Application of Invariant Feature-Based Multi-Reference Alignment on Lead-Lag Detection

Danni Shi
danni.shi@eng.ox.ac.uk
University of Oxford
Oxford, UK

Mihai Cucuringu
mihai.cucuringu@stats.ox.ac.uk
Department of Statistics, University of Oxford
Oxford, UK
The Alan Turing Institute
London, UK

Jan-Peter Calliess
jan-peter.calliess@oxford-man.ox.ac.uk
University of Oxford
Oxford, UK

ABSTRACT

In time series analysis, numerous methods have been developed to detect, measure and apprehend lead-lag relationships between variables, which refers to the time-delay of patterns in a time series relative to the other. Often, the extent of lead-lag between two time series is directly measured with a similarity metric. This approach suffers inconsistency and inaccuracy especially under high-noise regimes.

This work represents several frameworks of Multireference alignment (MRA), the estimation of a unified, de-noised latent signal from a population of time series with linearly-shifted and noisy patterns, and their application on lead-lag detection. The recovered latent signal is treated as the reference vector whose lead-lag metric is calculated against other time series. The relative lead-lag between two actual time series is then obtained as the difference between their lags against the reference vector. Results show that models with MRA-induced intermediates outperform the simplistic models relying on pairwise similarity in predicting the relative shifts between time series when the signal-to-noise ratio (SNR) is low. The effect of clustering the time series advancing the recovery of the latent signals is also investigated

The results of lead-lag detection can be used on trading stock returns. We developed a trading framework that utilises the relative shift estimations to construct financial signals from the leading time series and trade on multiple portfolios of the lagging time series.

CCS CONCEPTS

• **Computer systems organization** → **Embedded systems**; *Redundancy*; Robotics; • **Networks** → Network reliability.

KEYWORDS

High-dimensional time series, lead-lag, clustering, financial markets

ACM Reference Format:
Danni Shi, Mihai Cucuringu, and Jan-Peter Calliess. 2023. Application of Invariant Feature-Based Multi-Reference Alignment on Lead-Lag Detection. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym 'XX)*. ACM, New York, NY, USA, 2 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

1.1 Background

Time series forecasting is a major topic in data science and has been extensively researched from multiple perspectives. Forecasting financial time series is especially challenging due to low SNR and the deficiency of auto-correlation. Yet, in the financial market, it has been observed that the prices of certain groups of assets respond to events and factors quicker than others. In multivariate time series models, market factors can create similar patterns of change in the financial time series of different assets, each with a different time delay. This is what we call a lead-lag structure amongst the time series variables. We aim to construct a measure that estimates the extent and length of lead-lag between two stocks. In this paper, the extent of lead-lag is represented as the maximum cross-correlation between two time series up to a shift. The corresponding length of lead-lag is the number of timescale shifts (as we primarily works with discrete-time data in finance) to achieve the maximum cross-correlation.

Current methods of lead-lag detection often rely on computing a lead-lag score between a pair of time series (some refs and descriptions here), which are direct and effective in some use cases. However, several problems arise with this approach:

- Non-robust: At low SNR, pairwise measurement is significantly interfered by the noise in the data and hence become inaccurate
- Inconsistent: Pairwise lag predictions need to be synchronized for reliable applications. However, the lag between a pair of time series may not agree with the others. [1]
- Computationally expensive: The computation of the pairwise metric of N time series has a complexity $\mathcal{O}(N^2)$, making it non-scalable to large datasets.

Some algorithms aim to address the inconsistencies from pairwise measurements by forming a unifying vector of ranking and/or shifts (refs here). The recovered vector can be seen as a low-dimensional

[1] For example, estimations may show that variable X and Y each leads variable Z by 1 day, but that X lags Y by 1 day too.

Conference acronym 'XX, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/XXXXXXX.XXXXXXX

representation that closely summarise the original pairwise shift matrix. An SVD-based algorithm is shown to recover the ranking of a group of pairwise measurements with efficiency and robustness.

MRA is studied in multiple scientific domains such as structural biology, radar and image processing where observations are abundant yet noisy and cyclically shifted. Some methods tap on the signal features invariant under cyclic shifts and the Central Limit Theorem to construct an optimization problem[?]. The problem converges to an estimated latent signal and was shown to perform well at high noise levels as long as a sufficient number of observations are available.

1.2 Our contributions

In this paper, we propose a framework to estimate the relative shifts of multiple time series. We model financial time series in the same universe as the noisy and linearly shifted copies of several latent source signals. Clustering is performed to identify the time series coming from the same source signal. After that, we investigate the accuracy and limitations of 3 different MRA methods in terms of signal recovery and lead-lag detection. This part is implemented on synthetic data where the ground truths of latent signal and shifts are known. Furthermore, we separate the time series into groups of the same shift and devise a trading strategy that derive financial signals from the leading groups and trade on a basket of lagging stocks. To demonstrate the application in financial markets, we implement the clustering → lead-lag detection → trading pipeline on a dataset of 695 US equities' returns. With comparison to the baseline pairwisely evaluated model, the 3 MRA-based methods exhibit strong ability to estimate lead-lag relationships at high noise regimes and produce more significant portfolio returns from the trading strategy.

2 LEAD-LAG DETECTION BASED ON CROSS-CORRELATION

2.1 Pairwise Lag Measurement

2.2 Relative Lag Measurement with a Reference Signal

3 MULTIREFERENCE ALIGNMENT

3.1 Homogeneous MRA

The original method aims to recover the original signal from cyclically shifted copies of noisy observations while the financial data we work with contains non-cyclic shifts. We make an educated assumption that the shifts are small compared to the length of the time series so that a large proportion of the time series agrees with the cyclically shifted version.

3.2 Heterogeneous MRA

3.3 MRA Based on Correlation and Synchronization

4 EXPERIMENTS AND RESULTS

4.1 Synthetic Data

4.2 Financial Data

5 DISCUSSIONS

6 CONCLUSIONS

7 REFERENCES

ACKNOWLEDGMENTS

To Robert, for the bagels and explaining CMYK and color spaces.

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009