., McLaren, K. R., Beaumont, A. N., 4 2017. Forecasting compositional time series: A state space approach. International Journal of Forecasting 33, 502–512.

Spavound, S., Kourentzes, N., September 2022. Making forecasts more trustworthy. Foresight, 21–25.

Sroginis, A., 2021. The use of contextual information in demand forecasting. Ph.D. thesis, Lancaster University Management School, Lancaster, United Kingdom.

Svetunkov, I., 2021. smooth: Forecasting Using State Space Models. R package version 3.1.2.41023.

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

Svetunkov, I., 2022b. smooth: Forecasting Using State Space Models. R package version 3.2.0.
URL https://github.com/config-i1/smooth

Svetunkov, I., Chen, H., Boylan, J. E., 2022a. A new taxonomy for vector exponential smoothing and its application to seasonal time series. European Journal of Operational Research 304 (3), 964–980.

Svetunkov, I., Kourentzes, N., Ord, J. K., 2022b. Complex exponential smoothing. Naval Research Logistics 69 (8), 1108–1123.

Svetunkov, I., Pritularga, K., 2022. legion: Forecasting Using Multivariate Models. R package version 0.1.2.41001.
URL https://github.com/config-i1/legion

Taieb, S. B., Koo, B., 2019. Regularized Regression for Hierarchical Forecasting Without Unbiasedness Conditions. In: The 25th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. ACM, pp. 1–11.

Taieb, S. B., Taylor, J. W., Hyndman, R. J., mar 2020. Hierarchical Probabilistic Forecasting of Electricity Demand With Smart Meter Data. Journal of the American Statistical Association, 1–17.

Tashman, L. J., 2000. Out-of-sample tests of forecasting accuracy: an analysis and review. International Journal of Forecasting 16, 437–450.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. Journal of the Operational Research Society, Series B (Methodological) 58, 267–288.

Tibshirani, R. J., Taylor, J., 2011. The solution path of the generalized lasso. Annals of Statistics 39 (3), 1335–1371.

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.

Tsay, R. S., 2014. Multivariate Time Series Analysis: With R and Financial Application. John Wiley & Sons.

Tutton, R., August 2022. The sociology of futurelessness. Sociology.

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.

Wilms, I., Basu, S., Bien, J., Matteson, D. S., 2021. Sparse identification and estimation of large-scale vector autoregressive moving averages. Journal of the American Statistical Association 0 (0), 1–12.

Wilms, I., Croux, C., 3 2018. An algorithm for the multivariate group lasso with covariance estimation. Journal of Applied Statistics 45, 668–681.

Winters, P. R., 1960. Forecasting sales by exponentially weighted moving averages. Management Science 6, 324–342.

Xu, Q., Tsui, K. L., Jiang, W., Guo, H., 2016. A hybrid approach for forecasting patient visits in emergency department. Quality and Reliability Engineering International 32 (8), 2751–2759.

Yang, D., Goh, G. S., Jiang, S., Zhang, A. N., 2016. Forecast UPC-level FMCG demand, Part III: Grouped reconciliation. In: Proceedings - 2016 IEEE International Conference on Big Data, Big Data 2016. pp. 3813–3819.

Zhou, W., 2007. Asymptotic distribution of the largest off-diagonal entry of correlation matrices. Transactions of the American Mathematical Society 359 (11).

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