Dynamic forecast combination and problems in multivariate nonstationary time series

Maddie Rachael Smith, M.Sci (Hons.), M.Res

Submitted for the degree of Doctor of Philosophy at

Lancaster University.

September 2024

STOR-i

excellence with impact

Abstract

It is often the case that decision makers are presented with multiple forecasts for the same variable, produced perhaps from different forecasting models or experts. While attempting to identify a single 'best forecast' does offer a valid approach, favourable performance is often achieved through combining the N available forecasts in some way. As such, forecast combination presents a ubiquitous problem for decision makers in a wide array of fields, and provides the first focus of this thesis.

Although perhaps a seemingly simple problem, the combination of forecasts can prove difficult due to issues such as correlation between forecasts, and changing statistical properties of forecasters throughout time. In order to account for such nonstationary behaviour, it is necessary to combine the forecasts dynamically. In this thesis, we propose a dynamic linear model based procedure for combining point forecasts. This combination procedure works by creating a linearly weighted combination of forecasts, where the weights are allowed to evolve temporally in response to observed data. Following this, we consider the problem when one or more of the N forecasters fails to provide a prediction at time t. We discuss how this can also be interpreted as a sudden change in forecaster quality, and provide adaptive methods for dealing with this.

As the second focus of this thesis, we examine another problem in nonstationary time series, pertaining to the autoregressive process of order 1 (AR(1)). From two bivariate AR(1) processes, we construct a nonstationary oscillating stochastic process, for which we derive key theoretical properties.

Acknowledgements

Firstly, I would like to thank my academic supervisors Nicos Pavlidis and Adam Sykulski, without whom this research would simply not have happened. Your support, subject knowledge and guidance have been invaluable over the past three years, and I have learnt many a lesson that I will carry with me throughout my career. It would be futile to attempt to list all that you have taught me, so I will simply say that you have shown me how to be a better researcher, and for that I am immensely grateful. Nicos, thank you for introducing me to the world of Forecast Combination; I have truly enjoyed finding my 'niche' in this extensive research area. I must also thank you for your keen eye for typos, without which this thesis would undoubtedly feature many a spelling error. Adam, it has been a pleasure to work with you on the final chapter of this thesis. Undertaking such a theoretical project was a huge but welcome change, and I am really proud of the research we have produced. I am especially thankful for the many emails, meetings and hours that you have dedicated to this over the past few hectic weeks.

I would also like to say a huge thank you to Sofia Olhede from EPFL, for her invaluable contributions to the work featured in Chapter 5 of this thesis. Our meetings may have been few, but your impressive subject knowledge, and sense of humour, meant that they were an absolute pleasure.

I am incredibly grateful to have completed my PhD and MRes at the EPSRC funded STOR-i Centre for Doctoral Training, and I would like to thank the many STOR-i students and staff for making it such a pleasant experience. In particular, I would like

to thank my cohort, with whom I started this journey back in 2020. I know that my visits to the office were rare, but you always made me feel at home. I have no doubt that you will all go on to do amazing things, and I cannot wait to see them. I would especially like to thank my friend Lídia, who has supported me from the MRes exams right through to writing this thesis. Your mathematical expertise is astounding, your programming support is invaluable, and your cats are adorable.

By a stroke of luck, I was blessed to grow up with the best family known to man (trust me, I'm a scientist). I would therefore like to thank my parents, Janette and Dave, my sister, Hannah, and my grandparents, Grandma Pat, Grandad Jim, Nanna Bet and Grandad John. I know that none of you knew what a PhD was before I started one, and I know that most of you still do not. Despite this, I am truly grateful for your encouragement in everything that I do, however little you may understand it. I would also like to thank my Uncle Steve, who gave me the book that finally allowed programming to 'click' for me. Unfortunately, we lost Grandad Jim while I was finishing my PhD this year. Although he is no longer here to say it, I know that he, like you all, believed that I was capable of anything I put my mind to. I would not have made it this far without each of you, from Duke Street Nursery to Lancaster University.

Finally and above all, I would like to thank my soon-to-be husband, and partner of eleven years, Jason. Not a day goes by in which I do not appreciate your support, kindness and patience. Thank you for truly caring about my research, for always asking questions and offering advice when I needed it. You have always made me feel seen and heard, something that I do not take for granted as a woman in this field. Thank you for always listening to my problems, and wiping away my tears when things got too much (which happened more often than not in this past year). I am truly grateful that you were able to accompany me to my international conferences, and I must thank you for the endless cups of tea and digestive biscuits brought to my office. You are my other half, and this would not have been possible without you.

This thesis is dedicated to my grandad, James Monks; I love you most of all, Scarecrow.

Declaration

I declare that the work in this thesis has been done by myself and has not been submitted elsewhere for the award of any other degree. Chapters 1, 2, 3, 4 and 6 contain work supervised by Nicos Pavlidis and Adam Sykulski; Chapter 5 contains work supervised by Adam Sykulski with the help of Sofia Olhede. This thesis contains $37,707$ words.

Maddie Rachael Smith

Contents

List of Figures

List of Tables

List of Abbreviations

DLM	Dynamic Linear Model
AR	Autoregressive
MAPE	Mean Absolute Percentage Error
OLS	Ordinary Least Squares
TVP	Time-varying Parameter
MLE	Maximum Likelihood Estimation
BMA	Bayesian Model Averaging
DMA	Dynamic Model Averaging
LDF	Loss Discounting Framework
MSE	Mean Square Error
MAE	Mean Absolute Error
SMAPE	Symmetric Mean Absolute Percentage Error
AMOC	Atlantic Meridional Overturning Circulation
ECB	European Central Bank
GDP	Gross Domestic Product
HICP	Inflation
UNEM	Unemployment
MSFE	Mean Square Forecast Error
MPD	Model Probability Discounting
SGD	Stochastic Gradient Descent

KLIC Kullback-Leibler Information Criterion

SCS Splitting Conic Solver

PF Particle Filter

SA Simple Average

VAR Vector autoregressive

EEG Electroencephalography

Chapter 1

Introduction

Time series analysis is a fundamental area of study, with applications in numerous fields and research areas. When faced with a time series, we may ask: is there the presence of trends or seasonality? What is the dependence between observations? Can future values be predicted? Such questions can be answered through the development of suitable mathematical models, designed to describe the structure of the data. Time series models enable inference to be made on the values of key parameters, which in turn can be used to forecast future values of significant variables. The need for adequate variable predictions over time is of utmost importance to decision makers, and has led to a plethora of work dedicated to the development of time series forecasting models and methodologies, some of which can be complex and difficult to implement.

In practice, it is often the case that decision makers have access to not one, but several different time series. These can be recordings of distinct variables, or perhaps contemporaneous measurements of the same target variable made using different methods. Such data sets can be modelled using multivariate time series analysis methods, wherein the data is described in a joint way that considers the relationships and interdependencies between variables.

In this thesis, we consider two multivariate time series problems. Firstly, we focus

on the setting of N experts (or forecasters) providing individual point forecasts for the value of a time series of interest y_t at each time t. This collection of predictions can be considered as an N-dimensional multivariate time series, wherein the ith constituent series is given by the univariate time series of forecasts from the ith expert. We consider an online setting, such that the arrival of the N predictions at time t is closely followed by the observation y_t, and this process is repeated sequentially throughout time.

The overall aim in such a situation is to utilise the available information in order to produce the 'best' forecast for the target variable at each time. Thus, the practitioner has two primary options: attempt to identify the best expert, or aggregate the N available forecasts in some way, in order to provide a new, combined forecast.

While the former offers a valid approach, real data generating processes are often complex, and a single model is unlikely to capture the true behaviour of an observed time series; Wang et al. (2023) note that this is often due to the presence of time-varying trends, seasonality changes and structural breaks. Furthermore, forecasts from a particular method may afford us some additional useful information which is not expressed by the other forecasting models, since perhaps forecasts were produced from differing information sets. Thus, it is often more favourable to combine the forecasts in some way, in order to produce a 'combined' or 'ensemble' forecast. By combining several differing forecasts, we can mitigate against the negative effects arising from model misspecification of individual forecasters. Furthermore, we can aim to produce a forecast which contains more information than any constituent forecast, thus reducing forecast error variance if individual forecasters are noisy.

Indeed, the combination of forecasts has been shown to lead to more accurate forecasts than those provided by any single model in a wide variety of applications and settings, including retail (see Ma and Fildes (2021)), election forecasting (see Graefe et al. (2014)), and epidemiology (see Ray et al. (2022)). Moreover, forecast combinations have been shown to outperform individual forecasting methods in several forecasting

competitions. One of the most important findings from the recent M4 competition was that all top-performing methods were provided by combining mostly statistical methods; see Makridakis et al. (2020a).

The aforementioned references make a strong case for the combination of expert forecasts; however, the question is then how should this combination be carried out in practice? A wealth of literature has been dedicated to this problem, with its roots in the seminal work of Bates and Granger (1969), which aims to derive optimal combination weights in the $N = 2$ forecaster case. In the years since, research in the area has expanded significantly, with the development of ever more complex combination methods, machine learning procedures, and the discussion of probabilistic forecast combination; see Wang et al. (2023) for a thorough review of the literature. However, it is often found that simple combination schemes, such as taking the mean, often outperform more complex methods in practice. This phenomenon was termed the 'forecast combination puzzle' by Stock and Watson (2004), and has been the focus of much research. The majority of works on the subject attribute the puzzle to the difficulties that can arise through estimating combination weights; see Claeskens et al. (2016) and Blanc and Setzer (2020) for recent attempts to explain the problem.

Thus, the decision on how to combine the N expert forecasts remains far from trivial. Straightforward approaches like taking the mean clearly reduce estimation error, but is this really the most intuitive method for our problem setting? Firstly, the multivariate time series of forecasters may be nonstationary due to features such as changing expert quality throughout time; for example, perhaps a particular economic forecaster performs well when the market is stable, but less favourably when the environment is volatile. In order to model such behaviour, we will assume that each sequence of point forecasts has some underlying uncertainty, which is unknown and possibly changing dynamically. The question is then should the combination procedure also change throughout time, and if so, how should such a dynamic combination procedure be constructed?

Furthermore, it is often the case that experts are correlated; maybe similar methodologies have been implemented in order to produce their forecasts, or similar data sets utilised. This may lead to issues such as reduced diversification gains from the combination, an underestimate of the uncertainty of the combined forecast, or issues with multicollinearity if a regression-based approach is taken when combining.

Other challenges that can arise when combining forecasts include a lack of historical data and computational inefficiencies. In the case that one has access to information about the past performance of the N experts, one may wish to combine in a way that places a greater bearing on forecasters which have demonstrated superior performance previously. However, historical data is often limited in practical applications, making the quantification of past forecasting performance difficult. Furthermore, even if significant historical data is available, combination methods that utilise this can be complex and require considerable time to run. This is particularly detrimental in a high frequency setting, where forecasts and observed values are arriving in quick succession.

As the first contribution of this thesis, we address some of the above difficulties by proposing a novel dynamic linear model (DLM)-based methodology for combining point forecasts from N experts in an online manner. The proposed methodology is suitable for applications wherein the quality of the experts is changing throughout time, and can be applied even in the presence of little historical data and highly correlated forecasters. Due to the sequential nature of DLMs, the method is fast and computationally efficient, and therefore can be utilised even for a high number of forecasters N.

As the second focus of this thesis, we consider a different type of multivariate time series. Namely, we study bivariate AR(1) time series, which depict circular oscillations when plotted in the complex plane. Such oscillations frequently exist in nature, and often occur when the two components of the bivariate time series correspond to measurements in orthogonal spatial directions; for example, these could be measurements of northerly and easterly wind or ocean current velocities, see Gonella (1972) .

It is often the case that we observe not just one oscillation of a given frequency, but rather multiple oscillatory signals at different frequencies. Such behaviour occurs in a variety of applications, including ocean waves (see Chave et al. (2019)), the geomagnetic field (see Riegert and Thomson (2018)) and electroencephalogram (EEG) data (see Olhede and Ombao (2013)).

In this thesis, we propose a novel stochastic process for modelling two interacting oscillatory signals, simulated from bivariate AR(1) processes; thus, this is a four-dimensional vector autoregressive time series model of order 1. We propose a novel mechanism for generating nonstationary oscillations by 'locking' the phase difference between the two oscillations, such that interactions occur between distinct frequencies. We provide a thorough theoretical investigation of this novel model, which shall include deriving properties such as the autocovariance sequence and various frequency domain characteristics.

The structure of the remainder of this thesis is given as follows. In Chapter 2 we review some of the key forecast combination literature. While many methods for combining expert predictions exist, we focus on linearly weighted combinations, on accord of the associated interpretability. Chapter 3 introduces our proposed forecast combination method based on methodology from the DLM literature, and Chapter 4 considers extensions to deal with the case that one (or more) of the N experts fails to provide a forecast. Chapter 5 introduces our novel model for stochastic oscillations, and demonstrates how nonstationarity may be constructed. In Chapter 6, we review the novel contributions made in this thesis, and provide some possible avenues for further work.

Chapter 2

Literature review

Although we utilise methodologies from the dynamic linear model (DLM) literature in this thesis, this is not the focus of this literature review; a thorough introduction to the DLM methodology and relevant references for further reading are given in Section 3.2. Rather, here we consider the case that a decision maker has access to a set of N individual forecasts for a random variable of interest, and they wish to combine them in some way.

Forecast combination constitutes an extensive area of research, with its foundations rooted in the seminal work of Bates and Granger (1969). In the years since, the topic has garnered growing interest, with recent applications in epidemiology (Ray et al. (2022)), economics (Aastveit et al. (2019)) and the energy sector (Xie and Hong (2016)). An extensive and recent review of forecast combination techniques is given by Wang et al. (2023). Earlier thorough reviews of the topic are given by Clemen (1989) and Timmermann (2006).

The general forecast combination problem considers the following. A forecaster wishes to predict the value of the random variable y_t at time t, given only a vector of N individual forecasts for the variable at that time $\mathbf{f}_t = (f_{1,t}\, f_{2,t}\, \dots\, f_{N,t})$, the corresponding histories of the N forecasters, and the past realisations of the random variable.

The combined point forecast is given as a function of the constituent forecasts $\mathbf{f}_t$ and some parameters of the combination $\mathbf{w}_t$; it is denoted by $\hat{y}_t = C(\mathbf{f}_t; \mathbf{w}_t)$. A forecast combination is linear if the function $C(\mathbf{f}_t; \mathbf{w}_t)$ is linear. In this case, the parameters $\mathbf{w}_t$ can be thought of as weights, and the combined forecast is given by a weighted sum of the individual forecasts. On the other hand, when $C(\mathbf{f}_t; \mathbf{w}_t)$ is a nonlinear function, we have a nonlinear forecast combination method. Due to the interpretability of linear forecast combination methods, we will assume for the remainder of this thesis that the parameter vector $\mathbf{w}_t$ is given by a set of combination weights, and the combined forecast is given by a linearly weighted sum of the individual N forecasts (note, some of the methods in this chapter also introduce an intercept term to account for bias; this will be included explicitly in the parameter vector when necessary).

2.1 Simple point forecast combinations

Those familiar with the forecast combination literature will be aware that simple combination schemes have been shown to outperform more sophisticated methods time and time again, in an extensive range of studies. Even in the recent M4 competition (see Makridakis et al. (2020b)), simple combinations were shown to achieve relatively good forecasting performance when competing with more complex methods. This has lead to the consensus among the community that the forecasting performance of such simple combination schemes is difficult to surpass; see Kang (1986), Clemen (1989), Stock and Watson (2004) and Lichtendahl and Winkler (2020). In particular, assigning equal weights to forecasters, thereby taking the simple average, often outperforms more complex techniques in practice.

Makridakis and Winkler (1983) carried out an empirical study to investigate the accuracy of combined simple average forecasts from individual methods, and concluded that the accuracy of the combined forecast is improved by including a larger number of

individual methods in the combination. However, they also noted that the accuracy of the combined forecast is clearly dependent on which individual methods are included in the combination.

The key advantages of implementing simple average forecast combination are succinctly summarised by Palm and Zellner (1992). Firstly, it is incredibly straightforward to carry out, since all combination weights are equal and do not require estimation. Consequently, this mitigates any uncertainty associated with weight estimation. This is particularly beneficial in problems with a large number of individual forecasters, since both the computational load of estimating performance-based weights, and the associated uncertainties, are undesirable. Furthermore, such performance-based weighting schemes cannot be implemented when we have insufficient information regarding the past performance of individual forecasters, and hence simple averaging provides an appropriate solution. Palm and Zellner (1992) also note that simple average forecasts often average over any individual biases in component forecasters, which leads to reduced overall bias and variance in the combined result.

More recently, Gastinger et al. (2021) carried out a study with approximately $16,000$ time series from various sources. They examined the performance of several different combination procedures, including simple averaging, performance-based weights and machine learning methods. They found that, although different (and sometimes complex) combination strategies achieved the best forecasting accuracy for different data sets, the simple average method displayed more gains in improving accuracy on average.

Even when more complex weights should theoretically perform better, simple averaging often outperforms such weighting schemes in practice. Stock and Watson (2004) termed this phenomenon the 'forecast combination puzzle', and it continues to inspire a vast amount of literature. Smith and Wallis (2009) showed that, in a situation where the theoretically optimal combining weights are equivalent to the simple average, the simple average is expected to outperform the weighted average (where combination weights

have been estimated) systematically, as expected. They showed the detrimental effect of weight estimation on the MSFE when the optimal weights are close to equality in both simulation studies and an empirical analysis. A theoretical explanation for these empirical and simulation results is provided by Claeskens et al. (2016), where the weight estimation step is explicitly included in the optimal weight derivation. That is, they treat the weights as random and derive the resulting expectation and variance of the combined forecast. They show that, in general, estimation of the weights will lead to a biased combined forecast, with a possible increase in variance compared to the fixed-weight case. Claeskens et al. (2016) provide a useful visualisation of these effects for the special $N = 2$ case.

More recent works include Chan and Pauwels (2018), who propose a framework for the study of the theoretical properties of forecast combination, and Blanc and Setzer (2020), who analyse the forecast combination puzzle with concepts from statistical learning theory. The latter decompose the forecast error into bias, variance and irreducible error terms, and explain why simple averaging often provides a favourable bias-variance trade off in practice.

Despite the well-documented good empirical performance of simple average forecast combinations, there are some associated issues. As highlighted by Wang et al. (2023), perhaps most obvious is that the success of such schemes depends heavily on the quality of the forecasts being combined. Clearly, if all of the N available forecasts feature the same bias, then the combined forecast will also be biased. Such a situation may occur if all individual forecasts are produced using the same data set and similar forecasting models, and Thomson et al. (2019) therefore notes that the performance of simple averaging is improved when the component models are highly diverse. A further issue that may arise when implementing simple averaging is the inclusion of a particularly bad forecast in the combination. This can severely degrade the overall accuracy of the combined forecast, and therefore attention should be focused on selecting which

forecasts to combine. Despite these issues, simple average forecast combination remains a difficult benchmark to surpass, and is routinely used as a standard gauge against which to compare new combination techniques.

Other simple schemes, such as taking the median, trimmed mean or mode of forecasts, have also been examined in the literature; see Stock and Watson (2004) and Genre et al. (2013) for examples. Such methods are more robust to extreme forecasts than taking the simple average (Lichtendahl and Winkler (2020)); however, there is no general agreement on whether such techniques perform better.

2.2 Optimal linear combinations

Although simple combination methods are straightforward to implement and often exhibit favourable performance in empirical analyses, a more intuitive approach is to assign forecaster weights based on past performance. The question is therefore how should we assign such weights? A common approach in the literature is to construct the combined h-step ahead forecast at time t as a linear combination of the N individual h-step ahead forecasts at that time,

$$\hat{y}_{t+h|t} = \mathbf{w}'_{t+h|t}\mathbf{f}_{t+h|t},$$

where $\mathbf{f}_{t+h|t}$ is an N-dimensional vector of h-step ahead forecasts, and $\mathbf{w}_{t+h|t}$ is a vector of corresponding weights. To make things clearer, we shall drop the $t+h|t$ subscript, and instead denote the vector of forecasts for the time series at time t, given all available information up to the current time, by $\mathbf{f}_t$. Thus, the above notation is simplified to,

$$\hat{y}_t = \mathbf{w}'_t\mathbf{f}_t.$$

Hence, the prediction from Forecaster $i \in N$ for the value of the series y_t will be written as $f_{i,t}$. We note that the dependence on the available data is implicit in this notation.

The seminal work of Bates and Granger (1969) considers how best to linearly combine a pair of unbiased forecasts, such that the mean square error (MSE) loss of the combination is minimised. Let $e_{i,t}$ denote the error of the ith forecaster at time t; that is, $e_{i,t} = y_t - f_{i,t}$, $i = 1, 2$. Since both forecasters are unbiased, we know that the forecast errors are distributed with zero mean. Denote the variance of the ith forecaster error by σ_i^2, and let the covariance between the forecasters be denoted by $\rho\sigma_1\sigma_2$, where ρ is the correlation between the forecast errors and σ_i is the standard deviation of the ith forecast error. Thus, the error covariance matrix is given by

$$\mathrm{Cov}(e_{1,t}, e_{2,t}) = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix},$$

for all values of time t. In line with Bates and Granger (1969), let us assume the weights for the individual forecasts sum to one, such that the linear combination can be written as

$$\begin{aligned} \hat{y}_t &= w_{1,t} f_{1,t} + w_{2,t} f_{2,t}, \\ &= \alpha f_{1,t} + (1 - \alpha) f_{2,t}, \end{aligned}$$

where $\alpha \in [0, 1]$. The t subscript has been dropped on the coefficient α since the covariance matrix of the errors, and therefore the optimal weights, is constant throughout time. We can now write the forecast error for the combination, $e_t^c = y_t - \hat{y}_t$, as a weighted combination of the individual errors:

$$e_t^c = \alpha e_{1,t} + (1 - \alpha) e_{2,t}.$$

By construction this has zero mean, and variance given by

$$\begin{aligned}\sigma_c^2(\alpha) &= \text{Var}\bigg(\alpha e_{1,t} + (1-\alpha)e_{2,t}\bigg), \\ &= \alpha^2\sigma_1^2 + (1-\alpha)^2\sigma_2^2 + 2\alpha(1-\alpha)\rho\sigma_1\sigma_2. \end{aligned} \tag{2.2.1}$$

We can therefore find the weights which minimise the MSE of the combination by finding those which minimise this variance. Differentiating with respect to α and solving the first order condition, we find that the minimum variance of the combined forecast occurs when,

$$\alpha^* = \frac{\sigma_2^2 - \rho\sigma_1\sigma_2}{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2}. \tag{2.2.2}$$

Hence the optimal weights for all times t are given by,

$$w_1^* = \alpha^* = \frac{\sigma_2^2 - \rho\sigma_1\sigma_2}{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2}, \tag{2.2.3}$$

$$w_2^* = (1-\alpha^*) = \frac{\sigma_1^2 - \rho\sigma_1\sigma_2}{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2}. \tag{2.2.4}$$

We draw your attention to the fact that no positivity constraints are enforced on the weights in this derivation. In their recent paper, Radchenko et al. (2023) provide a useful visual analysis of when negative weights occur. They show that the weight $w_1^* < 0$ if $\rho > \sigma_2/\sigma_1$, and the weight $w_2^* < 0$ if $\rho > \sigma_1/\sigma_2$. On the other hand, it is clear from equations (2.2.3) and (2.2.4) that both weights are positive if the correlation ρ is zero (or negative). Radchenko et al. (2023) also note that, when the correlation equals the variance ratio, one of the forecasters is not included in the combination. This can be seen by rewriting equation (2.2.3) as,

$$w_1^* = \frac{1 - \rho(\sigma_1/\sigma_2)}{(\sigma_1/\sigma_2)^2 + 1 - 2\rho(\sigma_1/\sigma_2)}. \tag{2.2.5}$$

If we then set $\rho = \sigma_1/\sigma_2$, we have,

$$\begin{aligned} w_1^* &= \frac{1 - \rho(\sigma_1/\sigma_2)}{(\sigma_1/\sigma_2)^2 + 1 - 2\rho(\sigma_1/\sigma_2)}, \\ &= \frac{1 - (\sigma_1/\sigma_2)^2}{(\sigma_1/\sigma_2)^2 + 1 - 2(\sigma_1/\sigma_2)^2}, \\ &= 1, \end{aligned}$$

and consequently $w_2^* = 0$. Conversely, when $\rho = \sigma_2/\sigma_1$, we have $w_1^* = 0$ and $w_2^* = 1$.

The reformulation in equation (2.2.5) also aids understanding of the limiting case in which the variance of one of the forecasters approaches zero. Consider the case that the variance of Forecaster 1 tends towards zero, such that $\sigma_1/\sigma_2 \to 0$. It is clear from equation (2.2.5) that $w_1^* \to 1$, and thus the combination simply selects this forecast and sets the other weight to zero. Finally, Radchenko et al. (2023) note that a discontinuity is observed when $\sigma_1 = \sigma_2$ and $\rho = 1$. They show that when one goes from the situation where $\sigma_1 < \sigma_2$ to the situation where $\sigma_1 > \sigma_2$, the weight of Forecaster 1 flips from positive to negative. This indicates that near the discontinuity, the value of the weight w_1^* (and consequently w_2^*) is highly sensitive, with small variations in the parameters leading to significant shifts in w_1^*.

The expected squared error loss associated with using the above optimal weights to combine forecasts can be found by substituting the expressions for the optimal weights into the objective function given by equation (2.2.1),

$$\sigma_c^2(\alpha^*) = \frac{\sigma_1^2\sigma_2^2(1 - \rho^2)}{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2}.$$

In relation to our discussion in the previous section, it may be of interest to compare this with the expected squared error loss from the simple average combination. Denote the simple average combination $\hat{y}^{sa} = (1/2)(f_{1,t} + f_{2,t})$. The variance of the forecast

error under this combination scheme is therefore given by,

$$\sigma_{sa}^2 = \frac{1}{4}\sigma_1^2 + \frac{1}{4}\sigma_2^2 + \frac{1}{2}\sigma_1\sigma_2\rho.$$

Hence we can write the ratio of the expected squared error loss for the the derived optimal combination to the equally weighted combination as,

$$\frac{\sigma_{sa}^2}{\sigma_c^2(\alpha^*)} = \frac{(\sigma_1^2 + \sigma_2^2)^2 - 4\rho^2\sigma_1^2\sigma_2^2}{4\sigma_1^2\sigma_2^2(1-\rho^2)}. \tag{2.2.6}$$

It is possible to show that this ratio is always greater than or equal to one; that is, the expected squared error loss from optimally derived weights is less than that obtained from simple averaging. To see this, note that,

$$(a+b)^2 \geq 4ab,$$

and therefore,

$$(\sigma_1^2 + \sigma_2^2)^2 \geq 4\sigma_1^2\sigma_2^2.$$

Consequently, we have

$$(\sigma_1^2 + \sigma_2^2)^2 - 4\rho^2\sigma_1^2\sigma_2^2 \geq 4\sigma_1^2\sigma_2^2(1-\rho^2),$$

and therefore the fraction given in equation (2.2.6) is always greater than or equal to one. We see that this is an equality in the case that $\sigma_1 = \sigma_2$. This methodology was extended to deal with more than two individual forecasters by Newbold and Granger (1974). Assume we wish to linearly combine $N \geq 2$ unbiased forecasters, with $N \times N$ forecast error covariance matrix $\mathbf{\Sigma}_t$. Newbold and Granger (1974) assume that combination

weights are bounded and constrained to sum to one:

$$\mathbf{w}_t'\boldsymbol{\iota} = 1, \quad 0 \leq w_{i,t} \leq 1, \quad \text{for } i = 1, \ldots, N,$$

where $\boldsymbol{\iota}$ is an $N \times 1$ vector of 1s. As before, the forecast error of the combination is given by a weighted combination of the individual forecast errors,

$$e_t^c = w_{1,t}e_{1,t} + w_{2,t}e_{2,t} + \cdots + w_{N,t}e_{N,t},$$

which can be expressed in vector notation as $e_t^c = \mathbf{w}_t'\mathbf{e}_t$, where $\mathbf{e}_t = y_t\boldsymbol{\iota} - \mathbf{f}_t$. The variance of the combined forecast error can hence be written as,

$$\begin{aligned}\sigma_c^2(\mathbf{w}_t) = \text{Var}(\mathbf{w}_t'\mathbf{e}_t) &= E\{\mathbf{w}_t'\mathbf{e}_t\mathbf{e}_t'\mathbf{w}_t\} - E\{\mathbf{w}_t'\mathbf{e}_t\}^2 \\ &= \mathbf{w}_t'E\{\mathbf{e}_t\mathbf{e}_t'\}\mathbf{w}_t - (\mathbf{w}_t'E\{\mathbf{e}_t\})^2 \\ &= \mathbf{w}_t'\boldsymbol{\Sigma}_t\mathbf{w}_t.\end{aligned}$$

We wish to find the optimal weight vector that minimises this expression, $\mathbf{w}_t^*$, under the constraint that the sum of the weights is equal to one. Thus, we write the Lagrangian,

$$\mathcal{L}(\mathbf{w}_t, \lambda) = \mathbf{w}_t'\boldsymbol{\Sigma}_t\mathbf{w}_t - \lambda(\mathbf{w}_t'\boldsymbol{\iota} - 1).$$

The first order derivatives are then given by,

$$\frac{d}{d\mathbf{w}_t}\mathcal{L}(\mathbf{w}_t, \lambda) = \mathbf{w}_t'(\boldsymbol{\Sigma}_t' + \boldsymbol{\Sigma}_t) - \lambda\boldsymbol{\iota} = 2\boldsymbol{\Sigma}_t\mathbf{w}_t - \lambda\boldsymbol{\iota}, \tag{2.2.7}$$

$$\frac{d}{d\lambda}\mathcal{L}(\mathbf{w}_t, \lambda) = \mathbf{w}_t'\boldsymbol{\iota} - 1. \tag{2.2.8}$$

Setting equation (2.2.7) equal to zero and rearranging for the weight vector gives

$$\mathbf{w}_t = \frac{\lambda}{2}\boldsymbol{\Sigma}_t^{-1}\boldsymbol{\iota}. \tag{2.2.9}$$

This can be substituted into equation (2.2.8) in order to give an expression for the Lagrange multiplier, $\lambda = 2(\boldsymbol{\iota}'\boldsymbol{\Sigma}_t^{-1}\boldsymbol{\iota})^{-1}$. Finally, this expression for λ can be substituted into equation (2.2.9) in order to give us the optimal constrained weights under MSE loss for $N \geq 2$ unbiased forecasters:

$$\begin{aligned} \mathbf{w}_t^* &= \frac{\lambda}{2}\boldsymbol{\Sigma}_t^{-1}\boldsymbol{\iota}, \\ &= (\boldsymbol{\iota}'\boldsymbol{\Sigma}_t^{-1}\boldsymbol{\iota})^{-1}\boldsymbol{\Sigma}_t^{-1}\boldsymbol{\iota}. \end{aligned} \tag{2.2.10}$$

Henceforth, we will refer to these weights as the theoretically optimal covariance weights, as they are computed using the error covariance matrix. It can be shown that these weights are equivalent to the simple average weights when all forecast errors have the same variance σ^2 and correlation ρ. In this case, the inverse of the covariance matrix is given by

$$\begin{aligned} \boldsymbol{\Sigma}_t^{-1} &= \frac{1}{\sigma^2(1-\rho)}\left(\mathbb{I}_N - \frac{\rho}{1+(N-1)\rho}\boldsymbol{\iota}\boldsymbol{\iota}'\right), \\ &= \frac{1}{\sigma^2(1-\rho)(1+(N-1)\rho}((1+(N-1)\rho)\mathbb{I}_N - \rho\boldsymbol{\iota}\boldsymbol{\iota}'), \end{aligned}$$

where $\mathbb{I}_N$ denotes the $N \times N$ identity matrix. Inserting this expression into equation (2.2.10), we see that

$$\mathbf{w}_t^* = \frac{\sigma^2(1+(N-1)\rho)}{N}\frac{\boldsymbol{\iota}}{\sigma^2(1+(N-1)\rho)} = \left(\frac{1}{N}\right)\boldsymbol{\iota}.$$

The general solution for the optimal linear unconstrained weights under MSE loss was characterised by Timmermann (2006), where it is assumed that the outcome y_t and

the individual forecasts $\mathbf{f}_t$ follow a joint Gaussian distribution:

$$\begin{pmatrix} y_t \\ \mathbf{f}_t \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_{y,t} \\ \boldsymbol{\mu}_{\mathbf{f},t} \end{pmatrix}, \begin{pmatrix} \sigma^2_{y,t} & \boldsymbol{\sigma}'_{y\mathbf{f},t} \\ \boldsymbol{\sigma}_{y\mathbf{f},t} & \boldsymbol{\Sigma}_{\mathbf{f},t} \end{pmatrix} \right).$$

In this derivation, Timmermann (2006) wish to minimise the expected square forecast error,

$$E\{(y_t - \mathbf{w}'_t\mathbf{f}_t)^2\} = (\mu_{y,t} - \mathbf{w}'_t\boldsymbol{\mu}_{\mathbf{f},t})^2 + \sigma^2_{y,t} + \mathbf{w}'_t\boldsymbol{\Sigma}_{\mathbf{f},t}\mathbf{w}_t - 2\mathbf{w}'_t\boldsymbol{\sigma}_{y\mathbf{f},t}.$$

Once again, this is done by differentiating with respect to the weight vector,

$$\begin{aligned} \frac{\partial E\{(e^c_t)^2\}}{\partial \mathbf{w}_t} &= 2(\mu_{y,t} - \mathbf{w}'_t\boldsymbol{\mu}_{\mathbf{f},t})(-\boldsymbol{\mu}_{\mathbf{f},t}) + \mathbf{w}'_t(\boldsymbol{\Sigma}'_{\mathbf{f},t} + \boldsymbol{\Sigma}_{\mathbf{f},t}) - 2\boldsymbol{\sigma}_{y\mathbf{f},t}, \\ &= -2\mu_{y,t}\boldsymbol{\mu}_{\mathbf{f},t} + 2\boldsymbol{\mu}_{\mathbf{f},t}\boldsymbol{\mu}'_{\mathbf{f},t}\mathbf{w}_t + 2\boldsymbol{\Sigma}_{\mathbf{f},t}\mathbf{w}_t - 2\boldsymbol{\sigma}_{y\mathbf{f},t}. \end{aligned}$$

Setting this to zero, and assuming that $(\boldsymbol{\mu}_{\mathbf{f},t}\boldsymbol{\mu}'_{\mathbf{f},t} + \boldsymbol{\Sigma}_{\mathbf{f},t})$ is invertible, one can solve for the optimal weight vector,

$$\mathbf{w}^*_t = (\boldsymbol{\mu}_{\mathbf{f},t}\boldsymbol{\mu}'_{\mathbf{f},t} + \boldsymbol{\Sigma}_{\mathbf{f},t})^{-1}(\mu_{y,t}\boldsymbol{\mu}_{\mathbf{f},t} + \boldsymbol{\sigma}_{y\mathbf{f},t}),$$

as given in Timmermann (2006). A constant can be included in the linear combination, in order to account for bias in the individual forecasters (recommended under MSE loss by Granger and Ramanathan (1984)). In this case, Timmermann (2006) gives the optimal values for the constant and the combination weights as

$$\begin{aligned} w^*_{0,t} &= \mu_{y,t} - \mathbf{w}^*_t\boldsymbol{\mu}_{\mathbf{f},t}, \\ \mathbf{w}^*_t &= \boldsymbol{\Sigma}^{-1}_{\mathbf{f},t}\boldsymbol{\sigma}_{y\mathbf{f},t}. \end{aligned}$$

Although we have seen in this section that the theoretical optimal weights under

MSE loss are sometimes equivalent to the simple average weights, this is rarely the case in practice. When the set of forecasters have differing variances and correlations, knowledge of the forecast error covariance matrix is required in order to apply optimal weighting. In real world applications, this contributes to the forecast combination puzzle, as having to estimate the weights no longer guarantees reductions in MSE loss versus the simple average. In the next sections, we turn our attention to weight estimation.

2.3 Regression-based weights

Many different methods have been proposed for assigning weights in linear forecast combinations, including performance-based weights (Pawlikowski and Chorowska (2020)), criteria-based weights (Petropoulos et al. (2018)) and Bayesian weights (Bunn (1975)); see Wang et al. (2023) for a thorough review of such methods. An alternative and popular approach is to model the forecast combination under a linear regression framework, and estimate the weights using ordinary least squares (OLS). This idea was first proposed in the seminal work of Granger and Ramanathan (1984).

In their work, Granger and Ramanathan (1984) consider three possible regressions, wherein the response variable is given by the observations y_t, and the predictor variables are provided by the vector of individual forecasts $\mathbf{f}_t$. The idea is that the realisations of the target variable are regressed onto the N individual forecasts using all the data up to the current time t. As such, the method assumes constant weights, but allows the estimated weights to change in response to newly observed data. Each of the three regression formulae impose different restrictions on the combination weights, as shown below:

(i) $y_t = \mathbf{w}_t'\mathbf{f}_t + \epsilon_t, \quad \text{s.t.} \quad \sum \mathbf{w}_t = 1,$

(ii) $y_t = \mathbf{w}_t'\mathbf{f}_t + \epsilon_t,$

(iii) $y_t = w_{0,t} + \mathbf{w}_t'\mathbf{f}_t + \epsilon_t$.

Regression (i) arises from the assumption that all individual forecasters are unbiased. The fact that the weights are constrained to sum to one therefore ensures that the combined forecast is unbiased. Implementation of this regression model for $N = 2$ unbiased forecasters therefore estimates the 'optimal' weights derived by Bates and Granger (1969). On the other hand, both regressions (ii) and (iii) are unconstrained. The intercept term in regression (iii) enables adjustment in case of biased forecasters; Granger and Ramanathan (1984) argued that this was a superior method to the popular 'optimal' weight combination of Newbold and Granger (1974), since it enables the computation of an unbiased combined forecast even when constituent forecasts are biased.

Furthermore, since the 'optimal' covariance weights can be interpreted as a constrained least squares regression with no intercept term, the Granger and Ramanathan (1984) weights will result in a lower combined within sample MSE. This is acknowledged by De Menezes et al. (2000) in their review of practical guidelines for forecast combination; however, they caution against applying Granger and Ramanathan (1984)'s regression framework without proper care. In particular, De Menezes et al. (2000) note that, since the regression includes an intercept term in the predictor variables, the time series being forecast must be stationary (or made stationary) in order for the OLS regression to work. Moreover, De Menezes et al. (2000) note that issues can occur due to serial correlation in forecast errors, and the presence of multicollinearity in the case that the constituent forecasters are correlated. Difficulties can also arise in practical implementations where there is a large number of constituent forecasters N; see Stock and Watson (2006) for notes on combining many predictors.

2.4 Time-varying weights

The regression-based combination methods of Granger and Ramanathan (1984) assume that the combination weights are constant, and this framework is therefore impractical for cases in which individual forecaster quality evolves with time. In such cases, it is instead favourable to let the combination weights change accordingly. As discussed by Timmermann (2006), the general idea for such dynamic weighting schemes usually involves assigning a larger weight to whichever forecaster has most recently performed the best, or the application of an adaptive updating scheme wherein more importance is placed on the recent performance of individual forecasters, rather than their past behaviour.

2.4.1 Adaptive updating schemes for time-varying parameter estimation

The need for time-varying weights was considered by Granger and Newbold (1977), who suggested five different methods for adaptive weight estimation. These methods work by either considering historical forecasting performance over a fixed moving window, or introducing a parameter that discounts the effects of forecasting performance in the distant past. While the authors presented a total of five methods, two of these can be considered as simplifications, wherein correlations between forecast errors are ignored. Therefore, here we provide an overview of three of the suggested methods.

Firstly, Granger and Newbold (1977) suggested using the previous k observations to calculate the weights at each time point, such that the optimal weights are given by

$$\mathbf{w}_t = (\boldsymbol{\iota}'\boldsymbol{\Sigma}_t^{-1}\boldsymbol{\iota})^{-1}\boldsymbol{\Sigma}_t^{-1}\boldsymbol{\iota},$$

where the elements of the covariance matrix are given by

$$\Sigma_{ij,t} = \frac{1}{k} \sum_{\tau=t-k+1}^{t} e_{i,\tau} e_{j,\tau}. \tag{2.4.1}$$

This method assigns the most weight to forecasters who have displayed the best performance in the recent past; however, as noted by Diebold and Pauly (1987), the choice of window size k is arbitrary and will greatly impact the estimated weights.

Granger and Newbold (1977) also suggested an adaptive scheme wherein the estimated weights at time t are computed as a function of the estimated weights at the previous time $t-1$, in addition to the estimated forecast errors over the previous k observations,

$$w_{i,t} = \alpha w_{i,t-1} + (1-\alpha)\Bigg(\sum_{\tau=t-k+1}^{t} e_{i,\tau}^2 \Bigg) \Bigg/ \Bigg(\sum_{j=1}^{N} \Bigg(\sum_{\tau=t-k+1}^{t} e_{j,\tau}^2 \Bigg)^{-1} \Bigg), \quad i = 1, \ldots, N,$$

for some parameter $0 < \alpha < 1$. While this method may lead to smoother weight estimates, we note that covariance information is ignored in this case.

Finally, Granger and Newbold (1977) considered using the entire history of forecast errors to determine the weights, but assigned a discount factor $\lambda \geq 1$ such that more recent observations were given a greater weighting,

$$w_{i,t} = \Bigg(\sum_{\tau=1}^{t} \lambda^{\tau} e_{i,\tau}^2 \Bigg)^{-1} \Bigg(\sum_{j=1}^{N} \Bigg(\sum_{\tau=1}^{t} \lambda^{\tau} e_{j,\tau}^2 \Bigg)^{-1} \Bigg), \quad i = 1, \ldots, N.$$

Similar to the fixed window case, we note that the selection of λ is arbitrary, and the choice of this parameter will have a significant impact on the results.

Regardless of whether exponential discounting or rolling window methods are used to estimate the weights, placing more importance on recent observations leads to increased variability in the estimates. This is intuitive; the estimated weights will be influenced by variations in the recently observed data, and will not be largely affected by any

long term trends. Hence, as noted by Timmermann (2006), if the true underlying data generating process for the component forecasters is indeed covariance stationary, such methods will be overly sensitive to recently observed data, which in turn will result in unsatisfactory forecasting performance.

The above three methods for the estimation of time-varying combination weights are based on the original 'optimal' weights derived by Newbold and Granger (1974), and work by estimating the covariance matrix of the forecast errors at each time point. Diebold and Pauly (1987) compared the performance of such 'variance-covariance' methods with regression-based approaches for time-varying weights.

We recall that using the regression framework (i) of Granger and Ramanathan (1984) in the case of $N = 2$ forecasters leads to results numerically equivalent to the 'optimal' weights derived by Bates and Granger (1969) (given by equation (2.2.2)). Diebold and Pauly (1987) therefore surmised that the above 'variance-covariance' methods for time-varying weight estimation could be successfully extended into a regression context, which would benefit from the favourable characteristics of this framework (namely, the relaxation of the constraint that weights sum to unity, and an ability to deal with biases through an added intercept term).

To this end, Diebold and Pauly (1987) used weighted least squares (WLS) regression to minimise the matrix weighted average of square forecast errors throughout time, where more weight is assigned to recent observations. Furthermore, Diebold and Pauly (1987) also considered two regression-based systematically time-varying parameter models: namely, a deterministic evolution model for the combination weights and a stochastic parameter model. Their numerical examples indicated that these time-varying generalisations to the regression framework offered significant reductions in forecast error.

2.4.2 Explicitly modelling parameter time-variations

Modelling time variations in the weight parameters explicitly was also considered by Sessions and Chatterjee (1989) and LeSage and Magura (1992), who consider combination models of the form:

$$y_t = \mathbf{x}_t \boldsymbol{\beta}_t' + \epsilon_t, \tag{2.4.2}$$

$$\boldsymbol{\beta}_t = \mathbf{T}\boldsymbol{\beta}_{t-1} + \boldsymbol{\nu}_t. \tag{2.4.3}$$

Here, the vector of predictor variables $\mathbf{x}_t$ is given by the N individual forecasts $\mathbf{f}_t = (f_{1,t}\, f_{2,t} \ldots f_{N,t})$ in addition to an intercept term, such that we can write $\mathbf{x}_t = (1\, \mathbf{f}_t)'$. The corresponding vector of weights is given by $\boldsymbol{\beta}_t$. Equation (2.4.2) describes the evolution of the observed series as a linear function of the forecasts, with a disturbance term given by ϵ_t (here we choose not to provide a distribution for this noise term in order to keep the definition general, distributional assumptions are imposed in Chapter 3). Equation (2.4.3) describes the evolution of the regression coefficients (combination weights) throughout time, where the matrix $\mathbf{T}$ determines the form of this evolution and the disturbance term $\boldsymbol{\nu}_t$ describes the stochastic effects. When the matrix $\mathbf{T}$ is set to the identity matrix, $\mathbf{T} = \mathbf{I}$, the weights are modelled according to a random walk. This is the approach taken by LeSage and Magura (1992), and later Stock and Watson (2004), wherein the empirical performance of various combination methods is examined on data for output growth in seven OECD countries. This forecast combination framework is sometimes referred to as a time-varying-parameter (TVP) model in the literature; for example see Stock and Watson (2004).

Under certain assumptions, the above equations define a dynamic linear model (DLM), wherein combination weights can be recursively estimated using the Kalman filter (see Chapter 3). Stock and Watson (2006) note that, unlike the OLS regression method of Granger and Ramanathan (1984), this approach is appropriate for combining

high numbers of individual forecasts by virtue of the additional structure imposed on the weights.

However, this approach requires the user to specify the variance of the noise term $\boldsymbol{\nu}_t$, known as the system covariance. It is this that specifies the magnitude of the time variation in the weights, and knowledge of this variance is necessary in order to apply the standard Kalman filter to the problem (see Section 3.2.1). Of course, it is possible to estimate this quantity using standard maximum likelihood methods; however, this technique fails to take properly into account the uncertainty about the estimated parameter, and requires historical data in order to be implemented. Stock and Watson (2006) also note that the estimation of this covariance is particularly difficult in the case of many forecasters.

Sessions and Chatterjee (1989) circumvent the issues associated with an unknown system covariance by extending a method first proposed by Snyder (1985), which Sessions and Chatterjee (1989) entitle the Synder DLM (SDLM). This method replaces the error term $\boldsymbol{\nu}_t$ by $\boldsymbol{\alpha} u_t$, where u_t is a scalar variable distributed according to $u_t \overset{iid}{\sim} N(0, \sigma^2)$. Snyder (1985) derives updating equations for the unknown vector $\boldsymbol{\beta}_t$, where the algorithm is fully recursive for known $\boldsymbol{\alpha}$. In the case that $\boldsymbol{\alpha}$ is unknown, Snyder (1985) suggests estimation through maximum likelihood estimation (MLE) from forecast errors. Sessions and Chatterjee (1989) present an alternative method for the estimation of $\boldsymbol{\alpha}$ such that this can be carried out recursively, and apply this technique to the forecast combination problem. However, the proposed method for dealing with unknown system covariance is not applicable in the case that this parameter is changing throughout time.

LeSage and Magura (1992) propose a method for combining forecasts which uses the above time varying parameter framework combined with Gordon and Smith (1990)'s multiprocess mixture model. The aim of their method is to develop a combination procedure that effectively deals with outliers in the observed data, and abrupt struc-

tural shifts in the weight parameters. Such abrupt sudden shifts correspond to sudden changes in the accuracy of the individual forecasters, which LeSage and Magura (1992) note may be due to changes in external conditions, changes in the model used to provide the constituent forecasts, or changes in how the forecasts are reported. Despite the fact that the proposed method was tailored towards data featuring abrupt structural shifts, it did not improve upon simpler TVP-based combination methods in their empirical analysis.

2.5 Density combination

Although point forecast combination techniques are sufficient for many applications, a growing literature has been dedicated to the combination of density forecasts in recent years. Clearly, density forecasts provide not only a point estimate of the value of interest, but also a quantification of the associated uncertainty. This provides decision makers with a more informed understanding of risk, consequently enabling better decisions.

When working with point predictions, the objective of the combination is to improve the accuracy of the combined forecast. On the other hand, density combination aims to produce a 'good' distribution. Accordingly to Gneiting et al. (2007) and later Wang et al. (2023), the intent of probabilistic forecasting is 'to maximise the sharpness of the forecast distributions subject to calibration' based on the available information set. Therefore, in order to quantify the quality of a distribution, practitioners often use scoring rules which reward these desirable measures.

2.5.1 Linear pooling

Denote the density forecast provided by Forecaster i for the variable y_{t+1} at time t by $p_{i,t}(y)$. Here, y serves as a dummy variable representing all potential outcomes of

the random variable y_{t+1} under the forecast distribution. Assume we have N different density forecasts, such that $i \in \{1, \ldots, N\}$. In the literature, a common method for the aggregation of these forecasts is to take a mixture distribution with some mixture weights w_i,

$$p_t(y) = \sum_{i=1}^{N} w_i p_{i,t}(y),$$

known as a 'linear opinion pool' (dating back to at least Stone (1961)). This is intuitive, and mirrors the approach of taking a linear combination of point forecasts in the framework discussed previously. In order to ensure that the combined density is non-negative and integrates to one, the weights are taken to be non-negative and sum to one.

As in the point forecast combination setting, the question is therefore how to choose the weights? Of course, the most obvious approach would be to assign equal weights of $1/N$ to each distribution; see Wallis (2005) for a review. A more involved procedure for assigning mixture weights was later proposed by Hall and Mitchell (2007), who suggest using the weights which maximise the average logarithmic score of the combined density forecast,

$$\begin{aligned}\mathbf{w}^* =& \underset{\mathbf{w}}{\text{argmax}} \left\{ \frac{1}{T-1} \sum_{t=1}^{T-1} \ln p_t(y_{t+1}) \right\}, \\ =& \underset{\mathbf{w}}{\text{argmax}} \left\{ \frac{1}{T-1} \sum_{t=1}^{T-1} \ln \left(\sum_{i=1}^{N} w_i p_{i,t}(y_{t+1}) \right) \right\},\end{aligned}$$

where $p_{i,t}(y_{t+1})$ is the probability assigned to the realised outcome y_{t+1} by the forecast distribution from Forecaster i at time t. Following this, Conflitti et al. (2015) proposed a simple iterative algorithm to estimate such weights; we provide a more detailed description of these two methods in Section 4.2.

2.5.2 Bayesian Model Averaging

Bayesian Model Averaging (BMA) provides an alternative method for weighting density forecasts in a combination. For clarity, we shall write the dependence of the density forecasts on the previously observed data $\mathcal{D}_t$ explicitly, such that $p_{i,t}(y) = p_{i,t}(y|\mathcal{D}_t)$, where once again we have used y as a dummy variable to represent all potential outcomes of the random variable y_{t+1}. Assume each forecast i corresponds to a different model M_i, such that we can write $p_{i,t}(y|\mathcal{D}_t) = p_{i,t}(y|M_i, \mathcal{D}_t)$. In the BMA framework, the combined posterior probability forecast given the data $\mathcal{D}_t$ is written as,

$$p_t(y|\mathcal{D}_t) = \sum_{i=1}^{N} P(M_i|\mathcal{D}_t) p_{i,t}(y),$$

where $P(M_i|\mathcal{D}_t)$ denotes the posterior probability of model M_i given the data up to time t. This represents the weighted average of the posterior distributions for each model considered, with weights based on their posterior model probabilities. The posterior probability that model M_i is the true model is found using Bayes theorem,

$$P(M_i|\mathcal{D}_t) = \frac{P(M_i)P(\mathcal{D}_t|M_i)}{\sum_{i=1}^{N} P(M_i)P(\mathcal{D}_t|M_i)},$$

where $P(\mathcal{D}_t|M_i)$ is the marginal likelihood of model M_i given by,

$$P(\mathcal{D}_t|M_i) = \int_{\boldsymbol{\theta}_i} P(\mathcal{D}_t|\boldsymbol{\theta}_i, M_i) P(\boldsymbol{\theta}_i|M_i) d\boldsymbol{\theta}_i.$$

Here, $\boldsymbol{\theta}_i$ denotes the vector of parameters of model M_i, $P(\boldsymbol{\theta}_i|M_i)$ denotes the prior density of $\boldsymbol{\theta}_i$ under model M_i, $P(\mathcal{D}_t|\boldsymbol{\theta}_i, M_i)$ is the likelihood function of model M_i and $P(M_i)$ is the prior probability that M_i is the true model. We refer the reader to Koop (2003) for more details on implementing BMA.

Although BMA provides a conceptually straightforward approach to density combination, there are challenges that arise in practical implementation. Firstly, as detailed

by Wang et al. (2023), BMA assumes that the true model is included in the set of candidate models $\{M_1, \ldots, M_N\}$. This means that as the number of observations tends towards infinity, all but one of the posterior probabilities tend towards zero, with the weight of the single 'best' model tending towards unity. While this is appropriate in the case that the true model is included in the model set, this is often not the case in practice, and can lead to a combination that is suboptimal under logarithmic scoring (see Diebold (1991)).

Furthermore, BMA assigns fixed probabilities to each of the component models. As stated by Aastveit et al. (2019), BMA ignores the uncertainty of the weights attached to each model, which can be very large. In turn, this can lead to unstable combined forecasts when structural changes are present in the forecasting performance of the individual models.

With the aim of addressing this challenge of BMA, Dynamic model averaging (DMA) was proposed by Raftery et al. (2010) in order to deal with evolving model performance over time. This is done through the introduction of a forgetting factor $\alpha \in (0, 1]$ that discounts older data when computing the posterior model probabilities, allowing DMA to deal effectively with non-stationary settings. Specifically, Raftery et al. (2010) let $\pi_{t|t-1,i}$ represent the weight of model M_i at time t, based on data observed up to time $t-1$. When an observation is made at time t, the weight of each model is then updated, according to,

$$\pi_{t|t,i} = \frac{\pi_{t|t-1,i} p_{i,t-1}(y_t|\mathcal{D}_{t-1})}{\sum_{j=1}^{N} \pi_{t|t-1,j} p_{j,t-1}(y_t|\mathcal{D}_{t-1})},$$

where $p_{i,t-1}(y_t|\mathcal{D}_{t-1})$ represents the forecast density assigned to the realised outcome y_t by model M_i. The weight of model M_i at the next time $t+1$ is then adjusted using the forgetting factor α and a small constant c (typically in the range $0 < c \ll 1$), according

to,

$$\pi_{t+1|t,i} = \frac{(\pi_{t|t,i})^{\alpha} + c}{\sum_{j=1}^{N}[(\pi_{t|t,j})^{\alpha} + c]}.$$

The choice of α determines the rate of forgetting of past information, such that when α is close to one past observations have a longer-lasting influence on the model weights, making the system more stable and less responsive to recent changes in the data. The role of the constant c is to prevent degenerate weights and associated numerical problems such as division by zero. A very small c has minimal impact on the relative weighting of models; however, it can lead to models with poor performance being effectively excluded from the combination. Slightly larger values of c may be used if there is a desire to ensure that all models retain at least a minimal baseline weight, regardless of their performance.

Bernaciak and Griffin (2024) state that DMA has exhibited good empirical performance in economic applications, while avoiding the significant computational burden associated with alternative methods, such as sequential Monte Carlo. However, they show that the choice of forgetting factor α can have a significant impact on the performance.

In order to provide more robustness to parameter choice, Bernaciak and Griffin (2024) propose a loss discounting framework (LDF), wherein layers of 'meta-models' are defined which use different values of forgetting factor. The proposed flexible discounting scheme can be used for both dynamic model averaging and dynamic model selection. In order to carry out dynamic model averaging, Bernaciak and Griffin (2024) begin by defining a pool of forecast combination densities by applying DMA with different values of discount factor, and refer to these as 'meta-models'. The best meta-model average is then found by applying exponential discounting to the past performance of the meta-models.

2.6 Summary

The contributions presented in Chapter 3 and Chapter 4 of this thesis build upon the forecast combination literature. In particular, in Chapter 3 we introduce a point forecast combination procedure which utilises a similar time-varying parameter framework as that implemented by Sessions and Chatterjee (1989) and LeSage and Magura (1992). We formalise the procedure by providing a complete step-by-step guide for the combination of point forecasts within the DLM framework, and compare this method with key benchmarks. In particular, we assess our DLM-based procedure by comparing the forecasting performance with that of the simple average forecast, due to its favourable empirical performance. Furthermore, we also compare the forecasting performance with that of the optimal covariance weights given by equation (2.2.9), and the regression-based weights of Granger and Ramanathan (1984). This is done both in a stationary setting, and a dynamic setting using 'rolling' estimates. In Chapter 4, we consider density combination. We utilise a linear opinion pool to combine density forecasts, and assign mixture weights using an extension of the method proposed by Hall and Mitchell (2007); a more in depth review of this methodology is given in Section 4.2.

Chapter 3

DLM-based point forecast combination

In this chapter, we develop a dynamic linear model (DLM)-based point forecast combination method. This is motivated by our problem setting, wherein N individual experts provide (possibly correlated) point forecasts for some target variable of interest y_t at time t. The forecasts arrive sequentially, such that they must be combined in an online manner. Furthermore, we assume that there is some underlying uncertainty associated with each forecaster, which is unknown and dynamically changing throughout time. The proposed method assigns combination weights to each of the N forecasters, which are allowed to evolve as more information becomes available. We demonstrate this in simulations featuring correlated forecasters and changing forecaster quality throughout time, before considering two empirical applications.

3.1 Introduction

We seek a method for combining forecasts which allows combination weights to evolve in response to dynamic forecaster quality. Existing methods for explicitly modelling time variation in the combination weights include those of LeSage and Magura (1992)

and Sessions and Chatterjee (1989), which were introduced in Chapter 2 and will be discussed in more depth shortly. Our approach will use a DLM-based model for time-varying weights which enables sequential updating as more data is observed. In addition to the ability to adapt to changing forecaster quality, this approach is appealing due to its potential to overcome the common difficulties of correlated forecasters and a lack of historical data (small sample size). The latter can make the estimation of weights based on past performance difficult, particularly for methods such as the optimal weights of Newbold and Granger (1974), where a sufficient number of observations is required in order to estimate the covariance matrix. On the other hand, our DLM-based model allows suitable combination weights to be estimated from the outset. The sequential nature of DLMs also makes such models suitable for situations where forecasts are arriving with high-frequency, making this an applicable framework for our problem set up.

Although both LeSage and Magura (1992) and Sessions and Chatterjee (1989) utilise methodology from the DLM literature, this is not their primary focus. The former discuss how the simple DLM-based combination model given in their paper is undesirable for their purpose, since it assumes a single covariance structure for the weight parameter variation over the entire data sample. To remedy this, they propose a multiprocess mixture model, which derives five changepoint models from the basic model by adjusting the observation variance and the weight covariance. In order to avoid an exponential increase in the number of models under consideration at each time point, LeSage and Magura (1992) invoke an approximation to collapse the set of possible models back down to five at each time period.

Sessions and Chatterjee (1989) include a DLM-based combination method in their model evaluation, but the primary focus for the paper is given to an extension of a method first proposed by Snyder (1985), and the introduction of two ad-hoc methods for combining forecasts with non-stationary weights. Little information is given detailing

how to select necessary parameters for the DLM-based method, nor is its performance assessed for different parameter choices.

Thus, as the first contribution of this thesis we provide an explicit and coherent description of how to implement time-varying parameter forecast combination within the DLM framework. This type of forecast combination procedure is suitable for situations where forecasts are arriving with high frequency, and expert quality is changing dynamically throughout time. We provide advice on appropriate choices for prior parameters, and clear instructions for dealing with an unknown variance of the noise in the observation equation. Furthermore, a discount factor is introduced to deal with the case that the covariance of the noise in the evolution equation for the combination weights is also unknown.

We compare this combination method with various benchmarks in several simulation studies, and show an improvement on simple averaging for various metrics. We also consider two empirical applications, which help to highlight when this forecast combination method is suitable (and when it is not). To the best of our knowledge, a detailed guide on implementing DLM-based forecast combination, and an exploration of its performance in this way, has not been provided in the literature before.

Furthermore, we provide a formulation of the relevant filtering equations such that this forecast combination method can be easily integrated with methodology from the adaptive discount factor selection literature. The synthesis of such methods with DLM-based forecast combination is a novel contribution, and is discussed further in Chapter 4.

3.2 Background on DLMs

Dynamic linear models (DLMs) define a specific form of state-space model, wherein both the observation equation and the state equation are linear. Since their development in

an engineering setting in the early 1960's, DLMs have been implemented in a vast range of fields thanks to their straightforward approach to Bayesian modelling and forecasting. DLMs provide flexible modelling tools, and have recently been applied to areas including finance (Fisher et al. (2020)), environment (Ahn et al. (2017)) and epidemiology (Khan et al. (2021)). In the forecast combination literature, alongside the work of LeSage and Magura (1992) and Sessions and Chatterjee (1989), DLMs have also been implemented by Terui and van Dijk (2002) and Raftery et al. (2010). The following methodology in this chapter is largely in line with the work of West and Harrison (1997). Other valuable references in the area are given by Prado and West (2010) and Harvey (1990).

The general univariate DLM describes how a univariate time series y_t depends linearly on some underlying vector of size $n \times 1$, known as the state vector $\boldsymbol{\theta}_t$. The formal definition is given below.

Definition 3.2.1. *For each t, the general univariate DLM is defined by:*

$$\begin{aligned} &\textit{Observation equation:} \quad y_t = \boldsymbol{F}_t'\boldsymbol{\theta}_t + \nu_t, \qquad \nu_t \sim N(0, V_t), && (3.2.1)\\ &\textit{System equation:} \quad \boldsymbol{\theta}_t = \boldsymbol{G}_t\boldsymbol{\theta}_{t-1} + \boldsymbol{\omega}_t, \quad \boldsymbol{\omega}_t \sim N(\mathbf{0}, \boldsymbol{W}_t), && (3.2.2)\\ &\textit{Initial information:} \quad (\boldsymbol{\theta}_0|D_0) \sim N(\boldsymbol{m}_0, \boldsymbol{C}_0). \end{aligned}$$

The observation equation (3.2.1) describes the relationship between the observed time series y_t and the state vector $\boldsymbol{\theta}_t$ at time t. The $n \times 1$ vector $\boldsymbol{F}_t$ is chosen by the modeller, in order to define the nature of this linear dependence. The observational error term is denoted by ν_t, which is Gaussian distributed with mean zero and variance V_t.

The system equation (3.2.2) (also known as the state or evolution equation) describes how the unobservable state vector $\boldsymbol{\theta}_t$ evolves with time. Considering the two components of this equation in turn, we see that the deterministic part of the evolution is given by the transition from the state $\boldsymbol{\theta}_{t-1}$ to $\boldsymbol{G}_t\boldsymbol{\theta}_{t-1}$. Here, $\boldsymbol{G}_t$ is an $n \times n$ matrix,

defined by the modeller in order to describe the nature of this evolution. The second component of the system equation provides the stochastic part of the evolution. This is given by the $n \times 1$ noise vector $\boldsymbol{\omega}_t$, which defines the multivariate Gaussian distributed state evolution error with mean $\mathbf{0}$ and covariance $\boldsymbol{W}_t$.

As noted by West and Harrison (1997), the error sequences $\{\nu_t\}$ and $\{\boldsymbol{\omega}_t\}$ are assumed to be internally and mutually independent, and are independent of $(\boldsymbol{\theta}_0|D_0)$. This is a natural assumption, since clearly the observational error ν_t is a simple random perturbation in the measurement process; this only influences the immediate observation y_t. By contrast, the state evolution noise $\boldsymbol{\omega}_t$ influences the development of the system into the future. The assumption of independence between these two sources of stochastic behaviour clearly clarifies their separate roles in the model. We note that it is also possible to generalise the model in the case that we have known, non-zero means for either of the noise terms (see West and Harrison (1997) for further details).

Definition 3.2.1 also requires the assumption of a normal distribution for the state vector at time $t = 0$ given the available information at that time, D_0. We assume prior mean $\boldsymbol{m}_0$ and covariance $\boldsymbol{C}_0$. In the case that we are modelling the data as a continuation of a previously observed series, where we have chosen the time origin arbitrarily, one can think of the chosen prior as summarising the information from the past, where the state vector at the time origin $\boldsymbol{\theta}_0$ can be interpreted as the final state vector of the historical data. In the literature, it is often the case that a univariate DLM is written as its defining quadruple of the form,

$$\{\boldsymbol{F}_t, \boldsymbol{G}_t, V_t, \boldsymbol{W}_t\}.$$

When modelling an observed time series as a DLM, we are interested in several relevant distributions. West and Harrison (1997) note that at time t, all known information about a system is summarised by the posterior distribution of the state vector. For simplicity, consider the univariate DLM where all information about the system

at times $t \geq 1$ is provided by the observed time series; that is to say, the system is closed to external information. Assuming some prior information D_0, the available information about the system at time t is then given by $D_t = \{y_t, D_{t-1}\}$. Therefore, we denote the posterior distribution of the state vector at this time by $p(\boldsymbol{\theta}_t|D_t)$. We seek a method that enables this distribution to be determined throughout time. Furthermore, we often wish to estimate the prior distribution of the state vector at time t, given the information available at time $t-1$; for example, in applications which require one-step ahead forecasts of the state variable. Similarly, we may wish to estimate the forecasting distribution of the observed time series at some future time $t+h$, for $h \geq 0$; we shall consider the one-step-ahead case in this thesis, such that $h = 1$. We refer to these three relevant distributions as the posterior, prior and forecasting distributions respectively, and denote them by $p(\boldsymbol{\theta}_t|D_t)$, $p(\boldsymbol{\theta}_t|D_{t-1})$ and $p(y_t|D_{t-1})$.

3.2.1 The Kalman filter

It is expected that our inference on the posterior, prior and forecasting distributions will change throughout time as more data become available; recall, in a system closed to external information at times $t \geq 1$, the available information set D_t is updated with every new observation received. We therefore desire a sequential analysis where we are able to update our inference on the relevant distributions whenever a new observation is recorded.

The Kalman filter is a set of recursive equations which allow the aforementioned relevant distributions to be computed within a simple Bayesian framework. Whenever a new time series observation is made, the Kalman filtering equations are performed in order to compute the prior, posterior and forecasting distributions in closed-form. These are given as follows.

Theorem 3.2.2 (The Kalman Filter). *For the univariate DLM given by Definition*

3.2.1, let

$$(\boldsymbol{\theta}_{t-1}|D_{t-1}) \sim N(\boldsymbol{m}_{t-1}, \boldsymbol{C}_{t-1}),$$

for some mean $\boldsymbol{m}_{t-1}$ *and variance matrix* $\boldsymbol{C}_{t-1}$*. Then the prior, forecasting and posterior distributions are given for each* t *as follows:*

1. *Prior at* t*:*

$$(\boldsymbol{\theta}_t|D_{t-1}) \sim N(\mathbf{a}_t, \mathbf{R}_t), \quad \textit{where} \quad \mathbf{a}_t = \mathbf{G}_t\mathbf{m}_{t-1}, \quad \textit{and} \quad \mathbf{R}_t = \mathbf{G}_t\mathbf{C}_{t-1}\mathbf{G}_t' + \mathbf{W}_t.$$

2. *One-step-ahead forecasting distribution:*

$$(y_t|D_{t-1}) \sim N(f_t, Q_t), \quad \textit{where} \quad f_t = \mathbf{F}_t'\mathbf{a}_t, \quad \textit{and} \quad Q_t = \mathbf{F}_t'\mathbf{R}_t\mathbf{F}_t + V_t.$$

3. *Posterior at* t*:*

$$(\boldsymbol{\theta}_t|D_t) \sim N(\mathbf{m}_t, \mathbf{C}_t), \quad \textit{with} \quad \mathbf{m}_t = \mathbf{a}_t + \mathbf{A}_t e_t \quad \textit{and} \quad \mathbf{C}_t = \mathbf{R}_t - \mathbf{A}_t Q_t \mathbf{A}_t',$$
$$\textit{where} \quad \mathbf{A}_t = \mathbf{R}_t\mathbf{F}_t Q_t^{-1} \quad \textit{and} \quad e_t = y_t - f_t.$$

The closed-form analysis provided by the Kalman filter is one of the key advantages of DLMs compared to other Bayesian approaches. It allows real-time predictive densities and point forecasts to be obtained in a clear and uncomplicated manner. When applied to a sequence of observations, the filtering procedure provides a series of posterior distributions and predictive densities. The proof for the univariate Kalman filter, the multivariate extension and the relevant smoothing equations for a retrospective analysis can be found in West and Harrison (1997).

3.2.2 Dealing with unknown parameters

In order to apply the Kalman filter to a DLM, we require knowledge of the defining quadruple $\{\boldsymbol{F}_t, \boldsymbol{G}_t, V_t, \boldsymbol{W}_t\}$ at all times t. In general, the regression vectors $\boldsymbol{F}_t$ and the evolution matrices $\boldsymbol{G}_t$ are defined by the modeller in order to best describe the behaviour of the observed time series. They are chosen by considering some physical interpretation of the system, and can therefore be defined in a way that takes into account any seasonal, regression or trend components in the model (for more details, see Petris et al. (2009) and West and Harrison (1997)).

On the other hand, it is often the case that the observational variance V_t, and/or the system evolution covariance $\boldsymbol{W}_t$ are unknown. Since the modeller only has access to the observed time series y_t at time t and not the corresponding state variable $\boldsymbol{\theta}_t$, little may be known about the underlying stochastic variation of the system, or indeed the magnitude of the observational errors. Often, the direct quantification of these parameters is difficult, and the resulting estimates are grossly misspecified. In the following, we describe closed-form methods for dealing with unknown parameters which can be suitably integrated into the model combination framework.

Unknown observational variance

West and Harrison (1997) propose dealing with unknown observational variance by introducing a slight modification to the univariate DLM structure. Namely, they consider the defining quadruple now given by

$$\{\boldsymbol{F}_t, \boldsymbol{G}_t, V_t^* \sigma^2, \boldsymbol{W}_t^* \sigma^2\}.$$

We assume that the defining vectors $\boldsymbol{F}_t$ and $\boldsymbol{G}_t$ are known, in addition to the values V_t^* and $\boldsymbol{W}_t^*$. An unknown scale factor σ^2 has been introduced into the model, such that the observational variance is given by $V_t = V_t^* \sigma^2$, where V_t^* is the known constant

multiplier (similarly, the system evolution covariance is given by $\boldsymbol{W}_t = \boldsymbol{W}_t^* \sigma^2$). The introduction of the scale factor provides a scale-free model. In most cases, the known constant multiplier of the scale factor is set to $V_t^* = 1$, in which case the observational variance of the model is given by $V_t = \sigma^2$. We refer to σ^2 as the uncertain variance.

A closed-form filtering procedure for learning the uncertain variance can be developed by working with its reciprocal $\phi = 1/\sigma^2$, known as the observational precision. By assuming a gamma prior on this parameter,

$$(\phi|D_0) \sim G(n_0/2, n_0 s_0/2),$$

West and Harrison (1997) develop a fully conjugate Bayesian analysis, which allows the uncertain variance to be learned as the filtering procedure takes place.

The parameter n_0 describes the degrees of freedom at time $t = 0$, such that the mean of this prior distribution is given by,

$$E[\phi|D_0] = \frac{n_0}{n_0 s_0} = \frac{1}{s_0}.$$

Thus, the parameter s_0 provides a prior point estimate of the observational variance σ^2 at time $t = 0$. Under this prior assumption, and setting $V_t^* = 1$, the corresponding DLM is defined by

$$\begin{aligned}
&\text{Observation equation:} \quad y_t = \boldsymbol{F}_t' \boldsymbol{\theta}_t + \nu_t, \qquad \nu_t \sim N(0, \sigma^2), \\
&\text{System equation:} \qquad \boldsymbol{\theta}_t = \boldsymbol{G}_t \boldsymbol{\theta}_{t-1} + \boldsymbol{\omega}_t, \quad \boldsymbol{\omega}_t \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{W}_t^*), \\
&\text{Initial information:} \quad (\boldsymbol{\theta}_0|D_0, \phi) \sim N(\boldsymbol{m}_0, \sigma^2 \boldsymbol{C}_0^*), \quad (\phi|D_0) \sim G(\tfrac{n_0}{2}, \tfrac{n_0 s_0}{2}).
\end{aligned} \tag{3.2.3}$$

This model can be analysed using a closed-form filtering analysis similar to Theorem 3.2.2, with the addition of an updating step for the distribution of the unknown precision ϕ. In particular, when moving from time $t-1$ to time t, the updates to the

gamma parameters are given by

$$n_t = n_{t-1} + 1 \quad \text{and} \quad s_t = s_{t-1} + \frac{s_{t-1}}{n_t}\left(\frac{e_t^2}{Q_t} - 1\right).$$

Thus, the quantity s_t provides the posterior estimate of $\sigma^2 = 1/\phi$ at time t. In the one-step ahead forecasting equations (part (2) of Theorem 3.2.2), the time $t-1$ estimate for σ^2 is substituted into the expression for the conditional observational variance $V_t = \sigma^2$. In addition, the updating equation for the variance matrix $\boldsymbol{C}_t$ involves a change of scale to reflect the revised estimate of σ^2, that is,

$$\boldsymbol{C}_t = \frac{s_t}{s_{t-1}}(\boldsymbol{R}_t - \boldsymbol{A}_t \boldsymbol{A}_t' Q_t).$$

Finally, we note that in Theorem 3.2.2, the prior and posterior distributions for the state vectors, in addition to the one-step ahead forecasting distribution for the observed time series, are all normally distributed. In the sequential updating analysis of the model described by equation (3.2.3), these distributions are only normal when conditional on the unknown parameter σ^2. The unconditional distributions are instead Student's t-distributions; as an example, the prior for the state vector at time t is now given by $(\boldsymbol{\theta}_t|\mathcal{D}_{t-1}) \sim T_{n_{t-1}}(\boldsymbol{a}_t, \boldsymbol{R}_t)$, where T_n denotes the Student's t-distribution with n degrees of freedom. The updated posterior distribution is then given by $(\boldsymbol{\theta}_t|\mathcal{D}_t) \sim T_{n_t}(\boldsymbol{m}_t, \boldsymbol{C}_t)$, in comparison to the normal distributions defined in parts (a) and (c) of Theorem 3.2.2. A full summary of the modified updating equations for a DLM with unknown and constant observational variance can be found in Chapter 4 of West and Harrison (1997).

The above shows that by introduction of a gamma prior for the unknown precision parameter, a fully closed-form, conjugate analysis can be carried out in the case of unknown and constant observational variance. This enables sequential learning about the unknown parameter within a Bayesian updating framework, where inference on the unknown state vector is updated simultaneously.

Sometimes it is beneficial to consider the case where the uncertain variance is varying stochastically throughout time. An adapted closed-form updating sequence can be derived by introducing a variance discount factor β and a beta-distributed random variable γ_t, which represents random shocks at each time. These parameters are used to define the stochastic evolution of the precision. For brevity, we omit the exact details here and instead refer the interested reader to Chapter 4 of West and Harrison (1997). However, we do note that this analysis was incorporated into the forecast combination code developed for this thesis as a user-specified option.

Unknown system covariance

The appropriate specification of both the structure and magnitude of the system evolution covariance matrix $\boldsymbol{W}_t$ is critically important for successful modelling and forecasting in the DLM framework. The system evolution covariance controls the extent of the stochastic variation in the state equation, and is therefore responsible for determining the stability of the model over time.

In a classical framework, it is possible to estimate $\boldsymbol{W}_t$ using maximum likelihood methods, and then proceed with the filtering analysis using the maximum likelihood estimate. For some models and data sets, this method can be effective; however, such an approach fails to take properly into account the uncertainty about the estimated parameter (since we compute a point estimate). Furthermore, this approach is not particularly efficient in an online analysis; in the case that the optimal value of $\boldsymbol{W}_t$ changes with time, we are required to carry out the maximum likelihood estimation at each time step. We also note that this technique cannot be implemented in the case that we have no historical data.

We can instead consider a Bayesian approach to the problem. Under a Bayesian perspective, the unknown parameters are treated as random quantities; in the context of DLMs, the posterior distribution of interest is considered to be the joint conditional

distribution of the state vector and the unknown covariance $\boldsymbol{W}_t$, given the observed data. Markov chain Monte Carlo (MCMC) methods can be used to approximate the relevant distributions in the case that computations are not analytically tractable. For example, Petris et al. (2009) show how Gibbs sampling algorithms can be used effectively to simultaneously solve the filtering, smoothing and forecasting problems for a DLM with unknown system evolution covariance. However, such MCMC procedures are not ideally suited to recursive inference. Each time a new time series value is observed, the distribution of interest changes. This means that a new MCMC procedure must be performed at each time step, in order to sample from the updated posterior distribution. This is clearly computationally inefficient, particularly in applications that require an online analysis or where the data arrive with high frequency. As highlighted in Section 3.2.1, a key advantage of working with DLMs is the ability to carry out a closed-form, sequential analysis. Therefore, the requirement to run an MCMC simulation each time a new observation is recorded has an adverse effect on the overall benefits of such a model. Hence, we seek a method for performing a recursive, closed-form analysis in the case that system covariance $\boldsymbol{W}_t$ is unknown. To this end, we utilise the method of discounting (see Section 6.3 of West and Harrison (1997), and Prado and West (2010)), which we describe in the following.

In the system equation (3.3.1), it is seen that the covariance matrix $\boldsymbol{W}_t$ leads to an increase in uncertainty about the state vector as we move from time $t-1$ to time t; this can be equivalently interpreted as a loss of information from this transition. To see this, consider the posterior and prior distributions of the state vector at two subsequent time steps. According to Theorem 3.2.2, the posterior distribution of the state vector at time $t-1$ is normally distributed with covariance

$$\text{Cov}(\boldsymbol{\theta}_{t-1}|D_{t-1}) = \boldsymbol{C}_{t-1}.$$

By the evolution equation, this then results in a prior covariance for $\boldsymbol{\theta}_t$ given by,

$$\text{Cov}(\boldsymbol{\theta}_t|D_{t-1}) = \boldsymbol{G}_t\boldsymbol{C}_{t-1}\boldsymbol{G}_t' + \boldsymbol{W}_t.$$

In line with West and Harrison (1997), let $\boldsymbol{P}_t$ denote the first term in this covariance, that is

$$\boldsymbol{P}_t = \boldsymbol{G}_t\boldsymbol{C}_{t-1}\boldsymbol{G}_t' = \text{Cov}(\boldsymbol{G}_t\boldsymbol{\theta}_{t-1}|D_{t-1}).$$

Therefore, $\boldsymbol{P}_t$ can be interpreted as the appropriate prior covariance for a system with no stochastic evolution (that is, the DLM given by $\{\boldsymbol{F}_t, \boldsymbol{G}_t, V_t, \boldsymbol{0}\}$). By considering the covariance in this way, it can be seen that the role of $\boldsymbol{W}_t$ in the Kalman filtering equations is to increase the uncertainty from the ideal $\boldsymbol{P}_t$ to the realistic $\boldsymbol{R}_t = \boldsymbol{P}_t + \boldsymbol{W}_t$. To exploit this interpretation of $\boldsymbol{W}_t$, and to avoid the issues arising via MCMC methods, a popular choice in the DLM literature is to introduce a discount factor δ, which satisfies the condition $0 < \delta \leq 1$. Using this, we can write the implied covariance as a fraction of the ideal covariance in the case that no stochastic variation is present,

$$\text{Cov}(\boldsymbol{\theta}_t|D_{t-1}) = \boldsymbol{R}_t = \frac{1}{\delta}\boldsymbol{P}_t.$$

By considering that $\boldsymbol{R}_t = \boldsymbol{P}_t + \boldsymbol{W}_t$, we can therefore write

$$\boldsymbol{W}_t = \frac{1-\delta}{\delta}\boldsymbol{P}_t = \frac{1-\delta}{\delta}\boldsymbol{G}_t'\boldsymbol{C}_{t-1}\boldsymbol{G}_t. \tag{3.2.4}$$

Using this identification of the system evolution covariance, one can determine the whole sequence $\{\boldsymbol{W}_t\}$ from an initial $\boldsymbol{C}_0$ and discount factor δ. Hence, by inserting equation (3.2.4) into the Kalman filter, it is possible to carry out a sequential, closed-form analysis for a DLM with unknown $\boldsymbol{W}_t$.

Appropriate choice of a discount factor presents an interesting problem. Essentially,

the role of the discount factor is to control the local durability of the model (West and Harrison (1997)). A discount factor equal to 1 implies a zero system evolution covariance; in this case, the system model is globally reliable. Conversely, as $\delta \to 0$ and therefore $\boldsymbol{W}_t \to \infty$, the system model becomes completely unreliable. In general, application of a low value of discount factor indicates the need for a superior model.

The suggested optimal choice of discount factor varies in the literature. West and Harrison (1997) recommend a discount factor in the range $[0.9, 0.99]$ for routine analysis, as do Prado and West (2010). On the other hand, others consider $\delta \in \{0.95, 0.99\}$; for example, Koop and Korobilis (2012). Typically in the literature, the discount factor is simply set to a constant user-specified value, which is chosen a priori. Alternative and more complex methods for discount factor selection are discussed in Chapter 4.

3.3 Complete DLM-based forecast combination model framework

In this section we provide a user's guide for the implementation of a DLM-based point forecast combination framework, detailing step-by-step instructions for the practitioner. This framework allows forecast combination weights to be estimated sequentially as more data become available, in a computationally efficient manner. Combination weights are unconstrained and can take negative values (we demonstrated optimal weights can be negative in the $N = 2$ forecaster case in Chapter 2). To the best of our knowledge, a complete description of this methodology has not been provided in the literature before.

In order to define the forecast combination problem as a DLM, we assume the unconstrained regression framework with no intercept term suggested by Granger and Ramanathan (1984), such that the observed series y_t is modelled as a linear combination

of N individual forecasters,

$$y_t = w_{1,t}f_{1,t} + w_{2,t}f_{2,t} + \cdots + w_{N,t}f_{N,t} + \nu_t.$$

We note that the extension to include an intercept term is straightforward, and we advise this if bias is present in the individual forecasters. In order to avoid the requirement that the individual forecasters must be stationary (highlighted by De Menezes et al. (2000)), and to accommodate for changing forecaster quality throughout time, we let the regression coefficients evolve according to a random walk model:

$$\mathbf{w}_t = \mathbf{w}_{t-1} + \boldsymbol{\omega}_t,$$

where $\mathbf{w}_t = [w_{1,t}\, w_{2,t}\, \ldots\, w_{N,t}]'$. We emphasise here that we do not enforce the constraint that weights sum to one in our method, even though this may be desirable in the case that forecasters are known to be unbiased. We note that such weights could be achieved through an extension to constrained Kalman filtering, which is discussed briefly in Chapter 6.

Let us assume that the observational noise term ν_t follows a normal distribution with mean zero and variance V_t. Similarly, let us assume that the random walk noise term follows a multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix $\mathbf{W}_t$. Denote the vector of N expert forecasts at time t by $\mathbf{f}_t = [f_{1,t}\, f_{2,t}\, \ldots\, f_{N,t}]'$. If we also assume that, at time $t = 0$, there is a normal prior distribution on the regression parameters, with mean $\mathbf{m}_0$ and covariance $\mathbf{C}_0$, then we can define a DLM for the forecast combination problem with state vector $\mathbf{w}_t$ as follows,

$$\begin{aligned}
&\text{Observation equation:} && y_t = \mathbf{f}_t'\mathbf{w}_t + \nu_t, && \nu_t \sim N(0, V_t),\\
&\text{System equation:} && \mathbf{w}_t = \mathbf{w}_{t-1} + \boldsymbol{\omega}_t, && \boldsymbol{\omega}_t \sim N(\mathbf{0}, \boldsymbol{W}_t),\\
&\text{Initial information:} && (\mathbf{w}_0|D_0) \sim N(\boldsymbol{m}_0, \boldsymbol{C}_0).
\end{aligned}$$

By defining the forecast combination problem as a DLM in this way, we can compute the one-step-ahead forecasting distribution for the observed series y_t at each point in time through the application of the Kalman filtering equations. Moreover, application of the Kalman filter will enable prior and posterior estimates on the weight vector to be obtained throughout time. This is a DLM with defining quadruple:

$$\{\mathbf{f}_t, \mathbb{I}_N, V_t, \boldsymbol{W}_t\}.$$

Of course, the observational variance V_t and the system covariance $\mathbf{W}_t$ will be unknown in practice. We can deal with this by enforcing a gamma prior on the precision and including a discount factor δ for the system covariance matrix, as detailed in Section 3.2.2.

Thus, the complete DLM for the point forecast combination problem is given by,

$$\begin{aligned}
&\text{Observation equation:} \quad y_t = \mathbf{f}_t'\mathbf{w}_t + \nu_t, \qquad \nu_t \sim N(0, \sigma^2), \\
&\text{System equation:} \qquad\quad \mathbf{w}_t = \mathbf{w}_{t-1} + \boldsymbol{\omega}_t, \quad \boldsymbol{\omega}_t \sim N(\mathbf{0}, \sigma^2 \boldsymbol{W}_t^*), \qquad (3.3.1) \\
&\text{Initial information:} \quad (\mathbf{w}_0|D_0, \sigma^2) \sim N(\boldsymbol{m}_0, \sigma^2 \boldsymbol{C}_0^*), \quad (\phi|D_0) \sim G(\tfrac{n_0}{2}, \tfrac{n_0 s_0}{2}),
\end{aligned}$$

where the precision is denoted by $\phi = 1/\sigma^2$. Step-by-step instructions on the implementation of this forecast combination procedure are given in the following.

1. **Define prior parameters $\mathbf{m}_0$, $\mathbf{C}_0^*$, n_0, s_0, and discount factor δ.**

   (a) Simple averaging often displays favourable empirical performance, therefore a suitable prior for the weight vector is given by equal weights,

   $$\mathbf{m}_0 = [1/N \ 1/N \ ... \ 1/N].$$

   (b) An uninformative prior for the state evolution covariance reflects uncertainty in the optimal choice of combination weights and allows the filtering proce-

dure to converge to optimal weights quickly, therefore a suitable choice is given by the $N \times N$ matrix,

$$\mathbb{C}_0^* = \begin{bmatrix} 1 \times 10^7 & 0 & \dots & 0 \\ 0 & 1 \times 10^7 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \times 10^7 \end{bmatrix}.$$

(c) As is the convention in the literature, we recommend setting,

$$n_0 = 1, \quad s_0 = 1.$$

(d) In general, we recommend selecting a discount factor $\delta \in [0.9, 1)$. Individual guidance is provided for the simulation studies and empirical examples in this chapter, and more complex discount factor selection methods are discussed in Chapter 4.

2. **For times** $t = 1, \dots, T$**:**

(a) Compute the prior distribution for the weight vector,

$$(\mathbf{w}_t | D_{t-1}) \sim T_{n_{t-1}}(\mathbf{m}_{t-1}, \frac{s_{t-1}}{\delta}\mathbf{C}_{t-1}^*).$$

(b) Compute the one-step-ahead forecasting distribution,

$$(y_t | D_{t-1}) \sim T_{n_{t-1}}(\mathbf{f}_t'\mathbf{m}_{t-1}, s_{t-1}Q_t^*),$$

where

$$Q_t^* = \frac{1}{\delta}(\mathbf{f}_t'\mathbf{C}_{t-1}^*\mathbf{f}_t + \delta).$$

(c) Observe y_t and compute the forecast error,

$$e_t = y_t - \mathbf{f}_t' \mathbf{m}_{t-1}.$$

(d) Compute the posterior distributions,

$$(\boldsymbol{\theta}_t | D_t) \sim T_{n_t}(\mathbf{m}_t, s_t \mathbf{C}_t^*),$$
$$(\phi | D_t) \sim G(n_t/2, n_t s_t/2),$$

where the parameters for the precision distribution are updated recursively according to

$$n_t = n_{t-1} + 1, \quad s_t = \frac{n_t - 1}{n_t} s_{t-1} + \frac{e_t^2}{n_t Q_t^*},$$

and the parameters for the state vector distribution are given by the recursive equations:

$$\mathbf{m}_t = \mathbf{m}_{t-1} + \mathbf{A}_t e_t, \quad \mathbf{C}_t^* = \frac{1}{\delta}(\mathbb{I}_N - \mathbf{A}_t \mathbf{f}_t') \mathbf{C}_{t-1}^*,$$

with

$$\mathbf{A}_t = \frac{1}{\delta Q_t^*} \mathbf{C}_{t-1}^* \mathbf{f}_t = \frac{1}{\mathbf{f}_t' \mathbf{C}_{t-1}^* \mathbf{f}_t + \delta} \mathbf{C}_{t-1}^* \mathbf{f}_t.$$

The above formulation is given for a constant choice of discount factor δ. However, a simple extension can be made to consider δ_t at each time. With the equations in this format, it is straightforward to integrate the combination procedure with methods for adaptive discount factor selection, such as those of Irie et al. (2022) or Yusupova et al. (2023) (discussed in Chapter 4).

3.4 Simulation study

In this section, we undertake a simulation study to investigate the performance of our forecast combination procedure in a variety of settings, working with both independent and correlated forecasters, in static and dynamic environments. The purpose of this is both as a 'proof of concept' for our forecast combination method, and to assess its forecasting performance in comparison with established benchmarks. We demonstrate how our method can capture dynamics in the combination weights in an online and sequential framework, unlike other forecast combination methods.

In addition to our DLM-based method, we evaluate the performance of simple averaging, the optimal covariance weights of Newbold and Granger (1974), the regression-based weights of Granger and Ramanathan (1984), and taking the recent best forecaster. In dynamic scenarios, we also consider 'rolling' adaptations of the Granger and Ramanathan (1984) regression-based weights and the Newbold and Granger (1974) covariance weights. These are evaluated by implementing the estimation methods on the previous k observations at each time step (this was described in Section 2.4.1 for the optimal covariance weights). We compare the mean square error (MSE),

$$\text{MSE} = \frac{1}{n}\sum_{t=1}^{T}(y_t - \hat{y}_t)^2,$$

the mean absolute error (MAE),

$$\text{MAE} = \frac{1}{T}\sum_{t=1}^{T}|y_t - \hat{y}_t|,$$

and the symmetric mean absolute percentage error (SMAPE),

$$\text{SMAPE} = \frac{100}{T}\sum_{t=1}^{T}\frac{|y_t - \hat{y}_t|}{(|y_t| + |\hat{y}_t|)/2},$$

for each forecasting method tested.

The simulations in this section were inspired by financial applications. An 'underlying' time series of interest of length $T = 500$ was simulated according to an AR(1) model with mean level μ,

$$\tilde{y}_t = \phi\tilde{y}_{t-1} + (1-\phi)\mu + \epsilon_t, \quad \epsilon_t \overset{iid}{\sim} N(0, \sigma_\epsilon^2).$$

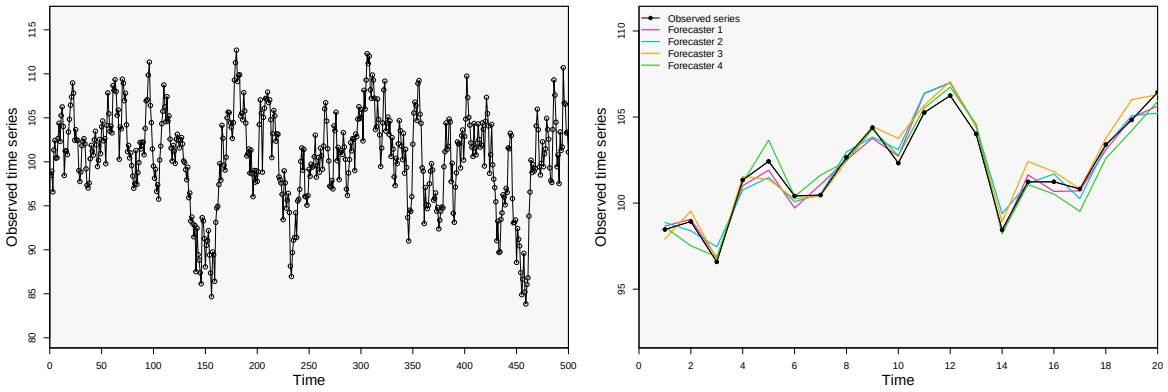
The observed series of interest y_t was then simulated by taking this underlying series and adding some 'local market effects',

$$y_t = \tilde{y}_t + \zeta_t, \quad \zeta_t \overset{iid}{\sim} N(0, \sigma_\zeta^2).$$

A set of $N = 4$ expert forecasts were simulated from a multivariate normal distribution, as though 'unaware' of the local market effects,

$$\mathbf{f}_t \sim N(\tilde{y}_t\boldsymbol{\iota}, \boldsymbol{\Sigma}_t),$$

where $\boldsymbol{\iota}$ is a 4-dimensional vector of 1s. In the following simulations, parameters were set to $\sigma_\epsilon^2 = 2.5^2$, $\sigma_\zeta^2 = 0.4^2$, $\phi = 0.9$, $\mu = 100$. These parameters were chosen based on a financial data set. Different simulation settings were defined through different choices of covariance matrix $\boldsymbol{\Sigma}_t$. In order to ensure robustness of the method, 100 replications of expert forecasts were simulated from the underlying time series $\tilde{y}_t$ for each choice of $\boldsymbol{\Sigma}_t$. The simulated time series y_t is shown in Figure 3.4.1a, and the first 20 observations along with a set of forecaster series (simulated from covariance matrix Σ_1 defined in Section 3.4.2) are shown in Figure 3.4.1b.



(a) Simulated time series y_t.

(b) The first 20 observations of y_t, and four forecasters.

Figure 3.4.1: Time series simulated according to an AR(1) plus noise model.

3.4.1 Oracle weight derivation

To begin, let us consider the estimation of the combination weights themselves, rather than the resulting combined forecast. While we require a combination method that provides accurate forecasts, we also desire interpretability. That is to say, we do not only wish to know the combined forecast for the variable of interest, but also which forecaster is the best at each point in time and likewise, which is the worst. For example, consider combining two forecasters of equal quality. By coincidence, assume that these two forecasters have produced identical forecasts at time t. Clearly, several different choices of combination weights would lead to the same combined forecast; however, only one choice would provide an accurate interpretation of the individual forecaster quality. For example, we could assign each a combination weight of $1/2$, or alternatively one forecaster could be assigned a weight of 1 and the other a weight of 0. In both cases, the combined forecast would be the same, and the resulting forecast error would be identical. If we were only interested in forecasting performance, either of these combinations would be sufficient. However, in the latter case we may incorrectly interpret these weights to mean that the forecaster with the weight of 1 was far superior than that with the zero weight. Only the combination where each forecaster is assigned

a weight of 1/2 provides the correct interpretation of the forecasters: namely, they are of equal quality.

With this in mind, we desired a method for computing what the optimal weight combination would be if we had access to the observed series y_t at each time. In Section 2.2, we provided the derivation of the covariance weights of Newbold and Granger (1974), which are optimal for the case of N unbiased forecasts with weights constrained to lie on the unit simplex. Similarly, we showed how Timmermann (2006) derived the generalised optimal weights for N forecasters, under the assumption of joint Gaussianity for the forecaster series and the observed y_t. In this section we follow a similar approach to the derivation of Timmermann (2006), but instead of imposing distributional assumptions on the observed data y_t, we treat this as a known value at each time point rather than a random variable. Thus, the assumption of joint multivariate Gaussianity is no longer required, and the derived optimal weights can be treated as the 'oracle' weights, in the sense that they are derived given knowledge of the observed data. This is done as follows.

Let us assume that the N-vector of forecasters is distributed with expectation and covariance matrix given by,

$$E\{\mathbf{f}_t\} = \boldsymbol{\mu}_t \quad \text{and} \quad \text{Cov}(\mathbf{f}_t) = \boldsymbol{\Sigma}_t.$$

We note that these moments are allowed to vary with time, and no other assumptions are made on the distribution of $\mathbf{f}_t$. We can find optimal weights by minimising the expected squared forecast error of the combination:

$$\begin{aligned} E\{(y_t - \mathbf{w}_t'\mathbf{f}_t)^2\} &= E\{y_t - \mathbf{w}_t'\mathbf{f}_t\}^2 + \text{Var}(y_t - \mathbf{w}_t'\mathbf{f}_t) \\ &= (y_t - \mathbf{w}_t'\boldsymbol{\mu}_t)^2 + \mathbf{w}_t'\boldsymbol{\Sigma}_t\mathbf{w}_t. \end{aligned}$$

This is done by differentiating with respect to the weight vector,

$$\begin{aligned}\frac{\partial E\{(y_t - \mathbf{w}_t'\mathbf{f}_t)^2\}}{\partial \mathbf{w}_t} &= -2(y_t - \mathbf{w}_t'\boldsymbol{\mu}_t)\boldsymbol{\mu}_t + (\boldsymbol{\Sigma}_t + \boldsymbol{\Sigma}_t')\mathbf{w}_t', \\ &= -2(y_t - \mathbf{w}_t'\boldsymbol{\mu}_t)\boldsymbol{\mu}_t + 2\boldsymbol{\Sigma}_t\mathbf{w}_t,\end{aligned}$$

where we have used the fact that the covariance matrix is symmetric. The optimal weight vector can then be found by setting this to zero and rearranging,

$$\begin{aligned}&- y_t\boldsymbol{\mu}_t + (\mathbf{w}_t'\boldsymbol{\mu}_t)\boldsymbol{\mu}_t + \boldsymbol{\Sigma}_t\mathbf{w}_t = 0 \\ \implies\ &\boldsymbol{\mu}_t(\boldsymbol{\mu}_t'\mathbf{w}_t) + \boldsymbol{\Sigma}_t\mathbf{w}_t = y_t\boldsymbol{\mu}_t \\ \implies\ &\mathbf{w}_t = (\boldsymbol{\mu}_t\boldsymbol{\mu}_t' + \boldsymbol{\Sigma}_t)^{-1}y_t\boldsymbol{\mu}_t, \qquad (3.4.1)\end{aligned}$$

where we have assumed that $(\boldsymbol{\mu}_t\boldsymbol{\mu}_t' + \boldsymbol{\Sigma}_t)$ is invertible. We shall refer to the weights given by equation (3.4.1) as the oracle weights, since they are evaluated with knowledge of y_t. Since we have not treated y_t as a random variable in our derivation, the oracle weights can be calculated for any observed series y_t, provided we have knowledge of the forecast vector mean and covariance matrix. Of course, such weights are meaningless in a practical forecast combination problem, since one would not have knowledge of y_t in advance of choosing the weights (if one did, then the entire forecast combination would be redundant). However, it is particularly useful in our simulation studies, as we can compare the weights estimated from our method with the oracle weights. Moreover, we can also use the oracle weights to compute a combined forecast at each time. The forecasting performance of our method and other forecast combination methods can then be compared with the forecasting performance of the oracle method, to provide insight on how well methods are performing.

We note that even in the case of unbiased forecasters with constant covariance matrix, the oracle weights will vary slightly throughout time, unless the observed series

y_t is itself constant. This differs from the optimal weights derived in Section 2.2, where constant forecaster quality implies constant optimal weights. However, the variations in the oracle weights are very small for typical signal to noise ratios. For example, consider the case of two forecasters of equal quality, which are unbiased and uncorrelated:

$$E\left\{ \begin{bmatrix} f_{1,t} \\ f_{2,t} \end{bmatrix} \right\} = \begin{bmatrix} y_t \\ y_t \end{bmatrix}, \quad \text{Cov}(f_{1,t}, f_{2,t}) = \begin{bmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{bmatrix}.$$

The oracle weights can then be written as:

$$\mathbf{w}_t = \frac{1}{2y_t^2 + \sigma^2} \begin{bmatrix} y_t^2 \\ y_t^2 \end{bmatrix} = \frac{1}{2 + \sigma^2/y_t^2} \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

Note that if $y_t^2 \gg \sigma^2$, then we have $\sigma^2/y_t^2 \approx 0$, and hence the weights are approximately $1/2$, as we would obtain using the methods in Section 2.2. We highlight that the oracle weights in this case are upper bounded by $1/2$, and will therefore sum to less than one. The forecast is therefore not the best linear unbiased estimator, but rather the best linear estimator that minimised the MSE.

3.4.2 Constant forecaster quality

The case of constant forecaster quality throughout time can be simulated by selecting a constant covariance matrix for the forecasters $\mathbf{\Sigma}_t = \mathbf{\Sigma}$. Here we consider the case of independent forecasters and correlated forecasters, described by the respective covariance

matrices

$$\boldsymbol{\Sigma}_1 = \begin{bmatrix} \sigma_1^2 & 0 & 0 & 0 \\ 0 & \sigma_2^2 & 0 & 0 \\ 0 & 0 & \sigma_3^2 & 0 \\ 0 & 0 & 0 & \sigma_4^2 \end{bmatrix}, \quad \text{and} \quad \boldsymbol{\Sigma}_2 = \begin{bmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 & 0 & 0 \\ \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 & 0 & 0 \\ 0 & 0 & \sigma_3^2 & \rho_{34}\sigma_3\sigma_4 \\ 0 & 0 & \rho_{34}\sigma_3\sigma_4 & \sigma_4^2 \end{bmatrix}.$$

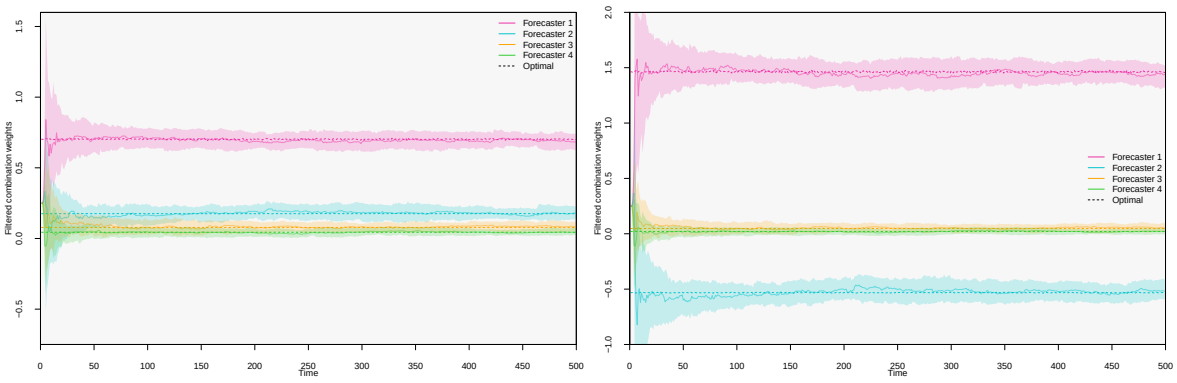
Covariance matrix $\boldsymbol{\Sigma}_1$ corresponds to the case that all forecasters are uncorrelated, meanwhile covariance matrix $\boldsymbol{\Sigma}_2$ induces correlations between Forecasters 1 and 2, and Forecasters 3 and 4. The variances were set to $\sigma_1^2 = 0.2$, $\sigma_2^2 = 0.4$, $\sigma_3^2 = 0.6$, $\sigma_4^2 = 0.8$, and the correlation coefficients were set to $\rho_{12} = 0.9$ and $\rho_{34} = 0.3$.

In order to apply our combination method, a discount factor of $\delta = 0.99$ was chosen. This choice lies within the range of values suggested by West and Harrison (1997) and also Koop and Korobilis (2012). In actuality, since the forecasters were simulated from a constant covariance matrix $\boldsymbol{\Sigma}$, we acknowledge that perhaps an even higher value of discount factor would be more suitable to the problem, since the oracle weights will be almost constant with time.

Estimated weights

The forecast combination procedure was carried out on each of the 100 replications, for the two choices of covariance matrix. For each replication, the estimated weight for each forecaster was computed at time t by taking the mean of the conditional posterior distribution $(\boldsymbol{\theta}_t|D_t, \sigma^2)$, and the corresponding one-step-ahead combined forecasts were obtained by taking the mean of the conditional forecasting distribution $(y_t|D_{t-1}, \sigma^2)$. The median estimated weights over the 100 replications and the interquartile ranges are plotted in Figure 3.4.2a and Figure 3.4.2b for the independent and correlated cases respectively. The coloured dashed lines show the oracle weights for each forecaster.

Both figures show initially wide interquartile ranges for the estimated weights. This



(a) Covariance matrix $\mathbf{\Sigma}_1$. (b) Covariance matrix $\mathbf{\Sigma}_2$.

Figure 3.4.2: Estimated combination weights for unbiased forecasters with constant covariance matrices. Optimal oracle weights shown by the dashed lines, median estimated across 100 repetitions shown by solid coloured lines, and interquartile ranges shown by corresponding shaded regions.

is a consequence of of the choice of prior for $\mathbf{C}_0^*$. To begin, there is large uncertainty on the weights, and consequently the filtered weights vary greatly between replications. As more data is observed, the uncertainty on the weights decreases and the variability between replications decreases, leading to narrower interquartile ranges. We see that it generally takes about 50 observations for the interquartile range to converge to a steady width.

The median filtered weights are close to the oracle weights in both cases. In the uncorrelated case (Figure 3.4.2a), all median weights are between 0 and 1. This is not a requirement of our method (unlike the optimal weights derived by Newbold and Granger (1974)). This can be seen by inspection of the weights in the correlated case (Figure 3.4.2b), where the weight associated with Forecaster 1 exceeds unity, and Forecaster 2 is assigned a negative weight. We refer back to the discussion in Chapter 2, and expect that this is a consequence of the high correlation between these two forecasters.

Forecasting performance

We compared the forecasting performance of our combination procedure with other popular methods in the literature, namely: simple averaging, regression-based weight

estimation (Granger and Ramanathan (1984)), and optimal covariance weights derived using an estimated covariance matrix (Newbold and Granger (1974)). Details of these methods can be found in Chapter 2. Since the simulated forecasters were unbiased, the regression equation without the intercept was applied (regression (ii) in Section 2.3). We also considered the forecast provided by the forecaster who had the smallest forecast error at the previous time point, since this is a method used in financial applications. We shall refer to this as the 'recent best' method. The combined forecasts achieved from applying the oracle weights were also computed, in order to provide a benchmark for the different combination methods.

The optimal covariance weights of Newbold and Granger (1974) require the covariance matrix to be estimated; similarly, the regression-based weights of Granger and Ramanathan (1984) require a linear regression to be carried out. Since we are interested in an online forecast combination method, we defined a training set of the first k observations to allow these quantities to be learned. Forecast performance of the different methods was then compared on the remaining $T - k$ observations. We considered two different sizes of training set, $k = 20$ and $k = 50$. The MSE, MAE and SMAPE were computed for each of the 100 replications, and the median values of these error metrics computed across the replications are reported in Table 3.4.1 (uncorrelated forecaster case) and Table 3.4.2 (correlated forecaster case) for the two different test sets.

For the uncorrelated forecasters, our method provided forecasts with lower median MSE, MAE and SMAPE across the replications for both training sets $k = 20$ and $k = 50$. We note that for the $k = 50$ case, the optimal covariance weights computed with an estimated covariance matrix, and the regression-based weights also performed well. The benefits of our DLM-based method become more apparent in the case $k = 20$, reflecting the ability to adapt quickly to new information as it becomes available. The simple approach of taking the forecast provided by the forecaster whom had the

smallest forecast error at the previous time point demonstrated the worst forecasting performance by all three error metrics, which is unsurprising given the benefits of forecast combination highlighted in Chapter 2. Similar results were obtained in the case of correlated forecasters. In all situations tested, our method performed best, with more gains achieved when a smaller training set was used.

Training	Combination Method	MSE	MAE	SMAPE
$k = 50$	Oracle	0.02799	0.13346	0.133%
	Our method	0.21778	0.37493	0.374%
	Optimal covariance weights	0.21973	0.37621	0.375%
	Regression-based weights	0.21988	0.37725	0.376%
	Simple average	0.25905	0.40598	0.405%
	Recent best	0.42477	0.50939	0.508%
$k = 20$	Oracle	0.02815	0.13344	0.133%
	Our method	0.21508	0.37241	0.372%
	Optimal covariance weights	0.23316	0.38581	0.385%
	Regression-based weights	0.24104	0.39264	0.392%
	Simple average	0.25396	0.40305	0.402%
	Recent best	0.42035	0.50470	0.504%

Table 3.4.1: Forecasting performance of five different forecast combination methods, in addition to the oracle combination weights (light grey row), on simulated uncorrelated forecasters. Values are provided by taking the median error metrics across 100 repetitions. Method(s) with the lowest MSE, MAE and SMAPE highlighted in green.

3.4.3 Gradually changing forecaster quality

Of course, a key advantage of our forecast combination method is that it is online and dynamic, allowing the combination weights to change over time in response to observed data. Dynamic quality of forecasters can be represented by allowing the covariance matrix to vary with time. Once again, two choices of covariance matrix were considered,

Training	Combination Method	MSE	MAE	SMAPE
$k = 50$	Oracle	0.020139	0.11272	0.112%
	Our method	0.20926	0.36743	0.367%
	Optimal covariance weights	0.21158	0.37063	0.370%
	Regression-based weights	0.21227	0.37102	0.370%
	Simple average	0.28538	0.42629	0.425%
	Recent best	0.44313	0.51727	0.516%
$k = 20$	Oracle	0.02023	0.11336	0.113%
	Our method	0.20715	0.36521	0.364%
	Optimal covariance weights	0.22429	0.37920	0.378%
	Regression-based weights	0.23226	0.38441	0.383%
	Simple average	0.28062	0.42393	0.423%
	Recent best	0.43506	0.51350	0.512%

Table 3.4.2: Forecasting performance of five different forecast combination methods, in addition to the oracle combination weights (light grey row), on simulated correlated forecasters. Values are provided by taking the median error metrics across 100 repetitions. Method(s) with the lowest MSE, MAE and SMAPE highlighted in green.

namely,

$$\mathbf{\Sigma}_{1,t} = \begin{bmatrix} \sigma_1^2 & 0 & 0 & 0 \\ 0 & \sigma_{2,t}^2 & 0 & 0 \\ 0 & 0 & \sigma_3^2 & 0 \\ 0 & 0 & 0 & \sigma_4^2 \end{bmatrix}, \quad \text{and} \quad \mathbf{\Sigma}_{2,t} = \begin{bmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_{2,t} & 0 & 0 \\ \rho_{12}\sigma_1\sigma_{2,t} & \sigma_{2,t}^2 & 0 & 0 \\ 0 & 0 & \sigma_3^2 & \rho_{34}\sigma_3\sigma_4 \\ 0 & 0 & \rho_{34}\sigma_3\sigma_4 & \sigma_4^2 \end{bmatrix}.$$

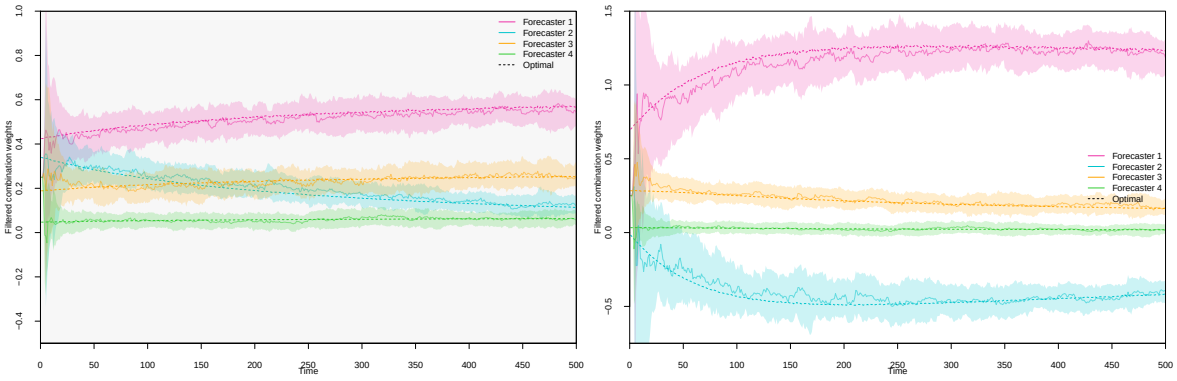
Covariance matrix $\mathbf{\Sigma}_{1,t}$ corresponds to the case that all forecasters are uncorrelated, and the variance of Forecaster 2 is changing with time. Covariance matrix $\mathbf{\Sigma}_{2,t}$ induces correlations between Forecasters 1 and 2 ($\rho_{12} = 0.9$) and Forecasters 3 and 4 ($\rho_{34} = 0.3$). The variance of Forecaster 2 was simulated such that it increased linearly from 0.2 to 0.8, and the other variances were set to $\sigma_1^2 = 0.4$, $\sigma_3^2 = 0.6$, $\sigma_4^2 = 1.2$.

Estimated weights

The combination DLM was defined using the same prior parameter choices as in the constant covariance matrix case, apart from the discount factor which was set to $\delta =$

0.97. This value of discount factor was chosen since we are expecting the optimal weights to change over time in accordance with the changing variance of Forecaster 2, and therefore a lower choice of discount factor is more suitable to accommodate these dynamics. This discount factor was not tuned to the problem, and it is likely that other values of discount factor would improve upon the forecasting performance detailed in this section.

Once again, the filtering procedure was carried out on each of the 100 forecaster vectors, for the two choices of covariance matrix. The median estimated weights over the 100 replications are plotted in Figure 3.4.3a and Figure 3.4.3b, for the uncorrelated and correlated forecasters, respectively. Interquartile ranges are shown by the shaded regions, and the oracle weights are shown by the dashed lines.



(a) Covariance matrix $\mathbf{\Sigma}_{1,t}$. (b) Covariance matrix $\mathbf{\Sigma}_{2,t}$.

Figure 3.4.3: Estimated combination weights for unbiased forecasters with time-varying covariance matrices. Optimal oracle weights shown by the dashed lines, median estimated across 100 repetitions shown by solid coloured lines, and interquartile ranges shown by corresponding shaded regions.

Changing the variance of Forecaster 2 over time clearly induced dynamics in the weights. In the uncorrelated case, the changes in the weights are gradual and relatively small, meaning that our choice of discount factor enables the estimated weights to 'track' the oracle weights well. On the other hand, the changing variance of Forecaster 2 in the correlated case has a more significant impact on the induced dynamics in the oracle weights around times $t \in [0, 150]$. The estimated weights from our method track

the oracle weights well for Forecasters 3 and 4 in this time period; however, the median estimated weights are not as close to the oracle weights for Forecasters 1 and 2. Here, we see that although the oracle weights lie within the relevant interquartile ranges of the estimated weights, the changes in the median weight appear 'too slow' to adequately track the oracle for the first part of the time series. It is expected that a lower choice of discount factor for the first 150 observations would improve upon this. This motivates the concept of time varying discount factors discussed in Chapter 4.

Forecasting performance

Once again, we compared the forecasting performance of our method with that of simple averaging, regression-based weights, optimal covariance weights and the recent best forecaster method. In order to accommodate the changing oracle weights, we also considered 'rolling' versions of the optimal covariance weights of Newbold and Granger (1974), and the regression-based weights of Granger and Ramanathan (1984), where the previous k observations were used to estimate the parameters at each time point. The former corresponds to the time-varying weights proposed by Granger and Newbold (1977), with covariance matrix given by equation (2.4.1).

The median MSE, MAE and SMAPE across the 100 replications, for the uncorrelated case, are shown in Table 3.4.3. For the $k = 50$ case, the lowest MSE was achieved by our method; the lowest MAE and SMAPE was achieved by the rolling covariance weights. The key difference between the two methods is that the DLM-based method enables sequential updating, whereas the rolling covariance weights requires the covariance matrix to be re-estimated at every time point. This is computationally inefficient, and would take considerable time if the number of forecasters N was to be increased. We also note that the value of discount factor chosen in our method was not tuned to the data, and therefore we expect that forecasting performance could be improved through more dedicated tuning of this parameter. Despite the lack of tuning,

our method displayed the lowest error metrics in the $k = 20$ case.

Training	Combination Method	MSE	MAE	SMAPE
$k = 50$	Oracle	0.08426	0.23141	0.231%
	Our method	0.28654	0.43010	0.430%
	Optimal covariance weights	0.29070	0.43264	0.432%
	Regression-based weights	0.29167	0.43421	0.434%
	Simple average	0.33777	0.46393	0.463%
	Recent best	0.66273	0.62065	0.622%
	Optimal covariance weights (rolling)	0.28666	0.42904	0.429%
	Regression-based weights (rolling)	0.29173	0.43214	0.432%
$k = 20$	Oracle	0.08363	0.23093	0.230%
	Our method	0.28327	0.42615	0.425%
	Optimal covariance weights	0.31077	0.44723	0.446%
	Regression-based weights	0.32329	0.45705	0.456%
	Simple average	0.33178	0.46071	0.459%
	Recent best	0.65136	0.61665	0.616%
	Optimal covariance weights (rolling)	0.31963	0.44971	0.449%
	Regression-based weights (rolling)	0.33531	0.46221	0.461%

Table 3.4.3: Forecasting performance of seven different forecast combination methods, in addition to the oracle combination weights (light grey row), on simulated uncorrelated forecasters. Changing forecaster quality. Values are provided by taking the median error metrics across 100 repetitions. Method(s) with the lowest MSE, MAE and SMAPE highlighted in green.

Similar behaviour was observed in the case of correlated forecasters (see Table 3.4.4). For the $k = 50$ case, the rolling covariance weights demonstrated the best performance for all three error metrics tested, with our method displaying an equal SMAPE. In the $k = 20$ case, our method displayed superior performance for all three error metrics.

3.5 Empirical analysis

Two empirical applications are considered in this section, in order to demonstrate the forecasting performance of our method in practice. Of course, issues can arise when working with real data, such as missingness in the observation series and also the forecasting series. The DLM filtering recursions deal with missing observations easily

Training	Combination Method	MSE	MAE	SMAPE
$k = 50$	Oracle	0.07872	0.22275	0.223%
	Our method	0.28112	0.42429	0.424%
	Optimal covariance weights	0.28636	0.42836	0.428%
	Regression-based weights	0.28940	0.42912	0.429%
	Simple average	0.39563	0.50300	0.502%
	Recent best	0.67302	0.62582	0.625%
	Optimal covariance weights (rolling)	0.28028	0.42426	0.424%
	Regression-based weights (rolling)	0.28621	0.42683	0.426%
$k = 20$	Oracle	0.07991	0.22584	0.225%
	Our method	0.28130	0.42439	0.423%
	Optimal covariance weights	0.33809	0.46206	0.461%
	Regression-based weights	0.34691	0.47138	0.470%
	Simple average	0.39042	0.49893	0.499%
	Recent best	0.66758	0.62348	0.622%
	Optimal covariance weights (rolling)	0.31407	0.44847	0.448%
	Regression-based weights (rolling)	0.33123	0.46000	0.459%

Table 3.4.4: Forecasting performance of seven different forecast combination methods, in addition to the oracle combination weights (light grey row), on simulated correlated forecasters. Changing forecaster quality. Values are provided by taking the median error metrics across 100 repetitions. Method(s) with the lowest MSE, MAE and SMAPE highlighted in green.

and effectively by simply propagating forward the prior state vector parameters, corresponding to stationary combination weights until more data is observed. The problem of missing forecasters is more complex, and is discussed in Chapter 4. The forecaster series in this section were aggregated to ensure no missing data were present.

3.5.1 Atlantic meridional overturning circulation

The Atlantic Meridional Overturning Circulation (AMOC) is a large system of ocean currents, which transports huge volumes of water around the globe. Warm water from the tropics is carried by surface currents into the North Atlantic, and likewise cold water is carried south by deep ocean currents several kilometres below the surface. The AMOC plays a crucial role in the global climate system, and contributes to the mild winter climate of North Western Europe. This significant impact on the global climate

means that the study and monitoring of the AMOC is of high importance to scientists, and many efforts have been made to model the phenomenon.

The AMOC is measured in Sverdrups (Sv), where 1Sv represents a transport of one million cubic metres of water per second (1Sv= $10^6 m^3/s$). In order to quantify this, it is necessary to take measurements that span a complete ocean basin. As such, observation is difficult. The RAPID/MOCHA/WBTS (RAPID) program was started in 2004 in order to monitor the AMOC at a latitude of 26.5°N. An array of instruments was deployed across 6500km of the Atlantic ocean, from Morocco to Florida, in order to measure ocean velocity, pressure, temperature and salinity. Since the installation of this basin-wide array, the AMOC has been continuously monitored (see McCarthy et al. (2015) for details).

Despite the large improvement in the observational network around the North Atlantic over recent decades, there remain significant gaps in the observational coverage (see Sanchez-Franks et al. (2021)). Ocean reanalyses allow observations from measurement instruments to be integrated with an ocean model based on physical equations, and can be used to provide reconstructions of the AMOC strength. Such reanalyses can be used to estimate the AMOC prior to 2004, and also at different latitudes. Although many ocean reanalyses use similar models and assimilated data, assimilation techniques vary, leading to differences in results. In this section, we apply our forecast combination method to four different AMOC reconstructions provided by ocean reanalyses, in order to dynamically weight the reanalyses and gain insight on their performance throughout time. We use the RAPID data as the true observed AMOC in our model. Our decision to implement our method on this data was motivated by Jackson et al. (2019), who assess the ability of several ocean reanalyses to capture the observed dynamics of the AMOC by comparing reanalyses constructions with data from the RAPID array.

Jackson et al. (2019) consider a reconstruction of the AMOC strength from the GloSea5 ocean reanalysis in their work, which combines observations of salinity, ocean

temperature, sea ice concentration and sea level from both in situ and satellite observations (MacLachlan et al. (2015)). The GloSea5 reanalysis is also considered by Sanchez-Franks et al. (2021), who note that this reanalysis underestimates the AMOC mean values in 2004 to 2010 and does not capture the strengthening apparent in the RAPID AMOC data in 2006. We wanted to see if this behaviour was reflected in the weights assigned by application of our model to the data.

The data

Data from the RAPID AMOC monitoring project are funded by the Natural Environment Research Council and are freely available from `www.rapid.ac.uk/rapidmoc`. The reanalyses data set used by Jackson et al. (2019) is available at: `https://zenodo.org/record/2598509#.Y-TC4BPP23J`. The raw RAPID data set features AMOC estimates every day from April 2nd 2004, taken twice a day at midnight and midday. Reanalysis data is provided on a monthly basis from January 1993, which is when satellite altimetry data became routinely available, to December 2016. The data features AMOC reconstructions from four different ocean reanalyses: GloSea5, ECCO V4 R3, GLORYS12v1 and GONDOLA100A.

In order to apply our method, five artificial time series were created by aggregating each series into monthly means and assigning these to the first day of each month. Only the time period for which both RAPID observations and reanalysis data were available was considered, leading to a time series of 141 observations from April 2004 to December 2016. The RAPID AMOC measurements and the corresponding reanalyses reconstructions are shown in Figure 3.5.1.

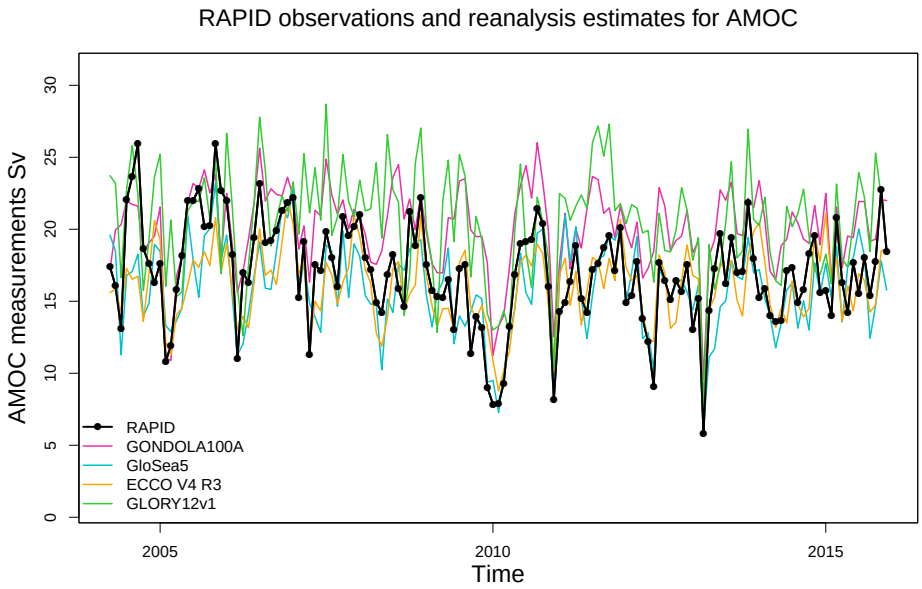


Figure 3.5.1: Monthly mean AMOC measurements from the RAPID array shown by black line, from April 2004 to December 2016. Reconstructions of the AMOC from four different ocean reanalyses: GloSea5, ECCO V4 R3, GLORYS12v1 and GONDOLA100A, shown by the coloured lines.

Choice of parameters

In order to implement our forecast combination model on the reanalysis data, a gamma prior distribution was assumed for the unknown observational precision,

$$(\phi|D_0) \sim G(n_0/2, n_0 s_0/2),$$

where initial parameter values were set to $n_0 = 1$ and $s_0 = 1$. The prior distribution of the state vector was then assumed to be,

$$(\boldsymbol{\theta}_0|\sigma^2, D_0) \sim N(\mathbf{m}_0, \sigma^2 \mathbf{C}_0^*),$$

where the parameters were set to

$$\mathbf{m}_0 = 1/4\boldsymbol{\iota}, \quad \mathbf{C}_0^* = 10^7 \mathbb{I}_4,$$

where $\boldsymbol{\iota}$ is a 4-dimensional vector of 1s as before. In order to deal with the unknown system evolution covariance, a discount factor δ was introduced. Sanchez-Franks et al. (2021) note that the GloSea5 reanalysis does not capture the strengthening of the AMOC in 2006, but performs well for other time periods. Hence, we expect there to be some evolution in the combination weights, corresponding to stochastic evolution in the state vector. With this in mind, we carried our the combination method for four different values of discount factor: $\delta = 0.900, 0.925, 0.950, 0.975$. All of these lie within the region of values recommended by West and Harrison (1997). We do not consider very high values (for example, $\delta = 0.99$) since these correspond to very little stochastic variation in the state vector.

Filtered weights and forecasting performance

The filtered combination weights for each choice of discount factor are shown in Figure 3.5.2 (a)-(d). We see that as the value of discount factor is increased, the filtered weights become less variable throughout time. By examining the case that $\delta = 0.975$, it is a little easier to identify any long-term trends in the filtered weights, which may be indicative of individual reanalysis performance. For example, the weight associated with the GLORY12v1 reanalysis (shown by the green line) is centred about zero for the majority of the time series, with a slight increase being exhibited from 2014 onwards (though this is not enough data to assume a long-term change from this point). This implies that the GLORY12v1 reanalysis does not adequately capture the behaviour of the AMOC at 26°N when compared to the other reanalyses included in the combination.

A sharp decrease in the weight associated with GloSea5 is exhibited just prior to 2006. The weight recovers from this quickly, before generally declining until around 2009. While we cannot say what may have caused the sudden decrease prior to 2006, it is possible that the steady declination of the weight from 2006 to 2009 is reflective of the inability of this reanalysis to capture the strengthening of the AMOC in 2006

as described by Sanchez-Franks et al. (2021). Similar behaviour is exhibited by the GONDOLA100A reanalysis. On the other hand, the ECCO V4 R3 reanalysis increases until around 2010, possibly indicating that this reanalysis captures the behaviour of the AMOC better in this period. We also see that some negative weights are assigned throughout the time series, implying correlations between the four reanalyses.

The forecasting performance of each choice of discount factor is compared in Table 3.5.1, where the MSE, MAE and SMAPE are provided. Measures of error are also provided for the popular forecast combination methods of simple averaging, regression-based weights and estimated Newbold and Granger (1974) weights, in addition to the 'recent best' method. Rolling windows of 50 observations were used to accommodate changing combination weights for the Newbold and Granger (1974) and regression-based methods.

The lowest value of MSE, MAE and SMAPE are highlighted in green in Table 3.5.1. We see that the lowest error metrics were achieved using our DLM-based forecast combination method with a discount factor of $\delta = 0.925$. Furthermore, the results show that for all values of discount factor tested, $\delta = 0.900, 0.925, 0.950, 0.975$, our DLM-based forecast combination method outperformed all other forecast combination methods considered.

3.5.2 European Central Bank survey of professional forecasters

The European Central Bank (ECB) survey of professional forecasters has been used widely in the forecast combination literature (see Magnus and Vasnev (2023), Radchenko et al. (2023), Matsypura et al. (2018)). The survey is concerned with forecasting three macroeconomic variables, namely: real GDP growth (GDP), inflation (HICP), and unemployment (UNEM). The survey began in 1999, and forecasters are asked to provide one- and two-year-ahead forecasts on a quarterly basis. In this em-

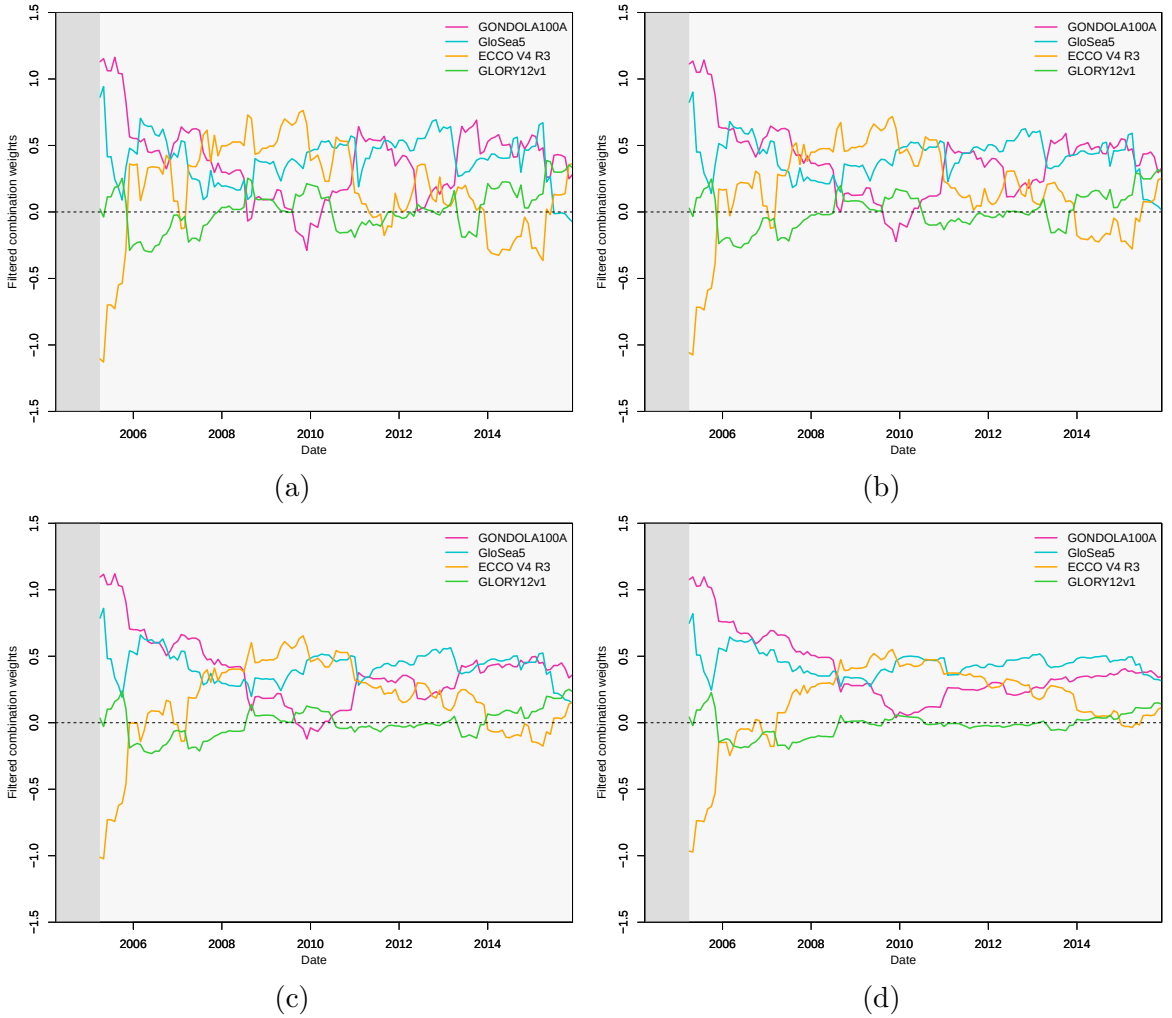


Figure 3.5.2: Estimated combination weights for GloSea5, ECCO V4 R3, GLORYS12v1 and GONDOLA100A ocean reanalyses, for different values of discount factor: (a) $\delta = 0.900$, (b) $\delta = 0.925$, (c) $\delta = 0.950$, (d) $\delta = 0.975$. Filtered weights have been omitted for a 12 month 'burn in' period to allow the posterior state variance to converge.

pirical analysis we will consider the one-year-ahead forecasts. Approximately 100 forecasters take part in the survey; however, most participating forecasters fail to provide forecasts for all time periods. The ECB survey data is freely available at `http://www.ecb.europa.eu/stats/prices/indic/forecast`, and is updated four times a year. Observed data of the GDP growth rate, HCIP and UNEM macroeconomic variables are available from the ECB at `http://sdw.ecb.europa.eu`.

In the forecast combination literature, it is generally the case that forecasters with infrequent responses are excluded from the analysis. Matsypura et al. (2018) suggest

Combination Method	MSE	MAE	SMAPE
Our method ($\delta = 0.900$)	3.718	1.432	9.31%
Our method ($\delta = 0.925$)	3.645	1.417	9.23%
Our method ($\delta = 0.950$)	3.658	1.426	9.31%
Our method ($\delta = 0.975$)	3.875	1.468	9.58%
Optimal covariance weights (rolling)	4.310	1.591	10.22%
Regression-based weights (rolling)	4.074	1.505	9.80%
Simple average	6.210	2.073	13.20%
Recent best	5.137	1.784	11.24%

Table 3.5.1: Forecasting performance of our forecast combination method for four values of discount factor ($\delta = 0.900$, $\delta = 0.925$, $\delta = 0.950$, $\delta = 0.975$), along with the forecasting performance of estimated optimal covariance weights, regression-based weights, simple average forecast combination and recent best. Smallest error metrics shown in green.

removing forecasters which fail to provide more than 6 years worth of forecasts (corresponding to 24 observations); similarly, Magnus and Vasnev (2023) only consider forecasters who have missed no more than 10 time periods. We also take this approach, which results in seven forecasters for the GDP data and six forecasters for the unemployment data. We do not consider the inflation data set, since this had the fewest available forecasts.

Despite removing the most infrequent forecasters from the analysis, some missing forecaster data was still present. In order for our forecast combination method to be applied, missing data was imputed using the last available value for that forecaster. Alternative methods for dealing with missing forecaster data are discussed in Chapter 4. The observed data and resulting forecaster series for the GDP growth rate is shown in Figure 3.5.3, from October 1999 to March 2023. The observed data and forecasts for unemployment are shown in Figure 3.5.4, from February 2000 to August 2023.

The effect of the COVID-19 pandemic is clearly noticeable in both data sets. In the case of the GDP growth rate, forecasters were unable to predict the significant drop in 2020, leading to drastically wrong forecasts for this period. On the other hand, it appears several forecasters severely overestimated the unemployment rate for the period

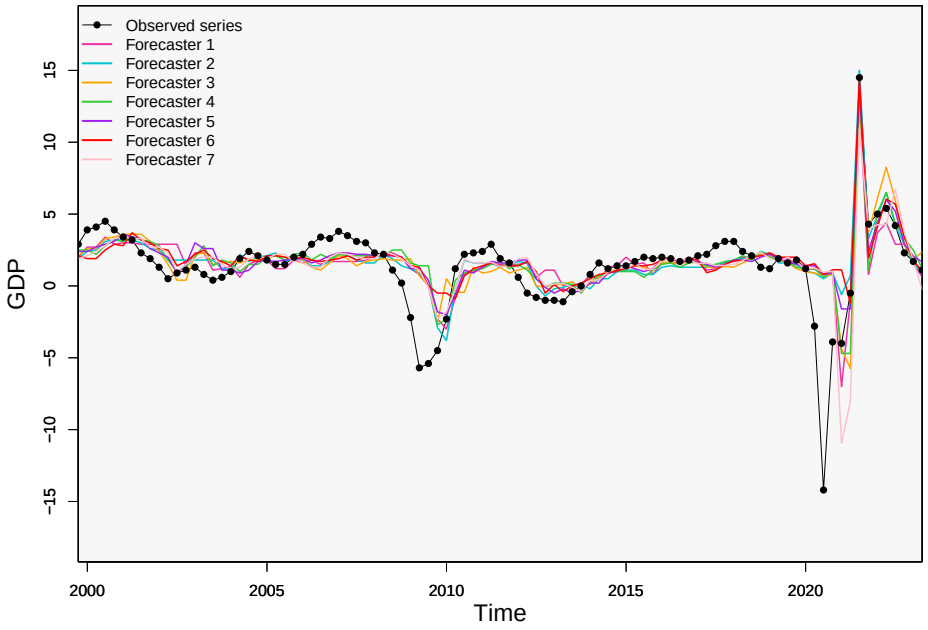


Figure 3.5.3: Observed GDP growth rate data from October 1999 to March 2023 shown by the black line, with observations given for each quarter. Corresponding one-year-ahead forecasts shown by coloured lines.

after the pandemic.

Choice of parameters

For both data sets, a gamma prior distribution was assumed for the unknown observational precision,

$$(\phi|D_0) \sim G(n_0/2, n_0 s_0/2),$$

where initial parameter values were set to $n_0 = 1$ and $s_0 = 1$. The prior distribution of the state vector was then assumed to be,

$$(\boldsymbol{\theta}_0|\sigma^2, D_0) \sim N(\mathbf{m}_0, \sigma^2 \mathbf{C}_0^*),$$

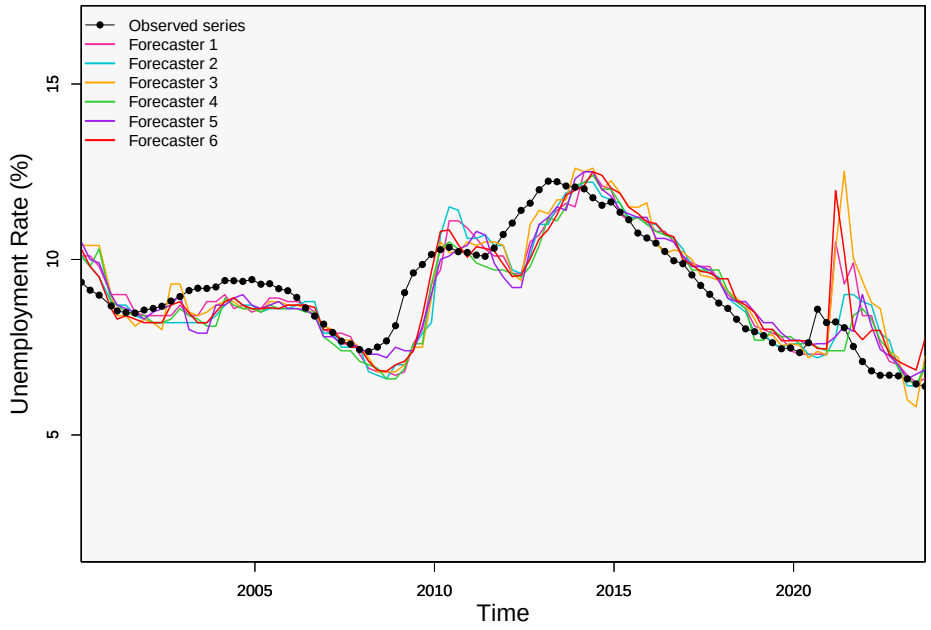


Figure 3.5.4: Observed unemployement rate data from February 2000 to August 2023 shown by the black line, with observations given for each quarter. Corresponding one-year-ahead forecasts shown by coloured lines.

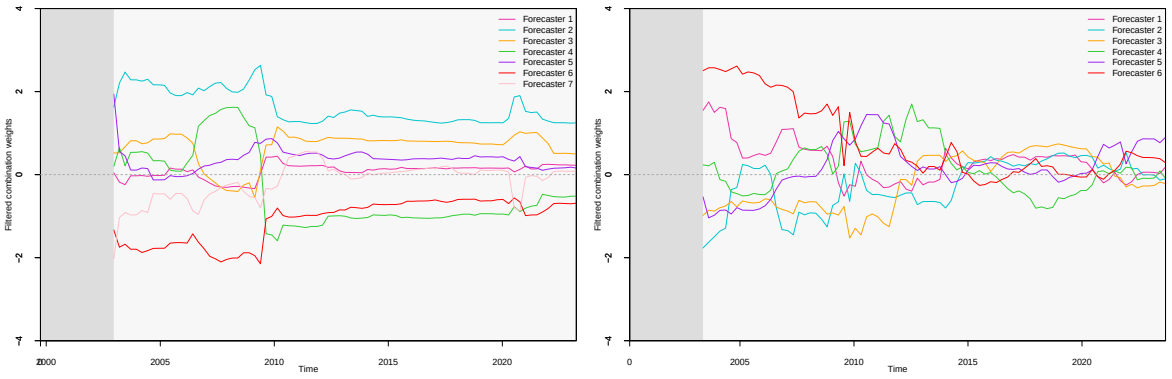
where the mean and covariance matrix were set to be the simple average weights and a noninformative prior respectively (see Section 3.3).

The choice of discount factor δ is difficult for this application due to the presence of the Covid-19 pandemic. Whilst we might expect a gradual change in forecaster quality throughout time like that observed for the AMOC, Figure 3.5.4 shows that certain forecasters severely overestimated the unemployment rate for the post-pandemic period. This implies a very sudden shift in forecaster quality at this point. We recall that sudden shifts in stochastic volatility of the weights can be accounted for by using a low value of discount factor; however, higher values of discount factor have been shown to perform better in more stable regions. Therefore, for this data set it may be preferable to use a time-varying discount factor. As methods for adaptive discount factor selection have yet to be discussed, we decided to select a constant value of discount factor throughout time and remove the data post 2020 when calculating error metrics. However, we note that forecasters also performed poorly in the late 2000s due to the financial crisis;

consequently, we believe that the use of our method with a constant discount factor might not be suitable to this data set. For the GDP data, the discount factor was set to $\delta = 0.999$; for the unemployment data set, the discount factor was set to $\delta = 0.935$. These values were chosen based on the data.

Filtered weights and forecasting performance

The filtered combination weights provided by our method are shown in Figures 3.5.5a and 3.5.5b for GDP growth rate and unemployment rate respectively. A 'burn in' period of 12 observations (corresponding to three years) has been omitted from the plot. Both plots show negative weights. This reflects the high correlations present between the ECB forecasters, as discussed by Radchenko et al. (2023). We also note that the weights for the GDP growth rate seem to stabilise after the 2008 financial crisis until the Covid-19 pandemic, implying that the forecasts were of consistent quality in this period.



(a) (b)

Figure 3.5.5: Filtered combination weights shown by the coloured lines for (a) GDP growth rate and (b) unemployment rate. A burn in period of 12 observations has been omitted.

The forecasting performance for the two data sets is compared with the 'rolling' optimal covariance weights method, the 'rolling' regression-based method, simple average forecast combination, and the recent best method in Table 3.5.2 and Table 3.5.3, for GDP growth rate and unemployment rate respectively. As mentioned previously, we

decided to remove observations after 2019 from the analysis in order to minimise the adverse affects of the Covid-19 pandemic. Rolling windows of 16 observations (corresponding to four years) were used to accommodate for changing weights in the optimal covariance weights method and the regression-based method. SMAPE was not included for the GDP data set since the observations were close to zero.

Combination Method	MSE	MAE
Our method ($\delta = 0.999$)	2.67	1.078
Optimal covariance weights (rolling)	11.07	1.667
Regression-based weights (rolling)	10.2	1.769
Simple average	2.27	0.991
Recent best	2.05	0.928

Table 3.5.2: Comparison of error metrics for different forecast combination methods applied to the ECB data set for GDP growth rate. Metrics were evaluated omitting the first $k = 16$ observations, corresponding to four years of data, and excluding data after 2019. SMAPE has not been included since values are close to zero.

Combination Method	MSE	MAE	SMAPE
Our method ($\delta = 0.935$)	0.625	0.637	6.41%
Optimal covariance weights (rolling)	1.103	0.743	7.95%
Regression-based weights (rolling)	0.773	0.641	6.60%
Simple average	0.564	0.566	6.02%
Recent best	0.430	0.467	4.94%

Table 3.5.3: Comparison of error metrics for different forecast combination methods applied to the ECB data set for unemployment rate. Metrics were evaluated omitting the first $k = 16$ observations, corresponding to four years of data, and excluding data after 2019.

Perhaps most interesting about these results is the stark contrast with both the simulations and the AMOC data set. The recent best method, which exhibited significantly poor forecasting performance in the previous studies, resulted in the lowest error metrics for both ECB data sets. Despite removing the post-pandemic period from the computation of the error metrics, our method was outperformed by both simple averaging and the recent best method. However, we do note that our method displayed

superior forecasting performance to the dynamic adaptions of the optimal covariance weights method and the regression-based method. This is unsurprising, given the small window size used to estimate the coefficients.

It is known that the ECB forecasters are very highly correlated as indicated from the filtered weights displayed in figures 3.5.5a and 3.5.5b (see also Radchenko et al. (2023) and Magnus and Vasnev (2023)). It is therefore possible that the superior performance of simple averaging can be attributed to the fact that we are trying to estimate individual combination weights from such highly correlated data. This feeds into the bias-variance trade off of the forecast combination puzzle, since it is likely that the gains obtained from estimating the combination weights do not outweigh the corresponding estimation error. Furthermore, methods that aim to estimate the combination weights are impacted greatly by the sudden 'shocks' of the financial crisis and the Covid-19 pandemic. This can make the use of past data to estimate the weights difficult, which would possibly explain the preferable performance of the simple averaging approach.

Although the recent best method performed well on this data set, we emphasise that this method is not robust to a sudden change in forecaster performance, and therefore not generally an appropriate choice. We expect that the positive performance of this method is a fluke result from the data set under consideration.

It is possible that superior performance from our method could be achieved by allowing a time varying discount factor to be used. This would allow for large changes in combination weights at highly volatile periods, while allowing for more consistent weight estimation in more stable periods. The concept of time varying discount factors is explored in Chapter 4.

3.6 Discussion

We have presented a step-by-step guide to combining point forecasts in a DLM framework. Furthermore, we have provided a novel formulation of the Kalman filtering equations specific to the forecast combination setting, which have been given in terms of discount factor δ. This is so that the forecast combination procedure can be combined with methodology from the adaptive discount factor selection literature discussed in Chapter 4.

The proposed DLM-based methodology enables combination weights to be estimated sequentially as more data become available. Consequently, this approach is highly applicable to situations where observed data and one-step-ahead forecasts are arriving in succession and with high frequency. By working within the DLM framework, inference on relevant distributions can be updated using closed-form Bayesian filtering. This is computationally efficient, contrasting with other methods wherein the covariance matrix of the forecasters must be estimated anew at each time step.

The way in which combination weights are updated sequentially also deals with the common issue of little historical data. By assigning a prior distribution to the combination weights with mean equivalent to simple averaging, we provide appropriate weight estimates from the outset. This is in contrast to methods such as those proposed by Newbold and Granger (1974) and Granger and Ramanathan (1984), which require a certain amount of information to compute initial weight estimates. The benefits of this were seen in the ECB empirical application, where we used the previous 16 observations to estimate the optimal covariance weights and the regression-based weights. Given the short time series, it would have been impractical to use a larger window size. In this application, the best performance was observed from methods which did not require parameters to be estimated at all, followed by our method, and finally the rolling covariance weights and regression-based weights.

As with the models of Sessions and Chatterjee (1989) and LeSage and Magura

(1992), we chose to model the evolution of the forecast combination weights as a random walk. This allows dynamic adaptation throughout time in response to forecasters' past performance, which is particularly useful in situations where the 'best' forecaster is changing throughout time. Unlike the aforementioned references, our method gives focus to the role of the discount factor parameter δ, which controls the stochastic volatility of the combination weights. We have demonstrated how an appropriate choice of discount factor allows gradually changing 'oracle' weights to be estimated in simulation studies; such behaviour corresponds to the case that the quality of a forecaster is changing gradually throughout time. We have also shown how the choice of δ plays a critical role in the forecasting performance of the method by comparing different choices of discount factor in an environmental empirical application.

In contrast to other more simple forecast combination methods (such as simple averaging), and more complex forecast combination methods (such as machine learning-based techniques), our method offers interpretability from the assigned forecast combination weights. For example, in the case of independent forecasters (and consequently, positive weights), a larger weight indicates superior forecaster performance. If we witness a positive weight decrease over time, this could be interpreted as the quality of the corresponding forecaster declining; likewise, an increasing positive weight could indicate an increase in forecaster quality. This can then be related to known information about the application; for example, a particular forecaster had difficultly predicting the increase in AMOC strength. If our method assigns negative weights, this indicates that correlations are present between the forecasters. This can make direct interpretations of forecaster quality more difficult (a more negative weight does not necessarily imply worse forecaster quality); however, it can provide valuable information about the degree of correlation between forecasters.

We conclude this chapter by considering some drawbacks of our DLM-based method. We saw in the ECB application that our method was outperformed by both simple aver-

aging and taking the recent best forecaster. This may be for two reasons: high levels of correlation between the forecasters, and the shocks to the data due to the financial crisis and the Covid-19 pandemic. The latter may have caused sudden jumps in forecaster quality (perhaps one forecaster did a significantly better job at predicting during these time periods); methods for dealing with this in the context of missing forecaster data are discussed in Chapter 4. In relation to the former, although our method does account for correlations between forecasters (by assigning negative combination weights), it is possible that the very high levels of correlation present in the ECB data negatively affected forecasting performance.

A high level of correlation between forecasters is common in many forecast combination applications, and the ECB data is no exception. Magnus and Vasnev (2023) show for the ECB data set that ignoring positive correlation can lead to confidence bands around forecast combinations that are much too narrow. Therefore, very high levels of correlation must be assumed between forecasters in order to correctly account for uncertainty. We showed in Chapter 2 that negative weights are optimal when combining two forecasters under certain conditions. Furthermore, our simulation studies showed how negative weights are assigned when correlations are introduced between forecasters. Radchenko et al. (2023) provide a thorough investigation of the negative weights that can emerge when combining forecasts. They study the theoretical conditions necessary for negative weights to emerge and find that negative weights are driven by high positive correlations.

Given that theoretically optimal weights can indeed take negative values, we developed our combination method such that the estimated weights would be unconstrained. However, the typical practice in the literature is to consider only convex combinations, and ignore or trim negative weights (set them to zero). Radchenko et al. (2023) show that the positive effect of trimming comes from reducing the variance of the estimated weights, but the threshold of zero is arbitrary and can be improved. In their empirical

analysis, Radchenko et al. (2023) range the threshold from $-\infty$ to 0 and look at the mean squared forecast error (MSFE) of the resulting combined forecast. They compare this with the MSFE achieved from taking the simple average. The recommendation of the paper is to implement trimmed weights using an optimal trimming threshold (rather than zero).

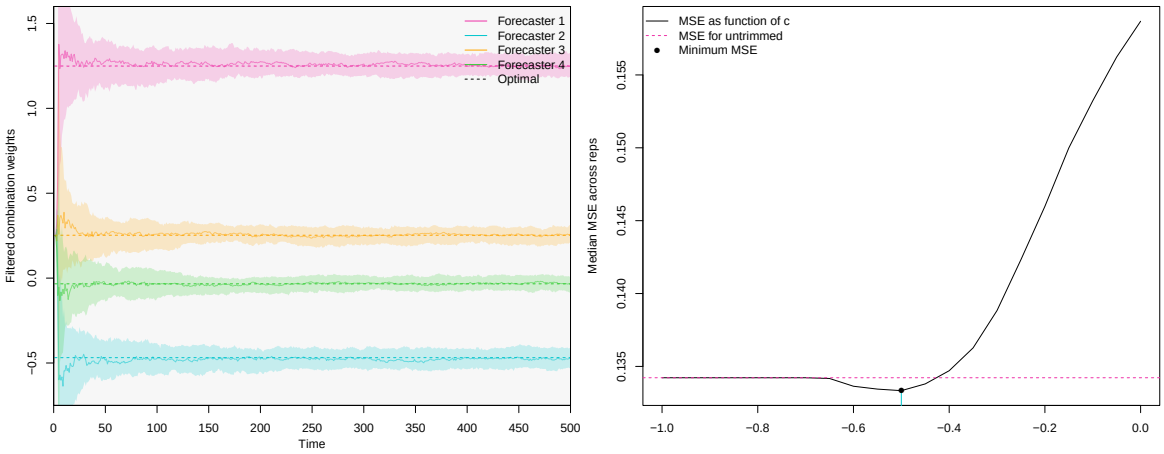
It is possible that the application of trimming procedures would lead to improved empirical performance of our DLM-based method for the ECB data set. For the sake of brevity, and issues that arise due to appropriate trimming threshold selection, a full empirical analysis of this has been omitted from this thesis. However, let us consider a simple toy example to investigate the effects of trimming highly correlated forecasters on MSE. We simulated 100 replications of four forecasters using the methodology from Section 3.4.2. The covariance matrix was set to,

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 & 0 & 0 \\ \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 & 0 & 0 \\ 0 & 0 & \sigma_3^2 & \rho_{34}\sigma_3\sigma_4 \\ 0 & 0 & \rho_{34}\sigma_3\sigma_4 & \sigma_4^2 \end{bmatrix},$$

with $\sigma_1^2 = 0.2$, $\sigma_2^2 = 0.4$, $\sigma_3^2 = 0.6$, $\sigma_4^2 = 0.8$. Correlation coefficients were set to $\rho_{12} = \rho_{34} = 0.9$, such that Forecasters 1 and 2 were highly correlated, and likewise for Forecasters 3 and 4.

The median filtered weights and interquartile ranges in the untrimmed case are shown in Figure 3.6.1a. We see that the large positive correlations between forecasters has lead to negative weights. Radchenko et al. (2023) utilise a two-step trimming procedure, where weights are first estimated and then trimmed according to

$$w_i^{TR1} = \alpha_1 \times \begin{cases} \hat{w}_i, & \hat{w}_i > -c, \\ -c, & \hat{w}_i \leq -c, \end{cases}$$



(a) Untrimmed filtered weights.

(b) MSE as a function of threshold parameter c.

Figure 3.6.1: Example of the effects of trimming negative weights.

where an additional scaling parameter α_1 is required to satisfy the constraint that the weights sum to one. This trimming procedure was implemented on the estimated weights, where the threshold c was allowed to range from -1 to 0 in steps of 0.05.

The median MSE across the replications for the trimmed weights is shown as a function of threshold parameter c in Figure 3.6.1b. The median MSE of our untrimmed weights is shown by the dashed pink line. We see that the MSE for the thresholded weights is lower for approximately choices of $c \in (-0.75, -0.45)$, with a minimum achieved at $c = -0.5$ (shown by the black dot). We note that this is close to the minimum filtered weight; consistent with Proposition 7.1 in Radchenko et al. (2023). Trimming the weights in this way decreases the variance of the estimated weights, leading to more accurate forecasts.

Of course, in practice the optimal trimming threshold is unknown. Radchenko et al. (2023) consider a data driven threshold in their empirical analysis, where the threshold is chosen based on the pseudo out of sample MSFE; however they found that this didn't improve forecasting performance when compared with the best fixed threshold for trimming.

It is possible that the combination of a trimming procedure with our method could

lead to improved empirical performance on the ECB data set. This could be incorporated as part of a two-step procedure, where weights are first estimated and then trimmed. Alternatively, there is the option to estimate and trim the weights at the same time as part of a one-step procedure. This could perhaps be achieved by utilising methodology from the constrained Kalman filtering literature; see, for example, Simon (2010) and Jiang and Zhang (2013). This is discussed in more detail in Chapter 6.

Chapter 4

Missing forecaster data

For some high frequency variable of interest y_t, we consider the case that N experts provide individual point forecasts for the observed value at each time. Each sequence of point forecasts has some underlying uncertainty, which is unknown and possibly changing dynamically, and the forecasts are often correlated. In this problem setting, a common difficulty that can arise is missing forecaster data. Since the data y_t is observed sequentially, after predictions are made, simple imputation techniques like linear interpolation cannot be applied. The question is therefore how should this missing data be dealt with; should only the available forecasters be combined? And if not, how should we impute the missing forecasters?

4.1 Introduction

In Chapter 3, we presented a DLM-based framework for combining point forecasts. We carried out simulation studies as both a proof of concept and to compare forecasting performance with other well-established forecast combination methods. We then applied the method to two empirical applications: the environmental AMOC data, and the ECB survey data.

In the aforementioned simulation studies, we assumed that our N forecasters pro-

vided data at every time point. Similarly, the empirical data was pre-processed such that no missing forecaster data was present before the forecast combination methods were applied. In order to ensure that each ocean reanalyses model provided predictions for the AMOC strength with the same frequency, we aggregated data. For the ECB data set, we chose to remove forecasters who failed to provide data for more than 10 time periods; any remaining missing data was then imputed by filling forward from the last provided forecast.

As demonstrated by the two empirical examples, missing forecaster data is a challenge that is frequently encountered in practice. There exists a multitude of reasons as to why this might be the case; perhaps different forecasting models have different temporal resolutions, or perhaps the missing data is more sporadic, resulting from human error. However, despite being a frequent occurrence in practice, the problem of missing forecasters in combinations is not widely studied in the literature. Capistrán and Timmermann (2009) consider the combination of forecasts with the 'entry and exit of experts' (also referred to as forecast combination with an unbalanced panel); however, in order to apply the least squares combination of Granger and Ramanathan (1984), missing forecasts are back-filled using the Expectation Maximisation algorithm. Therefore, such methods cannot be applied in an online setting. While Capistrán and Timmermann (2009) do consider least-squares combination methods without back-filling missing forecasters, they found that in approximately one-third of time periods, a subset of forecasters with suitably long histories of observed data could not be found. Consequently, the combined forecasts for these time points were given by the simple average, which can be viewed as neglecting the potentially useful historical data that is available.

In more recent work, Lahiri et al. (2017) discuss the challenges of comparing the performance of different combination algorithms when missing forecaster data is present. They argue that distinct combination methods implicitly impute missing forecasts in

different ways, and as a consequence the combination methods are applied to different data sets.

In order to understand how a given combination method can implicitly impute missing forecaster data, let us consider simple average point forecast combination. For N forecasters, let S_t^A denote the set of available forecasters at time t, and S_t^{NA} denote the set of forecasters that are unavailable. Denote the number of observed forecasts at time t as N_t, such that the number of elements in S_t^{NA} is given by $N - N_t$. If one chooses to compute the simple average combination of the available forecasters at time t, the weight assigned to Forecaster i is given by

$$w_{i,t}^{av} = \begin{cases} 0, & f_{i,t} \text{ is not available,} \\ 1/N_t, & \text{otherwise,} \end{cases}$$

where we use the superscript av to highlight the fact that these weights correspond to the simple average weights. The combined simple average (av) forecast is then given by

$$f_t^{av} = \sum_{i=1}^{N} w_{i,t}^{av} f_{i,t} = \frac{1}{N_t} \sum_{i \in S_t^A} f_{i,t}.$$

That is, the combined simple average forecast is equal to the mean of the available forecasters. In order to create a 'conceptually balanced' panel of forecasters, we must consider how this method imputes the missing forecasters. Denote the imputed value for some unobserved Forecaster j by $f_{j,t}^*$. The combined forecast can now be written as

$$f_t^{av} = \frac{1}{N} \left(\sum_{i \in S_t^A} f_{i,t} + \sum_{j \in S_t^{NA}} f_{j,t}^* \right).$$

Now, since the sum of the available forecasts is equal to the number of available forecasts

multiplied by the simple average combined forecast, $\sum_{i \in S_t^A} f_{i,t} = N_t f_t^{av}$, we can write,

$$
\begin{aligned}
f_t^{av} &= \frac{1}{N}\left(N_t f_t^{av} + \sum_{j \in S_t^{NA}} f_{j,t}^*\right), \\
f_t^{av} &= \frac{N_t}{N} f_t^{av} + \frac{1}{N} \sum_{j \in S_t^{NA}} f_{j,t}^*, \\
f_t^{av}\left(1 - \frac{N_t}{N}\right) &= \frac{1}{N} \sum_{j \in S_t^{NA}} f_{j,t}^*, \\
f_t^{av} &= \frac{1}{N - N_t} \sum_{j \in S_t^{NA}} f_{j,t}^*.
\end{aligned}
$$

This implies that applying the simple average method leads to imputed values $f_{j,t}^*$ ($j \in S_t^{NA}$), such that the average of these imputed values equals the average of the observed data. Lahiri et al. (2017) show that not all combination methods produce the same implicitly imputed values (and subsequently, the combination methods can be thought of as being applied to different data sets), and argue that this means that combined forecasts computed using different procedures are not comparable in the unbalanced panel setting.

Motivated by the sparse literature on the subject, we dedicate this chapter to a conceptual exploration of the missing forecaster problem. We have chosen not to include an empirical application, as we do not wish to focus on finding the best method for dealing with a particular data set with missing forecasts. Rather, we propose one possible method for imputing missing forecasters, and examine some of the implications of this.

One feature of the proposed approach is that it provides the practitioner with the choice to combine either point or density forecasts. We therefore also introduce two novel forecast combination methods tailored to deal with the output of the proposed imputation approach: an extension of the DLM-based point forecast combination procedure detailed in Chapter 3, and one density combination method. We aim to provide

insight on which methods are suitable for which kinds of missing forecaster data. To this end, we provide a conceptual simulation study which investigates two types of missingness.

4.2 Background and methodology

In Chapter 3, we described how the Kalman filter deals with missing data in the observed series y_t by simply propagating forward the state vector from the previous time point. This is straightforward to implement, and intuitive, since we have gained no new information with which to update the parameter estimates. However, there is no generally accepted method for dealing with missing values in the $\mathbf{F}_t$ vector, which in our model is given by the vector of N forecasts at time t, $\mathbf{f}_t = [f_{1,t} \ \dots \ f_{N,t}]$.

Of course, one approach for dealing with missing forecaster data is to only combine the available forecasts at each time. While this method is straightforward, and easy to implement in a simple combination setting such as simple averaging, it is not the most suitable for our problem. Firstly, we recall that there is an underlying uncertainty associated with each forecaster. Although the primary objective is to produce the most accurate combined forecast, it may also be of interest to the practitioner to quantify this uncertainty in some way. Quantification of this uncertainty would provide decision makers with a more comprehensive comparison between the different forecasters, and offer more insight into the uncertainty of the combined forecast; this in turn would enable more informed and balanced decision making. If we are to simply neglect a forecaster from the analysis when it fails to provide a forecast, we do not consider how this missing data might be related to the forecaster's uncertainty. For example, a forecaster repeatedly failing to provide a prediction due to human error may indicate that it is less reliable, and it is prudent to incorporate this information into our model in some way.

Furthermore, dealing with missing forecasters in this way raises the question of how should weights be allocated when a forecaster returns online. Is it sensible to simply restore a forecaster's weight back to its 'pre-missing data' level? From discussions with industry collaborators, we determined that it is possible that the quality of a particular forecaster will change after a prolonged period of missing data. This is intuitive; an upgrade to a forecasting system may lead to improved forecasts, meanwhile a new employee working on the forecasts may perform less favourably than their predecessor. For this reason, we desire an approach which requires forecasters to 'earn' their new weighting after a missing period, rather than simply be assigned their previous weight.

Secondly, it is not a trivial task to change the number of forecasters in the combination throughout time for some point forecast combination methods. The optimal covariance weights of Newbold and Granger (1974) require the covariance matrix between the forecasters to be estimated. If some forecasters fail to provide predictions for some time points, these cannot be included in the estimation. Hence, possibly useful information obtained from when forecasters have provided predictions in the past would be ignored. A similar problem is encountered in our DLM-based combination framework. Since the Kalman filter cannot be carried out when a forecast is missing, any forecasters with missing data would have to be neglected for the entire analysis. This would fail to utilise all the available information, leading to worse forecasts than those which could have been achieved at times when all forecasters were online.

Therefore, we propose that missing forecaster data must somehow be imputed; this also negates any issues for comparison of methods that may arise due to the reasons detailed by Lahiri et al. (2017). Of course, one option is to simply fill forward missing observations from the previously provided data point. Although this technique would preserve the online nature of the DLM-based combination method, it is not the most intuitive approach. Clearly, one expects the observed time series y_t, and likewise the associated forecasts, to change throughout time. Therefore, we seek an imputation

method that describes this evolution more accurately. More appropriate imputed values may be obtained via linear interpolation or smoothing; however, such methods can only be implemented in retrospect, after the missing forecaster comes back online. As we are considering forecaster data which is arriving sequentially, we instead seek a method that does not require knowledge of the next available forecast to make an imputation.

We propose modelling each individual forecaster series as a DLM. As the observed series will now be provided by the forecaster series itself, the Kalman filter provides a straightforward method for dealing with missing observations. Moreover, by virtue of the sequential nature of DLMs, the filtering of all N forecasters can be carried out in parallel, resulting in a computationally efficient method. If a suitable model is chosen for the forecaster series, fitting a DLM will provide imputed values which describe the true evolution of the series more effectively than simply filling forward the last observed value. Furthermore, modelling each forecaster as a DLM allows us to compute full forecasting distributions at each time point. This will enable the uncertainty of the forecasters to be quantified throughout time, with the effects of missing data directly reflected in the spread of the distribution. Therefore we are afforded a choice: should only the point forecasts (obtained by taking the expectations of the distributions) be combined, or should we now consider combining the full density forecasts? The latter may be more preferable if one wishes to quantify the uncertainty of the combined forecast.

In this chapter, we consider both approaches. For the DLM-based point forecast combination approach proposed in Chapter 3, we introduce the concept of adaptive discount factor selection. This was motivated by a desire to develop a method that can adapt quickly to sudden changes in forecaster quality (for example, caused by a forecaster going offline). We integrate an existing particle-learning based method for adaptive parameter selection with our DLM-based point forecast combination framework, which negates the need for the multiprocess mixture model proposed by LeSage

and Magura (1992). In the context of density forecast combination, we propose a novel method for assigning dynamic finite mixture model weights, based on maximising the log predictive score.

4.2.1 Proposed imputation methodology

In order to deal with missing forecaster data, we propose using a DLM to model each individual forecaster series. As a simple example, consider an observed time series y_t simulated according to an AR(1) plus noise model, as described in Chapter 3,

$$\begin{aligned} \tilde{y}_t &= \phi\tilde{y}_{t-1} + (1-\phi)\mu + \epsilon_t, \quad & \epsilon_t &\overset{iid}{\sim} N(0, \sigma_\epsilon^2), \\ y_t &= \tilde{y}_t + \zeta_t, & \zeta_t &\overset{iid}{\sim} N(0, \sigma_\zeta^2). \end{aligned}$$

We recall that this simulation was motivated by a financial application, wherein $\tilde{y}_t$ describes some underlying time series, and the observed series is given by adding some local market effects ζ_t. A set of N expert forecasts were then simulated from a multivariate normal distribution, as though 'unaware' of the local market effects,

$$\mathbf{f}_t \sim N(\tilde{y}_t\boldsymbol{\iota}, \boldsymbol{\Sigma}_t),$$

where $\boldsymbol{\iota}$ is a N-dimensional vector of 1s. Based on this simulating model, it is possible to describe each forecaster as a DLM. Recall, a univariate DLM is often written as its defining quadruple of the form $\{\mathbf{F}_t, \mathbf{G}_t, V_t, \mathbf{W}_t\}$. For Forecaster i, $i \in \{1, \ldots, N\}$, simulated from the above multivariate normal distribution, this defining quadruple is

given by,

$$\mathbf{F}_t = \begin{bmatrix} 1 & 0 \end{bmatrix}', \qquad \mathbf{G}_t = \begin{bmatrix} \phi & (1-\phi)\mu \\ 0 & 1 \end{bmatrix},$$

$$V_{i,t} = \Sigma_{ii,t}, \qquad \boldsymbol{W} = \begin{bmatrix} \sigma_\epsilon^2 & 0 \\ 0 & 0 \end{bmatrix}.$$

where $\Sigma_{ii,t}$ denotes the $[i,i]$th element of the forecaster covariance matrix $\boldsymbol{\Sigma}_t$ at time t. Of course, in practice the exact values of ϕ, μ, $\boldsymbol{\Sigma}_t$ and σ_ϵ^2 are unknown; however, it is straightforward to estimate these from the data. Alternatively, unknown observation and state evolution covariances can be estimated using the closed-form methods described in Chapter 3.

We emphasise that the choice of DLM is specific to the data set under consideration. In the previous chapter, we defined a universal DLM that could be applied to any point forecast combination problem (after dealing with missing forecaster data); here, the choice of DLM is to reflect the individual time series structure of Forecaster i, and is chosen by the modeller. Hence, we will not give further details on DLM selection for this purpose, and will refer to the individual DLM applied to Forecaster i as f_i-DLM. For more details on appropriate selection of DLMs for different types of time series, we refer the reader to West and Harrison (1997) and Petris et al. (2009), and note that many kinds of DLM can be constructed easily within the `DLM` package in R.

Once the defining quadruple for f_i-DLM has been specified, the Kalman filtering equations can be used to model the temporal evolution of the series of forecasts provided by the ith forecaster. In this case, the observed data is given by the forecaster series $f_{i,t}$, and not the observed series y_t. It is of importance to note that the observed series y_t is not taken into account in the modelling of the forecasters. The only information utilised in the f_i-DLM is the set of previous predictions provided by the ith forecaster.

We chose to model each individual forecaster as a DLM for two reasons. Firstly, DLMs deal with missing observed data by simply propagating forwards the parameters of the state vector from the previous time point, such that the posterior distribution of the state vector at time t is the same as the prior at time t. At time $t+1$, the prior mean and covariance of the state vector are then updated as usual, according to $\mathbf{a}_{t+1} = \mathbf{G}_{t+1}\mathbf{m}_t$ and $\mathbf{R}_{t+1} = \mathbf{G}_{t+1}\mathbf{C}_t\mathbf{G}'_{t+1} + \mathbf{W}_{t+1}$. Hence, the evolution of the state vector is still modelled throughout time, we just do not update our posterior distribution at each time point. Therefore, by applying a DLM to the forecaster series, any missing forecasts are dealt with implicitly as part of the filtering process. This is attractive since it negates the need for any offline imputation methods, whilst providing a more suitable alternative to simply filling forward the last provided forecast.

Secondly, modelling each individual forecaster as a DLM allows us to compute a full forecast distribution at each time t, rather than a simple point forecast. Recall, experts provide point forecasts at each time t, with no information regarding individual forecaster uncertainty. Application of the DLM-based point forecast combination method described in Chapter 3 does provide some interpretation of forecaster quality, reflected in the assigned combination weights. However, it is preferable to have an idea of individual forecaster uncertainty before the combination takes place. By taking the point forecasts provided by the experts, and applying the Kalman filter, we produce full forecasting distributions at each time. This allows the individual uncertainty of forecasters to be quantified in a more practical manner.

Here we provide a toy example which demonstrates the effects of modelling a forecaster series with missingness using a DLM. Under the above AR(1) plus noise process,

we simulated three forecasters with covariance matrix

$$\boldsymbol{\Sigma}_t = \begin{bmatrix} 1.0^2 & 0 & 0 \\ 0 & 1.5^2 & 0 \\ 0 & 0 & 1.7^2 \end{bmatrix}.$$

Artificially missing data was inserted for Forecaster 1, for times $t \in [90, 115]$. The Kalman filter was then applied to each forecaster series, with defining quadruple defined as above. In this set-up, density forecasts for Forecaster i at time t can be obtained from the posterior distribution of the state vector at this time. Figure 4.2.1 shows the observed data of interest y_t, along with the mean and 95% credible intervals of the density forecast for Forecaster 1 at each time. The missing data period is shown by the shaded grey region. We see that during the missing data period, the credible interval increases, reflecting the increased uncertainty in this period.

Thus, modelling the individual forecasters in this way enables a full density forecast to be computed at each time t, even in periods where the original forecaster may have failed to provide a forecast. This information can now be combined using point forecast combination methods (wherein the mean of the forecast distribution is taken as the point forecast), or density combination methods. In the case of point forecast combination, we require a method that allows for sudden changes in forecaster quality (reflective of a forecaster going offline, for example). With this in mind, we provide background on adaptive discount factor selection methods in the following, and present a novel methodology which incorporates an existing particle-learning based approach with our point forecast combination DLM. Furthermore, we present a novel method for combining density forecasts with dynamic weights. These methods are implemented on simulated data in Section 4.3.

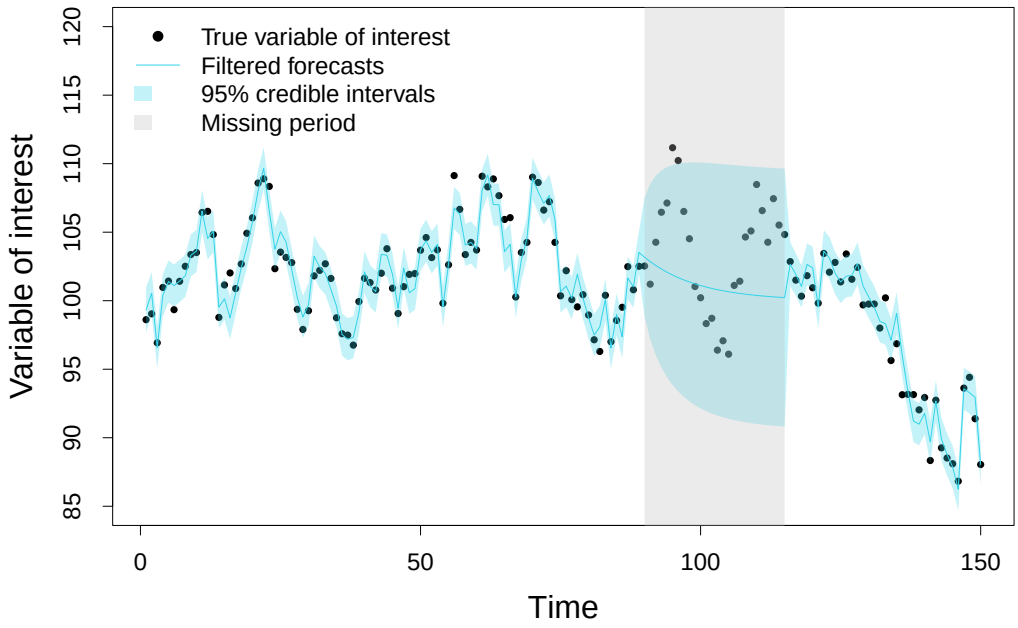


Figure 4.2.1: Simulated observed data y_t shown by black points. Artificially missing data was inserted for Forecaster 1 for times between $t \in [90, 115]$, shown by shaded grey region. Expectation of the filtered distribution shown by blue line; 95% credible intervals of filtered distribution shown by shaded blue region.

4.2.2 A dynamic discounting approach for point forecast combination

The role of the discount factor δ in our DLM-based point forecast combination method was introduced in Section 3.2.2. West and Harrison (1997) recommend a discount factor in the range $\delta \in [0.9, 0.99]$, while Koop and Korobilis (2012) consider $\delta \in \{0.95, 0.99\}$. In the simulations and empirical studies in Chapter 3, the discount factor was set to a constant value which was chosen a-priori. This is suitable in the case that the combination weights are changing gradually throughout time, at a constant rate. However, it is sometimes the case that different values of discount factor perform better at different times. Although we have proposed a method for imputing missing values for an offline forecaster, we still expect the quality of the said forecaster to decrease during the period of missing observations; this is reflected in the increase in the width of the

credible interval shown in Figure 4.2.1. In order to account for this very sudden change in forecaster quality, it is perhaps desirable to adopt a low value of discount factor at the start and end of the missing period. This would allow the combination weights to adapt quickly by responding to the recently observed data. Hence, in this section we describe an existing dynamic discount factor selection method from the literature, which we integrate with the DLM-based point forecast combination framework of Chapter 3.

Several alternatives to simply specifying a fixed value of discount factor a-priori have been proposed in the literature. For example, one approach is to define a grid of possible δ values, each with an associated probability, and update the posterior probability for each value whenever a new observation is made. For example, for the j-th choice of δ in the grid,

$$p(\delta_j|D_t) \propto p(y_t|D_{t-1}, \delta_j)p(\delta_j|D_{t-1}).$$

While intuitive, this method can lead to assigning significantly low probabilities to values of δ which may perform well in the future. This can in turn have an adverse effect on the predictive performance of the model, and therefore this method is often not favoured in practice. In order to prevent significantly low probability mass values being assigned to values of δ which may be favourable in the future, Zhao et al. (2016) utilise a power discount factor $\alpha \in (0, 1]$, and update the posterior probabilities such that,

$$p(\delta_j|D_t) \propto p(y_t|D_{t-1}, \delta_j)p(\delta_j|D_{t-1})^{\alpha}.$$

The introduction of the power discount factor α enables recent model performance to have a greater influence on the updated posterior distribution for δ. Zhao et al. (2016) consider a grid of possible values for both δ and α. Consequently, this method can become computationally expensive, therefore negating the computational savings

afforded from applying DLMs initially.

A more computationally efficient approach is provided by Yusupova et al. (2023), who propose a method for dynamic discount factor selection where δ_t is approximated through stochastic gradient descent (SGD). They exploit the idea that the optimal value of discount factor δ at time t is that which minimises the expectation of the one-step-ahead squared forecast error,

$$\delta_t^* = \underset{\delta}{\operatorname{argmin}} \frac{1}{2} E\{(y_t - \mathbf{F}_t' \mathbf{m}_{t-1})^2\}.$$

Although this optimisation cannot be solved directly Yusupova et al. (2023) note that an unbiased estimator is provided by the observed squared forecast error at time t,

$$e_t^2 = (y_t - \mathbf{F}_t' \mathbf{m}_{t-1})^2,$$

and approximate the optimal discount factor by computing a recursive formula for the derivative of this expression with respect to δ.

Although not proposed in the DLM literature, a further method for discount factor selection is provided by Irie et al. (2022). The authors augment a Poisson-Gamma state space model for web traffic data by introducing a time varying discount factor parameter. They develop an efficient particle-learning based estimation procedure that is suitable for sequential analysis, which allows the model to adapt quickly to structural changes.

Despite the fact that the approach of Irie et al. (2022) models the dynamics of a discount factor within a Poisson-gamma framework, the proposed non-linear state-space model for the evolution of δ can be directly implemented into a DLM. Given that this method is designed to work well with sudden structural changes, we decided to integrate this with our forecast combination procedure in order to explore the performance when a forecaster goes offline. To the best of our knowledge, this thesis provides the

first consideration of how such an adaptive discount factor selection approach may be integrated into the forecast combination framework.

In order to construct an appropriate particle-learning algorithm for the integration of Irie et al. (2022)'s method with our forecast combination framework, we will begin by describing the dynamics of the discount factor in line with Irie et al. (2022). Consider a logistic transformation of the form,

$$g_t = \text{logit}(\gamma_t) = \log\frac{\delta_t}{1-\delta_t},$$

where it is assumed that the transformed series, $g_1, g_2, \ldots, g_t, \ldots$, follows a first order autoregressive model,

$$g_t = (1-\phi)\mu + \phi g_{t-1} + \eta_t,$$

where $\eta_t \sim N(0, \sigma_\eta^2)$. Following the work of Irie et al. (2022), let us choose the prior,

$$g_0 \sim N(\mu, \sigma_\eta^2/\sqrt{1-\phi^2}),$$

such that the process g_t is stationary with marginal distribution identical to the prior g_0 for all t. It is then possible to implement a fully Bayesian approach by assuming priors on the AR(1) triplet, $(\mu, \phi, \sigma_\eta^2)$, and allowing the estimates to be updated as new data is received. Specifically, as per the work of Irie et al. (2022), let us consider another parameterisation given by $\phi_0 = (1-\phi)\mu$, $\phi_1 = \phi$, and $w = \sigma_\eta^{-2}$, and assume a normal-gamma hyperprior for these parameters,

$$p(\phi_0, \phi_1, w|D_0) = N(\phi_0, \phi_1|\tilde{m}_0, \tilde{C}_0/w)G(w|a_0/2, b_0/2).$$

The hyperparameters $\tilde{m}_0, \tilde{C}_0, a_0, b_0$ are user specified. Irie et al. (2022) recommend choosing these parameters such that a preference is given to high and stable values of

discount factor, since this is favoured in many cases. In our work, we choose to set these to $\tilde{m}_0 = [(1-0.9)\text{logit}(0.9), 0.9]$, $\tilde{C}_0 = 0.05^2\mathbb{I}_2$, $a_0 = 10$, $b_0 = 5$.

The state-space model is non-linear, and therefore particle-learning (Carvalho et al. (2010)) can be used in order to carry out filtering and estimation. We constructed a particle-learning procedure that incorporated the above model for the evolution of discount factor δ with the DLM-based point forecast combination procedure proposed in Chapter 3, such that both the parameters for the evolution of δ_t, and the parameters for the forecast combination DLM are updated at each time. This allowed us to output the value of δ_t at each time, in addition to the combination weight vector, for each particle. We note that the computational cost of this method is greatly increased when compared with a standard DLM with discounting.

4.2.3 Density combination

We recall that filtering the individual forecaster series leads to forecasting distributions at each time. Therefore, rather than combining the means of the distributions in a point forecast combination framework, we can instead choose to combine the distributions themselves.

Denote the density forecast provided by Forecaster i for the variable y_{t+1} at time t by $p_{i,t}(y)$, where we have introduced a dummy variable y as before, in order to represent all potential outcomes of the random variable y_{t+1} under the forecast distribution. As described in Chapter 2, a natural approach when combining density forecasts is to take a linear opinion pool. For a total of N individual forecasters, the linear opinion pool is defined by the finite mixture,

$$p_t(y) = \sum_{i=1}^{N} w_i p_{i,t}(y),$$

where weights are constrained to be non-negative and sum to one. The question is then

how to choose the combination weights? In the point forecast combination setting, optimal weights were described as those which minimised the expected mean square error. A similar notion is proposed in the density combination setting by Hall and Mitchell (2007), who suggest combining density forecasts using the weights which provide the most 'accurate' density forecast, in a statistical sense. We shall present a simple extension to this work which enables density forecasts to be combined dynamically, in order to account for changing forecaster quality.

Assume that the observed process is given by realisations from some unknown density $f(y)$. The combination weights which provide the most 'accurate' density forecast are therefore those which minimise the Kullback-Leibler information criterion (KLIC) between the true density $f(y)$ and the combined density $p(y)$. The KLIC is defined as

$$\begin{aligned} \text{KLIC}_t &= \int f_t(y_t) \ln \left\{ \frac{f_t(y_t)}{p_t(y_t)} \right\} dy_t, \quad \text{or} \\ \text{KLIC}_t &= E[\ln f_t(y_t) - \ln p_t(y_t)]. \end{aligned}$$

When the combined density is closer to the true density, the KLIC is smaller; the KLIC is equal to zero if and only if $f(y) = p(y)$. Hall and Mitchell (2007) state that in empirical applications, under certain regularity conditions, the KLIC can be consistently estimated by taking the sample mean over the available data,

$$\overline{\text{KLIC}} = \frac{1}{T} \sum_{t=1}^{T} [\ln f_t(y_t) - \ln p_t(y_t)]. \tag{4.2.1}$$

Hence, the optimal combination weights can be found by minimising the KLIC between the combined and true distributions, given by equation (4.2.1). Of course, in practice the true data generating distribution $f(y)$ is unknown; however, Hall and Mitchell (2007) note that in the forecast combination setting, this minimisation is equivalent to

finding the vector of weights $\mathbf{w}$ which maximise the concave cost function,

$$\begin{aligned}\Phi(\mathbf{w}) &= \frac{1}{T-1}\sum_{t=1}^{T-1}\ln p_t(y_{t+1}), \\ &= \frac{1}{T-1}\sum_{t=1}^{T-1}\ln\left(\sum_{i=1}^{N} w_i p_{i,t}(y_{t+1})\right),\end{aligned} \tag{4.2.2}$$

known as the log predictive score.

Although this approach is intuitively appealing, optimisation of the concave cost function can lead to increased computational expense, particularly in the case of many forecasters. Conflitti et al. (2015) tackle this by formulating an iterative algorithm for the computation of optimal weights, based on a 'minorisation-maximisation' strategy. The algorithm is given by

$$w_i^{(k+1)} = w_i^{(k)}\frac{1}{T-1}\sum_{t=1}^{T-1}\frac{\hat{P}_{ti}}{\sum_{i=1}^{N}\hat{P}_{tl}w_l^{(k)}},$$

where k denotes the iteration of the algorithm, and $\hat{P}$ is defined to be a $(T-1)\times N$ matrix with non-negative elements $\hat{P}_{ti} = p_{i,t}(y_{t+1})$. Under this notation, equation (4.2.2) can be written as,

$$\Phi(\mathbf{w}) = \frac{1}{T-1}\sum_{t=1}^{T-1}\ln(\hat{P}\mathbf{w})_t,$$

where $(\hat{P}\mathbf{w})_t$ denotes the product of the matrix $\hat{P}$ and the vector of combination weights $\mathbf{w}$, where the matrix $\hat{P}$ includes the densities for observations up to time t. In order to ensure that the estimated weights adhere to the non-negativity constraint, an appropriate initialisation must be chosen; Conflitti et al. (2015) recommend setting $w_i^{(0)} = 1/N$ for each forecaster. A stopping threshold must be defined such that the algorithm is terminated when the difference between the estimated weights on consecutive iterations is sufficiently small.

In order to account for changing forecaster quality, it is possible to optimise the log predictive score over windows of k observations, rather than the entire history of the time series. That is, the weights at time T will be given by those which maximise,

$$\Phi(\mathbf{w}_T) = \frac{1}{T-1-k} \sum_{t=T-1-k}^{T-1} \ln p_t(y_{t+1}),$$

for some window size k, where the subscript T has been introduced to reflect the fact that these weights are now time-varying. We found that the algorithm of Conflitti et al. (2015) can be modified in order to find such weights, by reducing the dimension of the matrix $\hat{P}$ such that it only includes the forecast densities for the previous $k-1$ observations under the combined distribution. The algorithm can then be implemented for each time t with a redefined matrix $\hat{P}^{(t)}$. However, we found in simulation studies that the assigned weights are incredibly sensitive to the choice of window size k, and therefore appropriate selection of this parameter presents a challenge in practical applications.

To avoid this, we instead propose computing the weights which maximise the following function at each time T,

$$\Phi(\mathbf{w}_T) = \sum_{t=1}^{T-1} \ln p_t(y_{t+1}) \alpha^{\Delta\tau}, \tag{4.2.3}$$

where $\Delta\tau = T - t$. Here, α denotes a forgetting factor that is between 0 and 1. The term $\alpha^{\Delta\tau}$ therefore describes the weight assigned to the observation at time t, where past observations retain more significance in the optimisation function as $\alpha \to 1$. In this case, rather than selecting an appropriate choice of window size k, the challenge is now selecting an appropriate choice of forgetting factor α. We find that this is a more intuitive problem than the former, since it is similar to the challenge of discount factor selection presented in our DLM-based point forecast combination framework.

4.2.4 Point combination vs density combination

To conclude this section, we highlight the key distinction between the point combination and the density combination approaches in this setting. If we choose to combine the N density forecasts in a point combination framework, whether applying a constant value of δ or adaptive δ_t, the effects of missing forecaster data are incorporated into the weight allocation indirectly. For example, when the ith forecaster fails to provide a forecast, an estimated value is provided by taking the expected value of the forecast distribution outputted by the f_i-DLM. It is likely that this value provides a less accurate prediction of the observed series than the available online forecasters, especially in the case that Forecaster i has been offline for a prolonged period. Therefore, when estimated combination weights are updated at this time using the DLM-based combination procedure, the weight assigned to the ith forecaster will drop (with the severity of this change influenced by the value of discount factor at that time). However, the decrease in the weight of the missing forecaster is not a guaranteed response to the forecaster going offline; it is perfectly possible that the predicted value for the missing period is not worse than the available online forecasters, and therefore the weight may not decrease at all. On the other hand, pursuing our proposed dynamic density combination method allows the effects of the missing forecaster to be incorporated into the combination weights explicitly. Since the entire forecasting density is taken into account in computation of the mixture weights, the increased spread of Forecaster i's distribution during the missing period will directly result in a lower assigned weight.

Whether these properties are beneficial or disadvantageous will depend on the type of missingness under consideration, and the observed data itself. This topic is explored further in the simulation studies in the following section.

4.3 Simulations

Missing forecaster data can arise for many different reasons, and take many different forms; perhaps one forecaster provides a prediction on a daily basis, whereas another provides one on every other day; perhaps a forecaster goes offline intermittently due to mechanical error; perhaps a forecaster is offline for an extended period while systems are upgraded. In the simulation studies in this section, we consider two different types of missingness: sporadic missing forecasts, and large periods of missing forecaster data. We consider how our proposed method for imputation deals with these two different types of missing data, and apply different combination methods to the resulting forecast densities. Strengths and weaknesses of each method are discussed.

Sporadic missing data may be encountered in practice due to problems such as lack of data availability, human error or network issues. For example, perhaps the data required for a forecasting model is incomplete or inaccurate for a given time, leading to a missing prediction. We chose to consider this type of missingness in our simulation studies in order to provide insight on how different combination methods deal with the case of single missing forecasts: how do the combination weights change when a forecaster fails to provide a forecast? Is responsiveness to recently observed data a bad characteristic of a combination method in this case? Are point forecast combination methods better at dealing with this type of missingness than density combination techniques?

Large periods of missing forecaster data may seem less intuitive at first glance, but can and do occur in practice; this is evidenced by the ECB data set considered in Chapter 3, wherein forecasters often failed to provide predictions for many consecutive time points. There are a variety of reasons as to why this type of missingness might occur; prolonged technical issues such as system failures or network outages, maintenance and upgrades to the forecasting system, and long-term absence of key personnel are all possible explanations.

For all simulations in this section, we generated an observed time series of length

$T = 150$ according to the AR(1) plus noise model first given in Chapter 3,

$$\begin{aligned} \tilde{y}_t &= \phi\tilde{y}_{t-1} + (1-\phi)\mu + \epsilon_t, \quad \epsilon_t \sim N(0, \sigma_\epsilon^2), \\ y_t &= \tilde{y}_t + \zeta_t, \qquad\qquad\qquad\quad \zeta_t \sim N(0, \sigma_\zeta^2), \end{aligned}$$

with parameter choices $\phi = 0.9$, $\sigma_\zeta^2 = 0.4^2$ $\sigma_\epsilon^2 = 2.5^2$, $\mu = 100$. We consider the case of $N = 3$ forecasters, simulated from a multivariate normal distribution as though 'unaware' of the local market effects,

$$\mathbf{f}_t \sim N(\tilde{y}_t\boldsymbol{\iota}, \boldsymbol{\Sigma}_t),$$

where $\boldsymbol{\iota}$ is a 3-dimensional vector of 1s. For both simulation studies, the forecaster covariance matrix was chosen to be constant throughout time and set to,

$$\boldsymbol{\Sigma} = \begin{bmatrix} 0.8^2 & 0 & 0 \\ 0 & 1.2^2 & 0 \\ 0 & 0 & 1.5^2 \end{bmatrix},$$

such that Forecaster 1 is the 'best' forecaster in terms of having the lowest variance. For both simulation studies, missing data was inserted only for Forecaster 1; both Forecaster 2 and Forecaster 3 provided data at all time points.

In order to deal with the missing forecaster data, all forecaster series were first filtered using an appropriate DLM, as described in Section 4.2. This allowed density forecasts for all forecasters to be obtained at each time, even for the periods where Forecaster 1 was offline. As shown in the toy example in Section 4.2 (see Figure 4.2.1), filtering the forecaster series in this way leads to an increase in the spread of the distribution when missing data occurs. The resulting forecasts were then combined using our standard DLM-based point forecast combination method, our DLM-based point forecast combination method integrated with the particle-learning based discount factor

selection method of Irie et al. (2022) (both by taking the means of the distributions), and the density forecast combination method with weights given by equation (4.2.3).

Point combination - parameter selection and details

In our simulation studies we consider our DLM-based point forecast combination method with a constant value of discount factor δ, and our DLM-based point forecast combination method with adaptive discount factor selection. For the constant discounting method, we implement discount factor values of $\delta = 0.93, 0.95, 0.99$. The number of particles in the particle-learning based method was set to $n_p = 100$. For all point forecast combination methods, priors for the combination DLM were set to $\mathbf{m}_0 = [1/3\ 1/3\ 1/3]$ and $\mathbf{C}_0 = 10^7\mathbb{I}_3$.

We chose to carry out 50 replications of the experiment; that is, for a given observed series y_t, 50 sets of forecaster series were simulated from the specified covariance matrix, with missing data inserted for Forecaster 1 at the same time points for all replications. Of course, since the adaptive discounting method uses particle-learning, this approach outputs $n_p = 100$ different estimates for the combination weights at each time, for a single replication. In order to compare estimated weights from the different methods, we decided to take the median weight vector across the particles at each time as the output for a given replication. This therefore allowed the median weights and interquartile ranges across the 50 replications to be computed and compared for the different combination methods.

Density combination - parameter selection and details

Our proposed method for assigning dynamic mixture weights to the density forecasts was applied for three different values of forgetting factor: $\alpha = 0.93, 0.95, 0.99$. This was done in order to draw a comparison with the DLM-based point forecast combination method with constant value of discount factor. The optimisation was carried out using

the `CVXR` package in R, using the Splitting Conic Solver (SCS).

4.3.1 Sporadic missing data

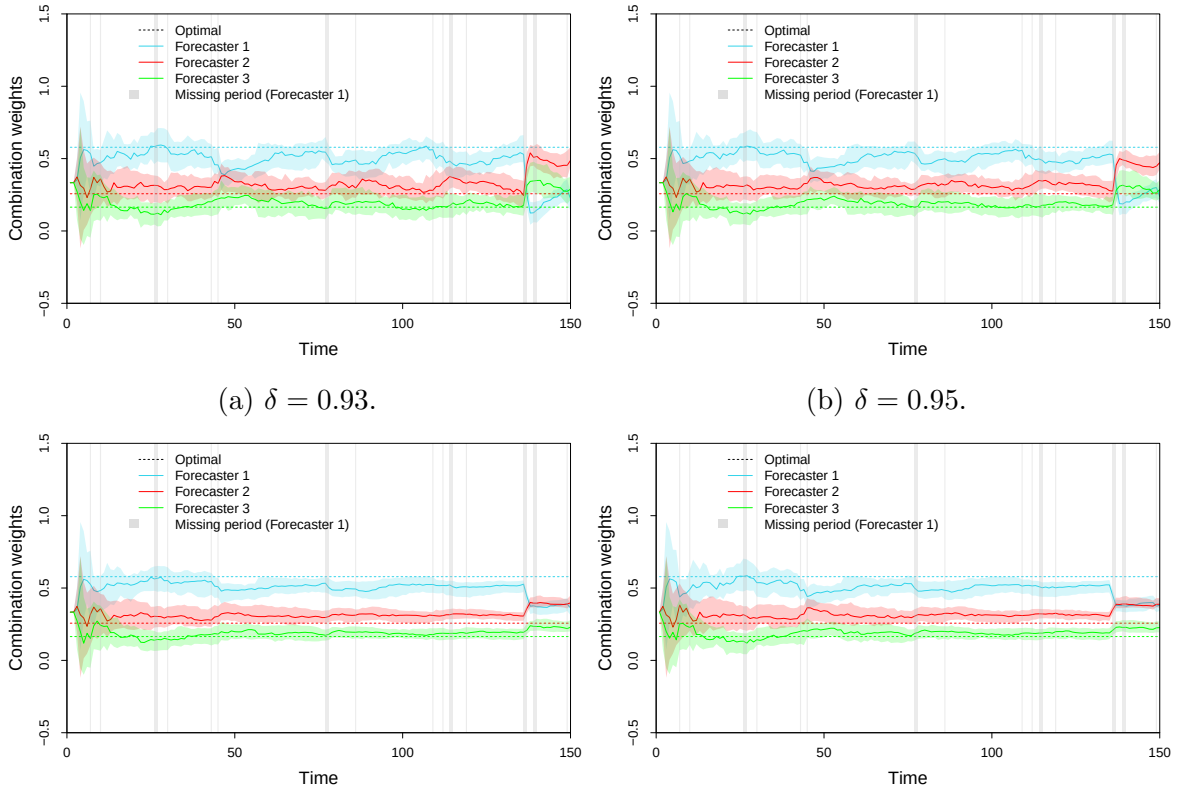
For the case of sporadic missingness, missing data was inserted randomly for Forecaster 1 for 20 observations throughout the time series. Figure 4.3.1 shows the median weights and interquartile ranges across the 50 replications for the different point forecast combination methods: (a) DLM-based combination with $\delta = 0.93$, (b) DLM-based combination with $\delta = 0.95$, (c) DLM-based combination with $\delta = 0.99$, and (d) DLM-based combination with particle-learning-based adaptive discounting. The missing data points for Forecaster 1 are shown by the shaded grey regions. We see that most missing observations are isolated, with at most two consecutive observations missing. The 'oracle' weights (as derived in Section 3.4.1) in the case that no forecasters had missing data are shown by the dashed lines.

We see in all cases that the presence of sporadically missing forecasts does not have a great influence on the combination weights for the majority of time points. Until approximately time $t = 140$, the weight assigned to Forecaster 1 remains the highest of the three combination weights, reflective of the low variance of this forecaster. Since Forecaster 1 is only ever offline for a maximum of two consecutive observations, this is a desirable feature of our imputation and combination methods. Clearly, it is unlikely that the quality of a forecaster will drastically decrease during a single period of missing data, and therefore methods which are overly sensitive would risk assigning a much lower weight than necessary when the forecaster returns online. The fact that our DLM-based point forecast combination method, both with a constant choice of discount factor and integrated with the adaptive discounting methodology, is not overly responsive to the missing observations indicates that our approach responds suitably to short and sporadic periods of missing data. Furthermore, the fact that the combination weights are not drastically impacted by the presence of missing observations implies that our

decision to impute missing forecasters by modelling each individual forecaster with a DLM is a sensible approach, for if the imputed values were acutely inadequate, then we would witness a sharp decrease in the corresponding forecaster weight.

Although all values of discount factor produced relatively stable combination weights, we note that the weights achieved from the combination with $\delta = 0.99$, and the combination with adaptive δ_t, responded the least to the missing observations. Recall, higher values of discount factor δ lead to more stable state vector estimates, which in our case describes combination weights which do not evolve much throughout time. The fact that our method with adaptive discount factor selection leads to such stable weights implies that a high value of δ_t is being chosen at each time step. As we do not expect the quality of the missing forecaster to drastically decrease in the missing period, we therefore expect these methods to display superior forecasting performance; this is compared in the following subsection.

Before we consider density forecast combination, we also note that all of the DLM-based point forecast combination methods struggle with a sharp decrease in the weight assigned to Forecaster 1 at approximately time $t = 140$. At this point, we see that Forecaster 1 has failed to provide two consecutive forecasts. We expect that this drastic decrease in weight is due to the presence of a shock in the observed series y_t. Recall, we have chosen to impute our missing forecaster by fitting a DLM to the series. Although this does provide some manner of modelling the evolution of y_t throughout time, clearly such models are incapable of capturing any shocks to the time series that have not yet been observed. Therefore, if indeed the observed time series y_t experiences a shock at approximately time $t = 140$, the estimated value for Forecaster 1 will be significantly inaccurate. Under the DLM-based point forecast combination framework, significantly inaccurate forecasts lead to low combination weights. This implies that perhaps point forecast combination methods are not the most robust option for combining our imputed time series.



(a) $\delta = 0.93$.

(b) $\delta = 0.95$.

(c) $\delta = 0.99$.

(d) Adaptive discount factor δ_t.

Figure 4.3.1: Combination weights evaluated using DLM-based point forecast combination. Optimal weights in the case that no missing data is present are shown by the dashed lines. Forecaster 1 missing in shaded grey regions.

We also consider application of our dynamic density combination method to the imputed forecaster series. Recall, this method assigns finite mixture weights to each distribution in the combination, based on maximising the exponentially weighted log predictive score given in equation (4.2.3). Figure 4.3.2 displays the median estimated mixture weights for the three forecasters across the 50 replications, along with the interquartile ranges. As before, the missing observations for Forecaster 1 are shown by the grey shaded regions.

It is important here to draw a clear distinction between our two approaches. Our DLM-based point forecast combination method aims to minimise the MSE of the combined forecast; on the other hand, our dynamic density combination method aims to maximise the log predictive score. Therefore, the oracle weights derived in Section

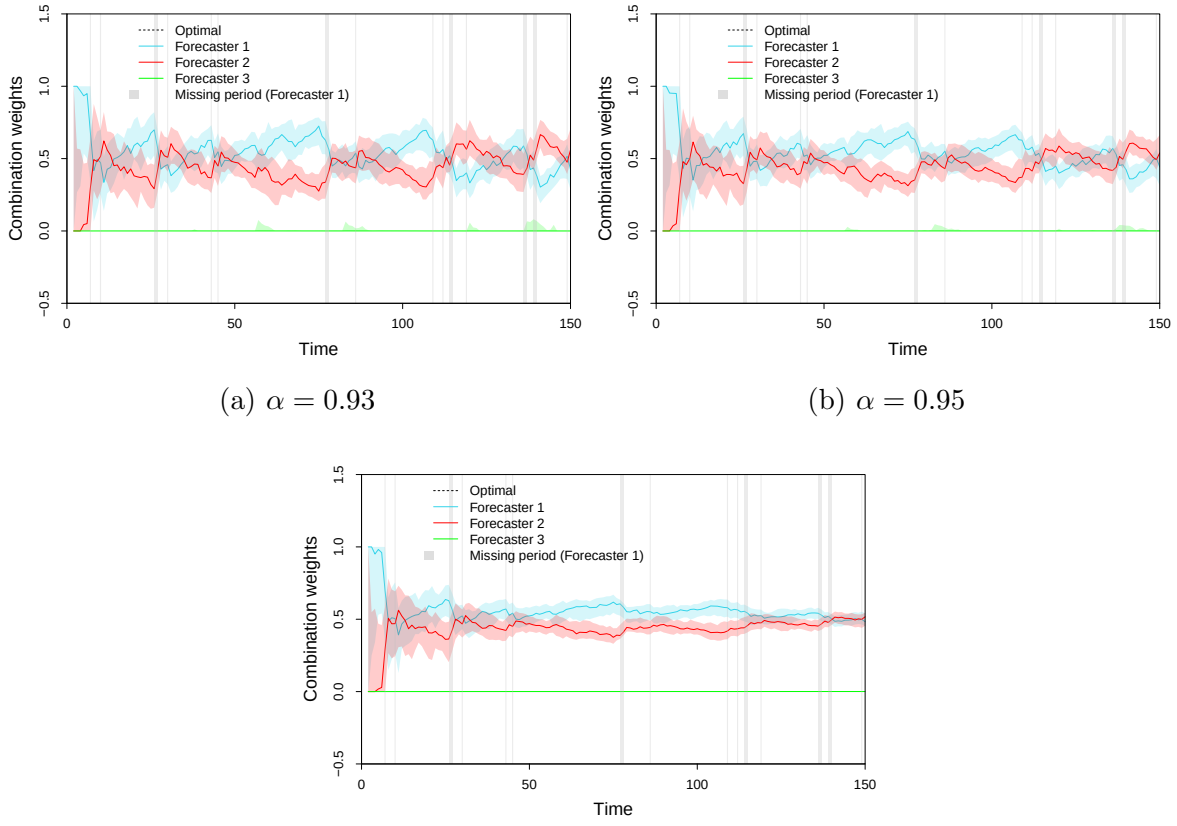
3.4.1 are not included in Figure 4.3.2, as these are not equivalent to the optimal finite mixture weights. We emphasise that the plots of the finite mixture weights have not been included such that direct comparisons can be drawn with the point forecast combination weights; rather, we wish to demonstrate the effects of sporadic missing data when combining density forecasts.

Figure 4.3.2 shows that Forecaster 3 has been assigned a zero weight for most replications, for all values of forgetting factor α. This is intuitive, since all forecast densities follow the same type of distribution, with Forecaster 3 having the widest spread. For all values of forgetting factor, the weights assigned to Forecaster 1 and Forecaster 2 are rather similar. As one would expect, weights are more variable for lower choices of forgetting factor, and more stable for higher choices. It does appear to be the case that the weight assigned to Forecaster 1 decreases when missing data is present, with the magnitude of this effect influenced by the choice of forgetting factor. This is reflective of the increase in variance of the forecasting distribution during the missing periods (recall Figure 4.2.1).

Unlike the DLM-based point forecast combination methods, the density combination method with different values of forgetting factor does not display a drastic decrease in the weight assigned to Forecaster 1 around $t = 140$. This is because the mixture weights are assigned based on the full distribution. Since the variance of the distribution of Forecaster 1 increases by the same amount for every individual missing period, the decrease in weight at time $t = 140$ is no greater than that at a different time period. Therefore, the finite mixture weights are less responsive to shocks in the observed time series, implying that the density combination approach is more robust.

Forecasting performance

As previously stated, our DLM-based point forecast combination procedure aims to minimise the MSE of the combined forecast. On the other hand, our dynamic density



(a) $\alpha = 0.93$

(b) $\alpha = 0.95$

(c) $\alpha = 0.99$

Figure 4.3.2: Combination weights evaluated using density forecast combination, designed to maximise the exponentially weighted log predictive score. Forecaster 1 missing in shaded grey regions.

combination procedure aims to maximise the log predictive score. Therefore, the interpretation of which method is 'best' will depend on which metric is of the most relevance to your application.

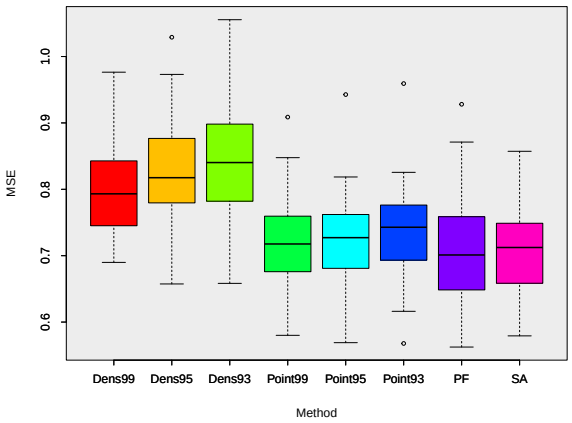
In this section, we compare the MSE, MAE, and SMAPE error metrics of the different methods. In order to obtain point forecasts from the density combination methods, we take the expectation of the combined mixture distribution. We include the simple average point combination as a benchmark. For the sporadic missing data case, the error metrics were computed using a data set of $T - k$ observations, where the first $k = 10$ observations were excluded as a 'burn in' period.

In addition to our usual error metrics, we also compare log-likelihoods of the meth-

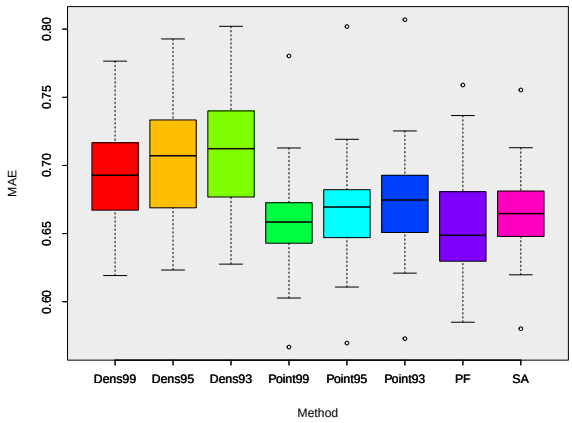
ods, where a higher log-likelihood describes a 'better' method. In order to compute the log-likelihood for the point forecast combinations, we recall that our combination procedure utilises the DLM framework. As detailed in Section 3.3, this provides access to Student's t-distributions for the combined forecast at time t, on account of the unknown observational variance V_t. In the case of our density combination, a mixture density can be obtained by simply taking the finite weighted mixture of the three forecaster densities. We also include the finite mixture density with equal weights as a benchmark; hereafter, this will be referred to as the simple average density.

Figure 4.3.3a displays box and whisker plots across the 50 replications, comparing (a) MSEs, (b) MAEs, (c) SMAPEs and (d) log-likelihoods, for the different methods considered in the previous subsection. We also include simple averaging (for MSE, MAE and SMAPE) and the simple average density (for the log-likelihood) as benchmarks. As expected, the point forecast combination approaches perform better on average for the MSE, MAE and SMAPE error metrics.

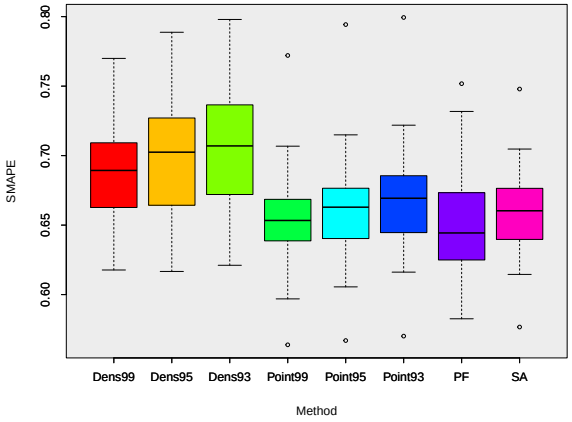
The lowest median MSE across the replications is achieved by the particle-learning based method for adaptive discount factor selection (denoted PF), with the median MSE across replications given by 0.701. This is followed by simple averaging (median MSE across replications given by 0.712) and DLM-based point forecast combination with $\delta = 0.99$ (median MSE across replications given by 0.718). Although the DLM-based point forecast combination with adaptive discounting does also achieve the lowest individual MSE across the 50 replications, we also note that this method has the widest interquartile range of all the point forecast combination approaches, implying that this is less robust over the replications. Finally, we acknowledge that although the density combination approach obtains higher median MSE across the replications for all values of forgetting factor, these are not substantially greater than the MSEs obtained by the point forecast approaches, and therefore such methods could still be applied in practice if more suited to one's application.



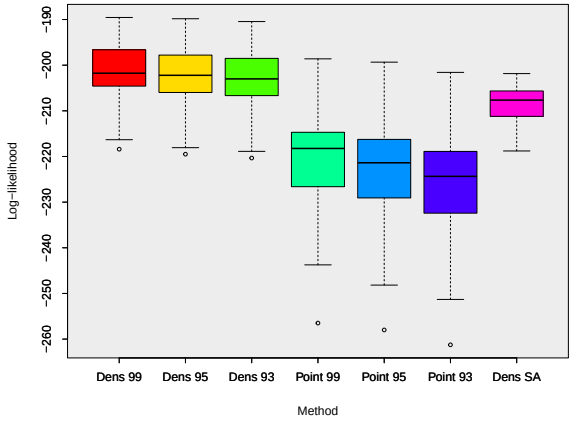
(a) Comparison of MSE across replications.



(b) Comparison of MAE across replications.



(c) Comparison of SMAPE across replications.



(d) Comparison of log-likelihood across replications.

Figure 4.3.3: Comparison of performance metrics across replications for the different methods implemented, featuring: point forecast combination with $\delta = 0.93, 0.95, 0.99$, point forecast combination with adaptive discounting, and density combination with $\alpha = 0.93, 0.95, 0.99$. For the MSE, MAE and SMAPE, simple averaging has also been included. For the log-likelihood, simple average density has been included.

MAEs across replications are displayed in Figure 4.3.3b. Here, the adaptive discounting approach displays the lowest median MAE (0.649), followed by our method with constant choice of discount factor $\delta = 0.99$ (0.658). Similar results are obtained in terms of SMAPE, shown in Figure 4.3.3c, with the lowest median SMAPE across replications given by the adaptive discounting method (0.644), followed by DLM-based point forecast combination with $\delta = 0.99$ (0.653). Hence, for the case of sporadic

missing data, the integration of our DLM-based point forecast combination procedure with the particle-learning based adaptive discount factor selection of Irie et al. (2022) demonstrates superior forecasting performance on average, when compared with the constant discount factor case. Despite simple averaging offering a benchmark that is difficult to outperform, we observed improved MSE, MAE and SMAPE across the 50 replications when utilising our DLM-based combination with adaptive discounting.

Log-likelihoods across the replications for the different methods are shown in Figure 4.3.3d. In this case, the highest log-likelihoods were achieved by the density combination methods, with the highest median log-likelihood obtained when $\alpha = 0.99$. This is as expected, since the combination weights were least responsive to the missing data in this case. Unlike the simple average point combination, the simple average density does not display comparably favourable performance.

4.3.2 Large missing data period

In order to consider a large missing data period, we introduced missing observations for Forecaster 1 for times $t \in [90, 115]$. Figure 4.3.4 shows the median weights and interquartile ranges across the 50 replications for the different point forecast combination methods: (a) DLM-based combination with $\delta = 0.93$, (b) DLM-based combination with $\delta = 0.95$, (c) DLM-based combination with $\delta = 0.99$, and (d) DLM-based combination with particle-learning-based adaptive discounting. Once again, the missing data period for Forecaster 1 is shown by the shaded grey region and optimal weights in the case that no forecaster had missing data are shown by the dashed lines.

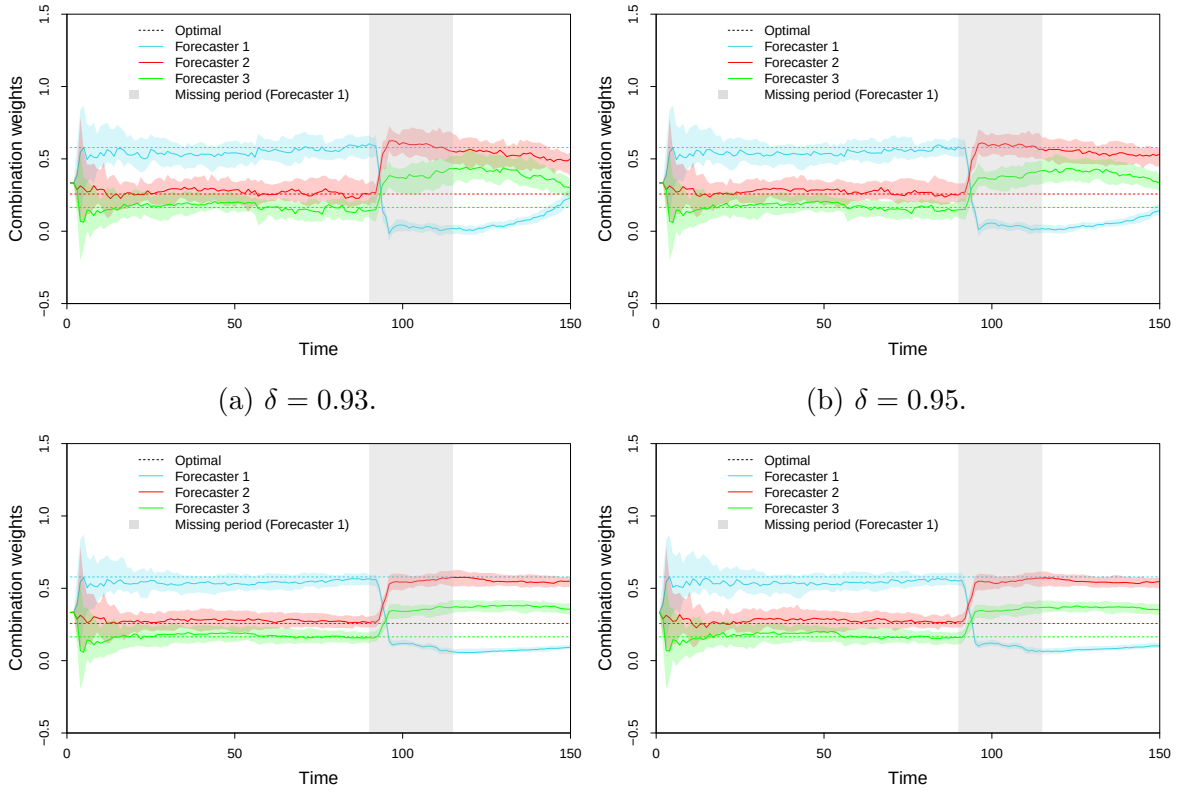
We see that all point forecast combination methods lead to a rapid decrease in the weight assigned to Forecaster 1 during the missing period; we also see that the weights of the other two forecasters increase in order to compensate for this. Given that the forecaster is offline for a significant period of time, it is sensible that such a drastic decrease in the combination weight should be exhibited: during the missing

data period, it is highly likely that the forecasts provided by Forecaster 1 will be less accurate than those provided by the online forecasters.

After approximately 10 missing observations, all four methods assign Forecaster 1 a weight of approximately zero. This implies that our particular imputation method is only useful for approximately the first 10 missing observations, after which point the gains of including the missing forecaster in the combination are negligible. It is possible that fitting a better DLM to the individual forecaster series would increase the number of missing observations for which this weight is non-zero. Furthermore, this would also be increased in the case that the available online forecasters were of sufficiently worse quality.

Most distinctions between the four methods can be drawn by considering the weights after Forecaster 1 returns online at time $t = 116$. From this time, the method begins to learn that Forecaster 1 is now performing better. Consequently, the combination weight assigned to this forecaster begins to gradually increase. The rate at which this weight increases varies for the different choices of discount factor. Recall that a higher value of discount factor δ means that historical information has a greater bearing on the combination weights. In this case, this means that the performance of Forecaster 1 in the missing data period continues to contribute to the combination weight even once the forecaster has returned online, resulting in a slow return to the pre-missing data level. Such behaviour can be observed in Figure 4.3.4c and Figure 4.3.4d, wherein the increase in the weight of Forecaster 1 towards the end of the time series is barely visible. This implies that the DLM with adaptive discounting is not selecting lower values of discount factor δ_t after the missing data period, despite the fact that this would lead to a quicker return to the optimal weights (and consequently improved forecasting performance). This was surprising, as the particle-learning approach developed by Irie et al. (2022) was created to deal with sudden jumps in the parameter value; we discuss possible reasons for the failings of this method in Section 4.4.

On the other hand, Figure 4.3.4a shows the weights in the case that $\delta = 0.93$. Here, the poor performance of Forecaster 1 in the missing period has less of an effect on the weight when the forecaster returns online, leading to a more visible increase in the weight of Forecaster 1 towards the end of the time series. Given that we simulated such that the covariance matrix of the forecasters $\mathbf{\Sigma}$ is constant with time, the optimal combination weights before the missing period are equal to the optimal combination weights after the missing period. Therefore, we expect the method with $\delta = 0.93$ to display superior forecasting performance when compared to the other point forecast combination techniques considered.



(a) $\delta = 0.93$.

(b) $\delta = 0.95$.

(c) $\delta = 0.99$.

(d) Adaptive discount factor δ_t.

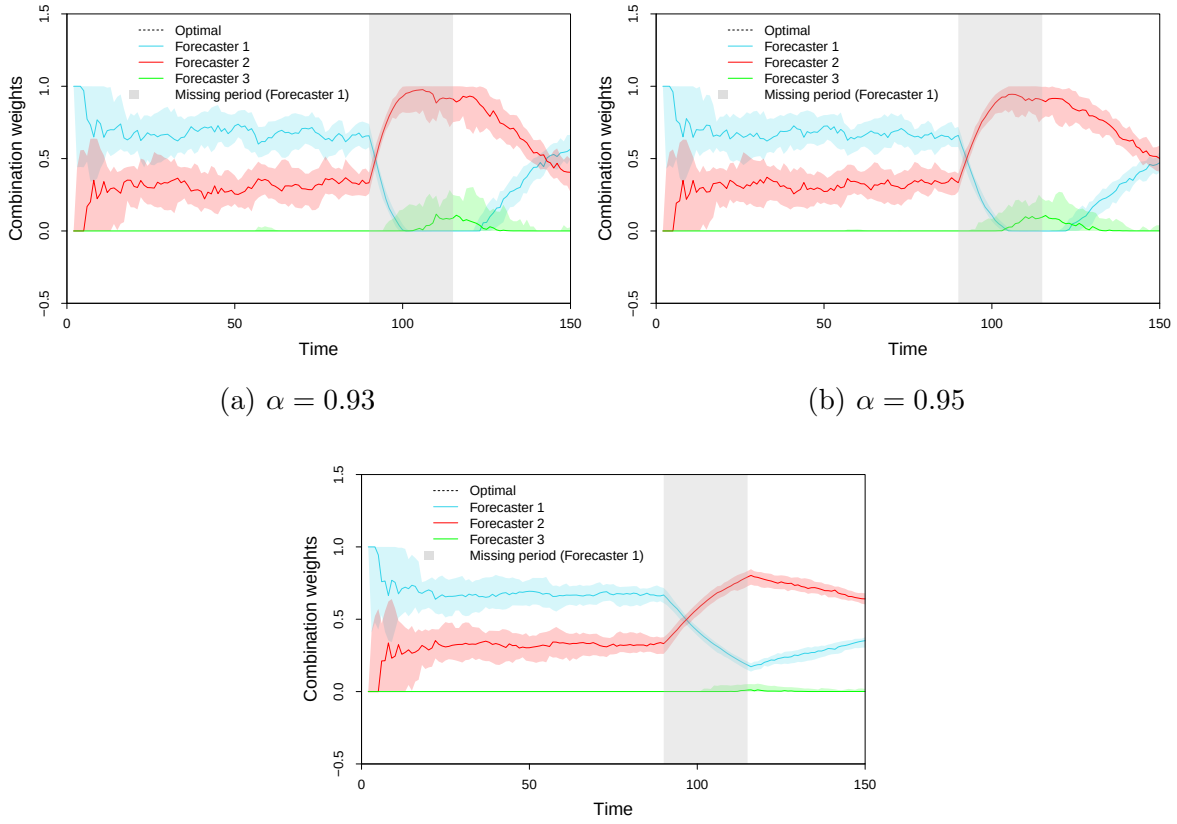
Figure 4.3.4: Combination weights evaluated using DLM-based point forecast combination. Optimal weights in the case that no missing data is present are shown by the dashed lines. Forecaster 1 missing in shaded grey region.

We also applied our density combination method to the simulation for the three choices of forgetting factor, $\alpha = 0.93, 0.95, 0.99$. The median estimated mixture weights

across the 50 replications, along with the interquartile ranges, are shown in Figure 4.3.5, with the missing data period shown by the shaded grey region. Again, we recall that the aim of this method is not to minimise the MSE of the combination, and therefore the oracle weights are excluded from the plots.

As was the case with the DLM-based point forecast combination methods, the weight assigned to Forecaster 1 decreases drastically in the missing period; however, we note that the rate at which this decrease occurs is quite different for the different values of forgetting factor α. In the case of $\alpha = 0.93$, shown in Figure 4.3.5a, the mixture weight assigned to Forecaster 1 decreases to zero after approximately 10 periods of missing data; this is inline with the point forecast combination methods. In order to counter this decrease and ensure that the sum of the mixture weights remains equal to one, the weights of the online forecasters increase in response. After the missing period, the weight of Forecaster 1 increases at a rate that is much quicker than any of the point forecast combination methods; by the end of the time series, the mixture weight has almost returned to it's pre-missing data level.

Figure 4.3.5c shows the case where the forgetting factor is set to $\alpha = 0.99$. Here, we see that the weight assigned to Forecaster 1 never drops to zero, even in the missing data period. The decrease in the weight is gradual, followed by a gradual but clear increase when the forecaster returns online. We emphasise that the rate at which the weight of Forecaster 1 recovers to it's pre-missing data level is much quicker for the density combination methods than their point forecast combination counterparts. This is a result of the fact that the spread of the density for Forecaster 1 decreases after the missing data period. While the DLM-based point forecast combination method must wait for the forecaster to prove itself enough to outweigh the effects of the previous poor performance, the density combination method incorporates the reduced variance of the missing forecaster into the assigned mixture weights. We explore how this affects forecasting performance in the following subsection.



(a) $\alpha = 0.93$

(b) $\alpha = 0.95$

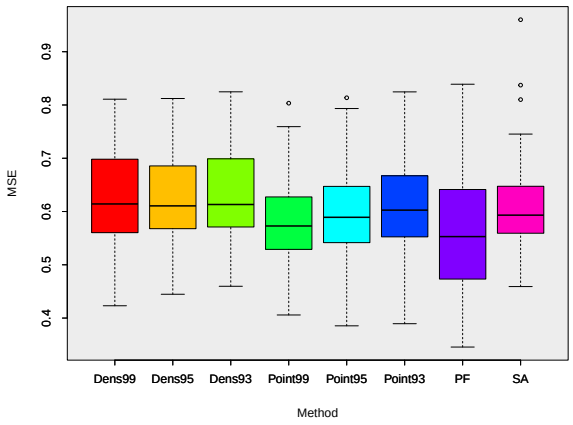
(c) $\alpha = 0.99$

Figure 4.3.5: Combination weights evaluated using density forecast combination, designed to maximise the exponentially weighted log predictive score. Forecaster 1 missing in shaded grey region.

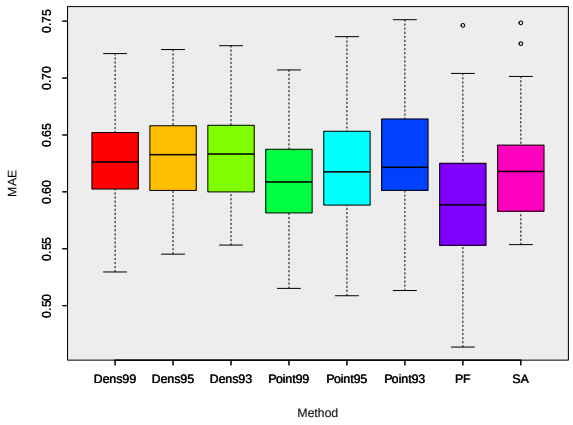
Forecasting performance

In contrast to the sporadic missing data case, wherein we compared forecasting performance across the time series as a whole, we now consider forecasting performance in three distinct time periods: before, during and after the missing data period. We remove the first $k = 10$ observations from the before period as a 'burn in'. Once again we consider the MSE, MAE, SMAPE and the log-likelihood of the different methods. For the three error based metrics, we also include the simple average results as a benchmark; for the log-likelihood, we include the simple average density.

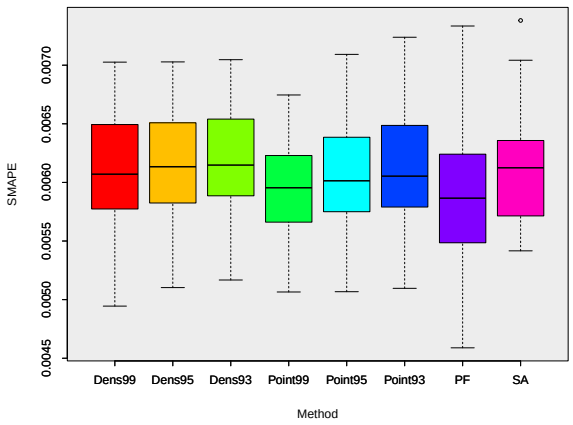
Figure 4.3.6 displays box and whisker plots across the 50 replications for the period **before** the missing data, comparing (a) MSEs, (b) MAEs, (c) SMAPEs and (d) log-



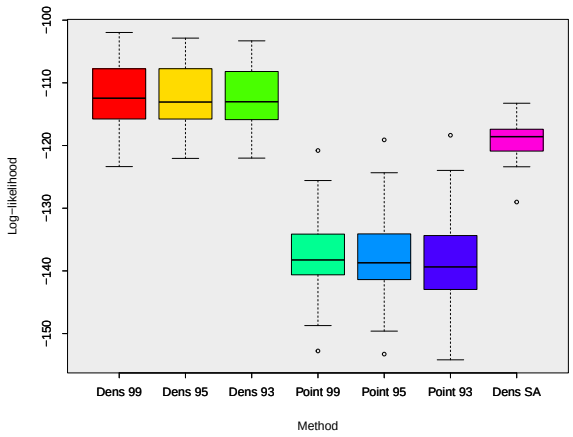
(a) Comparison of MSE across replications.



(b) Comparison of MAE across replications.



(c) Comparison of SMAPE across replications.

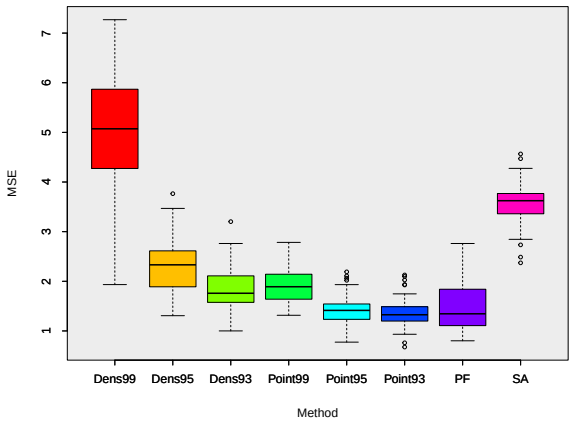


(d) Comparison of log-likelihood across replications.

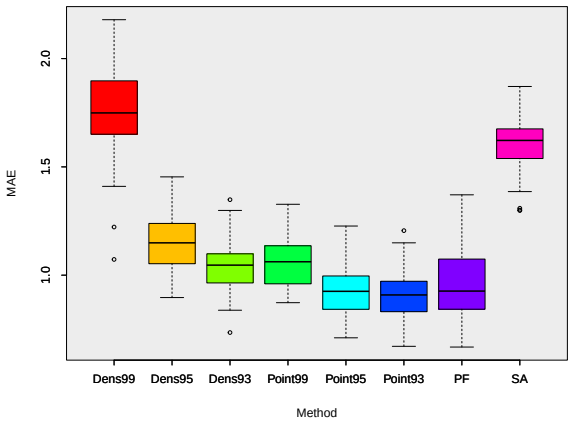
Figure 4.3.6: Comparison of performance metrics across replications for the different methods implemented for the period **before** the missing data, featuring: point forecast combination with $\delta = 0.93, 0.95, 0.99$, point forecast combination with adaptive discounting, and density combination with $\alpha = 0.93, 0.95, 0.99$. For the MSE, MAE and SMAPE, simple averaging has also been included. For the log-likelihood, simple average density has been included.

likelihoods, for the different methods. For this time period, results are largely similar to those obtained in the sporadic missing data simulation. The lowest median MSE, MAE and SMAPE are obtained by the DLM-based point forecast combination method with adaptive discounting, and the density combination methods provide the greatest log-likelihoods.

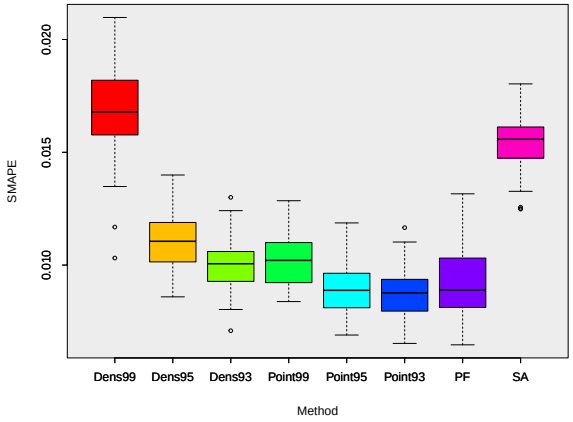
Consideration of the forecasting performance during and after the missing data



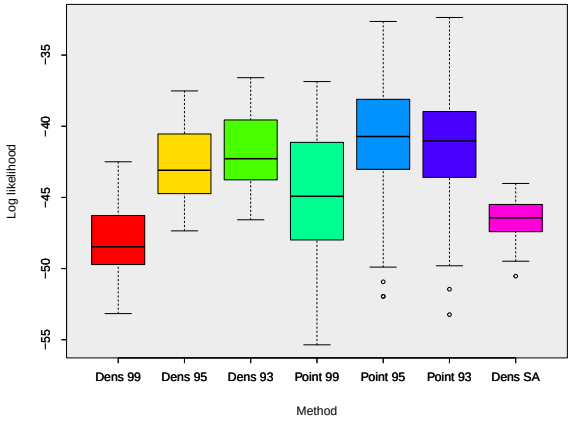
(a) Comparison of MSE across replications.



(b) Comparison of MAE across replications.



(c) Comparison of SMAPE across replications.



(d) Comparison of log-likelihood across replications.

Figure 4.3.7: Comparison of performance metrics across replications for the different methods implemented for the period **during** the missing data, featuring: point forecast combination with $\delta = 0.93, 0.95, 0.99$, point forecast combination with adaptive discounting, and density combination with $\alpha = 0.93, 0.95, 0.99$. For the MSE, MAE and SMAPE, simple averaging has also been included. For the log-likelihood, simple average density has been included.

period is more interesting. Figure 4.3.7 displays box and whisker plots across the 50 replications for the period **during** the missing data, comparing (a) MSEs, (b) MAEs, (c) SMAPEs and (d) log-likelihoods. Immediately, we observe that the density combination approach with $\alpha = 0.99$, and taking the simple average of the point forecasts, perform the worst in terms of the error metrics. We expect that the unfavourable performance of the density combination method with this choice of forgetting factor is due to the

fact that the mixture weight did not fall to zero in the missing data period. Likewise, we expect that simple averaging places too much weight on the missing forecaster, leading to unsatisfactory forecasting performance. The smallest median MSE across replications is achieved by our DLM-based point forecast combination method with $\delta = 0.93$ (1.326), closely followed by the adaptive discounting approach (1.345). We surmise that this is due to the fact that these methods allow the weight of Forecaster 1 to decrease quickly in the missing period. Similar behaviour is shown for the MAEs and SMAPEs.

Figure 4.3.7d shows the log-likelihoods for the different methods during the missing data period. Unlike in the sporadic missing data study, wherein the density combination methods always achieved greater log-likelihoods than the point combination methods, this is not the case here. We see that density combination with $\alpha = 0.99$ exhibits significantly unfavourable performance, once again likely on account of the slow decrease in mixture weight.

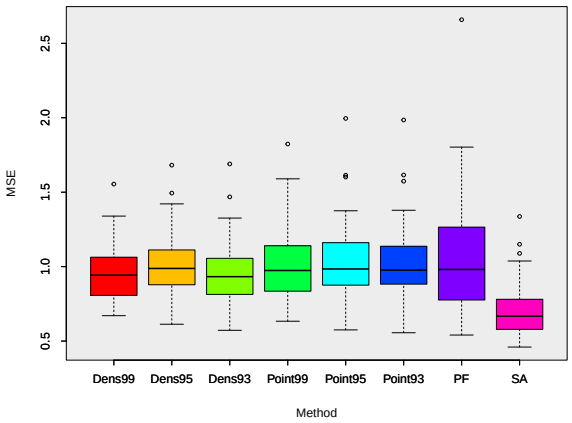
Figure 4.3.8 compares forecasting performance **after** the missing data period. Recall, the density combination methods allowed the weights to return to their pre-missing data levels at a much faster rate than their point combination counterparts. Consequently, we assumed that these methods would display superior forecasting performance in this period. This is indeed shown to be the case, as the median MSE across replications for the density combination method with $\alpha = 0.93$, and $\alpha = 0.99$, is lower than that of all of the DLM-based point forecast combination methods. This is particularly notable, since the density combination methods do not aim to minimise the MSE. In terms of MAE and SMAPE, all three density combination methods outperform the DLM-based point forecast combination methods across replications. We also note that the interquartile ranges are narrower for the density combination methods than the DLM-based point forecast combination methods, implying that these methods display more robust performance after the missing period. We do acknowledge that in this

simulation setting, taking the simple average point forecast combination outperformed all other methods in terms of MSE, MAE and SMAPE. We expect that this is due to the fact that this method assigns a significant weight of 1/3 to Forecaster 1 immediately as it returns online. Given that the quality of Forecaster 1 did not decrease during the missing period, this leads to improved forecasting performance. On the other hand, our methods were designed such that a forecaster is required to 'earn' their new weighting after a period of missing data. Hence, we expect that our combination methods would outperform simple averaging if the quality of Forecaster 1 decreased in the missing period. The greatest log-likelihood is obtained by the density combination with $\alpha = 0.93$, closely followed by the density combination with $\alpha = 0.95$. Once again, the simple average density benchmark did not exhibit comparably favourable performance.

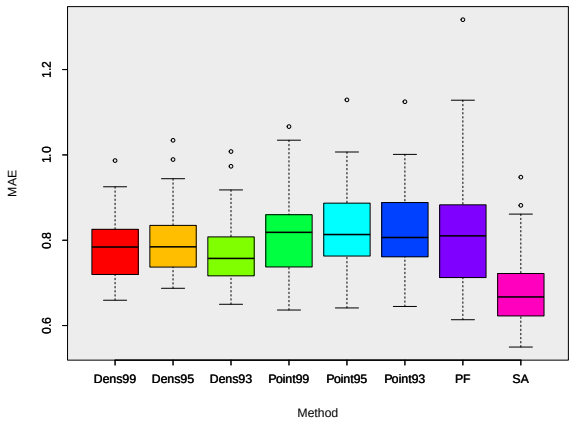
4.4 Discussion

In this chapter, we have proposed new dynamic forecast combination techniques, and carried out a conceptual study of the often encountered problem of missing forecaster data. Although forecast combination with unbalanced panels has been examined to a limited extent in the literature, methods for dealing with missing forecasters often resort to trimming those which fail to provide predictions for a sufficient number of time points. This leads to a loss of potentially useful information, and therefore some works instead consider imputing missing data by back-filling. However, such methods are unsuitable for use in online settings, where forecasters arrive sequentially prior to the observed data point y_t, and imputation methods often differ. Furthermore, the unbalanced panel literature mainly considers the case that forecasters fail to provide predictions on a sporadic basis, rather than for extended numbers of observations.

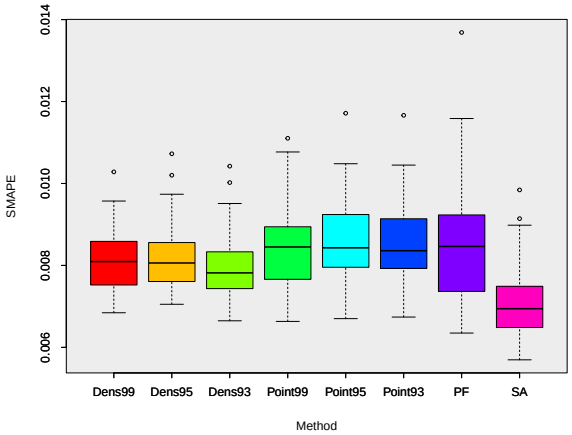
In order to deal with missing forecaster data, we have proposed modelling each individual forecaster using a DLM, where the particular DLM employed is chosen to



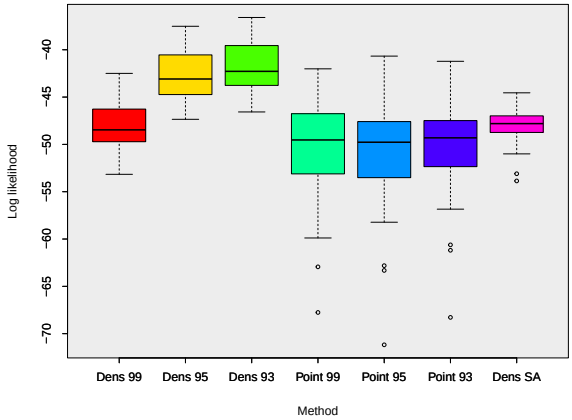
(a) Comparison of MSE across replications.



(b) Comparison of MAE across replications.



(c) Comparison of SMAPE across replications.



(d) Comparison of log-likelihood across replications.

Figure 4.3.8: Comparison of performance metrics across replications for the different methods implemented for the period **after** the missing data, featuring: point forecast combination with $\delta = 0.93, 0.95, 0.99$, point forecast combination with adaptive discounting, and density combination with $\alpha = 0.93, 0.95, 0.99$. For the MSE, MAE and SMAPE, simple averaging has also been included. For the log-likelihood, simple average density has been included.

reflect the structure of the forecaster series. In our simulation studies, we simulated such that forecasters could be modelled with an AR (1) plus noise DLM, but we note that this methodology can be applied to any form of time series for which a DLM can be defined. To the best of our knowledge, an investigation of the effects of imputing missing observations in this way has not been considered in the forecast combination setting before.

This method enables forecasting distributions to be computed for each forecaster at each time point, even in the case that a forecaster failed to provide a prediction; point forecasts can be obtained by taking the expectation of the distribution. Since using a DLM allows the temporal evolution of the forecaster series to be modelled, we expect the imputed values to be more accurate than those which would be obtained by a simple 'fill forwards' approach. Furthermore, due to the sequential nature of DLMs, this method allows missing forecasts to be imputed in an online framework, such that the newly imputed forecasts can be combined in a similarly online manner. By implementing this consistent approach to dealing with missing forecasts, we negate any issues in the comparison of methods that may arise due to differences in implicit imputations (see Lahiri et al. (2017)), and allow all potentially useful historical data to be included in our combinations.

In addition to providing a method for dealing with missing forecasts, we also discussed how modelling each individual forecaster as a DLM before the combination takes place enables their associated uncertainty to be quantified. This direct interpretation of forecaster uncertainty (as opposed to the indirect interpretation one might achieve from analysis of combination weights) may be of value to decision makers, in order to enable more informed decisions. When missing data does occur, modelling each forecaster as a DLM leads to an increase in the variance of the distribution of the missing forecaster, as shown in Figure 4.2.1.

Despite the fact that this imputation method offers a more accurate description of the temporal evolution of the forecast series than simply filling forward, we acknowledge that the quality of the imputations will likely be far lower than the quality of the observed forecasters. This means that a sudden change in forecaster quality is likely to be experienced when a forecaster goes offline, or similarly returns. To begin, we considered how to account for such sudden changes in quality by extending the DLM-based combination procedure, introduced in Chapter 3. Recall, the value of discount

factor δ in our method influences the stochastic variation of the combination weights throughout time. Therefore, in order to respond to sudden changes in forecaster quality, we would like to adopt a lower value of discount factor δ. To this end, we synthesised an existing particle-learning-based method for adaptive parameter selection with our DLM-based combination procedure, in order to create a DLM-based point forecast combination method with adaptive discounting. The new methodology can be used to combine point forecasts, without the need for a user specified discount factor δ.

Furthermore, we also proposed a novel method for dynamically combining density forecasts, based on maximising the exponentially weighted log predictive score at each time point. Since modelling a missing forecaster using a DLM results in an increased variance for the missing period, application of our proposed density combination technique allows this information to be directly fed into the finite mixture weights.

In our simulation studies, we considered two different types of missingness: sporadic missing data and a large period of missing data. In the former case, our DLM-based point forecast combination with adaptive discounting was shown to perform well in terms of MSE, MAE and SMAPE. This implies that suitable choices of δ_t were selected at each time. On the other hand, all DLM-based point forecast combination methods exhibited poor performance in the large period of missing data case, for the period after Forecaster 1 returned online. From analysis of the combination weights, we suspect that this is due to the slow rate at which the weights adapt after this period.

Given that the particle-learning method of Irie et al. (2022) is designed to handle abrupt changes in the value of δ_t, we were surprised by this poor performance. The unresponsiveness of the observed weights is due to the particle-learning approach selecting very high values of δ after the missing period. One possible reason for this is that the increase in the likelihood of observing the data y_t from assigning a lower value of δ_t is not significant enough to cause the particles to sample more of these lower values. It is possible that this is a consequence of the forecast combination setting, where the gains

in likelihood achieved from assigning different combination weights are low, since the quality of the forecasters are so similar. Perhaps if the forecasters had been defined with more distinct qualities, this method would have assigned lower values of δ_t after the missing period. Also, the hyperpriors selected for the method give rise to higher values of δ since these are often most suitable. It is possible that the distributions of the hyperparameters have become too narrow, meaning that it is extremely unlikely to sample a low value of δ_t.

On the other hand, the density combination approach with $\alpha = 0.93$ displayed superior performance after the large missing data period for all metrics. We believe this is due to the fact the mixture weights could evolve quickly through time, a direct consequence of the decrease in variance of the forecasting distribution after the missing period.

From our simulations, we conclude that the choice of which method is 'best' is informed by the particular application. Our DLM-based point forecast combination method with adaptive discount factor works well in the presence of sporadic missing data. On the other hand, our dynamic density combination method displayed superior performance after a forecaster had been offline for a long period of time, and we would recommend applying this method (with a relatively low value of forgetting factor such as $\alpha = 0.93$) in such situations. We would particularly advise against simple average point combination, as this performed especially poorly during the missing data period as a consequence of placing weights that are too large on the imputed values.

To conclude this chapter, we acknowledge some disadvantages of the proposed methods. Clearly, our DLM-based point forecast combination procedure with adaptive discounting greatly increases the computational complexity of the problem. Although an obvious benefit of this approach is the fact no discount factor needs to be specified, the improvements in MSE, MAE and SMAPE are perhaps not enough to warrant the increased computational time, especially if working in a high frequency setting.

In terms of our proposed density combination method, a clear challenge is appropriate selection of the forgetting factor α. We saw in our simulations that the choice of α greatly impacts forecasting performance in periods of missing forecaster data, and therefore choosing a suitable value for this parameter is no trivial task. We note that the benefits obtained from density combination after a large period of missing data appear to arise due to the fact that the uncertainty of the forecaster can be directly taken into account, therefore allowing weights to adapt quicker to a forecaster returning online than in the point combination approach. With this in mind, we expect that other density combination approaches for time varying weights would also show an improvement on the point forecast combination framework in this setting, such as dynamic model averaging (DMA) as proposed by Raftery et al. (2010). However, this method is also dependent on an appropriate choice of discount parameter. To mitigate these issues, one could investigate the loss discounting framework of Bernaciak and Griffin (2024) in the missing data setting. This is left to further research.

Chapter 5

The 'polyfoil' stochastic process

In this chapter, we propose a simple stochastic process for modelling interacting oscillations at different frequencies in multivariate time series. We term this the 'polyfoil' process, due to the shapes depicted when such time series are plotted in the complex plane. We show that nonstationarity can be induced by 'locking' the phase difference between the oscillations under expectation at general time t. Phase-locking the oscillations in this way leads to interactions between frequencies, and is motivated by applications where non-zero coherence values exist between distinct frequencies. For example, high coherences are often observed in the modelling of ocean waves (see Chave et al. (2019)) and the Earth's geomagnetic field (see Riegert and Thomson (2018)).

5.1 Introduction

The autoregressive process of order 1 (AR(1)) provides an effective model for describing the evolution of random processes in a wide variety of applications; indeed, we used this model in Chapter 3 to describe the evolution of the underlying time series of interest $\tilde{y}_t$. In this chapter, we shall use notation from the stochastic processes literature, and

define the AR(1) model as,

$$X_t = \phi X_{t-1} + \epsilon_t, \quad \epsilon_t \overset{iid}{\sim} (0, \sigma_\epsilon^2).$$

It is sometimes the case that we do not wish to model the evolution of a single time series X_t, but rather the interactions between a set of variables collected in an $p \times 1$ vector, $\mathbf{X}_t$. In such cases, one may apply the first order vector autoregressive model (VAR(1)), defined as,

$$\mathbf{X}_t = \mathbf{\Phi}\mathbf{X}_{t-1} + \boldsymbol{\epsilon}_t,$$

where $\mathbf{\Phi}$ an $p \times p$ matrix of autoregressive coefficients, and $\boldsymbol{\epsilon}_t$ is a vector generalisation of white noise. If we consider just the first time series in the vector, $X_{1,t}$, this can be written as,

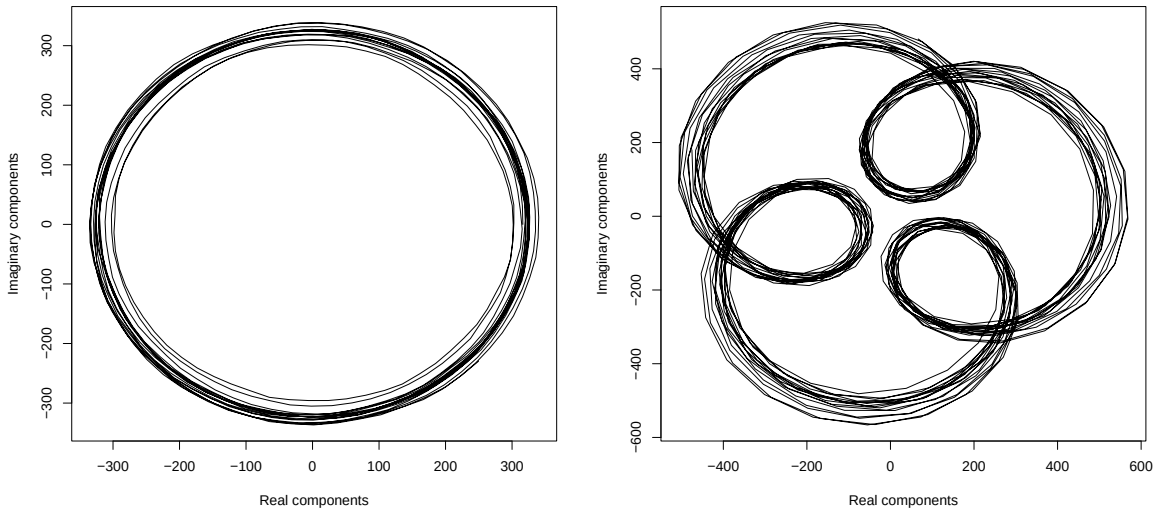
$$X_{1,t} = \phi_{1,1}X_{1,t-1} + \phi_{1,2}X_{2,t-1} + \cdots + \phi_{1,p}X_{p,t-1} + \epsilon_{1,t}, \quad \epsilon_{1,t} \overset{iid}{\sim} (0, \sigma_\epsilon^2).$$

Therefore, in a VAR(1) series each variable is regressed onto its own previous value, in addition to the values of the other $p-1$ series at the previous time point; see Hamilton (1994) for more details. In this chapter, we consider time series which can be described by combinations of bivariate AR(1) processes. That is, we assume that two univariate time series, X_t and Y_t, are observed simultaneously, and we can model these according to,

$$\begin{aligned} X_t &= a\cos(\theta)X_{t-1} - a\sin(\theta)Y_{t-1} + \epsilon_t^{\{1\}}, \quad \epsilon_t^{\{1\}} \overset{iid}{\sim} (0, \sigma_\epsilon^2), \\ Y_t &= a\cos(\theta)Y_{t-1} + a\sin(\theta)X_{t-1} + \epsilon_t^{\{2\}}, \quad \epsilon_t^{\{2\}} \overset{iid}{\sim} (0, \sigma_\epsilon^2). \end{aligned} \tag{5.1.1}$$

When the parameter $|a| < 1$, the above model describes a stationary process, which we shall assume in this chapter. Furthermore we constrain $\theta \in [-\pi, \pi)$ for identifiability due to the 2π-periodicity of the trigonemetric functions.

The bivariate representation is useful for practitioners, as it provides a convenient interpretation of the generating process for the time series. However, it is possible to instead consider the series as a complex AR(1) process, such that the real components are provided by the series X_t, and the imaginary components are provided by the series Y_t. This is a popular choice in many applications, including oceanography (see Gonella (1972)) and magnetic resonance imaging (see Rowe (2005)). As stated by Sykulski et al. (2016), this is due to the 'compactness and interpretability' of the complex representation. In particular, the complex-valued representation provides a more suitable framework for the study of signals where X and Y represent orthogonal components of the same process, for example westerly and northerly winds and ocean currents (see Gonella (1972) and Sykulski et al. (2017)).



(a) Circular oscillation with frequency $\theta = 0.1$.

(b) Coupled circular oscillations with frequencies $\theta = 0.1$ and $h\theta = 0.4$.

Figure 5.1.1: Simulated bivariate AR(1) series and polyfoil series of length $N = 1000$, plotted in the complex plane. Damping parameter set to $a = 0.99999$ for visualisation, and $\sigma_\epsilon^2 = 0.1$.

When it is assumed that the real and imaginary components of the noise are independent and identically distributed, as in equation (5.1.1), plotting such a bivariate AR(1) process in the complex plane depicts circular oscillations. This is shown in Figure 5.1.1a, where we have assumed Gaussian distributions for the noise terms and the value of parameter a has been set very close to one for visualisation purposes. One can think of such a series as oscillating at a given frequency, θ. In the signal processing community, such circular time series are often called 'proper' time series, for example see Schreier and Scharf (2010).

In practice, it is often the case that we observe not just one signal oscillating at a given frequency θ, but rather multiple oscillatory signals at different frequencies, which may also include harmonic frequencies; for example, see Elipot et al. (2016) and Sykulski et al. (2015). In such situations, it is desirable to study not just the frequencies of the individual oscillations, but also the interactions between different pairs of frequencies (if these exist). Coupling between frequency bands is observed in many applications, including electroencephalography (EEG) data (see Olhede and Ombao (2013)), and the Earth's geomagnetic field (see Riegert and Thomson (2018)). In this chapter, we shall consider a specific case of harmonic oscillations where an oscillation of frequency θ is jointly observed with an oscillation of frequency $h\theta$, where h is an integer. In this setting, we term the former oscillation the fundamental oscillation, and the latter the harmonic. Similarly, let us refer to θ and $h\theta$ as the fundamental and harmonic frequencies, respectively.

We propose a stochastic process for modelling two interacting oscillatory signals by superposing the fundamental oscillation with the harmonic. When realisations of this new oscillatory process are plotted in the complex plane, a 'polyfoil' shape is depicted; see Figure 5.1.1b for an example, where once again we have set the parameter a very close to one for visualisation purposes. We refer to such shapes as polyfoils as they are reminiscent of the trefoil and quatrefoil shapes commonly found in Gothic architecture,

whose form depicts overlapping circles of the same size. Such shapes can also be referred to as epicyloids, a geometric term first used in physics to help describe planetary motion under the geocentric model of the universe.

In this chapter, we will show that enforcing a 'locked' phase difference between the two oscillations, under expectation at general time t, allows us to construct a specific form of nonstationarity for our process. This distinguishes our research from other work on stochastic oscillations in the literature, where only a stationary oscillatory process is considered (for example, see the widely linear complex AR(1) process of Sykulski et al. (2016), which traces out elliptical, but stationary, oscillations in the complex plane).

Such polyfoil shapes have been observed in nature. Recently, Zheng et al. (2024) discussed the underlying cause of the 'flower-like' trajectories observed for drifting buoys at the surface of the ocean. The trajectories demonstrated in this paper exhibit a significant resemblance to the depictions of our polyfoil process (see Figure 5.1.1b and Figure 5.1.2). This chapter provides the first time series stochastic process model, to our knowledge, for generating such oscillatory patterns observed in nature.

The remainder of this chapter is organised as follows. We begin by formally constructing the stochastic polyfoil process, and show how this is a four-dimensional VAR(1) time series model. Following this, we provide some key background on the spectral representation of nonstationary processes in Section 5.3, before deriving key theoretical properties for the polyfoil model in Section 5.4. Finally, the theoretical results are compared with simulations in Section 5.5, and further work is discussed in Section 5.6.

5.2 Construction of the polyfoil process

The polyfoil process is constructed as a simple extension to the complex-valued AR(1) process, which is defined subsequently. To begin, let us define the fundamental oscilla-

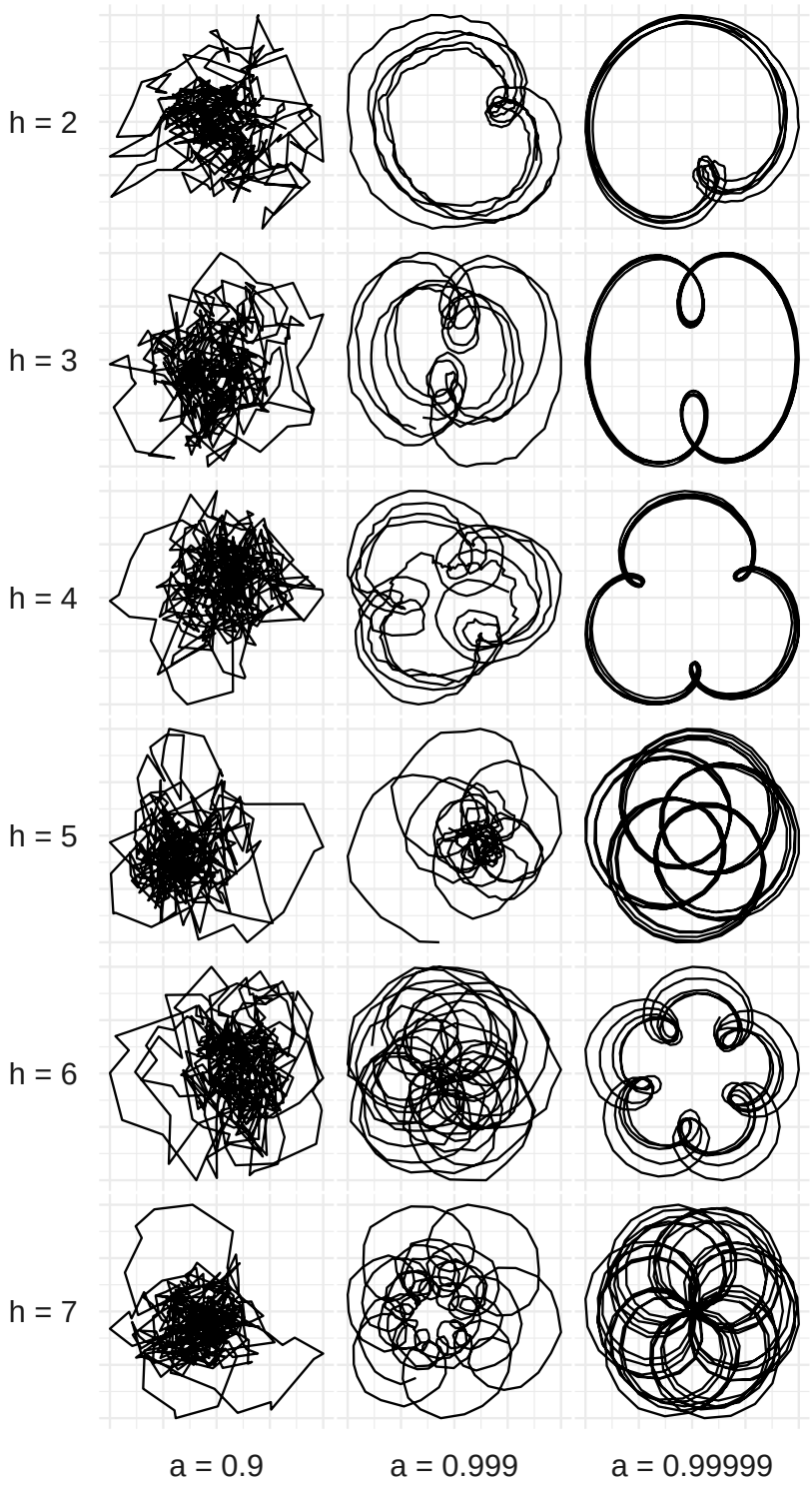


Figure 5.1.2: Simulated polyfoil shapes, corresponding to time series of length $N = 500$, for different values of damping parameter a and harmonic multiple h, for fixed noises variances and fundamental frequency. Fundamental frequency set to $\theta = 0.01(2\pi)$ for 'smooth' shapes, and noise variances set to $\sigma_\epsilon^2 = 0.1$ and $\sigma_\nu^2 = 0.05$. Real and imaginary components normalised.

tion in terms of a bivariate AR(1) process,

$$X_t^{\{1\}} = a\cos(\theta)X_{t-1}^{\{1\}} - a\sin(\theta)Y_{t-1}^{\{1\}} + \epsilon_t^{\{1\}}, \quad \epsilon_t^{\{1\}} \sim N(0, \sigma_\epsilon^2),$$
$$Y_t^{\{1\}} = a\cos(\theta)Y_{t-1}^{\{1\}} + a\sin(\theta)X_{t-1}^{\{1\}} + \epsilon_t^{\{2\}}, \quad \epsilon_t^{\{2\}} \sim N(0, \sigma_\epsilon^2),$$

where $|a| < 1, \theta \in [-\pi, \pi), \sigma_\epsilon^2 > 0$ and the superscript $\{1\}$ denotes this is the fundamental oscillation; the harmonic oscillation will subsequently have the superscript $\{2\}$. The constraint that $|a| < 1$ has been imposed in order to ensure stationarity, and the noise terms are independent of one another. Note now that we are also assuming Gaussian noise to simplify subsequent theoretical derivations, but this assumption can be relaxed with minor modification. As discussed in Section 5.1, realisations of this process depict circular oscillations when plotted in the complex plane, i.e. $Z_t^{\{1\}} = X_t^{\{1\}} + iY_t^{\{1\}}$ where $i = \sqrt{-1}$.

By defining the rotation matrix $R^{\{1\}}$,

$$R^{\{1\}} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix},$$

one may express the bivariate AR(1) series in vector form as follows,

$$\begin{bmatrix} X_t^{\{1\}} \\ Y_t^{\{1\}} \end{bmatrix} = aR^{\{1\}} \begin{bmatrix} X_{t-1}^{\{1\}} \\ Y_{t-1}^{\{1\}} \end{bmatrix} + \begin{bmatrix} \epsilon_t^{\{1\}} \\ \epsilon_t^{\{2\}} \end{bmatrix}.$$

The complex representation of the process is therefore given by,

$$\begin{aligned} Z_t^{\{1\}} &= (a\cos(\theta)X_{t-1}^{\{1\}} - a\sin(\theta)Y_{t-1}^{\{1\}} + \epsilon_t^{\{1\}}) + i(a\cos(\theta)Y_{t-1}^{\{1\}} + a\sin(\theta)X_{t-1}^{\{1\}} + \epsilon_t^{\{2\}}), \\ &= (X_{t-1}^{\{1\}} + iY_{t-1}^{\{1\}})(a\cos\theta + ia\sin\theta) + \epsilon_t^{\{1\}} + i\epsilon_t^{\{2\}}, \\ &= ae^{i\theta}Z_{t-1}^{\{1\}} + \nu_t^{\{1\}}, \end{aligned} \tag{5.2.1}$$

where $\{\nu_t^{\{1\}}\}$ is a sequence of independent and identically distributed complex-valued Gaussian noise with variance $2\sigma_\epsilon^2$. Thus, the complex autoregressive coefficient is expressed in terms of an amplitude $|a| < 1$, and phase $\theta \in [-\pi, \pi)$. The phase θ defines the angle of rotation of the process at each time step, and is often termed the spin parameter. The harmonic oscillation is defined in a similar way, with the key difference being the new frequency $h\theta$,

$$X_t^{\{2\}} = a\cos(h\theta)X_{t-1}^{\{2\}} - a\sin(h\theta)Y_{t-1}^{\{2\}} + \zeta_t^{\{1\}}, \quad \zeta_t^{\{1\}} \sim N(0, \sigma_\zeta^2),$$
$$Y_t^{\{2\}} = a\cos(h\theta)Y_{t-1}^{\{2\}} + a\sin(h\theta)X_{t-1}^{\{2\}} + \zeta_t^{\{2\}}, \quad \zeta_t^{\{2\}} \sim N(0, \sigma_\zeta^2),$$

where we constrain $|a| < 1, h \in \mathbb{Z}, h\theta \in [-\pi, \pi), \sigma_\zeta^2 > 0$. Note that the parameter a is common to both the fundamental and harmonic oscillations, but this assumption can be relaxed in extensions of the model. The complex representation of the harmonic process is therefore given by,

$$Z_t^{\{2\}} = ae^{ih\theta}Z_{t-1}^{\{2\}} + \nu_t^{\{2\}}, \tag{5.2.2}$$

where $\{\nu_t^{\{2\}}\}$ is a sequence of complex-valued Gaussian noise with variance $2\sigma_\zeta^2$.

We construct the polyfoil stochastic process by superposing the two oscillatory signals together, in order to define a new series Z_t,

$$Z_t = Z_t^{\{1\}} + Z_t^{\{2\}}.$$

The stochastic polyfoil is a five-parameter model and the particular shape depicted by realisations of the polyfoil process depend on the choice of the parameters $\{a, \theta, h, \sigma_\epsilon^2, \sigma_\zeta^2\}$ in equations (5.2.1) and (5.2.2), where we emphasise the constraint $|a| < 1$ for stationarity of each component, and $h\theta \in [-\pi, \pi)$ for identifiability due to 2π-periodicities. When the fundamental frequency θ is small $|\theta| << \pi$, this implies that the oscillation

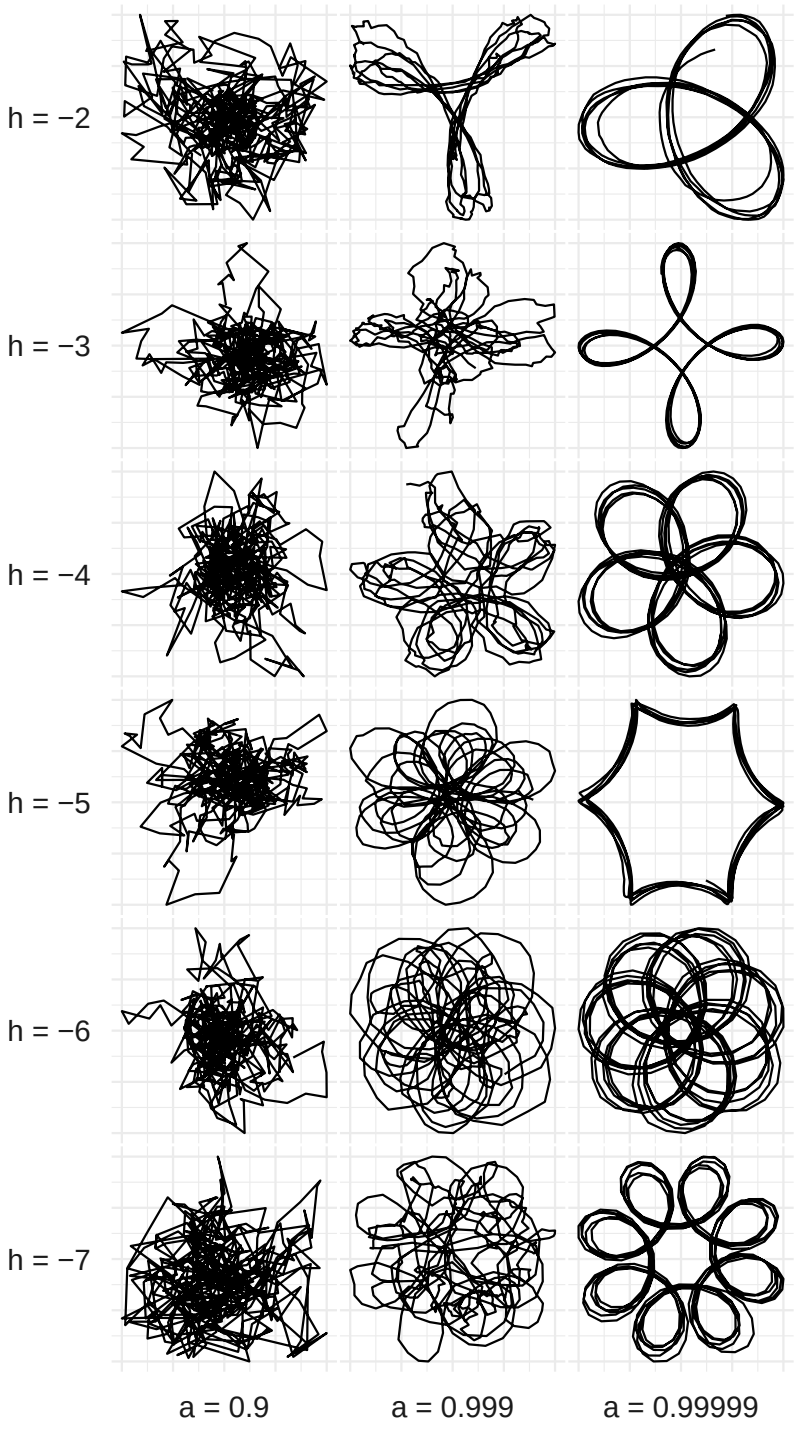


Figure 5.2.1: Simulated polyfoil shapes, corresponding to time series of length $N = 500$, for different values of damping parameter a and negative harmonic multiple h, for fixed noises variances and fundamental frequency. Fundamental frequency set to $\theta = 0.01(2\pi)$ for 'smooth' shapes, and noise variances set to $\sigma_\epsilon^2 = 0.1$ and $\sigma_\nu^2 = 0.05$. Real and imaginary components normalised to have comparable magnitude across polyfoils.

completes a small fraction of a rotation at each time t, and therefore the subsequent polyfoil is 'smooth'. This is shown in Figure 5.1.1b, where $\theta = 0.1$ and the polyfoil displays curved edges. This is also the case in Figure 5.1.2 and Figure 5.2.1, where the fundamental frequency has been set to $\theta = 0.01(2\pi)$. This implies that the process completes 1/100th of an oscillation at every time step, leading to the exhibited 'smooth' shapes. On the other hand, a larger choice of fundamental frequency leads to more angular polyfoils. The harmonic multiple h determines the number of 'loops' in the polyfoil (see Figure 5.1.2 for examples of polyfoils with positive harmonic frequency $h > 0$, and Figure 5.2.1 for examples with negative harmonic frequency $h < 0$); specifically, this is given by $|h - 1|$.

The shape of the polyfoil is also heavily influenced by the parameter a, which can be thought of as a damping parameter. When this parameter is very close to one, very little damping occurs, and as such the resulting polyfoil shape is almost completely deterministic. This is demonstrated in Figure 5.1.2 and Figure 5.2.1, where polyfoils become less stochastic as we increase the value of a. This is useful for visualisation purposes; however, such highly deterministic systems are unlikely to be found repeatedly in nature. We see that for lower values of damping parameter, simply viewing the trajectory of series does not provide sufficient information about the characteristics of the generating mechanism. For example, it is clearly impossible to determine the value of h by eye from the plots with $a = 0.9$ in Figure 5.1.2. Moreover, it is not possible to establish by eye whether or not the process is stationary. Thus, more sophisticated analysis techniques must be employed. In the following section, we introduce some key background on the spectral representation of nonstationary processes.

5.3 Background

Let $\{X_t\}_{t\in\mathbb{Z}}$ denote a discrete-time real-valued stochastic process, and let $F_X(X_{t_1}, X_{t_2}, \ldots, X_{t_N})$ denote the cumulative distribution function of the joint distribution of a length-N process $(X_{t_1}, X_{t_2}, \ldots, X_{t_N})$. The stochastic process is then said to be strictly stationary if we have,

$$F_X(X_{t_1}, X_{t_2}, \ldots, X_{t_N}) = F_X(X_{t_1+\tau}, X_{t_2+\tau}, \ldots, X_{t_N+\tau}),$$

for all $\tau, t_1, \ldots, t_N \in \mathbb{Z}$ and $N \in \mathbb{N}_0$. Second-order stationarity describes a weaker form of stationarity, wherein the mean of the distribution is constant throughout time and the covariance between two time points depends only on the lag τ, and not the time t.

Let us consider a mean zero time series X_t that may or may not be stationary. The dual-time autocovariance function of X_t is given by,

$$s(t, \tau) = \mathrm{Cov}(X_t, X_{t-\tau}) = E\{X_t X_{t-\tau}\};$$

see Olhede and Ombao (2013). If X_t is second order stationary, then $s(t, \tau)$ takes the simpler form of $\tilde{s}(|\tau|)$. It is often beneficial to express X_t in the frequency domain rather than the time domain. This can be done using the Cramér representation, given by,

$$X_t = \int_{-1/2}^{1/2} e^{i2\pi f t} dZ(f),$$

where $\{dZ(f)\}$ is a zero-mean orthogonal increments process. For stationary processes, let us define the spectrum $\tilde{S}(f)$ as the Fourier transform of the autocovariance sequence $\tilde{s}(|\tau|)$, where $f \in [-1/2, 1/2]$. The spectrum is then equal to the variance of the orthogonal increments process, $\mathrm{Var}(dZ(f)) = \tilde{S}(f)$. The spectrum provides a full representation for Gaussian stationary processes; however, for most time series, it does

not fully embody all relevant characteristics. For example, nonstationarity often leads to the presence of correlations between distinct frequencies, which are not captured by the spectrum. In order to characterise these correlations, Olhede and Ombao (2013), use the concept of Loève coherence.

Olhede and Ombao (2013) use the formalism of harmonisable processes in order to define the Loève coherence. Such processes describe time series which are generated by the superposition of random infinitesimal harmonic oscillators. The spectral representation of harmonisable processes is given by,

$$X_t = \int_{-1/2}^{1/2} e^{i2\pi ft} dZ(f),$$

where the increment random process $dZ(f)$ has zero mean and satisfies,

$$\mathrm{Cov}(dZ(f_1)dZ(f_2))) = E\{dZ(f_1)dZ^*(f_2)\} = S(f_1, f_2)df_1df_2.$$

The quantity $S(f_1, f_2)$ denotes a complex scalar value, referred to as the Loève spectrum, or dual-frequency spectrum, with $f_1, f_2 \in [-1/2, 1/2]$. This is related to the autocovariance of X_t by,

$$\begin{aligned} s(t,\tau) =& \mathrm{Cov}(X_t, X_{t-\tau}), \\ =& \int_{-1/2}^{1/2} S(f_1, f_2) e^{2i\pi[(f_1-f_2)t+f_2\tau]} df_1 df_2, \end{aligned} \tag{5.3.1}$$

and thus the Loève spectrum can be considered the Fourier pair of the nonstationary autocovariance function, just as the spectrum is the Fourier pair of the stationary autocovariance function. From the Loève spectrum it is possible to calculate the Loève

coherency,

$$\begin{aligned}\tau(f_1, f_2) =& \frac{S(f_1, f_2)}{\sqrt{S(f_1, f_1)S(f_2, f_2)}} \\ =& \rho^{1/2}(f_1, f_2)e^{-i\phi(f_1, f_2)}.\end{aligned} \tag{5.3.2}$$

Here, $\rho(f_1, f_2)$ denotes the Loève coherence at the frequency pair (f_1, f_2), given by $\rho(f_1, f_2) = |\tau(f_1, f_2)|^2$, and $\phi(f_1, f_2)$ denotes the corresponding Loève coherency phase. These quantities can be used to describe cross-dependencies between different frequencies in a nonstationary signal.

5.4 Theory

In this section we derive key theoretical results for the stochastic polyfoil process. We begin by presenting known results, including the autocovariance sequence for the stationary complex AR(1) process, and later use these to derive the autocovariance function for the polyfoil process. We demonstrate that the polyfoil process can be made nonstationary when the fundamental and harmonic oscillations are coupled in a particular way by locking their phase relationship. For the nonstationary polyfoil process, we derive the Loève spectrum and corresponding Loève coherency.

5.4.1 The complex AR(1) process

The polyfoil stochastic process is constructed by superposing two complex AR(1) processes, representing a fundamental oscillation and harmonic oscillation. When viewed in the complex plane, each process individually depicts circular oscillations, resulting from the fact that the real and imaginary components of the noise are assumed to be independent and identically distributed. The autocovariance sequence for such processes is already known and defined; we provide this here in the context of the fundamental

oscillation for reference in later work. We note that all proofs in this section can be easily extended to the harmonic oscillation by replacing the fundamental frequency θ with the harmonic frequency $h\theta$, and the noise term $\nu_t^{\{1\}}$ with $\nu_t^{\{2\}}$.

The complex representation of the fundamental oscillation is written as,

$$Z_t^{\{1\}} = ae^{i\theta} Z_{t-1}^{\{1\}} + \nu_t^{\{1\}}, \tag{5.4.1}$$

where $\nu_t^{\{1\}}$ is complex-valued Gaussian white noise. The autocovariance sequence for a complex-valued stationary process is given by $s_{t,\tau} = E\{Z_t Z_{t-\tau}^*\}$, and is independent of t. For the complex AR(1) the autocovariance is given in the following theorem.

Theorem 5.4.1. *For all lags $\tau \in \mathbb{Z}$, the autocovariance sequence $s_{t,\tau}^{\{1\}}$ of the fundamental oscillation described in equation (5.4.1) is given by,*

$$s_{t,\tau}^{\{1\}} = a^{|\tau|} e^{i\theta\tau} s_0^{\{1\}},$$

where $s_0^{\{1\}} = 2\sigma_\epsilon^2/(1-a^2)$ is the variance of the fundamental oscillation.

Proof. For non-negative lag τ, the covariance sequence $s_{t,\tau}^{\{1\}}$ is given by,

$$\begin{aligned}
s_{t,\tau}^{\{1\}} &= E\{Z_t^{\{1\}} Z_{t-\tau}^{\{1\}*}\} \\
&= E\{(ae^{i\theta} Z_{t-1}^{\{1\}} + \nu_t^{\{1\}}) Z_{t-\tau}^{\{1\}*}\} \\
&= ae^{i\theta} E\{Z_{t-1}^{\{1\}} Z_{t-\tau}^{\{1\}*}\} + E\{\nu_t^{\{1\}} Z_{t-\tau}^{\{1\}*}\} \\
&= ae^{i\theta} E\{Z_{t-1}^{\{1\}} Z_{t-\tau}^{\{1\}*}\} \\
&= ae^{i\theta} s_{\tau-1}^{\{1\}} \\
&= a^\tau e^{i\theta\tau} s_0^{\{1\}}.
\end{aligned}$$

Here, $s_0^{\{1\}}$ denotes the variance of the process, which can be found as follows. Assuming

stationarity, such that $\text{Var}(Z_t^{\{1\}}) = \text{Var}(Z_{t-1}^{\{1\}})$, we can write,

$$\begin{aligned} s_0^{\{1\}} = \text{Var}(Z_t^{\{1\}}) &= \text{Var}(ae^{i\theta} Z_{t-1}^{\{1\}} + \nu_t^{\{1\}}), \\ &= \text{Var}(ae^{i\theta} Z_{t-1}^{\{1\}}) + \text{Var}(\nu_t^{\{1\}}), \\ &= a^2 \text{Var}(Z_{t-1}^{\{1\}}) + 2\sigma_\epsilon^2, \\ &= a^2 s_0^{\{1\}} + 2\sigma_\epsilon^2, \end{aligned}$$

such that

$$s_0^{\{1\}} = 2\sigma_\epsilon^2/(1-a^2),$$

as required. The sequence for negative lags is then given by the simple relationship $s_{-\tau} = s_\tau^*$, where the superscript denotes the complex conjugate, such that we have $s_{t,\tau}^{\{1\}} = a^{|\tau|} e^{i\theta\tau} s_0^{\{1\}}$, for all $\tau \in \mathbb{Z}$. □

The autocovariance sequence for the harmonic oscillation can be derived in a similar way and is given in the following corollary.

Corollary 5.4.2. *The autocovariance sequence for the harmonic oscillation is given by,*

$$s_{t,\tau}^{\{2\}} = a^{|\tau|} e^{ih\theta\tau} s_0^{\{2\}},$$

where $s_0^{\{2\}} = 2\sigma_\zeta^2/(1-a^2)$ *is the variance of the harmonic oscillation.*

Proof. The proof follows the same reasoning as in Theorem 5.4.1. □

As a remark, to completely characterise the second-order properties of a complex-valued signal we must also define the *pseudo-covariance* or *relation* sequence $r_\tau = E\{Z_t Z_{t-\tau}\}$. When this sequence is non-zero at any lag then the process is called improper, otherwise it is proper; see Sykulski and Percival (2016) for more details.

In our case we have $r_\tau = E\{Z_t Z_{t-\tau}\} = 0$ for all $\tau \in \mathbb{Z}$, and hence the complex AR(1) process is proper. Given that the stochastic polyfoil process is constructed from the addition of two proper processes, it is therefore also proper. Examples of improper processes, where oscillations are elliptical, are given in Sykulski et al. (2016) and Sykulski et al. (2022).

As discussed in Section 5.3, correlations between distinct frequencies for nonstationary processes can be studied using the Loéve spectrum; this will be derived for the nonstationary polyfoil process later in this section. In order to aid in this derivation, we first consider how the stationary complex AR(1) process behaves in the frequency domain. The spectrum of such a process is given by the following theorem.

Theorem 5.4.3. *The spectrum of the fundamental oscillation described in equation 5.4.1 is given by*

$$f^{\{1\}}(\omega) = \frac{2\sigma_\epsilon^2}{1 - 2a\cos(\omega - \theta) + a^2}.$$

Proof. The spectrum is found by taking the Fourier transform of the autocovariance sequence given in Theorem 5.4.1. Define the unit step function,

$$u(\tau) = \begin{cases} 1 & \text{if } \tau \geq 0, \\ 0 & \text{if } \tau < 0. \end{cases}$$

. The Fourier transform of the autocovariance sequence is then given by,

$$\begin{aligned}
S^{\{1\}}(\omega) &= \frac{2\sigma_\epsilon^2}{(1-a^2)} \sum_{\tau\in\mathbb{Z}} a^{|\tau|} e^{i\theta\tau} e^{-i\omega\tau}, \\
&= \frac{2\sigma_\epsilon^2}{(1-a^2)} \left(\sum_{\tau\in\mathbb{Z}} a^{\tau} e^{i\theta\tau} u(\tau) e^{-i\omega\tau} + \sum_{\tau\in\mathbb{Z}} a^{-\tau} e^{i\theta\tau} u(-\tau) e^{-i\omega\tau} - 1 \right) \\
&= \frac{2\sigma_\epsilon^2}{(1-a^2)} \left(\frac{1}{1-ae^{-i(\omega-\theta)}} + \frac{1}{1-ae^{-i(-(\omega-\theta))}} - 1 \right), \\
&= \frac{2\sigma_\epsilon^2}{(1-a^2)} \left(\frac{1}{1-ae^{-i(\omega-\theta)}} + \frac{1}{1-ae^{i(\omega-\theta)}} - 1 \right), \\
&= \frac{2\sigma_\epsilon^2}{(1-a^2)} \left(\frac{1-ae^{i(\omega-\theta)} + 1 - ae^{-i(\omega-\theta)} - (1 - ae^{i(\omega-\theta)} - ae^{-i(\omega-\theta)} + a^2)}{(1-ae^{i(\omega-\theta)})(1-ae^{-i(\omega-\theta)})} \right), \\
&= \frac{2\sigma_\epsilon^2}{(1-a^2)} \left(\frac{1-a^2}{(1 - ae^{i(\omega-\theta)} - ae^{-i(\omega-\theta)} + a^2)} \right), \\
&= \frac{2\sigma_\epsilon^2}{1 - 2a\cos(\omega-\theta) + a^2},
\end{aligned}$$

as required. □

The spectrum for the harmonic oscillation is stated in the following corollary.

Corollary 5.4.4. *The spectrum of the harmonic oscillation is given by,*

$$S^{\{2\}}(\omega) = \frac{2\sigma_\zeta^2}{1 - 2a\cos(\omega - h\theta) + a^2}.$$

Proof. The proof follows the same reasoning as in Theorem 5.4.3. □

These expressions are utilised in the following subsection to derive the autocovariance sequence and the Loève spectrum for the polyfoil process.

5.4.2 Autocovariance sequence of the polyfoil process

Recall, the polyfoil process is constructed by superposing two complex AR(1) processes, representing a fundamental oscillation and a harmonic oscillation, such that we have $Z_t = Z_t^{\{1\}} + Z_t^{\{2\}}$. The autocovariance sequence of this process can be expressed as

the sum of the stationary autocovariance sequences from the circular oscillations, plus some nonstationary cross terms.

Lemma 5.4.5. *The autocovariance sequence for the polyfoil model $s_{t,\tau}$ is given by*

$$s_{t,\tau} = s_\tau^{\{1\}} + s_\tau^{\{2\}} + E\{Z_t^{\{1\}} Z_{t-\tau}^{\{2\}*}\} + E\{Z_t^{\{2\}} Z_{t-\tau}^{\{1\}*}\},$$

where $s_\tau^{\{1\}}$ is the autocovariance sequence of the fundamental oscillation and $s_\tau^{\{2\}}$ is the autocovariance sequence of the harmonic oscillation.

Proof.

$$\begin{aligned} s_{t,\tau} &= E\{Z_t Z_{t-\tau}^*\} \\ &= E\{(Z_t^{\{1\}} + Z_t^{\{2\}})(Z_{t-\tau}^{\{1\}*} + Z_{t-\tau}^{\{2\}*})\} \\ &= E\{Z_t^{\{1\}} Z_{t-\tau}^{\{1\}*} + Z_t^{\{1\}} Z_{t-\tau}^{\{2\}*} + Z_t^{\{2\}} Z_{t-\tau}^{\{1\}*} + Z_t^{\{2\}} Z_{t-\tau}^{\{2\}*}\} \\ &= s_\tau^{\{1\}} + s_\tau^{\{2\}} + E\{Z_t^{\{1\}} Z_{t-\tau}^{\{2\}*}\} + E\{Z_t^{\{2\}} Z_{t-\tau}^{\{1\}*}\}, \end{aligned}$$

as required. □

The above expression may be simplified by individual consideration of the cross terms, wherein the case of non-negative lag $\tau \geq 0$ and negative lag $\tau < 0$ are dealt with separately. The following lemma provides a simplified form of $E\{Z_t^{\{1\}} Z_{t-\tau}^{\{2\}*}\}$ for non-negative lag τ.

Lemma 5.4.6. *For non-negative lag $\tau \geq 0$, the expectation of $E\{Z_t^{\{1\}} Z_{t-\tau}^{\{2\}*}\}$ can be written as,*

$$E\{Z_t^{\{1\}} Z_{t-\tau}^{\{2\}*}\} = (ae^{i\theta})^\tau E\{Z_{t-\tau}^{\{1\}} Z_{t-\tau}^{\{2\}*}\}.$$

Proof.

$$\begin{aligned}
E\{Z_t^{\{1\}} Z_{t-\tau}^{\{2\}*}\} &= E\{(ae^{i\theta} Z_{t-1}^{\{1\}} + \nu_t^{\{1\}}) Z_{t-\tau}^{\{2\}*}\} \\
&= ae^{i\theta} E\{Z_{t-1}^{\{1\}} Z_{t-\tau}^{\{2\}*}\} + E\{\nu_t^{\{1\}} Z_{t-\tau}^{\{2\}*}\} \\
&= ae^{i\theta} E\{(ae^{i\theta} Z_{t-2}^{\{1\}} + \nu_{t-1}^{\{1\}}) Z_{t-\tau}^{\{2\}*}\} \\
&= (ae^{i\theta})^2 E\{Z_{t-2}^{\{1\}} Z_{t-\tau}^{\{2\}*}\} \\
&\vdots \\
&= (ae^{i\theta})^\tau E\{Z_{t-\tau}^{\{1\}} Z_{t-\tau}^{\{2\}*}\},
\end{aligned}$$

as required. □

A simplified form of $E\{Z_t^{\{2\}} Z_{t-\tau}^{\{1\}*}\}$ for non-negative lag τ can be obtained similarly. This is given in the following corollary.

Corollary 5.4.7. *For non-negative lag $\tau \geq 0$, the expectation of $E\{Z_t^{\{2\}} Z_{t-\tau}^{\{1\}*}\}$ can be written as,*

$$E\{Z_t^{\{2\}} Z_{t-\tau}^{\{1\}*}\} = (ae^{ih\theta})^\tau E\{Z_{t-\tau}^{\{2\}} Z_{t-\tau}^{\{1\}*}\}.$$

The proof follows the same reasoning as in Lemma 5.4.6.

Similarly, the following lemma provides a simplified form of $E\{Z_t^{\{1\}} Z_{t-\tau}^{\{2\}*}\}$ for negative lag τ.

Lemma 5.4.8. *For negative lag $\tau < 0$, the expectation of $E\{Z_t^{\{1\}} Z_{t-\tau}^{\{2\}*}\}$ can be written as,*

$$E\{Z_t^{\{1\}} Z_{t-\tau}^{\{2\}*}\} = a^{|\tau|} e^{ih\theta\tau} E\{Z_t^{\{1\}} Z_t^{\{2\}*}\}.$$

Proof.

$$\begin{aligned} E\{Z_t^{\{1\}} Z_{t-\tau}^{\{2\}*}\} &= E\{Z_t^{\{1\}}(ae^{-ih\theta} Z_{t-\tau-1}^{\{2\}*} + \nu_{t-\tau}^{\{2\}*})\} \\ &= ae^{-ih\theta} E\{Z_t^{\{1\}} Z_{t-\tau-1}^{\{2\}*}\} + E\{Z_t^{\{1\}} \nu_{t-\tau}^{\{2\}*}\} \\ &\vdots \\ &= (ae^{-ih\theta})^{|\tau|} E\{Z_t^{\{1\}} Z_t^{\{2\}*}\} \\ &= a^{|\tau|} e^{ih\theta\tau} E\{Z_t^{\{1\}} Z_t^{\{2\}*}\}, \end{aligned}$$

as required. □

The simplified form of $E\{Z_t^{\{2\}} Z_{t-\tau}^{\{1\}*}\}$ with negative lag τ is given in the following corollary.

Corollary 5.4.9. *For negative lag $\tau < 0$, the expectation of $E\{Z_t^{\{2\}} Z_{t-\tau}^{\{1\}*}\}$ can be written as,*

$$E\{Z_t^{\{2\}} Z_{t-\tau}^{\{1\}*}\} = a^{|\tau|} e^{i\theta\tau} E\{Z_t^{\{2\}} Z_t^{\{1\}*}\}.$$

The proof follows the same reasoning as in Lemma 5.4.8.

If the two complex AR(1) processes are independent, then trivially all cross terms in Lemma 5.4.5 vanish such that $E\{Z_t^{\{1\}} Z_{t-\tau}^{\{2\}*}\} = E\{Z_t^{\{2\}} Z_{t-\tau}^{\{1\}*}\} = 0$, and hence we create a stationary polyfoil model with autocovariance $s_{t,\tau} = s_\tau^{\{1\}} + s_\tau^{\{2\}}$ which is only dependent on τ and not t. However, we shall create a nonstationary polyfoil by modelling the cross-term $E\{Z_t^{\{1\}} Z_t^{\{2\}*}\}$ to be independent in amplitude between each component, but 'locked' in phase. Specifically, consider its polar representation:

$$\begin{aligned} E\{Z_t^{\{1\}} Z_t^{\{2\}*}\} &= E\{r_t^{\{1\}} e^{i\varphi_t^{\{1\}}} r_t^{\{2\}} e^{-i\varphi_t^{\{2\}}}\}, \\ &= E\{r_t^{\{1\}} r_t^{\{2\}} e^{i[\varphi_t^{\{1\}} - \varphi_t^{\{2\}}]}\}. \end{aligned}$$

We will assume that the amplitudes $r_t^{\{1\}}, r_t^{\{2\}}$ are independent of each other and the phase difference, such that,

$$E\{Z_t^{\{1\}} Z_t^{\{2\}*}\} = E\{r_t^{\{1\}}\} E\{r_t^{\{2\}}\} E\{e^{i[\varphi_t^{\{1\}} - \varphi_t^{\{2\}}]}\}.$$

We then phase-lock the phase difference at $t = 0$ such that $E\{e^{i[\varphi_0^{\{1\}} - \varphi_0^{\{2\}}]}\} = e^{i\phi_0}$, meaning that under expectation, the two oscillations are out of phase by angle ϕ_0 at time $t = 0$. As the two components oscillate at difference frequencies, this phase difference should be a function of time t. Specifically, with the first component we have that

$$E\{e^{i\varphi_t^{\{1\}}}\} = E\{e^{i\varphi_0^{\{1\}}}\} e^{i\theta t},$$

and with the second component

$$E\{e^{i\varphi_t^{\{2\}}}\} = E\{e^{i\varphi_0^{\{2\}}}\} e^{ih\theta t},$$

such that we assume the phase-locking under expectation at general time t is

$$\begin{aligned} E\{e^{i[\varphi_t^{\{1\}} - \varphi_t^{\{2\}}]}\} &= E\{e^{i[\varphi_0^{\{1\}} - \varphi_0^{\{2\}}]}\} e^{i\theta t} e^{-ih\theta t} \\ &= e^{i(\phi_0 + \theta(1-h)t)}, \end{aligned} \tag{5.4.2}$$

such that the expected phase difference periodically oscillates with period $t = 2\pi/(\theta(h-1))$, which is consistent with the $|h-1|$ loops observed in Figure 5.1.2.

Note that the amplitude $r_t^{\{1\}}$ can be written as,

$$r_t^{\{1\}} = \sqrt{(X_t^{\{1\}})^2 + (Y_t^{\{1\}})^2}.$$

By the stationarity and Gaussianity of X and Y, and the properties of the Rayleigh

distribution, it is therefore possible to express the expectation of this as,

$$E\{r_t^{\{1\}}\} = E\Big\{\sqrt{(X_t^{\{1\}})^2 + (Y_t^{\{1\}})^2}\Big\} = \sigma_\epsilon\sqrt{\frac{\pi}{2(1-a^2)}}.$$

Similarly, for $r_t^{\{2\}}$ we can write,

$$E\{r_t^{\{2\}}\} = \sigma_\zeta\sqrt{\frac{\pi}{2(1-a^2)}}.$$

These expressions can be modified for non-Gaussian processes in extensions of the model. The term $E\{Z_t^{\{1\}}Z_t^{\{2\}*}\}$ is then given by,

$$E\{Z_t^{\{1\}}Z_t^{\{2\}*}\} = \frac{\pi\sigma_\zeta\sigma_\epsilon}{2(1-a^2)}e^{i(\phi_0+\theta(1-h)t)}.$$

Note that we can similarly express,

$$E\{Z_t^{\{2\}}Z_t^{\{1\}*}\} = \frac{\pi\sigma_\zeta\sigma_\epsilon}{2(1-a^2)}e^{-i(\phi_0+\theta(1-h)t)},$$

by noting that,

$$\begin{aligned} E\{e^{i[\varphi_t^{\{2\}}-\varphi_t^{\{1\}}]}\} &= E\{e^{-i[\varphi_t^{\{1\}}-\varphi_t^{\{2\}}]}\}, \\ &= e^{-i(\phi_0+\theta(1-h)t)}, \end{aligned}$$

under our assumption of phase-locking. For non-negative lag, the cross expectations can therefore be written as,

$$\begin{aligned} E\{Z_t^{\{1\}}Z_{t-\tau}^{\{2\}*}\} &= (ae^{i\theta})^\tau E\{Z_{t-\tau}^{\{1\}}Z_{t-\tau}^{\{2\}*}\} = a^{|\tau|}e^{i\theta\tau}\frac{\pi\sigma_\zeta\sigma_\epsilon}{2(1-a^2)}e^{i(\phi_0+\theta(1-h)(t-\tau))}, \\ E\{Z_t^{\{2\}}Z_{t-\tau}^{\{1\}*}\} &= (ae^{ih\theta})^\tau E\{Z_{t-\tau}^{\{2\}}Z_{t-\tau}^{\{1\}*}\} = a^{|\tau|}e^{ih\theta\tau}\frac{\pi\sigma_\zeta\sigma_\epsilon}{2(1-a^2)}e^{-i(\phi_0+\theta(1-h)(t-\tau))}, \end{aligned}$$

and for negative lag we can write,

$$E\{Z_t^{\{1\}} Z_{t-\tau}^{\{2\}*}\} = a^{|\tau|} e^{ih\theta} E\{Z_t^{\{1\}} Z_t^{\{2\}*}\} = a^{|\tau|} e^{ih\theta\tau} \frac{\pi\sigma_\zeta\sigma_\epsilon}{2(1-a^2)} e^{i(\phi_0+\theta(1-h)t)},$$
$$E\{Z_t^{\{2\}} Z_{t-\tau}^{\{1\}*}\} = a^{|\tau|} e^{i\theta\tau} E\{Z_t^{\{2\}} Z_t^{\{1\}*}\} = a^{|\tau|} e^{i\theta\tau} \frac{\pi\sigma_\zeta\sigma_\epsilon}{2(1-a^2)} e^{-i(\phi_0+\theta(1-h)t)}.$$

Let $s_{t,\tau}^{\{3\}} = E\{Z_t^{\{1\}} Z_{t-\tau}^{\{2\}*}\} + E\{Z_t^{\{2\}} Z_{t-\tau}^{\{1\}*}\}$, such that this term encapsulates the nonstationary component of the autocovariance sequence. This term can be expressed in one expression for both non-negative and negative values of lag, as given in the following proposition.

Proposition 5.4.10. *For all lags $\tau \in \mathbb{Z}$, the nonstationary contribution to the autocovariance sequence of the polyfoil process with phase-locked oscillations, as defined by equation (5.4.2), is given by,*

$$s_{t,\tau}^{\{3\}} = \frac{\pi\sigma_\zeta\sigma_\epsilon}{2(1-a^2)} a^{|\tau|} \left(e^{i[\phi_0+\theta((1-h)t+h\tau)]} + e^{-i[\phi_0+\theta((1-h)t-\tau)]} \right).$$

Proof. For non-negative lag τ, the previous expressions give,

$$\begin{aligned} s_{t,\tau\geq 0}^{\{3\}} &= a^{|\tau|} e^{i\theta\tau} \frac{\pi\sigma_\zeta\sigma_\epsilon}{2(1-a^2)} e^{i(\phi_0+\theta(1-h)(t-\tau))} + a^{|\tau|} e^{ih\theta\tau} \frac{\pi\sigma_\zeta\sigma_\epsilon}{2(1-a^2)} e^{-i(\phi_0+\theta(1-h)(t-\tau))}, \\ &= \frac{\pi\sigma_\zeta\sigma_\epsilon}{2(1-a^2)} a^{|\tau|} \left(e^{i[\phi_0+\theta((1-h)t+h\tau)]} + e^{-i[\phi_0+\theta((1-h)t-\tau)]} \right). \end{aligned}$$

For negative lag τ, the previous expressions give,

$$\begin{aligned} s_{t,\tau<0}^{\{3\}} &= a^{|\tau|} e^{ih\theta\tau} \frac{\pi\sigma_\zeta\sigma_\epsilon}{2(1-a^2)} e^{i(\phi_0+\theta(1-h)t)} + a^{|\tau|} e^{i\theta\tau} \frac{\pi\sigma_\zeta\sigma_\epsilon}{2(1-a^2)} e^{-i(\phi_0+\theta(1-h)t)}, \\ &= \frac{\pi\sigma_\zeta\sigma_\epsilon}{2(1-a^2)} a^{|\tau|} \left(e^{i[\phi_0+\theta((1-h)t+h\tau)]} + e^{-i[\phi_0+\theta((1-h)t-\tau)]} \right). \end{aligned}$$

Thus we have $s^{\{3\}}_{t,\tau\geq 0} = s^{\{3\}}_{t,\tau<0}$, such that,

$$s^{\{3\}}_{t,\tau} = \frac{\pi\sigma_\zeta\sigma_\epsilon}{2(1-a^2)}a^{|\tau|}\left(e^{i[\phi_0+\theta((1-h)t+h\tau)]} + e^{-i[\phi_0+\theta((1-h)t-\tau)]}\right),$$

for all lag $\tau \in \mathbb{Z}$, as required. □

It can be seen that the nonstationary contribution to the autocovariance sequence $s^{\{3\}}_{t,\tau}$ obeys Hermitian symmetry be observing,

$$\begin{aligned} s^{\{3\}}_{t-\tau,-\tau} &= \frac{\pi\sigma_\zeta\sigma_\epsilon}{2(1-a^2)}a^{|-\tau|}\left(e^{i[\phi_0+\theta((1-h)(t-\tau)-h\tau)]} + e^{-i[\phi_0+\theta((1-h)(t-\tau)+\tau)]}\right) \\ &= \frac{\pi\sigma_\zeta\sigma_\epsilon}{2(1-a^2)}a^{|\tau|}\left(e^{i[\phi_0+\theta(1-h)t-\tau)]} + e^{-i[\phi_0+\theta((1-h)t+h\tau)]}\right) = s^{\{3\}*}_{t,\tau}. \end{aligned}$$

The autocovariance sequence for the nonstationary polyfoil process, defined under the assumption of phase-locked oscillations, is therefore given in the following theorem.

Theorem 5.4.11. *The autocovariance sequence for the polyfoil model with phase-locked oscillations, as defined by equation* (5.4.2)*, is given by,*

$$\begin{aligned} s_{t,\tau} =& \frac{2\sigma_\epsilon^2}{(1-a^2)}a^{|\tau|}e^{i\theta\tau} + \frac{2\sigma_\zeta^2}{(1-a^2)}a^{|\tau|}e^{ih\theta\tau} + \\ & \frac{\pi\sigma_\zeta\sigma_\epsilon}{2(1-a^2)}a^{|\tau|}\left(e^{i[\phi_0+\theta((1-h)t+h\tau)]} + e^{-i[\phi_0+\theta((1-h)t-\tau)]}\right). \end{aligned}$$

Proof. This follows directly from Lemma 5.4.5, Theorem 5.4.1, Corollary 5.4.2 and Proposition 5.4.10. □

5.4.3 Loève spectrum of the nonstationary polyfoil process

It is possible to derive an analytical form of the Loève spectrum for the nonstationary polyfoil process. Equation (5.3.1) shows that the Loève spectrum can be found by taking the double Fourier transform of the autocovariance function $s_{t,\tau}$, in terms of time t and lag τ, and this relationship holds true for complex-valued processes as well.

By linearity of the Fourier transform, this can be written as,

$$\begin{aligned} S(\omega,\nu) &= \sum_{\tau\in\mathbb{Z}}\sum_{t\in\mathbb{Z}} s_{t,\tau}e^{-i\omega\tau}e^{-i\nu t}, \\ &= \sum_{\tau\in\mathbb{Z}}\sum_{t\in\mathbb{Z}} (s_{\tau}^{\{1\}} + s_{\tau}^{\{2\}} + s_{t,\tau}^{\{3\}})e^{-i\omega\tau}e^{-i\nu t}, \\ &= \sum_{\tau\in\mathbb{Z}}\sum_{t\in\mathbb{Z}} s_{\tau}^{\{1\}}e^{-i\omega\tau}e^{-i\nu t} + \sum_{\tau\in\mathbb{Z}}\sum_{t\in\mathbb{Z}} s_{\tau}^{\{2\}}e^{-i\omega\tau}e^{-i\nu t} + \sum_{\tau\in\mathbb{Z}}\sum_{t\in\mathbb{Z}} s_{t,\tau}^{\{3\}}e^{-i\omega\tau}e^{-i\nu t}, \end{aligned}$$

where ω and ν can be considered to be a frequency and frequency-offset parameter respectively. For the sake of simplified notation, denote the above three terms by $S^{\{1\}}(\omega,\nu)$, $S^{\{2\}}(\omega,\nu)$ and $S^{\{3\}}(\omega,\nu)$, such that the double Fourier transform can be written as

$$S(\omega,\nu) = S^{\{1\}}(\omega,\nu) + S^{\{2\}}(\omega,\nu) + S^{\{3\}}(\omega,\nu). \tag{5.4.3}$$

The spectrum for the $s_{\tau}^{\{1\}}$ and $s_{\tau}^{\{2\}}$ components in terms of frequency ω are given by Theorem 5.4.3 and Corollary 5.4.4 respectively. Since these expressions are not dependent on t, we can use standard results of discrete time Fourier transform pairs in order to obtain,

$$\begin{aligned} S^{\{1\}}(\omega,\nu) &= \sum_{t\in\mathbb{Z}} \frac{2\sigma_{\epsilon}^2}{1-2a\cos(\omega-\theta)+a^2}e^{-i\nu t}, \\ &= \frac{4\sigma_{\epsilon}^2\pi}{1-2a\cos(\omega-\theta)+a^2}\sum_{-\infty}^{\infty}\delta\{\nu-2\pi k\}, \end{aligned} \tag{5.4.4}$$

and

$$\begin{aligned} S^{\{2\}}(\omega,\nu) &= \sum_{t\in\mathbb{Z}} \frac{2\sigma_{\zeta}^2}{1-2a\cos(\omega-h\theta)+a^2}e^{-i\nu t}, \\ &= \frac{4\sigma_{\zeta}^2\pi}{1-2a\cos(\omega-h\theta)+a^2}\sum_{-\infty}^{\infty}\delta\{\nu-2\pi k\}, \end{aligned} \tag{5.4.5}$$

where $\delta\{\cdot\}$ is the Dirac-delta function. We note that typically ω and ν only need to be considered in the range $[-\pi, \pi)$ or $[0, 2\pi)$, owing to the 2π-periodicity of $S(\omega, \nu)$ in both ω and ν, such that the Dirac-delta function is non-zero at $\nu = 0$ only. The third term is more complex due to the nonstationarity, and is given in the following proposition.

Proposition 5.4.12. *The Loève spectrum of the nonstationary component of the polyfoil is given by,*

$$\begin{aligned} S^{\{3\}}(\omega, \nu) = &\frac{\pi^2 \sigma_\zeta \sigma_\epsilon}{(1 - 2a\cos(\omega - h\theta) + a^2)} e^{i\phi_0} \sum_{k=-\infty}^{\infty} \delta\{\nu - \theta(1-h) - 2\pi k\} + \\ &\frac{\pi^2 \sigma_\zeta \sigma_\epsilon}{(1 - 2a\cos(\omega - \theta) + a^2)} e^{-i\phi_0} \sum_{k=-\infty}^{\infty} \delta\{\nu + \theta(1-h) - 2\pi k\} \end{aligned}$$

.

Proof. The Loève spectrum is given by taking the double Fourier transform of the

autocovariance sequence $s_{t,\tau}^{\{3\}}$ in terms of t and τ as follows,

$$\begin{aligned}
S^{\{3\}}(\omega,\nu) =& \sum_{\tau\in\mathbb{Z}}\sum_{t\in\mathbb{Z}} s_{t,\tau}^{\{3\}} e^{-i\omega\tau}e^{-i\nu t},\\
&\frac{\pi\sigma_\zeta\sigma_\epsilon}{2(1-a^2)}\sum_{\tau\in\mathbb{Z}}\sum_{t\in\mathbb{Z}} a^{|\tau|}\left(e^{i[\phi_0+\theta((1-h)t+h\tau)]}+e^{-i[\phi_0+\theta((1-h)t-\tau)]}\right)e^{-i\omega\tau}e^{-i\nu t}\\
=&\frac{\pi\sigma_\zeta\sigma_\epsilon}{2(1-a^2)}\sum_{\tau\in\mathbb{Z}}\sum_{t\in\mathbb{Z}} a^{|\tau|}e^{i[\phi_0+\theta((1-h)t+h\tau)]}e^{-i\omega\tau}e^{-i\nu t}+\\
&\frac{\pi\sigma_\zeta\sigma_\epsilon}{2(1-a^2)}\sum_{\tau\in\mathbb{Z}}\sum_{t\in\mathbb{Z}} a^{|\tau|}e^{-i[\phi_0+\theta((1-h)t-\tau)]}e^{-i\omega\tau}e^{-i\nu t},\\
=&\frac{\pi\sigma_\zeta\sigma_\epsilon}{2(1-a^2)}e^{i\phi_0}\sum_{\tau\in\mathbb{Z}} a^{|\tau|}e^{i\theta h\tau}e^{-i\omega\tau}\sum_{t\in\mathbb{Z}} e^{i\theta(1-h)t}e^{-i\nu t}+\\
&\frac{\pi\sigma_\zeta\sigma_\epsilon}{2(1-a^2)}e^{-i\phi_0}\sum_{\tau\in\mathbb{Z}} a^{|\tau|}e^{i\theta\tau}e^{-i\omega\tau}\sum_{t\in\mathbb{Z}} e^{-i\theta(1-h)t}e^{-i\nu t}\\
=&\frac{\pi\sigma_\zeta\sigma_\epsilon}{2(1-a^2)}e^{i\phi_0}\frac{1-a^2}{1-2a\cos(\omega-h\theta)+a^2}\sum_{t\in\mathbb{Z}} e^{i\theta(1-h)t}e^{-i\nu t}+\\
&\frac{\pi\sigma_\zeta\sigma_\epsilon}{2(1-a^2)}e^{-i\phi_0}\frac{1-a^2}{1-2a\cos(\omega-\theta)+a^2}\sum_{t\in\mathbb{Z}} e^{-i\theta(1-h)t}e^{-i\nu t},\\
=&\frac{\pi\sigma_\zeta\sigma_\epsilon}{2(1-2a\cos(\omega-h\theta)+a^2)}e^{i\phi_0}\sum_{t\in\mathbb{Z}} e^{i\theta(1-h)t}e^{-i\nu t}+\\
&\frac{\pi\sigma_\zeta\sigma_\epsilon}{2(1-2a\cos(\omega-\theta)+a^2)}e^{-i\phi_0}\sum_{t\in\mathbb{Z}} e^{-i\theta(1-h)t}e^{-i\nu t}.
\end{aligned}$$

Finally, using standard results from Fourier transform pairs, we may write,

$$\begin{aligned}
S^{\{3\}}(\omega,\nu) =& \frac{\pi^2\sigma_\zeta\sigma_\epsilon}{(1-2a\cos(\omega-h\theta)+a^2)}e^{i\phi_0}\sum_{k=-\infty}^{\infty}\delta\{\nu-\theta(1-h)-2\pi k\}+\\
&\frac{\pi^2\sigma_\zeta\sigma_\epsilon}{2(1-2a\cos(\omega-\theta)+a^2)}e^{-i\phi_0}\sum_{k=-\infty}^{\infty}\delta\{\nu+\theta(1-h)-2\pi k\},
\end{aligned}$$

as required. □

Note that the Dirac-delta function in $S^{\{3\}}(\omega,\nu)$ is non-zero at frequency off-set $\nu=\pm\theta(1-h)$ which is the difference in frequency between the two oscillations, thus showing how their interaction manifests in the Loève spectrum. The Loève spectrum

for the nonstationary polyfoil model, constrained in the range $\omega, \nu \in [-\pi, \pi)$, is given in the following theorem.

Theorem 5.4.13. *For a nonstationary polyfoil with phase-locked oscillations, as defined by equation (5.4.2), the Loève spectrum is given by,*

$$\begin{aligned} S(\omega,\nu) =& S^{\{1\}}(\omega,\nu) + S^{\{2\}}(\omega,\nu) + S^{\{3\}}(\omega,\nu), \\ =& \frac{4\sigma_\epsilon^2 \pi}{1 - 2a\cos(\omega - \theta) + a^2}\delta\{\nu\} + \\ & \frac{4\sigma_\zeta^2 \pi}{1 - 2a\cos(\omega - h\theta) + a^2}\delta\{\nu\} + \\ & \frac{\pi^2 \sigma_\zeta \sigma_\epsilon}{(1 - 2a\cos(\omega - h\theta) + a^2)} e^{i\phi_0}\delta\{\nu - \theta(1-h)\} + \\ & \frac{\pi^2 \sigma_\zeta \sigma_\epsilon}{(1 - 2a\cos(\omega - \theta) + a^2)} e^{-i\phi_0}\delta\{\nu + \theta(1-h)\}, \end{aligned}$$

where $\omega, \nu \in [-\pi, \pi)$.

Proof. The Loève spectrum of the nonstationary polyfoil process can be written as the sum of $S^{\{1\}}(\omega,\nu)$, $S^{\{2\}}(\omega,\nu)$, and $S^{\{3\}}(\omega,\nu)$, as given in equation (5.4.3). Expressions for $S^{\{1\}}(\omega,\nu)$ and $S^{\{2\}}(\omega,\nu)$ are given by equations (5.4.4) and (5.4.5) respectively. The Loève spectrum for the nonstationary component of the polyfoil process is given in Proposition 5.4.12. It can be shown that the delta function summations in these expressions simplify by considering how the frequencies are bounded. Given that we are working with discrete-time processes ($t \in \mathbb{Z}$) then we can restrict ourselves to $\omega, \nu \in [-\pi, \pi)$, such that they are within the Nyquist frequencies. Therefore the only contributions from the delta functions occur when $\nu = 0$ or $\nu = \pm\theta(1-h)$, and the full

Loève spectrum for the nonstationary polyfoil model is given by,

$$\begin{aligned}
S(\omega,\nu) =& S^{\{1\}}(\omega,\nu) + S^{\{2\}}(\omega,\nu) + S^{\{3\}}(\omega,\nu), \\
=& \frac{4\sigma_\epsilon^2\pi}{1-2a\cos(\omega-\theta)+a^2}\delta\{\nu\}+ \\
& \frac{4\sigma_\zeta^2\pi}{1-2a\cos(\omega-h\theta)+a^2}\delta\{\nu\}+ \\
& \frac{\pi^2\sigma_\zeta\sigma_\epsilon}{(1-2a\cos(\omega-h\theta)+a^2)}e^{i\phi_0}\delta\{\nu-\theta(1-h)\}+ \\
& \frac{\pi^2\sigma_\zeta\sigma_\epsilon}{(1-2a\cos(\omega-\theta)+a^2)}e^{-i\phi_0}\delta\{\nu+\theta(1-h)\},
\end{aligned}$$

as required. □

Let $\omega_1 = \omega$ and $\omega_2 = \omega + \nu$. Hence, the above expression shows that the Loève spectrum matrix is zero everywhere, apart from three parallel lines when $\omega_1, \omega_2 \in [-\pi, \pi)$. Namely,

(i) We have a contribution from the first two terms in the summation when the frequencies are equal $\omega_1 = \omega_2$ (that is, the frequency offset parameter ν is zero);

(ii) We have a contribution from the third term when $\omega_2 = [\omega_1 + \theta(1-h)]\text{mod}(2\pi)$;

(iii) We have a contribution from the fourth term when $\omega_2 = [\omega_1 - \theta(1-h)]\text{mod}(2\pi)$,

where $[x]\text{mod}(2\pi)$ is the modulo function placing x in the range $[-\pi, \pi)$ (as opposed to $[0, 2\pi)$ in slight abuse of convention). For the rest of this chapter we drop the $\text{mod}(2\pi)$ term for convenience, but note this should be added to always bring ω_2 in the range $[-\pi, \pi)$. These parallel lines are a characteristic of the fact that the polyfoil process is nonstationary under the assumption of phase-locked oscillations, which leads to interactions between distinct frequencies. These can be studied further by computing the Loève coherency, as is done in the following section.

5.4.4 Loève coherency of the nonstationary polyfoil process

The Loève coherency is given as a function of the Loève spectrum in equation (5.3.2). Clearly, since the Loève spectrum for the nonstationary polyfoil process is zero everywhere but the three aforementioned diagonals, this is also the case for the coherency. In the case where $\omega_1 = \omega_2$, the coherency is simply equal to 1. The Loève coherency along the line $\omega_2 = \omega_1 + \theta(1-h)$ is given in the following theorem.

Theorem 5.4.14. *The Loève coherency of the nonstationary polyfoil process along the line $\omega_2 = \omega_1 + \theta(1-h)$ is given by,*

$$\tau(\omega_1, \omega_1 + \theta(1-h)) = \frac{\pi\sigma_\zeta\sigma_\epsilon e^{i\phi_0}}{4(1-2a\cos(\omega_1 - h\theta) + a^2)} \frac{1}{\sqrt{(A+B)(C+D)}},$$

where we have defined

$$\begin{aligned} A &= \frac{\sigma_\epsilon^2}{1-2a\cos(\omega_1 - \theta) + a^2}, \\ B &= \frac{\sigma_\zeta^2}{1-2a\cos(\omega_1 - h\theta) + a^2}, \\ C &= \frac{\sigma_\epsilon^2}{1-2a\cos(\omega_1 - h\theta) + a^2}, \\ D &= \frac{\sigma_\zeta^2}{1-2a\cos(\omega_1 + \theta - 2h\theta) + a^2}. \end{aligned}$$

Proof. For the off-diagonal contribution along $\omega_2 = \omega_1 + \theta(1-h)$, we have

$$\tau(\omega_1, \omega_1 + \theta(1-h)) = \frac{S(\omega_1, \omega_1 + \theta(1-h))}{\sqrt{S(\omega_1, \omega_1)S(\omega_1 + \theta(1-h), \omega_1 + \theta(1-h))}}.$$

Let us begin by considering the numerator. Clearly all components of the Loève spectrum are equal to zero, apart from the contribution with the delta function $\delta\{\nu - \theta(1-$

$h)\}$, which is given by

$$S(\omega_1, \omega_1 + \theta(1-h)) = \frac{\pi^2 \sigma_\zeta \sigma_\epsilon}{(1 - 2a\cos(\omega_1 - h\theta) + a^2)} e^{i\phi_0} \delta\{0\}.$$

We can consider the two elements of the denominator in turn. Since these correspond to values of the spectrum where the two frequencies are equal, the contributions come from the first two terms in Theorem 5.4.13. For the first element we have,

$$\begin{aligned} S(\omega_1, \omega_1) = &\frac{4\sigma_\epsilon^2 \pi}{1 - 2a\cos(\omega_1 - \theta) + a^2} \delta\{0\} + \\ &\frac{4\sigma_\zeta^2 \pi}{1 - 2a\cos(\omega_1 - h\theta) + a^2} \delta\{0\}. \end{aligned}$$

Similarly, the second element is given by,

$$\begin{aligned} S(\omega_1 + \theta(1-h), \omega_1 + \theta(1-h)) = &\frac{4\sigma_\epsilon^2 \pi}{1 - 2a\cos(\omega_1 + \theta(1-h) - \theta) + a^2} \delta\{0\} + \\ &\frac{4\sigma_\zeta^2 \pi}{1 - 2a\cos(\omega_1 + \theta(1-h) - h\theta) + a^2} \delta\{0\}, \\ = &\frac{4\sigma_\epsilon^2 \pi}{1 - 2a\cos(\omega_1 - h\theta) + a^2} \delta\{0\} + \\ &\frac{4\sigma_\zeta^2 \pi}{1 - 2a\cos(\omega_1 + \theta - 2h\theta) + a^2} \delta\{0\}. \end{aligned}$$

We can compute the denominator by evaluating the product of these two terms. Using the above definitions,

$$\begin{aligned} A &= \frac{\sigma_\epsilon^2}{1 - 2a\cos(\omega_1 - \theta) + a^2}, \\ B &= \frac{\sigma_\zeta^2}{1 - 2a\cos(\omega_1 - h\theta) + a^2}, \\ C &= \frac{\sigma_\epsilon^2}{1 - 2a\cos(\omega_1 - h\theta) + a^2}, \\ D &= \frac{\sigma_\zeta^2}{1 - 2a\cos(\omega_1 + \theta - 2h\theta) + a^2}, \end{aligned}$$

the product can be written as,

$$\begin{aligned}S(\omega_1,\omega_1)S(\omega_1+\theta(1-h),\omega_1+\theta(1-h)) =&\Big(4A\pi\delta\{0\}+4B\pi\delta\{0\}\Big)\Big(4C\delta\{0\}+4D\delta\{0\}\Big),\\ =&\Big(4\pi\delta\{0\}(A+B)\Big)\Big(4\pi\delta\{0\}(C+D)\Big),\\ =&16\pi^2\delta\{0\}^2(A+B)(C+D).\end{aligned}$$

The denominator of the coherency is given by the square root of this product,

$$\begin{aligned}\sqrt{S(\omega_1,\omega_1)S(\omega_1+\theta(1-h),\omega_1+\theta(1-h))} =&\sqrt{16\pi^2\delta\{0\}^2(A+B)(C+D)},\\ =&4\pi\delta\{0\}\sqrt{(A+B)(C+D)}.\end{aligned}$$

The coherency for the off diagonal contribution when $\omega_2=\omega_1+\theta(1-h)$ is then given by,

$$\begin{aligned}\tau(\omega_1,\omega_1+\theta(1-h)) &= \frac{S(\omega_1,\omega_1+\theta(1-h))}{\sqrt{S(\omega_1,\omega_1)S(\omega_1+\theta(1-h),\omega_1+\theta(1-h))}},\\ &= \left(\frac{\pi^2\sigma_\zeta\sigma_\epsilon e^{i\phi_0}\delta\{0\}}{(1-2a\cos(\omega_1-h\theta)+a^2)}\right)\left(\frac{1}{4\pi\delta\{0\}\sqrt{(A+B)(C+D)}}\right),\\ &= \frac{\pi\sigma_\zeta\sigma_\epsilon e^{i\phi_0}}{4(1-2a\cos(\omega_1-h\theta)+a^2)}\frac{1}{\sqrt{(A+B)(C+D)}},\end{aligned}$$

as required. □

It is possible to find the Loève coherency along the line $\omega_2=\omega_1-\theta(1-h)$ by a similar procedure; this is given in the following corollary.

Corollary 5.4.15. *The coherency along the line $\omega_2=\omega_1-\theta(1-h)$ is given by*

$$\tau(\omega_1,\omega_1-\theta(1-h)) = \frac{\pi\sigma_\zeta\sigma_\epsilon e^{-i\phi_0}}{4(1-2a\cos(\omega_1-\theta)+a^2)}\frac{1}{\sqrt{(A+B)(\tilde{C}+\tilde{D})}},$$

where we have defined,

$$
\begin{aligned}
A &= \frac{\sigma_\epsilon^2}{1 - 2a\cos(\omega_1 - \theta) + a^2}, \\
B &= \frac{\sigma_\zeta^2}{1 - 2a\cos(\omega_1 - h\theta) + a^2}, \\
\tilde{C} &= \frac{\sigma_\epsilon^2}{1 - 2a\cos(\omega_1 - 2\theta + h\theta) + a^2}, \\
\tilde{D} &= \frac{\sigma_\zeta^2}{1 - 2a\cos(\omega_1 - \theta) + a^2}.
\end{aligned}
$$

Proof. The proof follows the same reasoning as in Theorem 5.4.14. □

5.4.5 Reparameterisation of the Loève coherency

Theorem 5.4.14 states that the coherency of the nonstationary polyfoil process along the line $\omega_2 = \omega_1 + \theta(1-h)$ is given by,

$$
\tau(\omega_1, \omega_1 + \theta(1-h)) = \frac{\pi\sigma_\zeta\sigma_\epsilon e^{i\phi_0}}{4(1 - 2a\cos(\omega_1 - h\theta) + a^2)} \frac{1}{\sqrt{(A+B)(C+D)}},
$$

where we have defined

$$
\begin{aligned}
A &= \frac{\sigma_\epsilon^2}{1 - 2a\cos(\omega_1 - \theta) + a^2}, \\
B &= \frac{\sigma_\zeta^2}{1 - 2a\cos(\omega_1 - h\theta) + a^2}, \\
C &= \frac{\sigma_\epsilon^2}{1 - 2a\cos(\omega_1 - h\theta) + a^2}, \\
D &= \frac{\sigma_\zeta^2}{1 - 2a\cos(\omega_1 + \theta - 2h\theta) + a^2}.
\end{aligned}
$$

This is a function of six parameters, namely: the damping parameter, the fundamental frequency, the harmonic multiple, the expected initial phase difference, and the noise terms of the fundamental and harmonic oscillations, $\{a, \theta, h, \phi_0, \sigma_\epsilon^2, \sigma_\nu^2\}$. If we introduce a reparameterisation, such that $\gamma_1 = \sigma_\epsilon\sigma_\zeta$ and $\gamma_2 = \sigma_\epsilon/\sigma_\zeta$, it is then possible to express

the coherency as a function of five parameters $\{a, \theta, h, \phi_0, \gamma_2\}$.

In order to implement the reparameterisation, denote,

$$\begin{aligned}
A^* &= 1 - 2a\cos(\omega_1 - \theta) + a^2, \\
B^* &= 1 - 2a\cos(\omega_1 - h\theta) + a^2, \\
C^* &= 1 - 2a\cos(\omega_1 - h\theta) + a^2, \\
D^* &= 1 - 2a\cos(\omega_1 + \theta - 2h\theta) + a^2,
\end{aligned}$$

such that the noise variance terms are separate. The product within the square root of the expression for coherency can then be written in terms of γ_1 and γ_2 as follows,

$$\begin{aligned}
(A+B)(C+D) =& AC + AD + BC + BD, \\
=& (\sigma_\epsilon^2)^2/A^*C^* + (\sigma_\epsilon^2\sigma_\zeta^2)/A^*D^* + (\sigma_\epsilon^2\sigma_\zeta^2)/B^*C^* + (\sigma_\zeta^2)^2/B^*D^*, \\
=& \gamma_1^2\gamma_2^2/A^*C^* + \gamma_1^2/A^*D^* + \gamma_1^2/B^*C^* + \gamma_1^2/(\gamma_2^2 B^*D^*).
\end{aligned}$$

Thus, the coherency can be written as,

$$\begin{aligned}
\tau(\omega_1, \omega_1 + \theta(1-h)) &= \frac{\pi\gamma_1 e^{i\phi_0}}{4\left(1 - 2a\cos(\omega_1 - h\theta) + a^2\right)} \\
&\quad \times \frac{1}{\sqrt{\gamma_1^2\gamma_2^2/A^*C^* + \gamma_1^2/A^*D^* + \gamma_1^2/B^*C^* + \gamma_1^2/\gamma_2^2 B^*D^*}}, \\
&= \frac{\pi e^{i\phi_0}}{4\left(1 - 2a\cos(\omega_1 - h\theta) + a^2\right)} \\
&\quad \times \frac{1}{\sqrt{\gamma_2^2/A^*C^* + 1/A^*D^* + 1/B^*C^* + 1/\gamma_2^2 B^*D^*}}.
\end{aligned}$$

Expressing the coherency in this way demonstrates that its magnitude is dependent on the ratio of the noise parameters $\gamma_2 = \sigma_\epsilon/\sigma_\zeta$, rather than the individual values. Furthermore, the above expression allows one to consider what happens to the coherency in limiting cases of the damping parameter, for example when $a = 0$ and $a \to 1$.

When $a = 0$, we note that $A^* = B^* = C^* = D^* = 1$ and hence,

$$\begin{aligned}\tau(\omega_1, \omega_1 + \theta(1-h)) &= \frac{\pi e^{i\phi_0}}{4} \frac{1}{\sqrt{\gamma_2^2 + 2 + 1/(\gamma_2^2)}}, \\ &= \frac{\pi e^{i\phi_0}}{4} \frac{1}{\sqrt{(\gamma_2 + 1/\gamma_2)^2}}, \\ &= \frac{\pi e^{i\phi_0}}{4(\gamma_2 + 1/\gamma_2)}.\end{aligned}$$

Thus, the magnitude of the coherency is dependent only on the ratio of the noises, and not the frequencies θ and $h\theta$. This is reflective of the fact that setting $a = 0$ generates a pure white noise process, since Z_t is no longer dependent on Z_{t-1}.

The case of $a \to 1$ requires more careful consideration. For simplicity, let us consider the value of the Loève coherency at the interaction point $\omega_1 = h\theta$ and $\omega_2 = \theta$. In this case, the above expressions simplify to

$$\begin{aligned}A^* &= 1 - 2a\cos(h\theta - \theta) + a^2, \\ B^* &= 1 - 2a + a^2, \\ C^* &= 1 - 2a + a^2, \\ D^* &= 1 - 2a\cos(h\theta - \theta) + a^2.\end{aligned}$$

Since we have chosen to set ω_1 and ω_2 to fixed values, we shall express the coherency as a function of the damping parameter a. In order to determine the behaviour of this function as $a \to 1$, denote $u(a) = 1 - 2a + a^2$ and $v(a) = 1 - 2a\cos(h\theta - \theta) + a^2$. We are therefore interested in finding,

$$\lim_{a\to 1}\tau(a) = \lim_{a\to 1}\frac{\pi e^{i\phi_0}}{4u(a)\sqrt{\gamma_2^2/u(a)v(a) + 1/v(a)^2 + 1/u(a)^2 + 1/\gamma_2^2 u(a)v(a)}}.$$

We begin by considering the behaviour of $v(a)$ and $u(a)$ in the limit. Evaluating these

expressions at $a = 1$ gives

$$u(1) = 1 - 2(1) + 1^2 = 0,$$
$$v(1) = 1 - 2\cos(h\theta - \theta) + 1^2 = 2 - 2\cos(h\theta - \theta).$$

From this, it is easy to see that the term $1/u(a)^2$ dominates the expression inside the square root as $a \to 1$. Therefore, the expression inside the square root can be approximated in the limit by,

$$\sqrt{\gamma_2^2/u(a)v(a) + 1/v(a)^2 + 1/u(a)^2 + 1/\gamma_2^2 u(a)v(a)} \approx \sqrt{1/u(a)^2} = 1/u(a).$$

Substituting this approximation into the expression for $\tau(a)$ then gives,

$$\tau(a) \approx \frac{\pi e^{i\phi_0}}{4u(a)} u(a) = \frac{\pi e^{i\phi_0}}{4}.$$

Thus, our final expression for the coherency evaluated at $(\omega_1, \omega_2) = (h\theta, \theta)$, as the parameter $a \to 1$, is given by,

$$\lim_{a \to 1} \tau(a) = \frac{\pi e^{i\phi_0}}{4}.$$

This behaviour is shown in Figure 5.4.1, where the magnitude of the reparameterised coherency at the interaction point $(h\theta, \theta)$ has been plotted as a function of $\gamma_2 = \sigma_\epsilon/\sigma_\zeta$, for different values of damping parameter a. We see that as $a \to 1$, the magnitude of the coherency tends towards $\pi/4$ for all values of γ_2.

5.5 Simulations

In the previous section, we showed that a nonstationary polyfoil process can be constructed by phase-locking the two oscillations, leading to interactions between distinct

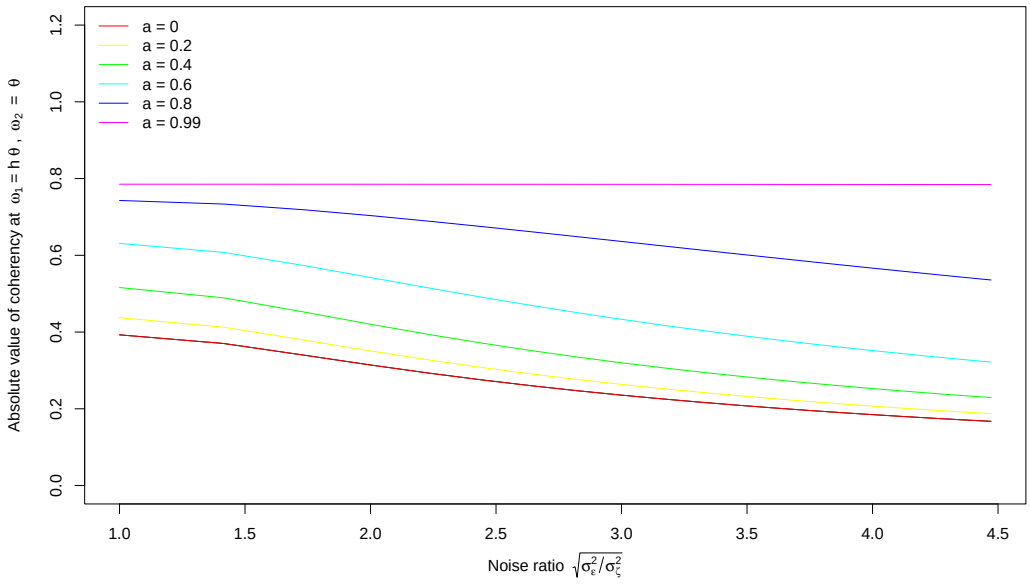


Figure 5.4.1: Magnitude of coherency at interaction point $(h\theta, \theta)$, function of $\gamma_2 = \sigma_\epsilon/\sigma_\zeta$. Fundamental frequency set to $\theta = 0.1\pi$, harmonic frequency set to 0.4π.

frequencies that can be observed in the Loève spectrum. This is in contrast to the stationary polyfoil process, which exhibits an observed Loève coherency matrix whose magnitude is given by the $N \times N$ identity matrix, where N is the number of observations in the time series; that is, there are no correlations between distinct frequencies.

We saw that the Loève spectrum for the nonstationary polyfoil was given by a sum of three components. The spectral density is therefore zero everywhere, apart from along three parallel lines: (i) there is a non zero contribution when the frequencies are equal, $\omega_1 = \omega_2$, (ii) there is a non zero contribution when $\omega_2 = \omega_1 + \theta(1-h)$, and (iii) there is a non zero contribution when $\omega_2 = \omega_1 - \theta(1-h)$. As shown in Section 5.4.4, this can then be used to derive a closed form expression for the Loève coherency, which is zero everywhere apart from along the aforementioned parallel lines.

In this section, we compare our theoretical results with simulations. In particular, we estimate the Loève coherency of simulated polyfoils with different parameter choices, and show that both the magnitude and phase of the estimated coherency correspond

to the theoretical derivations in Section 5.4. In order to estimate the Loève spectrum and subsequent Loève coherency from simulated data, we use the multitaper estimation method developed by Olhede and Ombao (2013); a brief description of this procedure is given in the following. In their work, Olhede and Ombao (2013) use their multitaper procedure to estimate the Loève coherence from several time aligned brain waves recorded from R trials. We also choose to consider many replications in our simulations, in order to provide better visualisations of the estimated Loève coherence.

In line with Olhede and Ombao (2013), denote X_t^r to be time series recorded at the rth trial. Olhede and Ombao (2013) define $h_t^{(k)}$ to be the k-th orthogonal taper that satisfies $\sum_t [h_t^{(k)}]^2 = 1$ and $\sum_t h_t^{(k_1)} h_t^{(k_2)} = 0$ if $k_1 \neq k_2$. The k-th orthogonal taper can then be used to define the k-th tapered Fourier coefficient at frequency f, which is given by,

$$x_k^{(r)}(f) = \sum_t h_t^{(k)} X_t^r \exp(-i2\pi ft), \quad f \in \left(-\frac{1}{2}, \frac{1}{2}\right).$$

For the frequency pair (f_1, f_2), Olhede and Ombao (2013) then define the k-th Loève periodogram to be

$$I_k^{(r)}(f_1, f_2) = x_k^{(r)}(f_1) x_k^{(r)*}(f_2),$$

where $(f_1, f_2) \in (-1/2, 1/2) \times (-1/2, 1/2)$. Thus, there is an estimate of the Loève periodogram for each taper $h_t^{(k)}$. In order to produce a suitable non-parametric estimator for the r-th trial, one can take the average of these estimates in order to produce the Loève multitaper spectral estimator,

$$\overline{I}^{(r)}(f_1, f_2) = \frac{1}{K} \sum_k I_k^{(r)}(f_1, f_2).$$

As stated, this estimator is for a single trial. In order to obtain an estimate for the

'population' Loève spectrum, that is, across the trails, Olhede and Ombao (2013) take the average of the trial specific estimators,

$$\overline{I}(f_1, f_2) = \frac{1}{R} \sum_r \overline{I}^{(r)}(f_1, f_2).$$

Furthermore, we recall that the coherency is given as a function of the Loève spectrum (see equation (5.3.2)). Olhede and Ombao (2013) therefore also estimate the trial-specific coherency using,

$$\hat{\tau}^{(r)}(f_1, f_2) = \frac{\overline{I}^{(r)}(f_1, f_2)}{\sqrt{\overline{I}^{(r)}(f_1, f_1)\overline{I}^{(r)}(f_2, f_2)}},$$

where the population coherency can be estimated by averaging across the R replications,

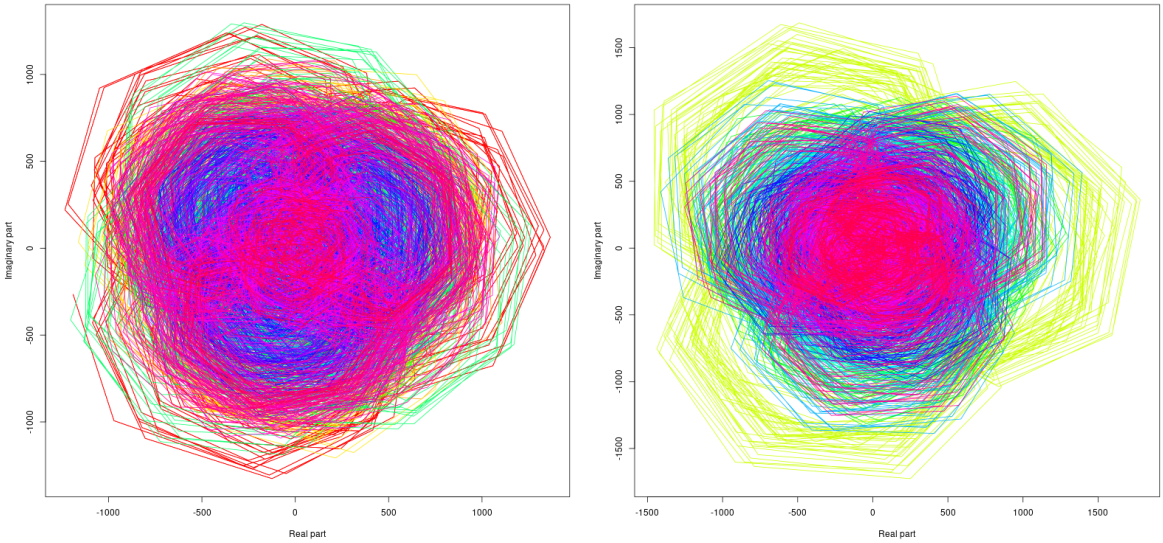
$$\hat{\tau}(f_1, f_2) = \frac{1}{R} \sum_r \hat{\tau}^{(r)}(f_1, f_2).$$

The polyfoils in this section were simulated from the Cholesky decomposition of the relevant covariance matrix; for stationary polyfoil simulation, we used the autocovariance sequence $s_{t,\tau} = s_\tau^{\{1\}} + s_\tau^{\{2\}}$, and for nonstationary polyfoil simulation, we used the expression given in Theorem 5.4.11. The values of θ, h, a, ϕ_0, σ_ϵ and σ_ζ are specified individually for each simulation. All polyfoils were simulated with length $N = 1000$, and $R = 100$ replications were carried out for each simulation. Estimates for the Loève spectrum were then obtained by averaging over the 100 trial specific estimators as above.

5.5.1 Magnitude of the Loève coherency

We begin by demonstrating the effects of nonstationarity on the magnitude of the Loève coherency. For visualisation purposes, we shall begin by assuming a relatively high value of damping parameter, $a = 0.999$. In order to ensure that a complete number of polyfoil

rotations are observed in the $N = 1000$ simulated time points, we must ensure that $N\theta/(2\pi)$ is an integer. Therefore, let us set the fundamental frequency to be $\theta = 2\pi/20$, and the harmonic multiple to be $h = 4$. Finally, let us set the standard deviations of the fundamental and harmonic noises to be $\sigma_\epsilon = 10$ and $\sigma_\zeta = 5$ respectively.

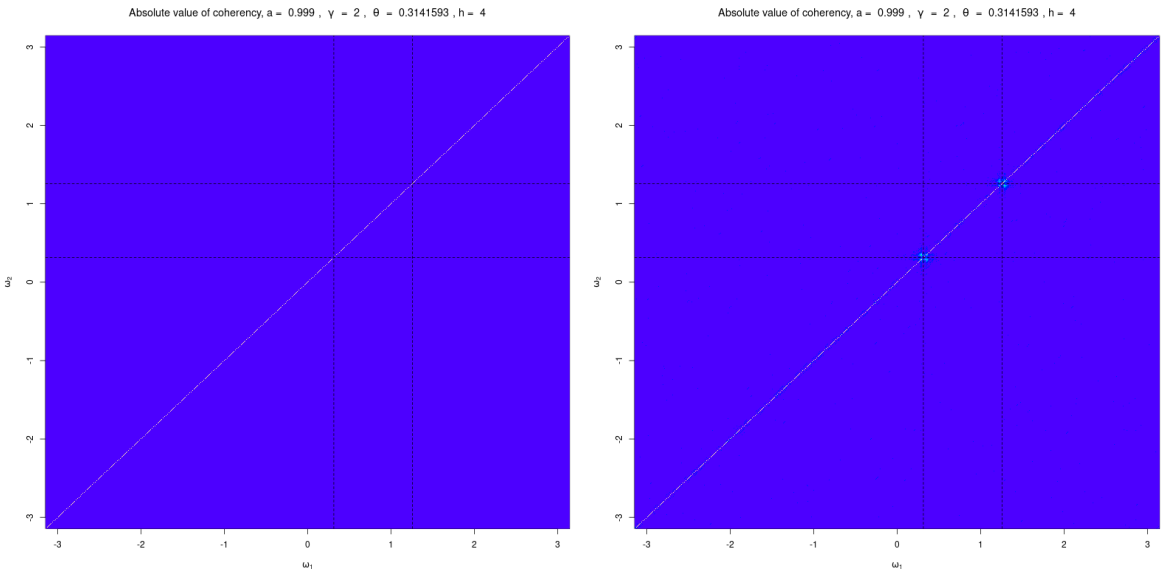


(a) Stationary polyfoils.

(b) Nonstationary polyfoils.

Figure 5.5.1: First 20 replications, polyfoils simulated using Cholesky decomposition of the relevant autocovariance sequence.

We simulated both stationary and nonstationary polyfoils with the aforementioned parameter choices. Recall, nonstationarity is induced by phase-locking the oscillations under expectation at general time t, according to equation (5.4.2). To begin, the phase parameter was set to $\phi_0 = 0.5$. Figure 5.5.1a shows 20 polyfoils simulated from the stationary autocovariance sequence, while Figure 5.5.1b shows 20 polyfoils generated from the nonstationary autocovariance sequence. When viewing the polyfoil replications in the complex plane as in Figure 5.5.1, there is no clear indication of which are stationary and which are not. In both cases, the polyfoils change between replications, both in terms of their size and orientation; this is a consequence of the fact the Cholesky method draws the initial position of each replication from a complex-valued Gaussian.

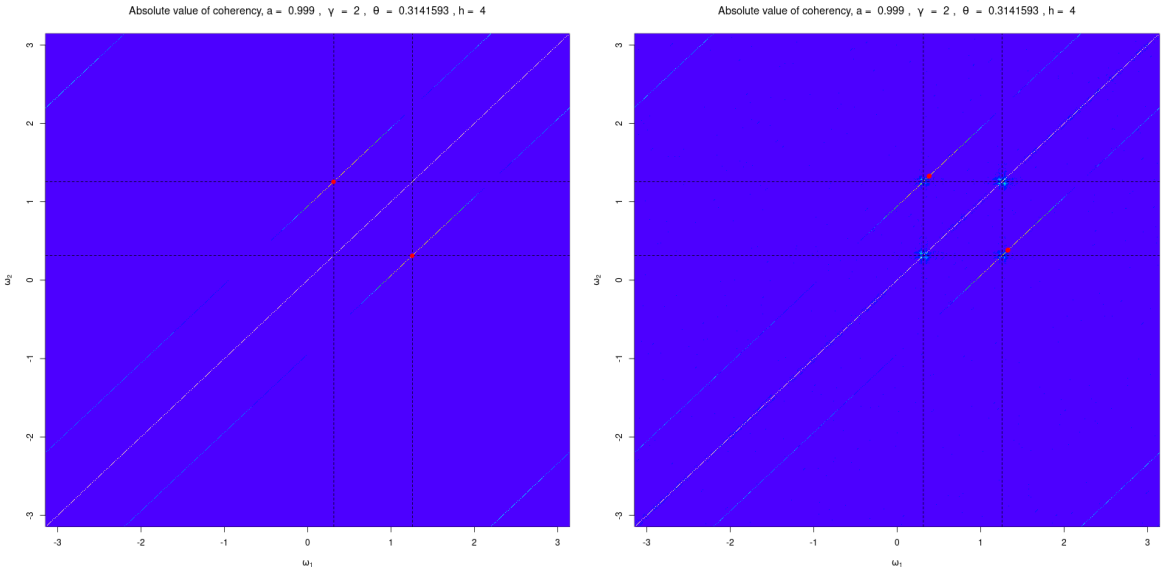


(a) Theoretical coherency matrix. (b) Estimated coherency matrix.

Figure 5.5.2: Magnitude of coherency matrix for stationary polyfoil. Position of θ and $h\theta$ marked by dashed black lines.

However, it becomes apparent which polyfoils are stationary and which are non-stationary when we estimate the Loève coherency using the multitaper method. The magnitude of the estimated coherency for the stationary polyfoil model is compared with the stationary theoretical coherency in Figure 5.5.2. The black dashed lines have been added to indicate the positions of θ and $h\theta$. We see that the estimated magnitude of the Loève coherency displays no interactions between distinct frequencies, as expected from the theory.

On the other hand, Figure 5.5.3 compares the magnitude of the estimated coherency for the nonstationary polyfoils with the theoretical result. Clearly, the nonstationary coherency displays interactions between distinct frequencies along the lines $\omega_2 = \omega_1 + \theta(1-h)$ and $\omega_2 = \omega_1 - \theta(1-h)$, as expected from the theory. Furthermore, the red points indicate the maximum values along these lines. We see that these are very close to the interaction points $(\theta, h\theta)$ and $(h\theta, \theta)$, which is where the true theoretical maximums occur. We also note the parallel lines in the top-left and bottom-right corners of these plots which are an effect of the 2π-periodicity of the coherency matrix, as discussed



(a) Theoretical coherency matrix. Maximum values on off-diagonal interactions marked by red points.

(b) Estimated coherency matrix. Maximum values on off-diagonal interactions marked by red points.

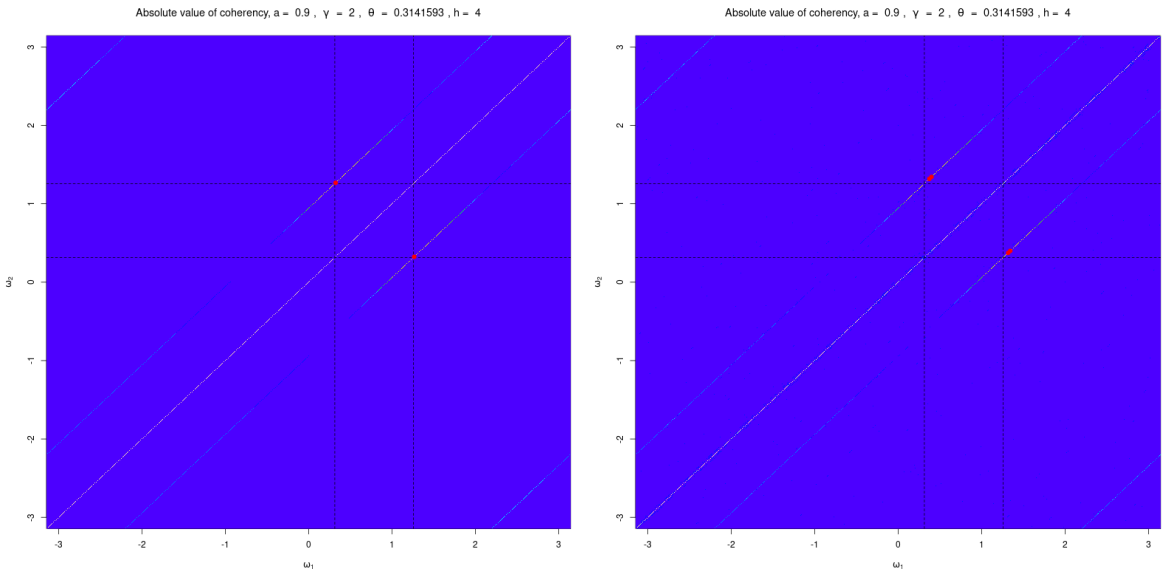
Figure 5.5.3: Magnitude of coherency matrix for nonstationary polyfoil, $a = 0.999$. Position of θ and $h\theta$ marked by dashed black lines.

previously.

Hence, while viewing the polyfoils in the complex plane does not indicate whether or not the process is stationary, this can instead be determined by computing the Loève coherency. Furthermore, the position of the off-diagonal interactions indicate both the fundamental frequency and the harmonic frequency.

Of course, the choice of $a = 0.999$ is quite high. The magnitude of the estimated Loève coherency of the nonstationary polyfoil process with $a = 0.9$ is compared with the theoretical result in Figure 5.5.4. We see that even with reduced choice of a, the interactions along the lines $\omega_2 = \omega_1 + \theta(1-h)$ and $\omega_2 = \omega_1 - \theta(1-h)$ are still clearly visible.

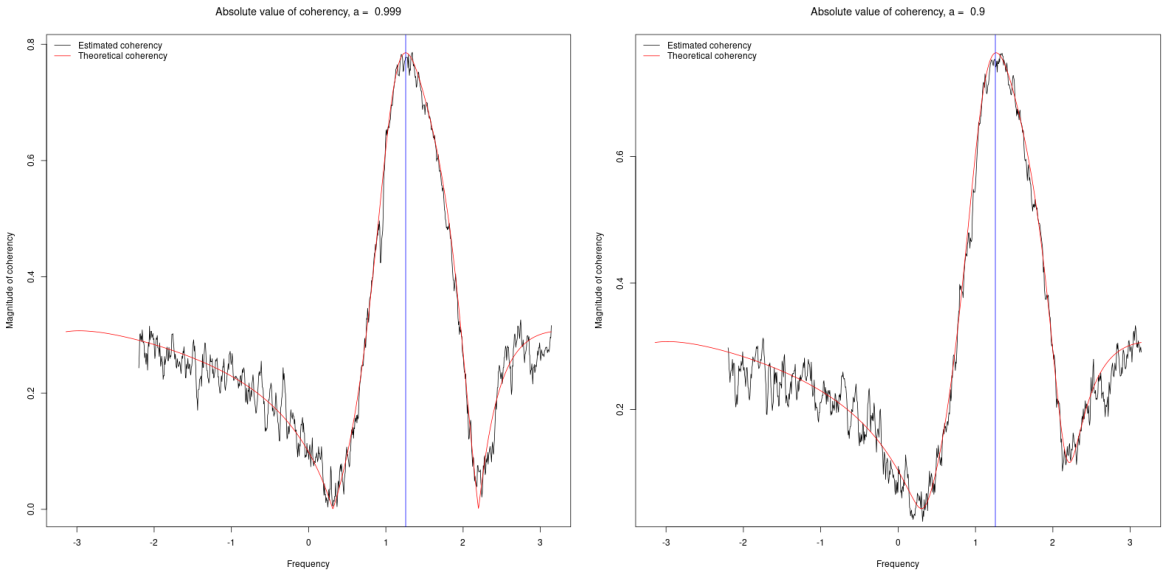
It is also possible to carry out a more precise analysis by comparing the theoretical results with the estimated coherency at a specific set of frequency pairs. We compared the magnitude of the estimated coherency along the line $\omega_2 = \omega_1 + \theta(1-h)$ with the theoretical expression, for $a = 0.9$ and $a = 0.999$. This is shown in Figure 5.5.5,



(a) Theoretical coherency matrix. Maximum values on off-diagonal interactions marked by red points.

(b) Estimated coherency matrix. Maximum values on off-diagonal interactions marked by red points.

Figure 5.5.4: Magnitude of coherency matrix for nonstationary polyfoil, $a = 0.9$. Position of θ and $h\theta$ marked by dashed black lines.



(a) Damping parameter $a = 0.999$.

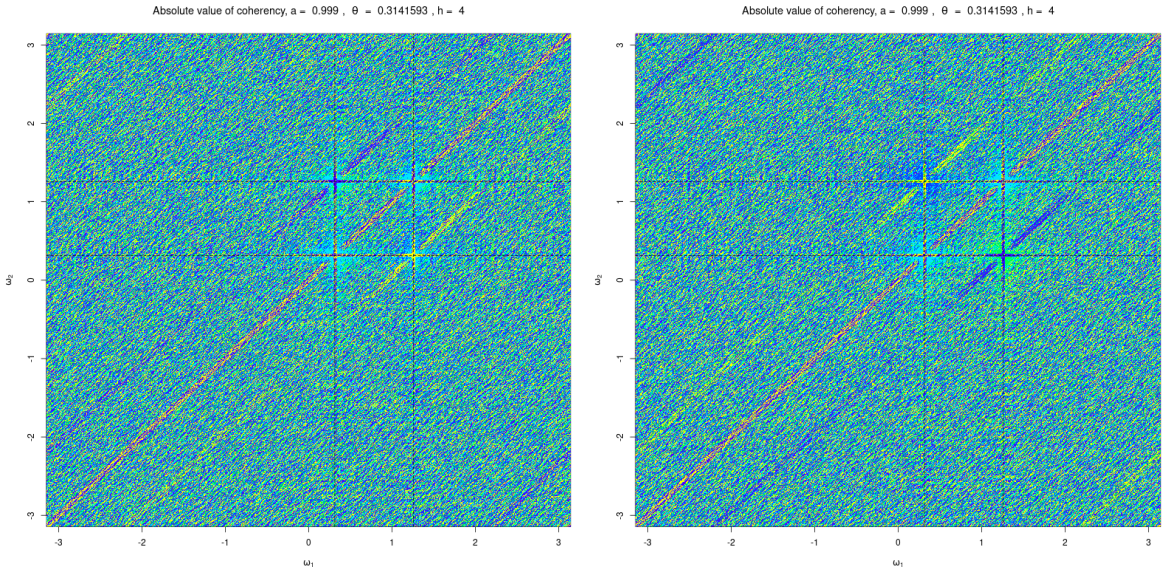
(b) Damping parameter $a = 0.9$.

Figure 5.5.5: Estimated magnitude of coherency for nonstationary polyfoils across replications, along line $\omega_2 = \omega_1 + \theta(1 - h)$. Theoretical magnitude of coherency shown by red line.

where the frequency corresponding to the theoretical maximum of the magnitude of the coherency is shown by the blue line; this is at $h\theta$. We see that the multitaper estimation method of Olhede and Ombao (2013) successfully captures the theoretically derived behaviour of the Loève coherency along this line.

5.5.2 Phase of the Loève coherency

The previous simulations consider the magnitude of the Loève coherency. However, when the polyfoil process is nonstationary, the Loève coherency is given by a complex quantity; hence, it is also possible to study the phase. Theorem 5.4.14 provides the expression for the Loève coherency of the nonstationary polyfoil model along the line $\omega_2 = \omega_1 + \theta(1-h)$; we see that this includes the term $e^{i\phi_0}$. Similarly, the expression for the Loève coherency along the line $\omega_2 = \omega_1 - \theta(1-h)$ includes the term $e^{-i\phi_0}$. Therefore, we expect the phase of the Loève coherency estimated using the multitaper method to be given by ϕ_0 and $-\phi_0$ along the aforementioned parallel lines.

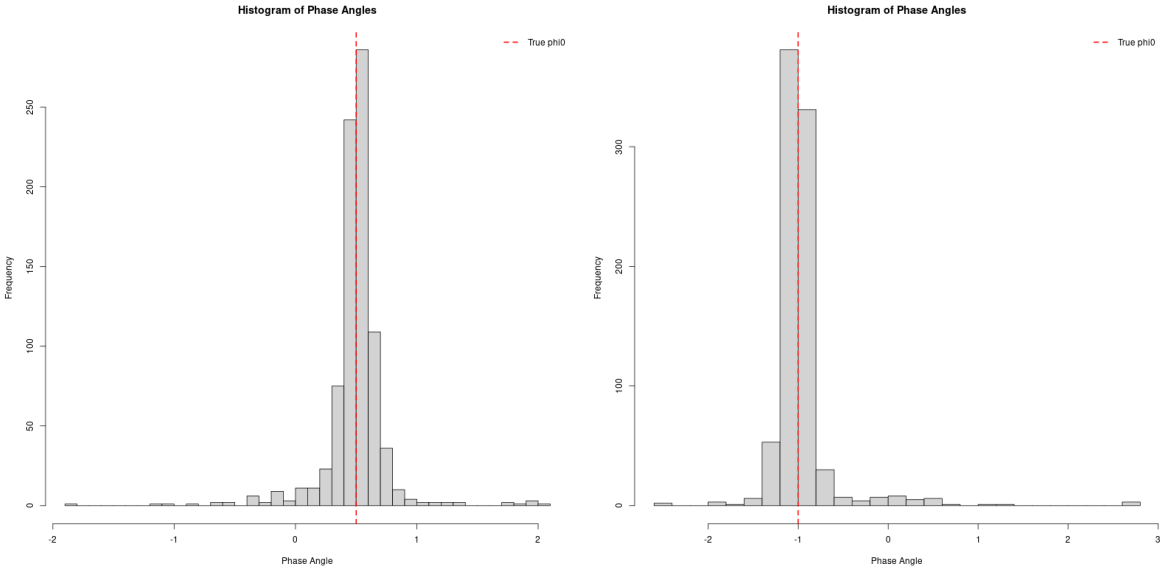


(a) Expected initial phase set to $\phi_0 = 0.5$.

(b) Expected initial phase set to $\phi_0 = -1$.

Figure 5.5.6: Estimated phase of coherency for nonstationary polyfoils with $a = 0.999$. Position of θ and $h\theta$ marked by dashed black lines.

We confirm that this is the case for different values of a and ϕ_0 in the following. To begin, the damping parameter was set to $a = 0.999$. Figure 5.5.6 shows (a) the phase of the estimated Loève coherency when $\phi_0 = 0.5$, and (b) the phase of the estimated Loève coherency when $\phi_0 = -1$. We see that when ϕ_0 is positive, the estimated coherency phase along the line $\omega_2 = \omega_1 + \theta(1-h)$ is also positive, while the phase along $\omega_2 = \omega_1 - \theta(1-h)$ is negative. When the initial expected phase difference is set to a negative number, the signs of these two lines switch in accordance with the theory.



(a) Expected initial phase set to $\phi_0 = 0.5$.

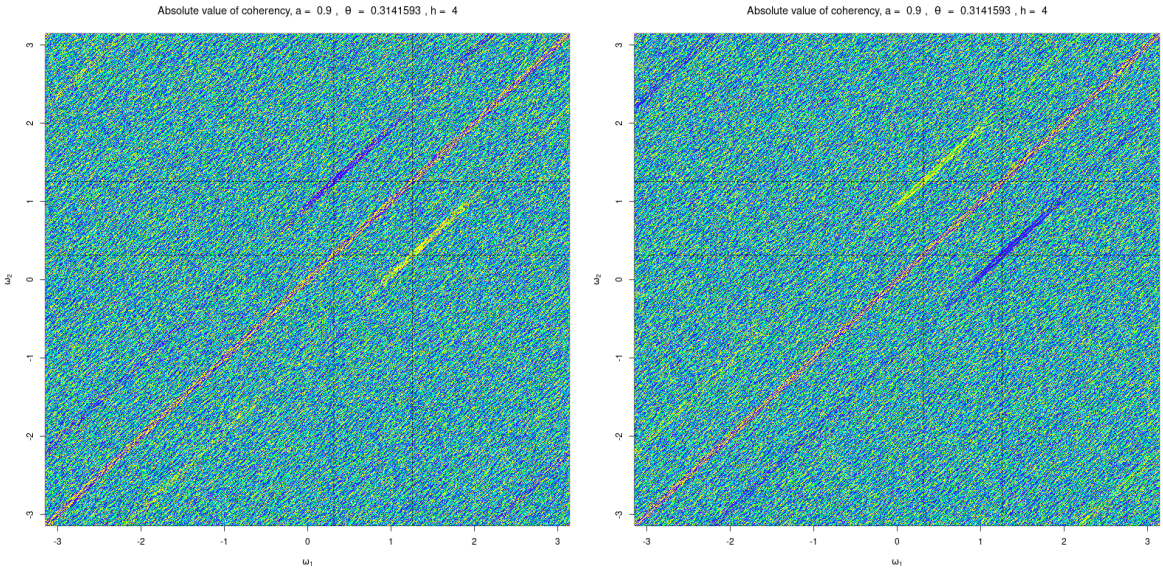
(b) Expected initial phase set to $\phi_0 = -1$.

Figure 5.5.7: Estimated phase of coherency for nonstationary polyfoils across replications, along line $\omega_2 = \omega_1 + \theta(1-h)$. True initial expected phase ϕ_0 shown by dashed red line.

We can also examine the spread of the phase of the estimated Loève coherency along the two parallel lines of interaction. Figure 5.5.7 shows a histogram of the mean phase values along the line $\omega_2 = \omega_1 + \theta(1-h)$, for (a) $\phi_0 = 0.5$, and (b) $\phi_0 = -1$. The true values of ϕ_0 are shown by the red dashed lines. We see that the estimated phase along the line of interaction is narrowly distributed around the theoretically true value in both cases, as desired.

We also considered the phase of the estimated Loève coherency for the nonstationary

polyfoil process with $a = 0.9$. Figure 5.5.8 depicts the phase of the estimated Loève coherency matrix when (a) $\phi_0 = 0.5$, and (b) for $\phi_0 = -1$. The spread of the phase of the estimated Loève coherency along the line $\omega_2 = \omega_1 + \theta(1 - h)$ for the two values of ϕ_0 is shown in Figure 5.5.9. The phase estimation remains stable for this more realistic value of a.



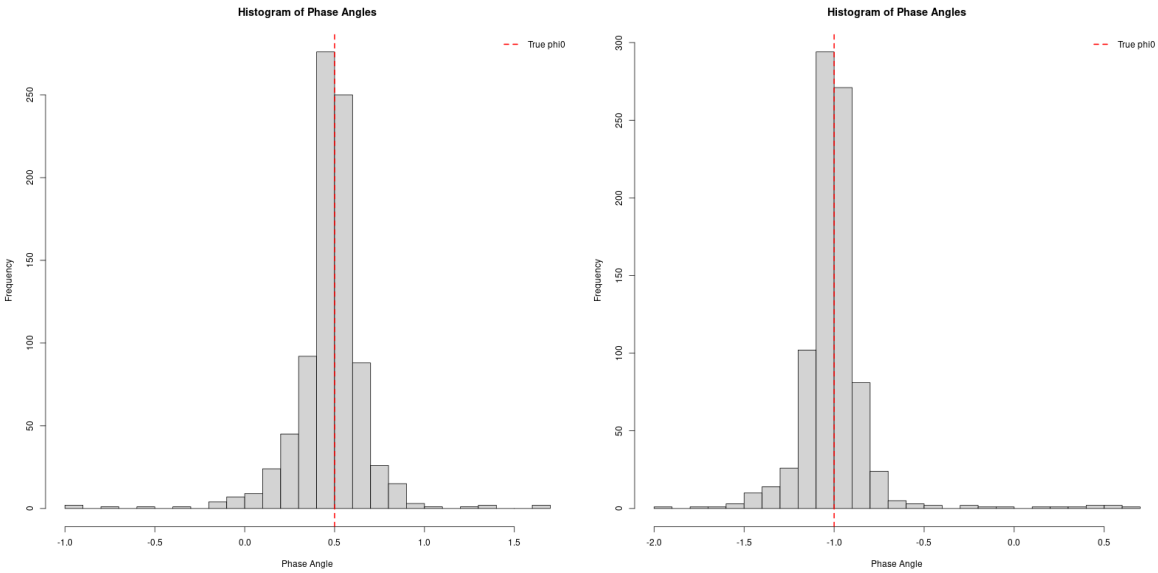
(a) Expected initial phase set to $\phi_0 = 0.5$.

(b) Expected initial phase set to $\phi_0 = -1$.

Figure 5.5.8: Estimated phase of coherency for nonstationary polyfoils with $a = 0.9$. Position of θ and $h\theta$ marked by dashed black lines.

5.6 Discussion

In this chapter, we proposed a stochastic process for modelling two interacting oscillatory signals of different frequencies. This is a logical extension to the complex-valued AR(1) process, and is generated by superposing two such processes together. By superposing two complex AR(1) processes in this way, we defined a novel stochastic process which depicts 'polyfoil' shapes when plotted in the complex plane (see Figure 5.1.2 for examples). The key novelty of the polyfoil stochastic process is that we can construct



(a) Expected initial phase set to $\phi_0 = 0.5$.

(b) Expected initial phase set to $\phi_0 = -1$.

Figure 5.5.9: Estimated phase of coherency for nonstationary polyfoils across replications, along line $\omega_2 = \omega_1 + \theta(1-h)$. True initial expected phase ϕ_0 shown by dashed red line.

nonstationarity by assuming that oscillations are phase-locked under expectation at general time t. This differs from other works on stochastic oscillations in the literature, for example the widely linear complex AR(1) process of Sykulski et al. (2016), which consider only stationary oscillations.

We demonstrated that when oscillations are not phase-locked, the polyfoil stochastic process is stationary. In this case, the autocovariance sequence for the process is simply given by summing the individual autocovariance sequences of the two oscillatory signals, and is therefore only dependent on the lag τ. On the other hand, we demonstrated that a nonstationary polyfoil process can be constructed by enforcing phase-locking on the oscillations. In this case, the autocovariance sequence features a nonstationary component $s_{t,\tau}^{\{3\}}$, which dependent on both lag τ and time t.

When a stochastic process is nonstationary, correlations may exist between distinct frequencies. In such cases, standard spectral analysis techniques do not fully characterise the process, and one must consider alternative methods. It is possible to observe

such interactions in the Loève spectrum, which is a complex-valued quantity given by taking the two-dimensional Fourier transform of the autocovariance function. In order to quantify the correlations between frequencies ω_1 and ω_2, one may compute the Loève coherency $\tau(\omega_1, \omega_2)$, which is a function of the Loève spectrum. The Loève coherence at (ω_1, ω_2) is then given by $\rho(\omega_1, \omega_2) = |\tau(\omega_1, \omega_2)|^2$, which lies in the range $[0, 1]$. Computation of the Loève coherency therefore provides insight on the characterisation of nonstationary processes.

In this chapter, we derived the analytical form of the Loève spectrum and the Loève coherency for our polyfoil stochastic process. We demonstrated that these quantities depend on six parameters, $\{a, \theta, h, \sigma_\epsilon^2, \sigma_\zeta^2, \phi_0\}$, and are zero-valued everywhere, apart from along the diagonal line $\omega_1 = \omega_2$, and the parallel lines $\omega_2 = \omega_1 + \theta(1 - h)$ and $\omega_2 = \omega_1 - \theta(1-h)$. The maximum values of the magnitude of the coherency along the off-diagonal lines correspond to the interaction points $(h\theta, \theta)$ and $(\theta, h\theta)$. By introducing a reparameterisation such that $\gamma_2 = \sigma_\epsilon/\sigma_\zeta$, we showed that the coherency can instead be written in terms of five parameters $\{a, \theta, h, \gamma_2, \phi_0\}$. This allowed us to consider what happens to the magnitude of the coherency for limiting cases of the damping parameter a, for fixed θ, h, and γ_2. We saw that when the damping parameter $a = 0$, the magnitude of the coherency depends only on the ratio of the noises γ_2, which reflects the fact that the polyfoil model defines a pure white noise process in this case. Furthermore, this reparameterisation may also be beneficial in the future for practical applications, since fewer parameters now require estimation.

Simulations were used to demonstrate the distinction between the stationary and nonstationary polyfoil process, and to show agreement with the theoretically derived results. For each choice of parameters, we simulated 100 realisations of the stochastic process. In order to motivate the need for Loève spectrum analysis, we began by plotting the first twenty polyfoils simulated from both the stationary and nonstationary autocovariance sequences in the complex plane. We saw that there was no clearly

visible distinction between the two, indicating that one cannot identify the stationarity of a polyfoil process by simply plotting the realisations in this way. We therefore used the multitaper method of Olhede and Ombao (2013) in order to estimate the Loève spectrum and Loève coherency of the simulated processes. This estimation showed how nonstationary processes display nonzero values in the off-diagonals of the Loève spectrum (and Loève coherency) matrix, and demonstrated that the theoretical derivations correctly captured the exhibited behaviour.

To conclude this chapter, we highlight some key areas for further work. The results presented in this chapter provide a substantial step in performing parametric inference fitting the nonstationary polyfoil model to data, and clearly this is the immediate next step in our future work. It is possible to estimate the parameters $\{a, \theta, h, \sigma_\epsilon^2, \sigma_\zeta^2\}$ (but not ϕ_0) from the stationary model, as the stationary autocovariance/spectrum contains all these parameters, even for a nonstationary process. However, it would be interesting to investigate whether the inference on these parameters could be improved by including the extra information available in the nonstationary model.

To begin, one could consider maximum likelihood methods in order to estimate the parameters of our polyfoil process in the time domain. Under the assumption of complex Gaussian error sequences, the polyfoil process can be modelled as a complex-valued Gaussian process. Thus, the likelihood can be computed based on the assumption that a vector of N complex-valued observations $\mathbf{Z}$ follows a multivariate complex-valued normal distribution,

$$\mathbf{Z} \sim N(\mathbf{0}, \mathbf{\Sigma}),$$

where the pseudo-covariance or relation matrix is 0 for polyfoil oscillations as discussed earlier. Here $\mathbf{\Sigma}$ is an $N \times N$ covariance matrix between all pairs of observations, which could be computed using the expression for autocovariance of the nonstationary polyfoil model given in Theorem 5.4.11. The log-likelihood of the observed vector $\mathbf{Z}$ is then given

by the log-likelihood of the multivariate normal distribution,

$$\log L(\mathbf{Z}) = \frac{N}{2}\log(2\pi) - \frac{1}{2}\log(\det(\mathbf{\Sigma})) - \frac{1}{2}\mathbf{Z}^{\dagger}\mathbf{\Sigma^{-1}}\mathbf{Z}.$$

Inversion of the matrix $\mathbf{\Sigma}$ is computationally expensive in the event that the number of observations N is large, or in the case that we choose to perform inference over many replications (as in the simulations in Section 5.5). In order to address this, frequency domain methods for inference have been proposed. One such method is the 'Whittle' likelihood (Whittle (1953)), which is a frequency domain approximation to the maximum likelihood. However, this procedure requires stationarity, which is not the case for the polyfoil process. Therefore, one possible avenue for further work would be to consider a dual-frequency version of the Whittle likelihood, with the aim of developing a more computationally efficient inference procedure for our model.

Another important avenue for future work is to consider more general extensions to the polyfoil process. In this chapter, we assumed that the damping parameter a was common between the fundamental and harmonic oscillations. Alternatively, it would be possible to consider different damping parameters for each. Furthermore, it was assumed in our derivations that the noise terms were uncorrelated. Our model could be extended by introducing correlations between these variables.

Finally, we seek to find an application dataset to test our model in the real-world. An obvious candidate is the polyfoil patterns observed recently in Zheng et al. (2024) in oceanographic trajectories of particle motion.

Chapter 6

Conclusions and further work

In this thesis, we have considered two distinct nonstationary multivariate time series problems: the combination of N expert forecasts, and the modelling of two interacting oscillations using a novel nonstationary stochastic process. There exists a wealth of literature concerning the former, wherein a variety of different methods for forecast combination have been developed; some of these are outlined in Chapter 2. It is often the case that different methods are suited to different applications, and therefore it is not uncommon for forecasting methodologies to be developed with a specific operation in mind. Rather than a specific application, our research was instead motivated by the consideration of a specific forecasting setting. We assumed predictions were provided sequentially by a set of N forecasters, in advance of the variable of interest y_t being observed. This necessitated the development of a computationally efficient combination method which could be implemented in an online manner. Furthermore, we assumed that forecaster quality was changing with time, and that correlations existed between forecasters.

For Chapter 5, we considered a distinctly different nonstationary multivariate time series problem. Namely, we developed a novel stochastic process for modelling two interacting oscillations of different frequencies, θ and $h\theta$. When realisations of such

a process are plotted in the complex plane, the corresponding trajectory maps out a 'polyfoil' shape, and as such we defined our process as the polyfoil process. While such stochastic oscillations have been studied in the literature, this is often under the assumption of stationarity. Instead, we constructed a specific form of nonstationarity by phase-locking the fundamental and harmonic oscillations.

While clearly disparate in terms of their focus and subject areas, both problems require data to be described in a joint way, which incorporates relationships and interdependencies across variables. Furthermore, both problems exhibit nonstationarity. In the forecast combination problem, nonstationarity can arise from time-varying forecaster quality, and must be dealt with through dynamic combination techniques, while in Chapter 5 a nonstationary oscillatory process was constructed. In this chapter, we conclude our work by summarising key contributions. The research presented in Chapters 3 and 4 is quite open-ended, and therefore we also highlight some possible avenues for further work that have not yet been suggested. On the other hand, the direction of further work concerning the polyfoil stochastic process is more explicit and has already been discussed in Section 5.6, hence we will not deal with this further.

6.1 Key contributions

The first contribution of this thesis was presented in Chapter 3, wherein we introduced a Dynamic Linear Model (DLM)-based procedure for combining N forecasts in an online manner, assuming no missing forecaster data. The use of DLMs to combine forecasts has been notionally investigated in the literature by LeSage and Magura (1992) and Sessions and Chatterjee (1989), both of whom choose to model the evolution of the combination weights as a random walk. However, the contribution presented in Chapter 3 of this thesis is the first to formalise the procedure by providing a complete step-by-step guide for the combination of point forecasts within the DLM framework. We explicitly state

the roles of key parameters in the model, and provide clear methods for dealing with unknown values, in addition to suitable prior parameter choices. In order to do this, we give focus to the role of the discount factor parameter δ, and provide a reformulation of the Kalman filtering equations in terms of this.

The DLM-framework enables inference on relevant distributions to be updated sequentially as more data become available, meaning that it is suitable for applications in which forecasts arrive in quick succession. Distributions are updated by a closed-form procedure, which outputs a forecast distribution for the variable of interest at each time t, in addition to a distribution for the vector of combination weights (combination weight estimates can be obtained by taking the expectation). Since weights are modelled according to a random walk, dynamics in the combination weights can be captured in order to reflect changing forecaster quality over time. Unlike other performance-based weighting schemes, such as the regression based weights of Granger and Ramanathan (1984) and the optimal covariance weights of Newbold and Granger (1974), our DLM-based forecast combination method provides weight estimates from the outset; this makes it a suitable choice in applications with little historical information. Furthermore, unlike simple weighting schemes such as simple averaging, and more complex weighting schemes such as machine learning based methods, our method offers an interpretability regarding forecaster quality from the assigned weights; for uncorrelated forecasters, a higher weight reflects a superior forecast quality.

Chapter 4 of this thesis considers the problem of missing forecaster data. Despite being a frequent feature of empirical data (as evidenced by the AMOC and ECB data sets), the question of how to deal with a forecaster going 'offline' has been scarcely considered in the literature. Motivated by Lahiri et al. (2017) (who discuss the challenges of different combination algorithms implicitly imputing missing forecasters in different ways), and the difficulties associated with just combining the available forecasters, we decided that the best approach would to be impute missing forecaster data prior to

combining. However, in our online setting, simple imputation techniques like linear interpolation and smoothing cannot be applied. Instead, for the second contribution of this thesis we proposed dealing with missing forecasts by modelling each individual forecaster as a DLM.

Fitting a DLM to each forecaster series enables missing forecaster data to be dealt with implicitly, as part of the filtering procedure. Application of the Kalman filter enables the computation of forecasting distributions for each forecaster throughout time, from which imputed values may be obtained in missing periods. Providing that a suitable DLM is chosen, such that the temporal evolution of the forecaster series is adequately described, the imputed values will be more accurate than simply 'filling forward'. Moreover, in periods of missing data, the spread of the forecast distribution will increase in order to reflect the uncertainty associated with the imputed forecast.

We discussed how filtering the forecaster series in this way provides the practitioner with a choice: should point forecasts be combined, or density forecasts? We noted that, although the imputed values will be more accurate than those offered by simply 'filling forward', it is unlikely that this will provide a better forecast in the point forecast framework than the available, online forecasters. Consequently, we expect the quality of the missing forecaster to suddenly drop at the start of the missing period, and increase when it returns online. In order to deal with this, we introduced a procedure for point forecast combination with adaptive discount factor δ_t. This was achieved by integrating the particle-learning based parameter estimation procedure of Irie et al. (2022) with our DLM-based point forecast combination methodology introduced in Chapter 3.

We saw that this method performed well in the case of sporadic missing data, obtaining the lowest MSE, MAE and SMAPE. On the other hand, this method struggled to respond to improved forecaster performance after very large periods of missing data.

We also considered the combination of density forecasts in this chapter. As an extension of Hall and Mitchell (2007), we proposed computing dynamic mixture weights by

maximising the exponentially weighted log predictive score at each time. This allowed the increase in uncertainty of the forecasting distribution caused by missing data to be explicitly incorporated into the combination. We saw how this resulted in improved forecasting performance after a forecaster had been offline for an extended period of time when compared to the point combination based approaches.

As the final contribution of this thesis, we presented the novel stochastic 'polyfoil' process. This was developed as a natural extension to the complex valued AR(1) process, and defined by summing together two individual complex valued AR(1) processes with different spin parameters, representing a fundamental and a harmonic oscillation. Unlike other work on stochastic oscillations in the literature, our method allowed us to induce nonstationary stochastic oscillations by setting the expected initial phase difference between the fundamental and harmonic oscillation to a constant value, which we termed 'phase-locking'.

Nonstationarity leads to interactions between distinct frequencies that can be observed in the Loève spectrum. For the nonstationary polyfoil process, we derived closed-form expressions for the autocovariance sequence, Loève spectrum and Loève coherency. These theoretical results were shown to align with simulations in Section 5.5.

6.2 Further work

Our DLM-based forecast combination method, introduced in Chapter 3, was shown to outperform almost all benchmarks in a range of simulation studies for different scenarios: independent forecasters, correlated forecasters, stationary forecaster quality, and changing forecaster quality. Furthermore, our method outperformed all benchmarks when applied to the AMOC data set, for all values of discount factor δ used in the analysis.

However, we noted that our DLM-based forecast combination method was outper-

formed by the 'recent best' method and simple averaging in terms of the MSE, MAE and SMAPE error metrics for the ECB data set. We discussed how this may be a consequence of the high levels of correlation present between the forecasters, and provided a simple toy example investigating the effects of trimming negative weights on forecasting performance. An obvious avenue for further work would be a more thorough exploration of the effects of trimming in both simulations and empirical applications. In particular, it would be interesting to see whether implementing a trimming procedure such as in Radchenko et al. (2023) would improve the forecasting performance of our method on the ECB data set.

It is possible to implement trimming as part of a two-step procedure, wherein weights are first estimated, and then trimmed, or a one-step procedure, wherein weights are estimated and trimmed in a single step via constrained optimisation. In the toy example in Chapter 3, we demonstrated how a two-step procedure could be implemented. An obvious extension would be to apply this procedure to a large scale simulation study, and investigate the effects of trimming on forecasting performance for different levels of correlation and different trimming thresholds.

An alternative method for trimming would be to carry out constrained Kalman filtering, which would enable the computation of non-negative weights as part of the forecast combination procedure itself. In particular, the constrained Kalman filter considers the case where the system satisfies the equality constraints,

$$\mathbf{D}\boldsymbol{\theta}_t = \mathbf{d},$$

or the inequality constrains

$$\mathbf{D}\boldsymbol{\theta}_t \leq \mathbf{d},$$

where $\mathbf{D}$ is a known matrix and $\mathbf{d}$ is a known vector; recall the state vector $\boldsymbol{\theta}_t$ is given by

the vector of forecast combination weights in our application. There are several different methods for modifying the Kalman filter in order to accommodate such constraints; Simon (2010) provides a survey of linear and nonlinear algorithms for this purpose. As detailed by Simon (2010), one possible method is to project the unconstrained state vector estimate $\hat{\boldsymbol{\theta}}_t$ onto the constrained surface. For the case of equality constraints, the constrained estimate can then be written as,

$$\tilde{\boldsymbol{\theta}}_t = \left\{ \underset{\boldsymbol{\theta}}{\operatorname{argmin}} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_t)' \mathbf{W} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_t) : \mathbf{D}\boldsymbol{\theta}_t = \mathbf{d} \right\},$$

where $\mathbf{W}$ is a positive definite weighting matrix. The solution to this is then given by,

$$\tilde{\boldsymbol{\theta}}_t = \hat{\boldsymbol{\theta}}_t - \mathbf{W}^{-1}\mathbf{D}'(\mathbf{D}\mathbf{W}^{-1}\mathbf{D}')^{-1}(\mathbf{D}\hat{\boldsymbol{\theta}}_t - \mathbf{d}).$$

As stated by Simon (2010), if the process and measurement noises are Gaussian, and the positive definite weighting matrix is set to $\mathbf{W} = (\mathbf{C}_t)^{-1}$, then we obtain the maximum probability estimate of the state subject to state constraints at time t. If instead we set $\mathbf{W} = \mathbb{I}_2$, we obtain the least squares estimate of the state vector subject to the constraints. For the case of inequality constraints, the constrained estimate for the state vector can be written as,

$$\tilde{\boldsymbol{\theta}}_t = \left\{ \underset{\boldsymbol{\theta}}{\operatorname{argmin}} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_t)' \mathbf{W} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_t) : \mathbf{D}\boldsymbol{\theta}_t \leq \mathbf{d} \right\}.$$

Finding this estimate is a quadratic programming problem, and can be solved used interior point approaches or active set methods.

It would be possible to incorporate this projection estimation into our forecast combination procedure in order to ensure that the estimated combination weights are non-negative; however, we note that this would lead to increased computational complexity. A comparison of this one-step trimming procedure with the two step trimming proce-

dure described in Chapter 3 could be undertaken in future work.

In Chapter 4, challenges and extensions to the the adaptive discounting combination approach and the dynamic density combination method were given. In this section, we instead highlight a possible avenue for further work regarding the modelling of the individual forecasters.

We proposed modelling each forecaster as a univariate DLM, wherein only the previous forecasts from that forecaster were taken into account. That is, the previous forecasts from the other $N-1$ experts are not included in the model. This can be carried out in parallel, and is computationally efficient; however, it is natural to question whether all N forecasters could be modelled together in a multivariate framework instead. A multivariate DLM would take into account the correlations present between the forecaster series. When a particular forecaster goes offline, this would be particularly beneficial, since we would still gain some information with which to update the distribution for the missing forecaster.

The full definition of a general multivariate DLM is given by West and Harrison (1997). In short, the model equations for a multivariate DLM are given by,

$$\begin{aligned} \mathbf{y}_t &= \boldsymbol{F}_t'\boldsymbol{\theta}_t + \boldsymbol{\nu}_t, \qquad \boldsymbol{\nu}_t \sim N[\mathbf{0}, \mathbf{V}_t], \\ \boldsymbol{\theta}_t &= \boldsymbol{G}_t\boldsymbol{\theta}_{t-1} + \boldsymbol{\omega}_t, \quad \boldsymbol{\omega}_t \sim N[\mathbf{0}, \boldsymbol{W}_t]. \end{aligned}$$

Therefore, the key difference from the univariate case is that the observed data, and the corresponding observational noise, is now a vector. West and Harrison (1997) provide the updating, forecasting and filtering equation for such a DLM in the case that the components of the defining quadruple are known. However, West and Harrison (1997) state that in general, 'there is no neat conjugate analysis available to enable the sequential learning' of the unknown covariance matrix $\mathbf{V}_t$ in the above model. This is in contrast with the univariate case, wherein a sequential analysis can be applied by

enforcing a gamma prior on the precision.

However, some methods do exist for a closed-form multivariate analysis in specific circumstances. Quintana and West (1987) develop the class of 'Matrix Normal DLMs', which make the assumption that all individual component time series $y_{i,t}$, $i \in \{1, \ldots, N\}$, follow univariate DLMs with common $\mathbf{F}_t$ and $\mathbf{G}_t$. This allows a fully conjugate analysis to be carried out, even in the case of an unknown multivariate covariance structure. The cross-sectional structure across time series at time t is introduced by an $N \times N$ covariance matrix:

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_{1,2} & \ldots & \sigma_{1,N} \\ \sigma_{1,2} & \sigma_2^2 & \ldots & \sigma_{2,N} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1,N} & \sigma_{2,N} & \ldots & \sigma_N^2 \end{pmatrix},$$

where $\sigma_{i,j}$ determines the covariance between the series $y_{i,t}$ and $y_{j,t}$, for all i and j, $(i = 1, \ldots, N; j = 1, \ldots, N; i \neq j)$. By introducing a state matrix $\boldsymbol{\Theta}_t$, whose columns are the state vectors of the individual DLMs, Quintana and West (1987) show that the individual model equations can be written jointly in matrix notation; see also West and Harrison (1997). In this multivariate framework, the observational error is a $N \times 1$ vector, defined to follow a multivariate normal distribution with covariance matrix $V_t\boldsymbol{\Sigma}$, where V_t is an observational scale factor common across all N series. The evolution error is now given by a matrix $\boldsymbol{\Omega}$, which follows a matrix normal distribution (see Dawid (1981) and West and Harrison (1997)). Under this framework, West and Harrison (1997) describe how the state matrix $\boldsymbol{\Theta}_t$ and the covariance matrix $\boldsymbol{\Sigma}$ can be learned about jointly in a closed-form manner, using the class of matrix normal/ inverse Wishart distributions.

Of course, the assumption that all univariate series share common $\mathbf{F}_t$ and $\mathbf{G}_t$ may be restrictive in certain applications. However, in our case where we wish to model each

individual forecaster as a DLM, it is perfectly possible that these should be modelled in the same way. Consider, for example, our simulations in Chapter 4 where each forecaster is modelled as an AR plus noise process. Since all forecasters are modelling the same variable of interest y_t, it seems likely that a common choice of $\mathbf{F}_t$ and $\mathbf{G}_t$ would be suitable. Therefore, one possible extension to the work presented in Chapter 4 of this thesis could be the investigation of jointly modelling individual forecasters in order to deal with missing data. It is possible that modelling the forecasters in a multivariate framework would lead to better imputed values in missing data periods, reducing the need for sudden changes in the combination weights.

Bibliography

Aastveit, K. A., Mitchell, J., Ravazzolo, F., and van Dijk, H. (2019). The evolution of forecast density combinations in economics. *Oxford Research Encyclopedia of Economics and Finance*.

Ahn, K.-H., Yellen, B., and Steinschneider, S. (2017). Dynamic linear models to explore time-varying suspended sediment-discharge rating curves. *Water Resources Research*, 53(6):4802–4820.

Bates, J. M. and Granger, C. W. J. (1969). The combination of forecasts. *Operational Research Quarterly*, 20(4):451–468.

Bernaciak, D. and Griffin, J. E. (2024). A loss discounting framework for model averaging and selection in time series models. *International Journal of Forecasting*, 40(4):1721–1733.

Blanc, S. M. and Setzer, T. (2020). Bias–variance trade-off and shrinkage of weights in forecast combination. *Management Science*, 66(12):5720–5737.

Bunn, D. W. (1975). A Bayesian approach to the linear combination of forecasts. *Journal of the Operational Research Society*, 26(2):325–329.

Capistrán, C. and Timmermann, A. (2009). Forecast combination with entry and exit of experts. *Journal of Business & Economic Statistics*, 27(4):428–440.

Carvalho, C. M., Johannes, M. S., Lopes, H. F., and Polson, N. G. (2010). Particle learning and smoothing. *Statistical Science*, 25(1):88–106.

Chan, F. and Pauwels, L. L. (2018). Some theoretical results on forecast combinations. *International Journal of Forecasting*, 34(1):64–74.

Chave, A. D., Luther, D. S., and Thomson, D. J. (2019). High-q spectral peaks and nonstationarity in the deep ocean infragravity wave band: Tidal harmonics and solar normal modes. *Journal of Geophysical Research: Oceans*, 124(3):2072–2087.

Claeskens, G., Magnus, J. R., Vasnev, A. L., and Wang, W. (2016). The forecast combination puzzle: A simple theoretical explanation. *International Journal of Forecasting*, 32(3):754–762.

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

Conflitti, C., De Mol, C., and Giannone, D. (2015). Optimal combination of survey forecasts. *International Journal of Forecasting*, 31(4):1096–1103.

Dawid, A. P. (1981). Some matrix-variate distribution theory: Notational considerations and a Bayesian application. *Biometrika*, 68(1):265–274.

De Menezes, L. M., W. Bunn, D., and Taylor, J. W. (2000). Review of guidelines for the use of combined forecasts. *European Journal of Operational Research*, 120(1):190–204.

Diebold, F. X. (1991). A note on Bayesian forecast combination procedures. In *Economic Structural Change*, pages 225–232, Berlin, Heidelberg. Springer Berlin Heidelberg.

Diebold, F. X. and Pauly, P. (1987). Structural change and the combination of forecasts. *Journal of Forecasting*, 6(1):21–40.

Elipot, S., Lumpkin, R., Perez, R. C., Lilly, J. M., Early, J. J., and Sykulski, A. M. (2016). A global surface drifter data set at hourly resolution. *Journal of Geophysical Research: Oceans*, 121(5):2937–2966.

Fisher, J. D., Pettenuzzo, D., and Carvalho, C. M. (2020). Optimal asset allocation with multivariate Bayesian dynamic linear models. *The Annals of Applied Statistics*, 14(1):299 –338.

Gastinger, J., Nicolas, S., Stepić, D., Schmidt, M., and Schülke, A. (2021). A study on ensemble learning for time series forecasting and the need for meta-learning. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

Genre, V., Kenny, G., Meyler, A., and Timmermann, A. (2013). Combining expert forecasts: Can anything beat the simple average? *International Journal of Forecasting*, 29(1):108–121.

Gneiting, T., Balabdaoui, F., and Raftery, A. E. (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2):243–268.

Gonella, J. (1972). A rotary-component method for analysing meteorological and oceanographic vector time series. *Deep-Sea Research*, 19:833–846.

Gordon, K. and Smith, A. F. M. (1990). Modeling and monitoring biomedical times series. *Journal of the American Statistical Association*, 85(410):328–337.

Graefe, A., Armstrong, J. S., Jones, R. J., and Cuzán, A. G. (2014). Combining forecasts: An application to elections. *International Journal of Forecasting*, 30(1):43–54.

Granger, C. and Newbold, P. (1977). *Forecasting Economic Time Series*. Academic Press Inc, 1 edition.

Granger, C. W. and Ramanathan, R. (1984). Improved methods of combining forecasts. *Journal of Forecasting*, 3(2):197.

Hall, S. G. and Mitchell, J. (2007). Combining density forecasts. *International Journal of Forecasting*, 23(1):1–13.

Hamilton, J. D. (1994). *Time Series Analysis*. Princeton University Press, Princeton, NJ.

Harvey, A. C. (1990). *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge University Press.

Irie, K., Glynn, C., and Aktekin, T. (2022). Sequential modeling, monitoring, and forecasting of streaming web traffic data. *The Annals of Applied Statistics*, 16(1):300 –325.

Jackson, L. C., Dubois, C., Forget, G., Haines, K., Harrison, M., Iovino, D., Köhl, A., Mignac, D., Masina, S., Peterson, K. A., Piecuch, C. G., Roberts, C. D., Robson, J., Storto, A., Toyoda, T., Valdivieso, M., Wilson, C., Wang, Y., and Zuo, H. (2019). The mean state and variability of the North Atlantic circulation: A perspective from ocean reanalyses. *Journal of Geophysical Research: Oceans*, 124(12):9141–9170.

Jiang, C.-Y. and Zhang, Y.-A. (2013). Some results on linear equality constrained state filtering. *International Journal of Control*, 86(12):2115–2130.

Kang, H. (1986). Unstable weights in the combination of forecasts. *Management Science*, 32(6):683–695.

Khan, F., Ali, S., Saeed, A., Kumar, R., and Khan, A. W. (2021). Forecasting daily new infections, deaths and recovery cases due to COVID-19 in Pakistan by using Bayesian dynamic linear models. *PLOS ONE*, 16(6):1–14.

Koop, G. (2003). *Bayesian Economics*. Wiley, 1 edition.

Koop, G. and Korobilis, D. (2012). Forecasting inflation using dynamic model averaging. *International Economic Review*, 53(3):867–886.

Lahiri, K., Peng, H., and Zhao, Y. (2017). Online learning and forecast combination in unbalanced panels. *Econometric Reviews*, 36(1-3):257–288.

LeSage, J. P. and Magura, M. (1992). A mixture-model approach to combining forecasts. *Journal of Business & Economic Statistics*, 10(4):445–452.

Lichtendahl, K. C. and Winkler, R. L. (2020). Why do some combinations perform better than others? *International Journal of Forecasting*, 36(1):142–149. M4 Competition.

Ma, S. and Fildes, R. (2021). Retail sales forecasting with meta-learning. *European Journal of Operational Research*, 288(1):111–128.

MacLachlan, C., Arribas, A., Peterson, K. A., Maidens, A., Fereday, D., Scaife, A. A., Gordon, M., Vellinga, M., Williams, A., Comer, R. E., Camp, J., Xavier, P., and Madec, G. (2015). Global seasonal forecast system version 5 (glosea5): a high-resolution seasonal forecast system. *Quarterly Journal of the Royal Meteorological Society*, 141(689):1072–1084.

Magnus, J. R. and Vasnev, A. L. (2023). On the uncertainty of a combined forecast: The critical role of correlation. *International Journal of Forecasting*, 39(4):1895–1908.

Makridakis, S., Spiliotis, E., and Assimakopoulos, V. (2020a). The M4 competition: 100,000 time series and 61 forecasting methods. *International Journal of Forecasting*, 36(1):54–74. M4 Competition.

Makridakis, S., Spiliotis, E., and Assimakopoulos, V. (2020b). The M4 competition: 100,000 time series and 61 forecasting methods. *International Journal of Forecasting*, 36(1):54–74. M4 Competition.

Makridakis, S. and Winkler, R. L. (1983). Averages of forecasts: Some empirical results. *Management Science*, 29(9):987–996.

Matsypura, D., Thompson, R., and Vasnev, A. L. (2018). Optimal selection of expert forecasts with integer programming. *Omega*, 78:165–175.

McCarthy, G., Smeed, D., Johns, W., Frajka-Williams, E., Moat, B., Rayner, D., Baringer, M., Meinen, C., Collins, J., and Bryden, H. (2015). Measuring the Atlantic meridional overturning circulation at 26°N. *Progress in Oceanography*, 130:91–111.

Newbold, P. and Granger, C. W. J. (1974). Experience with forecasting univariate time series and the combination of forecasts. *Journal of the Royal Statistical Society. Series A (General)*, 137(2):131–165.

Olhede, S. C. and Ombao, H. (2013). Modeling and estimation of covariance of replicated modulated cyclical time series. *IEEE Transactions on Signal Processing*, 61(8):1944–1957.

Palm, F. C. and Zellner, A. (1992). To combine or not to combine? Issues of combining forecasts. *Journal of Forecasting*, 11(8):687–701.

Pawlikowski, M. and Chorowska, A. (2020). Weighted ensemble of statistical models. *International Journal of Forecasting*, 36(1):93–97. M4 Competition.

Petris, G., Petrone, S., and Campagnoli, P. (2009). *Dynamic Linear Models with R*. Springer New York, NY.

Petropoulos, F., Kourentzes, N., Nikolopoulos, K., and Siemsen, E. (2018). Judgmental selection of forecasting models. *Journal of Operations Management*, 60(1):34–46.

Prado, R. and West, M. (2010). *Time Series: Modeling, Computation, and Inference*. Chapman & Hall/CRC, 1st edition.

Quintana, J. and West, M. (1987). Multivariate time series analysis: New techniques applied to international exchange rate data. *The Statistician*, 36:275–281.

Radchenko, P., Vasnev, A. L., and Wang, W. (2023). Too similar to combine? On negative weights in forecast combination. *International Journal of Forecasting*, 39(1):18–38.

Raftery, A. E., Kárný, M., and Ettler, P. (2010). Online prediction under model uncertainty via dynamic model averaging: Application to a cold rolling mill. *Technometrics*, 52(1):52–66.

Ray, E. L., Brooks, L. C., Bien, J., Biggerstaff, M., Bosse, N. I., Bracher, J., Cramer, E. Y., Funk, S., Gerding, A., Johansson, M. A., Rumack, A., Wang, Y., Zorn, M., Tibshirani, R. J., and Reich, N. G. (2022). Comparing trained and untrained probabilistic ensemble forecasts of COVID-19 cases and deaths in the united states. Preprint.

Riegert, D. L. and Thomson, D. J. (2018). Non-stationarity and offset coherence information in geomagnetic applications. In *2018 IEEE Statistical Signal Processing Workshop (SSP)*, pages 179–182.

Rowe, D. B. (2005). Modeling both the magnitude and phase of complex-valued fmri data. *NeuroImage*, 25(4):1310–1324.

Sanchez-Franks, A., Frajka-Williams, E., Moat, B. I., and Smeed, D. A. (2021). A dynamically based method for estimating the Atlantic meridional overturning circulation at 26°N from satellite altimetry. *Ocean Science*, 17(5):1321–1340.

Schreier, P. J. and Scharf, L. L. (2010). *Statistical Signal Processing of Complex-Valued Data: The Theory of Improper and Noncircular Signals*. Cambridge University Press.

Sessions, D. N. and Chatterjee, S. (1989). The combining of forecasts using recursive techniques with non-stationary weights. *Journal of Forecasting*, 8(3):239–251.

Simon, D. (2010). Kalman filtering with state constraints: A survey of linear and nonlinear algorithms. *Control Theory & Applications, IET*, 4:1303 – 1318.

Smith, J. and 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(2):272–276.

Stock, J. H. and Watson, M. W. (2004). Combination forecasts of output growth in a seven-country data set. *Journal of Forecasting*, 23(6):405–430.

Stock, J. H. and Watson, M. W. (2006). Chapter 10 forecasting with many predictors. volume 1 of *Handbook of Economic Forecasting*, pages 515–554. Elsevier.

Stone, M. (1961). The opinion pool. *The Annals of Mathematical Statistics*, 32(4):1339–1342.

Sykulski, A., Olhede, S., and Sykulska-Lawrence, H. (2022). The elliptical Ornstein–Uhlenbeck process. *Statistics and Its Interface*, 16(1):133–146.

Sykulski, A. and Percival, D. (2016). Exact simulation of noncircular or improper complex-valued stationary Gaussian processes using circulant embedding. In *2016 IEEE International Workshop on Machine Learning for Signal Processing, MLSP 2016 - Proceedings*. IEEE Computer Society. 26th IEEE International Workshop on Machine Learning for Signal Processing, MLSP 2016 - Proceedings ; Conference date: 13-09-2016 Through 16-09-2016.

Sykulski, A. M., Olhede, S. C., and Lilly, J. M. (2016). A widely linear complex autoregressive process of order one. *IEEE Transactions on Signal Processing*, 64(23):6200–6210.

Sykulski, A. M., Olhede, S. C., Lilly, J. M., and Danioux, E. (2015). Lagrangian time series models for ocean surface drifter trajectories. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 65(1):29–50.

Sykulski, A. M., Olhede, S. C., Lilly, J. M., and Early, J. J. (2017). Frequency-domain stochastic modeling of stationary bivariate or complex-valued signals. *IEEE Transactions on Signal Processing*, 65(12):3136–3151.

Terui, N. and van Dijk, H. K. (2002). Combined forecasts from linear and nonlinear time series models. *International Journal of Forecasting*, 18(3):421–438.

Thomson, M. E., Pollock, A. C., Önkal, D., and Gönül, M. S. (2019). Combining forecasts: Performance and coherence. *International Journal of Forecasting*, 35(2):474–484.

Timmermann, A. (2006). Chapter 4 forecast combinations. volume 1 of *Handbook of Economic Forecasting*, pages 135–196. Elsevier.

Wallis, K. F. (2005). Combining density and interval forecasts: A modest proposal. *Oxford Bulletin of Economics and Statistics*, 67(s1):983–994.

Wang, X., Hyndman, R. J., Li, F., and Kang, Y. (2023). Forecast combinations: An over 50-year review. *International Journal of Forecasting*, 39(4):1518–1547.

West, M. and Harrison, J. (1997). *Bayesian Forecasting and Dynamic Models (2nd Ed.)*. Springer-Verlag, Berlin, Heidelberg.

Whittle, P. (1953). Estimation and information in stationary time series. *Arkiv för Matematik*, 2(5):423 – 434.

Xie, J. and Hong, T. (2016). Gefcom2014 probabilistic electric load forecasting: An integrated solution with forecast combination and residual simulation. *International Journal of Forecasting*, 32(3):1012–1016.

Yusupova, A., Pavlidis, N. G., and Pavlidis, E. G. (2023). Dynamic linear models with adaptive discounting. *International Journal of Forecasting*, 39(4):1925–1944.

Zhao, Z. Y., Xie, M., and West, M. (2016). Dynamic dependence networks: Financial time series forecasting and portfolio decisions. *Applied Stochastic Models in Business and Industry*, 32(3):311–332.

Zheng, Y., Wu, W., Wang, M., Zhang, Y., and Du, Y. (2024). Different trajectory patterns of ocean surface drifters modulated by near-inertial oscillations. *Environmental Research Letters*.