Anlaysing Irithimcs' Investor Expectation Forecasts using the Hilbert-Huang Transform

An initial investigation into insights to be gained from the decomposition of unusual non-stationary financial time series

Ffinlo Wright

A dissertation submitted in partial fulfilment
of the requirements for the degree of Master of Science
in

Applied Statistics and Datamining

at the

University of St Andrews

University of
St Andrews

Supervisor: **Dr Michail Papathomas**
External Supervisor: **Dr Grant Fuller**
Date of Submission: **20/08/2023**

Abstract

We introduce the Hilbert-Huang Transform (HHT) for analysing the unusual financial time series of investor expectations provided by Irtihmics. As financial markets are reflexive with respect to investors' views and expectations, anticipating the expectations of others is key to successful investing — Irithmics presents a solution to this problem by leveraging machine learning to generate forecasts of investor expectations for publicly traded companies. We propose the use of the HHT as it is a fully adaptive, data-driven method designed for nonlinear, non-stationary time series analysis and so, does not require any distributional assumptions to be made about the unusual expectation data. The HHT can decompose any time series into a finite number of Intrinsic Mode Functions (IMFs) that capture the fluctuations contained in the original series at different time scales. This project is an initial investigation into the possible insights that can be gained from this method. Therefore, we present a broad analysis that explores the frequency profiles obtained from HHT, the statistical significance of the IMFs, and the local correlations between investor expectations and stock price. This approach allows us to visualise the dynamics between expectations and price across multiple time scales. The HHT is shown to be a powerful approach, but not without its limitations. We discuss the shortcomings of the methodology encountered during our investigation and suggest multiple future lines of inquiry to fully realise the potential benefits of the HHT for analysing unusual financial data.

Acknowledgements

We would like to thank Dr Grant Fuller, and Irithmics, for offering and supervising this project. We are grateful for the opportunity to investigate such an interesting data set and for the generous amount of time taken by Dr Fuller to provide us with much-appreciated guidance and advice at all stages of this project.

We would also like to thank Dr Michail Papathomas for supervising this project and supporting our investigation by advising on the new mathematical concepts that we encountered and keeping us on track to produce a polished report.

Data Access Statement

The data provided by Irithmics for this project is made available in the supplementary material submitted alongside this report.

The stock price data was obtained using the Refinitiv Workspace (1) with an access licence provided by the University of St Andrews Library. We have included this price data in the supplementary material alongside this report (see Appendix A).

Word Count: 14,911

Contents

Chapter 1

Introduction

1.1 Anticipating the Anticipations of Others

Financial markets are complex institutions that facilitate the trade of numerous securities by millions of individual investors. The specific portfolio of securities held by an investor dictates their exposure to market risk and profit-making opportunities. Therefore, to maximise profits and minimise potential losses, everyday investors choose to allocate and reallocate their portfolio capital and risk through making trades in the financial markets. These decisions are made based on each investor's perceived utility to be gained from holding or trading an asset. As this utility depends on the difference between the present and the future, what an investor expects to happen in the future dictates how they make capital allocation decisions in the present. The current price of a security is determined by what investors- each with their own expectations about the future- are willing to buy or sell it for. For this reason, the famous early twentieth-century economist John Maynard Keynes postulated that successful investors devote their "double intelligences to anticipating what average opinion expects the average opinion to be." (2, p. 79)

The efficient market hypothesis asserts that "doublethe price of an asset reflects the consensus evaluation of the market based upon the information available to the market regardless of private information held by [an] investor." (3, p. 359) However, it is unreasonable to assume that investors assimilate all information into their decisions and choices — especially if they must consider the anticipations of others. In recognition of this reality, Herbert A. Simon (4) proposed the concept of bounded rationality, suggesting that when making decisions, investors only consider the information they find salient at the time. Therefore, to anticipate the anticipations of others, an investor would need to know not just all the underlying information available, but also what information other investors find relevant when making decisions.

Irithmics is a financial technology company that attempts to overcome this Sisyphean task by leveraging advanced machine learning techniques to forecast market expectations. This is made possible by three main concepts. Firstly, through inverse optimisation, it is possible to estimate an investor's views about an asset's future utility from the amount of portfolio capital assigned to it — its portfolio weighting. (5) As large, so-called institutional investors make both regulatory and voluntary disclosures about their portfolios, Irithmics can use inverse optimisation to estimate their views and expectations about assets within their portfolios. Secondly, similar investors (those with similar portfolios, strategies, mandates, regulations etc.) are likely to use similar information to contextualise their investment decisions and make similar decisions when encountering the same market environments and constraints. This phenomenon is known as homophily within social network analysis. Due to the large amounts of capital under management, homophily between institutional investors affects market supply and demand (and by extension, prices). Thirdly, financial markets are reflexive (6) — as interacting with financial markets based on views held about the market changes the market, and changes in the market alter views about the market. This property of reflexivity amplifies the shifts in supply and demand caused by the homophily between institutional investors, as their investment decisions, based on their future expectations, cause changes in market conditions for other, smaller investors. Therefore, by combining the three concepts of inverse optimisation, investor homophily, and market reflexivity, Irithmics can estimate market expectations about the future based on capital allocation choices in the present.

For this project, Irithmics provided us with expectation forecasts produced by their ML model for each constituent company of the FTSE 100 index as of June 2023. The data contained expectations for 2020, 2021, and 2022. For each year and each company, there were multiple data files,

each centred on an announcement date. The FTSE 100 companies are legally mandated to make quarterly announcements about their performance, and the dates of these reports are known by the market well in advance. Anticipating the content of, and market reaction to, these announcements is a major driver of capital allocation decisions by institutional investors. Therefore, these dates are of vital importance when modelling market expectations. We provide a detailed overview of the data used in this investigation in Section 2.1.

1.2 Why use the Hilbert-Huang Transform?

At the start of this project, we were cautioned by Irithmics to be highly cautious of using traditional parametric techniques to analyse the expectation data. Because, as it was the product of a reflexive process, we would be unwise to make any distributional assumptions about the data. Furthermore, it is already well-understood that financial time series exhibit nonlinear and non-stationary dynamics, placing further restrictions on the statistical methods that would be suitable for analysing this unusual data. There is also substantial evidence that financial time series are driven by factors acting on different time scales, from rapid short-term fluctuations due to information shocks to longer-term economic trends. (7) The same is obviously true for the data provided by Irithmics, as it is an estimate of the expectations of the investors that drive the dynamics observed in traditional financial time series — investors make both longer-term strategic and shorter-term tactical allocations of capital in response to changing information and market conditions. (8) Therefore, we required an adaptive, non-parametric, nonlinear, and non-stationary approach for analysis.

One alternative approach in adaptive time-series analysis is offered by the Hilbert-Huang Transform (HHT). The HHT is a fully adaptive method that can decompose any time series into a set of oscillating component series, known as Intrinsic Mode Functions (IMFs), using an Empirical Mode Decomposition (EMD) algorithm. Using the Hilbert transform, the instantaneous frequencies and amplitudes of each IMF may be extracted. By construction, each IMF corresponds to fluctuations in the original series at different timescales. Therefore, the HHT method allows us to analyse the dynamics of, and between, non-stationary time series at different time scales. The HHT is similar to the universally accepted approach for signal analysis: the Fourier Transform (FT), which represents a stationary signal as a superposition of sine waves with different frequencies and amplitudes. (9) Unlike the FT, the HHT does not require stationarity or an *a priori* assumption of the basis function, as the EMD algorithm is completely data-driven and adaptive.

The HHT was first proposed as a method for non-stationary and non-linear time series analysis by Huang et al. in 1996, (10) and since has been applied to a wide variety of time series problems, such as seismology, (11) sunspot dynamics, (12, 13) and even the relationship between the weather and cardiovascular mortality. (14) Non-stationary adaptations of the FT, most notably the short-time Fourier Transform and Wavelet analysis(15), have been developed and applied to financial time series for decades. The first application of HHT to financial time series was in 2003 when Huang et.al. (16) applied the method to mortgage rate data.Since then, the HHT has been used for improving ML forecasts for stock-market indices, (7) and exploring the correlation structures between financial time series. (17)

1.3 Aims and Structure

The primary aim of this project is to explore the application of the HHT to the analysis of Irithmics' expectation data. We also apply the HHT to the stock price and price change time series that correspond with the generation dates of the Irithmics expectation forecasts. We do this for two main reasons. Firstly, applying the HHT to more standard financial time series provides a benchmark for the methodology that complements the analysis of the more unusual expectation data. Secondly, decomposing the price series allows us to explore the correlations between expectations and stock price at different time scales using a method adapted originally from Chen et al. (18). The stock price is not directly used in the Irithmics ML model — however, the idea that markets are reflexive is a core assumption of their approach to forecasting expectations. We show, by using HHT, that there is evidence of strong local correlations between expectations and price across time scales with complex lead-lag dynamics specific to each forecast. However, we are only able to present interim results for the correlation analysis due to the time constraints of this project.

This report is structured as follows. In Chapter 2 we provide the methods used in our investigation. Due to the unusual nature of the expectation forecasts, we begin by offering an explanation of the structure and our interpretation of the data provided by Irithmics. We also explain how

we processed the data to remove forecast overlap by merging the data sets. Then in Section 2.2, we provide a mathematical overview of both the HHT, and our adapted method for calculating correlations between non-stationary series. These theoretical sections are followed by Section 2.3 where we apply HHT to an arbitrarily chosen set of forecasts and the corresponding stock price and price change time series. Here, we explain how we performed our analysis by first decomposing the series and then, showing how the HHT can adaptively filter a time series, examining the instantaneous frequencies, calculating cross-correlations and performing an information significance test on the IMFs. In Chapter 3, we present the results obtained from applying the HHT and the analysis methods applied to the example set in Section 2.3, to thirty-two of the longest merged data sets. This is followed by Chapter4 where we provide a discussion of our results, the challenges encountered during this project, and how HHT may be used in future investigations to improve ML methods and to obtain more conclusive results than those found in this initial investigation into the method. Conclusions are then provided in Section 4.4.

All of the code used to carry out our analysis and generate the plots used in this report are supplied as a supplementary material folder. Further details can be found in Appendix A.

Chapter 2

Methods

2.1 Irithmics Expectation Forecasts

The Irithmics' forecasts of investor expectations form unusual time series data. Each data set contains multiple expectation forecasts, and each forecast contains estimates of expectations for some future day t, but also each forecast is generated on a different day, g. Therefore, we consider these to be two-dimensional time series. In this section, we lay out the theoretical challenges associated with this unusual data, and how we overcame them.

2.1.1 Visualising Expectation Time Series

Forecast-time series

In Figure 2.1 we have plotted the first, thirtieth and sixtieth expectation forecasts contained in the Irithmics data file for Unilever that was centred around an announcement on 22 July 2021 (shown by a dashed vertical line). The file contained sixty-four forecasts generated on each financial trading day between 22 April and 22 July 2021. Each of these time series is a snapshot in time of what the Irithmics Machine Learning (ML) algorithm expects investor expectations to be for a given future date. From this plot, we can see that each forecast-time series does not cover the whole forecast date range and that there is significant variability in the expectations of investor behaviour for any given date. There are some broadly similar structures, with peaks and troughs in expectations occurring at similar points in the forecast range — but their amplitudes vary wildly, showing how investor expectations changed for that date at some point during the generation period.

Generation-time series

We can observe this variation in expectations for a given forecast date by plotting the Irithmics score for that forecast date against the date it was generated to form another time series (Figure 2.2). We call these generation-time series. In this example generation-time series, we can see how investor expectations changed from being slightly positive in late April to extremely positive in June, then decreasing in the lead-up to the announcement on 22 July. This series also illustrates the fact that mean expectations change with time — therefore the expectations are intuitively non-stationary. Similar series can be plotted for any date in the forecast range.

Colour map representation of expectation forecasts

Both one-dimensional time series approaches have their own advantages for exploring the expectation data provided by Irithmics. On one hand, the forecast-time series show how their ML algorithm expects investor expectations to change given the information available to it on that generation date, while on the other hand, the generation-time series shows how expectations about a given date change between the algorithm's forecasts based upon newly available information between generation days. However, it is difficult to ascertain why expectations change over time from both perspectives. Changes in the expectation score generated for a given date may result from investors bringing forward or delaying their intended course of action — not necessarily a fundamental change in their expectations. Therefore looking at either one-dimensional time series in isolation fails to explain the dynamics of how expectations are actually changing over time. So instead, we can plot the expectation data as a colour map with the Irithmics expectation score as

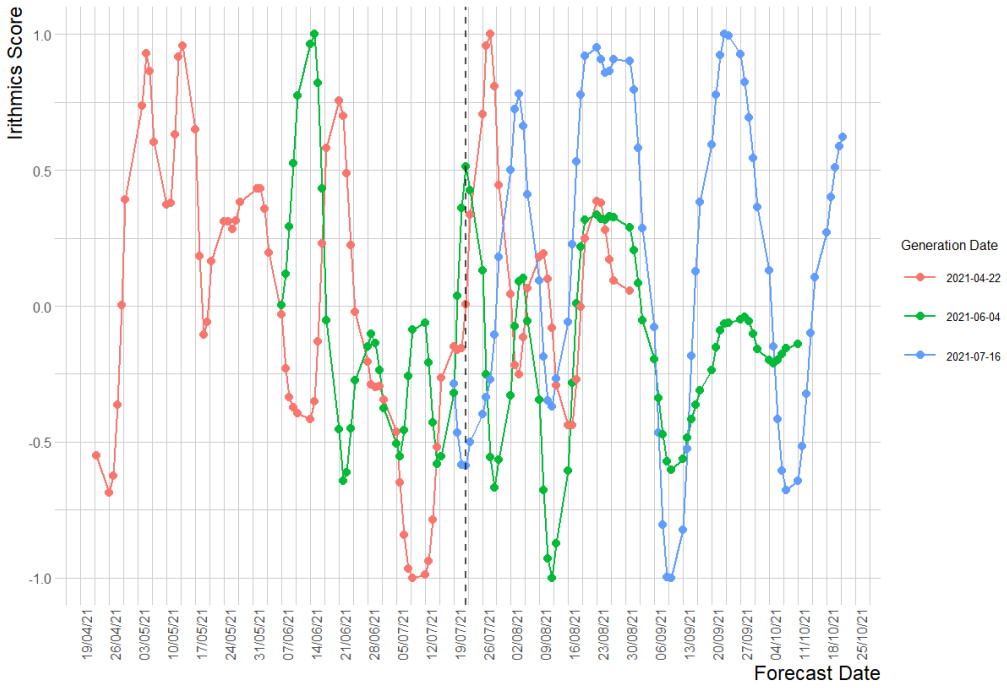


Figure 2.1: Forecast-time series of expectations for ULVR generated thirty trading days apart. The dashed black line shows an announcement on 22 July 2021.

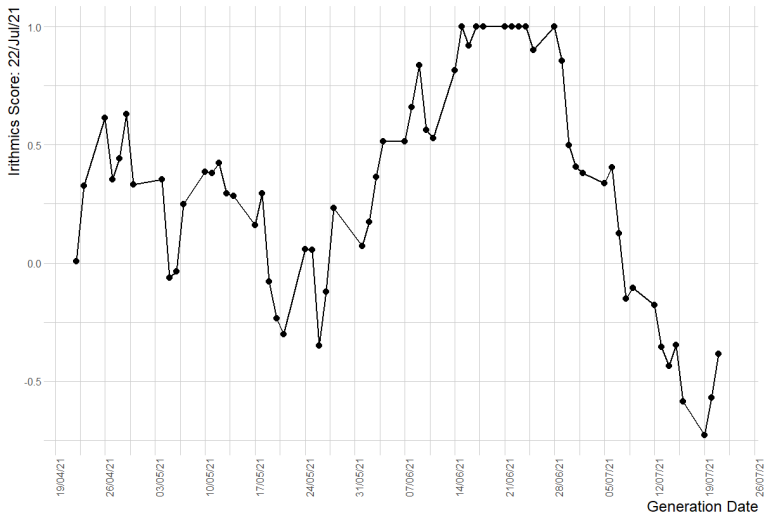


Figure 2.2: ULVR example generation-time series for the announcement date 22/07/2021.

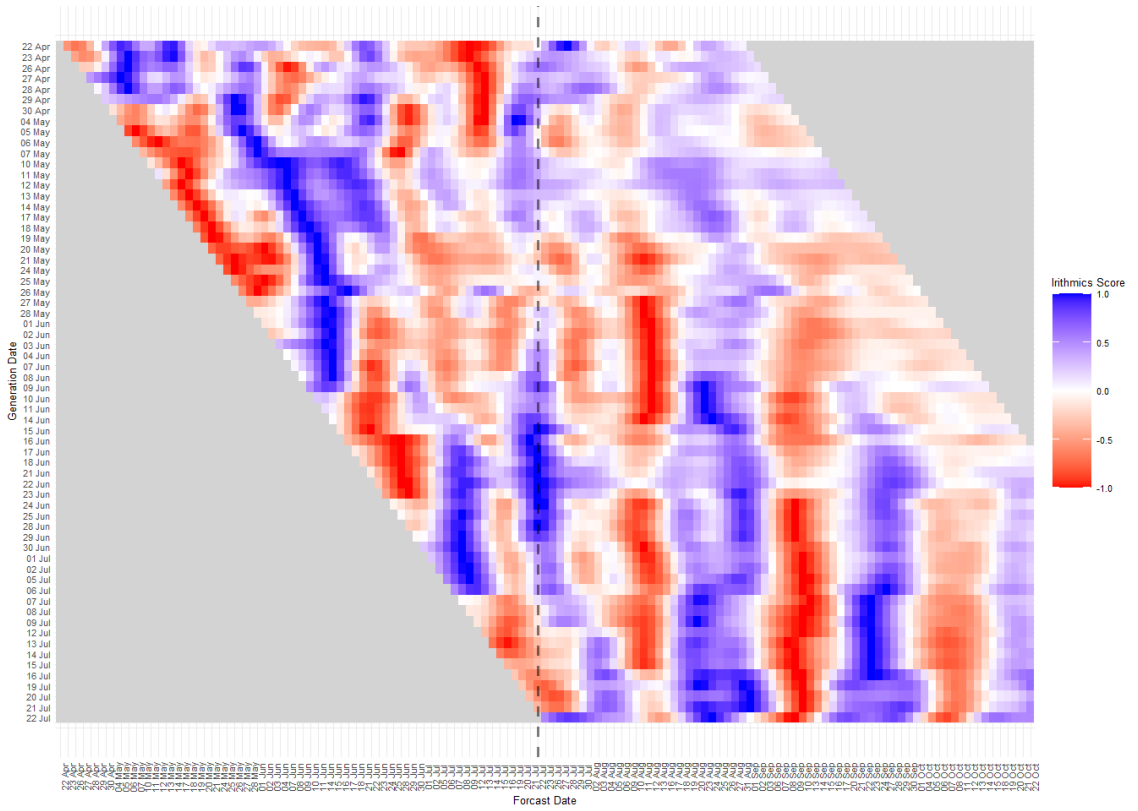


Figure 2.3: A colour map of expectation scores centred on the announcement date 22/07/21 (Shown as a dashed line). Blue signifies positive expectations (optimism), and red signifies negative expectations (pessimism).

the colour gradient, and each forecast "stacked" on each other so that the generation date is on the y-axis and the forecast date is on the x-axis (Figure 2.3).

This colour map gives a much richer visual understanding of how expectations change within and between different forecasts generated by Irithmics. One can get a sense of how persistent positive (or negative) expectations are for a given date, and when they change, one can see if investors have simply changed their minds, or if they have delayed (or brought forward) their intended course of action. This information is difficult to interpret when examining the forecast-time or generation-time series in isolation, highlighting why simple models, such as Auto-Regressive Moving Average (ARMA) models, would be unsatisfactory in capturing the true dynamics of these forecasts. Changes in expectations on a given date depend not only on the previous investor expectations but also on the previously held expectations about days both before and after the date of interest.

The Forecast Distance Problem

The colour map visualisation of expectation forecasts also draws attention to another difficulty of performing comparative statistical analysis implicit in the one-dimensional time series examples — each forecast contains different dates. In Figure 2.3, we can see that each subsequent forecast 'slides' forward by a day. This is because we cannot forecast the past, and on the other, the algorithm can produce a prediction for one more future trading day. However, due to the centring of the data set on the announcement date, forecasts toward the end of the generation period are shorter than those made at the start. Therefore, a direct comparison between forecast-time series over the whole generation period presents both theoretical and practical challenges. Similarly, vertical generation-time series also vary in length across the data set due to new forecasts 'sliding' forward by a day.

The greatest conceptual problem when considering the comparisons between forecasts and other financial time series is that the temporal distance between the expectation score for a given day in the forecast range and its generation date inherently changes between forecasts. To illustrate this, consider the forecast $F_g(t)$, generated on some day g. Now consider another forecast, $F_{g+10}(t)$, generated ten days later on generation day $g + 10$. A generation-time series, $G_\tau(g)$, for some

arbitrary day, τ, would contain the expectation scores, $f_{g,\tau}$ and $f_{g+10,\tau}$, taken from $F_g(t)$ and $F_{g+10}(t)$, respectively. Suppose that $\tau = g + 11$, which means that $f_{g,\tau}$ was generated eleven days before τ, whilst $f_{g+10,\tau}$ was only generated one day before τ. Therefore, we naturally associate more certainty in the nearer estimate, $f_{g+10,\tau}$, than the further estimate,$f_{g,\tau}$. Therefore, the two values are qualitatively different, which causes a conceptual issue when we draw comparisons between different generation-time series (e.g. between $G_\tau(g)$ and $G_{\tau+1}(g)$, or with the stock price time series, $S(g)$).

Put in another, less formal way, we intuitively know that our own expectations about what we will do tomorrow are less subject to change than our expectations about what we will do in eleven days' time. We may have an idea or planned course of action, but we are naturally much less certain about what it will be, and we have the time to change our minds based on learning new information. This intuitive fact is no different for institutional investors, as they can reallocate their capital portfolios as their expectations change more easily for days further into the future. This means that we must be careful when comparing the generation-time series of expectations, as earlier observations in the series are qualitatively different from later observations. A similar problem presents itself for forecast-time series as each successive value of the series is for a date further into the future from the generation day, and so has less certainty associated with it. Traditional time series analysis does not suffer from this forecast distance problem, as usually time series are constructed by taking like-for-like observations of a variable over a sampling period. Therefore, we propose a slight change in perspective to overcome the forecast distance problem.

'Reading along the diagonal': $g + h$ time series

Instead of reading across the x-axis or down the y-axis of Figure 2.3, consider 'reading along the diagonal' and constructing a time series, of expectations about tomorrow, $g+1$, for each generation day g. Essentially, this is the series of the first values of each forecast $F_g(t)$. Now consider the adjacent diagonal — the day after tomorrow — $g + 2$ for each day g. Like the generation-time series, these diagonal series depend on the generation day, g, but their expectation scores all have the same forecast distance when they were generated. This avoids the issue of different degrees of certainty being embedded in each observation of the series. In this project, we refer to these diagonal time series as the $g + h$ expectations. Figure 2.4 presents some example $g + h$ time series. Unlike the generation and forecast-time series, the announcement date is no longer fixed, appearing to move backwards through the series as h increases.

The advantages of diagonal $g + h$ series are that the forecast distance is the same for all observations within a series and that they have a consistent length for almost the whole data set. Furthermore, this allows us to make meaningful comparisons between the $g+h$ series and the stock price series $S(g)$ as both series depend on g. Similar comparisons between forecast-time series $F_g(t)$ and $S(g)$ would be meaningless as t is a hypothetical time increment, whereas g corresponds to real temporal changes between observations. Whilst generation-time series, $G_\tau(g)$ also depend on g they suffer from the forecast distance problem, invalidating any comparisons with $S(g)$.

Figure 2.5 is the colour map formed by plotting the $g + h$ series as vertical lines so that h, not t, is along the x-axis. This colour map is the result of shifting the forecasts that makeup Figure 2.3 to the left so that they are aligned on the y-axis. This means that each forecast day, t, now forms a line $g = h + t$ that intercepts the generation date axis at $g = t$. This means that the once vertical generation time forecasts are now diagonal lines across the plot. We have included a dashed line for the 22 July 2021 announcement date as a reference point for this transformation.

2.1.2 Merging Overlapping Expectation Data

The centring of each forecast set on an announcement date is an inherent part of Irithmic's approach to expecting expectations. They use reverse optimization (5) to infer investor expectations about the future from changes to portfolio weightings in the present. Portfolio re-weighting is done in anticipation of the market response to corporate announcements — and so the expectation forecasts are intrinsically linked to these dates. However, for each company, these announcements are not evenly distributed across the financial year. During data exploration, we found many cases of significant overlap in the generation periods contained in different files for the same company — meaning that the previous quarter's announcement (and some preceding forecasts) appeared among the initial generation days for some of the forecast sets.

This presented two major problems for our analysis. Firstly, announcement days are disruptive information events that expectations are centred around, and so failing to account for their presence at the beginning of a data set would lead to false conclusions about the dynamics of investor

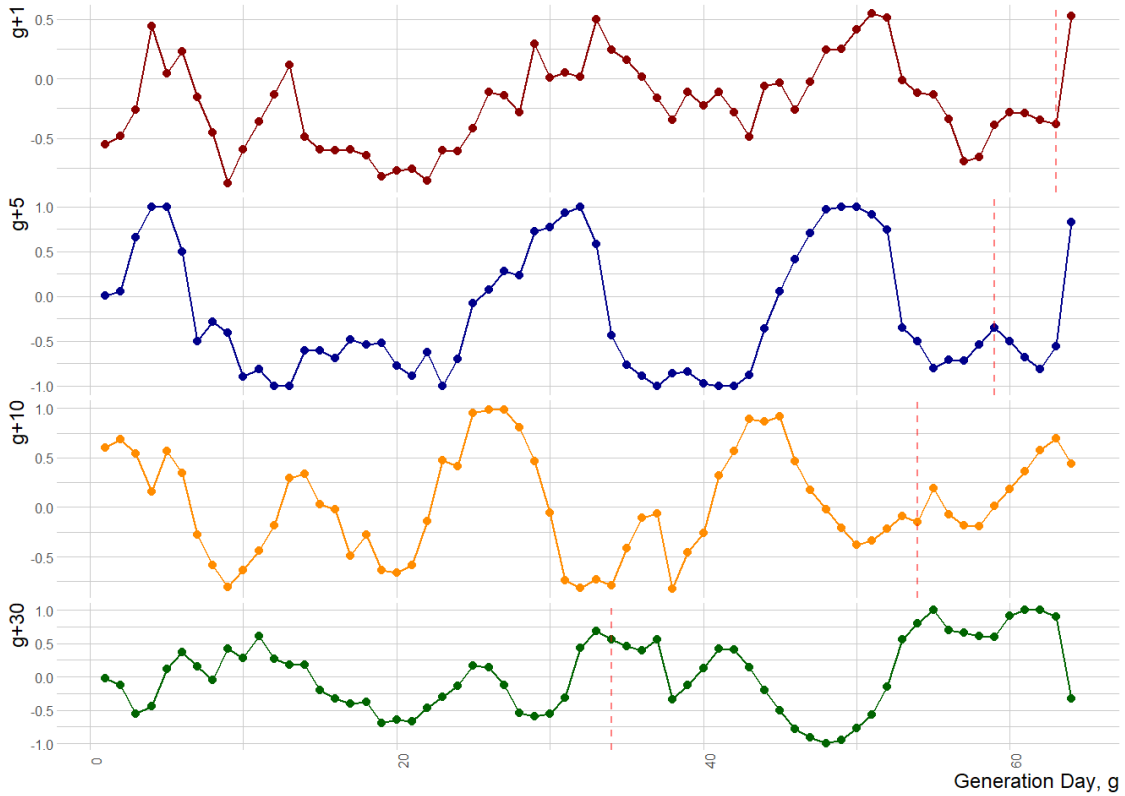


Figure 2.4: ULVR example $g + h$ time series: $g + 1, g + 5, g + 10, g + 30$. The red dashed line represents the announcement on 22 July 2021.

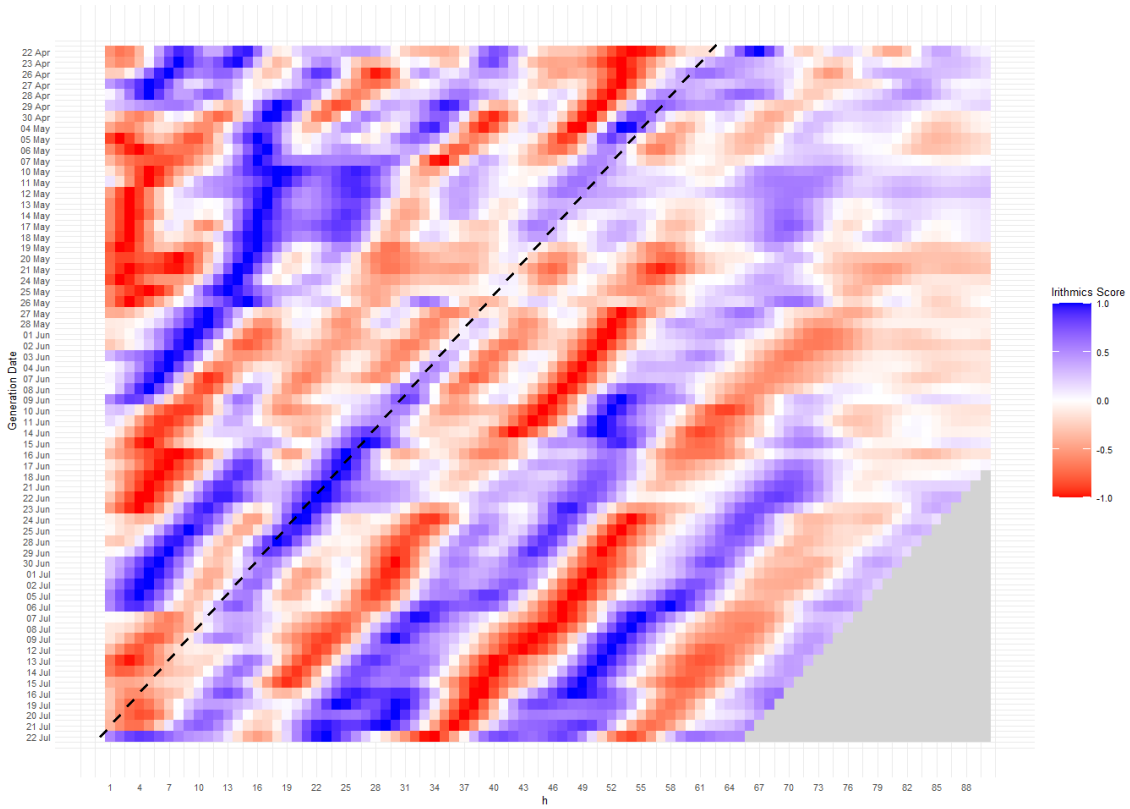


Figure 2.5: This $g + h$ expectations colour map is the result of shifting the forecasts in Figure 2.3 shifted so that they are aligned against the y-axis

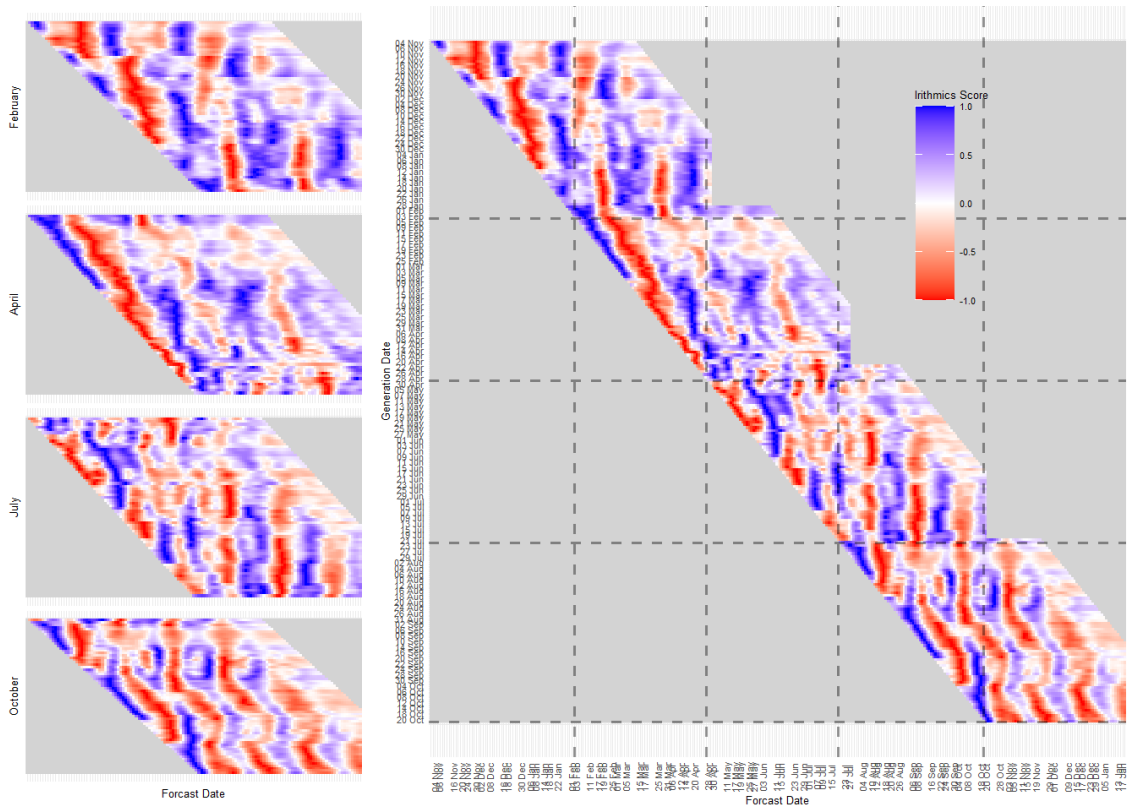


Figure 2.6: Merged colour maps of 2021 expectation forecasts for ULVR. The quarterly forecasts are shown on the left, and the result of the merging algorithm is plotted on the right. Dashed lines have been added to both axes to indicate when the announcement dates were.

	Total Number of Files	Mean Number of Files per Company	Mean Generation Period Length	Max Generation Period Length
Before Merging	937	10	63	67
After Merging	273	5	109	254

Table 2.1: Summary of the number of data files and forecasts before and after merging.

expectations. Secondly, multiple overlapping forecasts for the same company mean that each forecast is not unique and poses issues when drawing comparisons between the data sets. Therefore, we merged the overlapping forecast sets to remove the duplicated forecasts. A visual representation of this merging process is presented in Figure 2.6. The merging algorithm we developed for this data first identifies which generation dates are overlapped between two or more forecast sets for a given company and then stitches them together, removing the overlap. The code for this algorithm can be found in Appendix A. The degree of overlap varied both between and within companies, from no overlapping forecast sets to nearly fifty per cent of one forecast set appearing in another. Table 2.1 shows that, on average, half of the data files provided for each company were overlapped and so were together merged by our algorithm. Thirty-two of the merged forecasts have generation periods over two-hundred days long resulting from the combination of four individual data files. Therefore, these long merged forecasts capture a whole financial year (250 trading days) of expectations.

In this project, we have only presented results for the thirty-two longest merged forecasts because they contain the same number of internal announcements and are of similar lengths. This ensured that results obtained for different merged forecasts were comparable. Using these merged data sets in combination with 'reading along the diagonal' allowed us to use much longer time series for our analysis. Merging the forecast sets meant that announcement dates were no longer the last generation date. Instead, each series now contained 'internal' announcements that allowed us to view how the investor expectation forecasts changed before and after an announcement. This was impossible when considering the original forecast sets in isolation.

Before analysing an example forecast, we must first provide a mathematical introduction to the Hilbert-Huang Transform.

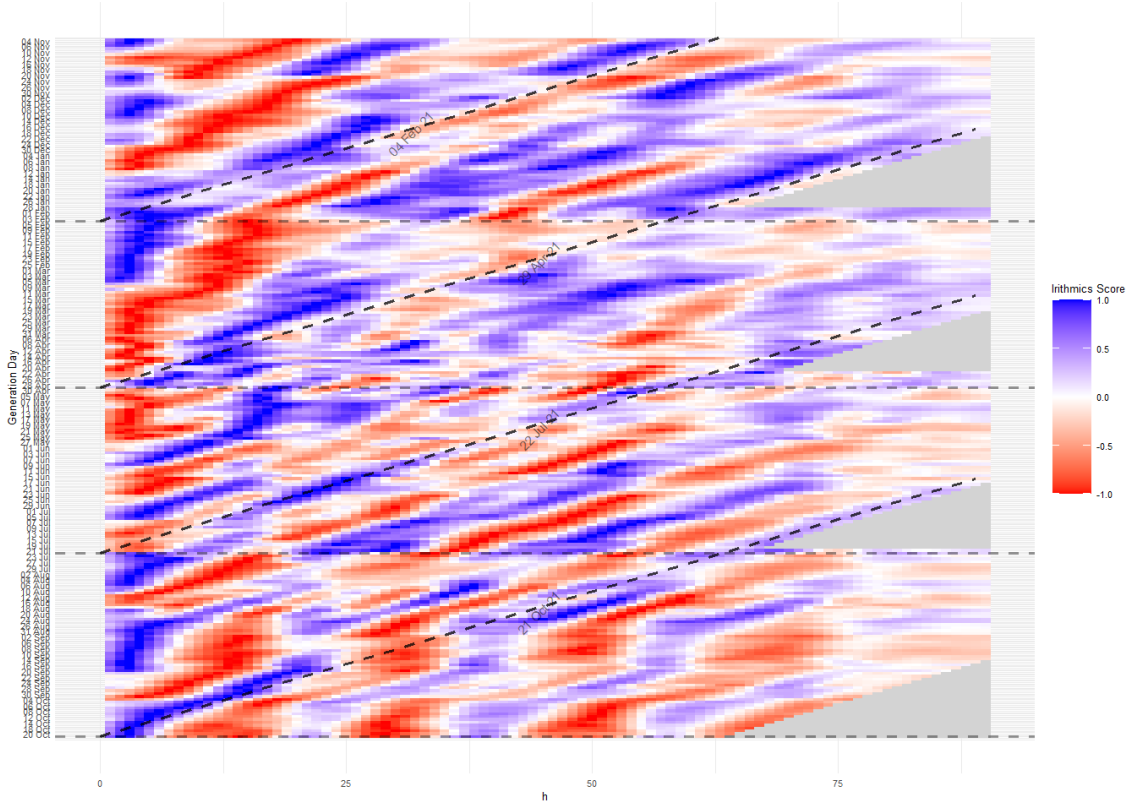


Figure 2.7: 'Reading along the diagonal', $g + h$ series, of the merged expectation forecasts represented as a colour map. Announcement dates are represented as dashed lines.

2.2 Mathematical Overview

In this section we provide a mathematical overview of the Hilbert-Huang Transform (HTT) and the approach used in this paper to calculate cross-correlation coefficients.

2.2.1 Non-stationary time series

Given that the HHT is a method designed for handling non-stationarity time series it is worth reviewing the definition before proceeding.

A time series $X(t)$ is defined as weakly stationary if, for all t,

$$\begin{aligned} E(|X(t)^2|) &< Inf, \\ E(X(t)) &= \mu, \\ Cov(X(t_1), X(t_2)) = Cov(X(t_1 + \tau), X(t_2 + \tau)) &= Cov(t_1 + t_2), \end{aligned} \tag{2.1}$$

where $E(\cdot)$ is the expected value defined as the ensemble average of the series and $Cov(\cdot)$ is the covariance function. This means that the series has a constant mean, μ, and has constant covariance for all time lags τ. A time series is considered non-stationary if it does not fulfil these conditions. (10)

Financial data fails to meet the conditions of weak stationarity due to trends in the data causing the expected value to be time-dependent and the fact that volatility — the variance of the log returns of an asset — is also time-dependent. To accommodate these traits, Generalised Auto-Regressive Conditional Heteroscedasticity (GARCH) models are often employed for financial time series analysis. (19) In GARCH models the short-term conditional volatility is time-varying, while the long-run unconditional volatility remains constant. This was introduced to economists by Engle (20) over forty years ago in 1982. While GARCH models, and the multitude of proposed adaptations, can capture time-varying volatility, they still require us to make distributional assumptions about the data. There is a vast body of literature addressing this problem for traditional financial time series, however, the expectation data used in this project is not typical. The HHT is a fully adaptive method, circumventing the need to make distributional assumptions about the data under investigation.

2.2.2 The Hilbert-Huang Transform

The key part of the HHT is the EMD algorithm that can decompose any series into IMFs. Without applying EMD, the second step of HHT, applying the Hilbert transform, would not provide a meaningful analytical signal. 16 Therefore, we focus on the EMD method before covering the Hilbert Transform.

The Empirical Mode Decomposition: Intrinsic Mode Functions

For any time series $X(t)$, EMD can be used to decompose it iteratively into a finite sequence of time-dependent oscillating components $IMF_1(t), IMF_2(t), ..., IMF_n(t)$, and a non-oscillatory residual term, $r_n(t)$. (7) The original series can be recomposed by summing these components:

$$X(t) = \sum_{j}^{n} IMF_j(t) + r_n(t). \tag{2.2}$$

The EMD algorithm ensures that, by construction, each $IMF_j(t)$ is an Intrinsic Mode Function (IMF). An IMF must satisfy the following criteria: (10)

1. The number of extrema (maxima and minima) and the number of zero crossings must be either equal or differ at most by one;
2. At any point t the mean value of the envelope defined by the local maxima and the envelope defined by the local minima is zero.

The first condition means that each IMF occupies a narrow frequency band with no riding waves, because it imposes the condition that the IMF series must cross the zero line between each turning point contained in the series. The second condition imposes that the IMF is locally symmetric, as the local mean of the data is zero. Huang et al. (10) note that defining a 'local mean' involves a 'local time scale', which is impossible to define — therefore, as a surrogate, the envelope between local maxima and local minima is used to force the required symmetry (see Figure 2.8 for a visual representation of this concept).

HHT Step 1: the EMD Algorithm

The following list outlines the EMD algorithm mathematically. Figure 2.8 gives a visual representation of the second and most important, step of the algorithm — the sifting step. This outline is adapted from the one provided by Nava et al., (17) and the plots in Figure 2.8 are taken from their paper as well. We have included this plot as we were unable to produce our own toy example within the time constraints of this project, but during our research process, we found this illustration and explanation to be the most helpful in aiding our understanding of the EMD process.

The EMD algorithm to decompose $X(t)$ iterates over the following steps:

1. INITIALIZE STEP: Set the residual term to the series to be decomposed $r_0 = x_t$ and set IMF index $k = 1$.
2. SIFTING STEP: Extract IMF_k:
   (a) initialise by setting target series $H_0(t) = r_{k-1}(t)$ and set an iteration counter $i = 1$;
   (b) find the local maxima and local minima of $H_{i-1}t$;
   (c) Using interpolation methods such as cubic splines to connect local extrema:
      i. create the upper envelope $V_u(t)$ by interpolating between maxima;
      ii. create the lower envelope $V_l(t)$ by interpolating between minima;
   (d) calculate the mean value from between these envelopes as $\mu_{i-1}(t) = (V_u(t) + V_l(t))/2$;
   (e) obtain new target series by subtracting the envelope mean from the target series, $H_i(t) = H_{i-1}(t) - \mu_{i-1}(t)$;
   (f) VERIFICATION STEP: check if the new target series $H_i(t)$ satisfies the IMF conditions:
      i. number of extrema is equal to the number of zero crossings ± 1 ;
      ii. AND envelope mean $\mu_{i-1}(t) = 0$;
   (g) IF either i. and ii. are FALSE, increase counter $i = i + 1$ and repeat the sifting process from step 2. b. using the new $H_i(t)$;

(h) ELSE, set $IMF_k(t) = H_i(t)$ and set residual term $r_k(t) = r_{k-1}(t) - IMF_k(t)$.

3. RESIDUAL EVALUATION STEP:

   (a) IF $r_k(t)$ is either constant, monotonic or contains at most one maximum or minimum, STOP the process;

   (b) Otherwise, set $k = k + 1$ and repeat SIFTING STEP from 2.

Stopping Criteria

Theoretically, the sifting process stops when the IMF properties are fulfilled. However, with real-world data, the local mean can only be approximated to zero, so the sifting algorithm would never converge. Therefore a stopping criterion is required. In their original paper, Huang et al. (10) defined the stopping criterion by limiting the standard deviation of two consecutive sifting results to 0.2. The standard deviation of sifting results is calculated as follows:

$$SD = \sum_{t=0}^{T} \left[\frac{|H_{1(k-1)}(t) - H_{1k}(t)|^2}{H^2_{1(k-1)}(t)} \right] \tag{2.3}$$

However, this was found to be vulnerable to sifting the signal too much and consequently losing all amplitude variation. Therefore, a much simpler stopping criteria was formulated for when the sifting algorithm does not converge. For this stopping condition, the number of extrema and zero crossings are recorded after each sifting step. If they remain unchanged for s consecutive iterations, then the algorithm designates $H_i(t)$ as an IMF and proceeds to the 'Residual Evaluation Step'. The best range for s was found to be between 3 and 8. (21) For this project we used $s = 6$.

In addition, it can be advantageous to stop the decomposition process at a set number of IMFs — however, some oscillatory information will remain in the residual term $r_k(t)$. (21)

Mode Mixing Problem

The original EMD algorithm presented above was found to suffer from the so-called mode mixing problem in which the extracted IMFs may fail to represent the true decomposition correctly. Sometimes an extracted IMF would contain fragments of another IMF with a different oscillatory frequency. This pheneomenon is known as the mode-mixing problem, and causes the decomposition to be inconsistent. (22) To overcome this problem, noise-assisted Ensemble Empirical Mode Decomposition (EEMD) was proposed. The EEMD approach first makes multiple copies of the target series and each copy is slightly distorted by some low-amplitude white noise. The EMD algorithm is then applied to each of these copies and the ensemble mean for each IMF is returned. This means that for large ensemble sizes, any instances of mode mixing are averaged out. In this paper, we use the Complete Ensemble Mode Decomposition with Adaptive Noise (CEEMDAN) to implement EMD and overcome the mode mixing problem.

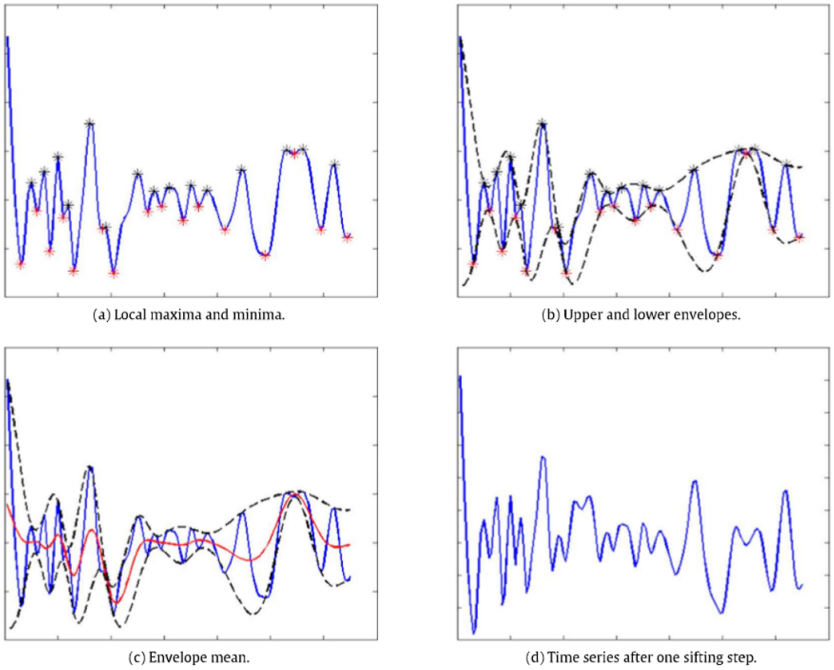
Complete Ensemble Empirical Mode Decomposition with Adaptive Noise

EEMD resolves the mode mixing problem. However, the added noise does not completely cancel out in the averaging process for any finite ensemble size, meaning that the resulting decomposition no longer sums to the original series. (21) Torres et.al., (23) proposed the CEEMDAN modification of EEMD where the averaging over the ensemble is done separately for each IMF component, before extracting the next IMF. The resulting decomposition is termed 'complete' as the added noises cancel out between each step, and so like with the original EMD algorithm, the IMFs sum to the original series. CEEMDAN was also found to be much more computationally efficient than EEMD as it required only half the sifting iterations.

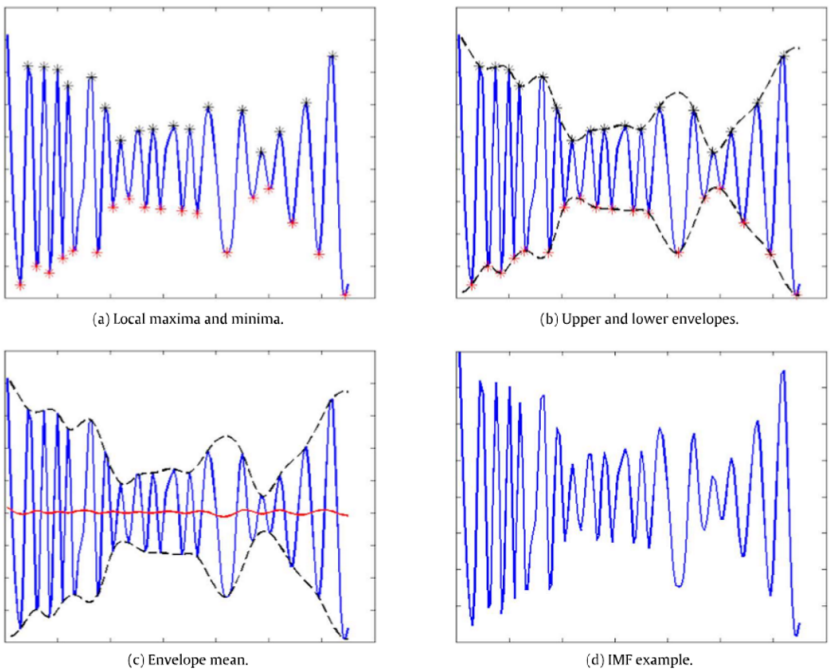
This is just one of many adaptations of EMD that solve the mode mixing problem. (22) However, these are only slight modifications of the original EMD algorithm — therefore, in this project, we refer to EMD in the general sense even though we chose to use the CEEMDAN modification.

HHT Step 2: the Hilbert Transform

The second step of the HHT is to extract the analytical signal for each IMF by applying the Hilbert Transform. An oscillating real-valued function can be viewed as the projection of an orbit in the



(a) An initial sifting step of the EMD algorithm. The resulting series (bottom right) does not satisfy the IMF conditions so the sifting process is repeated.



(b) A final sifting step. The envelope mean (red line, bottom left) is essentially zero so the resulting series is an IMF.

Figure 2.8: Graphical examples of the EMD sifting step produced by Nava et al. (17) Subplot 2.8a shows an initial sifting step that does not produce a valid IMF. Subplot 2.8b shows a final sifting step and the resulting IMF. For each subplot, the top left and right panels show the identification of maxima and minima of $H_{i-1}(t)$ and joining them using cubic splines to obtain $V_u(t)$ and $V_l(t)$. The bottom left panel shows the envelope mean, $\mu_{i-1}(t)$, in red. The bottom right panel shows the resulting series $H_i(t)$ obtained by subtracting the envelope mean from $H_{i-1}(t)$.

complex plane onto the real axis. (7) For an arbitrary time series $X(t)$ the Hilbert Transform is given by,

$$Y(t) = \mathcal{H}[X](t) := \frac{1}{\pi} \int_{-\infty}^{+\infty} \frac{X(s)}{t-s} ds, \tag{2.4}$$

where the improper integral is defined as the Cauchy principal value. (7) $Y(t)$ provides the complementary imaginary part of $X(t)$ to form the analytical signal, $Z(t)$, in the complex plane with Cartesian and polar co-ordinate representations as:

$$Z(t) = X(t) + iY(t) = a(t)e^{i\theta(t)} \tag{2.5}$$

in which

$$a(t) = Mod|Z(t)| = \sqrt{X^2(t) + Y^2(t)},$$

$$\theta(t) = Arg(Z(t) = \arctan \frac{Y(t)}{X(t)}.$$

We refer to $a(t)$ as the instantaneous amplitude and to $\theta(t)$ as the instantaneous phase of the transformed time series as they are time-dependent. The instantaneous frequency is defined as the derivative with respect to time of the instantaneous phase:

$$\omega(t) = \frac{\mathrm{d}\theta(t)}{\mathrm{d}t}. \tag{2.6}$$

This instantaneous frequency describes local frequency changes within a time series and so is useful for non-stationary signals. Discussion on the derivation and 'controversy' surrounding this definition can be found in Huang et al. (10) The Hilbert transform can be applied to any time series — however, the resulting instantaneous frequencies may be negative and bear no relationship with the real oscillation of the data. It is only by first applying EMD to extract oscillating IMFs with zero mean that the analytical signal obtained by the Hilbert Transform is physically meaningful. (16) This realisation is the crux of the HHT methodology.

Similarity of the HHT and the Fourier Transform

As the Hilbert transform is only meaningful for oscillatory series, we do not apply it to the non-oscillatory residual term, $r_n(t)$. After applying the Hilbert transform on each IMF, the original time series can be expressed as the sum of the real parts and the residual:

$$X(t) = Re\left\{\sum_{j=1}^{n-1} a_j(t) e^{i\int \omega_j(t)\mathrm{d}t}\right\} + r_n(t). \tag{2.7}$$

With this representation, it is worth noting the similarity to the Fourier decomposition of the same signal:

$$X(t) = Re\left\{\sum_{j=1}^{n} a_j e^{i\theta_j}\right\}. \tag{2.8}$$

Unlike in the HHT the amplitude a_j and phase θ_j of the Fourier decomposition are constants. Therefore, the HHT can be considered as a time-dependent generalisation of the Fourier Transform that can locally describe the properties of a series at any time point t. (13) Furthermore, the Fourier Transform requires that the trend be removed before decomposing the signal; however, the data-driven approach of EMD does not require trend removal, and so requires fewer assumptions to be made about the data. (10)

2.2.3 Time Series Filtering using HHT

By construction, the EMD algorithm first identifies and extracts the finest structures and then iteratively extracts IMFs with lower frequencies corresponding to longer and longer timescales. We can use this feature and the fact that the original series can be recomposed as the linear combination of the IMFs and the residual, $r_n(t)$ (Equations 2.2, 2.7), to construct low and high-pass filters for the original series. Considering the often noisy and complex nature of financial data, this time scale separation is a valuable feature of the HHT approach. (17)

A low-pass filter excludes the high-frequency fluctuations of a time series. Therefore, to construct a low-pass filter we do not include the first few IMFs. For a series that has been decomposed into n IMFs, and $k < n$, a low-pass filter is defined as:

$$X_L^k(t) = \sum_{j=n}^{k} IMF_j(t) + r_n(t). \tag{2.9}$$

Similarly, if we only select the high-frequency IMFs, we can construct a high-pass filter:

$$X_H^k(t) = \sum_{j=1}^{k} IMF_j(t). \tag{2.10}$$

Such a filter allows us to consider only the short-term fluctuations of the series. In Section 4.2.1 we show that such a high-pass filter can be used to estimate a time-dependent volatility measure — variability. Another interesting application, but not one that is explored in this project, is using this filtering technique to remove the time series trend by setting $k = n$, effectively discarding the residual term $r_n(t)$.

2.2.4 Cross-correlations using the IMFs

As IMFs are seemingly stationary, as by definition they have zero mean, we can assume they are locally stationary, and so we can calculate meaningful cross-correlation estimates between the series. (17) However, also by definition, they are still time-dependent and have time-varying instantaneous amplitudes and frequencies (Equation 2.7). As a result, calculating correlations between whole series may obscure real local correlations between IMFs. Therefore, Chen et al. (18) proposed the so-called Time Dependent Intrinsic Correlation that uses an adaptive rolling window to calculate the correlation between two nonlinear non-stationary time series. In this investigation, we calculate the correlation between expectations and price using both the standard Pearson cross-correlation method and an adapted version of the simplification of the TDIC proposed by Nava et al. (17) We call this adaptation the Time Dependent Lagged Cross-Correlation (TDLCC).

As we assume that our original time series are non-stationary, both methods present an improved approach to exploring the inter-series relationships as EMD has adaptively transformed our original series into a set of time series that satisfy traditional statistical assumptions. Furthermore, by calculating correlations between paired IMFs of similar frequencies relevant lead-lag relationships may be uncovered at specific time scales that would be otherwise lost when only considering the original through usual non-stationary time series analysis.

To formally explain the correlation methods used in this paper, let us consider two time series $X(t)$ and $Y(t)$ with, $t = 1, 2, 3, ..., N$. Which have been decomposed via EMD into a series of IMFs, $IMF_i^X, IMF_j^Y; i, j = 1, ..., n$.

Pearson Cross-correlation

We can compute the Pearson correlation coefficients between two component series IMF_i^X, IMF_j^Y in the standard way as follows:

$$\rho_{i,j}^{XY} = \frac{1}{N} \sum_{t=1}^{N} \frac{(IMF_i^X(t) - \overline{IMF_i^X})(IMF_j^Y(t) - \overline{IMF_j^Y})}{\sigma_i^X \sigma_j^X} \tag{2.11}$$

where $\overline{IMF_i^X}$ denotes the sample mean over time of IMF_i^X and σ_i^X denotes the sample standard deviation of IMF_i^X, and similarly for $Y(t)$. Note that the IMFs do not have to be the same, although, large correlations only really occur between IMFs with similar frequencies (see Section 3.3).

While the IMFs may be considered locally stationary, the residual, $r_i(t)$, is not and so the correlation coefficient between residuals is just the measure of linear dependency of the trends of each series indicating their direction of co-movement. Therefore, the correlation coefficient is likely to be high, and should not be compared directly with the correlation coefficients found for the IMFs — but it remains relevant for the analysis as it quantifies the similarity between series' trends. (17)

Time-dependent Intrinsic Correlation

In their 2010 paper, Chen et al. (18) proposed the TDIC for comparing two non-stationary time series based upon the use of EMD. This method calculates the Pearson correlation between IMFs using an adaptive window whose length depends on the instantaneous periods $T_i^X(t)$ and $T_j^Y(t)$ of IMF_i^X and IMF_i^Y. The instantaneous period of series $X(t)$ is defined as,

$$T_i^X(t) = \frac{1}{\omega_i^X(t)}, \tag{2.12}$$

where $\omega_i^X(t)$ is the instantaneous frequency derived in Equation 2.6. $T_i^Y(t)$ is obtained similarly. The minimum adaptive window size at a time point, t, is set as $W_{t,i} = max(T_i^X(t), T_i^Y(t))$. Setting the minimum window size to the instantaneous period ensures that the correlation is calculated over at least one whole period of the IMF for any point t — so, the series is stationary within the adaptive window. (18) The Pearson correlation of the two series is calculated by setting t as the centre of the adaptive window and modifying Equation 2.11 to be:

$$\rho_i^{XY}(t) = \frac{1}{\tau_u - \tau_l - 1} \sum_{\tau_l = t - \frac{W_{t,i}}{2} + 1}^{\tau_u = t + \frac{W_{t,i}}{2}} \frac{(IMF_i^X(\tau) - \overline{IMF_i^X})(IMF_i^Y(\tau) - \overline{IMF_i^Y})}{\sigma_i^X \sigma_i^X}. \tag{2.13}$$

The lower limit of the sum is $\tau_l = t - \frac{W_{t,i}}{2} + 1$ so when $t = \frac{W_{t,i}}{2}, \tau_l = 1$. And the coefficient $\frac{1}{\tau_u - \tau_l - 1} = \frac{1}{W_{t,i}}$, is just the reciprocal of the adaptive window size.

They then extend the TDIC to include a lag term λ, which is common when calculating windowed cross-correlations as the windowing may miss important lead-lag inter-series relationships. The problem with this method is that the extra window size variable makes reporting and interpreting results much more difficult and requires a three-dimensional plot (18).

Time-Dependent Lagged Cross-correlation

Due to the constraints of this project, we have chosen to implement a simplified version of the TDIC with a fixed window size, as proposed by Nava et al. (17) Instead of an adaptive window based on the instantaneous period, T_i^X, a suitable window size is chosen based on the mean period, $\overline{T}_i^X$, of each series.

As IMF_i^X oscillates around zero, its mean period can be approximated by dividing the length of the series by the number of maxima, m_i^X, observed within the series:

$$\overline{T}_i^X = \frac{N_i^X}{m_i^X}. \tag{2.14}$$

The fixed window size is set as $W_i = max(\overline{T}_i^X, \overline{T}_i^Y)$. This trade-off retains some of the adaptive nature of the TDIC, as the window size depends on the average periodicity of the series but no longer requires the window size to be time-dependent. This simplification is less rigorous than the TDIC, and does not technically guarantee that the series is stationary within the window. However, this method significantly improves upon assuming stationarity for the whole series.

In their simplified approach, Nava et. al. choose to calculate the cross-correlation over the window cast backwards from time point t (e.g. for zero lag the the sum would be between $\tau = [t - W + 1, t]$) so that each series had the same endpoint but different starting points depending on the window size. (17) However, we found that for large window sizes the correlation coefficient could not be calculated before the first announcement date for the merged expectation forecasts. As we are interested in the internal features of the forecast series due to these announcements, we decided to use a centred window, like in the original TDIC (Equation 2.13). This ensured correlation coefficients were calculated around the first announcement date, but means that we were unable to calculate correlations for the first and last $\frac{W_i}{2}$ observations in the series.

Therefore, the modified TDLCC equation used in this project is defined as:

$$\rho_{i,j}^{XY}(t, \lambda_i) = \frac{1}{W_i - \lambda_i} \sum_{\tau = t - \frac{W_i}{2} + 1}^{t + \frac{W_i}{2} - \lambda_i} \frac{(IMF_i^X(\tau) - \overline{IMF_i^X})(IMF_j^Y(\tau + \lambda_i) - \overline{IMF_j^Y})}{\sigma_i^X \sigma_j^X} \tag{2.15}$$

In their paper, Nava et.al. set $|\lambda_i| \leq max(\overline{T}_i^X, \overline{T}_i^Y)$ and $W_i = max(|\lambda_i|)$ so that there were no repetitive lagged patterns in the correlation matrix. (17) However, we found that this caused computational issues for maximum positive lag cases as $W_i = \lambda_i$ implies a division by zero in Equation 2.15. This contributed to our decision to adapt their approach to calculate the TDLCC using a centred window. To avoid lagged patterns in the correlation plot, we set the lag to be $|\lambda_i| \leq \frac{W_i}{2}$ — therefore avoiding any issues with division by zero.

Although this is based on the simple Pearson correlation measure, the use of EMD and a rolling window adapts to the local nature of the data and provides a dynamic correlation measure across time scales. (17)

2.3 Applying HHT to an Example Data Set

To investigate the potential benefits of using the HHT to analyse the expectation data provided by Irtithmics — and to illustrate the methods used in this project — we present an applied example for one arbitrarily chosen merged forecast set. For the sake of both the brevity of this project and clarity for the reader, we decided against introducing another example data. So we will continue to use the same example merged data set from Section 2.1, for Unilver (ULVR) containing the expectation forecasts generated between 04/11/2020 and 20/10/2021.

We obtained the daily stock price data for ULVR using Refinitiv Workspace. 1 This price data and the code used to obtain it can be found in Appendix A. For the price series, used in this section, we selected the closing price for the trading days 05/11/2020–21/10/2021, as these corresponded to the values of the $g + 1$ diagonal series. This meant that when we would be calculating the correlation between the expectations for a given day, and the price that was observed on that day. This should capture the reflexivity between expectations about a day, and the price on that day. An alternative method could have been to consider a longer price series than the expectation series, however, that was not possible considering the time constraints of this project. We also apply the HHT methods to the price change series calculated from the difference between the closing price on day g and the closing price on day $g + 1$.

A note on logarithms

Following the example of Huang et.al, when applying the HHT to financial data, we do not take the logarithm of our price data as is usual in financial statistical analysis. This transformation is usually done so that the target series is seemingly stationary. (16) As the HHT is designed for nonlinear, non-stationary time series, this usual transform is not necessary. Exploratory analysis confirmed that taking logarithms only affected the absolute amplitudes of the IMFs, but their instantaneous frequencies were unchanged. Additionally, as the Irithmics data contains negative values by design, taking the logarithm of those forecasts would be impossible.

2.3.1 HHT Step 1: Extracting IMFs

For this paper, we chose to use the R package `Rlibeemd` developed by Luuko et al. (21) to implement our chosen EMD algorithm, CEEMDAN, instead of coding our own implementation of the algorithm. The package is very well documented, and as it is written in C, we found that it runs much more efficiently than similar EMD packages or our own attempts at implementing the algorithm. We wrote code to decompose all of the merged data sets at once utilizing parallel computing and stored the processed data. This data processing allows analysis to be done across a large number of time series without rerunning the decomposition — increasing the efficiency and reproducibility of our analysis. We then selected the IMFs of the ULVR example set from the processed data set.

In Figure 2.9, we present the IMFs abotained for the $g + 1$ diagonal expectations series, and the corresponding stock price close and price change time series. Some features are immediately apparent: amplitude and frequency spikes around the announcement dates, and broadly similar shapes between the lower frequency IMFs (IMFs 5-7) of the different series. We only present the time series of the IMFs obtained from the $g+1$ expectations for clarity, however, we also performed the EMD on all the $g + h$ diagonal series contained in the example forecast set.

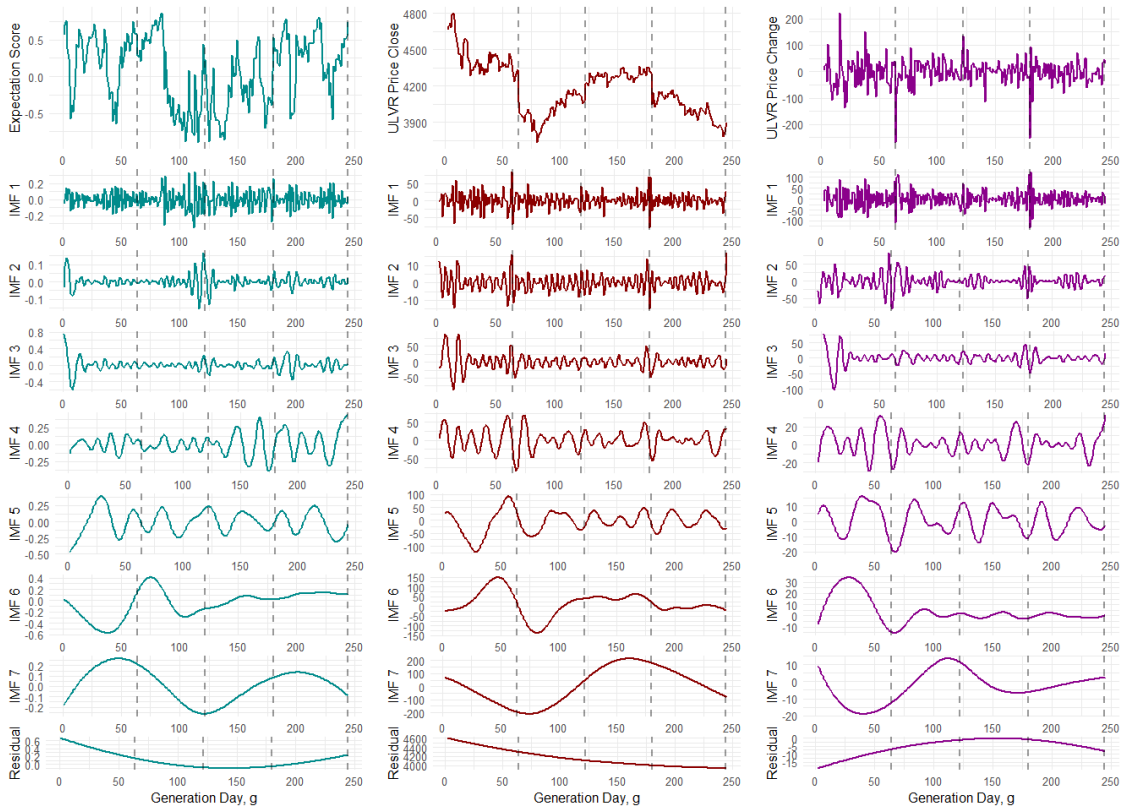


Figure 2.9: The original series and the extracted IMFs for the $g+1$ expectations (left), ULVR stock price close (middle), and the change in stock price (right). Note that the x-axis is in relative time units g as opposed to dates. Announcements have been added in the usual way.

2.3.2 Filtering using EMD in Practice

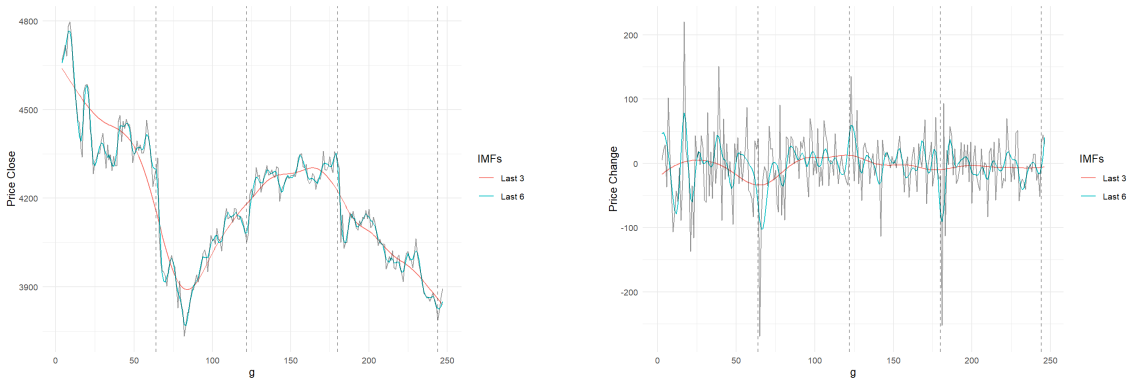
Filtered Stock Price Time Series

To demonstrate the ability to filter out the high-frequency oscillations using EMD, we applied the low-pass filter (Equation 2.9) using the IMFs of the ULVR stock price and price change series. The red filtered series in Figure 2.10 contains only the residual and the last two IMFs and was obtained by setting $k = 6$, in Equation 2.9. The blue series was obtained by setting $k = 3$, and contains all of the IMFs except for IMF_1 and IMF_2. In Figure 2.10a the last six component series almost completely match the original series, and the last three seem to provide a convincing estimate of the mean trend of the stock price series. However, the low pass filter has not been as effective in capturing the dynamics of the price change series. By consulting the y-axes of the IMF series presented in Figure 2.9, we find that for IMF_1, the stock price series oscillates between $[-50, 50]$, whereas for the price change series IMF_1 oscillates between $[-100, 100]$. This range is fifty per cent of the range of the original price change series $[-200, 200]$, whereas the range of IMF_1 for the price is only about five per cent of the original stock price series range. Therefore, we might expect a high-pass filter as defined in Equation 2.10 to be more suitable for capturing the dynamics of the price change series — suggesting that its dynamics are primarily over short time periods.

Filtered Expectation Forecasts

Using the IMFs of each of the $g+h$ expectation series, we can apply low and high pass filters to the whole ULVR example set of expectation forecasts. The resulting colour maps are presented in Figure 2.11. The low-pass filtered set (Figure 2.11d) shows broad periods of positive and negative expectations but not the shorter regions of negative or positive expectations within these larger regions. Figure 2.11c contains the information, that when added to the low-pass colour map, results in a filtered colour map that is almost identical to the original set of expectations (Figure 2.11e).

These filtered colour maps show that the first step of the HHT method successfully separates out the information contained in the expectation forecasts at different timescales. This separation could be used to visualise how longer-term strategic and shorter-term tactical allocations of investor



(a) ULVR Stock Price

(b) ULVR Price Change

Figure 2.10: Linear combinations of the example ULVR stock price and price change series IMFs overlaid on the original series (grey). The red lines are the linear combination of the last three component series $X_L^6 = \sum_{j=7}^{6} IMF_j + r_7$. The blue lines are the linear combination of the last six component series $X_L^3 = \sum_{j=7}^{3} IMF_j + r_7$. These are examples of low-pass filters (Equation 2.9) with $k = 6$ and $k = 3$ respectively.

capital are represented by the expectation forecasts produced by Irithmics. Therefore, the filtering method is a major result of this project. However, we have chosen not to include further examples of its application beyond the ULVR data set as the method's utility is specific to analysing individual data sets — more examples without market context would be meaningless. So we leave further exploration of time series filtering using HHT to a future investigation.

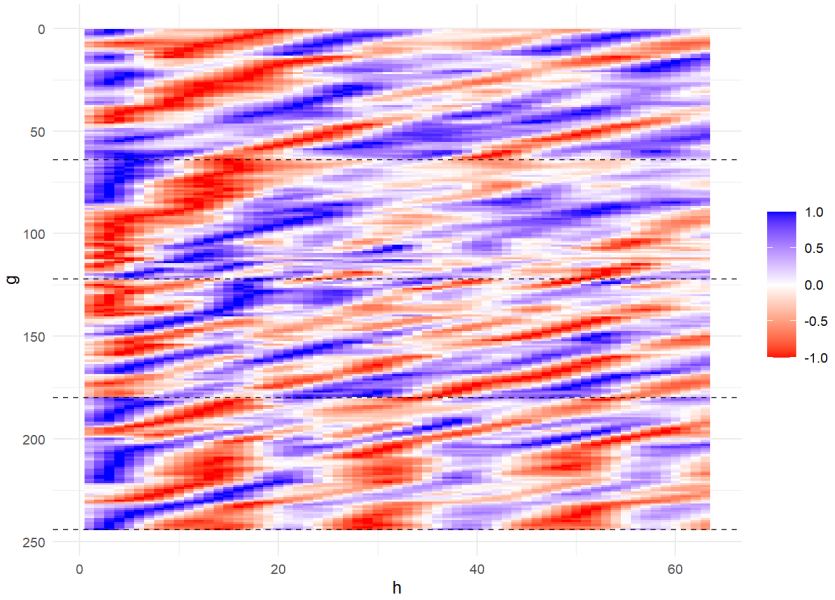
2.3.3 HHT Step 2: Applying the Hilbert Transform

The second part of the HHT is to apply the Hilbert Transform to obtain analytical signals for each IMF (Equation 2.5). We obtain the instantaneous amplitude (IA) and instantaneous frequency (IF) for each time point g as described in Section 2.2.2. We used the `EMD` package to implement the Hilbert transform and apply numerical differentiation to obtain the IF. (24) The IA and IF time series for each IMF are shown in Figure 2.12 and 2.13 respectively. These plots illustrate the non-stationary dynamics preserved by EMD within each IMF as both variables change with time. Visually there is seemingly some similarity between the relative amplitudes of the series.

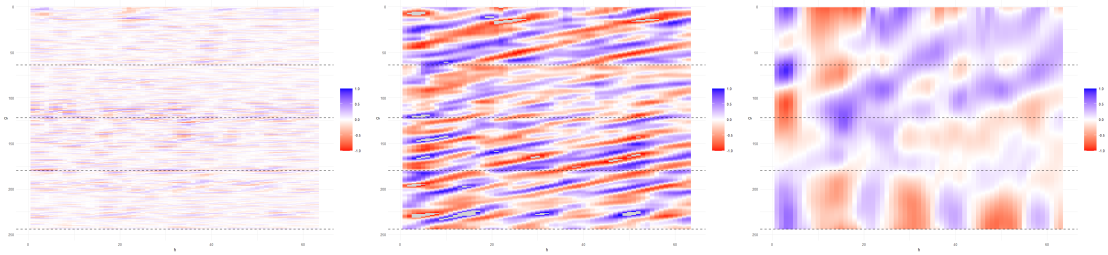
2.3.4 Spectral Analysis: Instantaneous Frequency Plots Example

Many of the papers that use the HHT for data analysis construct the Hilbert spectrum from the IF and IA of the IMF components to explore the time-frequency-energy dynamics of a target series (for applications of this approach to financial data see Huang et al. (16,) and for examples from the natural sciences see Bowman and Lees (11), Vecchio et al. (12), and Barnhart and Eichinger (13)). However, we found that this holistic method did not provide any meaningful insights into the expectation data. Therefore, instead of presenting the Hilbert spectra for each series, we followed the example set by Leung and Zhao (7) to explore the spectral dynamics of financial time series and plotted the pairs of instantaneous frequencies obtained for the same generation day, g.

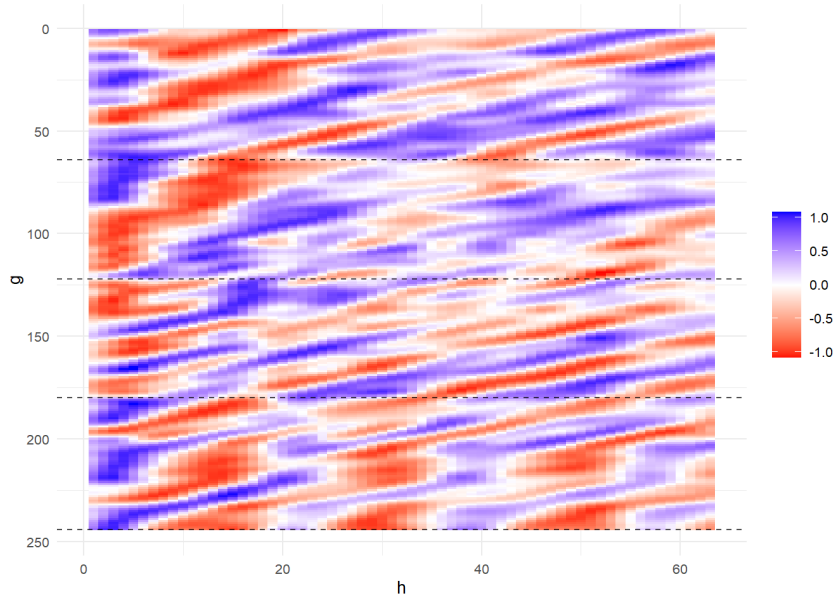
For any time series decomposed via the HHT there is an instantaneous frequency profile made up from their IMFs which allows us to compare the non-stationary dynamics of different time series. (7) In Figure 2.14 we have plotted the pairs of IFs for each of our example series. The dashed black line is a reference line of identical frequencies, and the black crosses indicate the cluster means for each IMF. The cluster means for all the IMFs are close to the reference line of identical frequencies. The frequency pairs associated with the first IMF seem to form a patternless scatter, whereas for the other IMFs they are more tightly clustered on the reference line. This suggests that for these lower frequency IMFs there is a linear relationship between the IMF frequency profiles. There is a visibly closer (and expected) linear relationship between the instantaneous frequencies of price and price change IMFs.



(a) Original $g + h$ expectation colour map.



(b) $IMF_1 + IMF_2$ (c) $IMF_3 + IMF_4 + IMF_5$ (d) $IMF_6 + IMF_7 + Residual$



(e) Low-pass filtered expectations color map: $X_L^3 = \sum_{j=7}^{3} IMF_j + r_7$.

Figure 2.11: Using the IMFs of the $g + h$ expectations to apply low-pass filtering to the whole expectation forecast colour map. Excluding the first two IMFs returns a similar colour map to the original series. The colour map2.11e can also be obtained through the linear combination of 2.11c and 2.11d.

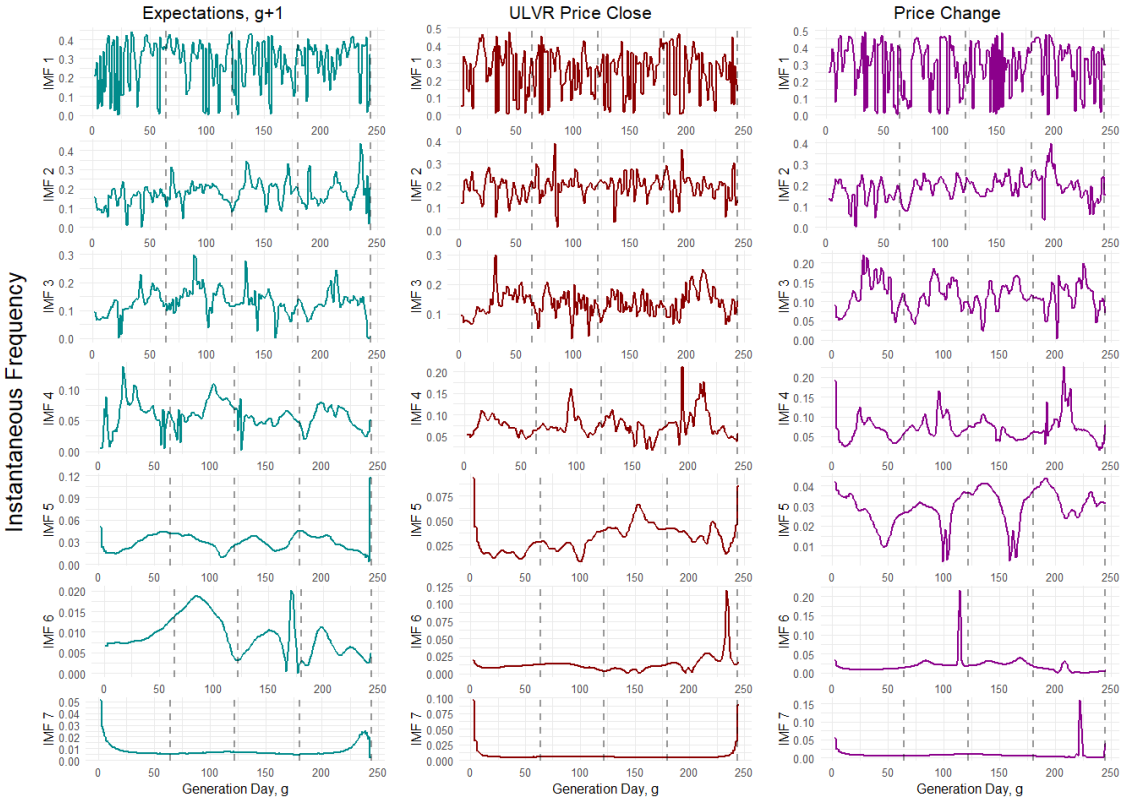


Figure 2.12: Instantaneous frequencies for the three example series. We can see that the corresponding IMF for each series have similar frequency ranges, but quite different series.

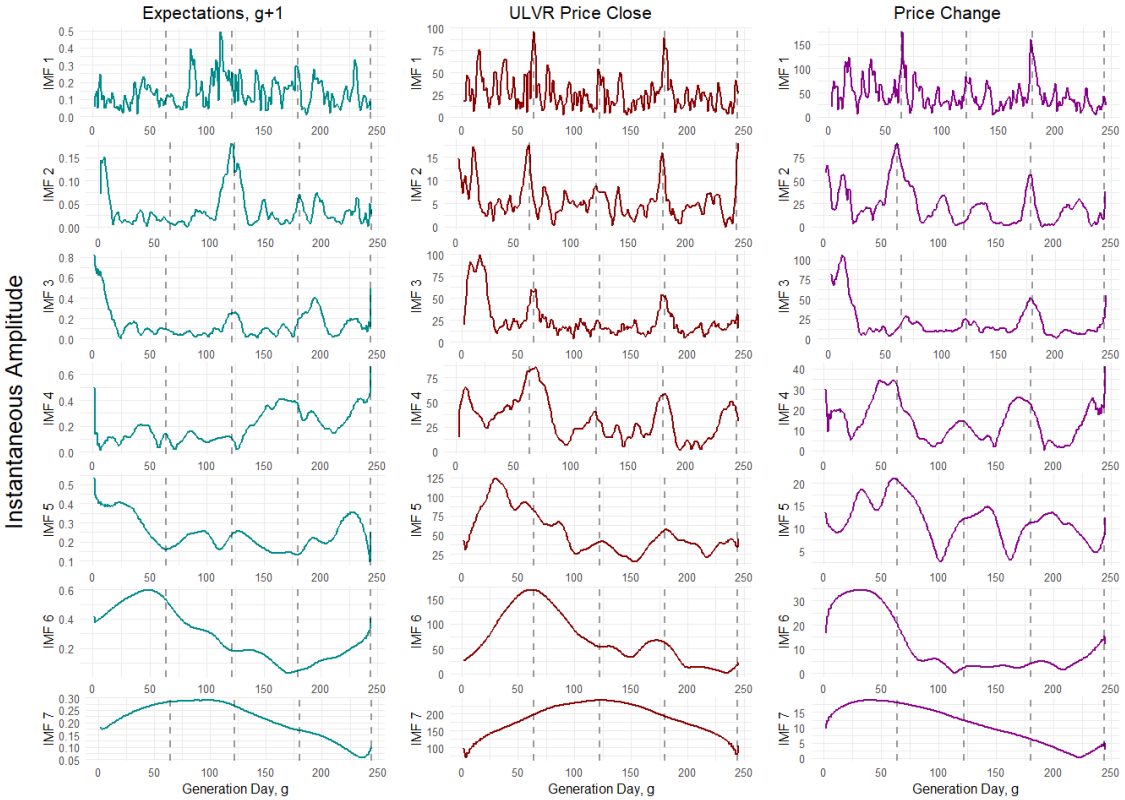
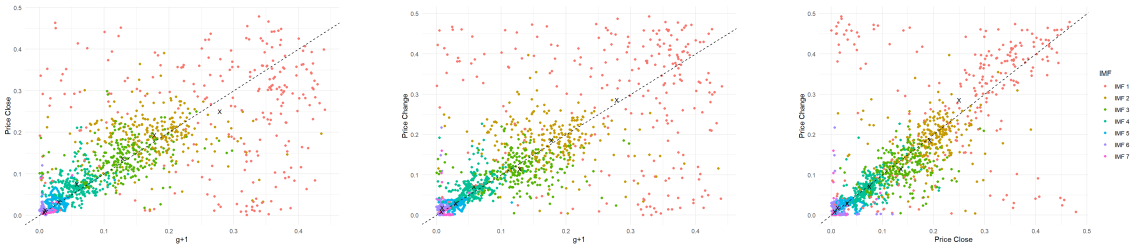


Figure 2.13: Instantaneous amplitudes for the three example series. Visually there seems to be evidence of lagged similarities between the relative amplitudes of the IMFs.



(a) $g + 1$ expectations versus ULVR price close.

(b) $g + 1$ expectations versus price change.

(c) ULVR price close versus price change.

Figure 2.14: Instantaneous frequency plots for the three example series. The black crosses are mean values for each IMF cluster.

2.3.5 Pearson and Time-Dependent Lagged Cross-Correlation Example

In this section, we investigate the cross-correlation between the $g+1$ series and the corresponding price close and price change series and do not consider the correlations between the price series and the other $g+h$ expectation series. This was a product of both practical and theoretical necessity, as defining what time period to compare the relationship between the $g+h$ expectations and the observed stock price series suffered from the forecast distance problem laid out in Section 2.1.1.

Pearson Cross-correlation example

First, as a benchmark to compare the TDLCC against, we compute the Pearson cross-correlation between the IMFs using Equation 2.11. The cross-correlations of the $g+1$ expectations with stock price and the stock price change are presented as correlation matrices in Figure 2.15. We can see that the greatest correlations occur on the diagonal between the IMFs with similar frequencies. Correlations are greater for the low-frequency IMFs for both price and change. The strong residual correlations are expected given the shape of the residual time series in Figure 2.9. We also must be careful not to draw any conclusions from the correlations between the residuals and the other IMFs, as the residual term is not assumed to be locally stationary.

(a) $g + 1$ expectations versus price close

(b) $g + 1$ expectations versus price change

Figure 2.15: Pearson cross-correlations calculated using the IMFs of the example ULVR series.

Time-Dependent Lagged Cross-Correlation Example

In Section 2.2.4 we defined our adaptation of the TDLCC using a fixed window size relative to the mean periods, $\overline{T}$, of the IMFs used in the calculation. We then derived the lag, λ to be half the window size to avoid repetitive correlation structures in the output. However, for very low-frequency IMFs, and the non-oscillatory residual term $r(t)$, these definitions are not suitable as their mean periods are often as long as, or longer, than the original series. Therefore, we also added the condition that the window size, $W \leq 90$, and correspondingly, $|\lambda| \leq 45$. Whilst necessary, we must acknowledge that this condition violates the assumption that at least one complete period is used to calculate the correlation coefficient for the IMFs. We also set a minimum window size of $W \geq 6$ so that we were not calculating correlations over three-day windows for the highest frequency IMFs. We chose six days as it is roughly twice the average mean period of the first

IMF, and so should still capture meaningful correlations if they exist. (18) To explain what these conditions mean for the window size in practice, we present the mean periods, windows, and lags used in the calculation of the TDLCC for the Unilver example series in Table 2.2.

Table 2.2: Rounded mean periods for each IMF, the corresponding window sizes, and the maximum lags used to calculate TDLCC.

(a) $g+1$ versus Price Close

IMF Series	$\overline{T}_{g+1}$	$\overline{T}_{Price}$	W_i	λ_i
IMF_1	3	3	6	3
IMF_2	5	5	6	3
IMF_3	7	7	7	3
IMF_4	14	11	14	7
IMF_5	35	27	35	17
IMF_6	81	49	81	40
IMF_7	122	244	90	45
Residual	—	—	90	45

(b) $g+1$ versus Price Change

IMF Series	$\overline{T}_{g+1}$	$\overline{T}_{Change}$	W_i	λ_i
IMF_1	3	3	6	3
IMF_2	5	5	6	3
IMF_3	7	8	8	4
IMF_4	14	14	14	7
IMF_5	35	27	35	17
IMF_6	81	49	81	40
IMF_7	122	244	90	45
Residual	—	—	90	45

TDLCC Example Results and Discussion

The results of applying TDLCC to the example ULVR series are presented as colour maps of the correlation coefficients in Figure 2.16. These plots support the inferences drawn from the Pearson cross-correlation matrices in Figure 2.15. The low-frequency IMFs and Residual have consistent bands of positive and negative correlation for $\lambda = 0$ for almost the whole observation period, and so, they exhibit relatively large whole-series Pearson correlations. For the higher frequency IMFs, there are fewer persistent correlation patterns — especially for IMF_1. However, for IMFs 2 - 4, which have periods from five to fourteen days, or in other words, one to three trading weeks, we can see many examples of long periods of persistent, large correlations and exhibit abrupt breakpoints and inversions from positive to negative correlation, or vice-versa. There is also some evidence to suggest that these changes in local correlation occur in the proximity of an announcement date — which would support the hypothesis that such announcements affect the reflexive market dynamics between expectations and stock prices.

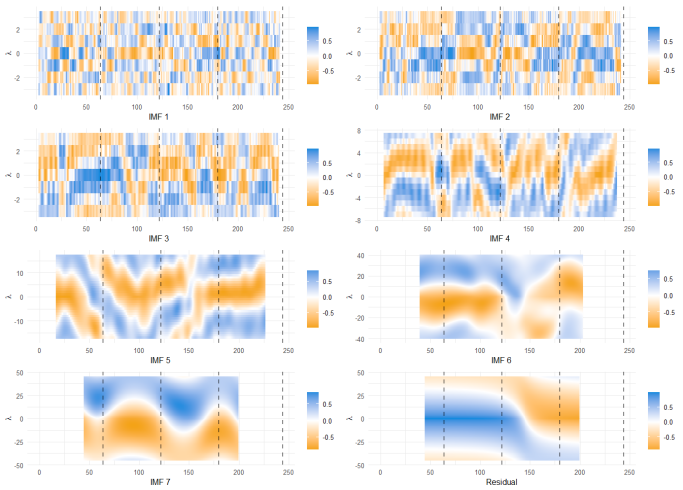
Finally, it is worth reiterating that the residual series are not stationary, and so the correlation coefficients calculated for these series do not represent a true relationship but only a relative correlation between the series trends. However, this residual correlation remains a valuable analytical tool. (17) For example, if the obvious point in the residual correlations in Figure 2.16a is associated with particular changes in either series or even the information available to the markets, then the residual correlation shows how this affected the underlying trend relationship between the two series.

2.3.6 IMF Significance Test Example

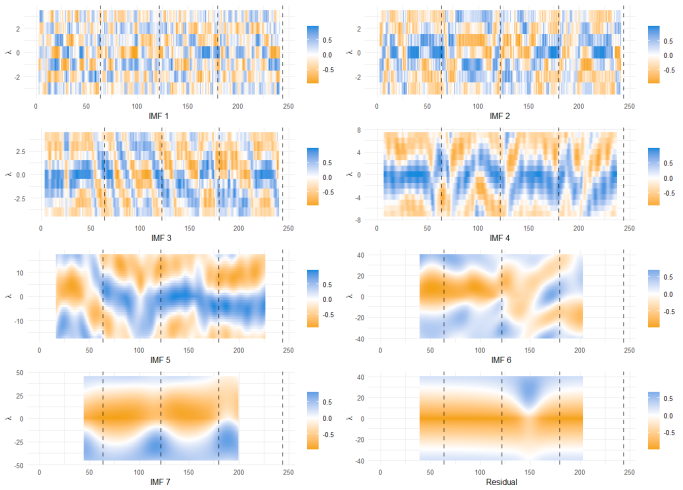
In this section, we introduce a significance test for the information content of the IMFs proposed by Wu and Huang. (25) We first explain the methodology of the test before illustrating it by applying it to the example ULVR series.

The Wu Significance Test

The analysis conducted so far using the IMFs obtained from the HHT method is only valuable if we can be confident that the IMFs contain meaningful information about the original series. We can test whether the information contained in each of the IMFs is statistically different from if they were just white noise using a graphical test developed by Wu and Huang. Through Monte Carlo simulation methods and theoretical mathematics, Wu and Huang found that the energy-density function of the IMFs of white noise is chi-squared distributed. From this, they derived the energy-density-spread function and the spread lines for a graphical significance test. The spread lines indicate the region on a log energy density, $\ln E$, against log mean period, $\ln \overline{T}$, plot where we expect the IMFs of a decomposed white noise signal to be plotted. Therefore, any IMF with an



(a) $g + 1$ expectations versus Price Close.



(b) $g + 1$ expectations versus Price Change.

Figure 2.16: TDLCC for ULVR example series calculated using Equation 2.15. Dashed lines are the announcement dates during the generation period. The x-axis contains the generation day g that was used as the centre of the window. For the larger window sizes, we were unable to calculate correlations for the first and last $\frac{W_i}{2}$ values — he low-frequency IMFs have shorter colour maps as a result.

energy density profile that lies inside this region could be considered the result of the decomposition of a white noise signal — therefore, containing no information by the definition of white noise. (25)

The method for the Wu significance test is as follows:

1. Decompose the target (normalised) data set into IMFs using EMD;
2. Calculate the spread function for various percentiles;
3. Select a confidence-limit level and determine the upper and lower spread lines;
4. Finally, compare the energy density of the IMFs with the spread functions. If the IMF is plotted above the upper, or below the lower, spread line, we may conclude that it is statistically different than white noise at the selected confidence level.

Along with the theoretical distribution, Wu and Huang also proposed a Gaussian approximation of the energy density distribution of white noise with the corresponding spread lines defined as:

$$y = -x \pm k\sqrt{\frac{2}{N}}e^{x/2} \tag{2.16}$$

where $y = \ln \overline{E}$, $x = \ln \overline{T}$, k is the quantile of the standard normal distribution for the desired confidence-limit level, and N is the number of observations in the sample. So, for the one and five

per cent spread lines, k would equal -2.326 and -1.645 respectively. These normally approximated spread lines only significantly diverge from the theoretical spread lines for values of $x > 6$ which applies to IMFs with mean period, $\overline{T} = e^6 = 403.429$ days. This is much greater than even the longest merged data set (Table 2.1), and through initial investigation of the whole data set, we found that the data used in this investigation does not approach this periodicity. Therefore, in this project, we use these approximated spread lines instead of deriving the theoretical spread lines from the distribution derived by Wu and Huang. As the approximated spread lines define a slightly wider region than the theoretical spread lines, (25)we may find more IMFs to not be statistically significant than if we had employed the true distribution. However, for the purposes of this initial investigation into the application of the HHT, the simplification of the methodology is preferred.

Normalisation

Wu and Huang make a passing reference to 'normalizing' (25, p. 1607)the target series when laying out their testing methodology; initial analysis found that not scaling the data before calculating energy density caused large differences in which IMFs were considered to be significant. Consultation of the literature surrounding the application of the HHT and EMD for time series analysis uncovered only a few examples of the Wu significance test being used, but with no mention of how they normalised their series before applying the test (for examples, see Barnhart and Eichinger (13), Vecchio et al. (12), and Masselot et al. (14)).

There are multiple normalisation techniques, but two of the most common are z-score normalisation, where a variable is scaled according to its mean and standard deviation, and min-max scaling where a variable is mapped to a predefined range (commonly 0 to 1). (26) As the HHT method is designed for non-stationary data we do not believe that the z-score scaling method is a suitable normalisation technique as the unconditional mean and variance are undefined. (19) Therefore, we use the min-max method for normalising our series before decomposition. This is defined by the following equation:

$$X^*(t) = \frac{X(t) - min_X}{max_X - min_X}. \tag{2.17}$$

This ensures that the values of $X^*(t)$ are in the range $[0, 1]$ and that all other relative features are maintained. The frequency (or period) profile of the series remain unchanged, but the absolute value of the instantaneous amplitudes of the decomposed IMF are different. Min-max normalisation fails when one does not know the minimum and/or maximum values of $X(t)$, making it problematic in forecasting applications. (26) However, as we are applying this test to historical data (and the fact that the Irithmics expectations forecasts are bound between $[-1, 1]$) we know the maximum and minimum values for our target data series during the observation period.

Applying normalisation to the target series before EMD ensures that the energy densities of the IMFs are comparable to the energy densities used to define the white-noise energy density distributions. Wu and Huang found that the product of the energy density, E_n and mean period, $\overline{T}_n$ is a constant and that if the white noise is normalized, then without loss of generality, this constant can be unity. (25) Therefore, by taking logarithms, this gives the following equation to describe the relationship between energy density and mean period:

$$\ln E_n + \ln \overline{T}_n = 0. \tag{2.18}$$

Min-max normalising a series ensures that for the decomposed IMFs $E_{max} = 1$, and $\ln E_{max} = 0$ thus satisfying the condition used in the derivation of the energy density distribution for all mean periods $\overline{T}_n$. Alternatively, one could specify a different constant and "shift" the white noise distribution up or down the y-axis to fit the data. However, for consistency with the literature, we will proceed by normalising the series prior to EMD.

Applying the test to the example series: Results and Discussion

We have already calculated the mean periods for each IMF and these are reported in Table 2.2, and these are unchanged by min-max normalisation. The energy density of any series is defined as the inner product of that series. Therefore, the energy density of IMF_n^* obtained from the decomposition of a min-max scaled series with N observations over time g is defined as, (25)

$$E_n = \frac{1}{N} \sum_{g=1}^{N} [IMF_n^*(g)]^2. \tag{2.19}$$

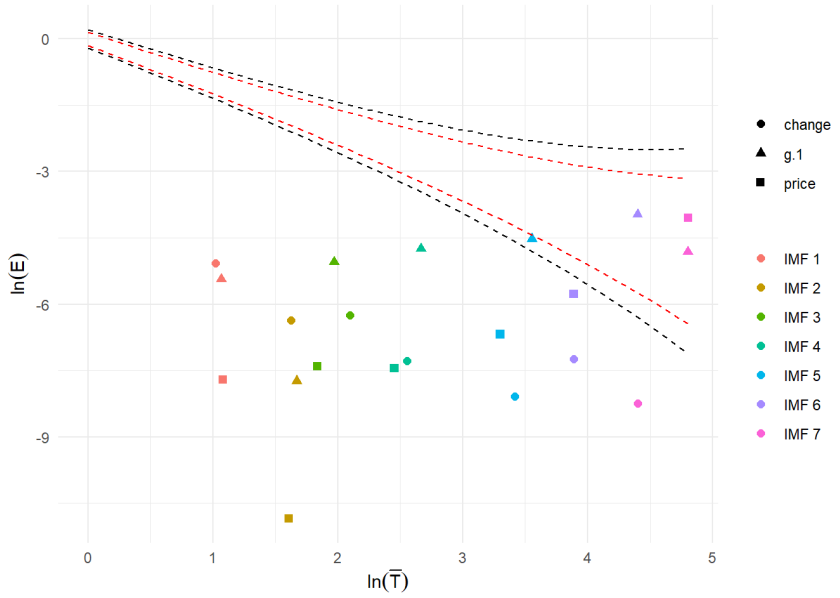


Figure 2.17: Wu IMF significance test for $g + 1$ (triangle), ULVR price close (square), and ULVR price change (circle). Black dashed spread lines represent the normal approximation for the 1% significance level, and the red dashed line represent the 5% level. IMFs 1-5 for each of the series are significant at the 5% level.

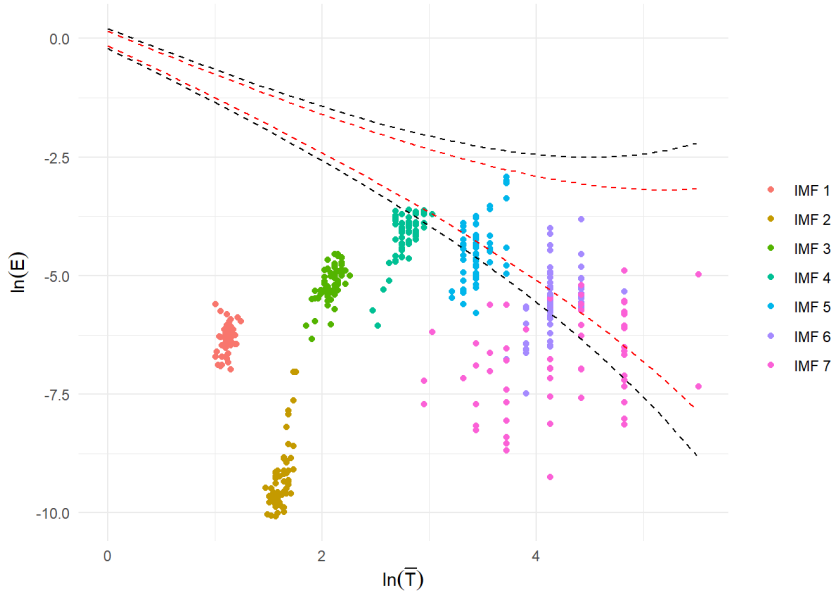


Figure 2.18: Wu IMF significance test for all $g + h$ expectations series for the merged ULVR data set. We can see that the first four IMFs for every series are significant at the 5% level.

Applying this equation, and taking logarithms of the energy densities and mean periods, allows us to carry out the Wu significance test for our example series (Figure 2.17). We can clearly see that for $g + 1$, price close and price change the first five IMFs are all significant at the 5% level. In fact, all of the IMFs for price change are significant at the 1% level. We can also see that for the $g + 1$ expectations, IMF_5 is not significant at the 1% level and borders the 5% lower spread line. By this test, both IMF_6 and IMF_7 of the $g + 1$ series are found to not be statistically different from the sixth and seventh IMFs obtained from white noise. Applying these observations to the IMF time series in Figure 2.9 suggests that, despite capturing the structural shape of the original series, the low-frequency IMFs could have come from white noise. The significance of the high-frequency IMFs indicates that the amplitude and frequency structures observed in their time series are not random noise, but are in fact statistically meaningful fluctuations of expectations and price over short time scales.

Figure 2.18 shows the results of applying the Wu test to each $g + h$ series, capturing the information significance of the IMFs for the whole merged ULVR data set. This shows that the findings for the $g + 1$ expectations are consistent for the first four IMFs which are all significant at the 5% level. However, there is evidence that for some time steps h the low-frequency IMFs do in fact contain information that is statistically different from white noise. Therefore, we cannot definitively conclude that any of the decomposed IMFs are purely white noise.

Chapter 3

Results

Here, we present the results obtained from applying the HHT to all the expectation forecasts with two hundred or more observations obtained from the merging process laid out in Section 2.1.2. Choosing only this subset of the data ensures that the results presented here are similar to the examples presented in Section 2.3. Therefore, each series used in the results was decomposed into seven IMFs and a residual term, and four announcement dates are included in each generation period. However, we must stress that the HHT can be applied to any length of time series, but we have chosen to limit this investigation to the longer series for practical reasons.

There are thirty-two sets of merged forecasts with two hundred or more observations taken from 22 different FTSE 100 companies. We applied the HHT to obtain IMFs and analytical signals for each of the $g + h$ series contained in each merged forecast set. We obtained the corresponding stock price time series using Refinitiv Workspace (1) in the same way as described in Section 2.3.

In this section, we first present the results from applying the Wu significance test to all of the merged forecasts, followed by the instantaneous frequency plots obtained from the analytical signal. Finally, we report cross-correlation results.

3.1 Wu Significance Test

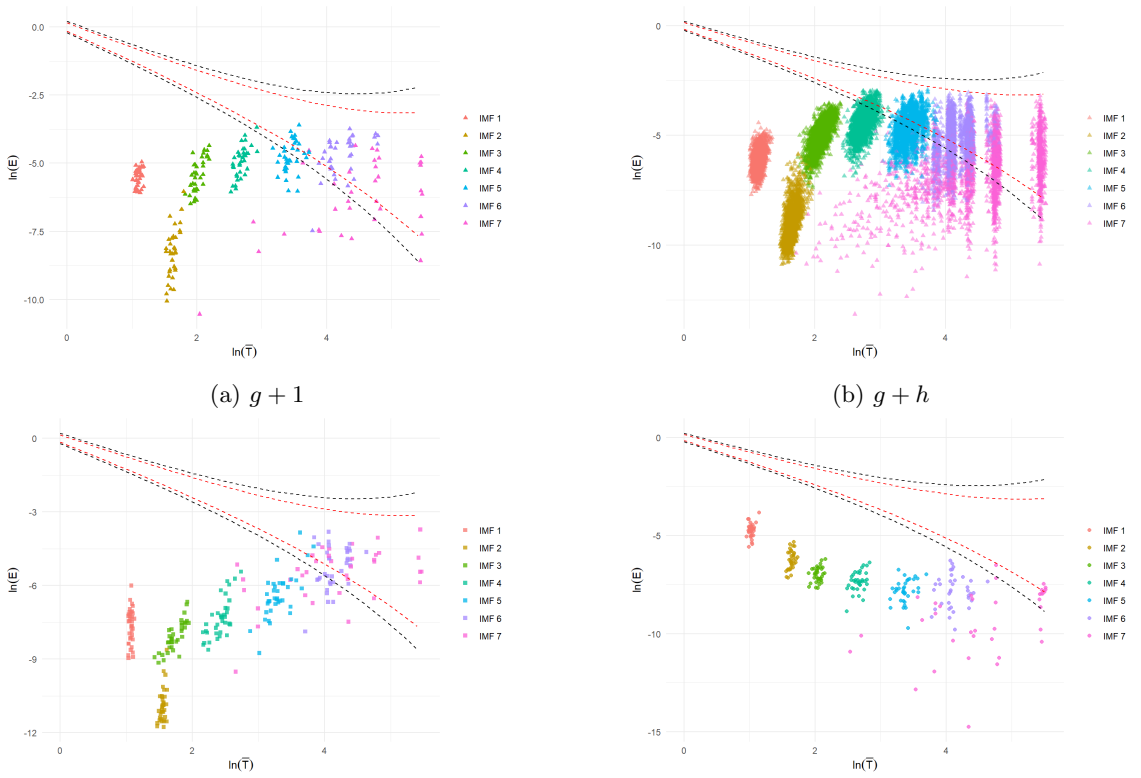
3.1.1 Results

In Figure 3.1, we report the results of four different applications of the Wu significance test: (1) For all $g + 1$ expectations, (2) all $g + h$ expectations, (3) all stock price series, (4) all stock price change series. These results are similar to the ULVR examples in Section 2.3.6. The price IMFs (3.1c), the $g + 1$ IMFs (3.1a), and the $g + h$ IMFs (3.1b) all exhibit the same relative patterns in the distribution of log energy density across log mean period. For each of these series, we see that IMF_2 has a much lower energy density than the other IMFs. Almost all of the first five IMFs for price close are significant at the 1% level, whilst for the expectation series there is some evidence that the fourth and fifth IMFs may not be significantly different from white noise at the 5% level. There are also multiple examples of the sixth and seventh IMFs plotted between the spread lines, suggesting that they do not contain information that is statistically different from a white noise decomposition. These results imply that for expectation and price time series over the generation period, the first five IMFs contain statistically significant information about the time series' dynamics on timescales between three and thirty-five days.

The Wu significance test for price change suggests a different energy density distribution with respect to the mean oscillatory period of the IMFs compared to both price and expectations. From Figure 3.1d we can see that almost all of the price change IMFs are significant at the 5% confidence level, with only a few of the seventh IMFs not being significant at the 1% level. But more importantly, the energy density of each IMF appears to decrease as the mean period increases, whereas for price and expectations energy density seems to remain constant or slightly increase as the mean period increases (with the obvious exception of the second IMF that has a much lower energy density than the other IMFs).

3.1.2 Discussion

We observe large variability in the periodicity of the seventh IMF of the expectation and price series (pink triangles or squares). This could result from the method used to count maxima for calculating the mean period over counting and resulting in misleading mean period values. However, we also



(a) $g + 1$

(b) $g + h$

(c) Price close

(d) Price change.

Figure 3.1: Wu significance test results for the thirty-two merged forecasts. The black and red dashed spread lines represent the 1% and 5% confidence-limit levels respectively.

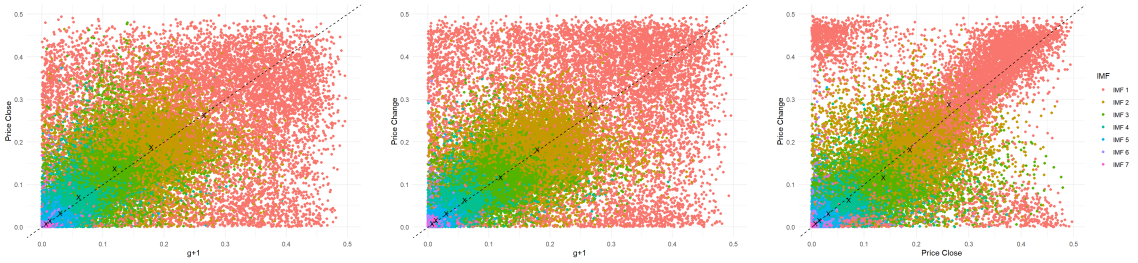
observe some trends in the outlier values of IMF_7. The observations with shorter mean periods also appear to have a lower energy density than the other IMFs at the same periodicity as the outliers are usually plotted below the other clusters of IMF observations. The energy density for these outliers is comparable with the energy densities observed for IMF_2, which are also outliers when compared with the energy densities of the other IMFs. Comparing the plot of all $g + 1$ expectations (Figure 3.1a) with both the example plot of ULVR $g + h$ expectations 2.18) and the plot of all $g + h$ expectations (Figure 3.1b), does not give us any reason to believe that the unusually significant IMF_7 occur for only one particular time step h, as there is a large increase in the number of seemingly significant IMF_7 when all expectation forecasts are plotted together. Although, the number of seemingly significant IMF_7 appear to be similar between the $g + 1$ expectations and the outliers observed for the stock price decomposition (Figure 3.1c). A more thorough exploration of these outliers was not possible given the time constraints of this project, but such an investigation may yield an improved implementation of the Wu test to the expectation data, uncover problems with the EMD method, or find that these observations are not outliers and do contain useful information about the original series' dynamics.

The decreasing relationship between energy density and mean period observed for the price change IMFs (Figure 3.1d) is much more similar than the distributions of the stock price and expectation IMFs to results reported by other applications of the Wu test. The results for the decomposition of time series of the Southern Oscillation Index (a tropical Pacific sea-level pressure index), (25) and for monthly sunspot data, (13) find that the energy density of each IMF is much greater, and the IMFs are all distributed along or above the upper 1% spread line. To the best of our knowledge, the Wu significance test has not been performed on stock price data decomposed via EMD in an academic publication — therefore, we have been unable to validate whether the energy density distributions obtained in this project are the expected energy dynamics of stock price time series, or if we have made a mistake in our application of the testing methodology laid out in Section 2.3.6. Assuming that the results reported here are valid, we may conclude that in general, the IMFs obtained from stock price and expectation forecast data have statistically lower energy densities with respect to their periodicity than the energy densities we would expect from the decomposition of white noise. As energy density is the mean of the squared values of the series (Equation 2.19), this suggests that the IMFs obtained from the normalized series have both less

extreme and lower amplitudes than IMFs obtained from the decomposition of white noise or other physical series. Further exploration of these observations is beyond the scope of this project and is left for future investigations into the application of the HHT on financial time series data.

3.2 Spectral Analysis Results

3.2.1 Instantaneous Frequency Scatterplots



(a) $g+1$ versus price close. (b) $g+1$ versus price change. (c) Price close versus price change.

Figure 3.2: Paired instantaneous frequency plots for the thirty-two long merged forecasts. The IMF frequency pairs are differentiated by colour.

In Figure 3.2 we present the results of plotting all pairs of instantaneous frequencies of each IMF for the same generation day g across all thirty-two long merged forecast generation periods. As with the ULVR example, we have three plots showing the instantaneous frequency relationships of the $g+1$ expectations with stock price and change in stock price and the frequency relationship between price and change. There is still visual evidence of a monotonic relationship between the instantaneous frequencies of the IMFs — but this is no surprise as, by definition, the first IMF has a higher frequency than the second and so on. Compared with the ULVR example frequency plots (Figure 2.14), the clusters of frequencies for each IMF seem to exhibit more spread around the reference line of identical instantaneous frequency. The price versus price change frequency plot (Figure 3.2c) shows a much closer relationship than either plot containing the frequency profile of the $g+1$ series. A notable feature is the cluster of IMF_1 instantaneous frequency pairs in the top left corner that is not present on the plots of frequency pairs containing $g+1$ expectations. This visual difference suggests that the spectral dynamics between price and price change differ from their dynamics with the corresponding $g+1$ expectation series.

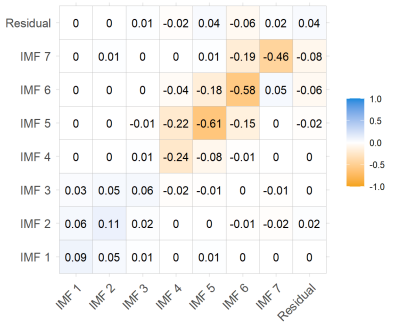
3.3 IMF Cross-Correlation Results

Following the method laid out in Section 2.3.5 we calculated both the Pearson cross-correlation (Equation 2.11) and the TDLCC (Equation 2.15) for the $g+1$ expectation series from each of the thirty-two long merged expectation forecasts, and the corresponding stock price and price change series.

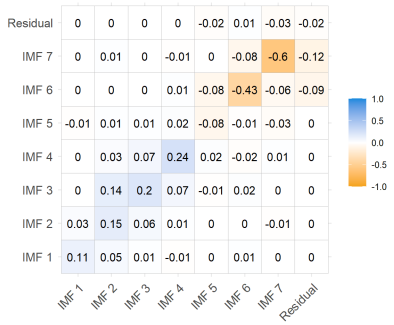
3.3.1 Pearson Cross-correlation

In Figure 3.3, we present colour maps of the median correlation coefficients for each pair of IMFs between the $g+1$ expectations and the corresponding stock price and price change series for the thirty-two long merged forecasts. Compared to the example IMF cross-correlations in Figure 2.15, there is even less evidence of correlations between different IMFs. As expected, the largest correlations are observed on the diagonal. Therefore, in Figure 3.4, we present the distributions of the correlation coefficients taken from the diagonals of the correlation matrices for each pair of series. From these histograms, we find that the unexpectedly low median correlation coefficient for the residuals resulted from a bimodal distribution where the residual trend correlation is either strongly positive or strongly negative. This suggests that the trend relationship between expectations and stock price is specific to each set of forecasts. Therefore, general conclusions about the relationship between the co-movement of expectation and price trends cannot be made.

The histograms confirm that the median correlation values represent the distributions of the IMF correlations relatively well. Therefore, we can conclude that when considered across a long

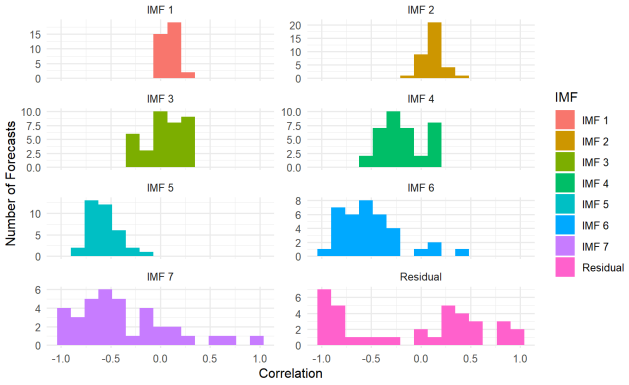


(a) $g + 1$ and price close.

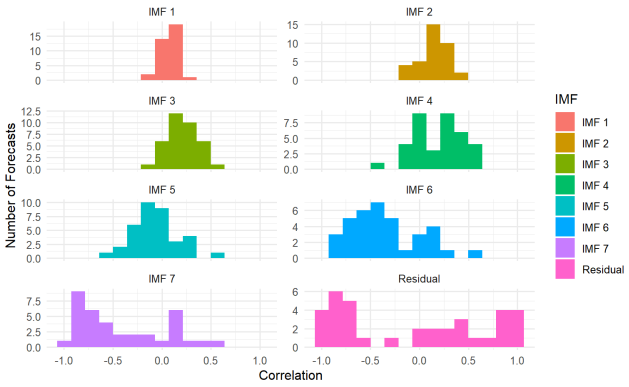


(b) $g + 1$ and price change.

Figure 3.3: Median of the Pearson correlation coefficients between the $g + 1$ series and the stock price, and price change, series calculated for each of the thirty-two merged forecasts.



(a) $g + 1$ and price close.



(b) $g + 1$ and price change.

Figure 3.4: Histograms of Pearson correlation coefficients separated by IMF pair calculated using Equation 2.11. These values were obtained from the diagonals of the correlation matrices for each of the thirty-two merged forecasts.

generation period, there is a small positive correlation between $g + 1$ and stock price for the first three IMFs, and a weak to moderate negative correlation for the last four IMFs. For price change, there is a weak positive correlation for the first four IMFs and a weak to moderate negative correlation for the last three IMFs. This suggests that there is only a weak relationship between expectations and the corresponding stock price and price changes, even at different time scales. However, as this correlation measure is calculated over the whole series, it may be artificially low due to changes in the correlation dynamics within the generation periods.

3.3.2 Time-dependent Lagged Cross-correlation

In their paper, Nava et al. (17) presented the results of their TDLCC between the S&P 500, IPC and VIX indices as colour maps of the sample median of the cross-correlation matrices obtained across 184 trading days. They investigated intraday correlations between the indices, so they were able to take cross-correlations over a fixed observation period between 09:40 and 16:00 for each day. In this project, we investigated the interday correlations over multiple different relative generation periods. This meant that for each merged forecast, the exact number of observations was different, and the relative location within the forecast of announcement dates was also different. Furthermore, as the lags, λ, used to calculate the cross-correlations were dependent on the mean periods specific to the IMFs being compared, they too would vary between different merged forecasts. Due to the brief nature of this project, we have not had the time to develop and implement a satisfactory adaptation of the method used by Nava et al. for summarising our TDLCC results obtained for each merged forecast. Therefore, in this section, we only present and discuss interim results for three arbitrarily chosen merged forecasts (Figure 3.5). We have also included the TDLCC colour maps obtained for all thirty-two merged forecasts in Appendix B so that the reader may consider other examples.

TDLCC Results and Discussion

The three randomly chosen examples are the merged series for Lloyds Banking Group (November 2021- October 2022), Barclays PLC (November 2019 - October 2020), and Intertek Group (September 2020 - July 2021). Despite both being financial firms, there is no evidence of similarity between the correlations of Lloyds and Barclays. Intertek Group is an assurance and product testing company — so it occupies a very different market sector than either of the banks, or Unilever. This small selection of examples, and those contained in Appendix B, illustrate the aforementioned difficulties associated with summarising the TDLCC coefficients as the colour maps for the IMFs have quite different lengths due to the varying fixed window sizes specific to each IMF series used to calculate their correlation.

The interim results presented here support the inferences drawn for the Unilver example analysed in Section 2.3.5. There is little evidence of correlation between the first IMFs of $g + h$ and stock price or price change. However, for the other IMFs, there is further evidence of persistent correlation structures at different lags. These persistent correlations appear to have sudden breaks and inversions (quickly switching from strongly negative to strongly positive), occurring close to an internal announcement date. These changes in the correlation structures within the series explain the low median global correlation coefficients of the diagonals of the correlation matrices in Figure 3.3 as the strong negative and positive local correlations are averaged out to a much lower whole-series Pearson correlation.

This is particularly true for the residual trend correlations, as we can see for both the Lloyds and Barclays examples — the residual series exhibit distinct internal inversions in the direction of trend correlation between expectations, price, and price changes.

These TDLCC results suggest that persistent correlation structures across all of the IMFs can last over multiple announcement dates or even invert multiple times within the observation period between announcements. This is a notable benefit of considering the merged forecasts constructed during this project, as it prompts us to ask questions about how similar a given region of the observation period is to those that came before or after it — we may not have considered this if we only analysed one sixty-four day generation period in isolation.

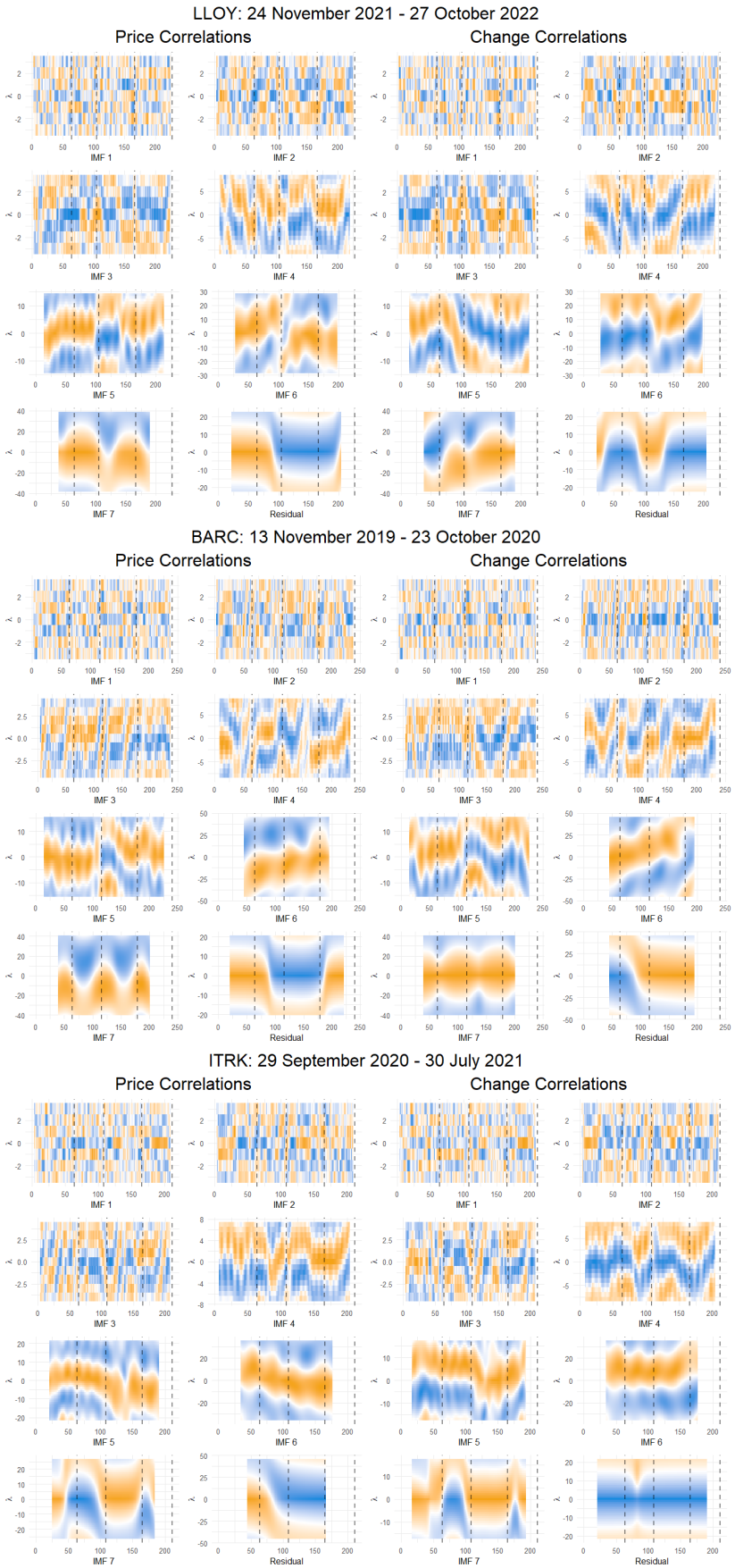


Figure 3.5: Randomly selected examples of TDLCC for Lloyds Banking Group (LLOY), Barclays PLC (BARC) and Intertek Group (ITRK). Legends have been omitted for plotting convenience, as in the other plots blue indicates a positive correlation and orange indicates a negative correlation.

Chapter 4

Discussions and Conclusions

4.1 Summary of Results

This project aimed to investigate the application of the Hilbert-Huang Transform to forecasts of investor expectations generated by Irithmics' machine learning model. By also applying the HHT to the corresponding stock price and stock price changes, we hoped to use the method to provide insight into the reflexivity of financial markets at different time scales. The primary method for investigating the relationship between investor expectations and stock prices was to use the TDLCC to visualise the local correlation structures between the time series at different time scales. There were four main advantages offered by the HHT method:

1. The first step of the HHT — Empirical Mode Decomposition — is a completely adaptive, data-driven algorithm that has been designed for handling nonlinear non-stationary data. This meant that it did not require any assumptions to be made about the expectation data. Using traditional time series analysis and modelling methods, such as Auto-Regressive Moving Average (ARMA) or GARCH models, would have required that we make distributional assumptions about the unusual expectation forecast data. (19)

2. The Intrinsic Mode Functions obtained from EMD capture information about changes in the original series at different time scales due to the different frequencies of each IMF. Furthermore, by definition, they were locally stationary, thus allowing us to investigate correlations between the series, and across different time scales — which is an expansion of the concept of correlations to a higher-dimensional level. (17)

3. The IMFs can be used to construct low and high-pass filters for the original series that are fully adaptive to the characteristics of the original time series (see Section 2.2.3). This allows for the visualisation of shorter and longer-term changes in expectations and price.

4. The second step of the HHT — applying the Hilbert Transform to the IMFs — returned a time-dependent analytical signal and allowed for the calculation of instantaneous frequencies (Equations 2.5 and 2.6). This analytical signal can be considered as a generalized version of the Fourier Transform of the original series (Equation 2.8) which does not require the series trend to be removed before decomposition, contains fewer basis functions, and does not require assumptions about linearity, stationarity, or choosing appropriate basis functions. [Barnhart and Eichenger]

The most promising insights uncovered by this project are directly related to the EMD method. Firstly, the Wu significance test suggests that most of the IMFs obtained by EMD contain information statistically significant from white noise (Figure 3.1). Secondly, even though they are only preliminary results, the TDLCC plots presented here (Figure 3.5) and in Appendix B, indicate that there are persistent local correlation structures between investor expectations, stock price and price change at different time scales — this opens up the possibility of further investigations into market reflexivity.

However, we cannot draw definitive conclusions from our results as we have not been able to perform traditional statistical tests. This was due to the limited time available for this project, the unusual data under consideration, and our choice to investigate the novel analysis method of HHT. Further investigation using the HHT is required before formal conclusions can be drawn.

4.1.1 Discussion of Contradictory Results

The results presented in this project seem to imply contradictory conclusions. First, consider the Wu significance tests. From these graphical tests, we might conclude that for the expectation series, the first four IMFs contain information different from a white noise signal and that there is substantial evidence that the last three IMFs are not statistically different from white noise (Figures 3.1a, 3.1b). However, when we consider the TDLCC plots (Figure 3.5, and the instantaneous frequency scatter plots (Figure 3.2), we find that for IMF_1 in particular, there is little evidence of a discernible correlation or frequency relationship between the high-frequency series. Instead, the TDLCC correlation results imply that strong local correlations exist between the low-frequency IMFs for the different series. If these low-frequency component series are statistically no different from the decomposition of white noise — as suggested by the Wu test — then intuitively, we would not expect to observe any correlation between the series.

Consulting the ULVR example IMFs in Figure 2.9 provides evidence for both contradictory conclusions. On the one hand, the low-frequency IMFs (4-7) are visually similar for large sections of the generation period, giving credence to the large local correlations observed in the TDLCC plots (Figure 2.16). But, as shown in Section 2.3.2, a linear combination of these low-frequency component series approximates the long-term trend of the original series. On the other hand, for the high-frequency IMFs, there are examples of high amplitudes and frequencies occurring in the proximity of announcement dates — which suggests that these IMFs are capturing the fluctuations in the original time series resulting from these announcements. Furthermore, as the low-frequency IMFs are visually similar to simple sinusoidal functions, which could explain the large positive and negative local correlations.

This apparent contradiction might indicate a flaw in our application of the EMD algorithm. In theory, the EMD algorithm only iterates to the next IMF once the IMF criteria are met and only stops decomposing a series into IMFs when the residual term is constant, monotonic, or has at most one maxima or minima. However, as explained in Section 2.2.2, in practice, stopping criteria are imposed to stop over sifting. In the `Rlibeemd` implementation of the CEEMDAN algorithm used in this paper, in addition to the stopping criteria, the number of IMFs to decompose the series into is specified. By default, this is set to the maximal number of IMFs, $\log_2 N$. (21)We chose to use this value as it would ensure that no oscillatory information was left in the residual term and that for series of similar length we would have the same number of IMFs for calculating cross-correlations. However, we believe this might have caused some series to be decomposed into more IMFs than necessary, resulting in over-sifting and converting otherwise meaningful IMFs into meaningless fluctuations with constant amplitude. (17)

To check this hypothesis, we would need to re-run our analysis using more strict stopping criteria to prevent over-sifting — but this may lead to high-frequency IMFs that do not fulfil the IMF criteria of zero envelope mean, which is a base assumption for the application of the cross-correlation equations. We could also conduct the analysis using fewer IMFs — however, some information would likely be left in the residual. Using the same stopping criteria but extracting fewer IMFs should leave the results for the high-frequency IMFs unchanged as they are extracted from the original series first, and so they are independent of the number of IMFs the series is decomposed into. Unfortunately, due to the time constraints of this project, undertaking this validation and tuning process was not possible and is left to a future investigation.

This project has only been an initial investigation into the applicability of HHT analysis for Irithmics' expectation forecasts. The results presented here are promising, and suggest that the use of the data-driven EMD algorithm can lead to valuable insights about changes in investor expectations, and the reflexive relationship between these expectations and the stock price of a company. In particular, we have found via the Wu significance tests that the first four component IMFs are statistically significant, suggesting those series may give insight into changes in expectations on a three to fourteen-day time scale. In combination with the interim results of the TDLCC analysis, we have some evidence to suggest that the correlation relationships between expectations, stock price and price change exhibit persistent local correlation structures that are obscured when the correlation is calculated for the whole observation period. It is unfortunate that our initial investigation is only able to offer qualitative, and predominantly graphical results. However, we have proven through this investigation that the HHT is a promising method for financial time series analysis, but that promise is not without its limitations.

4.2 Challenges Encountered in the Application of the HHT

As the HHT is a relatively novel methodology, the body of literature available to establish how to best apply the method to our data and interpret our results is sparse — the works pertaining to financial data even more so. This made obtaining and reporting results a challenging process for this project.

To further explain the challenge of delivering meaningful results from applying HHT to the expectation data consider the following brief example of variability — an initially promising line of inquiry that was dropped from this project due to challenges when applying it to the expectation data.

4.2.1 Variability

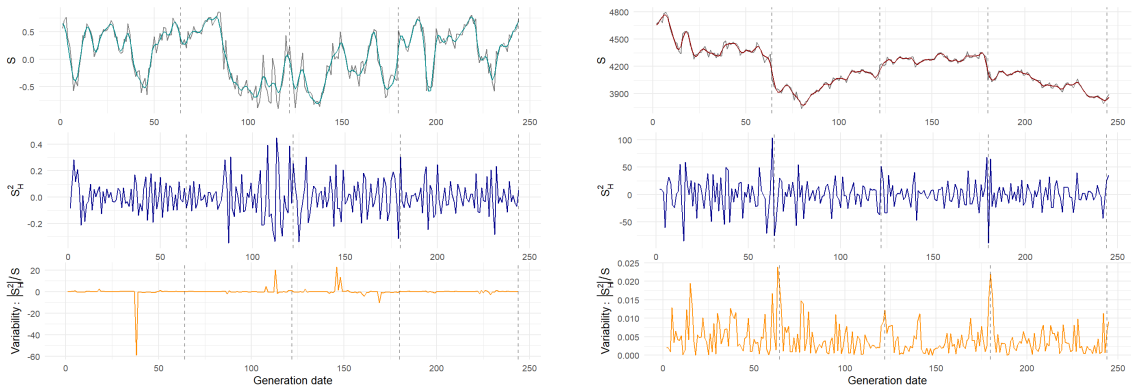
Huang et al., in one of the first known applications of the HHT to financial time series analysis, proposed a novel measure of time-dependent volatility — called variability — which is calculated using the IMFs obtained from EMD. They were motivated by the same idea that motivated the GARCH model - volatility should be a function of time. (16) As already discussed, the commonly used GARCH model accommodates time-dependent volatility by defining the conditional variance of a time series as a function of past shocks. (19) Variability takes advantage of the adaptive low and high pass filtering outlined in Section 2.2.3 to obtain an intrinsic time-dependent measure of the volatility of the original series.

Variability is defined as the ratio between the absolute value of the IMF component(s) and the original series at any time: (16)

$$V(t) = \frac{|S_H^k(t)|}{S(t)}, \tag{4.1}$$

where S_H^k is the high-pass filtered series up to IMF_k defined by Equation 2.10.

Therefore, the variability measure uses the idea shown in Section 2.3.2, that a low-pass filtered series approximates the local trend of the original signal, and the high-frequency IMFs excluded by the filter contain the short-term fluctuations around this trend. Figure 4.1 shows two examples of variability calculated using the first two IMFs (e.g. $k = 2$) for the ULVR $g+1$ expectations and the corresponding stock price example series used throughout Section 2.



(a) $g + 1$ (b) Price Close

Figure 4.1: Variability plots for ULVR example series. The top panel of each subfigure shows the original series, S, (grey) with the low-pass filtered series for $k = 2$, calculated using Equation 2.9, which describes the local trend of each series. The second panel shows the high-pass filtered series $S_H^2 = IMF_1 + IMF_2$. The third panels show plots of the variability calculated using Equation 4.1). The dashed lines are the announcement dates for the generation period.

The problem with applying this variability measure to the expectation data is immediately apparent. When applied to the price series in Figure 4.1b, the variability series is very similar to what we would expect the conditional volatility series obtained via a GARCH model to be. We are able to identify noticeable peaks in variability around the announcement dates — supporting observations made throughout this project about the relationship between these dates and the fluctuations in the IMFs. Furthermore, the magnitude of the variability, with values of between 0 and 0.025 (or alternatively, 2.5%) is also commensurate with the magnitude of traditional volatility

measures. This is not the case for the variability of $g + 1$ expectations. On the one hand, the expectations forecast series is bounded between $[-1, 1]$, and so, contains both negative values and values close, or possibly equal, to zero. On the other hand, the relative magnitude of S_H^2 to S is much greater than when variability is calculated for the price series. Taken together, these two facts can result in unreasonably large absolute values of variability as seen in Figure 4.1a.

During this project, we were unable to devise a satisfactory adaptation of the variability measure for the expectation data. Normalizing the data (using a method like the min-max normalisation technique used in 2.3.6) would solve the issue of the negative variability values, however, the large relative magnitude of S_H^2, and division by values close to zero, will remain issues for obtaining variabilities comparable to those from other financial series. Perhaps this approach to measuring the variability of a series is just unsuitable for divergent data. However, the simplicity and convenience of the price variability example for obtaining a volatility measure suggest that this approach has significant potential for financial time series analysis. The development of a suitable adaptation for the expectation series is worth further exploration.

4.2.2 Spectral Analysis

As mentioned in Section 2.3.4, we found that the Hilbert spectra produced from the instantaneous frequencies and amplitudes obtained through the HHT did not yield meaningful insights into the expectations or price forecasts. When applied elsewhere, the Hilbert spectrum is often compared with Fourier or Wavelet spectra derived for the same signal. (16,13) In these papers, the authors find that the Hilbert spectra have a higher frequency resolution due to the HHT representing the analytical signal in significantly fewer basis functions than the other decomposition methods based on the Fourier transform. The increased frequency resolution makes identifying noticeable features in the spectra much easier — but this is only a benefit when one knows what they are looking for. Due to the time constraints of this project, we did not have the opportunity to research and apply other decomposition methods to the Irithmics expectation data. To fully ascertain the potential of the HHT, further inquiry into the application of alternative decomposition methods and time series spectral analysis is required.

Therefore, within this project, we did not have the opportunity to make full use of the second step of the HHT methodology, and so, the spectral analysis results presented Section 3.2.1 are somewhat superficial when compared with our other results. It was to be expected from the definition of the IMFs that IMF_1 would have both the largest instantaneous frequency and the greatest variation in frequency ranges — it is unsurprising, then, that the corresponding pairs form an almost patternless scatter. The similar instantaneous frequencies between the other IMFs, which is the reason for the clustering around the reference line in Figures 2.14 and 3.2, is an observation that is confirmed by the similar mean periods for each IMF observed in the Wu Significance test results (Figure 3.1) and in the window sizes used for calculating the TDLCC in Section 3.3.2.

4.3 Improved Machine Learning using the HHT

This investigation has highlighted the fact that HHT is a powerful method for uncovering underlying relationships within and between time series that may otherwise be missed when considering the time series as a whole. But, this investigation has also found that summarising the insights gained from HHT analysis in a quantifiable manner is a challenging endeavour. However, other papers have found that the HHT is a valuable method for feature generation that significantly enhances the performance of traditional Machine Learning (ML) models. Leung and Zhao (7) found that using the HHT on stock market indices to obtain the analytical signal, instantaneous frequencies, and instantaneous amplitudes and using these as training inputs for generic random forest, support vector regression, and Long Short-Term Memory (LSTM) neural network models, led to a four to six times reduction in the mean-squared prediction error compared to the same models trained on the original series. Nava et al. (27) also used EMD with support vector regression models to forecast financial time series. In another, non-financial, example, Masselot et.al. (14) proposed the so-called EMD-regression to improve traditional regression models by first decomposing an input time series variable into IMFs and then fitting a Lasso regression model to the new variables. This approach acknowledges the complexities of real-world data by allowing relatively simple models to be fitted to non-stationary time series, and in doing so EMD-regression allows for the influence of all time scales on the response variable to be assessed where traditional regression models only depict the dominant timescale.

The insights about applying HHT to Irithmics's expectation data explored in this project, combined with the HHT-assisted ML approaches outlined here, is a natural next step for a future project. Such a combination may be able to realise the goal of quantifying the effect announcements have on investor expectations at different time scales or provide a model for the reflexive dynamics between a company's stock price and investor expectations.

4.4 Conclusion

This project offered an initial exploration into the application of the Hilbert-Huang Transform for analysing forecasts of investor expectations provided by Irithmics. Due to the unusual type of financial data, we chose to use the HHT as it is an adaptive method suitable for nonlinear non-stationary time series, and so did not require any distributional assumptions to be made about the expectation data. Before undertaking our analysis, we established that there is a qualitative difference between the expectation values within a forecast generated on a given day and the series of expectation values for a specific future trading day. We called this the forecast distance problem. Therefore, we introduced the concept of the $g+h$ expectation series obtained by 'reading along the diagonal.' We also identified that there was an overlap between many of the forecast generation periods, so we carried out data processing to form merged sets of expectation forecasts.

We selected thirty-two merged forecast sets containing two-hundred or more forecast series to explore the use of HHT. We performed EMD using the noise-assisted CEEMDAN modification of the algorithm to decompose the $g+h$ series for each merged forecast into the seven Intrinsic Mode Functions (IMFs) and a residual term. In the second step of HHT, we applied the Hilbert transform to obtain the analytical signal of the original series. We also applied the HHT to the corresponding stock price and stock price change time series to explore possible dynamic relationships between the $g+1$ investor expectations and stock price. To do this, we introduced the concept of the Time Dependent Lagged Cross-Correlation (TDLCC) which allowed us to quantify and visualise complex local correlation structures between the expectation and stock price series that are obscured when considering only the original time series. The TDLCC results suggest that corporate announcements can cause changes to the local correlation structures between expectations and stock price. By applying the Wu IMF Significance Test, we found that the first four IMFs of the $g+h$ series contained statistically significant information at the 5% significance level, which suggests that the original expectation and price series exhibit noticeable fluctuations on time scales varying between three days to three trading weeks.

Unfortunately, due to multiple factors, including the unusual data, the novel analysis method, the time constraints and the ambition of this project, we have not been able to present results from quantitative statistical tests. Therefore, this project can only be considered as a proof of concept and potential for the use of the HHT for the analysis of Irithmics' data. However, our initial results are promising, and we suggest that combining the HHT with Machine Learning methods might allow a future project to advance the initial steps made here and realise the full potential of the HHT for understanding the dynamics of investor expectations.

Bibliography

1. Datastream International and Refinitiv Workspace. FTSE100 Constituent Companies' Daily Price Close c.2020–2023. (Visited on 08/18/2023).
2. Keynes JM. The General Theory of Employment, Interest and Money. ISN ETH Zurich, 1936. URL: https://www.files.ethz.ch/isn/125515/1366_keynestheoryofemployment.pdf (visited on 08/18/2023).
3. Teall JL. Chapter 12 - Market Efficiency. In: *Financial Trading and Investing (Third Edition)*. Ed. by Teall JL. Third Edition. Academic Press, 2023:359–402. DOI: https://doi.org/10.1016/B978-0-323-90955-6.00012-4. URL: https://www.sciencedirect.com/science/article/pii/B9780323909556000124.
4. Simon HA. Bounded Rationality. In: *Utility and Probability*. Ed. by Eatwell J, Milgate M, and Newman P. London: Palgrave Macmillan UK, 1990:15–18. DOI: 10.1007/978-1-349-20568-4_5.
5. Two Sigma Client Solutions Team. Estimating Global Investor Views with Reverse Optimization. 2022. URL: https://www.twosigma.com/articles/estimating-global-investor-views-with-reverse-optimization/ (visited on 08/18/2023).
6. Soros G. Soros: General Theory of Reflexivity. 2009. URL: https://www.ft.com/content/0ca06172-bfe9-11de-aed2-00144feab49a (visited on 08/18/2023).
7. Leung T and Zhao T. Financial time series analysis and forecasting with Hilber-Huang transform feature generation and machine learning. *Applied Stochastic Models in Business and Industry*. 2021;37:993–1016.
8. Irithmics. About. URL: https://www.irithmics.com/about/.
9. Baccigalupi A and Liccardo A. The Huang Hilbert Transform for evaluating instantaneous frequency evolution of transient signals in non-linear systems. *Measurement*. 2016;86:1–13.
10. Huang NE, Shen Z, Long SR, et al. The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. *Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*. 1998;454:903–995.
11. Bowman DC and Lees JM. The Hilbert-Huang Transform: A High Resolution Spectral Method for Nonlinear and Nonstationary Time Series. *Seismological Research Letters*. 2013;84:1074–1080.
12. Vecchio A, Laurenza M, Meduri DG, Carbone V, and Storini M. The Dynamics of the Solar Magnetic Field: Polarity Reversals, Burtterfly Diagram, and the Quasi-Biennial Oscillations. *The Astrophysical Journal*. 2010;749:1–10.
13. Barnhart BL and Eichinger WE. Analysis of Sunspot Variability Using the Hilbert-Huang Transform. *Solar Physics*. 2011;269:439–449.
14. Masselot P, Chebana F, Bélanger D, et al. EMD-regression for modelling multi-scale relationships, and application to weather-related cardiovascular mortality. *Science of The Total Environment*. 2018;612:1018–1029.
15. Silverman BW, Vassilicos JC, and Ramsey JB. The contribution of wavelets to the analysis of economic and financial data. *Philosophical Transactions of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*. 1999;357:2593–2606.
16. Huang NE, Wu ML, Qu W, Long SR, and Shen SSP. Applications of Hilbert-Huang transform to non-stationary financial time series analysis. *Applied Stochastic Models in Business and Industry*. 2003;19:245–268.
17. Nava N, Matteo TD, and Aste T. Dynamic correlations at different time-scales with empirical mode decomposition. *Physica A: Statistical Mechanics and its Applications*. 2018;502:534–544.

18. Chen X, Wu Z, and Huang NE. The Time-Dependent Intrinsic Correlation Based on the Empirical Mode Decomposition. *Advances in Adaptive Data Analysis*. 2010;2:233–265.

19. Harris R and Sollis R. Applied Time Series Modelling and Forecasting. Wiley, 2003.

20. Engle RF. Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation. *Econometrica*. 1982;50:987–1007.

21. Luukko P, Helske J, and Räsänen E. Introducing libeemd: a program package for performing the ensemble empirical mode decomposition. *Computational Statistics*. 2016;31:545–557.

22. Yeh JR, Shieh JS, and Huang NE. Complementary Ensemble Empirical Mode Decomposition: A Novel Noise Enhanced Data Analysis Method. *Advances in Adaptive Data Analysis*. 2010;02:135–156.

23. Torres ME, Colominas MA, Schlotthauer G, and Flandrin P. A complete ensemble empirical mode decomposition with adaptive noise. In: *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2011:4144–4147. DOI: 10.1109/ICASSP.2011.5947265.

25. Wu Z and Huang NE. A study of the characteristics of white noise using the empirical mode decomposition method. *Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*. 2004;460:1597–1611.

26. Ogasawara E, Martinez LC, Oliveira D de, Zimbrão G, Pappa GL, and Mattoso M. Adaptive Normalization: A novel data normalization approach for non-stationary time series. In: *The 2010 International Joint Conference on Neural Networks (IJCNN)*. 2010:1–8. DOI: 10.1109/IJCNN.2010.5596746.

27. Nava N, Di Matteo T, and Aste T. Financial Time Series Forecasting Using Empirical Mode Decomposition and Support Vector Regression. *Risks*. 2018;6.

28. Erdiş A, Bakir MA, and Jaiteh MI. A method for detection of Mode-Mixing problem. *Journal of Applied Statistics*. 2021;48:2847–2863.

R Packages Used

11. Bowman DC and Lees JM. The Hilbert-Huang Transform: A High Resolution Spectral Method for Nonlinear and Nonstationary Time Series. *Seismological Research Letters*. 2013;84:1074–1080.

21. Luukko P, Helske J, and Räsänen E. Introducing libeemd: a program package for performing the ensemble empirical mode decomposition. *Computational Statistics*. 2016;31:545–557.

24. Kim D and Oh HS. EMD: A Package for Empirical Mode Decomposition and Hilbert Spectrum. *The R Journal*. 2009;1:40–46.

29. Wickham H, Averick M, Bryan J, et al. Welcome to the tidyverse. *Journal of Open Source Software*. 2019;4:1686.

30. Ooms J. The jsonlite Package: A Practical and Consistent Mapping Between JSON Data and R Objects. *arXiv:1403.2805 [stat.CO]*. 2014.

31. Auguie B. gridExtra: Miscellaneous Functions for "Grid" Graphics. R package version 2.0.0. 2015. URL: http://CRAN.R-project.org/package=gridExtra.

32. Rudis B. hrbrthemes: Additional Themes, Theme Components and Utilities for 'ggplot2'. R package version 3.4.0. 2020. URL: https://github.com/hrbrmstr/hrbrthemes.

33. Daniel F, Ooi H, Calaway R, Microsoft Corporation, and Weston S. foreach. R package version 2.5.0. 2022. URL: https://github.com/RevolutionAnalytics/foreach.

34. Kassambara A and Pati I. ggcorrplot. 2022. URL: http://www.sthda.com/english/wiki/ggcorrplot-visualization-of-a-correlation-matrix-using-ggplot2.

35. Hyndman R, Athanasopoulos G, Bergmeir C, et al. forecast: Forecasting functions for time series and linear models. R package version 8.21. 2023. URL: https://pkg.robjhyndman.com/forecast/.

36. Hyndman RJ and Khandakar Y. Automatic time series forecasting: the forecast package for R. *Journal of Statistical Software*. 2008;26:1–22.

37. Kassambara A. ggpubr: 'ggplot2' Based Publication Ready Plots. R package version 0.6.0. 2023. URL: https://rpkgs.datanovia.com/ggpubr/.

38. Firke S, Denney B, Haid C, Knight R, Grosser M, and Zadra J. janitor: Simple Tools for Examining and Cleaning Dirty Data. R package version 2.2.0. 2023. URL: https://github.com/sfirke/janitor.

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

40. Daniel F, Microsoft Corporation, Weston S, and Tenenbaum D. doParallel: Foreach Parallel Adaptor for the 'parallel' Package. 2022. URL: https://github.com/RevolutionAnalytics/doparallel.

Appendix A

R Code Used

We include the R code and data used to produce this report in the supplementary material folder uploaded alongside this document. Here, we explain how to navigate this folder. The information contained in this Appendix is also contained inside the `README.txt` file. The most important folder to consult if one is curious about how the plots and analysis contained wihtin this report were produced is the "markdown_html_files" folder, which contains the knitted Rmarkdown files used to make this report.

- data_folder:
  - Contains all the raw JSON data files supplied by Irithmics. Each file contains a set of forecasts and is identified by its London Stock Exchange ticker, "ZZZZ.XLON", and the announcement date that the file is centred on. This announcement date is also the last forecast generated within that file.
- data_processing_and_merging:
  - data_merging_functions.R:
    - `stitich_function` — merges two data files by matching generation dates and deleting the overlap from the first file.
    - `file_merge` — a function that uses the `stitich_function` to perform all viable merges between a list of dataframes.
    - `ticker_merge` — a function that uses `file_merge` to perform all possible merges for a specified company.
  - data_merging_script.R — performs all possible merges and outputs "merged_raw_data.rds".
  - ceemdan_unscaled_scaled.R — a script that applies the `CEEMDAN_function` to every merged data set using socketed parallelised foreach loops. This ensured that all other parts of our report were reproducible and saved time when producing plots. Produces both normalised scaled and un-normalised IMFs for use in other .R or .Rmd files.
- functions:
  - CEEMDAN_function.R — function to perform CEEMDAN with specified parameters and output a data frame that is used across the other files. This increased reproducibility.
  - data_merging_functions.R — duplicate of above
  - significance_test_finctions.R:
    - `min_max` — performs min-max normalisation.
    - `mean_period_func` — calculates the mean period by counting maxima and dividing the length of the series by the number of maxima.
    - `energy_density_func` — calcualtes the energy density by taking the sum of squares of an IMF dataframe.
    - `confidence_intervals_func` — calculates the Wu test spread lines using the normal approximation for a given confidence-limit level.
  - step_ahead_expectation_functions.R:

    * `expectation_time_extract` — converts a normal data file into a reading along the diagonal data file.
    * `hm_plot_exp_space` — plots the diagonal data as a colour map.
    * `add_announcemnts_diag` — adds the diagonal announcements dates to a transformed colour map

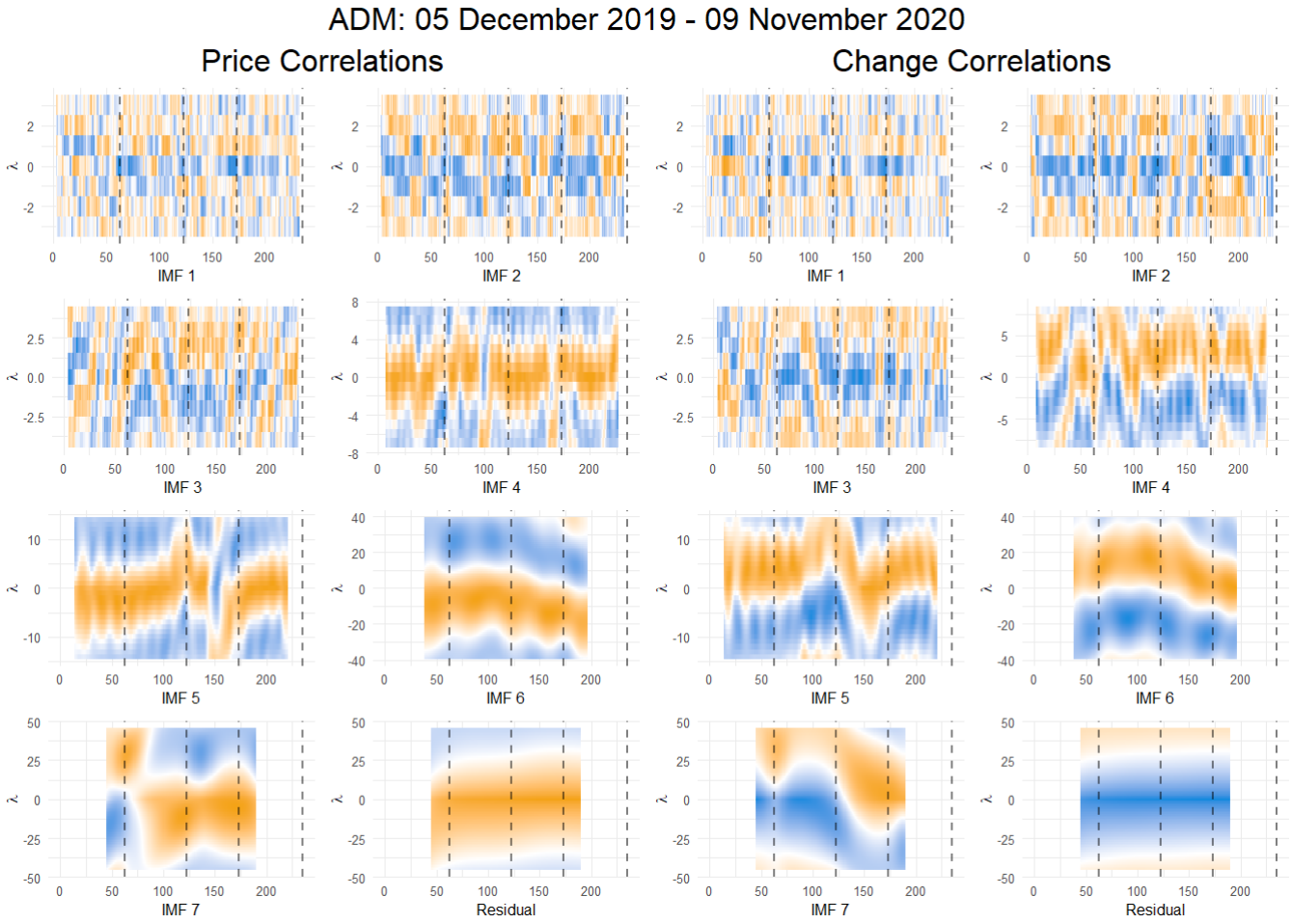
- markdown_html_files:
  - Contains the HTML files produced by the markdown documents used to generate plots and results for this report. It also contains a document describing the process of preparing the price data using Refinitiv Workspace.
- price_data_prep:
  - Price_Data_Preparation.Rmd — markdown that prepares and extracts the dates required for acquiring price data using Refinitiv.
  - long_series_date_range.csv — file input to in Refinitiv Workspace to get prices for the date range of each company.
- Refinitiv_data:
  - This folder contains the excel documents used with Refinitiv Workspace to get price data. It also contains the .rds file produced by Price_Data_Preparation.Rmd to store the price data.
- The rest of the folder is made up of the Rmarkdown files used to produce the HTML documents in the markdown_html_files folder. This is also the project's main directory, and so contains almost all the .Rds objects produced and stored by different scripts contained in the above folders.

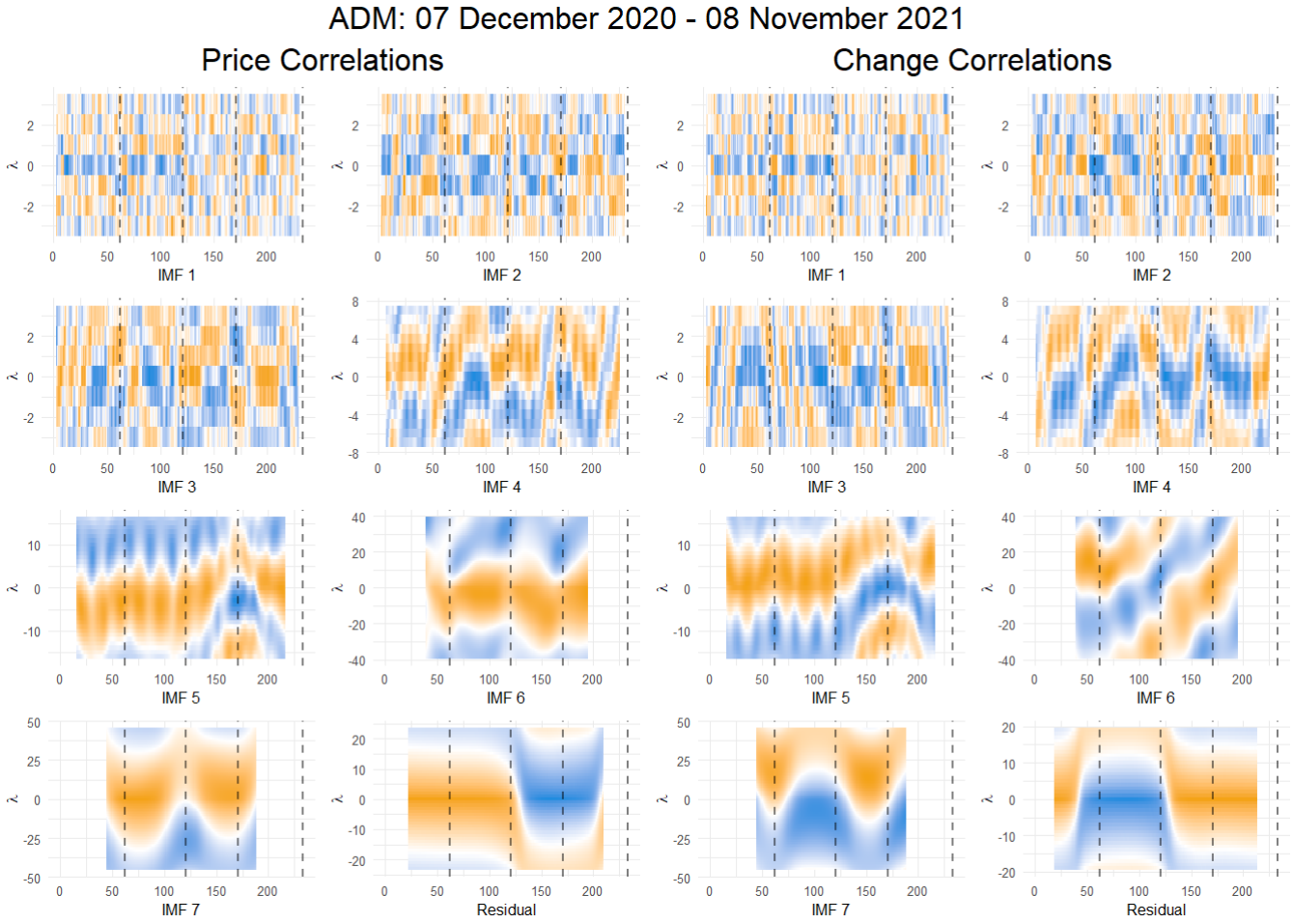
Appendix B

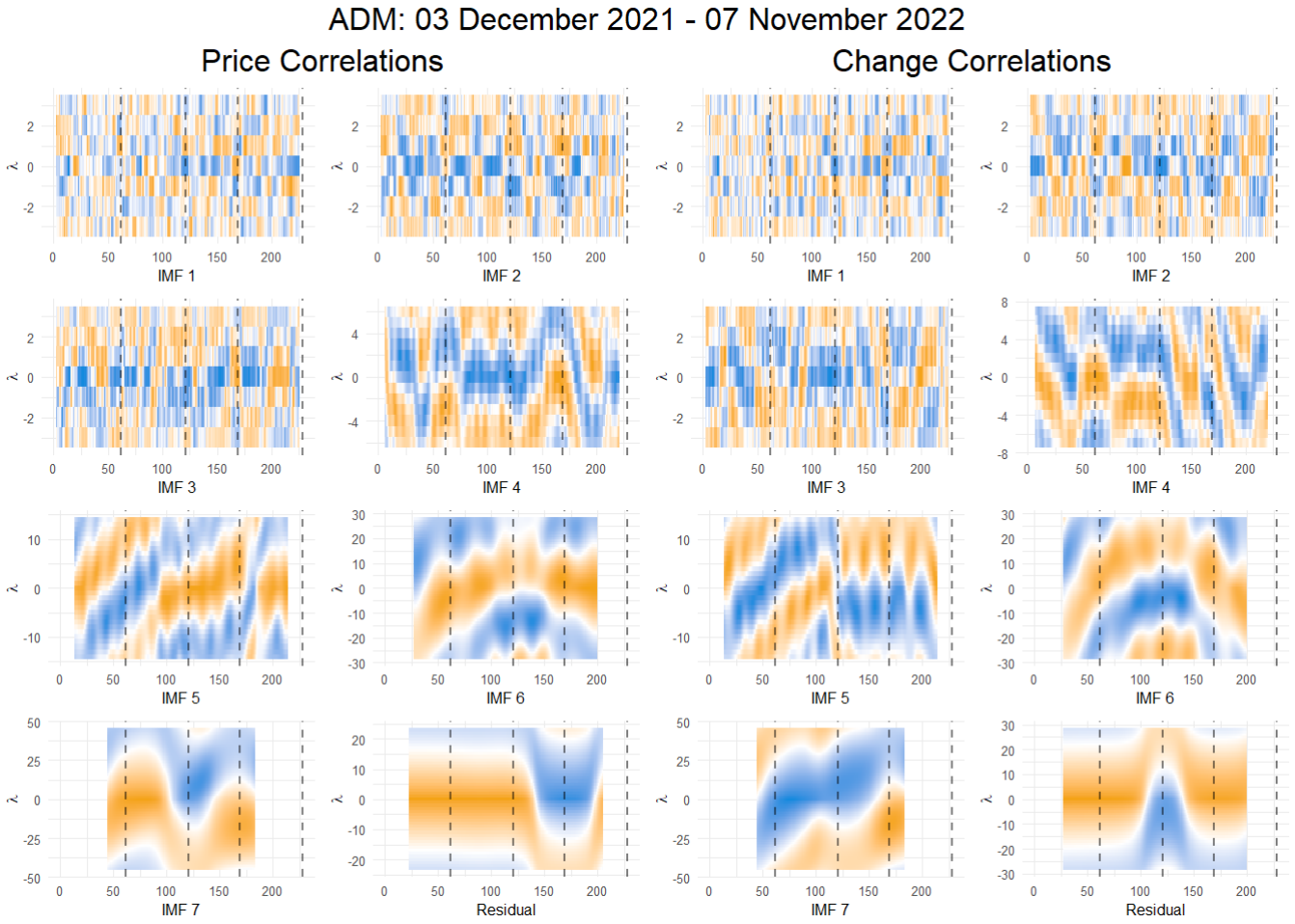
Supplementary Time Dependent Lagged Cross-Correlation Results

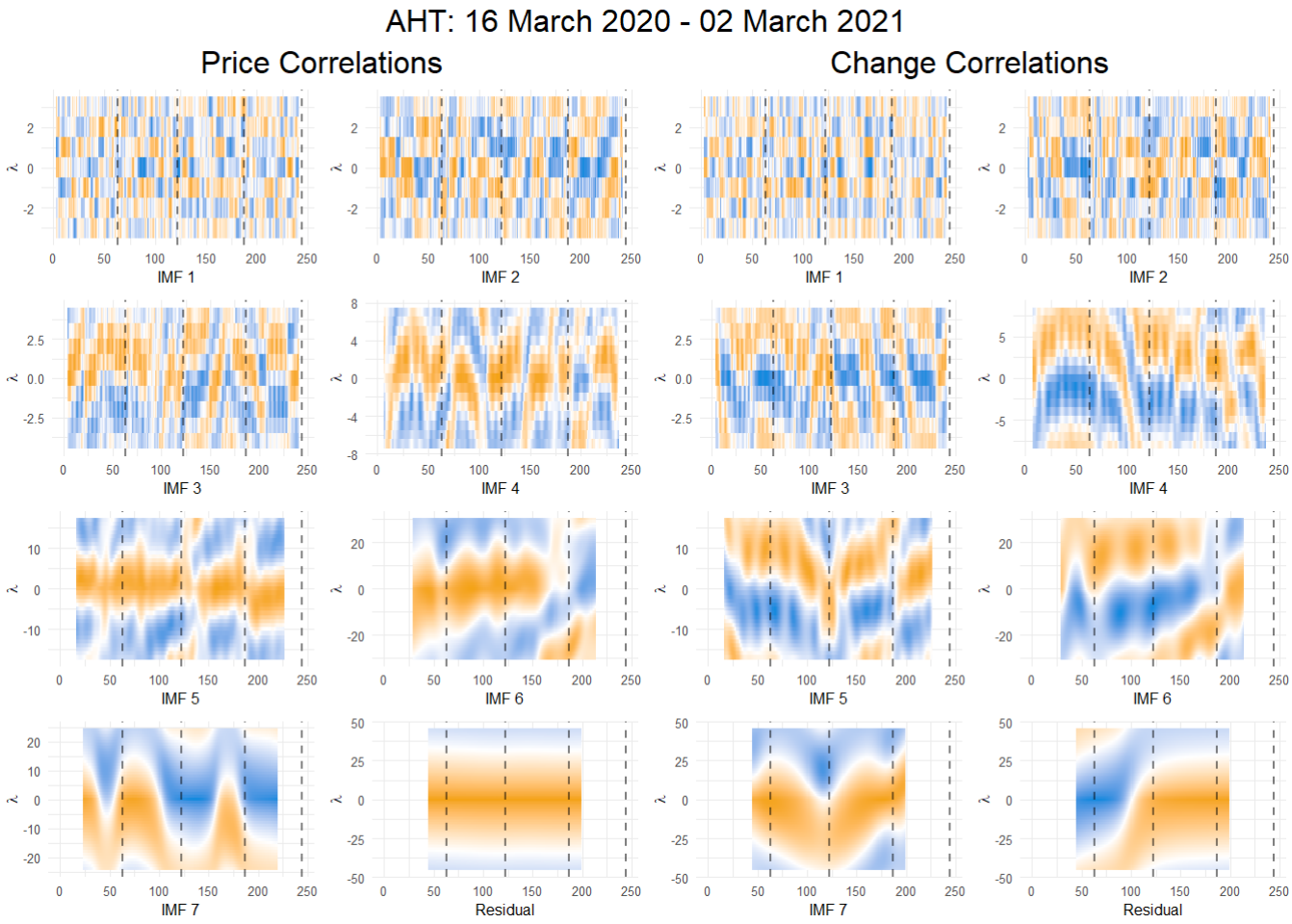
This appendix contains all TDLCC colour maps for the thirty-two merged forecast data sets so that the reader of this project may consult further examples of the correlation measure proposed in this project. Each forecast is identified by the company ticker and the generation date range for the forecast. The correlations between $g + 1$ expectations price are printed on the left, and correlations between $g + 1$ expectations and price change are printed on the right.

The code used to produce these colour maps can be found in Appendix A as a Rmarkdown script, or as an HTML file.


ADM: 05 December 2019 - 09 November 2020
Price Correlations
Change Correlations
IMF 1
IMF 2
IMF 3
IMF 4
IMF 5
IMF 6
IMF 7
Residual

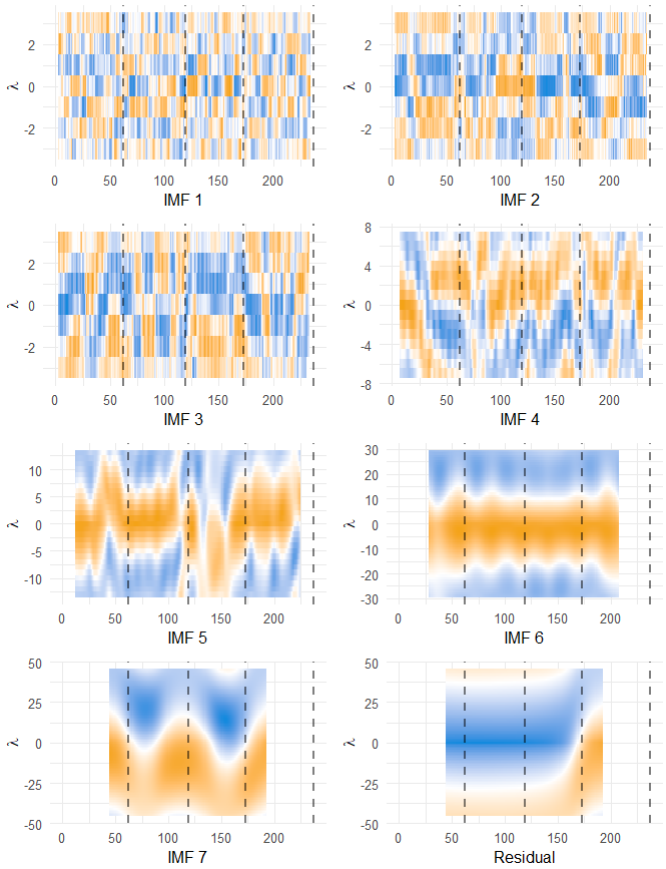

ADM: 07 December 2020 - 08 November 2021
Price Correlations
Change Correlations
IMF 1
IMF 2
IMF 3
IMF 4
IMF 5
IMF 6
IMF 7
Residual


ADM: 03 December 2021 - 07 November 2022
Price Correlations
Change Correlations
IMF 1
IMF 2
IMF 3
IMF 4
IMF 5
IMF 6
IMF 7
Residual
λ

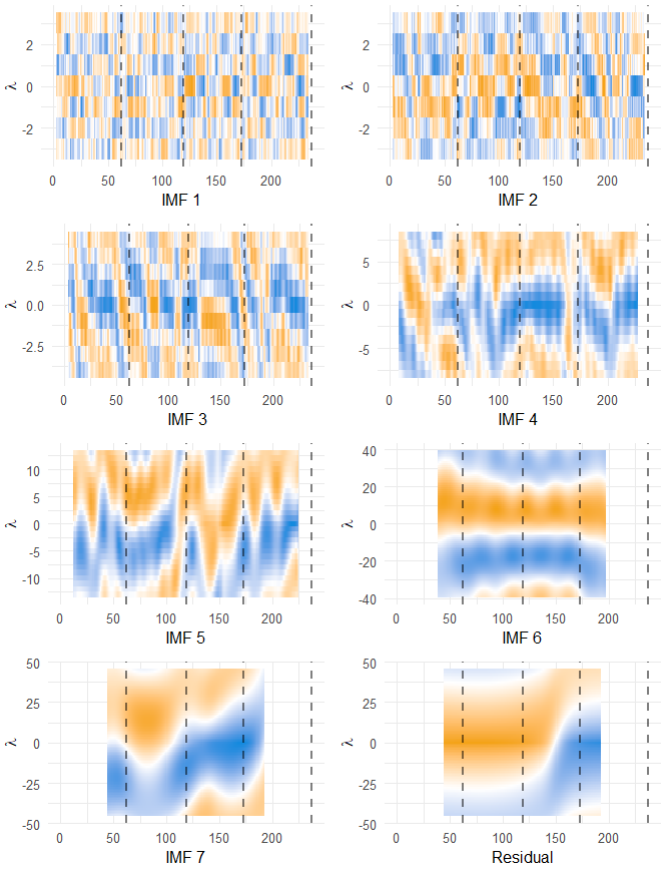

AHT: 16 March 2020 - 02 March 2021
Price Correlations
Change Correlations
IMF 1
IMF 2
IMF 3
IMF 4
IMF 5
IMF 6
IMF 7
Residual
λ

AV: 04 December 2020 - 11 November 2021

Price Correlations

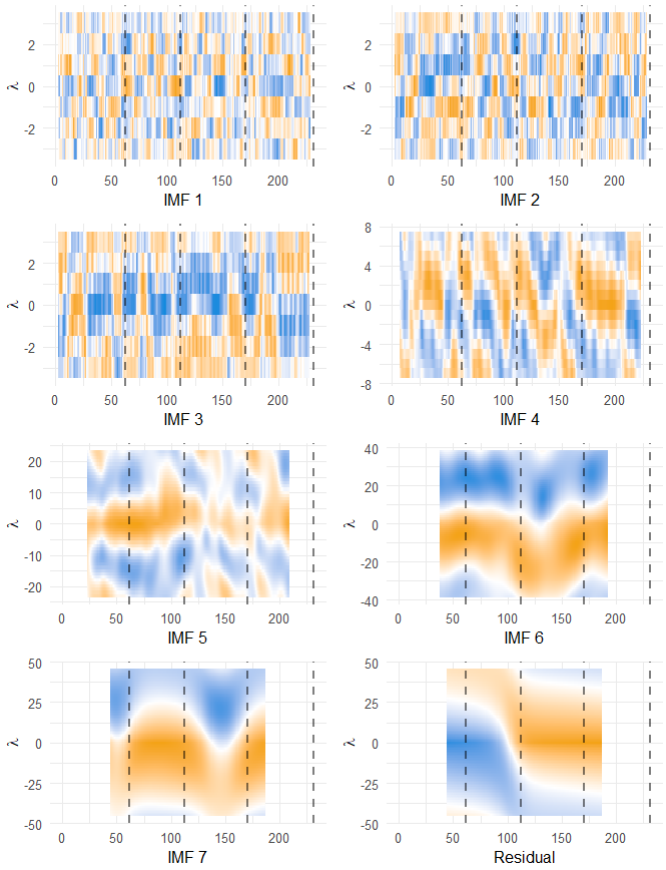


Change Correlations

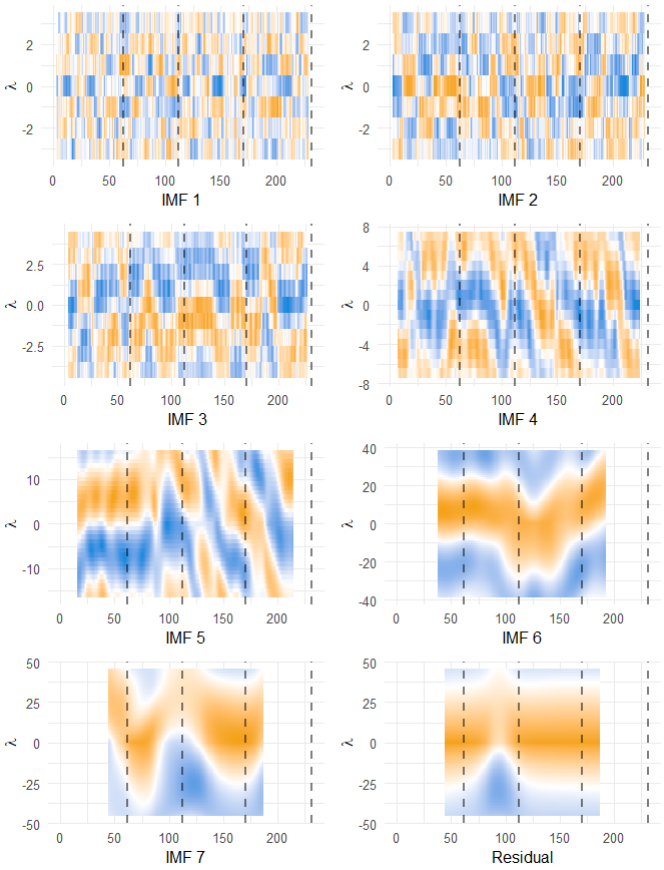


AV: 02 December 2021 - 09 November 2022

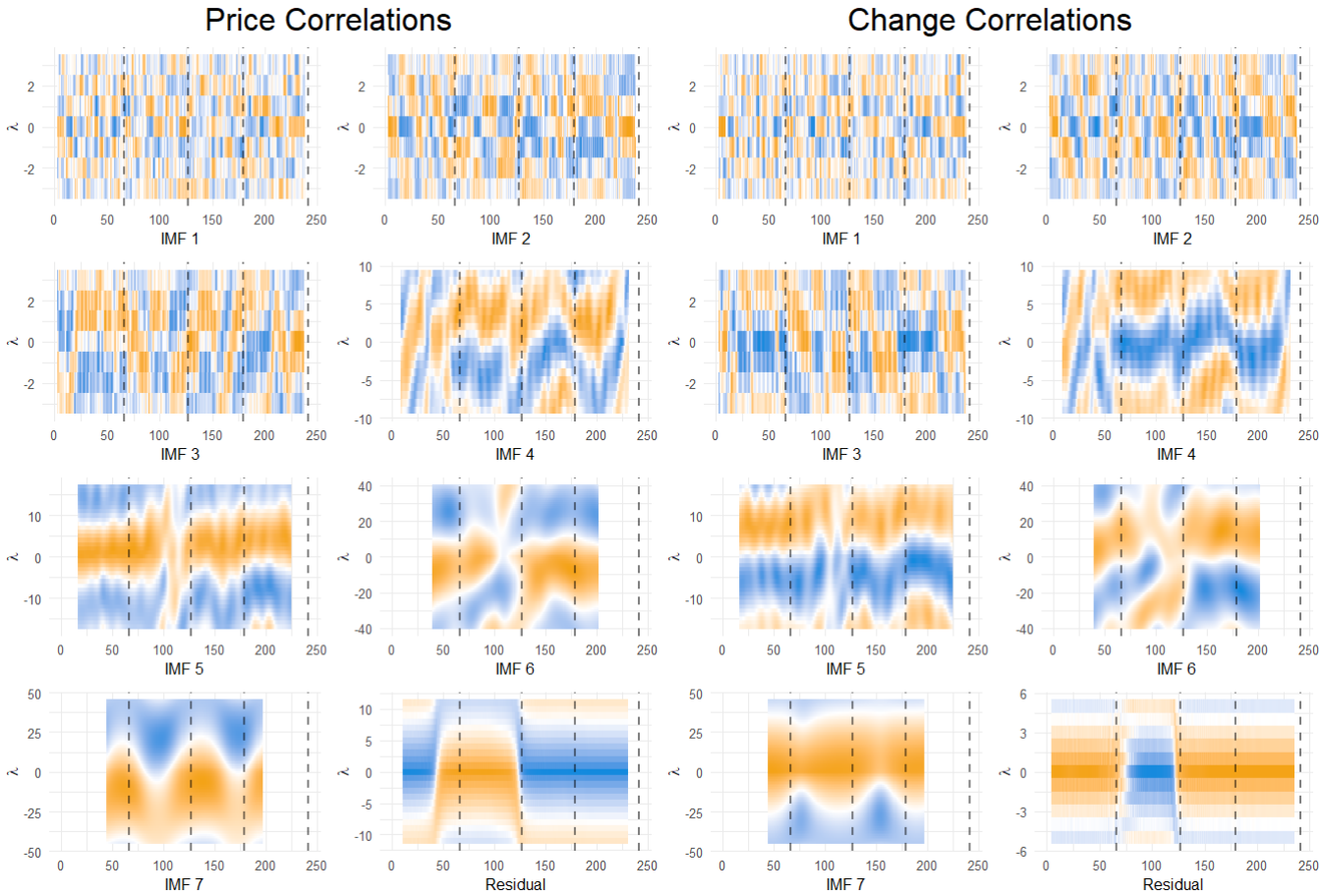
Price Correlations



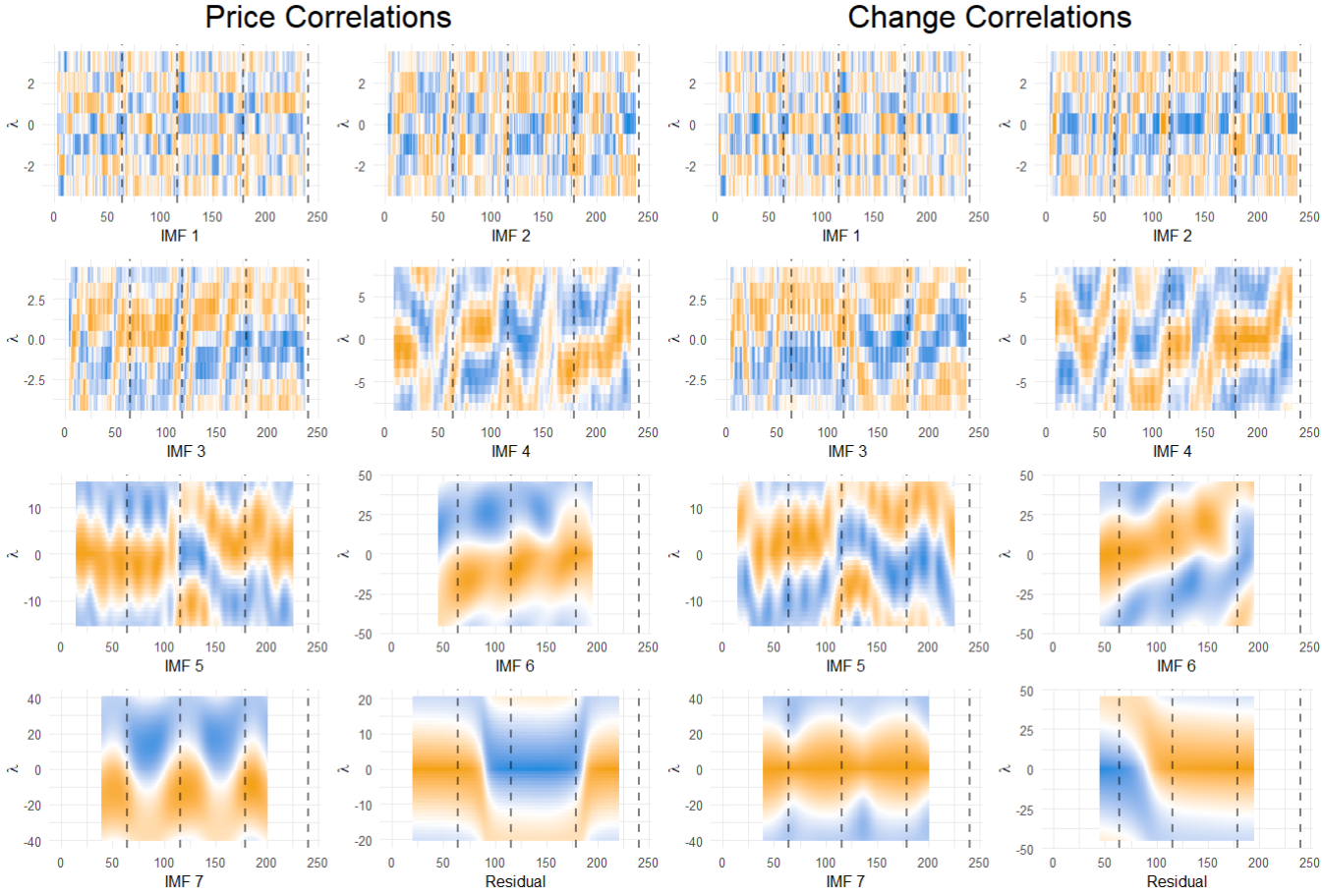
Change Correlations



AZN: 12 August 2021 - 29 July 2022

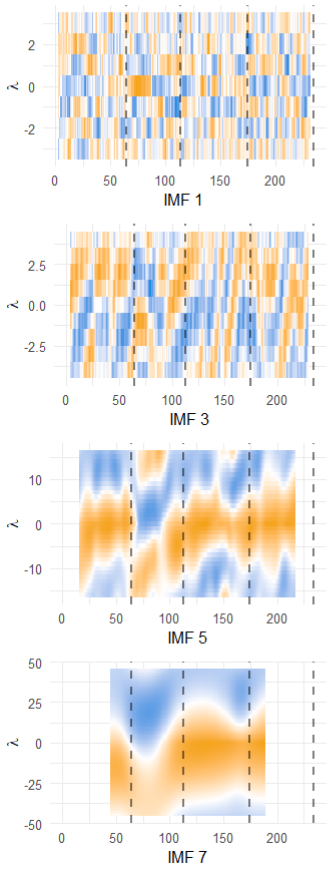


BARC: 13 November 2019 - 23 October 2020

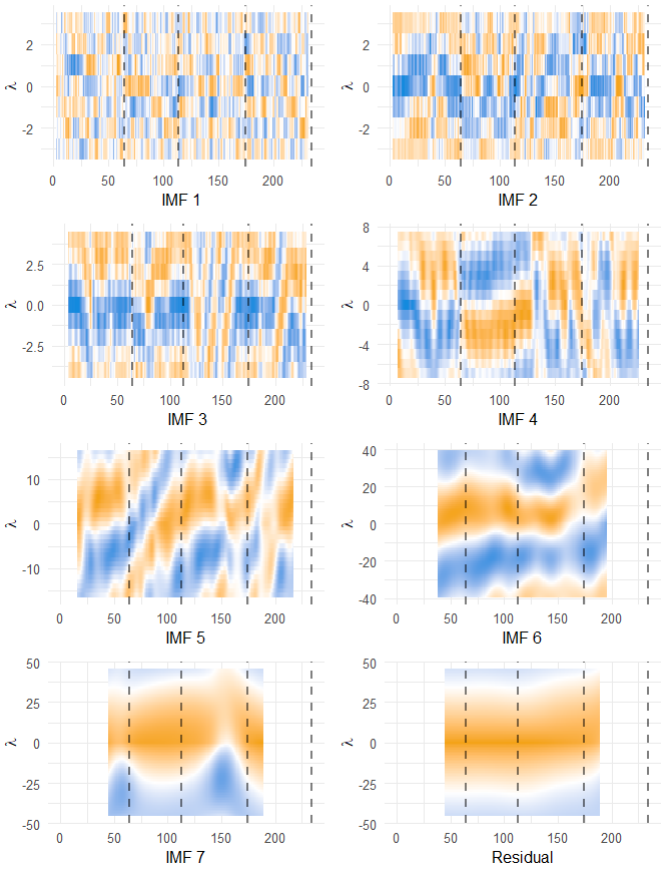


BARC: 18 November 2020 - 21 October 2021

Price Correlations

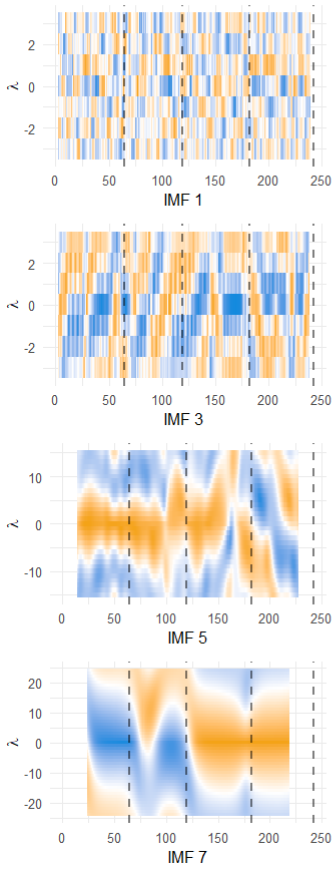


Change Correlations

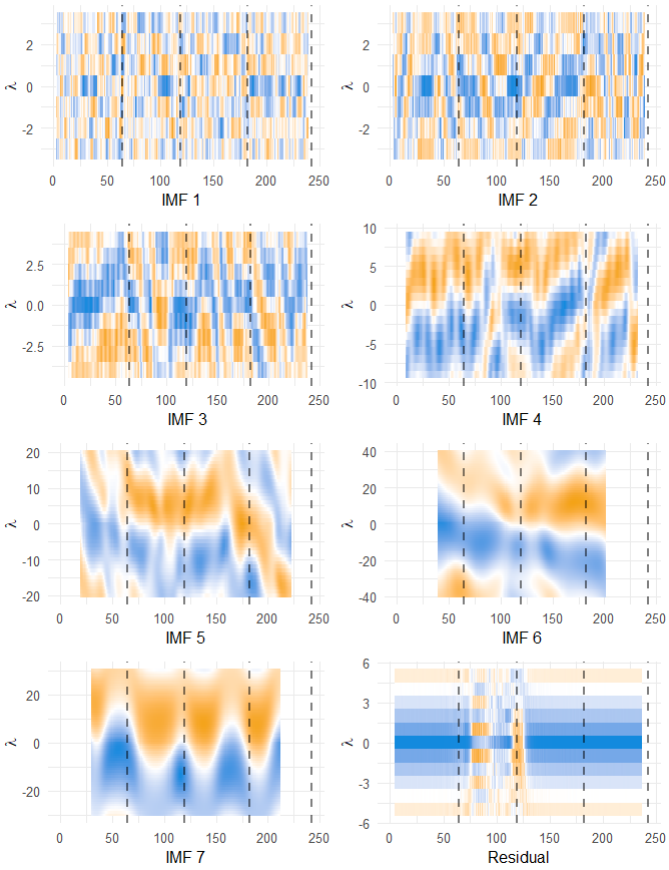


BP: 08 November 2021 - 01 November 2022

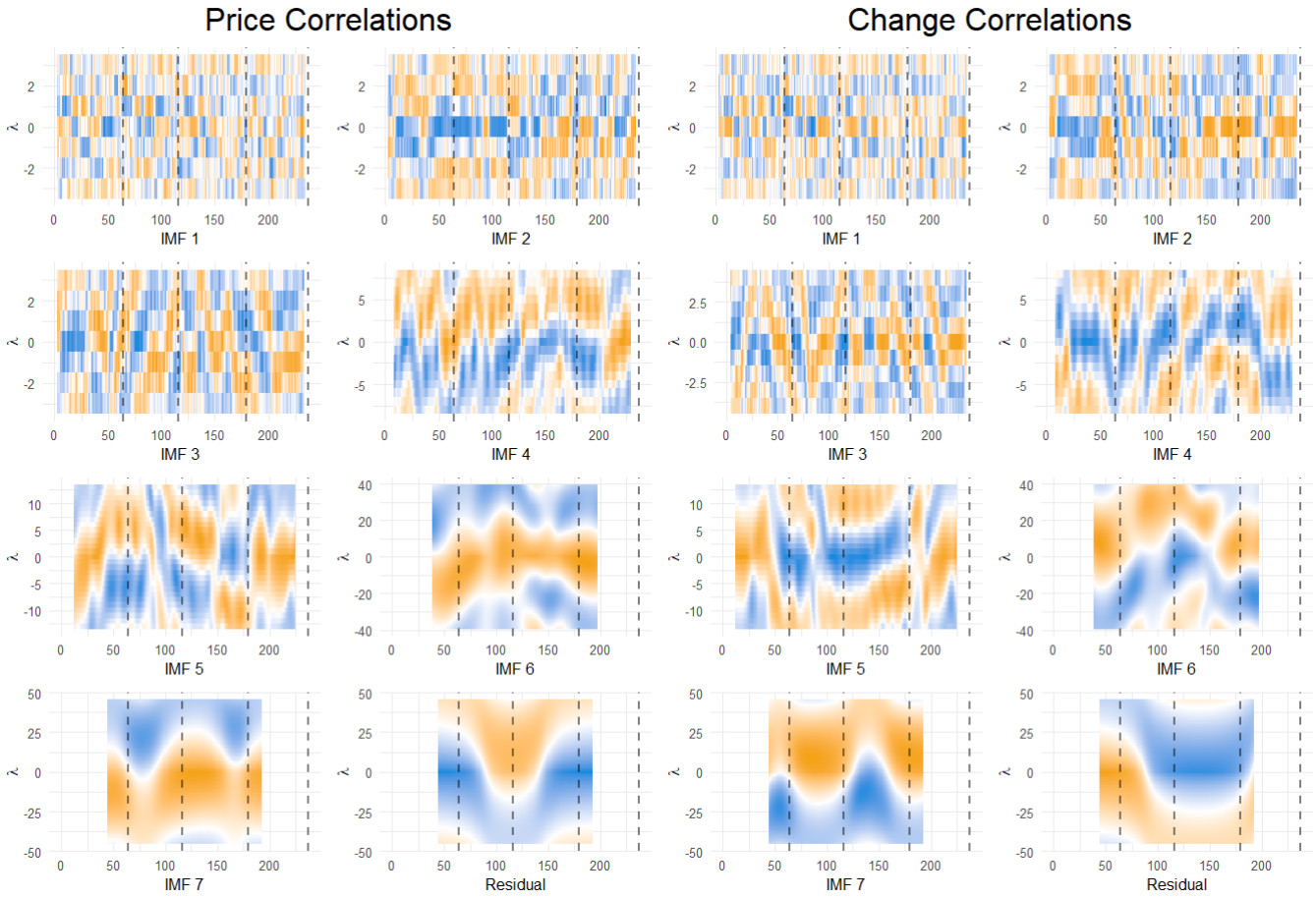
Price Correlations



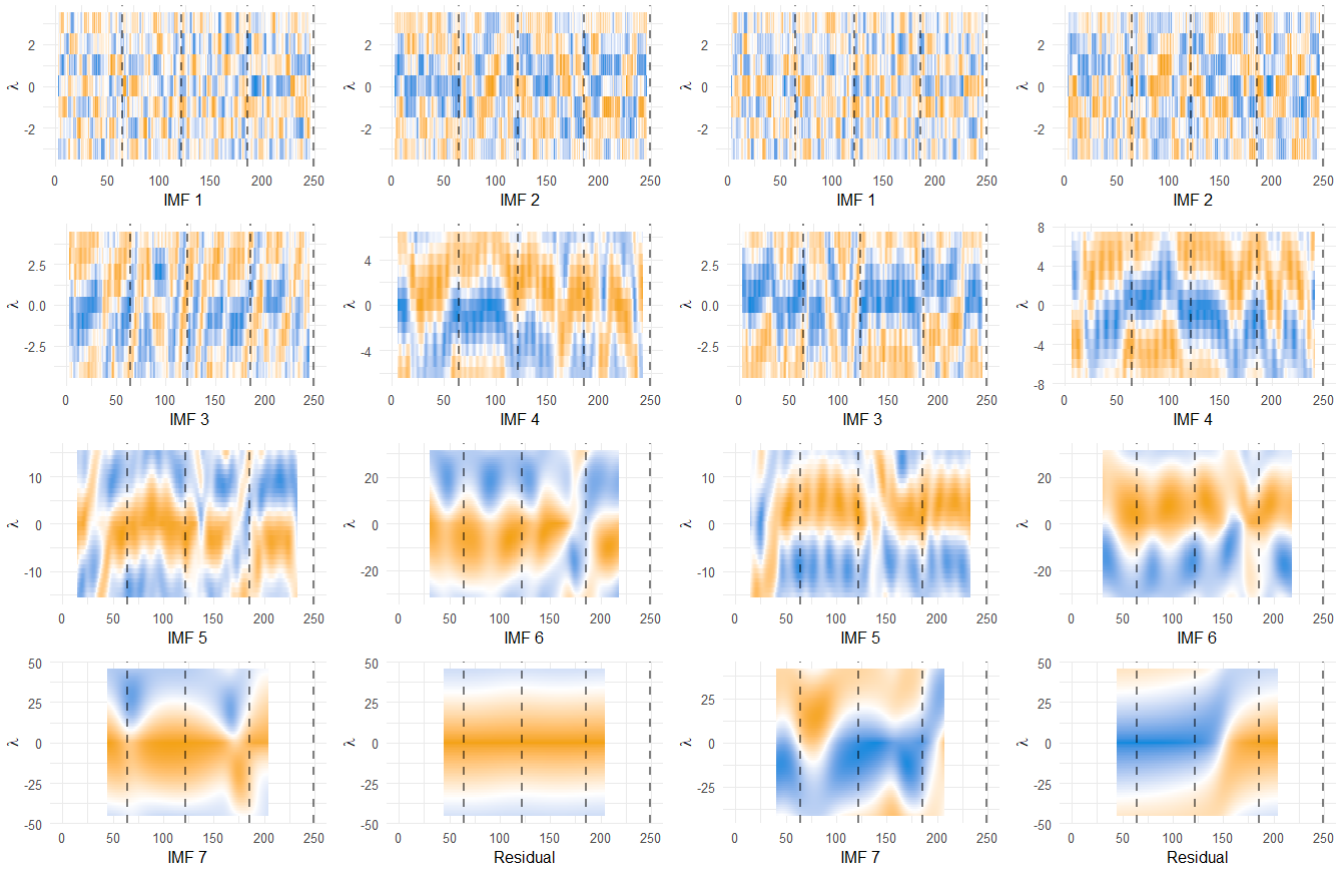
Change Correlations



CCH: 22 November 2021 - 08 November 2022



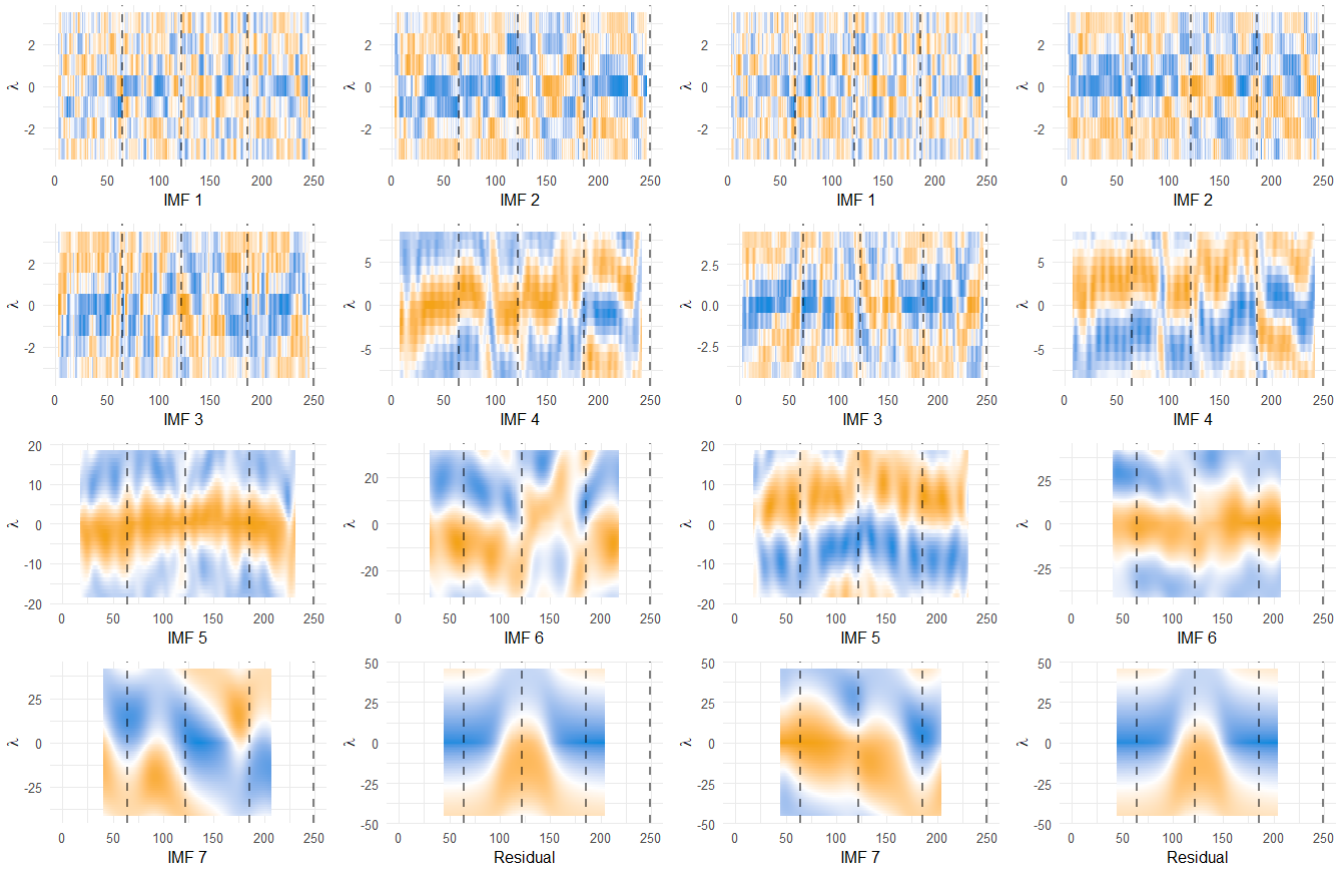
GSK: 05 November 2019 - 28 October 2020



GSK: 03 November 2020 - 27 October 2021

Price Correlations

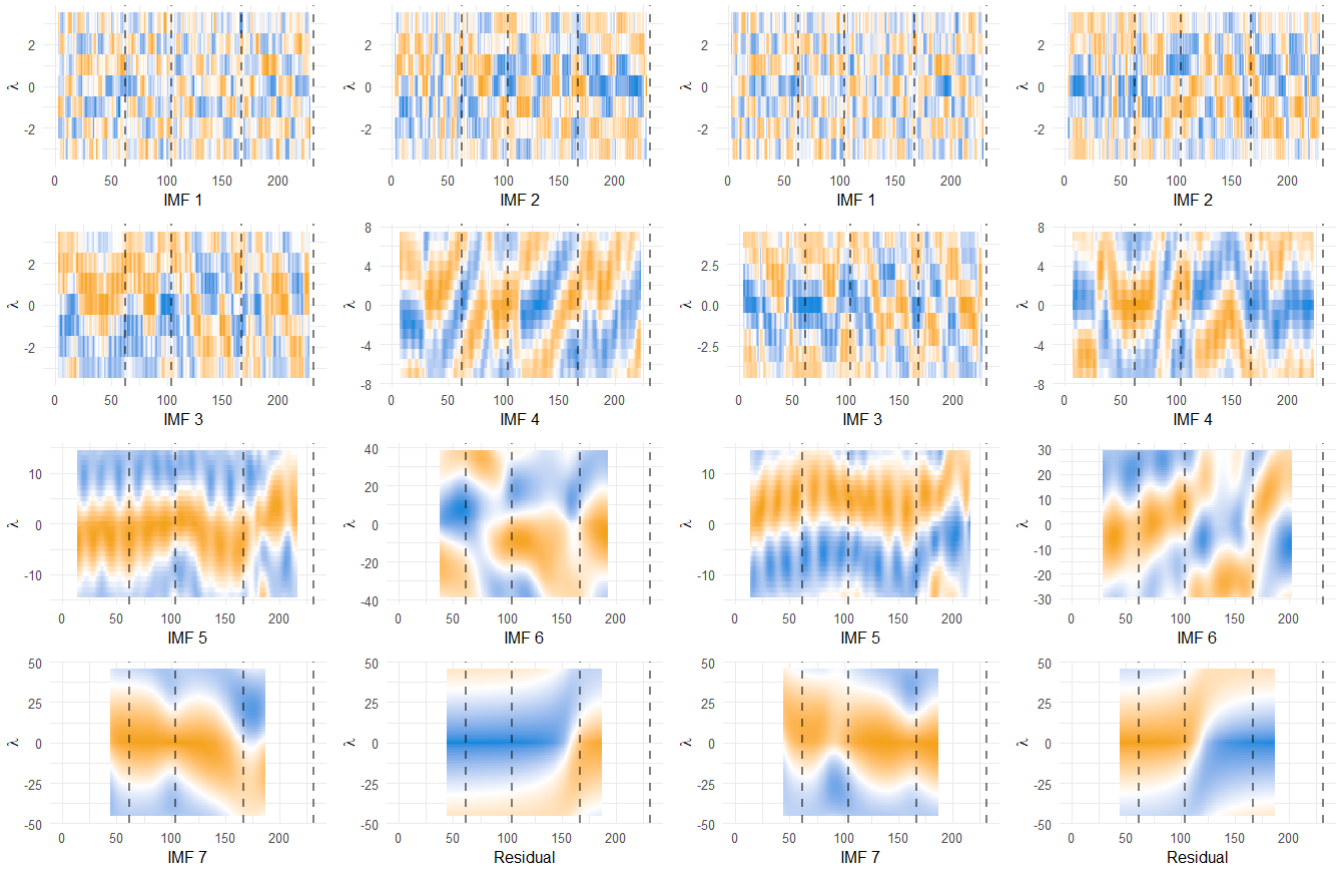
Change Correlations



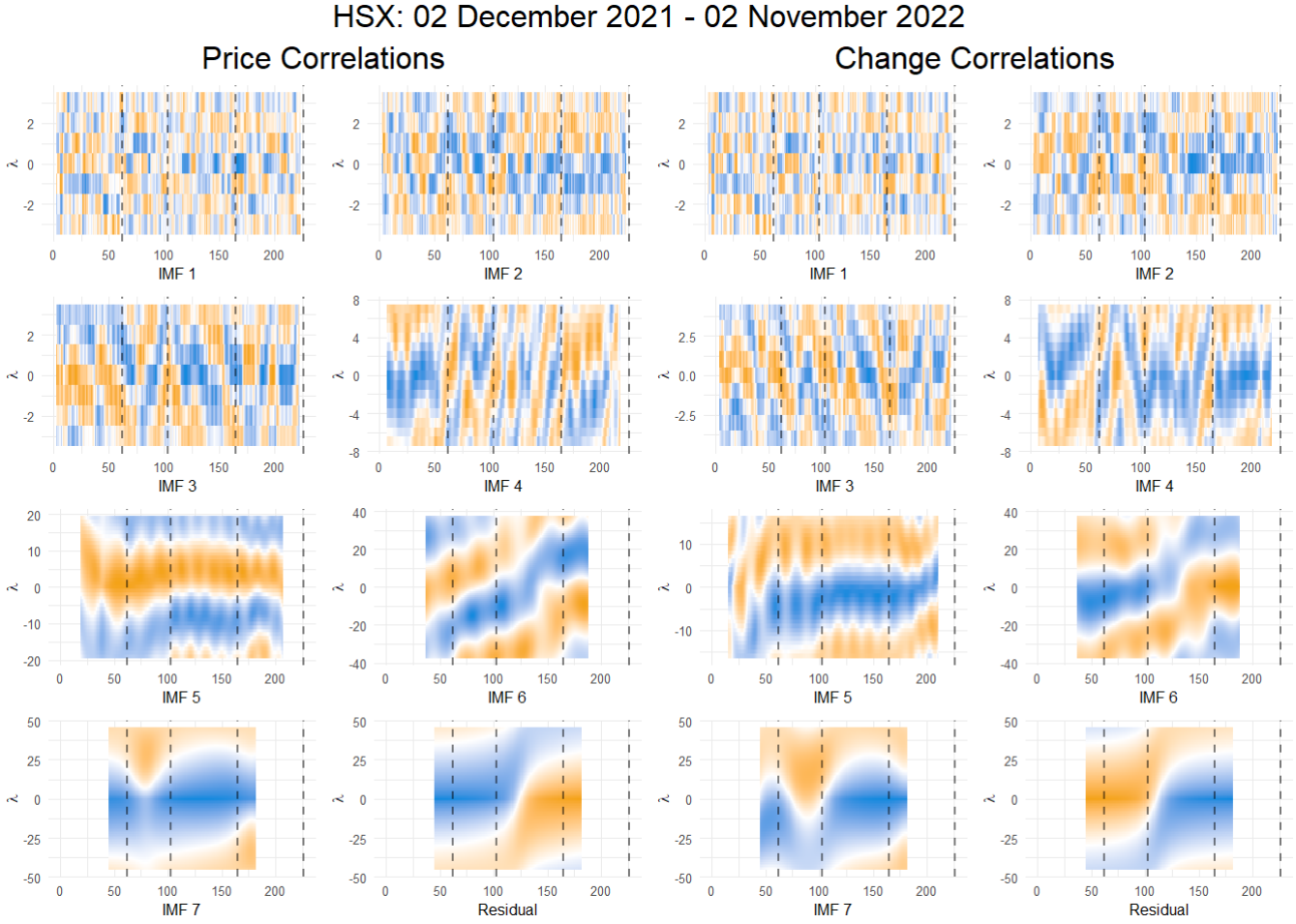
HSX: 03 December 2020 - 02 November 2021

Price Correlations

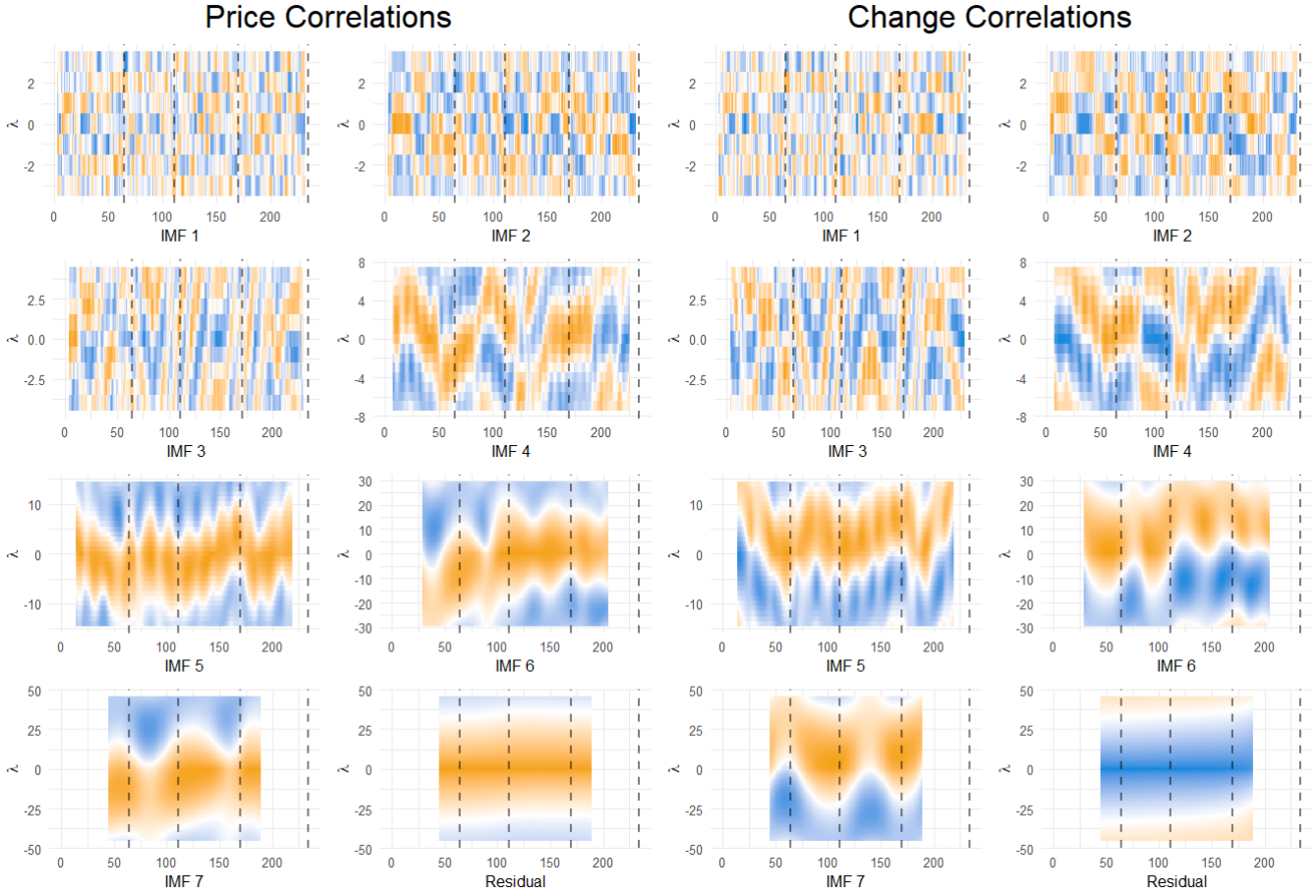
Change Correlations

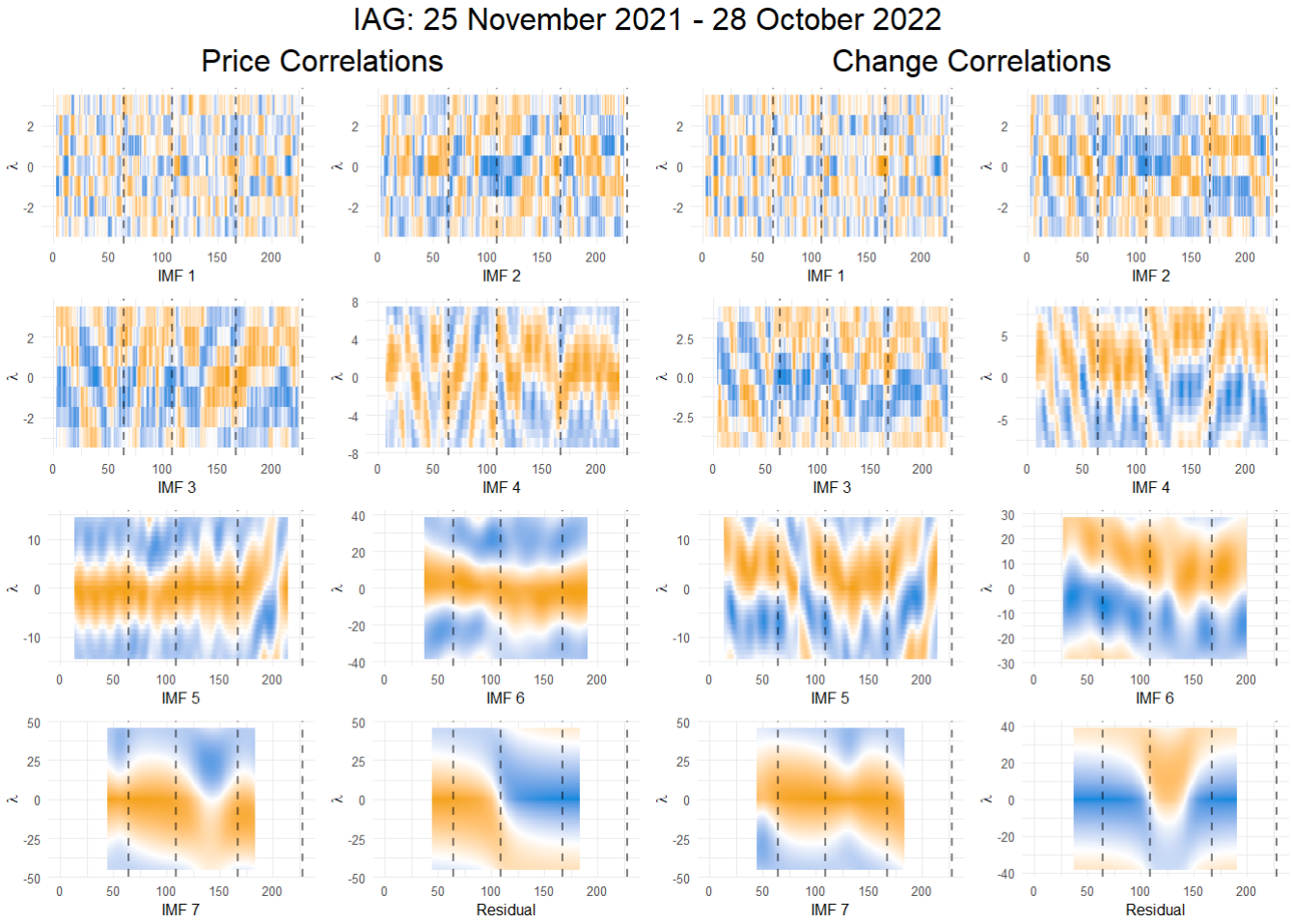


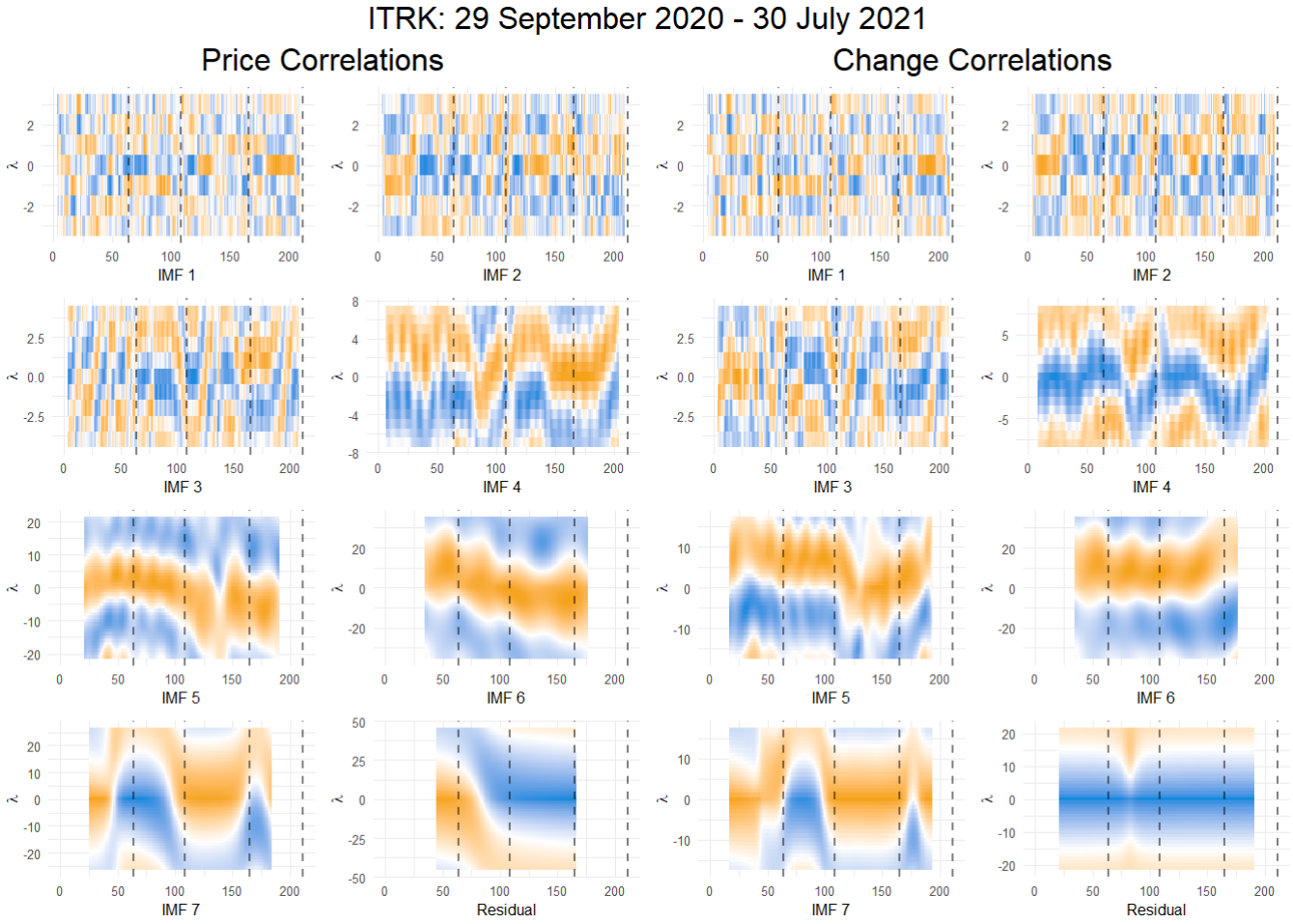
HSX: 02 December 2021 - 02 November 2022



IAG: 28 November 2019 - 30 October 2020

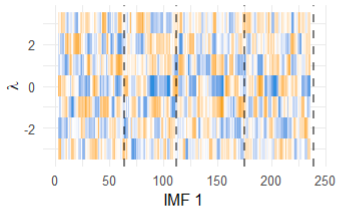


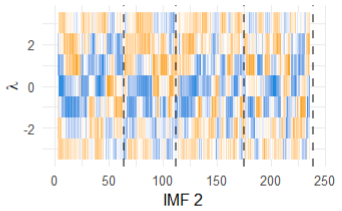

IAG: 25 November 2021 - 28 October 2022
Price Correlations
Change Correlations
IMF 1
IMF 2
IMF 3
IMF 4
IMF 5
IMF 6
IMF 7
Residual
λ

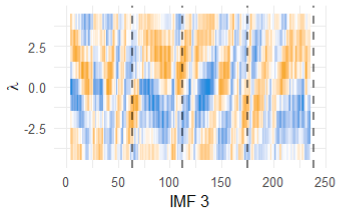

ITRK: 29 September 2020 - 30 July 2021
Price Correlations
Change Correlations
IMF 1
IMF 2
IMF 3
IMF 4
IMF 5
IMF 6
IMF 7
Residual
λ

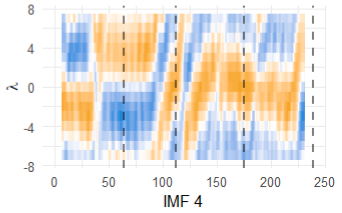
LLOY: 20 November 2019 - 29 October 2020

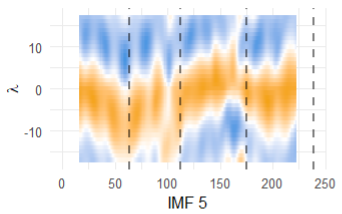
Price Correlations

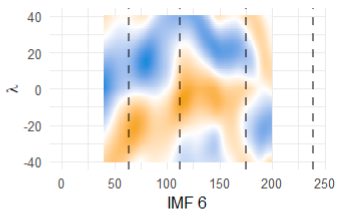


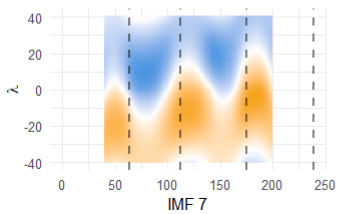


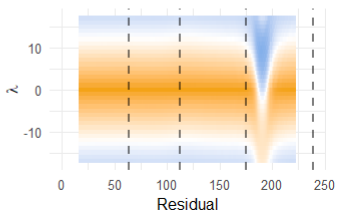




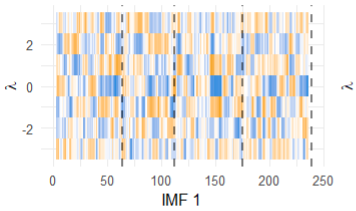


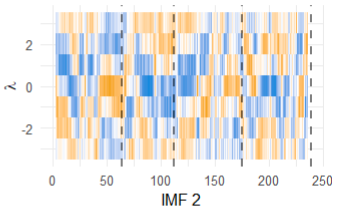


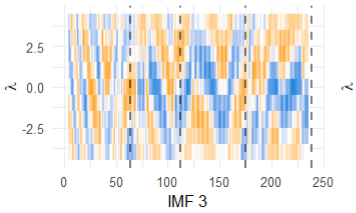


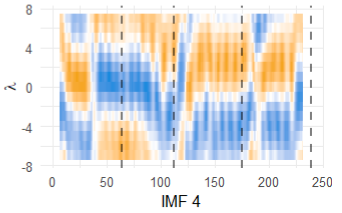


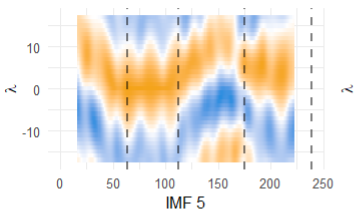
Change Correlations

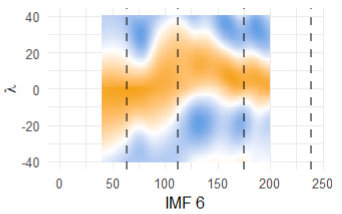


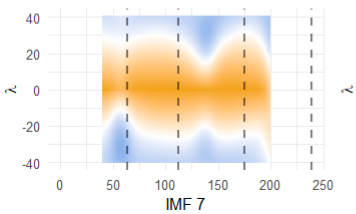


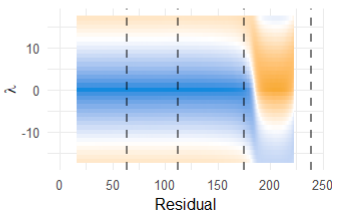






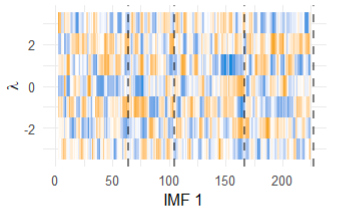


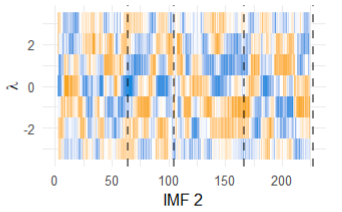


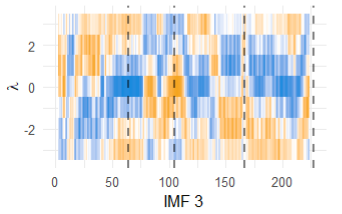


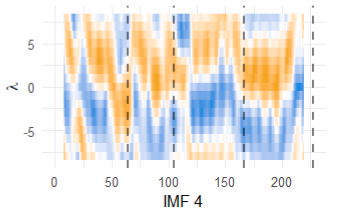
LLOY: 24 November 2021 - 27 October 2022

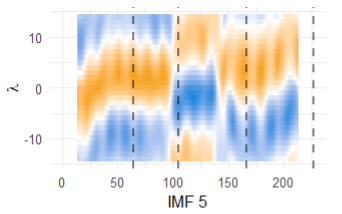
Price Correlations

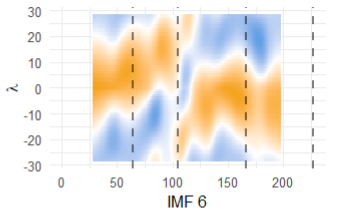


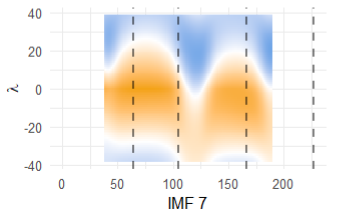


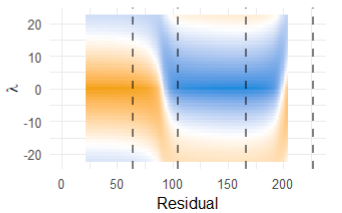




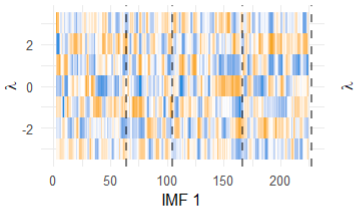


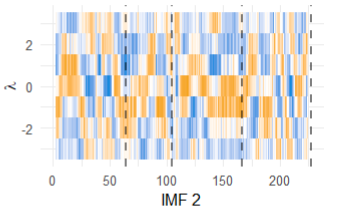


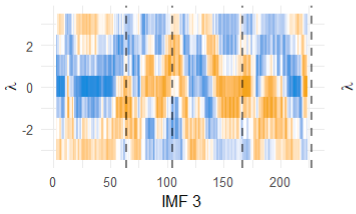


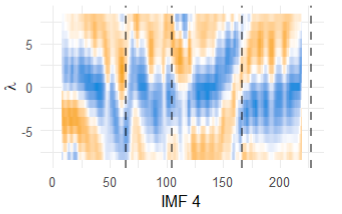


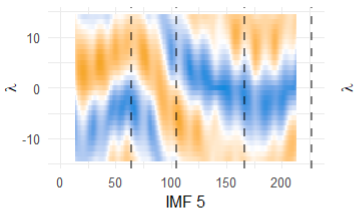
Change Correlations

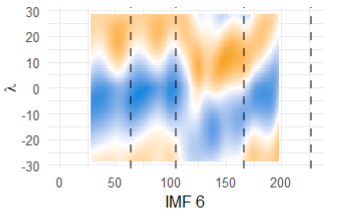


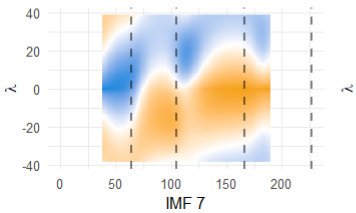


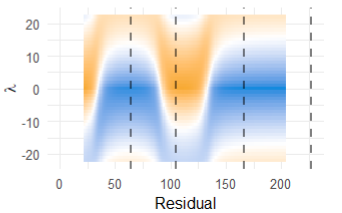








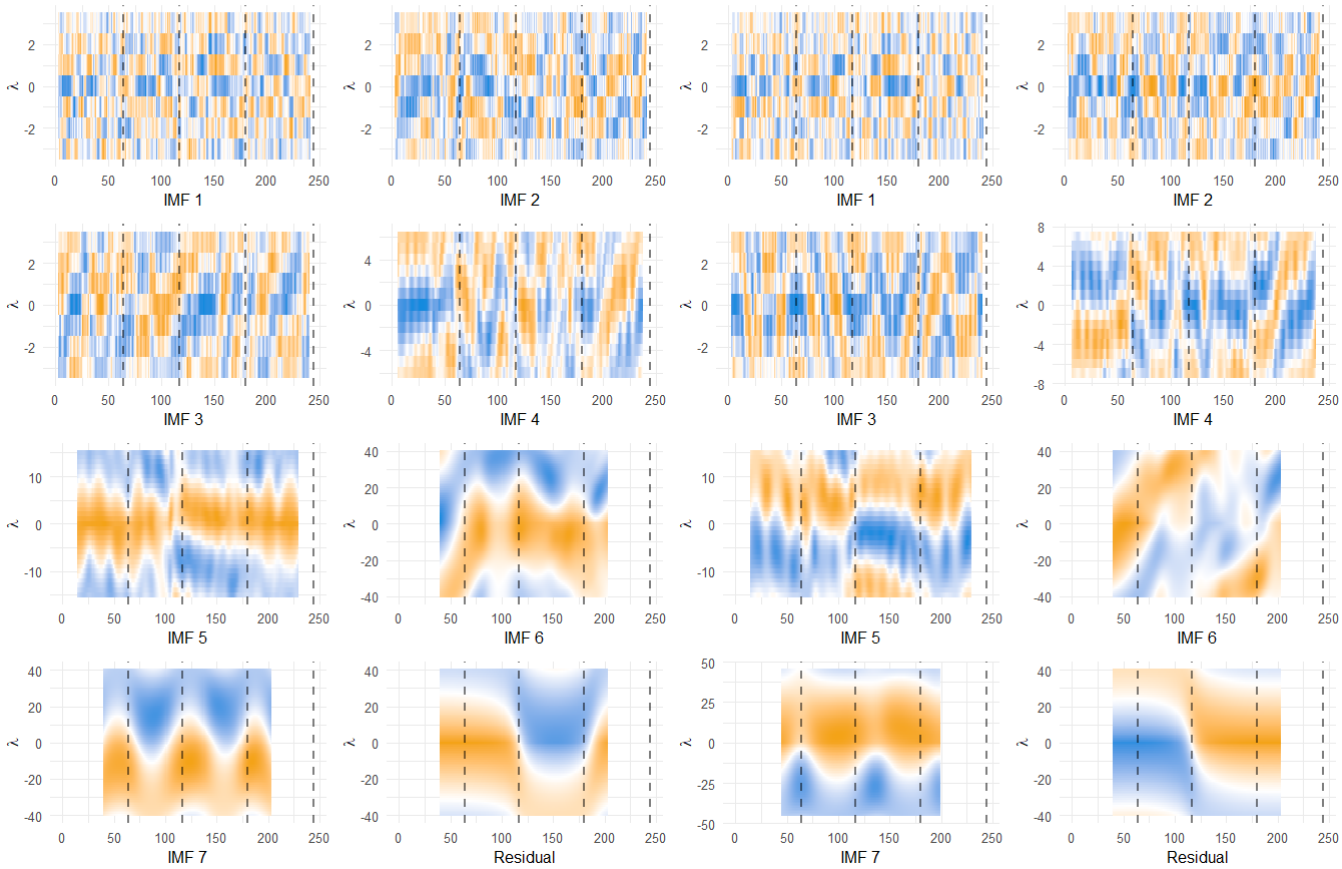




NWG: 14 November 2019 - 30 October 2020

Price Correlations

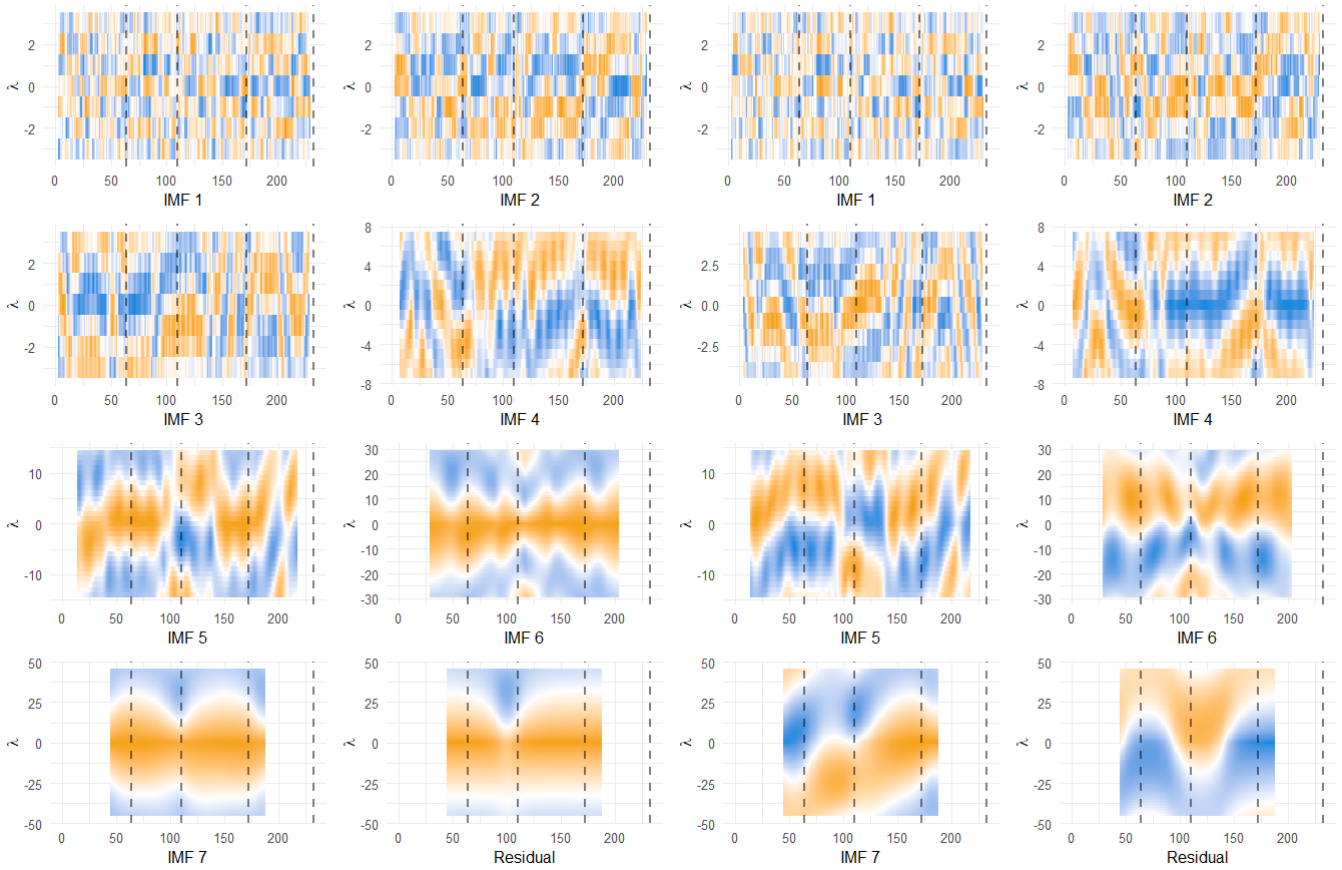
Change Correlations



NWG: 18 November 2021 - 28 October 2022

Price Correlations

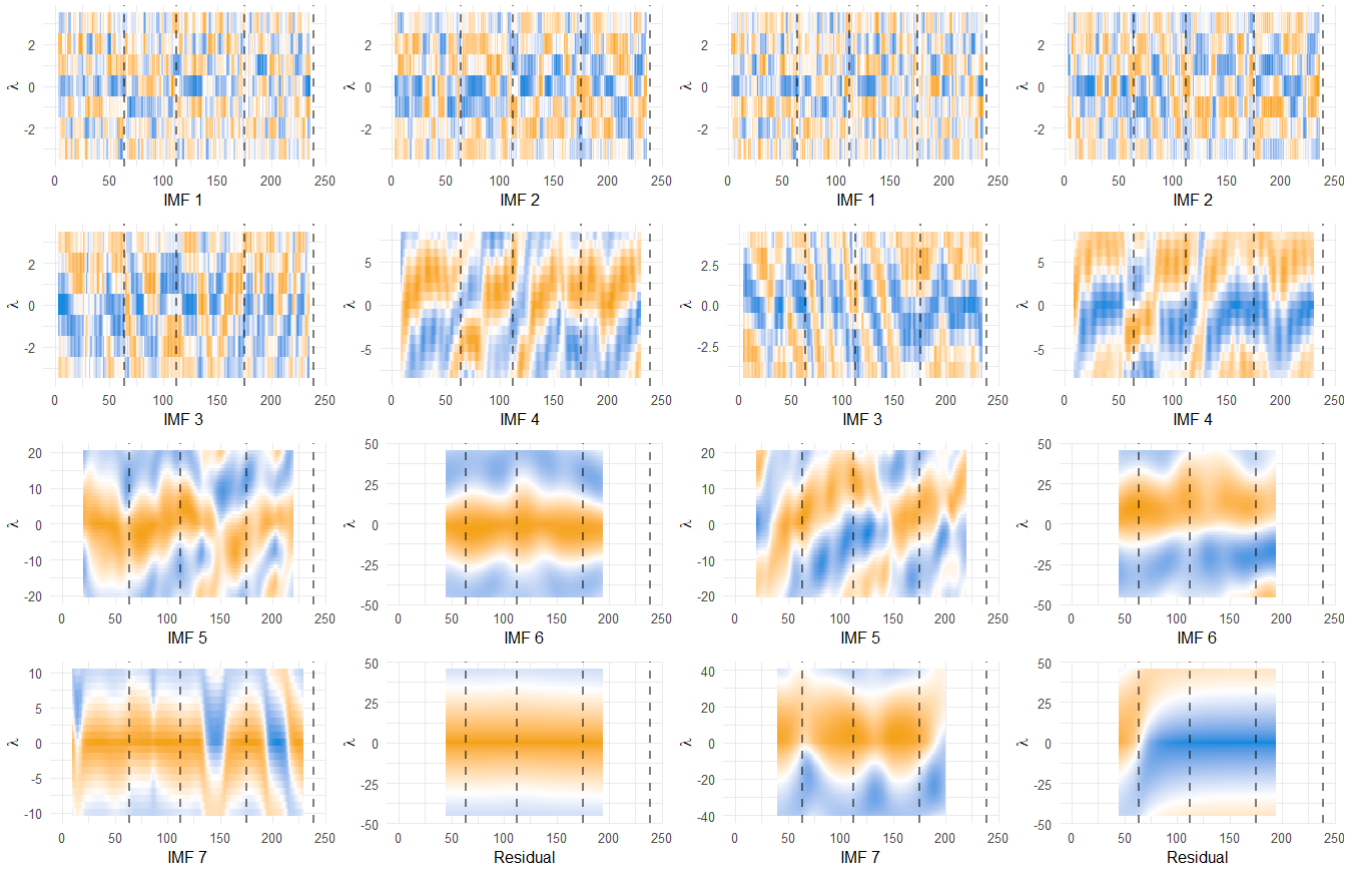
Change Correlations



REL: 13 November 2019 - 22 October 2020

Price Correlations

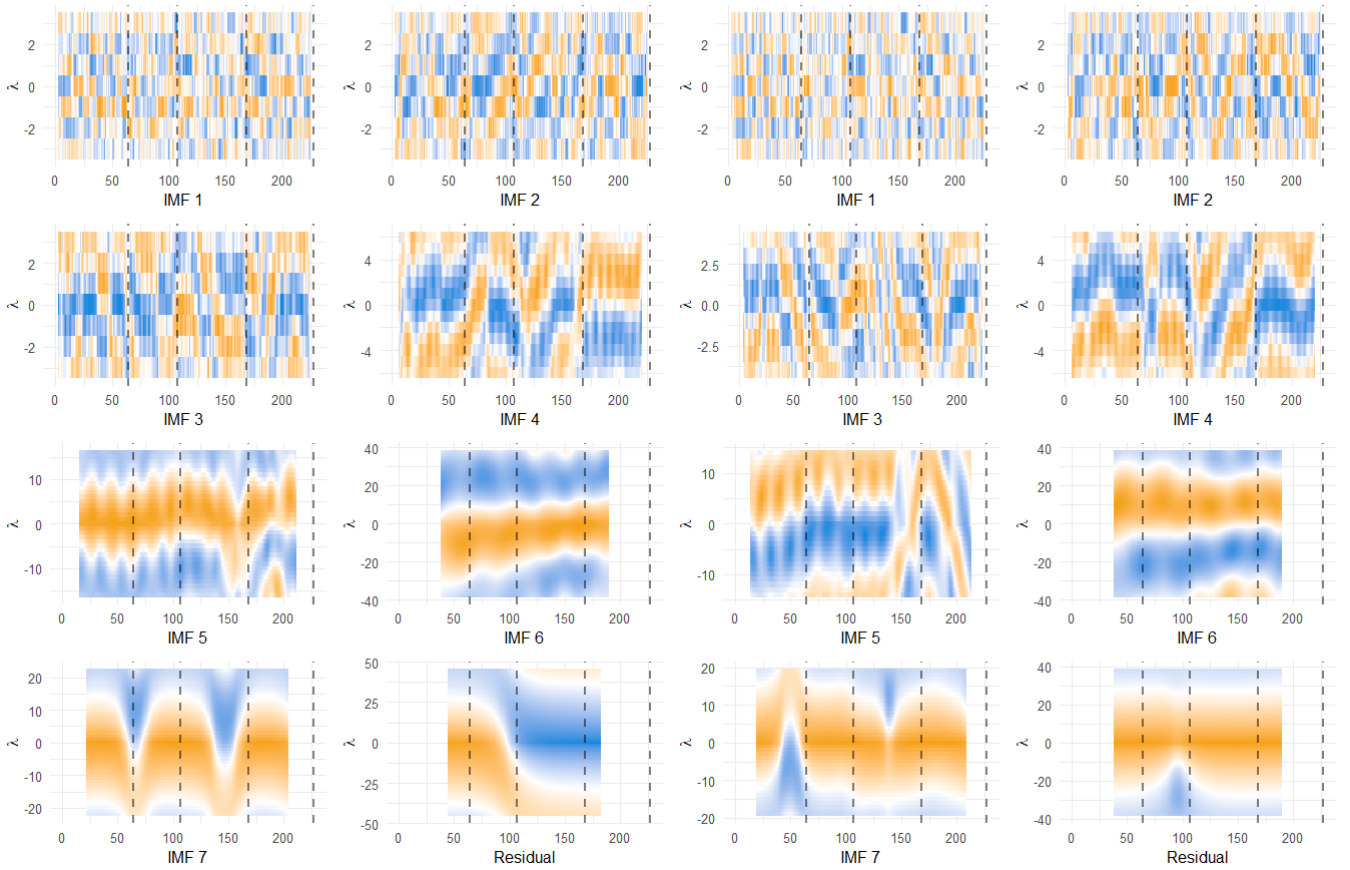
Change Correlations



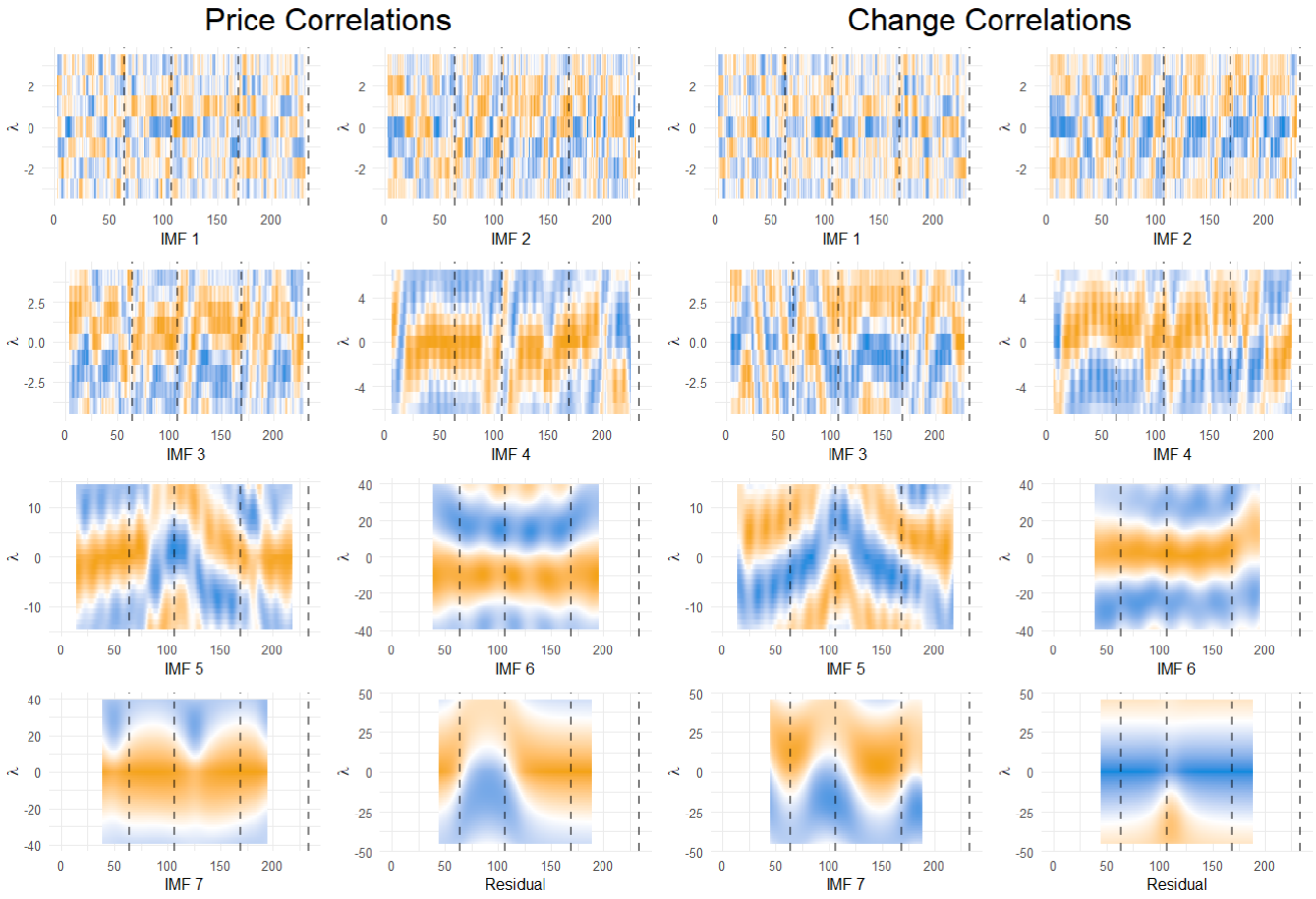
RKT: 27 November 2019 - 20 October 2020

Price Correlations

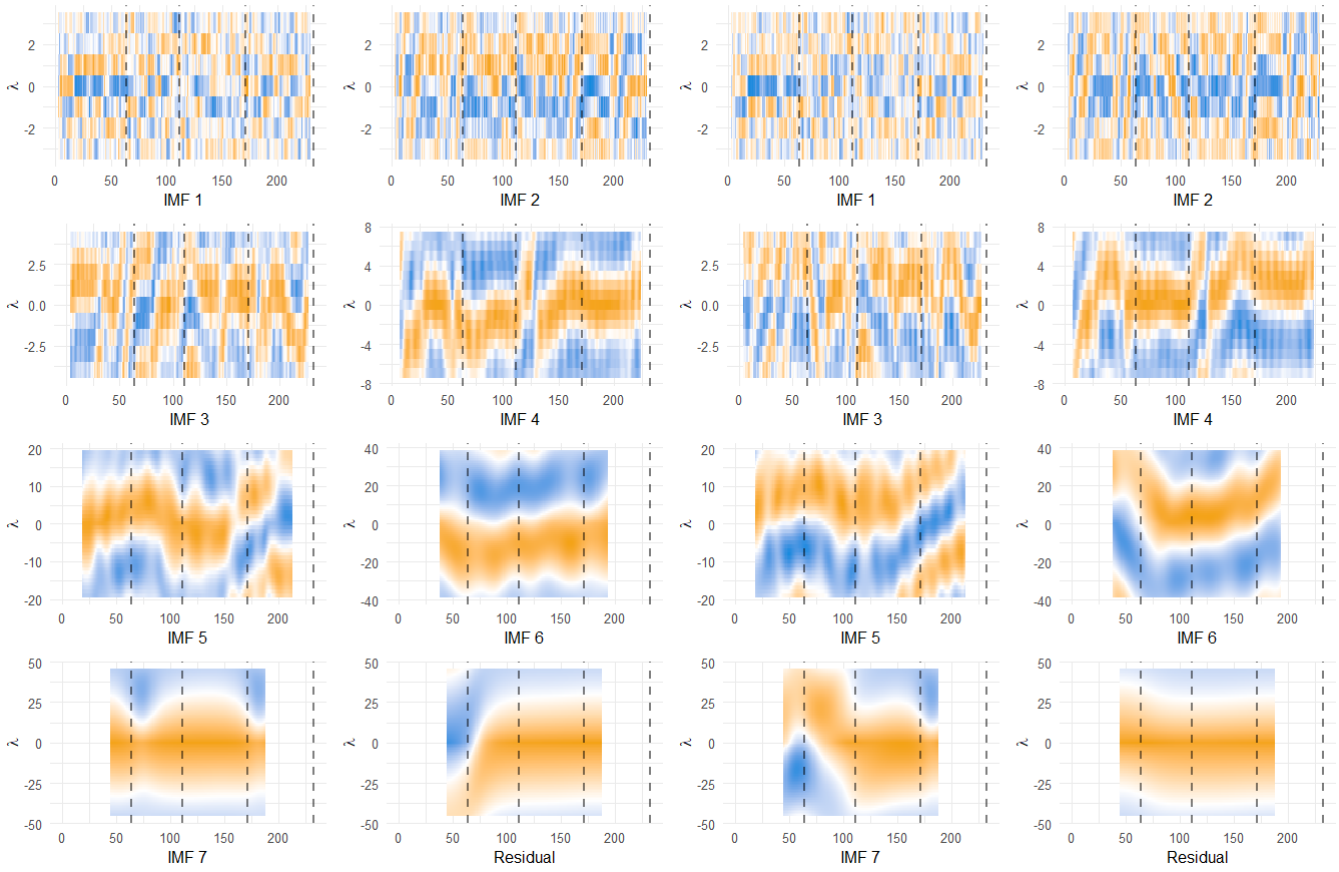
Change Correlations



RKT: 24 November 2020 - 26 October 2021

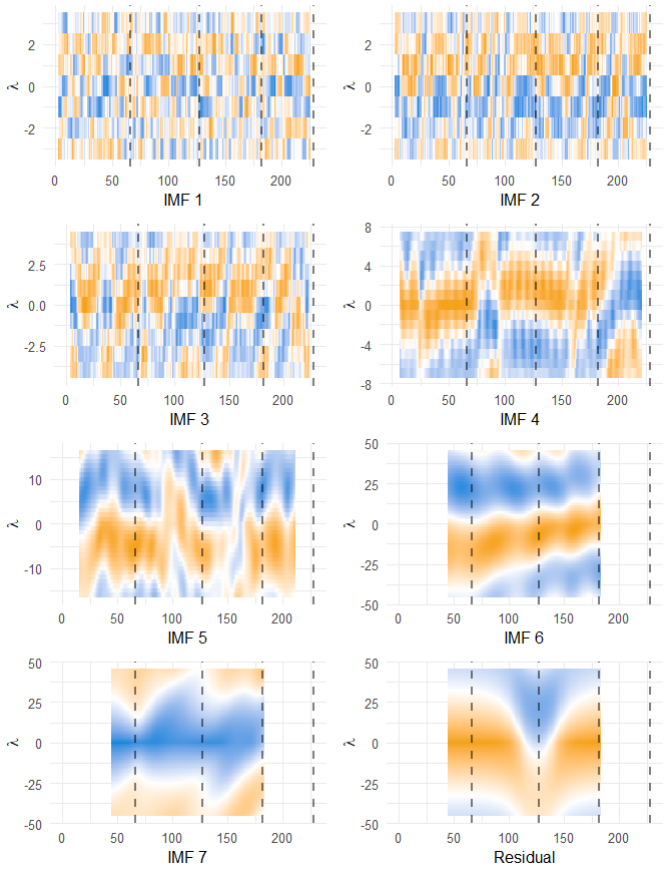


RKT: 17 November 2021 - 26 October 2022

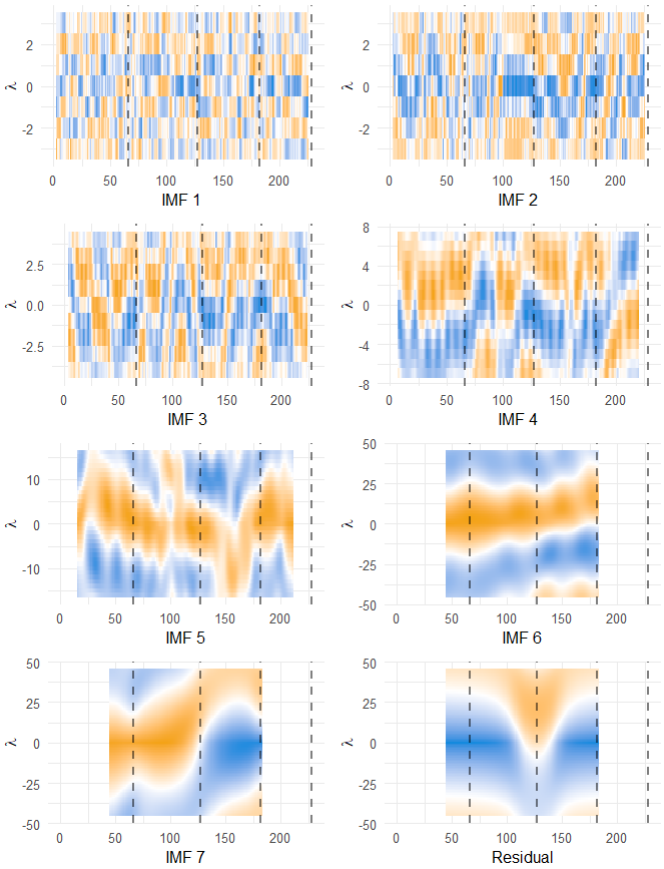


RR: 11 September 2020 - 05 August 2021

Price Correlations

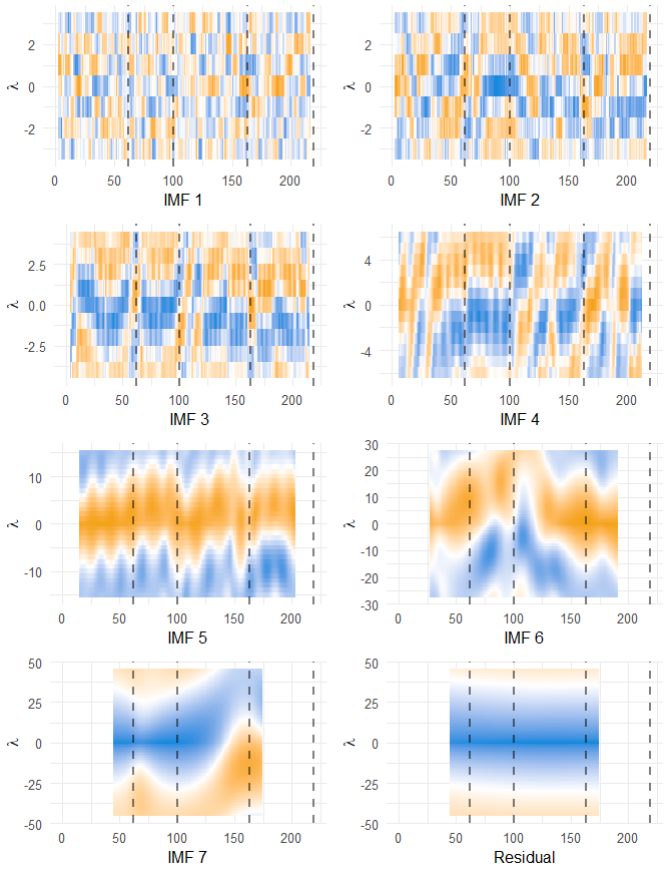


Change Correlations

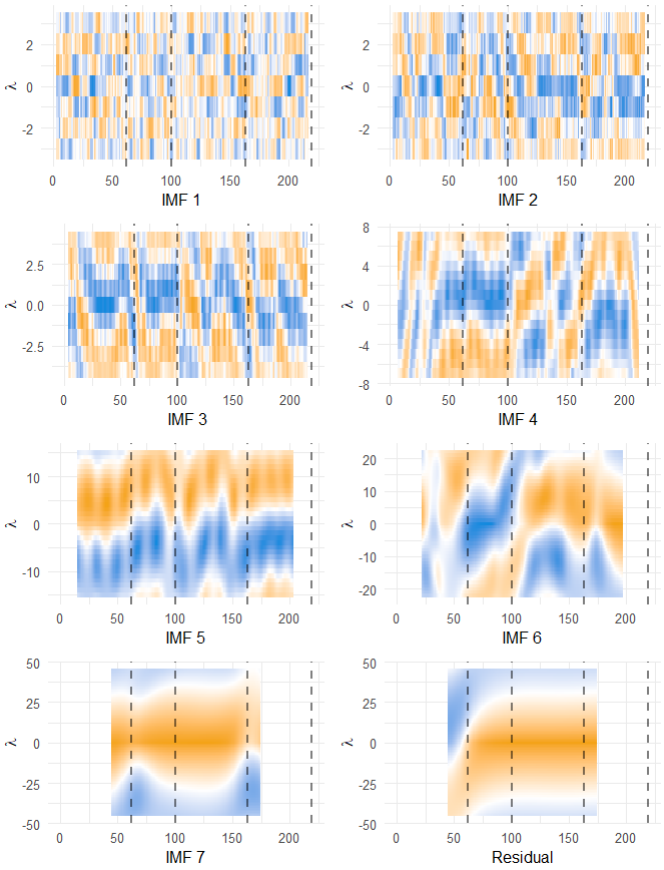


SDR: 04 December 2020 - 18 October 2021

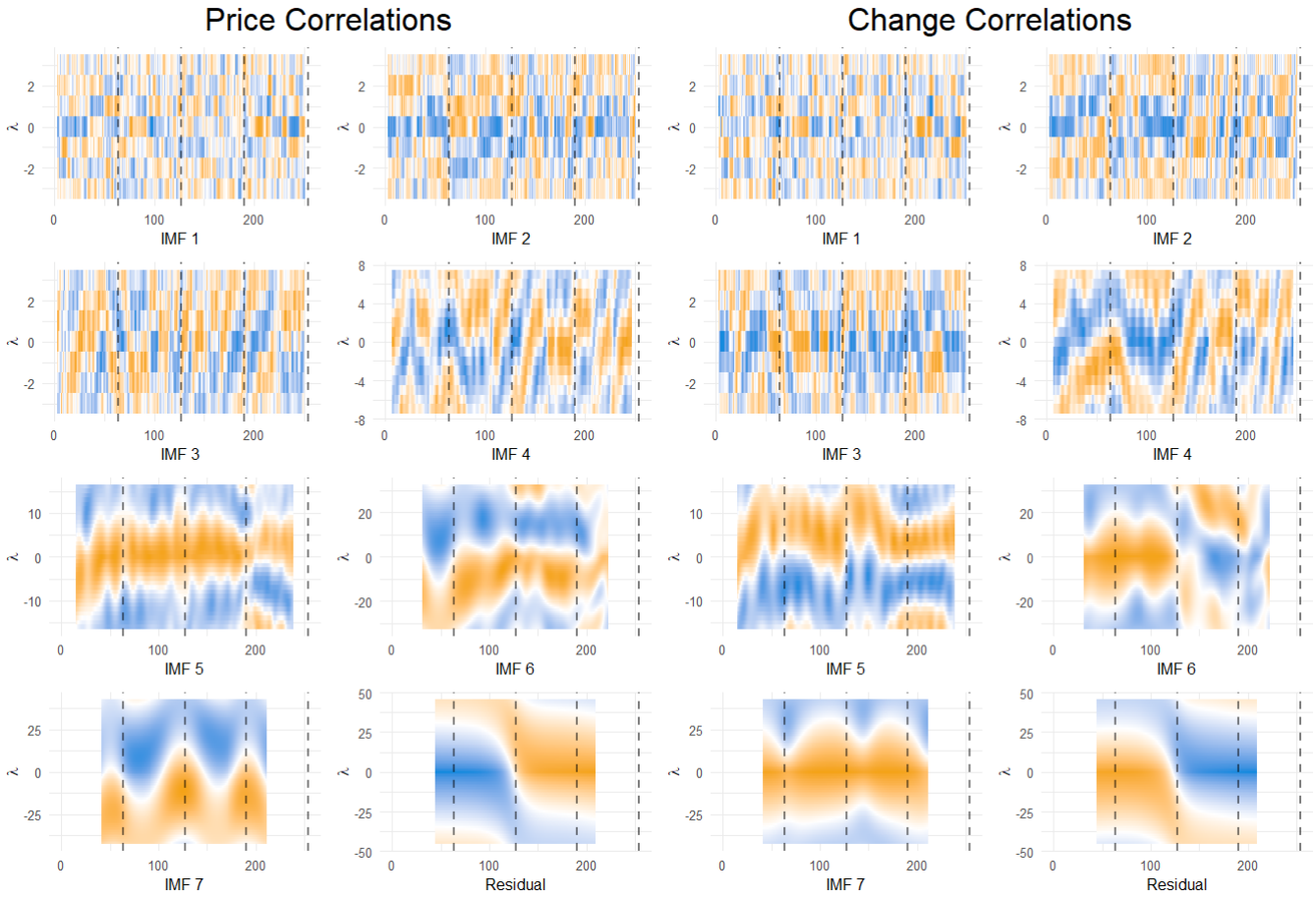
Price Correlations



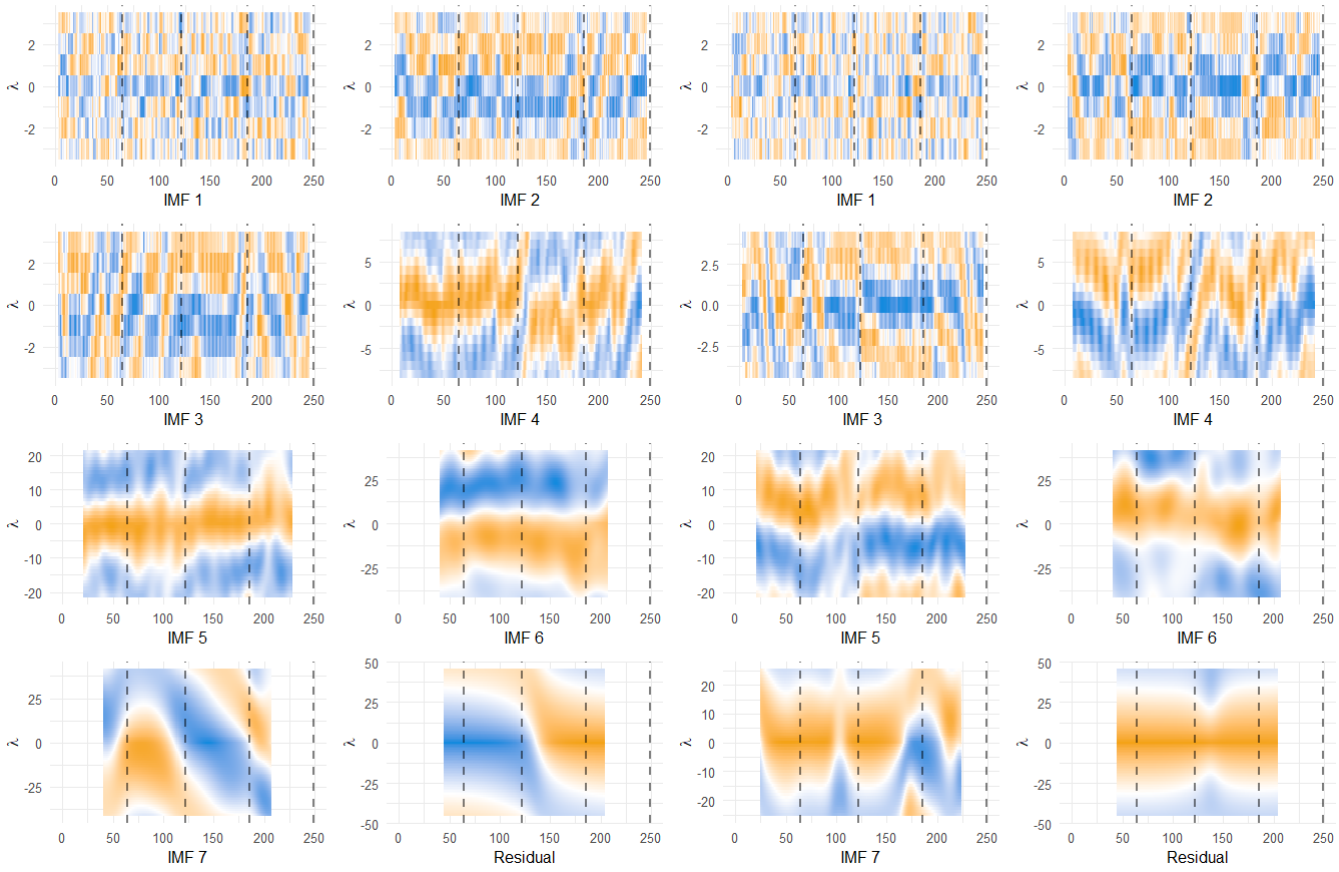
Change Correlations



SHEL: 30 October 2019 - 29 October 2020

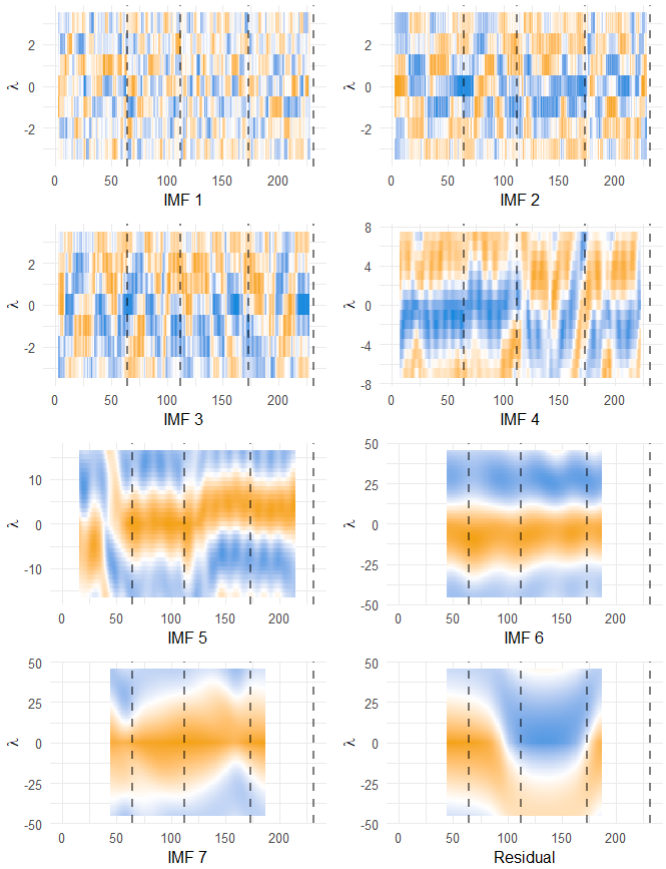


SHEL: 04 November 2020 - 28 October 2021

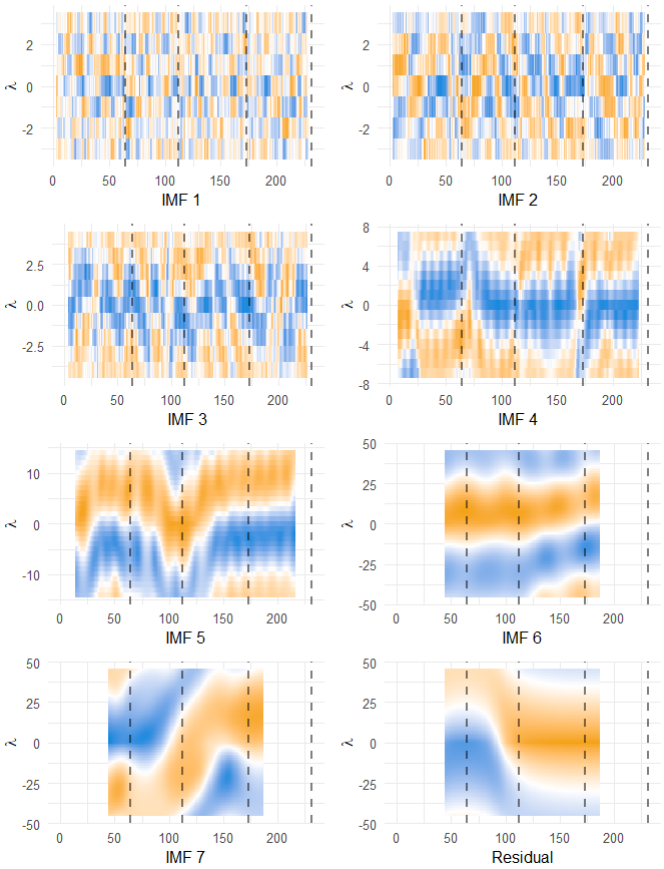


STAN: 17 November 2021 - 26 October 2022

Price Correlations

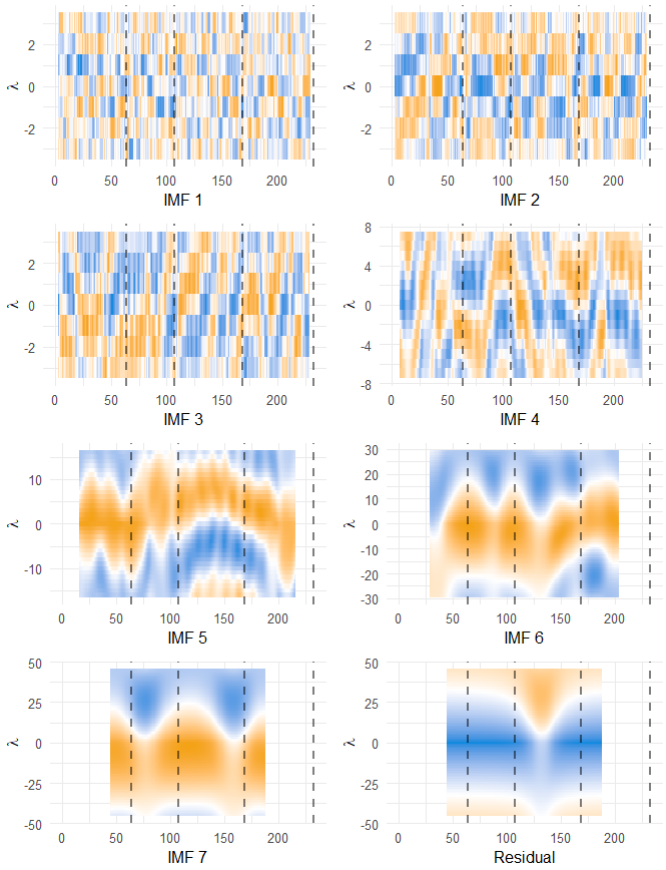


Change Correlations

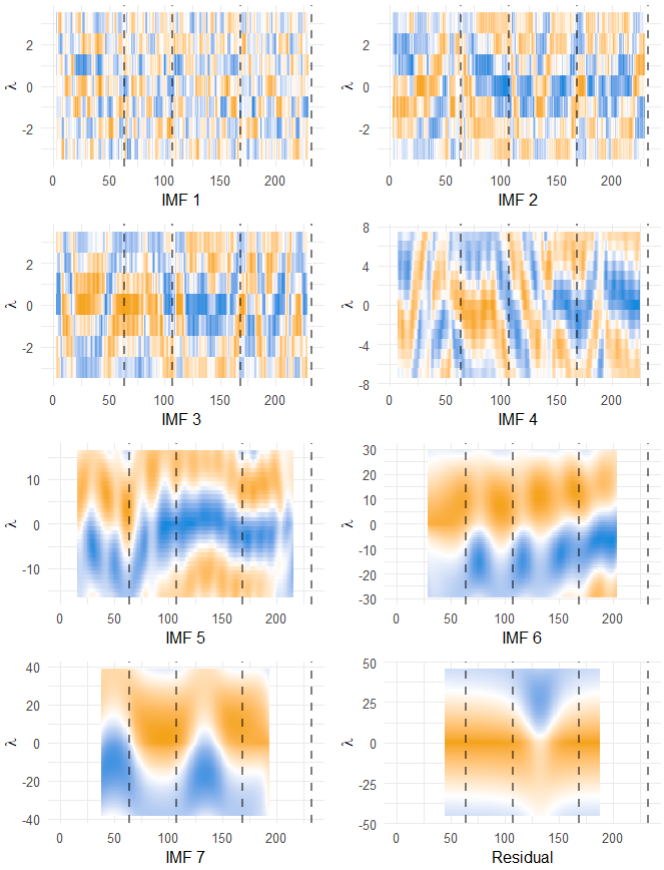


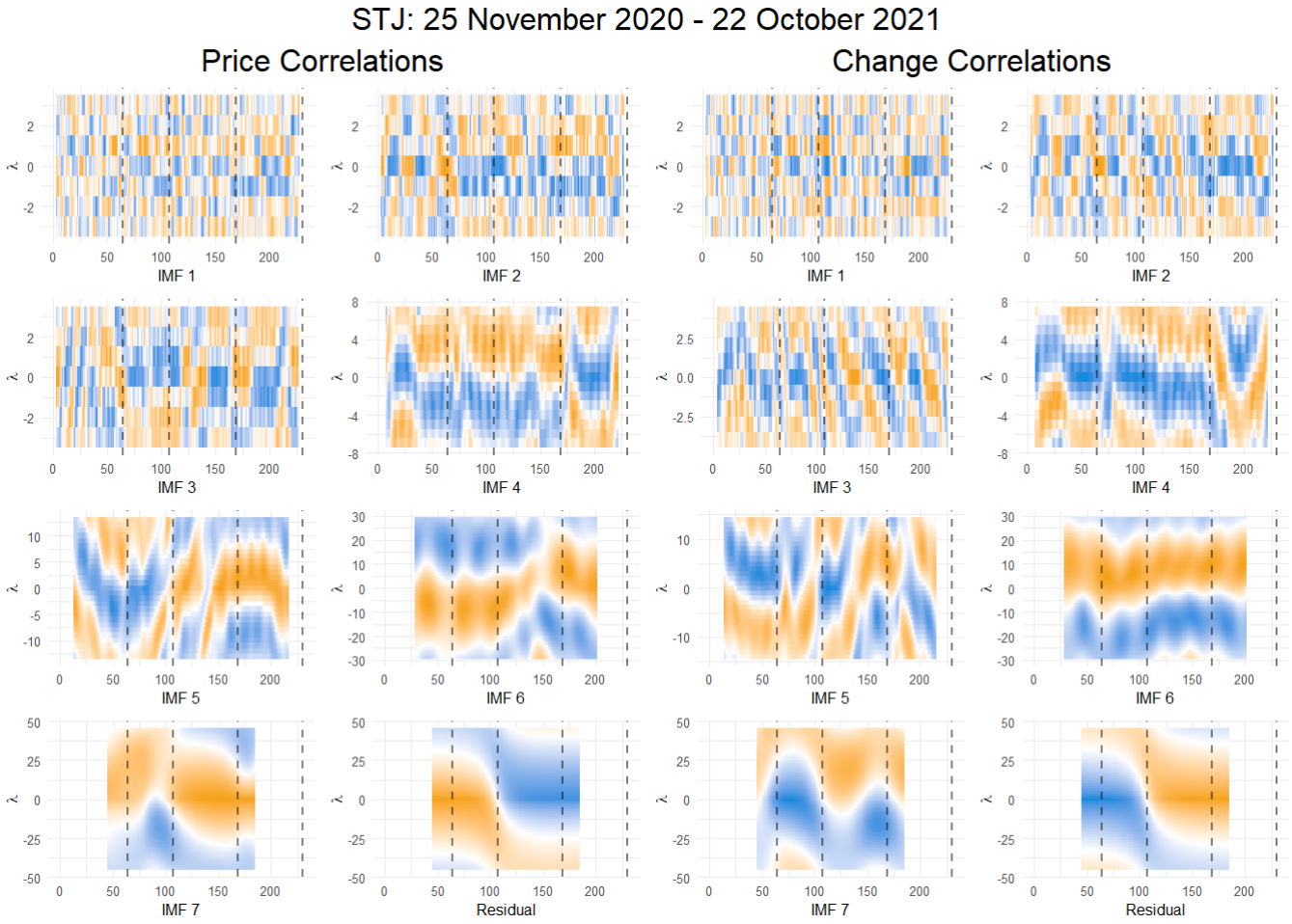
STJ: 27 November 2019 - 27 October 2020

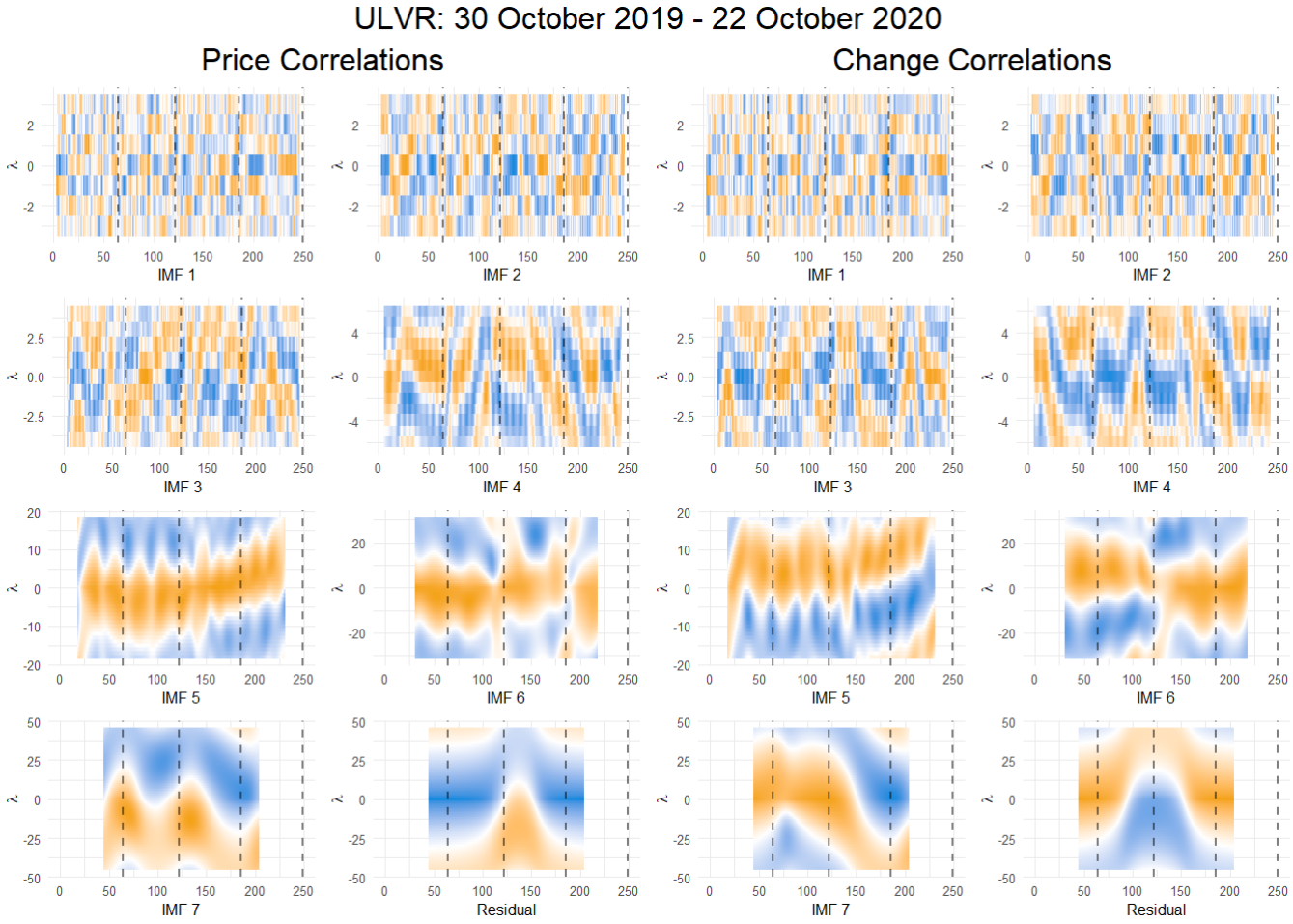
Price Correlations

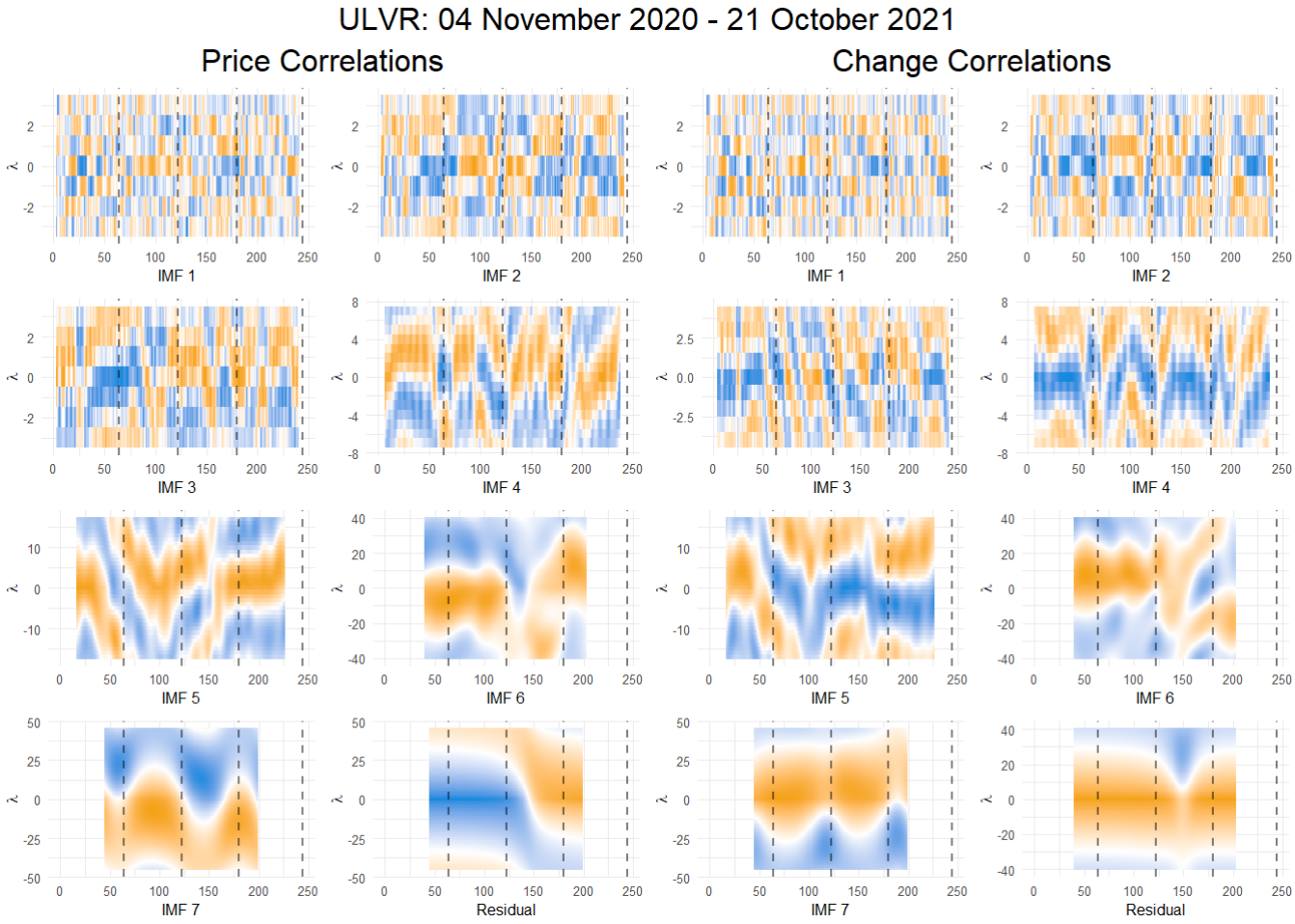


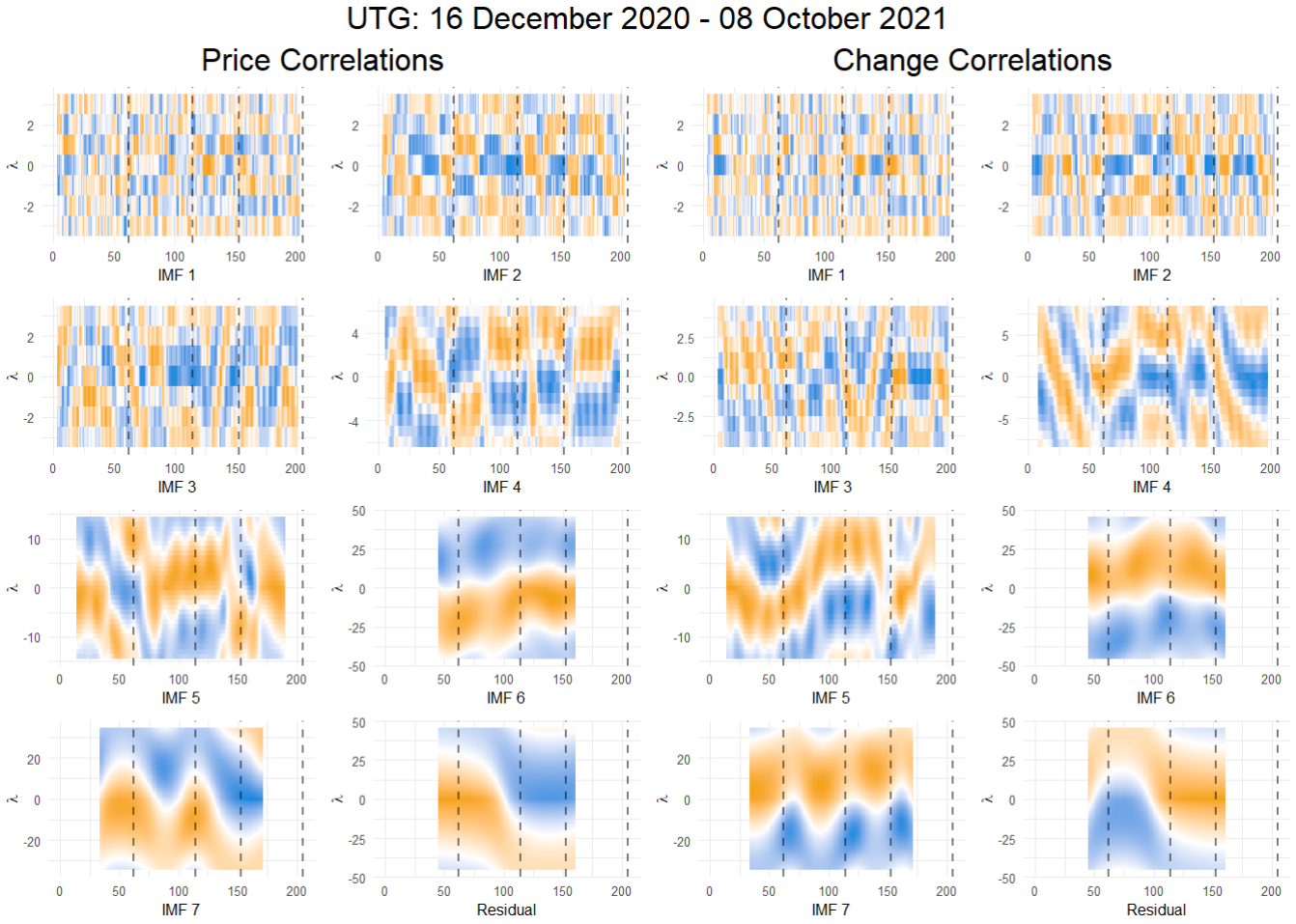
Change Correlations




STJ: 25 November 2020 - 22 October 2021
Price Correlations
Change Correlations
IMF 1
IMF 2
IMF 3
IMF 4
IMF 5
IMF 6
IMF 7
Residual
λ


ULVR: 30 October 2019 - 22 October 2020
Price Correlations
Change Correlations
IMF 1
IMF 2
IMF 3
IMF 4
IMF 5
IMF 6
IMF 7
Residual
λ


ULVR: 04 November 2020 - 21 October 2021
Price Correlations
Change Correlations
IMF 1
IMF 2
IMF 3
IMF 4
IMF 5
IMF 6
IMF 7
Residual


UTG: 16 December 2020 - 08 October 2021
Price Correlations
Change Correlations
IMF 1
IMF 2
IMF 3
IMF 4
IMF 5
IMF 6
IMF 7
Residual