Université Libre de Bruxelles
Faculté des Sciences
Département d'Informatique

# Spectral Factor Model for Time Series Learning

**Abhilash Alexander Miranda**


Thèse présenté en vue de
l'obtention du grade académique de
Docteur en Sciences

Année académique
2011-2012

# Spectral Factor Model for Time Series Learning

Abhilash Alexander Miranda

Prof. Gianluca Bontempi of Machine Learning Group, Département d'Informatique, Université libre de Bruxelles is the promoter of this thesis. The other members of the jury are:

- Prof. Christine De Mol (Département de Mathématique and European Center for Advanced Research in Economics and Statistics, Université libre de Bruxelles)
- Prof. Maarten Jansen (Département de Mathématique and Département d'Informatique, Université libre de Bruxelles)
- Prof. Guy Latouche (Méthodes algorithmiques en probabilité, Département d'Informatique, Université Libre de Bruxelles)
- Prof. Michel Verleysen (Machine Learning Group, Université catholique de Louvain)

## Acknowledgements

I dedicate this work to my brothers, Anoop and Anand, for their unconditional love.

My deepest gratitude goes to Prof. Gianluca Bontempi. As my advisor and colleague, it was his belief in my aptitude that constantly fuelled this work. The generous freedom he gave me, his patience, selfless and timely advices, deep understanding of my strengths and weaknesses, his questions and careful reviews meant I always craved for higher quality and greater challenges.

The quality of this thesis improved immensely due to the reviews and suggestions from the honourable members of the thesis jury.

Project TANIA of Région Wallonne funded this work; Prof. Luc Barvais and François Clément were very supportive.

The fraternity at the Machine Learning Group and my office mates Olivier Caelen, Olivier Cailloux and Maud Vidick meant a life-jacket to me. Since my undergraduate years, I am very privileged to have Dr. P. P. Mohanlal as a mentor which aided me considerably in my pursuit of higher education.

I thank the respected members of my Committé d'Acompaignement for their advices to see this thesis to completion. The wonderful team at the Secretariat of the Département d'Informatique ensured no administrative paperwork interferes with research work; merci à vous!

The entire D Square NV team greatly encouraged me towards the finish line.

Infant thoughts of this thesis were first inked in one serviette over a lunch with Yann-Aël Le Borgne towards the end of the scheduled funding. Prof. Bontempi encouraged me posing the "show the experiments" challenge and monitored closely. Catharina Olsen spent numerous evenings and weekends containing the overdrive of my algebraic intuitions and never complained. Pieter Schuddink was always eager to jump into any depths of the analysis. Olivier Caelen took pains in proofreading parts of the thesis; Maryam Bagheri, Tim Bosschaerts, and Eduard Céspedes Borràs gave useful tips. But the fault for any and all lapses in this thesis lies solely with me.

Numerous great hearts instilled confidence in me to stay afloat in this work during the tempests in my research and personal lives. I cannot list them all here; but my utmost appreciation goes to Marcin Kaminski, François Clément, Manju Hiranmay, Catharina Olsen, Eduard Céspedes Borràs, Fabien Rogister, Patrick Meyer, Zhang Bingbing and Peter Lalvani for helping me stay on track.

Finally, but most importantly, I wish to sincerely admit that the hard work during these years is my parents' and they wish I take all the credit; thank you Amma and Acha!

# Contents

# List of Figures

# List of Tables

# Abstract

Today's computerized processes generate massive amounts of streaming data. In many applications, data is collected for modeling the processes. The process model is hoped to drive objectives such as decision support, data visualization, business intelligence, automation and control, pattern recognition and classification, etc. However, we face significant challenges in data-driven modeling of processes. Apart from the errors, outliers and noise in the data measurements, the main challenge is due to a large dimensionality, which is the number of variables each data sample measures. The samples often form a long temporal sequence called a multivariate time series where any one sample is influenced by the others. We wish to build a model that will ensure robust generation, reviewing, and representation of new multivariate time series that are consistent with the underlying process.

In this thesis, we adopt a modeling framework to extract characteristics from multivariate time series that correspond to dynamic variation-covariation common to the measured variables across all the samples. Those characteristics of a multivariate time series are named its 'commonalities' and a suitable measure for them is defined. What makes the multivariate time series model versatile is the assumption regarding the existence of a latent time series of known or presumed characteristics and much lower dimensionality than the measured time series; the result is the well-known 'dynamic factor model'. Original variants of existing methods for estimating the dynamic factor model are developed: The estimation is performed using the frequency-domain equivalent of the dynamic factor model named the 'spectral factor model'. To estimate the spectral factor model, ideas are sought from the asymptotic theory of spectral estimates. This theory is used to attain a probabilistic formulation, which provides maximum likelihood estimates for the spectral factor model parameters. Then, maximum likelihood parameters are developed with all the analysis entirely in the spectral-domain such that the dynamically transformed latent time series inherits the commonalities maximally.

The main contribution of this thesis is a learning framework using the spectral factor model. We term learning as the ability of a computational model of a process to robustly characterize the data the process generates for purposes of pattern matching, classification and prediction. Hence, the spectral factor model could be claimed to have learned a multivariate time series if the latent time series when dynamically transformed extracts the commonalities reliably and maximally. The spectral factor model will be used for mainly two multivariate time series learning applications: First, real-world streaming datasets obtained from various processes are to be classified; in this exercise, human brain magnetoencephalography signals obtained during various cognitive and physical tasks are classified. Second, the commonalities are put to test by asking for reliable prediction of a multivariate time series given its past evolution; share prices in a portfolio are forecasted as part of this challenge.

For both spectral factor modeling and learning, an analytical solution as well as an iterative solution are developed. While the analytical solution is based on low-rank approximation of the spectral density function, the iterative solution is based on the expectation-maximization algorithm. For the human brain signal classification exercise, a strategy for comparing similarities between the commonalities for various classes of multivariate time series processes is developed. For the share price prediction problem, a vector autoregressive model whose parameters are enriched with the maximum likelihood commonalities is designed. In both these learning problems, the spectral factor model gives commendable performance with respect to competing approaches.

# Résumé

Les processus informatisés actuels génèrent des quantités massives de flux de données. Dans nombre d'applications, ces flux de données sont collectées en vue de modéliser les processus. Les modèles de processus obtenus ont pour but la réalisation d'objectifs tels que l'aide à la décision, la visualisation de données, l'informatique décisionnelle, l'automatisation et le contrôle, la reconnaissance de formes et la classification, etc. La modélisation de processus sur la base de données implique cependant de faire face à d'importants défis. Outre les erreurs, les données aberrantes et le bruit, le principal défi provient de la large dimensionnalité, i.e. du nombre de variables dans chaque échantillon de données mesurées. Les échantillons forment souvent une longue séquence temporelle appelée série temporelle multivariée, où chaque échantillon est influencé par les autres. Notre objectif est de construire un modèle robuste qui garantisse la génération, la révision et la représentation de nouvelles séries temporelles multivariées cohérentes avec le processus sous-jacent.

Dans cette thèse, nous adoptons un cadre de modélisation capable d'extraire, à partir de séries temporelles multivariées, des caractéristiques correspondant à des variations - covariations dynamiques communes aux variables mesurées dans tous les échantillons. Ces caractéristiques sont appelées «points communs» et une mesure qui leur est appropriée est définie. Ce qui rend le modèle de séries temporelles multivariées polyvalent est l'hypothèse relative à l'existence de séries temporelles latentes de caractéristiques connues ou présumées et de dimensionnalité beaucoup plus faible que les séries temporelles mesurées; le résultat est le bien connu «modèle factoriel dynamique». Des variantes originales de méthodes existantes pour estimer le modèle factoriel dynamique sont développées : l'estimation est réalisée en utilisant l'équivalent du modèle factoriel dynamique au niveau du domaine de fréquence, désigné comme le «modèle factoriel spectral». Pour estimer le modèle factoriel spectral, nous nous basons sur des idées relatives à la théorie des estimations spectrales. Cette théorie est utilisée pour aboutir à une formulation probabiliste, qui fournit des estimations de probabilité maximale pour les paramètres du modèle factoriel spectral. Des paramètres de probabilité maximale sont alors développés, en plaçant notre analyse entièrement dans le domaine spectral, de façon à ce que les séries temporelles latentes transformées dynamiquement héritent au maximum des points communs.

La principale contribution de cette thèse consiste en un cadre d'apprentissage utilisant le modèle factoriel spectral. Nous désignons par apprentissage la capacité d'un modèle de processus à caractériser de façon robuste les données générées par le processus à des fins de filtrage par motif, classification et prédiction. Dans ce contexte, le modèle factoriel spectral est considéré comme ayant appris une série temporelle multivariée si la série temporelle latente, une fois dynamiquement transformée, permet d'extraire les points communs de façon fiable et maximale. Le modèle factoriel spectral

sera utilisé principalement pour deux applications d'apprentissage de séries multivariées : en premier lieu, des ensembles de données sous forme de flux venant de différents processus du monde réel doivent être classifiés; lors de cet exercice, la classification porte sur des signaux magnétoencéphalographiques obtenus chez l'homme au cours de différentes tâches physiques et cognitives; en second lieu, les points communs obtenus sont testés en demandant une prédiction fiable d'une série temporelle multivariée étant donnée l'évolution passée; les prix d'un portefeuille d'actions sont prédits dans le cadre de ce défi.

À la fois pour la modélisation et pour l'apprentissage factoriel spectral, une solution analytique aussi bien qu'une solution itérative sont développées. Tandis que la solution analytique est basée sur une approximation de rang inférieur de la fonction de densité spectrale, la solution itérative est basée, quant à elle, sur l'algorithme de maximisation des attentes. Pour l'exercice de classification des signaux magnétoencéphalographiques humains, une stratégie de comparaison des similitudes entre les points communs des différentes classes de processus de séries temporelles multivariées est développée. Pour le problème de prédiction des prix des actions, un modèle vectoriel autorégressif dont les paramètres sont enrichis avec les points communs de probabilité maximale est conçu. Dans ces deux problèmes d'apprentissage, le modèle factoriel spectral atteint des performances louables en regard d'approches concurrentes.

# Chapter 1

# Introduction

The yearning to master a complicated process often prompts us to build its model. We wish to have a model that is simple but consistent with the characteristics of volumes of data obtained from the process. A model might serve several purposes: it could aid in representing and reviewing available data and to generate more data of similar characteristics. We also prefer the flexibility to evaluate the loyalty of the model to the characteristics of the given data and subsequently alter it if need be. Such a well-founded model should ultimately enable us to rein on the process.

The data typically comes as a set of samples and each sample is constituted by a set of **measured variables**. The characteristics of the measured variables might not be simple to comprehend. Hence, we wish the model to have a latent simplicity. To that end, we might conveniently demand the model to use a lower number of **latent variables** than the number of measured variables. Such a simplified interpretation of the process with a fewer number of underlying unobserved latent variables than the number of measured variables is called a **latent variable model** [11]. It is hoped that the data could be represented and reviewed with ease in terms of the latent variables.

In many applications, the set of measured variables of a sample is dependent on those of its preceding samples. The result is variation for a measured variable and covariation between the measured variables with respect to time and such data is called a **multivariate time series** [102]. The temporal variation-covariation across the measured variables of a multivariate time series is termed its **dynamic characteristics**.

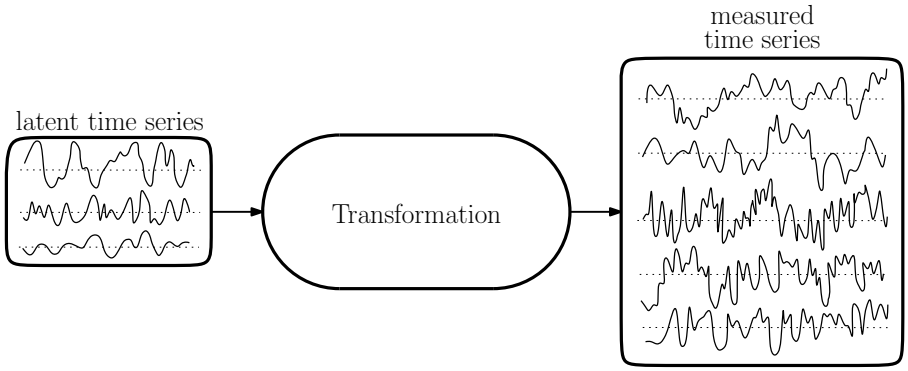


**Figure 1.1:** Three latent time series are dynamically transformed to five measured time series.

A model built on the dynamic characteristics of a time series is a reasonably flexible and accurate depiction of the underlying process.

Figure 1.1 illustrates a number of measured time series variables being generated by a transformation of a lower number of variables of the latent time series. It is this, possibly complicated, transformation that is to be modeled using the simplicity of the lower number of latent variables. Obviously, in doing so, we ought to be aware of the challenge that neither the transformation nor the number and characteristics of the latent variables are known. The modeling challenge is even greater because the lower number of variables might not be able to inherit the entire dynamic characteristics of the measured time series.

The scope of the problem of latent variable modeling of the dynamic characteristics of a multivariate time series is wide. It is natural, then, to restrict its scope as well as make it practically interesting. To this end, the dynamic characteristics common to any two measured variables [117] is deemed interesting for the modeling problem; those characteristics are termed the **commonalities**. In this thesis, commonalities will be defined in Definition 4.1 as the component cross-covariance functions of a weakly stationary multivariate time series.

For the still unknown model and latent time series, it is assumed that the latent variables are dynamically transformed to maximally inherit the commonalities of the measured time series according to some suitable metric. Thus, first we seek a **modeling framework** that defines the ingredients and scope of the latent variable model. The framework develops solutions for data-driven estimation of the **parameters** that control the transformation.

Apart from confirming our understanding of the process which generated the time series used to build the model, what could we do with a data-driven model? Suppose a model represents a collection of time series with similar commonalities. Then, the model could be used for **classification** of any new time series as belonging to that collection or not. This will be done based on similarities and dissimilarities of the commonalities of the new time series with those of the time series already available. Another utility could be in consistently generating future time series samples that bear characteristics similar to the time series we had used to build the model. Such a prospect might allow **prediction** of the time series given its past samples. It should be noticed that both of these applications involve applying the latent variable model towards unseen time series suggesting its ability to **learn**. Enabling the parameters of the transformation to predict and classify multivariate time series based on their commonalities implies a **learning framework** which is the main broad contribution of this thesis.

In Section 1.1 of this introductory chapter, a brief overview of the latent variable model with relevant references is given.

In Section 1.2, two practical examples to emphasize the motivation for a latent variable model with **dynamic transformation** are animated; these examples also form the experiments of the thesis. By these examples, interpretation of the commonalities in a multivariate time series as well as the learnability of a dynamic factor model are attempted. The basic assumptions that hold the latent variable model together and the basic strategy to arrive at a suitable model are listed. The technique of independent component analysis is complementary to the factor modeling pursued in this thesis; it is reviewed briefly.

In Section 1.3, the motivation for choosing the **dynamic factor model** as the latent

variable model is stated. Its structure, it is discussed there, will transform the latent time series to measured time series dynamically while the transformation maximally inherits the commonalities from the measured time series according to some suitable metric. Using its frequency-domain counterpart called the **spectral factor model** for modeling and learning purposes is then vouched for.

In Section 1.4, the modus-operandii of learning a multivariate time series based on the commonalities is elaborated in layman terms. The modeling framework is introduced there; it is mentioned there how it is intended to bin discrete frequencies in subbands and maximize the inheritance via dynamic transformation of the measured commonalities within each of those subbands. The learning framework is also introduced there; the strategies for prediction and classification of multivariate time series using the spectral factor model are illustrated.

In Section 1.5, a brief review of existing works elsewhere in the growing literature of multivariate time series analysis that have similar objectives as that of the spectral factor model is conducted. Existing methodologies from diverse fields such as control systems, econometrics, biomedical signal processing, geology, etc., that are related to the ones used during various stages in the development of spectral factor model in this thesis are recapped there.

In Section 1.6, a pithy statement of objectives and in Section 1.7 a summary list of the main and supporting contributions of the thesis are provided. In Section 1.8, the organization of the thesis is outlined while in Section 1.9 a list of publications that motivated and aided this thesis are listed. In Section 1.10, a very essential summary of this notation-rich thesis is presented.

## 1.1 Latent variable model: An overview

A model of a process with a set of underlying variables held responsible for generating or representing a set of measured variables is the basic notion behind the latent variable model. Among the premiers to voice this notion loudly was Spearman in [114] and a series of works that followed. His research in psychology argued that there exists a statistical quantity called the 'general factor' that remains same in the scores of all mental tests on humans; whereas there is a 'unique factor' that varies with the tests. This idea has evolved over a century. Today it is all too common to conduct such tests where discrete responses to questions in the form of personality statements are assumed to be expressions of latent personality traits [121]. There also exist problems pursued in the sciences where it is necessary to assume that latent variables are not a continuum but discrete or categorical in nature. They are mainly of two types: First, the mixture model involves associating measured samples to a finite set of latent variables by estimating probabilities of the associations [82]. And the second type is the latent class model which pursues discrete latent variables that when presumed known or available amounts to locally independent measured variables [81]; this could be treated as a special case of the mixture model.

What qualifies as a latent variable model in this thesis in light of the above possibilities is the one which maps continuous latent variables to continuous measured variables. However, the important requirement stipulated is that there ought to be very few of the former in comparison to the latter. By this requirement, as envisioned by [55], the hope is "to attain scientific parsimony or economy of description." Considerable research has progressed in this arena known as **factor modeling** with an aim

to explain correlations in the measured variables by a much lower number of latent variables: It was shown in [107] that three latent factors are generally sufficient in accounting for voltage variations, especially those relevant to electrocardiogram, recorded on the surface of a human body. In [69], yield rates of a large portfolio of stocks were shown to have a fewer number of latent factors corresponding to industry-wide common activities; whereas there were market factors unique to each of stock. Via factor analysis, six latent features out of twelve standard measured features were extracted for forecasting weather phenomenon in [9]. Factor models will be explored further in Chapter 3.

It will not escape our notice that in the seminal applications of factor models reviewed above, time dependency of the data was ignored for latent variable analysis. But this thesis focuses on the type of continuous latent variables that are to be modeled based on correlated samples in a multivariate time series. For the purpose of learning from such data, the classical factor model above will be insufficient and, instead, a **dynamic factor model** is required. Before entering into a detailed discussion on the dynamic factor model and its salient features in Section 1.3, the next section serves practical motivations for it.

## 1.2 Latent variable model: Two examples

In order to assert the context for a latent variable model for multivariate time series we discuss two practical examples below.

### Classification of brain activities

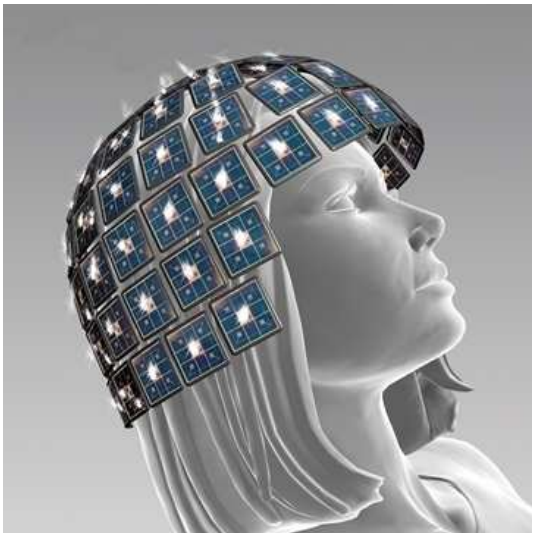

**Figure 1.2:** Illustration of the measuring of MEG signals via sensors positioned around the head [5].

Consider the scenario of a comfortably chaired computer gamer who makes smooth movements of a joystick by moving one of her wrists depending on the demands of a

game. Under same experimental conditions, it might be assumed that activities in her brain are similar every time she makes the same wrist movement. Suppose we wish to do some experiments to know what could be going on in her brain for every wrist movement she makes. Remember that such experiments are very common these days and international conferences and competitions are conducted to learn more about brain activities [3]. The biggest beneficiaries of such studies include patients of neurological disorders [105, 115].

For the experiments, magnetoencephalography (MEG) signals from a human brain could be measured. These signals are based on magnetic fields induced by currents due to synchronized neuronal activities. Their recording is non-invasively performed via extremely sensitive magnetic sensors, as depicted in Figure 1.2; in reality, the sensors of an MEG scanner are encased in a well-isolated cavity in which the head is positioned comfortably. The signals have a temporal resolution of under a millisecond [42] and methods are available to attribute the readings from the sensors to designated spatial spots of the brain. Suppose ten signals attributed to ten spatial spots of the brain are measured. We know that these signals depend on one another mutually, i.e., activities in one part of the brain are influenced by activities in other parts. Figure 1.3 shows real signals from one such experiment [1]. We could perhaps observe various types of similar characteristics among any two measured MEG signals, i.e., delayed or inverted patterns, similar peaks and troughs but with one signal more fluctuating than other, etc. As a result, these signals could be considered temporally dependent on one another, i.e., current brain activity at a spot is influenced by current and previous activities at all spots.

For making a wrist movement based on some prompt, hypothesize the existence of only two **latent** activities in the brain of the gamer. This hypothesis could be based on a subjective opinion of an expert or mere guess. What they neurologically are is not relevant here. Nevertheless, assume that these two latent activities to be, e.g., (i) her cognition of the demands of the game and (ii) her reactions to move her wrist. In addition, suppose the general characteristics, e.g., averages, ranges, and other statistics, of these two fictitious latent signals of cognition and reaction are known.

The assumptions made so far are, firstly, the existence of a set of low-dimensional latent signals and, secondly, that their statistical characteristics are known. In addition, thirdly, assume that when the gamer has to make a particular wrist movement, the presumed latent cognition and reaction sequences undergo a particular transformation that gets expressed as mutual and temporal dependence seen in the ten measured signals. Although this assumption compounds to limiting the characteristics of the measured sequences as well. But it is a fair assumption because there ought to be a number of time dependent characteristics common to the ten MEG signals which are part of the same brain that collectively results in her making a particular type of movement of the joystick or another. For this reason, it is opined that the latent signals of cognition and reaction manifest themselves as the ten measured MEG signals consisting of a large amount of common variation-covariation, i.e., **commonalities**, corresponding to her brain activities. Then, essentially, **cross-correlations between the measured variables equals commonalities**. Obviously, there will be variations of the signals unaccounted by the commonalities, which will be unique or **idiosyncratic** characteristics pertaining to each of the measured signals and independent of the commonalities. Hence, the gamer making a wrist movement may be regarded as, the fourth in the list of model assumptions, that the latent signals transforming them-

selves maximally imparting desired commonalities to the measured signals according to some suitable metric; whereas, the fifth assumption is that any unaccounted variations of the measured signals are just undesired and independent noise.

**Note 1.1.** *From a data generation perspective, transformation of latent variables imparts commonalities to measured variables. From a modeling perspective, transformation of latent variables inherits commonalities from measured variables.*

To summarize, the model assumptions are

1. there exist generative latent variables of lower-dimensionality than the measured variables,
2. the statistical characteristics of the latent variables are known,
3. the transformation of the latent signals limit the modeled characteristics of the measured variables,
4. the transformation should maximally impart measured cross-correlation characteristics, and
5. the non-transformable characteristics are independent noise unique to each measured variable.

The presumed characteristics of the cognition and reaction latent variables stay same throughout the game; the gaming conditions will stay the same but challenges will differ. Then, it could be inferred that the common characteristics of the MEG signals during one wrist movement switch to a different class if and only if she changes the wrist movement to another class. This is a valid inference because one part of the brain behaves differently from another to various cognition and reaction challenges of the game she is playing. As a result, any class differences of the movement will manifest in the dynamic characteristics of the measured MEG signals. So a particular class of characteristics of the measured signals during a particular class of wrist movements is attributed to a corresponding class of transformation the latent variables undergo in imparting the commonalities.

The objective of this experiment is modeling multivariate time series for classification of wrist movements. Transformations corresponding to each cognition-reaction challenge are to be estimated and one class of transformations from another are to be distinguished. An approach could be to estimate, from all possible transformations, one that maximizes the **likelihood** to have generated the measured signals. Then, the estimate could be constrained further by requiring the presumed latent signals to maximally inherit, according to some suitable metric, the commonalities of the measured signals upon their transformation. It is now clear that the two steps:

1. estimate a maximum likelihood transformation based on model assumptions and
2. estimate the maximum likelihood transformation that inherits commonalities maximally as per a suitable metric.

Suppose we estimate the optimal latent variable model of the cognition-reaction process corresponding to each classified example of wrist movements. Then, as shown in Figure 1.4, for two classes of example measured MEG signals, we should be able to **classify** a test measured signal as belonging to a class of movements by computing how similar the commonalities of the test measured signal are to those in the classified examples. Obviously, the intrigue lies in classifying the measured signals without actually knowing or seeing the particular wrist movement she had performed.

### Prediction of share prices

We take financial market as our next example where, suppose, the interest is in investing in a portfolio of shares of six companies, e.g., as shown in Figure 1.5, from various sectors of economic activities in a country. Suppose we know a successful investor who believes that investors are driven to purchase or sell shares based on perceived values of three underlying latent variables, viz., general political climate, consumer sentiments, and investor confidence. Of course, none of these fictitious latent variables could be metered objectively in practice. We wish to validate this belief before buying his advice. Note that as in the previous example, it is the number of latent variables and their presumed characteristics that is our concern and not their real physical or financial interpretations. If the investor's belief has merit, we could think of those latent variables to transform investment activities in the share market that manifest as changes in the share prices. Also, the latent variables when transformed must impart as much of the common dynamic characteristics, i.e., commonalities, demonstrated by the measured share prices.

In practice, even the best investors cannot consistently outsmart the market. And, our investor acquaintance above could blame any unexplainable fluctuations in the share prices on the dynamic characteristics of the share prices that the latent variables cannot inherit. These fluctuations could be **idiosyncratic** characteristics unique to each of those shares. However, if the transformation of the latent variable to the commonalities as we envisaged is true, we might be able to explain evolving tendencies of share prices. Therefore, in order to validate existence and influence of the commonalities, we could go by traditional investor wisdom to assess past behavior to bet on future: We could gather a **training** set of share prices of a sufficiently long evolution of various shares of the portfolio. We could then then estimate a **dynamic transformation** that is optimal in the sense of having the **maximum likelihood** to have generated the training series. Subsequently, we could search among the maximum likelihood transformations one that will **maximally inherit**, according to some suitable metric, the commonalities of the share price evolution process. We could use a predictor that is based on minimizing temporal tendencies to err in predicting the training series. The set of parameters of such a predictor will be a function of the optimal dynamic transformation. Then, given a current evolution of the share prices, it should be possible to **predict** their future evolution with a reasonable accuracy.

## Independent components versus latent factors

The thesis, as discussed so far, involves estimating a generative model where a set of latent variables are transformed to a larger number of measured variables based on the latter's characteristics. To estimate the transformation matrix, the maximal inher-

itance of the mutually dependent variation-covariation characteristics was the criterion considered.

In a complementary setting, there exists a wide body of literature called independent components analysis or blind source separation [27, 24]. Independent component analysis is often called 'non-Gaussian factor analysis' [61]. In contrast to the objective of factor analysis, the objective in independent components analysis is to identify mutually 'independent' latent variables.

One of its working philosophy is due to the central limit theorem whereby any transformation of the latent variables will be maximally non-Gaussian if it equals one of the independent latent variables; hence, latent variables are considered non-Gaussian [60]. In contrast, factor analysis stresses on dependencies and Gaussians are readily accepted as the latent variables.

In another working philosophy of the independent components analysis, higher predictability of a latent series component than that of any dynamic transformation of the latent series components is exploited to sequentially identify the latent variables [26]. In dynamic factor analysis as presented in this thesis, higher cross-correlations via commonalities aid predictability. On the other hand, in this thesis, the variation-covariation characteristics of the latent variables, their mutual dependence or independence, will be assumed known.

Moreover, in this thesis, the transformation of the latent variables will be assumed a linear process; therefore, the measured variables are also assumed linear processes. The focus in this thesis is in estimating a transformation for the latent variables rather than identifying the latent variables themselves as done in a blind source separation problem.

## 1.3 Dynamic and spectral factor models

As the two examples above highlight, the processes that are of our interest generate data samples such that each measured variable is free to influence the preceding samples of itself and other variables. This emphasizes that the order in the sequence of occurrences of the measured samples is rather important and it must be indexed appropriately. It is convenient to attribute the index of the sequence to discrete instants of time. This is the reason we call such a sequential collection of correlated data samples a **time series**.

In many processes we measure a set of variables at the same instant. This implies that every sample of the data is formed by the same ordered set of multiple variables. Such a collection of samples is referred to as **multivariate** data.

This thesis focuses on learning from **multivariate time series** where any measured variable in a data sample is influenced by, in general, the rest of the variables in the sample and all the variables of all the preceding samples. Such an influence could be quantified as a function of the **lag**, which is the number of time instants by which two samples differ. So, when a multivariate time series is said to display dynamic characteristics, the term **dynamic** attributes its characteristics to be **lag-dependent**.

In the context of multivariate time series, the driving assumption is that a lower number of latent variables are transformed to a number of measured variables resulting in a latent variable model as illustrated in Figure 1.1. A practical motivation for that assumption is that a fewer number of variables will aid simplicity in interpretation, mod-

eling, and computation. Note, however, that the true latent variable transformation is unknown and estimating it is part of the objective of this thesis.

Recall that the characteristics of a measured time series are to be modeled. But how could simplicity in modeling be aided when unknowns such as latent time series and variable transformation are injected into the model? In that respect, either or both the transformation and the characteristics of the latent time series could be assumed unknown. Remember, we wish to strictly control the underlying process which the latent time series represents and prefer it to have characteristics not as complicated as those of the measured time series. Moreover, if possible, expert opinion on the latent time series could be invited. Hence, it will be assumed that the latent variable characteristics are known and the transformation is unknown.

To enhance simplicity even further, the latent variables will be assumed a multivariate time series with lag-independent characteristics whereas it is the transformation that is dynamic and unknown. The challenge then is to estimate the dynamic transformation that best generates the measured time series from the latent time series. In this framework, the latent variables upon transformation are assumed to impart the dynamic characteristics to the measured time series. Hence, given a dataset of measured time series, such a framework implies estimating the ideal transformation that could yield the desired dynamic characteristics. This is illustrated in Figure 1.6, where the 'desired time series' is enabled to capture the desired dynamic characteristics pertaining to the measured time series; whereas the 'undesired time series' is the difference between the measured time series and the desired time series. The set of parameters $\theta$ of the dynamic transformation are retained for reference.

**Note 1.2.** *Figure 1.1 depicts the unknown true transformation that generates the measured time series from a latent time series of unknown characteristics. Whereas an appropriate dynamic transformation of Figure 1.6 has to be estimated based on the measured time series and the presumed characteristics of the latent time series.*

Remember that the desired dynamic characteristics of the measured time series are its commonalities. As introduced earlier and through the examples, maximally capturing the commonalities is tantamount to learning. It has been decided to keep the latent time series characteristics known, lag-independent, and simple; they are the underlying factors of the model. The model which consults the measured time series to dynamically transform the factors to maximize the commonalities is named the **dynamic factor model**. This concept is illustrated in Figure 1.7, where the term **idiosyncrasies** refers to the undesired time series that retains no commonalities. Hence, a dynamic factor model is a multivariate time series model which dynamically transforms a latent time series of predetermined characteristics to maximally, in some suitable metric sense, inherit the common dynamic characteristics of a set of measured multivariate time series. It accepts measured time series as input and outputs commonalities, idiosyncrasies, and the optimal model parameters.

One possible dynamic characteristic of the measured variables is periodicity. There could be many periodic dynamic characteristics in the measured time series. A periodicity corresponds to a **frequency**, which is associated with the number of time series samples that constitutes the period. Decomposing the measured time series into component frequencies is intuitively simple, analytically rich, and practically useful. Such a decomposition of a time series across all possible frequencies is called the **spectral**

**analysis** [99]. An inverse synthesis of frequency components to time-domain is also possible through spectral analysis. This is a motivation to understand the influence of various frequencies in the dynamic characteristics of the measured time series. As depicted in Figure 1.8, such a frequency spectral analysis of dynamic factor model would require analyzing measured and latent time series, commonalities and idiosyncrasies, and dynamic transformation all in the spectral or frequency-domain. It will be called the **spectral factor model**, which may be contrasted with the time-domain equivalent in Figure 1.7. In that respect, Figure 1.9 depicts the frequency spectral equivalent of the dynamic factor model. Note that Figure 1.9 has the same input and outputs as the dynamic factor model in Figure 1.7 for they are subjected to spectral analysis and its inverse, respectively.

## 1.4 Learning by maximizing spectral commonalities

The appeal of the frequency-domain approach in many fields of study are mainly due to the computational advantages and the physical interpretation it offers [97, 23]. Many time-domain processing requirements of a time series may be easily realized in the frequency-domain; the software and hardware implementation of such processing is widely available [63]. These further motivate, in addition to the theoretical appeal, the development of a spectral factor model for learning from multivariate time series.

The spectral components correspond to an infinite continuum of frequencies, but samples from a discrete time series are practically limited. This limits and motivates targeting just a set of discrete frequencies. But uncertainty is encountered in balancing resolution and precision of the spectral components at these discrete frequencies. To tackle the challenge, spectral components in small non-overlapping bands of frequencies may be considered. In these **frequency subbands**, spectral factor modeling might be performed by assigning probabilities to various discrete spectral components of the measured time series. The aim is to estimate a probabilistic spectral factor model that is the most likely to affiliate the measured spectral components. For this purpose, model parameters that will maximize the likelihood of simultaneous occurrences of all the measured spectral components within a subband will be probed. From all possible maximum likelihood spectral factor models, the one which maximally, in some suitable metric sense, inherits the measured commonalities on the dynamically transformed factors could be chosen. Recall that commonalities are cross-correlations of the measured variables. Later in the thesis, their inheritance by the dynamic factor transformation will be defined as a very simple and intuitive function of **all cross-correlations of the measured variables over all lags**.

Figure 1.10 illustrates the strategy for **maximum likelihood maximum commonalities** spectral factor model estimation. The spectral components of the presumed latent spectra and the given measured spectra are divided into frequency subbands. For each subband, maximum likelihood estimation of parameters of the spectral factor model will be performed. Two distinct maximum likelihood estimation methods will be demonstrated: The first method is an **analytical estimation** which gives an explicit formula for the optimal parameters. The second method is an **iterative estimation** starting with initial guesses of the model parameters that are updated till they converge to possible optimal parameters. Further, for each of those methods, techniques to extract those parameters that will maximize the commonalities are devised.

Commonalities of the measured time series maximally inherited in some suitable

metric sense by the dynamic factors allow the model to learn a process. **Classification** of multivariate time series measured from various distinct processes is the first of our two learning applications. In Section 1.2, the example of classification of MEG signals involved in maneuvering a joystick via wrist movements was discussed in detail. The various classes there could be regarded as dynamically transformed latent signals corresponding to various visual prompts on a computer monitor. There will be several example time series in a class. For each such example dataset obtained from any two processes deemed to be distinct by an expert, two classes of spectral factor model examples are built. The models are considered to have learned the example processes upon maximally inheriting their commonalities on their respective maximum likelihood parameters according to some suitable metric. Then, in order to decide which of any two possible processes a new unclassified measured time series belongs to, the commonalities of the new dataset need to compared with those of the two classes of spectral factor models. Based on the discussions so far, the commonalities will determine the dynamic transformation. In that regard, the new test measured time series will be assigned to the class to which its estimated dynamic transformation has the most proximity to. Such a strategy for the classification exercise requires a comparator of the dynamic transformations as shown in Figure 1.11. This method could be extended to associate a time series as belonging to one of any number of identified classes of processes.

**Prediction** of multivariate time series is chosen as the other learning application. Once knowledge of the characteristics and the number of latent time series variables are presumed, a spectral factor model based on a training set of measured time series could be estimated. Based on the optimal dynamic transformation that maximally, according to some suitable metric, inherits commonalities of the measured time series, a multivariate time series predictor could be built . The example of a portfolio of share prices that was discussed in Section 1.2 is used for prediction experiments later in the thesis. For a given length of training time series, a number of latent time series less than the number of the measured time series are experimented with to build the spectral factor models. Using their parameters, a prediction framework based on minimizing the prediction error given past samples is built. As shown in Figure 1.12, a future evolution could be charted for a given current evolution. The prediction accuracy will be validated using the true share prices whenever it becomes available.

## 1.5 Dynamic and spectral factor models in literature: A brief review

It must be mentioned at this juncture that the concept of commonalities and dynamic factor model is not very new. In one of the earliest formal studies about dynamic factor model, its estimation in the Fourier domain was famously attempted by [100] for advanced control systems and [104] for macroeconomic forecasting. An idea similar to commonalities was promoted by [104] in econometrics literature as "common shocks." Like in this thesis, they too state the relation between the spectral density functions via a likelihood function of the discrete Fourier transform components within disjoint frequency subbands. Then, they obtain maximum likelihood parameter estimates via Fletcher-Powell optimizations and standard hypothesis testing procedures. However, they stop short of going much farther than the possibility of infinitely many uncon-

strained solutions for the spectral factor transformation matrix.
Another line of approach is an approximate dynamic factor model with finite lags as was developed in [118]. There, the estimation was performed as the principal components of an expanded set of static factors; their aim was prediction of macroeconomic variables. Their prediction equations take the form of vector autoregression where the estimated static factor components may be directly plugged in without having to estimate the Fourier domain parameters as in fully dynamic models.
Recently, [36] changed the landscape of research in this domain substantially with their generalized dynamic factor model, e.g., they spell out the extent of flexibility allowed for idiosyncrasies and derived the convergence properties of the model parameters as the number of samples and measured variables grow. They focus on forecasting macroeconomic variables and the work forms a series of highly acclaimed and rigorous treatment of the subject. There are agreements between the parts of the approach to the problem in this thesis and theirs in (i) concluding that the principal components of the spectral density matrix gives the analytical solution (ii) the idiosyncrasies could be mildly cross-correlated. However, the ideas introduced in this thesis are quite different from theirs; e.g., an iterative estimation procedure and a time series classification strategy are provided. Moreover, while this thesis focuses in the multivariate time series modeling and learning frameworks, they focus on prediction of latent commonalities.
In §7.8 of [111] a maximum likelihood estimation estimation of dynamic factor model in the spectral domain much like in this thesis is pursued. They use it for analyzing function magnetic resonance imaging data. Their final analytical solution overlaps with the one developed in this thesis and in [36]. But they seem not to share any qualms regarding the non-analytical nature of the log-likelihood function and does not see such a model from the classification or prediction perspectives. They do not provide an iterative solution strategy either.
Among the front-runners of the dynamic factor model was [90] who wanted to estimate the latent trajectory of a patient's state based on vital signals. He rewrote the dynamic factor model parameters as a Markovian state model whose estimation was carried out via Kalman filter principles.
Now, let us divert the attention to a spectral domain method whose priority was multivariate time series classification rather than prediction. In [66], sample spectral densities are compared for classifying and clustering episodes of multivariate time series. Their experiments involved discriminating between time series generated by earthquakes and those by explosions. However, they do not consider existence of a low-dimensional latent time series and, as a result, were able to design disparity measures that work by comparing the full-rank sample spectral densities. This thesis uses the information contained in a rank-deficient maximum-likelihood maximum-commonalities spectral factor transformation matrix to perform classification.

## 1.6 Objectives

Based on discussions on the motivation and the premise of this thesis so far, its objectives are broadly divided into developing

1. *a multivariate time series latent variable modeling framework*
   To meet this objective, dynamic and spectral factor models as well as commonalities are formally introduced and defined in Chapter 4. The maximum likelihood

maximum commonalities spectral factor model is derived in Chapter 5. An analytical form as well as an iterative procedure for estimating such a spectral factor model are developed.

2. *a multivariate time series maximum commonalities learning framework*
   This objective is achieved by providing multivariate time series classification and prediction algorithms in Chapter 6, which exploit the maximum commonalities parameters of the spectral factor model.

## 1.7 Contributions

The following is the list of main contributions of this thesis:

- ▷ The most original contribution of this thesis is the development of a commonalities-based classification metric in (6.4) that compares overlap of spectral factor model subspaces to distinguish multivariate time series processes.
- ▷ The second most important contribution is the utilization of the estimated commonalities in developing a multivariate time series prediction strategy via classical vector autoregression on current and past samples; it is detailed in Section 6.3.

The following is the list of supporting contributions of this thesis, which are improvements, interpretations, or alternatives to existing work in the literature:

- ▷ Derived an analytical solution for spectral factor model in (5.10) using low-rank approximation theorem.
- ▷ Derived an iterative solution for spectral factor model in Section 5.2 using the Expectation - Maximization algorithm whose converged parameters that maximally inherit the commonalities are extracted by applying the Gauss - Markov theorem in Section 5.2.3.
- ▷ Obtained the mild cross-correlation property of the idiosyncrasies in Property 5.1 via Weyl's theorem.
- ▷ Used Wirtinger relaxations for maximizing log-likelihood in Chapter 5.

## 1.8 Organization

A non-technical overview of the thesis was presented so far. In the two chapters that follow, the basics on which this thesis is built is presented.

- In Chapter 2, an essential overview of multivariate time series analysis is provided; very essential time-domain and frequency-domain analyses are presented there.
- In Chapter 3, parametric estimation methods for probabilistic models concisely and as required is discussed.

With much groundwork done with the aforementioned chapters, the two chapters that follow introduce and develop the dynamic factor model framework to suit the learning framework objective of this thesis.

- In Chapter 4, a technical introduction and motivation for the concepts of dynamic and spectral factor models as well as commonalities and their maximization are provided.

- In Chapter 5, an analytical method and an iterative method for maximum likelihood maximum commonalities spectral factor model are derived.

Subsequent to the development of the dynamic factor model, the learning framework is provided.

- In Chapter 6, a time series learning framework is built using the inherited commonalities by explicitly stating algorithms for classification and prediction of multivariate time series analysis.

The contributions are tested and possible extensions are discussed in the last two chapters:

- In Chapter 7, the methodology and results of multivariate time series classification and prediction experiments are presented.

- In Chapter 8, improvements and plans for further research and applications are mentioned.

## 1.9 List of relevant publications

(I) Miranda, A. A., Olsen, C., Bontempi, G.: Fourier spectral factor model for prediction of multidimensional signals, *Signal Processing*, 91(9):2172-2177, Elsevier, 2011.
This paper presents the vector autoregressive prediction of a multivariate measured time series on the current and past samples as developed in Section 6.3 using the autocovariance of the maximally inherited commonalities. It demonstrates prediction of yield rate of a six-variate share portfolio with substantially better accuracy than standard vector autoregression; the daily prices of those yield rates are used for experiments in this thesis.

(II) Miranda, A. A., Bontempi, G., Schuddinck, P.: Fourier spectral factor model for classification of high-dimensional MEG signals, *Under review in Biomedical Signal Processing and Control*, Elsevier, 2011.
This paper presents the commonalities-driven classification strategy developed in Section 6.2 for multivariate time series; the magnetoencephalography experiments conducted for Section 7.1 are also presented.

(III) Miranda A. A, Caelen O., Bontempi, G.: Machine learning for automated polyp detection in computed tomography colonography, *Biomedical Image Analysis and Machine Learning Technologies*, Medical Information Science Reference, 2009.
This paper compares a number of classifiers well-known in machine learning that perform well despite severe imbalance in the class representation and unreliable features. That classification problem may be compared with that in Section 6.2 to understand that popular robust classifiers designed towards identically and independently distributed data are not directly usable for a multivariate time series classification problem.

(IV) Miranda, A. A., Le Borgne, Y.-A., Bontempi, G.: New routes from minimal approximation error to principal components, *Neural Processing Letters*, 27(3):197-207, Springer, 2008.
This well-cited paper discusses the classical principal components analysis from a layman perspective. Principal subspaces, eigenvalue decomposition, trace minimization are recurrent themes in this thesis and are presented in simple terms in the paper.

(V) Miranda, A. A., Whelan, P. F.: Fukunaga-Koontz transform for small sample size problems, *Proceedings of the IEE Irish Signals and Systems Conference*, pp. 156-161, Dublin (2005)
This paper discusses a strategy for comparing the principal subspaces due to the autocorrelation matrices of two classes of multivariate data in a common full-rank space. The features of this paper such as real-valued projections, euclidean distance measures, binary classification, etc., are serious shortcomings for comparing multiple spectral factor subspaces and to overcome them the classification metric in (6.4) was developed.

## 1.10 Notations

Herein, notations and conventions used in this thesis are introduced. Unfortunatelely, terms whose proper definitions will show up in later chapters only will be mentioned here. Nevertheless, it is important to read this section carefully for grasping the treatment of technical aspects later.

The following convention of using Latin characters is adhered to: Incremental variables such as indices are denoted using $i$, $j$, $k$, and $l$.

**Note 1.3.** *From Chapter 4 onwards, certain alphabets are appointed to imply the same variable for the rest of the thesis. These are, respectively, $q$ and $r$ for the latent dimensionality and observed dimensionality. The letters $v$, $x$, $y$, and $z$ are used for transformed, latent, measured, and idiosyncratic variables; but they will have an appropriate meaning depending on whether it appears in Roman, sans-serif or boldface fonts.*

Use of $t$ for time indices and $h$ for time delays are reserved throughout. The aforementioned conventions imply that both scalars and vectors are denoted in small-case. Linear algebra drives much of the contributions and a rectangular matrix is always in capital-case as in $X$.

Ideas from the basics of probability and stochastic processes are used liberally. A sans-serif font such as in $\mathsf{x}$ is used to denote a random variable and its realization will be in Roman font as in $x$.

**Note 1.4.** *Random variables and vector random variables, either real-valued or 'complex-valued', will be denoted in the same fashion using a sans-serif font; the context will make their distinction clear.*

Also, the sans-serif font will be used to denote common mathematical operations or functions such as $\log$ for natural logarithm, $\mathsf{p}$ for a probability density function, $\mathsf{S}$ for spectral density function, etc.

The standard practice of using a blackboard bold font to denote number sets, e.g. set of complex numbers $\mathbb{C}$, set of integers $\mathbb{Z}$, etc. are followed. However, a calligraphic font will be used to denote a group of items such as two classes $\mathcal{C}_1$ and $\mathcal{C}_2$ and the Gaussian family of probability densities $\mathcal{N}$.

Certain Greek alphabets will denote the same variable, function, or metric throughout, e.g., $\mu$ for mean and $\Gamma$ for autocovariance function matrix.

Subscripts are used for indices in two capacities: First, they denote indices as in $x_t$ for the $t$-th time sample of $x$. Second, they denote a component of a vector or a matrix. E.g., $\mathsf{x}_k$ is the $k$-th component of random variable $\mathsf{x}$ and $X_{ij}$ is the $i,j$-th element of matrix $X$. This gives the possibility to interpret nested subscripts appropriately. E.g., $x_{k_t}$ is the $t$-th time sample of $x_k$ and the inner subscript, i.e., $k$ in $x_{k_t}$, will be always interpreted as the component index and the outer subscript, i.e., $t$ in $x_{k_t}$, as the sequence index.

**Note 1.5.** *Other than its usual interpretation as scalar exponent, superscript on a function or an operator will denote the operand, e.g., $\mu^{\mathsf{y}}$ denotes mean of the random variable* $\mathsf{y}$.

Fourier analysis is a persistent theme in this thesis and boldface, e.g., $\mathbf{x}(\omega_k)$, implies discrete Fourier transform components.

Presented below is a table of certain frequently used symbols and notations:

| | |
|---|---|
| $u'$, $U'$ | transpose of vector $u$ or matrix $U$ |
| $\bar{u}$, $\overline{U}$ | complex conjugate of scalar or vector $u$ or matrix $U$ |
| $u^*$, $U^*$ | complex conjugate transpose of scalar or vector $u$ or matrix $U$ |
| $\|u\|$, $\|\mathcal{U}\|$ | absolute value of scalar $u$; cardinality of set $\mathcal{U}$ |
| $U_{i:j}$ | matrix formed by columns $i, i+1, \ldots, j-1, j$ of matrix $U$ |
| $\det(U)$ | determinant of real or complex-valued square matrix $U$ |
| $\{\mathsf{u}_t\}$ | time series due to sequence of random variables $\mathsf{u}_t$ $\forall t \in \mathbb{Z}$ |
| $u_t$ | realization of a time series $\{\mathsf{u}_t\}$ at instant $t$ |
| $\mathbf{u}(\omega_j)$ | discrete Fourier transform due to $\{u_t\}$ at frequency $\omega_j$ |
| $\mathsf{P}(\mathcal{U})$ | probability of the event $\mathcal{U}$ |
| $\mathsf{p}^{\mathsf{u}}$ | probability density of (a possibly vector) random variable u |
| $p$ | order of vector autoregression |
| $\mathsf{E}^{\mathsf{u}}$ | expectation with respect to $\mathsf{p}^{\mathsf{u}}$ |
| $\mu^{\mathsf{u}}$ | mean of (vector) random variable u |
| $\Gamma^{\mathsf{u}}$ | variance (covariance matrix) of (vector) random variable u |
| $\Gamma^{\mathsf{u},\mathsf{v}}$ | cross-covariance (matrix) of (vector) random variables u and v |
| *acvf* | autocovariance function |
| $\gamma_h^{\mathsf{u}}$ | *acvf* of univariate $\{\mathsf{u}_t\}$ at lag $h$ |
| $\Gamma_h^{\mathsf{u}}$ | *acvf* of (univariate or multivariate) $\{\mathsf{u}_t\}$ at lag $h$ |
| $\mathsf{i}$ | imaginary operator |
| $I_q$ | identity matrix of size $q \times q$ |
| $\widetilde{\mathsf{diag}}(U)$ | setting off-diagonal elements of $U$ to zero |
| $\|U\|_F$ | Frobenius norm of matrix $U$ |
| $\mathcal{F}$ | Fourier transformation; discrete Fourier transform |
| iid | independently and identically distributed |
| $\langle x \rangle$ | *A posteriori* mean of x |
| L | log-likelihood function |
| $\mathcal{D}$ | a dataset |
| $\tau$ | length of a time series realization |
| $\hat{\jmath}$ | number of frequency subbands |
| $\kappa$ | number of relevant nearest neighbors |
| $W$ | dynamic factor transformation matrix |
| $\mathbf{W}$ | spectral factor transformation matrix |
| $\widehat{\mathbf{W}}$ | maximum likelihood $\mathbf{W}$ |
| $\widetilde{\mathbf{W}}$ | maximum commonalities $\widehat{\mathbf{W}}$ |
| $\mathsf{S}^{\mathsf{u}}$ | spectral density function of $\{\mathsf{u}_t\}$ |
| $\check{\mathsf{S}}^{\mathsf{u}}$ | sample spectral density function of $\{\mathsf{u}_t\}$ |
| $\widehat{\mathsf{S}}^{\mathsf{u}}$ | maximum likelihood $\mathsf{S}^{\mathsf{u}}$ |
| $\widetilde{\mathsf{S}}^{\mathsf{u}}$ | maximum commonalities $\widehat{\mathsf{S}}^{\mathsf{u}}$ |

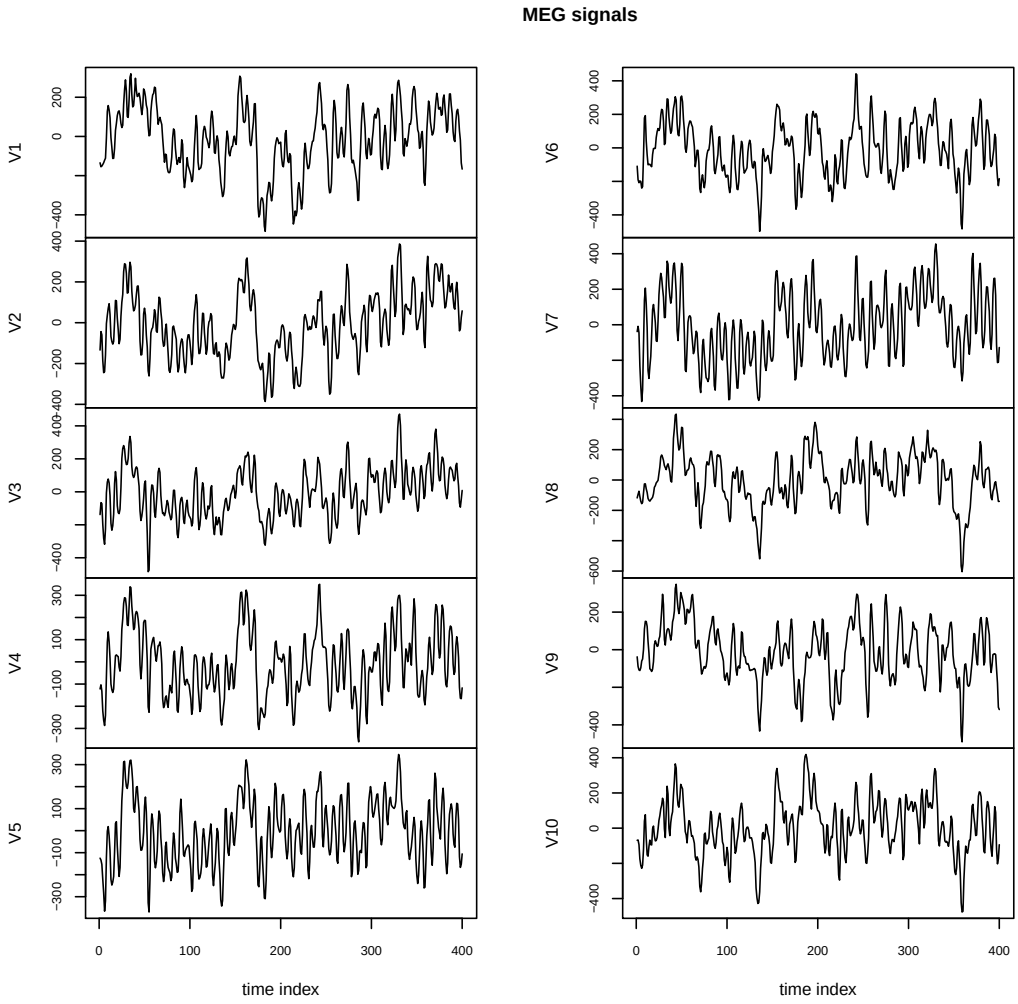


**Figure 1.3:** MEG signals corresponding to ten spatial spots V1-V10 of the brain upon a particular movement of the wrist.

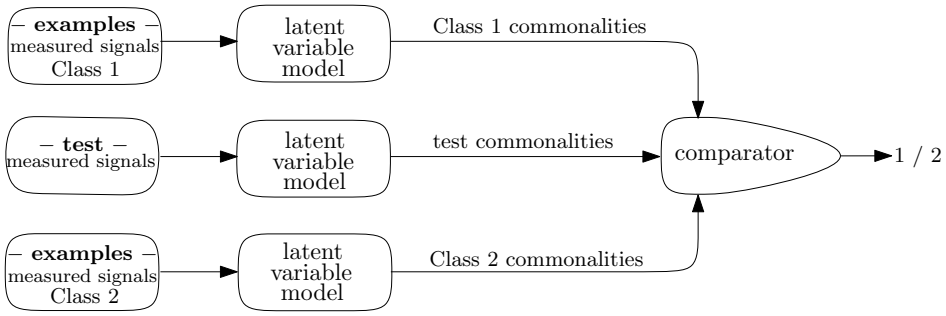


**Figure 1.4:** Among the two classes, a test measured signal is associated to the one to which its commonalities are closest to.

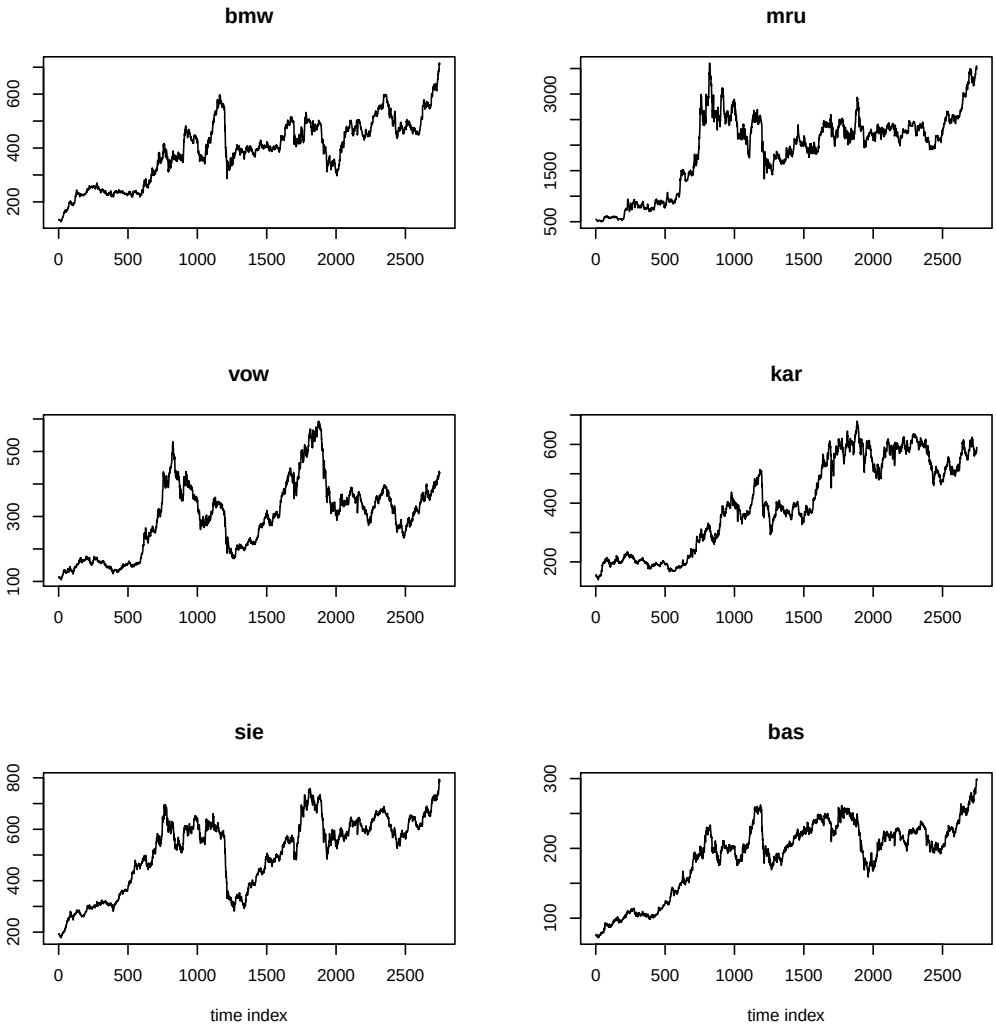


**Figure 1.5:** Daily stock prices in Deutsche Mark of six German companies between 01/01/1983 - 30/12/1993 [6].

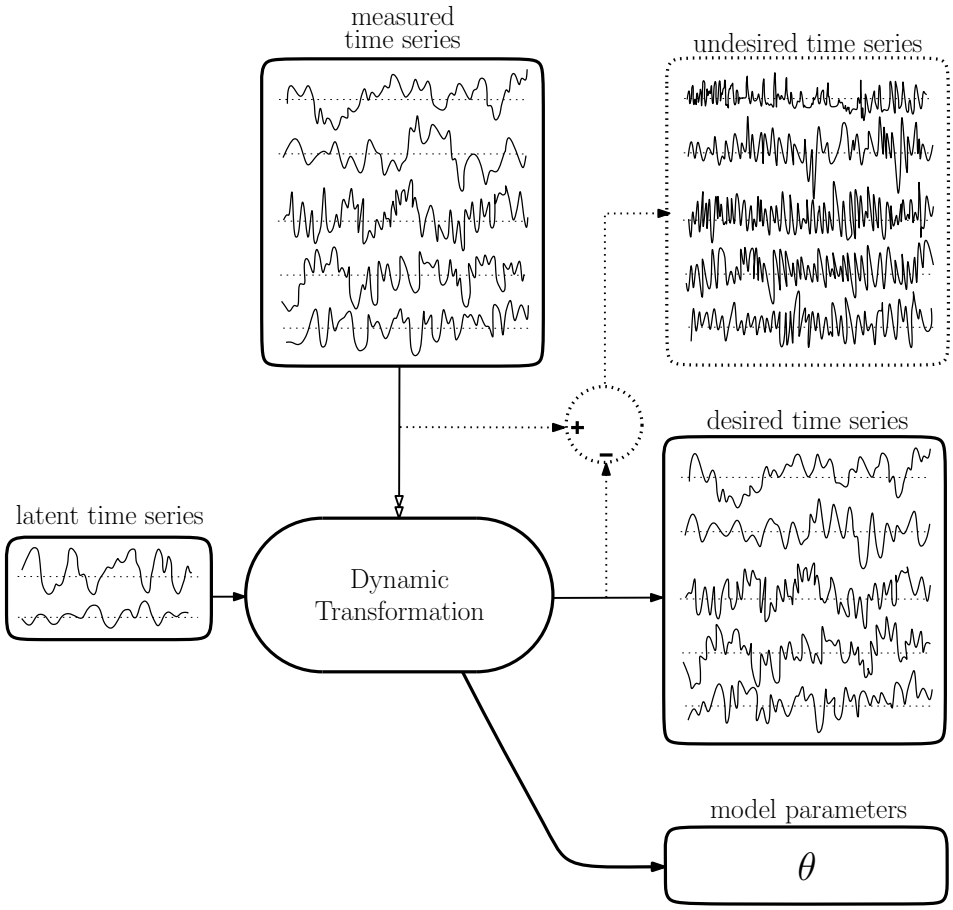


**Figure 1.6:** Dynamic transformation, whose parameters are summarized by $\theta$, of the latent time series will consult the measured time series to decompose the latter into desired and undesired time series.

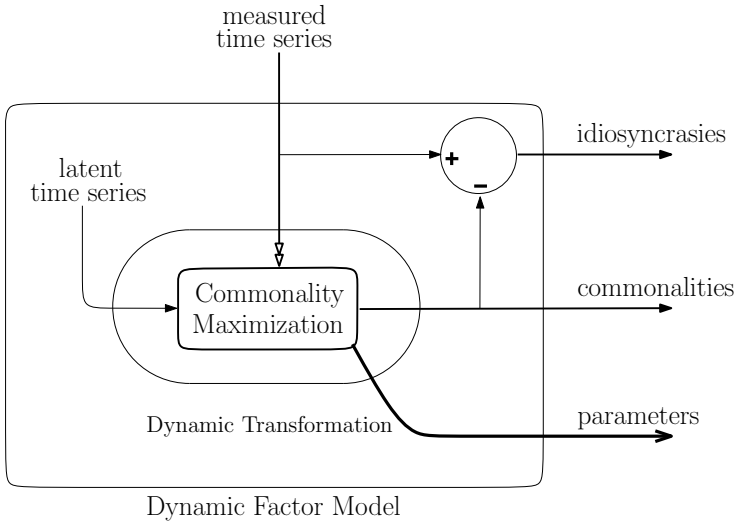


**Figure 1.7:** According to some suitable metric, the dynamic factor model allows the dynamic transformation to maximally inherit the commonalities from the measured time series; their difference forms the idiosyncratic time series.

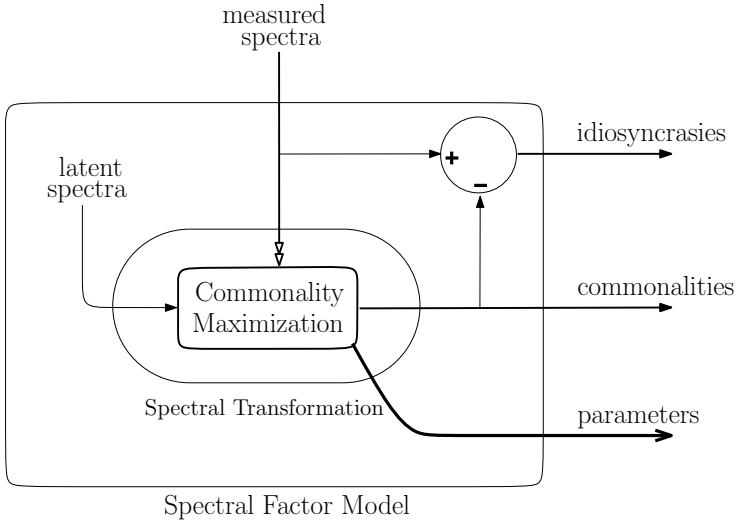


**Figure 1.8:** The spectral factor model expresses dynamic factor model in frequency-domain.

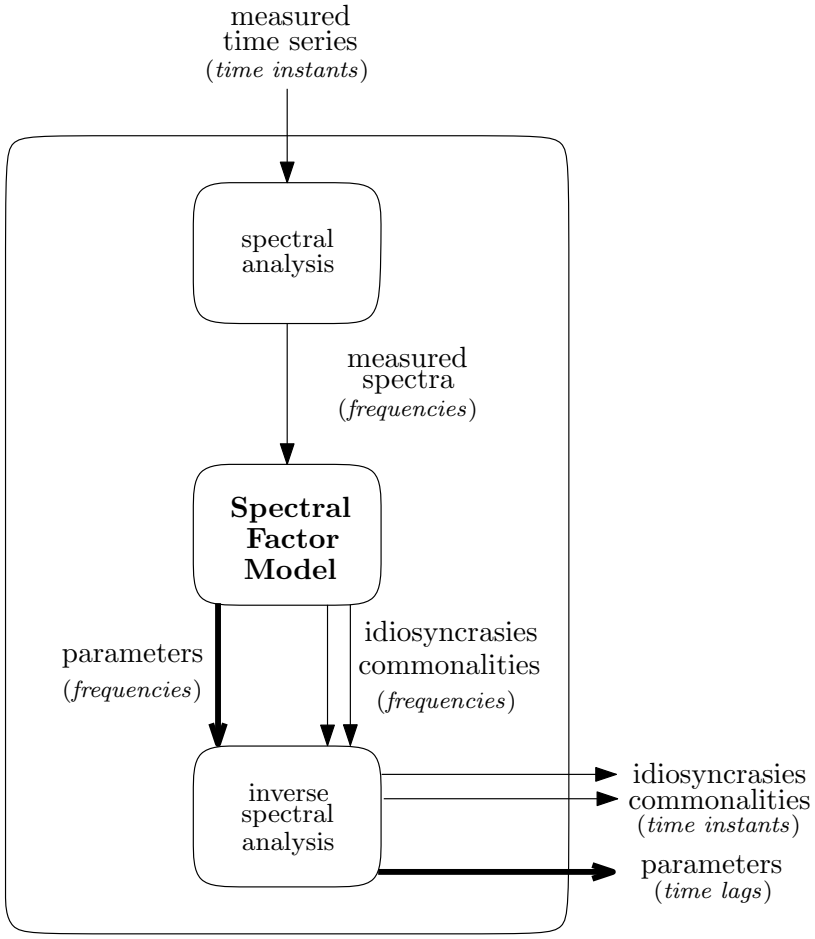


**Figure 1.9:** An equivalent of the dynamic factor model is built by sandwiching the spectral factor model between spectral analysis and its inverse operations.

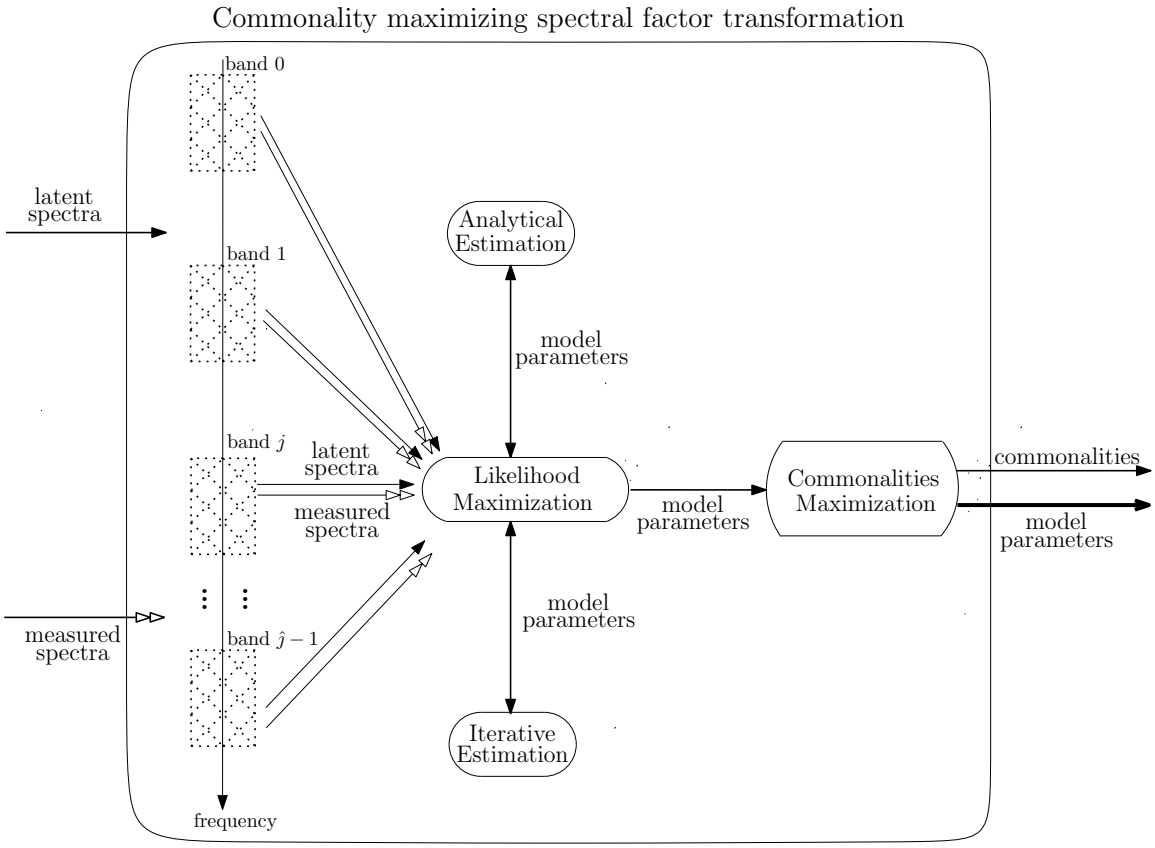


**Figure 1.10:** Maximized commonalities for a finite $\hat{\jmath}$ number of individual frequency bands are obtained from amongst the family of maximum likelihood spectral factor model parameters analytically and iteratively.

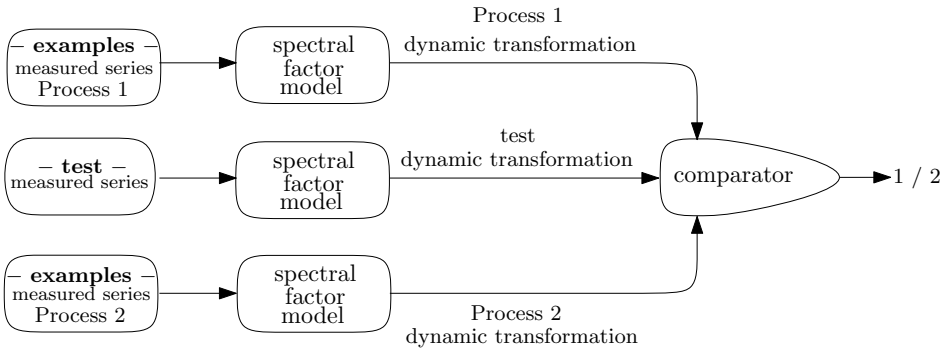


**Figure 1.11:** A test time series is associated to a class of time series if that class has the closest proximity, in terms of the commonalities of its examples, among all classes to the test time series.

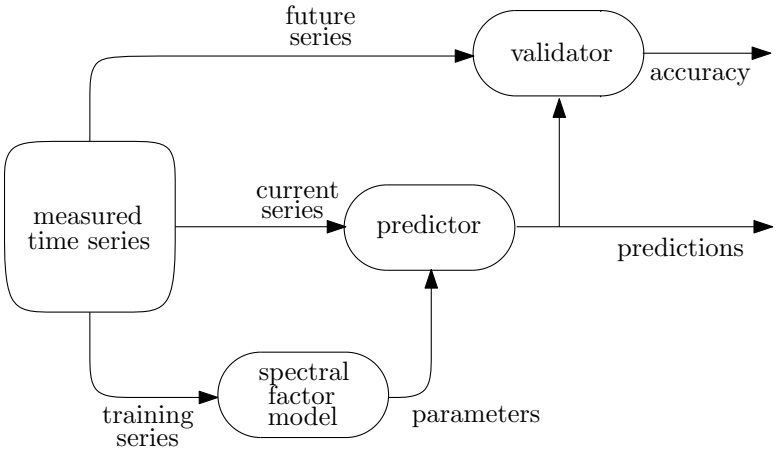


**Figure 1.12:** A predictor for the measured time series is built using parameters pertaining to maximally inherited commonalities of a training series. Accuracy of predictions based on current samples of the measured time series as evidence is compared with its future samples.

# Chapter 2

# Multivariate time series analysis: Some essential notions

An overview of a modeling and learning framework for multivariate time series was presented in Chapter 1. In this chapter, some notions on **multivariate time series** analysis in time and frequency domains are succinctly introduced; tools and conventions used herein are essential to appreciate the contributions later in the thesis. Although they are widely available in textbooks, they have been adapted appropriately to suit this thesis.

In Section 2.1, a multivariate time series model and the concept of **weak stationarity** are formally defined; only those time series that are weakly stationary are considered throughout this thesis. Weak stationarity requires defining the **autocovariance function** of a time series. Autocovariance characteristics of a few relevant types of stationary multivariate time series are presented.

In Section 2.2, frequency or spectral domain concepts belonging to Fourier analysis are introduced. The motion of a **simple pendulum** is used as an example to motivate the presentation of **Fourier series**; this is subsequently extended to the **Fourier transform**. Clearly, this example to introduce Fourier analysis is a detour to a **continuous time process**; but it will enhance understanding of spectral domain tools, notations, and definitions.

In Section 2.3, the discrete time process is introduced as a limiting case of continuous time processes; this leads to **discrete time Fourier transform**. Discrete time Fourier transform gives a periodic and continuous spectrum and it underpins important developments in subsequent sections. There, **discrete Fourier transform** of a **discrete time process** is also discussed.

In Section 2.4, after having defined time and spectral domain characteristics of a deterministic process, spectral analysis of stationary time series is presented. The two most important ideas that need to be taken from this chapter are presented next: First, the relation between autocovariance function and **spectral density function** is simply that of a Fourier transform. Second, the probability distribution of discrete Fourier transform components of a linear process is **complex-Gaussian** within small **subbands** of frequencies. The first idea is a direct application of Fourier analysis derived in earlier sections. For the second idea, the asymptotic theory of spectral

estimates is involved. In Section 2.5, therefore, the asymptotic probability distribution function of discrete Fourier transform components is provided without proof.

## 2.1 Temporal analysis of stationary processes

In this section, an introductory review of the time-domain or temporal analysis of time series is performed. It starts by adapting some definitions from the literature of random processes [29, 68, 86, 47]; the presented definitions might be termed differently by various authors elsewhere in the literature.

An infinite sequence of random variables forms a random process. Then, a **vector random process** is defined as an infinite sequence of vector random variables that is a set of random variables maintaining the same order of the vector components in every realization.

**Definition 2.1.** *A (**multivariate**) **time series** of a (vector) random process is a connected subsequence formed by its constituent (vector) random variables.*

A time series is called so because the index of the sequence is often attributed to time instants. If $\mathsf{y}_t$ is the random variable at time instant $t \in \mathbb{Z}$ of a random process of interest, then $\{y_t\}$, $t = 1, \ldots, \tau$ shall be called a $\tau$-length realization of a time series whose $t$-th **sample** is $y_t$. An $r$-dimensional vector random process generates an infinite sequence $\{\mathsf{y}_t\}$ of $r$-dimensional vector random variables $\mathsf{y}_t = [\mathsf{y}_{1_t} \cdots \mathsf{y}_{r_t}]'$, where $\mathsf{y}_{i_t}$, $i = 1, \ldots, r$ are the component random variables at time instant $t \in \mathbb{Z}$. Figure 2.1 selects a realization of a $\tau$-length $r$-variate time series whose $t$-th sample is the vector $y_t \in \mathbb{R}^r$.

It may now be implied that when referring to the term 'process' in Chapter 1 in a broad sense, it meant the vector random process underlying a multivariate time series. On similar lines, the term 'model' there referred to the joint probability distribution function of the samples of the time series; this is because it is a set of random variables that is dealt with. Then, the model of a process corresponding to a $\tau$-length $r$-variate time series requires evaluating the $\tau \times r$-dimensional joint probability distribution function $\mathsf{P}(\mathsf{y}_1 \leq c_1, \ \ldots, \ \mathsf{y}_\tau \leq c_\tau)$ for any constant vector $c_t \in \mathbb{R}^r$, $t = 1, \ldots, \tau$, where P denotes probability and the comparison of vectors are component-wise. Of course, direct evaluation of such a probability distribution is very unwieldy. Therefore, restricting the scope of the studies and bringing forth assumptions to simplify the process is inevitable for modeling a process generating a multivariate time series.

Let a few useful terms associated with random variables be first defined [95].

**Definition 2.2.** *The **probability density function** $\mathsf{p}^{\mathsf{u}}$ of a random variable $\mathsf{u}$ is defined as $\mathsf{p}^{\mathsf{u}}(a) = \frac{\mathsf{d}}{\mathsf{d}a}\mathsf{P}(\mathsf{u} \leq a)$ $\forall a \in \mathbb{R}$, wherever the derivative exists.*

In the above definition, $\mathsf{p}^{\mathsf{u}}(a)$ is any positive finite real number wherever the derivative does not exist. Then the joint probability density function $\mathsf{p}^{\mathsf{u}_1,\ldots,\mathsf{u}_r}$ of $r$ random variables $\mathsf{u}_1, \ldots, \mathsf{u}_r$ may be given by $\mathsf{p}^{\mathsf{u}_1,\ldots,\mathsf{u}_r}(a_1, \ldots, a_r) = \partial^r \mathsf{P}(\mathsf{u}_1 \leq a_1, \ \ldots, \ \mathsf{u}_r \leq a_r) / \partial a_1 \cdots \partial a_r$ $\forall a = [a_1 \ \cdots \ a_r]' \in \mathbb{R}^r$ and $\mathsf{p}^{\mathsf{u}_1,\ldots,\mathsf{u}_r}(a_1, \ldots, a_r)$ is any positive finite real number wherever the derivative does not exist.

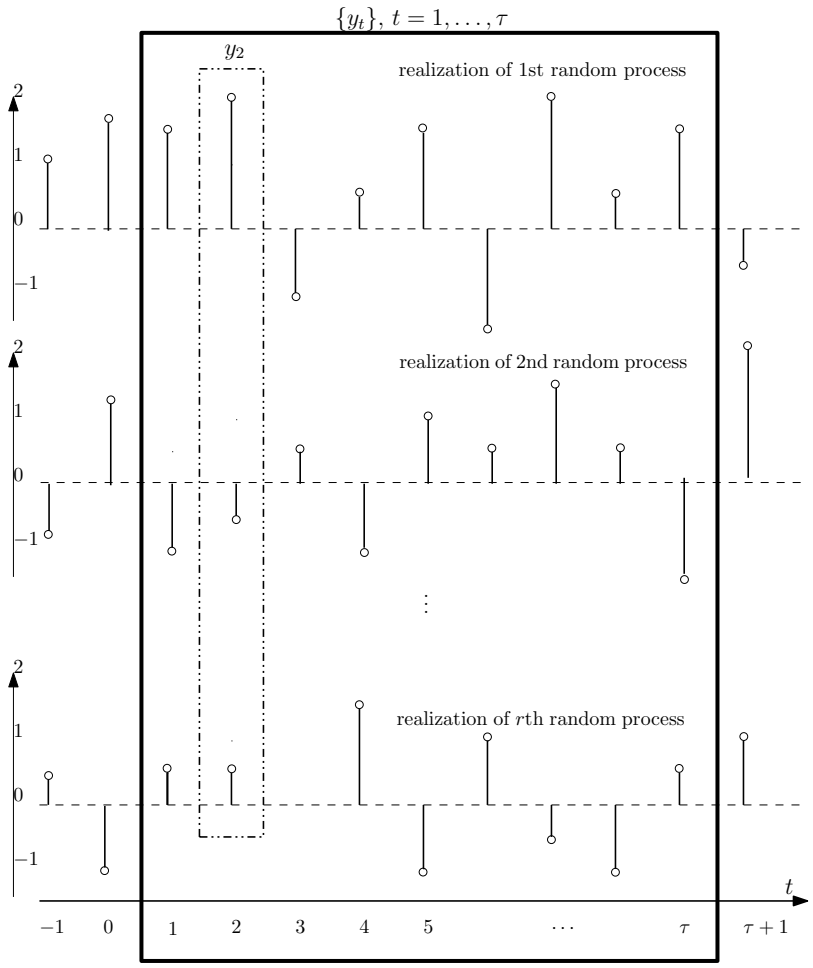


**Figure 2.1:** The second sample $y_2$ of a realization of a $\tau$-length $r$-variate time series $\{\mathsf{y}_t\}$ is highlighted.

**Definition 2.3.** *The **multivariate probability density function** $\mathsf{p}^{\mathsf{u}}$ of an $r$-dimensional vector random variable $\mathsf{u} = [\mathsf{u}_1 \ \cdots \ \mathsf{u}_r]'$ is defined as the joint probability density function of its $r$ component random variables, i.e., $\mathsf{p}^{\mathsf{u}}(a) = \mathsf{p}^{\mathsf{u}_1,\ldots,\mathsf{u}_r}(a_1, \ldots, a_r) \ \forall a = [a_1 \cdots a_r]' \in \mathbb{R}^r$.*

**Definition 2.4.** *For an $r$-dimensional vector random variable $\mathsf{u}$ whose probability density function $\mathsf{p}^{\mathsf{u}}(b)$ exists $\forall b \in \mathbb{R}^r$, the **expectation** of a function $g(\mathsf{u})$ with respect to $\mathsf{p}^{\mathsf{u}}$ is defined as $\mathsf{E}^{\mathsf{u}}[g(\mathsf{u})] = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(b) \ \mathsf{p}^{\mathsf{u}}(b) \ \mathrm{d}b$.*

**Definition 2.5.** *The **mean** $\mu^{\mathsf{u}} \in \mathbb{R}^r$ of an $r$-dimensional vector random variable $\mathsf{u}$ is defined as its expectation with respect to its $r$-variate probability density function, i.e., $\mu^{\mathsf{u}} = \mathsf{E}^{\mathsf{u}}[\mathsf{u}]$.*

**Definition 2.6.** *The **variance (covariance matrix)** $\Gamma^{\mathsf{u}}$ of the ($r$- dimensional vector) random variable* $\mathsf{u}$ *is defined as the expectation, with respect to its (multivariate) probability density function* $\mathsf{p}^{\mathsf{u}}$*, of the (outer) product of the (vector) random variable with itself about its mean $\mu^{\mathsf{u}}$, i.e., $\Gamma^{\mathsf{u}} = \mathsf{E}^{\mathsf{u}}\,[(\mathsf{u} - \mu^{\mathsf{u}})\,(\mathsf{u} - \mu^{\mathsf{u}})']$.*

**Definition 2.7.** *The **cross-covariance** $\Gamma^{\mathsf{u},\mathsf{v}}$ between the (vector) random variables* $\mathsf{u}$ *and* $\mathsf{v}$ *is defined as the expectation, with respect to their joint (multivariate) probability density function* $\mathsf{p}^{\mathsf{u},\mathsf{v}}$*, of the (outer) product of the random variables about their respective means $\mu^{\mathsf{u}}$ and $\mu^{\mathsf{v}}$, i.e., $\Gamma^{\mathsf{u},\mathsf{v}} = \mathsf{E}^{\mathsf{u},\mathsf{v}}\,[(\mathsf{u} - \mu^{\mathsf{u}})\,(\mathsf{v} - \mu^{\mathsf{v}})']$.*

**Definition 2.8.** *The cross-covariance between any two constituent (vector) random variables of the same (vector) random process is called the **autocovariance function** between the (vector) random variables.*

Therefore, the autocovariance function (**acvf**) between the $r$- dimensional vector random variables $\mathsf{y}_t$ and $\mathsf{y}_s$ $\forall t, s \in \mathbb{Z}$ is

$$\Gamma^{\mathsf{y}_t,\mathsf{y}_s} = \mathsf{E}^{\mathsf{y}_t,\mathsf{y}_s}[(\mathsf{y}_t - \mu^{\mathsf{y}_t})(\mathsf{y}_s - \mu^{\mathsf{y}_s})']. \tag{2.1}$$

**Note 2.1.** *When it specifically concerns a univariate time series $\{\mathsf{y}_t\}$ and not a multivariate time series, its acvf will be denoted by $\gamma^{\mathsf{y}_t,\mathsf{y}_s}$. Then, for a multivariate time series $\{\mathsf{y}_t\} = \{[\mathsf{y}_{1_t}\,\mathsf{y}_{2_t} \cdots \mathsf{y}_{r_t}]'\}$, the $(i,j)$-th element of its acvf $\Gamma^{\mathsf{y}_t,\mathsf{y}_s}$ may be written as $\gamma^{\mathsf{y}_{i_t},\mathsf{y}_{j_s}}$, which according to Definition 2.7, may be interpreted as the cross-covariance between $\mathsf{y}_{i_t}$ and $\mathsf{y}_{j_s}$ $\forall i, j \in 1, \ldots, r$ and $\forall t, s \in \mathbb{Z}$.*

**Definition 2.9.** *A (vector) time series $\{\mathsf{y}_t\}$ $\forall t \in \mathbb{Z}$ is **weakly stationary** if the mean $\mu^{\mathsf{y}_t}$ is a constant (vector) $\mu^{\mathsf{y}}$ and its acvf $\Gamma^{\mathsf{y}_t,\mathsf{y}_s}$ between the (vector) random variables $\mathsf{y}_t$ and $\mathsf{y}_s$ $\forall s \in \mathbb{Z}$ depends on $s$ and $t$ only through $s - t$.*

The variable $h = s - t$ of the *acvf* $\Gamma_h^{\mathsf{y}_t,\mathsf{y}_s}$ will be referred to as the **lag**. It follows from Definition 2.9 that the *acvf* between $\mathsf{y}_{t+h}$ and $\mathsf{y}_t$ of a weakly stationary time series $\{\mathsf{y}_t\}$ is

$$\Gamma_h^{\mathsf{y}} \triangleq \Gamma^{\mathsf{y}_{t+h},\mathsf{y}_t} = \mathsf{E}^{\mathsf{y}_{t+h},\mathsf{y}_t}[(\mathsf{y}_{t+h} - \mu^{\mathsf{y}})(\mathsf{y}_t - \mu^{\mathsf{y}})'] \quad \forall h \in \mathbb{Z}. \tag{2.2}$$

It is easy to verify that $\gamma_h^{\mathsf{y}_i,\mathsf{y}_j} = \gamma_{-h}^{\mathsf{y}_j,\mathsf{y}_i}$, which gives rise to the following property of the *acvf*:

**Property 2.1.** *A weakly stationary acvf is transpose symmetric about $h = 0$, i.e.,*

$$(\Gamma_h^{\mathsf{y}})' = \Gamma_{-h}^{\mathsf{y}}. \tag{2.3}$$

In this thesis, the focus is on time series that are weakly stationary and the main references on that topic are [102, 99, 111, 19, 20]. Now, take a look at a few examples of weakly stationary multivariate time series.

**Property 2.2.** *A weakly stationary $r$-variate time series $\{\mathsf{z}_t\}$ is* ***idiosyncratic*** *if any two components $\mathsf{z}_{i_t}$ and $\mathsf{z}_{j_{t+h}}$ $\forall h \in \mathbb{Z}$, $i \neq j$ of its corresponding vector random variable $\mathsf{z}_t = [\mathsf{z}_{1_t} \cdots \mathsf{z}_{r_t}]$ $\forall t \in \mathbb{Z}$ have zero cross-covariance, i.e., $\gamma_h^{\mathsf{z}_{i_t},\mathsf{z}_{j_{t+h}}} \triangleq \gamma_h^{\mathsf{z}_i,\mathsf{z}_j} \in \mathbb{R}$ is zero whenever $i \neq j$ $\forall i,j \in 1,\ldots,r$ and $h \in \mathbb{Z}$.*

**Note 2.2.** *The diagonal elements of an acvf $\Gamma^{\mathsf{u}}$ of the vector random variable $\mathsf{u}_t = [\mathsf{u}_{1_t}\, \mathsf{u}_{2_t} \cdots \mathsf{u}_{r_t}]'$ will be written simply as $\gamma_h^{\mathsf{u}_i} \triangleq \gamma_h^{\mathsf{u}_i,\mathsf{u}_i}$ $\forall i \in 1,\ldots,r$.*

The *acvf* of an $r$-variate idiosyncratic time series $\{\mathsf{z}_t\}$ due to vector random variable

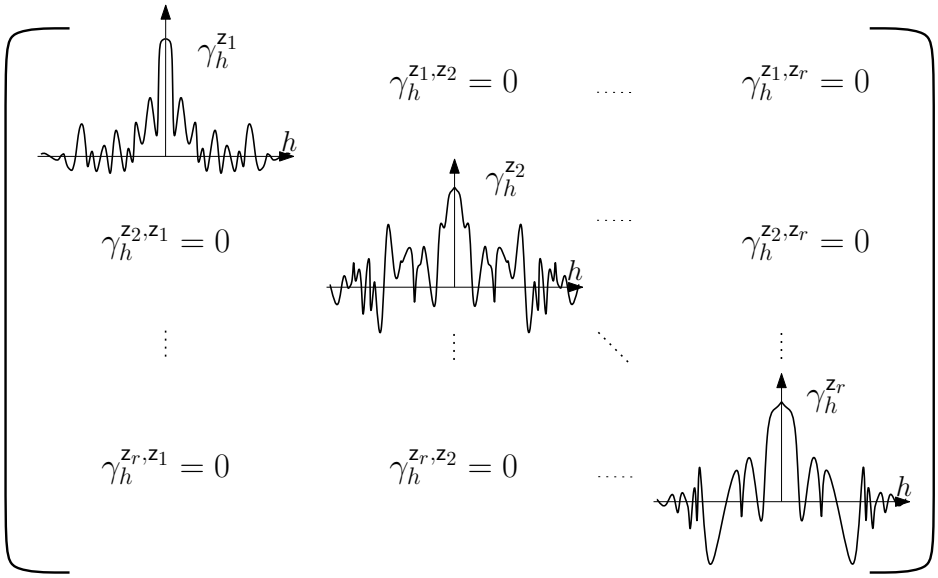


**Figure 2.2:** The structure of the $r \times r$ *acvf* matrix $\Gamma_h^{\mathsf{z}}$ of an $r$-variate idiosyncratic time series $\{\mathsf{z}_t\}$ shows zeros off-diagonal due to no cross-correlation between its components. The plots are hypothetical and interpolated for, an otherwise discrete $h$, for clarity.

$\mathsf{z}_t = [\mathsf{z}_{1_t} \ldots \mathsf{z}_{r_t}]'$ has an $r \times r$ matrix structure shown in Figure 2.2. Following Note 2.2, the cross-covariance $\gamma_h^{\mathsf{z}_i,\mathsf{z}_j}$ of Property 2.2 may be written as the *acvf* $\gamma_h^{\mathsf{z}_i}$ of $\mathsf{z}_i$ whenever $i = j$, whereas $\gamma_h^{\mathsf{z}_i,\mathsf{z}_j} = 0$ is zero otherwise. This means that the off-diagonal elements of such an *acvf* are always zero. Let a special case of an idiosyncratic time series whose each diagonal element of the *acvf* is an impulse function be now defined.

**Definition 2.10.** *For a weakly stationary (vector) time series $\{\mathsf{x}_t\}$, if (the components of) $\mathsf{x}_t$ are independently and identically distributed $\forall t \in \mathbb{Z}$, then $\{\mathsf{x}_t\}$ is said to be (multivariate)* ***white noise****.*

Note that the *acvf* $\Gamma_h^{\mathsf{x}}$ of a $q$-variate white noise $\{\mathsf{x}_t\}$ is $\Gamma_h^{\mathsf{x}} = 0_q$ $\forall h \neq 0$ and $\det(\Gamma_0^{\mathsf{x}}) \neq 0$. Definition 2.10 implies that mean-subtracted white noise components may be defined completely by their component variances $\sigma_{i_h}^2 = \sigma_i^2$ $\forall i = 1,\ldots,q$ so that $\Gamma_h^{\mathsf{x}} = \mathrm{diag}(\sigma_1^2,\ldots,\sigma_q^2)$ $\forall h \in \mathbb{Z}$, which is said to be **isotropic** if $\sigma_i = \sigma$,

$i = 1, 2, \ldots, q$. A special case of the zero-mean isotropic white noise is the following: If the component random variables of a $q$-variate white noise $\{\mathsf{x}_t\}$ have zero means and unit variances so that $\Gamma_h^{\mathsf{x}} = \mathrm{diag}(1, \ldots, 1) \in \mathbb{R}^{q \times q}\, \forall h \in \mathbb{Z}$, then $\{\mathsf{x}_t\}$ is termed a **zero-mean unit-variance white noise**. The white noise is used as the ingredient in many weakly stationary time series process models because of the simplicity of its *acvf*. Defined below is one such model; refer §11.1 of [20] or §9.2 of [79] among many references for its details.

**Definition 2.11.** *For a $q$-variate zero-mean white noise $\{\mathsf{x}_t\}\ \forall t \in \mathbb{Z}$ and matrices $C_j \in \mathbb{R}^{q \times q}\ \forall j \in \mathbb{Z}$ with absolutely summable elements, a **linear process** is defined as the $q$-variate time series*

$$\mathsf{u}_t = \sum_{j \in \mathbb{Z}} C_j \mathsf{x}_{t-j}, \tag{2.4}$$

*which is weakly stationary with zero mean and acvf*

$$\Gamma_h^{\mathsf{u}} = \sum_{j \in \mathbb{Z}} C_{j+h} \Gamma_0^{\mathsf{x}} C_j'. \tag{2.5}$$

## 2.2 Spectral analysis of continuous processes

The purpose of this section is to introduce certain Fourier analysis concepts required for this thesis.

**Note 2.3.** *This section deviates momentarily to discuss continuous time processes; the time index is $t \in \mathbb{R}$; everywhere else in this thesis $t \in \mathbb{Z}$.*

Consider the motion of a simple pendulum as an example of a periodic continuous process. It is assumed for simplicity that the mass of the string attached to the bob of the pendulum is negligible. The oscillation is restricted to a plane so that the constant string length and an angle, viz., the instantaneous angle that the string forms with respect to its equilibrium position, are sufficient to describe its motion. It is also assumed that the amplitude $\alpha$, which is the maximum displacement of the bob from its position of equilibrium, is very small relative to the length of the pendulum. Refer to Figure 2.3; let $\tau$ be the time period of oscillation so that $\tau^{-1}$ is the frequency of oscillation. The standard association of $2\pi$ radians to be equivalent to one complete oscillation may be made. Let $\phi$ radians be the part of $2\pi$ radians of an oscillation the pendulum has completed at time $t = 0$; its sign depends on the choice of direction of reference of the bob's trajectory. If it is assumed that the pendulum is undamped by any kinds of friction and disturbances, then the pendulum's displacement with respect to the equilibrium position of the string at time $t \in \mathbb{R}$ is $y_t = \alpha \cos(2\pi \frac{t}{\tau} + \phi)$. Basic trigonometric identities enable writing $y_t$ in various combinations of sinusoids, e.g.,

$$y_t = a \cos(2\pi t/\tau) + b \sin(2\pi t/\tau), \tag{2.6}$$

where $a = \alpha \cos(\phi)$ and $b = -\alpha \sin(\phi)$.

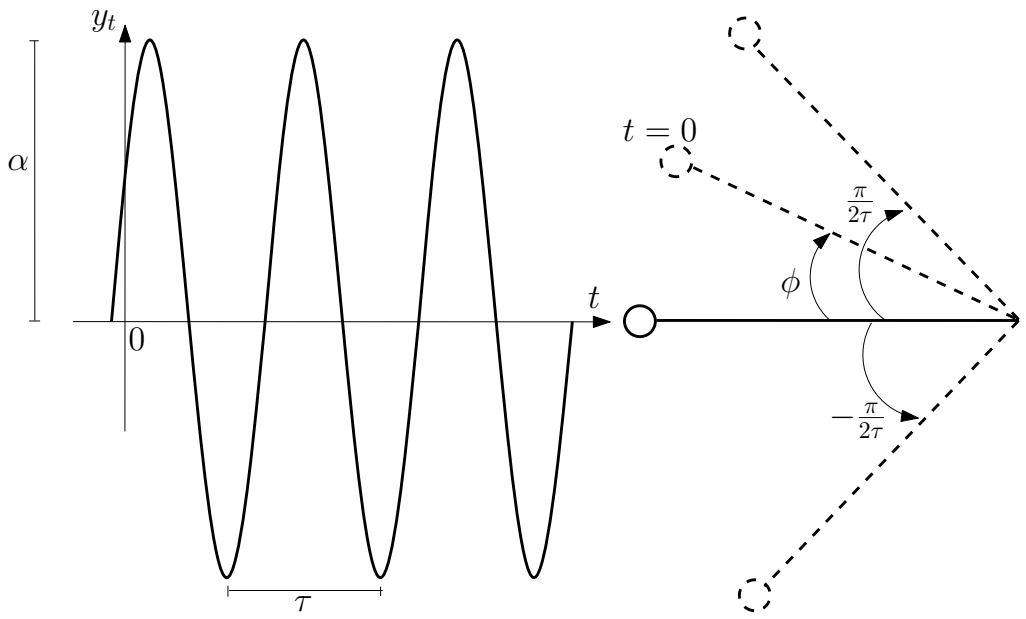


**Figure 2.3:** The motion of the simple pendulum registers a continuous function based on the displacement of its bob.

Just seen is the decomposition of a basic equation of oscillation into two sinusoids of frequency $\tau^{-1}$. Consider that $y_t$ was expressed as a weighted sum of two basis functions; this is because the sinusoids here are orthogonal functions, i.e., $\int_{-\frac{1}{2}\tau}^{\frac{1}{2}\tau} \cos(2\pi t/\tau) \sin(2\pi t/\tau)\, \mathrm{d}t = 0$. The necessity of orthogonal functions in many problems is analogous to the necessity of orthogonal coordinate axes in expressing the position of a point in a Cartesian plane.

The decomposition of a time-domain process to its frequency components is known as **Fourier (spectral) analysis** and the definitions presented in this and Section 2.3 on this topic can be found in references such as [99, 44, 93, 89]. Fourier analysis is based on one of the most important contributions to the sciences originally formalized by Joseph Fourier in 1807 that any 'well-behaved' deterministic continuous periodic function $y_t$ could be expressed as a sum of orthogonal functions if and only if the orthogonal functions are sinusoids, where a 'well-behaved' function satisfies the following condition:

**Definition 2.12.** *A function $y_t \, \forall t \in \mathbb{R}$ is said to be **absolutely summable** if*

$$\int_{-\infty}^{\infty} |y_t| \, \mathrm{d}t < \infty. \tag{2.7}$$

The unique decomposition of such a deterministic periodic function $y_t$ into a possibly infinite number of sinusoids is called its Fourier series representation: $y_t = \frac{a_0}{2} + \sum_{n=1}^{\infty}(a_n \cos(2\pi nt/\tau) + b_n \sin(2\pi nt/\tau))$, where $a_m = \frac{2}{\tau} \int_{-\frac{1}{2}\tau}^{\frac{1}{2}\tau} y_t \cos(2m\pi t/\tau)\, \mathrm{d}t$, $m = 0, 1, 2, \ldots$ and $b_l = \frac{2}{\tau} \int_{-\frac{1}{2}\tau}^{\frac{1}{2}\tau} y_t \sin(2l\pi t/\tau)\, \mathrm{d}t$, $l = 1, 2, 3, \ldots$. It is often con-

venient to use Euler identity $e^{\mathsf{i}\theta} = \cos(\theta) + \mathsf{i}\sin(\theta)$ to reach the following definition which holds for complex-valued functions also.

**Definition 2.13.** *The **Fourier series** representation of a deterministic continuous periodic function $y_t \in \mathbb{C}\ \forall t \in \mathbb{R}$ satisfying (2.7) is*

$$y_t = \sum_{m=-\infty}^{\infty} c_m e^{\mathsf{i}2\pi mt/\tau} \tag{2.8}$$

*where $c_m \in \mathbb{C}\ \forall m \in \mathbb{Z}$ is*

$$c_m = \frac{1}{\tau}\int_{-\frac{1}{2}\tau}^{\frac{1}{2}\tau} y_t e^{-\mathsf{i}2\pi mt/\tau}\mathsf{d}t. \tag{2.9}$$

Note that if $c_m = \overline{c}_m\ \forall m \in \mathbb{Z}$, then the function $y_t \in \mathbb{R}$, else $y_t \in \mathbb{C}$.
Involve frequency spacing $\Delta u = 2\pi/\tau$ and write the Fourier series coefficients as $c_m = \frac{\Delta u}{2\pi}\int_{-\frac{1}{2}\tau}^{\frac{1}{2}\tau} y_t\ e^{-\mathsf{i}mt\Delta u}\mathsf{d}t$. Substituting these coefficients back into the series summation gives $y_t = \sum_{m=-\infty}^{\infty}\frac{\Delta u}{2\pi}\int_{-\frac{1}{2}\tau}^{\frac{1}{2}\tau} y_t e^{-\mathsf{i}nt\Delta u}\ \mathsf{d}t\, e^{\mathsf{i}mt\Delta u}$. Suppose

$$\mathbf{y}(n\Delta u) = \int_{-\frac{1}{2}\tau}^{\frac{1}{2}\tau} y_t e^{-\mathsf{i}nt\Delta u}\mathsf{d}t,$$

then $y_t = \sum_{m=-\infty}^{\infty}\frac{\Delta u}{2\pi}\,\mathbf{y}(n\Delta u)\ e^{\mathsf{i}mt\Delta u}$. As $\tau \to \infty$ or $\Delta u \to 0$,

$$y_t = \frac{1}{2\pi}\int_{u=-\infty}^{\infty}\mathbf{y}(u)e^{\mathsf{i}tu}\mathsf{d}u.$$

This is regarded as the inverse relation of a very important transform in mathematics that is defined below. The term $\mathbf{y}(u)$ in the above development is the result of limiting $\Delta u \to 0$ in $\mathbf{y}(n\Delta u)$ and acquires the following definition:

**Definition 2.14.** ***Fourier transform** of a function $y_t\ \forall t \in \mathbb{R}$ satisfying (2.7) is*

$$\mathbf{y}(u) = \int_{-\infty}^{\infty} y_t e^{-\mathsf{i}tu}\mathsf{d}t. \tag{2.10}$$

## 2.3 Spectral analysis of discrete processes

In the Fourier transform relation of (2.10), a continuous function defined for $t \in \mathbb{R}$ is dealt with. Consider the continuous time function $y_t\ \forall t \in \mathbb{R}$ such that $y_t = 0$ whenever $t \neq m\Delta\tau\ \forall m \in \mathbb{Z}$ for some constant $\Delta\tau > 0$. This is equivalent to sampling the continuous function $y_t\ \forall t \in \mathbb{R}$ at discrete instants separated by $\Delta\tau$ and zero at all other instants. Since only discrete instants are relevant here from the Fourier transform

perspective, $y_t$ would be called a discrete time function. Therefore, the discrete time Fourier transform using (2.10) becomes

$$\mathbf{y}(u) = \sum_{m=-\infty}^{\infty} y_{m\Delta\tau} e^{-\mathrm{i}um\Delta\tau}.$$

Denote $y_m \triangleq y_{m\Delta\tau}$ and refer to it as the $m$-th sample. Then a sufficient condition for the existence of such a relation is $|\mathbf{y}(u)| < \infty$, i.e., $|\sum_{m=-\infty}^{\infty} y_m e^{-\mathrm{i}um\Delta\tau}| \leq \sum_{m=-\infty}^{\infty} |y_m| \, |e^{-\mathrm{i}um\Delta\tau}| < \infty$. This results in the absolute summability condition

$$\sum_{m=-\infty}^{\infty} |y_m| < \infty. \tag{2.11}$$

Using angular frequency $\omega$ as $u\Delta\tau = 2\pi\omega$ in the above development results in the following definition of Fourier transform for discrete time functions.

**Definition 2.15.** *The **discrete time Fourier transform** of a complex-valued discrete function $y_m \; \forall m \in \mathbb{Z}$ satisfying (2.11) is*

$$\mathbf{y}(\omega) = \sum_{m=-\infty}^{\infty} y_m e^{-\mathrm{i}2\pi\omega m}. \tag{2.12}$$

Although real-valued discrete functions were being discussed, the discrete time Fourier transform is valid for complex-valued functions also. Furthermore, since $e^{\mathrm{i}2\pi k} = 1 \, \forall k \in \mathbb{Z}$, the following property holds:

**Property 2.3.** *The discrete time Fourier transform has **unit periodicity**, i.e.,*

$$\mathbf{y}(\omega) = \mathbf{y}(k + \omega) \, \forall k \in \mathbb{Z}. \tag{2.13}$$

Another easily verifiable property, which holds true for any absolutely summable discrete or continuous function, is due to the following theorem; refer §22.1 of [41] or Chapter 3 of [72]:

**Theorem 2.1.** *According to the **Plancherel-Parseval theorem** for the discrete time Fourier transform $\mathbf{y}(\omega) \, \forall \omega \in [-\frac{1}{2}, \frac{1}{2}]$ of the function $y_m \, \forall m \in \mathbb{Z}$,*

$$\sum_{m\in\mathbb{Z}} |y_m|^2 = \int_{-\frac{1}{2}}^{\frac{1}{2}} |\mathbf{y}(\omega)|^2 \, \mathrm{d}\omega. \tag{2.14}$$

Just as the discrete time Fourier transform was defined being valid for complex-valued discrete functions, the Fourier series discussed earlier in (2.8) and (2.9) is applicable to any complex valued periodic function defined over any continuous domain. Hence, replacing $(y, -\frac{t}{\tau}) \to (\mathbf{y}, \omega)$ in (2.8) makes it equivalent to (2.12). In other words, the discrete time Fourier transform of a sequence of equally spaced samples of a real function is also a Fourier series whose coefficients form the sequence. Therefore, allowing

the same replacement in (2.9) gives the differential $\frac{1}{\tau}\mathrm{d}t \to -\mathrm{d}\omega$ and the integral limits $t = \pm\frac{1}{2}\tau \to \omega = \mp\frac{1}{2}$ resulting in the following inverse of the relation in (2.12):

**Definition 2.16.** *The **inverse discrete time Fourier transform** of a complex-valued continuous function $\mathbf{y}(\omega)\,\forall\omega \in \mathbb{R}$ is defined as*

$$y_m = \int_{-\frac{1}{2}}^{\frac{1}{2}} e^{\mathrm{i}2\pi m\omega}\mathbf{y}(\omega)\mathrm{d}\omega. \tag{2.15}$$

The Fourier series gave discrete frequency components of a continuous time process and the discrete time Fourier transform gave continuous frequency components of a discrete time process. On the other hand, the following transform gives discrete frequency components of a finite realization of a discrete time process:

**Definition 2.17.** *The **discrete Fourier transform** of a series $\{y_t\}$, $t = 1, \ldots, \tau$ is defined as*

$$\mathbf{y}(\omega_j) = \frac{1}{\sqrt{\tau}}\sum_{t=1}^{\tau} y_t e^{-\mathrm{i}2\pi\omega_j t} \tag{2.16}$$

*at discrete frequencies $\omega_j = \frac{j}{\tau}$, $j = 0, \ldots, \tau - 1$ and the **inverse discrete Fourier transform** at the discrete instants is defined as*

$$y_t = \frac{1}{\sqrt{\tau}}\sum_{t=1}^{\tau} \mathbf{y}(\omega_j) e^{\mathrm{i}2\pi\omega_j t}. \tag{2.17}$$

The equivalent of Theorem 2.1 for the discrete Fourier transform is as follows [17]:

**Property 2.4.** *According to the **Plancherel-Parseval theorem** for the discrete Fourier transform $\mathbf{y}(\omega_j)$ of the sequence $y_t\,\forall j, t = 1, \ldots, \tau$,*

$$\sum_{t=1}^{\tau} |y_t|^2 = \sum_{j=1}^{\tau} |\mathbf{y}(\omega_j)|^2. \tag{2.18}$$

In this thesis, for a given finite length realization of a multivariate time series, certain asymptotic properties of the discrete Fourier transform will be used to define, derive, and optimize the dynamic transformation of the latent time series into commonalities. These asymptotic properties will be discussed in Section 2.5. The Plancherel-Parseval theorem will enable measuring and containing the commonalities that are retained during the transition between the time-domain and the frequency-domain. In Section 2.4, how the frequency-domain analysis finds utility in a stationary process will be discussed. Specifically, in Theorem 2.2, it will be learned how the Fourier transform relates two important statistical properties of a time series.

## 2.4 Spectral analysis of stationary processes

Suppose the pendulum motion expressed in (2.6) is subject to random amplitude $\alpha$ and phase $\phi$ disturbances so that $(a, b)$ become uncorrelated zero-mean unit-variance random variables $(\mathsf{a}, \mathsf{b})$. Moreover, the discrete time domain is considered so that the equation of motion in (2.6) takes the form $\mathsf{y}_t = \mathsf{a}\cos(2\pi\omega t) + \mathsf{b}\sin(2\pi\omega t)\ \forall t \in \mathbb{Z}$; it will be called the '*perturbed pendulum*.' Its mean $\mu^{\mathsf{y}} = \mathsf{E}^{\mathsf{a},\mathsf{b}}[\mathsf{y}_t] = 0$. The *acvf* is $\gamma_h^{\mathsf{y}} = \mathsf{E}^{\mathsf{a},\mathsf{b}}[\mathsf{y}_{t+h}\mathsf{y}_t]$, which due to non-correlated a and b takes the form $\gamma_h^{\mathsf{y}} = \mathsf{E}^{\mathsf{a},\mathsf{b}}[\mathsf{a}^2\cos(2\pi\omega(t+h))\cos(2\pi\omega t)] + \mathsf{E}^{\mathsf{a},\mathsf{b}}[\mathsf{b}^2\sin(2\pi\omega(t+h))\sin(2\pi\omega t)]\ \forall h \in \mathbb{Z}$. Since it was assumed that $\mathsf{E}^{\mathsf{a}}[\mathsf{a}^2] = \mathsf{E}^{\mathsf{b}}[\mathsf{b}^2] = 1$, what one gets[1] is $\gamma_h^{\mathsf{y}} = \cos(2\pi\omega h)$; spectral analysis of such an *acvf* could be found in [111, 19]. Since $\mu^{\mathsf{y}}$ and $\gamma_h^{\mathsf{y}}$ are independent of $t$, it is found that $\{\mathsf{y}_t\}$ is weakly stationary. And, since weakly stationary $\{\mathsf{y}_t\}$ does not satisfy (2.11), its Fourier transform simply does not exist.

Using Euler's identity, the *acvf* of the perturbed pendulum is written as a summation $\gamma_h^{\mathsf{y}} = \sum_{i=1}^{k} \alpha_i g(\omega_i)$, where $g(\omega) = e^{\mathrm{i}2\pi\omega h}$, $k = 2$, $\omega_1 = -\omega$, $\omega_2 = \omega$, and $\alpha_1 = \alpha_2 = \frac{1}{2}$. But such a summation with a general $g(\omega)$ has an integral representation $\sum_{i=1}^{k} \alpha_i g(\omega_i) = \int g(\omega)\ \mathrm{d}\mathfrak{S}^{\mathsf{y}}(\omega)$, where $\mathfrak{S}^{\mathsf{y}}(\omega) \triangleq \sum_{i=1}^{k} \alpha_i\, 1(\omega_i \leq \omega)$ is a monotonically increasing function bounded between $\mathfrak{S}^{\mathsf{y}}(-\infty) = 0$ and $\mathfrak{S}^{\mathsf{y}}(\infty) = 1$, and $1(\omega_i \leq \omega)$ is the step function which jumps from zero to unity at $\omega = \omega_i$.

However, due to periodicity of $g(\omega) = e^{\mathrm{i}2\pi\omega h}$ in the above example of a perturbed pendulum, the *acvf* is essentially represented in the integral form $\gamma_h^{\mathsf{y}} = \int_{-\frac{1}{2}}^{\frac{1}{2}} e^{\mathrm{i}2\pi\omega h}\ \mathrm{d}\mathfrak{S}^{\mathsf{y}}(\omega)$, where $\mathfrak{S}^{\mathsf{y}}(\omega)$ is a monotonically increasing function bounded between $[-\frac{1}{2}, \frac{1}{2}]$ while $\mathfrak{S}^{\mathsf{y}}(-\frac{1}{2}) = 0$ and $\mathfrak{S}^{\mathsf{y}}(\frac{1}{2}) = \gamma_0^{\mathsf{y}}$. The reader is referred to [99, 20] for the details of this representation and other properties that $\mathfrak{S}^{\mathsf{y}}(\omega)$ adheres to. The notion carried forward is that whenever the derivate $\mathsf{s}^{\mathsf{y}}(\omega) = \frac{\mathrm{d}}{\mathrm{d}\omega}\mathfrak{S}^{\mathsf{y}}(\omega)$ exists, it is possible to write the *acvf* as

$$\gamma_h^{\mathsf{y}} = \int_{-\frac{1}{2}}^{\frac{1}{2}} e^{\mathrm{i}2\pi\omega h}\ \mathsf{s}^{\mathsf{y}}(\omega)\,\mathrm{d}\omega. \tag{2.19}$$

But if there are discontinuities in $\mathfrak{S}^{\mathsf{y}}(\omega)$, e.g., the perturbed pendulum, it will not be possible to write the *acvf* according to (2.19) because $\mathsf{s}^{\mathsf{y}}(\omega)$ does not exist.

Now refer back to (2.15) to see its analogy with (2.19) which requires that a condition

$$\sum_{h=-\infty}^{\infty} |\gamma_h^{\mathsf{y}}| < \infty, \tag{2.20}$$

equivalent to (2.11) be satisfied by $\gamma_h^{\mathsf{y}}$. This enables the following theorem and definition; refer §4.3 of [20]:

[1] Using the trigonometric identity $\cos(\theta_1 - \theta_2) = \cos\theta_1\cos\theta_2 + \sin\theta_1\sin\theta_2$

**Theorem 2.2.** *For acvf $\gamma_h^{\mathsf{y}}$ of a weakly stationary time series $\{\mathsf{y}_t\}\ \forall t \in \mathbb{Z}$ satisfying (2.20), according to Herglotz's theorem, the* ***spectral density function*** $\mathsf{s}^{\mathsf{y}}(\omega)$ *at the frequency $\omega \in \mathbb{R}$ exists and is defined as the discrete time Fourier transform of its acvf, i.e.,*

$$\mathsf{s}^{\mathsf{y}}(\omega) \triangleq \sum_{h=-\infty}^{\infty} \gamma_h^{\mathsf{y}} e^{-\mathrm{i}2\pi\omega h}. \tag{2.21}$$

In this context, it may be noted that the perturbed pendulum does not satisfy the absolute sum condition because $\sum_{h=-\infty}^{\infty} |\gamma_h^{\mathsf{y}}| = \sum_{h=-\infty}^{\infty} |\cos(2\pi f h)| = \infty$ and it does not have a spectral density function.
In light of (2.21) and similar to Property 2.3, the following property of the spectral density function is arrived at:

**Property 2.5.** *The spectral density function has* ***unit periodicity****, i.e.,* $\mathsf{s}^{\mathsf{y}}(\omega) = \mathsf{s}^{\mathsf{y}}(k+\omega)\, \forall k \in \mathbb{Z}$.

For an $r$-variate time series $\{\mathsf{y}_t\}\ \forall t \in \mathbb{Z}$, where $\mathsf{y}_t = [\mathsf{y}_{1_t} \cdots \mathsf{y}_{r_t}]'$, let $\gamma_h^{\mathsf{y}_i,\mathsf{y}_j}$ be the $(i,j)$-th element of its autocovariance matrix $\Gamma_h^{\mathsf{y}}$. Referring to §1.1.2 of [102], the condition equivalent to (2.20) for vector random variables becomes

$$\sum_{h=-\infty}^{\infty} |\gamma_h^{\mathsf{y}_i,\mathsf{y}_j}| < \infty, \tag{2.22}$$

which is valid $\forall i,j \in 1,\ldots,r$ in defining the matrix of spectral density function $\mathsf{S}^{\mathsf{y}}(\omega) \in \mathbb{C}^{r\times r}$ whose $(i,j)$-th element is $\mathsf{s}^{\mathsf{y}_i,\mathsf{y}_j}$. Then, due to the development of Property 2.1 and the relation (2.21), the following is easily got:

**Property 2.6.** *The spectral density function* $\mathsf{S}^{\mathsf{y}}(\omega)$ *is* ***Hermitian symmetric*** *about $\omega = 0$, i.e.,*

$$(\mathsf{S}^{\mathsf{y}}(\omega))' = \mathsf{S}^{\mathsf{y}}(-\omega) = \overline{\mathsf{S}}^{\mathsf{y}}(\omega). \tag{2.23}$$

Referring to Theorem 2.7.1 of [18], Theorem 4.4.1 of [39], and [92, 34], another important property of the $r$-variate time series follows:

**Property 2.7.** *If $\{\mathsf{y}_t\}\ \forall t \in \mathbb{Z}$ is a linear process, then $|\mathsf{S}^{\mathsf{y}}(\omega)| \neq 0\ \forall \omega \in [0,1]$.*

The above discussion is very relevant to the intention in this thesis to assess the commonalities of a multivariate time series via its spectral density function. For the purpose of learning multivariate time series based on the commonalities, the hope is to take the following approaches: Firstly, two multivariate time series are compared by evaluating how similar the components of their spectral density functions are. Secondly, the future evolution of a multivariate time series is predicted by estimating the *acvf*, via its spectral density function, that inherits maximum commonalities.

## 2.5 Asymptotic properties of linear processes

In practical problems, it is infeasible to have a dataset consisting of an infinite collection of samples to compute true statistical properties such as mean, *acvf*, variance, etc. Characteristics of a time series have to be estimated from a given finite length subset of its realization. With a limited number of samples, sample estimates may be questioned for their reliability. The field of study of asymptotic statistics strives to design properties, procedures, tests, and estimators in the limit that the sample size becomes large [122, 58]. A broad review of the asymptotic techniques will not be resorted to; however, presented below are the essentials for the thesis's purposes.

Consider the scenario in which, due to computational or limited access to data, time series characteristics have to be gleaned from one realization forming a finite length data stream. For a weakly stationary time series, these characteristics include its mean and *acvf* for which, first, the following asymptotic properties referring to §11.2 of [20] are presented:

**Theorem 2.3.** *For a finite $\tau$-length realization $\{y_t\}$ $t = 1, \ldots, \tau$ of a weakly stationary time series $\{\mathsf{y}_t\}$ $t \in \mathbb{Z}$ whose acvf $\Gamma_h^{\mathsf{y}}$ satisfies (2.22), the* ***sample mean***

$$\hat{y} = \frac{1}{\tau} \sum_{t=1}^{\tau} y_t \tag{2.24}$$

*converges in a mean square sense to the population mean $\mu^{\mathsf{y}}$.*

**Theorem 2.4.** *For a finite $\tau$-length realization $\{y_t\}$ $t = 1, \ldots, \tau$ of a weakly stationary time series $\{\mathsf{y}_t\}$ $t \in \mathbb{Z}$ with sample mean $\hat{y}$, the $r \times r$* ***sample acvf***

$$\widehat{\Gamma}_h^{\mathsf{y}} = \begin{cases} \frac{1}{\tau} \sum_{t=1}^{\tau-h} (y_{t+h} - \hat{y})(y_{t+h} - \hat{y})' & 0 \le h \le \tau - 1, \\ \frac{1}{\tau} \sum_{t=-h+1}^{\tau} (y_{t+h} - \hat{y})(y_{t+h} - \hat{y})' & -\tau + 1 \le h < 0 \end{cases} \tag{2.25}$$

*converges in probability to the population acvf $\Gamma_h^{\mathsf{y}}$.*

With the sample *acvf* $\widehat{\Gamma}_h^{\mathsf{y}}$ for finite lags, the best hope is for estimates of the spectral density function $\mathsf{S}^{\mathsf{y}}(\omega_k)$ at finite discrete frequencies $\omega_k = \frac{k}{\tau}$ $\forall |k| = 0, \ldots, \tau - 1$ via inverse discrete Fourier transform. For an otherwise continuous spectral density function $\mathsf{S}^{\mathsf{y}}(\omega)$, $0 \le \omega < 1$, those estimates at discrete frequencies is an approximation of $\mathsf{S}^{\mathsf{y}}(\omega_k)$ dependent on how good the sample estimation $\widehat{\Gamma}_h^{\mathsf{y}}$ is. Therefore, in what follows, described is the asymptotic property of $\mathsf{S}^{\mathsf{y}}(\omega)$ near any target frequency $\omega_j = j/\hat{\jmath}$ $\forall j = 0, \ldots, \hat{\jmath} - 1$, or $0 \le \omega_j < 1$ and $\hat{\jmath} \ll \tau$.

It starts by splitting a period of $\omega \in [0, 1)$ of the spectral density function into $\hat{\jmath}$ non-overlapping frequency bands. Suppose there is a total of $\tau = n\hat{\jmath}$ discrete frequencies that are considered for the splitting so that each band will have $n$ discrete frequencies. By the 0-th frequency band represented by the target frequency $\omega_0 = 0$, implied are $n$ discrete frequencies $\omega_{0,l} > 0$, $l = 1, \ldots, n$ closest to 0. By the $j$-th frequency band $\omega_{j,l}$ $\forall l = 1, \ldots, n$; $j = 1, \ldots, \hat{\jmath} - 1$, implied are $n$ frequencies closest to the target frequency $\omega_j = j/\hat{\jmath}$ and between $\omega_j - b$ and $\omega_j + b$, where $2b = n/\hat{\jmath}$ is called the

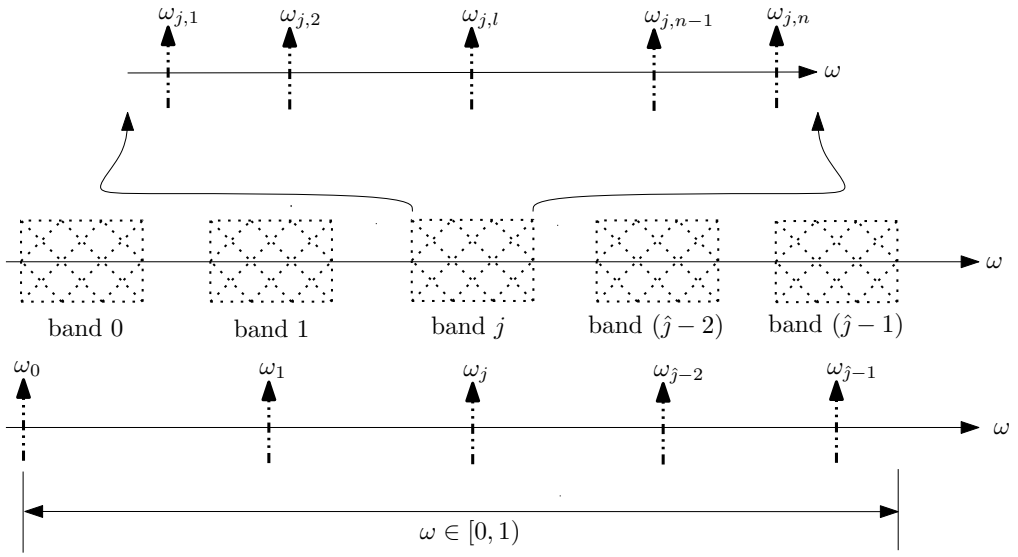


**Figure 2.4:** The scheme of splitting the frequency range $\omega = [0, 1)$ into $\hat{j}$ non-overlapping subbands containing $n$ discrete frequency components each.

bandwidth. Suppose $2b < \omega_1 - b < \omega_{\hat{j}-1} + b < 1$ and $n \ll \hat{j}$ is choosen so that the bandwidth is very low.

For proceeding further, the following definitions are needed; refer to [45]:

**Definition 2.18.** *An $r$-dimensional '**complex-valued vector random variable**' $\xi = [\xi_1 \cdots \xi_r]' = \Re(\xi) + \mathrm{i}\Im(\xi) \in \mathbb{C}^r$ is defined as the $2r$-variate vector random variable $\eta = [\Re(\xi_1)\, \Im(\xi_1) \cdots \Re(\xi_r)\, \Im(\xi_r)]' \in \mathbb{R}^{2r}$ formed by its real and imaginary components.*

As established in [45], the covariance matrix $\Gamma^\xi$ of an $r$-variate complex valued vector random variable $\xi$ is isomorphic, i.e., equivalent upto a row and a column permutation, to the covariance matrix $\Gamma^\eta$ of its corresponding $2r$ component vector random variable $\eta$ via

$$\Gamma^\xi \approxeq 2\Gamma^\eta;\ (\Gamma^\xi)^{-1} \approxeq \frac{1}{2}(\Gamma^\eta)^{-1};$$

whereas the means are isomorphic via $\mu^\xi \approxeq \mu^\eta$. Also, it was shown there that $\det(\Gamma^\xi) = 2^r(\det(\Gamma^\eta))^{\frac{1}{2}}$ and $\xi^*\Gamma^\xi\xi = \eta'\Gamma^\eta\eta$. Then, following the convention of a Gaussian distribution of an $r$-variate random variable u with mean $a$ and covariance matrix $B$ denoted by

$$\mathcal{N}(\mathsf{u}|a, B) = \frac{\exp\left(-\frac{1}{2}(\mathsf{u}-a)'B^{-1}(\mathsf{u}-a)\right)}{(2\pi)^r\ (\det(B))^{\frac{1}{2}}}, \tag{2.26}$$

the following definition could be arrived at:

**Definition 2.19.** *The $r$-variate '**complex Gaussian probability density**' of a complex valued random variable* $\mathsf{u}$ *with mean $a \in \mathbb{C}^r$ and covariance matrix $B \in \mathbb{C}^{r\times r}$ is defined as*

$$\mathcal{N}_{\mathbb{C}}(\mathsf{u} \mid a, B) = \frac{\exp\left(-(\mathsf{u}-a)^* B^{-1}(\mathsf{u}-a)\right)}{\pi^r \det(B)}. \tag{2.27}$$

Now an essential theorem for this thesis is presented; refer Theorem 4.4.1 of [18], §C.2 of [111], and [53].

**Theorem 2.5.** *The discrete Fourier transform components of a realization of an $r$-variate linear process at frequencies $\omega_{j,l}$ such that $\lim_{\hat{\jmath}\to\infty} |\omega_j - \omega_{j,l}| \to 0 \; \forall l \in 1,\ldots,n$; $\forall j = 1,\ldots,\hat{\jmath} \gg n$ are iid samples of an $r$-dimensional 'complex-valued vector random variable' $\mathsf{y}_j$ at frequency $\omega_j \in [0,1]$ with a probability density*

$$\mathsf{p}^{\mathsf{y}(\omega_j)}(u) = \begin{cases} \mathcal{N}_{\mathbb{C}}(u \mid 0, \mathsf{S}^{\mathsf{y}}(\omega_j)), & \omega_j \in (0,1) \\ \mathcal{N}(u \mid 0, 2\,\mathsf{S}^{\mathsf{y}}(\omega_j)), & \omega_j \in \{0,1\}. \end{cases} \tag{2.28}$$

Theorem 2.5 furnishes a probabilistic model for discrete Fourier transform samples obtained from a finite realization of a time series of a linear process. The theorem simply recommends that discrete Fourier transform components within a 'small' bandwidth near a target frequency $\omega_j$ is Gaussian with the covariance matrix equal to the spectral density function $\mathsf{S}^{\mathsf{y}}(\omega_j)$; at zero frequency the covariance matrix is twice $\mathsf{S}^{\mathsf{y}}(0)$.

In order to use this theorem, the following procedure is adhered to: Given $\tau$ samples of a time series realization, first compute the $\tau$-length discrete Fourier transform $\mathbf{y}(\omega_k)$, $k = 0,\ldots,\tau-1$. Then, $n$ discrete Fourier transform components $\mathbf{y}(\omega_{j,l})$, $l = 1,\ldots,n$; $j = 0,1,\ldots,\hat{\jmath}-1$, contained in the $j$-th subband may be assigned as

$$\omega_{j,l} : |\omega_j - \omega_{j,l}| \leq n\tau^{-1}; \quad n \ll \hat{\jmath}. \tag{2.29}$$

For the $j$-th frequency band, the sample covariance matrix is computed as

$$\widehat{\mathsf{S}}^{\mathsf{y}}(\omega_j) = \frac{1}{n}\sum_{l=1}^{n} (\mathbf{y}(\omega_{j,l}) - \hat{\mathbf{y}}(\omega_j))(\mathbf{y}(\omega_{j,l}) - \hat{\mathbf{y}}(\omega_j))^*, \tag{2.30}$$

and

$$\hat{\mathbf{y}}(\omega_j) = \frac{1}{n}\sum_{l=1}^{n} \mathbf{y}(\omega_{j,l})$$

is the sample mean of the discrete Fourier transform $\mathbf{y}(\omega_k)$ and $\omega_j - b < \omega_k < \omega_j + b$. To ensure robustness of the estimate $\widehat{\mathsf{S}}^{\mathsf{y}}(\omega_j)$, typically, one would also want to maintain

$$n \geq r^2, \tag{2.31}$$

refer [106, 12].

It could be shown, as done in §4.2 of [18] or §12.4 of [25] that for a linear process

$$\lim_{n\to\infty} \mathsf{E}^{\mathsf{y}}[\widehat{\mathsf{S}}^{\mathsf{y}}(\omega \mid n)] = \mathsf{S}^{\mathsf{y}}(\omega) \quad \forall \omega \in [0,1). \tag{2.32}$$

Hence, while maintaining $\hat{\jmath} \gg n$, a sufficiently large $n$ should provide an unbiased estimate of $\mathsf{S}^{\mathsf{y}}(\omega_j)$ through (2.30). This process is given in Algorithm 1, where care should be taken to ensure that there are sufficient subbands as required by Theorem 2.5.

**Algorithm 1**: Prepare discrete Fourier transform subbands

**Input**: $\mathcal{D} = \{y_t\}, t = 1, \ldots, \tau$; $y_t \in \mathbb{R}^r$; $n$; $\hat{\jmath}$;
**Output**: $\{\widehat{\mathsf{S}}^{\mathsf{y}}(\omega_j)\}$; $\{\mathbf{y}(\omega_{j,l})\}$; $j = 1, \ldots, \hat{\jmath}$; $l = 1, \ldots, n$;
compute $y_t \overset{\mathcal{F}}{\longleftrightarrow} \mathbf{y}(\omega_k)$; $\omega_k = \frac{k}{\tau}$, $k = 0, \ldots, \tau - 1$ using (2.16);
assign $\mathbf{y}(\omega_{j,l})$; $j = 1, \ldots, \hat{\jmath}$; $l = 1, \ldots, n$ using (2.29);
estimate $\widehat{\mathsf{S}}^{\mathsf{y}}(\omega_j)$ using (2.30);

## 2.6 Summary

This chapter introduced certain frequently sought after notions pertaining to time and frequency domain analyses of time series. These notions include *acvf*, spectral density function, discrete Fourier transform, white noise, etc. The relation between the spectral density function and the *acvf* was recapped on. Also introduced were some of the notations adhered to for the remaining chapters. As discussed, the spectral density function of a stationary time series is the Fourier transform of its autocovariance function. The discrete Fourier transform components of a linear process within a small bandwidth around a target frequency is approximately complex-Gaussian with mean zero and covariance matrix equaling the spectral density function at the target frequency.

Our goal for this thesis is to model and learn a measured multivariate time series by dynamically transforming a low-dimensional latent time series. The hope is to use classical probabilistic modeling concepts introduced in the next chapter to achieve this goal. Most of those concepts will be based on fitting popular probability density function models on time and lag independent data; but it is time series data that is dealt with. In order to elicit a similar and manageable probability density function that applies to a wide class of time series, the asymptotic theory of discrete Fourier transform components was approached. This is because those components within a small bandwidth may be considered as realizations of a complex-valued Gaussian vector random variable. This enables the possibility of applying standard probabilistic modeling techniques, as reviewed in the next chapter, to multivariate time series modeling.

# Chapter 3

# Analytical and iterative factor modeling

Modeling a process which generated a given multivariate time series dataset for the purpose of learning motivates this thesis. In Chapter 2, the essential time series analysis tools that are needed in the modeling were presented; whereas in this chapter the elements of building the model itself will be discussed. Models with parameters that could be tuned to fit the statistical characteristics of the dataset at hand will be chosen; this tuning is called **parametric estimation**. It may be seen as a limitation because it warrants assumptions on the type of the data generation process involved. However, essential precautions will be taken by modeling on a dataset that is representative enough of the process. Moreover, in order to avoid any overfitting, learning methods which caution when wide deviations from the assumptions on the model and data are detected will be used.

The treatment of this chapter from the rest of the thesis has one main difference; here, any temporal correlation of the data samples in the given dataset is ignored. Yet, later on in the thesis, the parametric modeling techniques presented herein with the time series techniques of Chapter 2 will be utilized to achieve the thesis objectives.

In Section 3.1, a well-founded modeling strategy based on the **principle of maximum likelihood** will be introduced [96]. The principle assumes that the data has been generated by a known class of probability distributions whose parameters are to be estimated such that the **likelihood** of observing the data is maximized.

In Section 3.2, the concept of a **linear model** whose parameters are linear combinations of the data samples will be introduced. The derivation of the optimal parameters will be summarized by the **Gauss-Markov theorem**. The ideas of **unbiased** and **efficient** parameters defined there are desirable properties for any parametric model.

In Section 3.3, the **factor model** is presented. While the linear model of Section 3.2 utilizes some **measured variables** of the dataset to explain themselves or other measured variables, the factor model is remarkably different. The latter assumes existence of a fewer number of unmeasured **latent variables** responsible for generating all measured variables of a given dataset. The transformation from latent variables to measured variables is assumed to be non-random but unknown; this transformation will account for the covariations in the data. However, in the measured data, there will be deviations unexplained by such a generative model. Those deviations will be assumed unique to each of the measured variables and the variables that absorb these unique

deviations will be called **unique factors**. This characteristic of the factor model is actually facilitated by imposing a diagonal structure on the covariance matrix of the unique factors; whereas the latent variables are transformed such that they absorb the common variation of the measured variables. The transformed latent variables are, hence, called **common factors**.

Note that the parameters of the factor model are (i) **transformation matrix** of the latent variables to the measured variables and (ii) **variances of the unique factors**. The principle of maximum likelihood cannot yield these two sets of parameters independently. By assuming knowledge or guessing one of them, an estimate for the other parameter could be found.

In Section 3.4.1, the **principal factor model** first estimates the covariance matrix of the unique factors in order to estimate the transformation matrix; the reverse procedure is followed in **principal component factor model** of Section 3.4.2.

In Section 3.5, the **Expectation - Maximization** (EM) algorithm for maximum likelihood estimation of the factor model will be first narrated in an original manner. It is an iterative scheme established by [33]. The expression for **complete log-likelihood** of the measured variables as well as the latent variables are written out. However, its analytical tediousness in direct maximization is realized. This is overcome by probing its lower bound. It turns out that the local maximum of the lower bound is attained whenever the complete log-likelihood converges to the log-likelihood of the measured variables. Hence, starting with a set of guessed parameters, iteratively maximizing the complete log-likelihood converges towards the standard log-likelihood. Writing the lower bound of the complete log-likelihood scheme in an *a posteriori* expectation format and maximizing it for the optimal parameters is the crux of the EM algorithm.

In Sections 3.6 and 3.7, the scheme for using the EM algorithm for iteratively estimating factor model parameters is presented; it is partly along the lines of [14]. In doing so, the expression for the log-likelihood in the complete log-likelihood form is first written out; the latter is conducive for use with the algorithm. In the E-step of the algorithm presented in Section 3.7.1, the *a posteriori* mean and covariance of the latent variables are derived. In the M-step of Section 3.7.2, *a posteriori* expectation format of the log-likelihood is maximized; the parameters of the factor model, viz., transformation matrix of the common factors and covariance matrix of the unique factors, are thereby estimated.

## 3.1 Maximum likelihood model

This section starts by presenting some notions and usages that will help in explaining the characteristics of the data to be modeled. The primary assumption is that the data is a collection of samples of some relevant variables measured in an experiment; this collection will be denoted by $\mathcal{D}$ and will be called simply the **dataset**. Let $\mathcal{D}$ be constituted by $n$ data samples written in the sequence $\mathcal{D} = \{y_l\}$, $l = 1, \ldots, n$ and $y_l$ be called the $l$-th **sample**.

In this chapter, unlike in Chapter 2, any sequential dependence of the value or occurrence of a data sample on next or any other sample is ignored. So there is no need to sort the data samples based on the time of data acquisition or any other criteria. But an index to identify the samples individually may be used.

Let the samples in $\mathcal{D}$ be realizations of the sequence $\mathrm{y}_l$ of random variables. Let $\mathrm{p}^{\mathrm{y}_l}$ denote density functions of their respective probability distributions. Throughout this

chapter, it is assumed that the samples encompassing $\mathcal{D}$ adhere to the characteristic defined below [108]:

**Definition 3.1.** *A dataset $\{y_l\}$, $l = 1, \ldots, n$ due to its respective random variables $\mathsf{y}_l$ is said to be **independently and identically distributed or iid** if its joint probability density function is $\prod_{l=1}^{n} \mathsf{p}^{\mathsf{y}}(y_l)$, where $\mathsf{y}$ is a random variable such that $\mathsf{p}^{\mathsf{y}}(y_l) = \mathsf{p}^{\mathsf{y}_l}(y_l)\ \forall l = 1, \ldots, n$.*

Based on Definition 3.1, $\mathcal{D}$ is interpreted as $n$ realizations of a random variable $\mathsf{y}$ whose probability density function is $\mathsf{p}^{\mathsf{y}}$. Therefore, any realization $y$ of $\mathsf{y}$ has a corresponding probability density $\mathsf{p}^{\mathsf{y}}(y)$. Extending $\mathsf{p}^{\mathsf{y}}(y)$ based on Definition 3.1 to $\mathcal{D}$ leads to the joint probability density of the dataset, which is simply denoted by $\mathsf{p}^{\mathsf{y}}(\mathcal{D})$, and is given by

$$\mathsf{p}^{\mathsf{y}}(\mathcal{D}) \triangleq \prod_{l=1}^{n} \mathsf{p}^{\mathsf{y}}(y_l). \tag{3.1}$$

In order to maintain simplicity for the model which generated $\mathcal{D}$, a set of parameters $\theta$ will be introduced for the model. In the context of Definition 3.1, $\theta$ refers to the set of parameters of the probability density function of $\mathsf{y}$. However, $\theta$ is unknown and has to be estimated. In this setup, $\mathsf{p}^{\mathsf{y}|\theta}(\mathcal{D} \mid \theta)$ will be termed as the likelihood of the dataset whilst the parameters $\theta$ are available. Note that the distribution is over $\mathsf{y}$ and the likelihood is a function of non-random $\theta$.

At a set of parameters $\theta$, due to (3.1), the likelihood of the dataset $\mathcal{D}$ consisting of iid samples $y_l$, $l = 1, \ldots, n$ factorizes as

$$\mathsf{p}^{\mathsf{y}|\theta}(\mathcal{D} \mid \theta) = \prod_{l=1}^{n} \mathsf{p}^{\mathsf{y}|\theta}(y_l \mid \theta). \tag{3.2}$$

In order to find an appropriate model for a given dataset, the intention is to utilize the following statistical methodology [96].

**Definition 3.2.** *According to the **principle of maximum likelihood**, an optimal set of parameters $\hat{\theta}$ for the model corresponding to a dataset $\mathcal{D}$ is the set of parameters $\theta$ for which the likelihood of $\mathcal{D}$ is maximized, i.e.,*

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\ \mathsf{p}^{\mathsf{y}|\theta}(\mathcal{D} \mid \theta). \tag{3.3}$$

A modification to (3.3) is made now: Since probability density is a non-negative function, the logarithm of the likelihood is maximized to arrive at the same solution for the optimal parameters as per Definition 3.2. Such an analysis of the exponential family of probability density functions will lead to substantial simplification [96]. Hence, (3.3) may be rewritten equivalently as

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\, \mathsf{L}(\theta), \tag{3.4}$$

where

$$\mathsf{L}(\theta) = \log_e \mathsf{p}^{\mathsf{y}|\theta}(\mathcal{D} \mid \theta). \tag{3.5}$$

## 3.2 Linear model

In Section 3.1, an appropriate estimate $\hat{\theta}$ of $\theta$ is considered as parameter for a model given the dataset $\mathcal{D}$. It is assumed that $\theta$ is a non-random quantity. Now consider the estimator $\Theta$ of $\theta$, i.e., $\Theta$ is a random variable. It is hoped that $\Theta$ gives a reasonably good estimate of the true set of parameters $\theta$ given $\mathcal{D}$. Denoting mean of $\Theta$ by $\mu^{\Theta}$ and variance by $\gamma^{\Theta}$, the following properties indicate the quality of $\Theta$; refer §4.4 of [70] and §10.3 of [94]:

**Property 3.1.** *An estimator $\Theta$ of $\theta$ is **unbiased** if $\mu^{\Theta} = \theta$.*

**Property 3.2.** *An estimator $\Theta$ of $\theta$ is **efficient** if $\Theta = \underset{\widetilde{\Theta} \in \mathcal{C}}{\operatorname{argmin}} \gamma^{\widetilde{\Theta}}$, where $\mathcal{C}$ is the class of all unbiased estimators of $\theta$.*

A modeling strategy in Section 3.1 with a dataset constituted by iid samples was considered. Another popular parametric modeling paradigm called **linear model** involves treating a set of $r$-variate iid samples $y_1, \ldots, y_n$ as dependent on a set of $q$-variate iid samples $x_1, \ldots, x_n$, where $n > q$. The simplest of linear models regresses $x_l$ towards $y_l$ $\forall l = 1, \ldots, n$ through the relation

$$y_l = W x_l + z_l, \tag{3.6}$$

where $W \in \mathbb{R}^{r \times q}$ is a linear function of $x_l \in \mathbb{R}^q$ and $z_l \in \mathbb{R}^r$ is the error in the regression [119, 109]. The model parameters $\theta$ discussed above refer to $W$ here. Suppose the modeling errors $z_l$ are realizations of the vector random variable z, then the measurements $y_l$ may also be treated as realizations of the vector random variable y. The linear model may then be effectively written as

$$\mathsf{y} = W x + \mathsf{z}, \tag{3.7}$$

for any $x \in \mathbb{R}^q$. By restricting the quality of the regression error z, the following theorem defines a popular linear model; for details one may refer to §6.2.1 of [62], §7.1 of [38] or §8.1 of [51] among plenty of references in the literature: Suppose each realization $y_l \in \mathbb{R}^r$ of the vector random variable y is related to $x_l \in \mathbb{R}^q$, $l = 1, \ldots, n$ through (3.6) or (3.7) where $z_l$ are due to zero mean uncorrelated Gaussian vector random variable z. According to the Gauss-Markov theorem, an efficient estimator of $W$ is given by

$$\widehat{W} = [y_1 \cdots y_n] X' \left(X X'\right)^{-1}, \tag{3.8}$$

where $X = (x_1 \cdots x_n) \in \mathbb{R}^{q \times n}$ has $\operatorname{rank}(X) = q$. Then, the error estimate for the $l$-th sample becomes $\hat{z}_l = (\hat{z}_{1_l} \cdots \hat{z}_{r_l})' = y_l - \widehat{W} x_l \; \forall l = 1, \ldots, n$. The unbiased estimator of the covariance matrix $\Gamma^{\mathsf{z}} = \operatorname{diag}(\gamma^{\mathsf{z}_1}, \ldots, \gamma^{\mathsf{z}_r})$ of z is given by

$$\gamma^{\mathsf{z}_k} = \frac{1}{n-q} \sum_{l=1}^{n} \hat{z}_{k_l}^2 \quad \forall k = 1, \ldots, r. \tag{3.9}$$

## 3.3 Factor model

The linear model reviewed in Section 3.2 could be seen as the $q$-variables of the iid samples $x_l$, $l = 1, \ldots, n$ together explaining the $r$-variables of each and every iid sample $y_l$, where both these sets of measured variables are available as part of the dataset $\mathcal{D}$. Suppose $y_l$ and $x_l$ , $l = 1, \ldots, n$ are treated to be due to vector random variables y and x, respectively, so that y is the result of the transformation $W$x, where $W \in \mathbb{R}^{r \times q}$. Then, the challenge is to explain y when x is unavailable or inaccessible in $\mathcal{D}$. One way to proceed is by assuming latent existence of the $q$-dimensional vector random variable x in generating the $r$-dimensional vector random variable y. In that context, y is named the set of **measured variables** and x the **latent variables**.
It is wished to pursue here a parametric model by involving the probability density function; this will help extract the statistical characteristics of the dataset in a finite number of parameters. Hence, if the probability density function of x is assumed known in the transformation $\mathsf{y} = W\mathsf{x}$, then $W$ serves as the parameter that needs to be estimated from the measured y.
However, the model $\mathsf{y} = W\mathsf{x}$ is very restrictive because it assumes that any randomness in y is due to x whose characteristics are assumed. The model is relaxed by introducing an $r$-dimensional random variable z uncorrelated with x and designated to absorb all deviations in y that cannot be retained by

$$\mathsf{v} \triangleq W\mathsf{x}. \tag{3.10}$$

Thus, the measured variables y are split into the common factors $\mathsf{v} = W\mathsf{x}$ and unique factors z; the following is such a model [80]:

**Definition 3.3.** *A **factor model** is defined as*

$$\mathsf{y} = W\mathsf{x} + \mathsf{z}, \tag{3.11}$$

*where* y *and* z *are* $r$*-dimensional vector random variables,* x *is a* $q$*-dimensional vector random variable,* $W \in \mathbb{R}^{r \times q}$ *is a non-random transformation matrix, and*

$$\mu^{\mathsf{x}} = 0,\ \mu^{\mathsf{y}} = 0,\ \mu^{\mathsf{z}} = 0, \tag{3.12}$$

$$\Gamma^{\mathsf{x}} = I_q, \tag{3.13}$$

$\Gamma^{\mathsf{z}}$ *is diagonal, and* (3.14)

$$\Gamma^{\mathsf{x},\mathsf{z}} = 0. \tag{3.15}$$

Given a dataset of realizations of y, the parameters of the factor model that need to be estimated are $W$ and the covariance matrix $\Gamma^{\mathsf{z}}$ of z. The factor model, in contrast to the linear model of (3.6), does not observe any realizations of x.
The following essential result for the moments of a function of a vector random variable is summarily provided; refer Chapter 6 of [35]: For $\mathsf{v} = W\mathsf{x}$, $\mu^{\mathsf{v}} = W\mu^{\mathsf{x}}$ and $\Gamma^{\mathsf{v}} = W\Gamma^{\mathsf{x}}W'$. Therefore, applying (3.13) gives

$$\Gamma^{\mathsf{v}} = WW'. \tag{3.16}$$

In the factor model, the condition (3.15) of zero correlation between x and z is crucial. Naturally, it leads to $\Gamma^{\mathsf{v},\mathsf{z}} = 0$. Therefore, taking the second-order moments on both sides of (3.11) gives

$$\Gamma^{\mathsf{y}} = \Gamma^{\mathsf{v}} + \Gamma^{\mathsf{z}} = WW' + \Gamma^{\mathsf{z}}. \tag{3.17}$$

Due to (3.14), the $r$ components of z are uncorrelated, or, all cross-covariances between the $r$ components of y are inherited by only the covariance matrix $\Gamma^{\mathsf{v}} = WW'$ of $\mathsf{v} = W\mathsf{x}$ and not by $\Gamma^{\mathsf{z}}$. This can be seen as each component of z inheriting only a part of the variance unique to its corresponding component in y. Hence, the components of z are called the **unique factors**. Since no part of the covariance of y common to all its components are held by z but instead by the transformation $\mathsf{v} = W\mathsf{x}$, v is called the **common factors**.
Note that in the factor model, the $rq$ elements of $W$ and $r$ diagonal elements of $\Gamma^{\mathsf{z}}$ are to be estimated. Since (3.17) is just one equation with two unknowns, i.e., $W$ and $\Gamma^{\mathsf{z}}$, it cannot be solved uniquely; more conditions and assumptions may be placed to restrict possible solutions.

## 3.4 Maximum likelihood factor model

Well-known is the following assumption towards a proper solution of the factor model parameters, e.g., refer §3.5 of [62]: The measured variables follow a Gaussian distribution with parameters $\theta = \{\mu^{\mathsf{y}}, \Gamma^{\mathsf{y}}\}$, i.e.,

$$\mathsf{p}^{\mathsf{y}|\theta}(y \mid \theta) = \mathcal{N}(y \mid \mu^{\mathsf{y}}, \Gamma^{\mathsf{y}}), \tag{3.18}$$

as defined in (2.26).

Given samples $y_l$ , $l = 1, \ldots, n$ of the measured variables y, the principle of maximum likelihood as per Definition 3.2 could be used to estimate an optimal set of parameters according to (3.4). The maximum likelihood parameters $\widehat{\mu}^{\mathsf{y}}$ and $\widehat{\Gamma}^{\mathsf{y}}$ of the mean and covariance matrix of the Gaussian distribution in (3.18) are the sample mean and sample covariance matrix, respectively, i.e.,

$$\widehat{\mu}^{\mathsf{y}} = \frac{1}{n}\sum_{l=1}^{n} y_l, \tag{3.19}$$

$$\widehat{\Gamma}^{\mathsf{y}} = \frac{1}{n}\sum_{l=1}^{n} (y_l - \widehat{\mu}^{\mathsf{y}})(y_l - \widehat{\mu}^{\mathsf{y}})'. \tag{3.20}$$

Then (3.20) may be substituted in (3.17) to get

$$\widehat{\Gamma}^{\mathsf{y}} = WW' + \Gamma^{\mathsf{z}}. \tag{3.21}$$

However, it gives no clue regarding the maximum likelihood $W$ and $\Gamma^{\mathsf{z}}$, which are the parameters of interest to the factor model. In what follows, two relevant methods which derive appropriate solutions on the basis of the general maximum likelihood solution are briefly presented.

### 3.4.1 Principal factor model

One of the approaches to finding possible solutions to the maximum likelihood factor model of the $r$-dimensional measured variables y using $q$-dimensional common factors x and $r$-dimensional unique factors z as per Definition 3.3 starts with a good guess $\widehat{\Gamma}^{\mathsf{z}}$ of $\Gamma^{\mathsf{z}}$. The approach is called the principal factor model. One may refer to §10.2 of [54] or §6.3 of [46] to know how this guess could be made reliable; the details which are unnecessary for the objective of the present discussion are skipped. Substituting $\widehat{\Gamma}^{\mathsf{z}}$ in (3.21) gives

$$\widehat{\Gamma}^{\mathsf{y}} = WW' + \widehat{\Gamma}^{\mathsf{z}}. \tag{3.22}$$

The problem then is to estimate a $W$ such that $WW' = \widehat{\Gamma}^{\mathsf{y}} - \widehat{\Gamma}^{\mathsf{z}}$ subject to some quality criterion. Suppose columns $u_1, \ldots, u_r$ of $U \in \mathbb{R}^{r\times r}$ are the eigenvectors of $\widehat{\Gamma}^{\mathsf{y}} - \widehat{\Gamma}^{\mathsf{z}}$ whose corresponding eigenvalues $d_1^2 \geq \cdots \geq d_r^2 > 0$ constitute the diagonal elements of a diagonal matrix $D^2$ from top-left to bottom-right. If the eigenvalue-eigenvector decomposition of $WW' = UD^2U'$ and the subscript $1:q$ is used to refer the first $q$ column indices of a matrix, then the optimal transformation matrix of the principal factor model is

$$\widehat{W} = U_{1:q}D_{1:q}. \tag{3.23}$$

If necessary, the estimation between $\widehat{\Gamma}^{\mathsf{z}}$ and $\widehat{W}$ may alternate iteratively.

### 3.4.2 Principal component factor model

Another approach, for which [103] is referred to, involves first estimating $W$ and then the covariance matrix $\Gamma^{\mathsf{z}}$ of the unique factors. In order to estimate $\Gamma^{\mathsf{v}}$ and $W$, first, it has to be reminded that $WW'$ is of rank $q$. Second, note that the relation (3.21) may be thought as $WW'$ approximating the variance-covariance of the measured variables y as contained in $\Gamma^{\mathsf{y}}$. There could be infinitely many ways $WW'$ could approximate $\Gamma^{\mathsf{y}}$ and an approximation with respect to the Frobenius norm $\|\widehat{\Gamma}^{\mathsf{y}} - \Gamma^{\mathsf{v}}\|_F$ seems reasonable and standard practice; refer §10.2 of [54] and and §2.12 of [103]. In that context, the following theorem is used; refer Lecture 5 of [120]:

**Theorem 3.1.** *For full rank matrix $A \in \mathbb{C}^{r\times r}$ with eigenvectors $u_1, \ldots, u_r$ whose corresponding eigenvalues are $\alpha_1 \geq \cdots \geq \alpha_r$, matrix $\widetilde{A} \in \mathbb{C}^{r\times r}$ with $\mathsf{rank}(\widetilde{A}) = q < r$ defined as*

$$\widetilde{A} = [u_1 \ldots u_q]\,\mathsf{diag}(\alpha_1, \ldots, \alpha_q)\,[u_1 \ldots u_q]^* \tag{3.24}$$

*is such that*

$$\|A - \widetilde{A}\|_F = \inf_{\substack{B\in\mathbb{C}^{r\times r} \\ \mathsf{rank}(B)=q}} \|A - B\|_F = \alpha_{q+1}. \tag{3.25}$$

Due to Theorem 3.1, the optimal approximation of $\Gamma^{\mathsf{y}}$ using $WW'$ in the Frobenius norm sense is declared as $\widehat{W}\widehat{W}' = E_{1:q}\Lambda^2_{1:q}E'_{1:q}$, where columns of $E$ are eigenvectors of $\widehat{\Gamma}^{\mathsf{y}}$ whose corresponding eigenvalues in decreasing order form the diagonal of $\Lambda^2$.

Therefore, if the subscript $1:q$ to refer to the first $q$ column indices of a matrix, the estimate of $W$ sought is given by

$$\widehat{W} = E_{1:q}\Lambda_{1:q}. \tag{3.26}$$

Since $\Gamma^{\mathsf{z}}$ ought to be diagonal due to uncorrelated z, taking into account (3.21), an approximate solution for $\Gamma^{\mathsf{z}}$ is

$$\widehat{\Gamma}^{\mathsf{z}} \approx \widetilde{\mathrm{diag}}(\widehat{\Gamma}^{\mathsf{y}} - \widehat{W}\widehat{W}'), \tag{3.27}$$

where $\widetilde{\mathrm{diag}}$ refers to setting the off-diagonal elements to zero.

## 3.5 EM algorithm

Now a presentation of the Expectation-Maximization (EM) algorithm is attempted; as stated in the introduction of this chapter, it is a popular iterative method for maximum likelihood estimation.

**Note 3.1.** *In this section,* x *is assumed a discrete and univariate random variable; this is to avoid any unnecessary analytical complications otherwise leading to equivalent conclusions. E.g. the summation over* x *has to be replaced by an integration for continuous* x*. And, a summation or integration across all dimensions of* x *is to be applied had* x *been a vector random variable.*

Note the following lemma (refer §4.5 of [49]):

**Lemma 3.1.** *If a random variable* x *is* ***marginalized*** *from its joint distribution* $\mathsf{p}^{\mathsf{x},\mathsf{y}}$ *with the random variable* y*, the result is the distribution of* y*, i.e.,*

$$\mathsf{p}^{\mathsf{y}}(y) = \sum_{x} \mathsf{p}^{\mathsf{x},\mathsf{y}}(x, y). \tag{3.28}$$

The definition of log-likelihood in (3.5) may be rewritten through Lemma 3.1 as

$$\mathsf{L}(\theta) = \log_e \sum_{x} \mathsf{p}^{\mathsf{y},\mathsf{x}|\theta}(\mathcal{D}, x \mid \theta); \tag{3.29}$$

the maximization of this expression for the log-likelihood is intractable due to the summation inside the logarithm. In order to evade this situation, a dummy function $\eta(\mathsf{x})$ such that

$$\sum_{x} \eta(x) = 1; \quad \eta(\mathsf{x}) > 0 \tag{3.30}$$

is introduced and the complete log-likelihood

$$\mathsf{L}(\theta, \eta) = \log_e \sum_{x} \eta(x) \frac{\mathsf{p}^{\mathsf{y},\mathsf{x}|\theta}(\mathcal{D}, x \mid \theta)}{\eta(x)} \tag{3.31}$$

is formed. The purpose in introducing $\eta(x)$ is to seek possibilities to maximize $\mathsf{L}(\theta, \eta)$ in lieu of $\mathsf{L}(\theta)$. In that pursuit, as shown in Section B.1.1, the logarithm may be brought inside the summation, i.e.,

$$\mathsf{L}(\theta, \eta) \geq \sum_x \eta(x) \log_e \frac{\mathsf{p}^{\mathsf{y},\mathsf{x}|\theta}(\mathcal{D}, x \mid \theta)}{\eta(x)}. \tag{3.32}$$

Referring to Section B.1.2, it is possible to decompose the complete log-likelihood as

$$\mathsf{L}(\theta, \eta) \geq \mathsf{L}(\theta) + \mathsf{K}(\theta, \eta), \tag{3.33}$$

where

$$\mathsf{K}(\theta, \eta) = \sum_x \eta(x) \log_e \frac{\mathsf{p}^{\mathsf{x}|\mathsf{y},\Theta}(x \mid \mathcal{D}, \theta)}{\eta(x)}. \tag{3.34}$$

Now think of two iterative steps:

**Step 1 – Find optimal $\eta$ for a fixed $\theta$:** For a particular $\theta = \theta_i$, let $\hat{\eta}_i = \underset{\eta}{\text{argmax}}\ \mathsf{L}(\theta_i, \eta)$. Since local increase of $\mathsf{L}(\theta, \eta)$ is guaranteed by locally maximizing its global lower bound $\mathsf{L}(\theta) + \mathsf{K}(\theta, \eta)$, $\hat{\eta}_i = \underset{\eta}{\text{argmax}}\ \mathsf{K}(\theta_i, \eta)$; one may refer [87] for more details. By differentiating (3.34) with respect to $\eta(x)$, it may be found that

$$\hat{\eta}_i = \mathsf{p}^{\mathsf{x}|\mathsf{y},\Theta}(x \mid \mathcal{D}, \theta_i). \tag{3.35}$$

However, $\mathsf{K}(\theta_i, \hat{\eta}_i) = 0$ whereby

$$\mathsf{L}(\theta_i, \hat{\eta}_i) = \mathsf{L}(\theta_i). \tag{3.36}$$

Note that for a Gaussian density for y, the conditional probability for $\hat{\eta}_i$ in (3.35) is tractable.

**Step 2 – Find optimal $\theta$ for a fixed $\eta$:** Having found the locally optimal $\eta$ for a fixed $\theta$, the locally optimal $\theta$ for a fixed $\eta = \hat{\eta}_i$ is pursued. Based on (3.33) and (3.36), it may be written that

$$\theta_{i+1} = \underset{\theta}{\text{argmax}}\ \mathsf{L}(\theta, \hat{\eta}_i) \tag{3.37}$$

Note that (3.36) ensures that likelihood $\mathsf{L}(\theta_i)$ is approached in every $i$-th iteration whenever $\mathsf{L}(\theta_i, \hat{\eta}_i)$ is maximized to obtain the $i+1$-th estimate $\theta_{i+1}$; in other words, the iterations converge to a local maximum of $\mathsf{L}(\theta_i)$.

## E and M steps

The two steps arrived at above are now compiled. Suppose there is an initial guess $\theta_0$ of $\theta$. Then a local maximization of likelihood may be performed such that, in the $i$-th iteration, where $i = 1, 2, \ldots$, has the two steps:

**Step 1:** estimate $\hat{\eta}_i$ based on $\theta_i$, and

**Step 2:** locally maximize the likelihood to obtain $\theta_{i+1}$.

These steps of an iteration become explicit if we note as shown in Section B.1.3 that

$$\theta_{i+1} = \underset{\theta_i}{\text{argmax}}\ \mathsf{E}^{\mathsf{x}|\mathsf{y},\theta}\big[\log_e \mathsf{p}^{\mathsf{y},\mathsf{x}|\theta}(\mathcal{D}, x \mid \theta_i)\big]. \tag{3.38}$$

Hence, the $i$-th iteration involves:

**Step 1 (Expectation):** Evaluating the expectation $\mathsf{E}^{\mathsf{x}|\mathsf{y},\theta}\big[\log_e \mathsf{p}^{\mathsf{y},\mathsf{x}|\theta}(\mathcal{D}, x \mid \theta_i)\big]$, and

**Step 2 (Maximization):** Maximizing $\mathsf{E}^{\mathsf{x}|\mathsf{y},\theta}\big[\log_e \mathsf{p}^{\mathsf{y},\mathsf{x}|\theta}(\mathcal{D}, x \mid \theta_i)\big]$ locally with respect to $\theta_i$.

## 3.6 Basic setup of EM algorithm for factor modeling

The difference between the linear model and the factor model is obvious and wide. While access to the samples $x_n$ and $y_n$ in (3.6) of the linear model is available, x in (3.11) is assumed inaccessible in the factor model. Hence, for the factor model, the conditional distribution $\mathsf{p}^{\mathsf{y}|\mathsf{x}}$ of y given x is of interest.
Firstly, using the properties of the conditional distribution, e.g. refer §8 in Chapter 1 of [110], it could easily be shown for the factor model that $\mathsf{E}^{\mathsf{y}|\mathsf{x}}[\mathsf{y}|\mathsf{x}] = W\mathsf{x} + \mathsf{E}^{\mathsf{z}|\mathsf{x}}[\mathsf{z}]$ $= W\mathsf{x}$ because z is independent of x and has zero mean.
Secondly, the conditional variance $\Gamma^{\mathsf{y}|\mathsf{x}}$ is $\mathsf{E}^{\mathsf{y}|\mathsf{x}}[\mathsf{y}\mathsf{y}'|\mathsf{x}] - (\mathsf{E}^{\mathsf{y}|\mathsf{x}}[\mathsf{y}|\mathsf{x}])\ (\mathsf{E}^{\mathsf{y}|\mathsf{x}}[\mathsf{y}|\mathsf{x}])'$. On expansion based on (3.11), $\Gamma^{\mathsf{y}|\mathsf{x}} = \mathsf{E}^{\mathsf{y}|\mathsf{x}}\ [(W\mathsf{x} + \mathsf{z})(W\mathsf{x} + \mathsf{z})'|\mathsf{x}] - (W\mathsf{x})(W\mathsf{x})'$. Upon term-by-term expansion, due to the independence of z and x and since $\mathsf{E}^{\mathsf{z}}[\mathsf{z}] = 0$, the only surviving term will be $\mathsf{E}^{\mathsf{z}|\mathsf{x}}[\mathsf{z}\mathsf{z}' \mid \mathsf{x}]$, which becomes $\mathsf{E}^{\mathsf{z}}[\mathsf{z}\mathsf{z}'] = \Gamma^{\mathsf{z}}$.
It is also well-known that the distribution of a Gaussian random vector conditioned on another is itself Gaussian; one may refer of §4.8 of [48] or Theorem 3.10.1 of [15] among many methods to verify it. Therefore, the Gaussian probability density of y | x with parameters $\theta = \{W, \Gamma^{\mathsf{z}}\}$ for the factor model may be written as

$$\mathsf{p}^{\mathsf{y}|\mathsf{x},\theta}(y \mid x, \theta) = \mathcal{N}(y \mid W\mathsf{x}, \Gamma^{\mathsf{z}}). \tag{3.39}$$

Note that $\theta$ in $\mathsf{p}^{\mathsf{y}|\mathsf{x},\theta}$ refers to the availability of the set of parameters; the distribution is conditioned only on x. Based on the discussions in Section 3.1, the conditional probability density $\mathsf{p}^{\mathsf{y}|\mathsf{x},\theta}$ underpins the likelihood of the factor model. As with (3.1), the dataset $\mathcal{D}$ is considered to consist of the iid samples $y_l$, $l = 1, \ldots, n$ of y. The likelihood of the dataset is

$$\mathsf{p}^{\mathsf{y}|\mathsf{x},\theta}(\mathcal{D} \mid \mathsf{x}, \theta) = \prod_{l=1}^{n} \mathsf{p}^{\mathsf{y}|\mathsf{x},\theta}(y_l \mid \mathsf{x}, \theta). \tag{3.40}$$

Using Theorem B.1 known as Bayes theorem, $\mathsf{p}^{\mathsf{y},\mathsf{x}|\theta}(\mathcal{D}, \mathsf{x} \mid \theta) = \mathsf{p}^{\mathsf{y}|\mathsf{x},\theta}(\mathcal{D} \mid \mathsf{x}, \theta)\mathsf{p}^{\mathsf{x}}(\mathsf{x})$. If it is assumed that the distribution $\mathsf{p}^{\mathsf{x}}(x)$ to be independent of $\theta$, then (3.38) of the EM-algorithm reduces to iteratively solving

$$\theta_{i+1} = \underset{\theta_i}{\text{argmax}}\ \mathsf{E}^{\mathsf{x}|\mathsf{y},\theta}\big[\log_e \mathsf{p}^{\mathsf{y}|\mathsf{x},\theta}(\mathcal{D} \mid \mathsf{x}, \theta_i)\big]. \tag{3.41}$$

From the above three equations, it may be written that

$$\theta_{i+1} = \underset{\theta_i}{\text{argmax}}\ \mathsf{E}^{\mathsf{x}|\mathsf{y},\theta}[\mathsf{f}(\theta_i, \mathsf{x})], \tag{3.42}$$

$$\begin{aligned}\mathsf{f}(\theta_i, \mathsf{x}) &= \log_e \mathsf{p}^{\mathsf{y}|\mathsf{x},\theta}(\mathcal{D} \mid \mathsf{x}, \theta_i) \\ &= \sum_{l=1}^{n} \log_e \mathcal{N}(y_l \mid W_i\mathsf{x}, \Gamma_i^{\mathsf{z}}),\end{aligned} \tag{3.43}$$

where the parameters

$$\theta_i \triangleq \{W_i, \Gamma_i^{\mathsf{z}}\} \tag{3.44}$$

correspond to the $i^{th}$ iteration.

## 3.7 Two steps of EM algorithm for factor modeling

In light of the discussion in Section 3.5 and the parameter update equations of (3.42), it may be stated that the $i$-th iteration of the factor model estimation alternates between:

1. **Expectation-step** Evaluate the expectation $\mathsf{E}^{\mathsf{x}|\mathsf{y},\theta}\big[\mathsf{f}(\theta_i, \mathsf{x})\big]$, and
2. **Maximization-step** Update $\theta_{i+1} \leftarrow \theta_i$ by maximizing $\mathsf{E}^{\mathsf{x}|\mathsf{y},\theta}\big[\mathsf{f}(\theta_i, \mathsf{x})\big]$ with respect to $\theta_i$.

To proceed note that in (3.43) that $\mathsf{f}(\theta_i, \mathsf{x}) = -n \log_e(\det(\Gamma_i^{\mathsf{z}})) - \sum_{l=1}^{n} \mathsf{M}(y_l, W_i\mathsf{x}, \Gamma_i^{\mathsf{z}})$, where for any compatible vectors $a$, $b$ and matrix $C$

$$\mathsf{M}(a, b, C) = (a-b)'C^{-1}(a-b), \tag{3.45}$$

whose expansion gives $\mathsf{M}(y_l, W_i\mathsf{x}, \Gamma_i^{\mathsf{z}}) = y_l'(\Gamma_i^{\mathsf{z}})^{-1}y_l - 2y_l'(\Gamma_i^{\mathsf{z}})^{-1}W_i\mathsf{x} + \mathsf{tr}((\Gamma_i^{\mathsf{z}})^{-1}W_i\mathsf{x}\mathsf{x}'W_i')$. Note the presence of terms with random variables $\mathsf{x}$ and $\mathsf{x}\mathsf{x}'$ in $\mathsf{M}(y_l, W_i\mathsf{x}, \Gamma_i^{\mathsf{z}})$. Therefore, the EM algorithm, as a result of this expansion of $\mathsf{M}(y_l, W_i\mathsf{x}, \Gamma_i^{\mathsf{z}})$, will involve alternating between:

1. **Expectation-step** Evaluate, for $l = 1, \ldots, n$,

$$\begin{aligned}\langle x \rangle_{i,l} &\triangleq \mathsf{E}^{\mathsf{x}|\mathsf{y},\theta}[\mathsf{x} \mid y_l, \theta_i], \\ \langle xx' \rangle_{i,l} &\triangleq \mathsf{E}^{\mathsf{x}|\mathsf{y},\theta}[\mathsf{x}\mathsf{x}' \mid y_l, \theta_i],\end{aligned} \tag{3.46}$$

   where $\langle x \rangle_{i,l} \in \mathbb{R}^q$ and $\langle xx' \rangle_{i,l} \in \mathbb{R}^{q \times q}$, and

2. **Maximization-step** Update $\theta_{i+1} \leftarrow \theta_i$ by maximizing $\mathsf{f}(\theta_i, \mathsf{x})$ with respect to $\theta_i$, where $\mathsf{x}$ and $\mathsf{x}\mathsf{x}'$ are replaced by their corresponding *a posteriori* estimates, i.e., in (3.43)

$$\begin{aligned}\mathsf{x} &\Longleftarrow \langle x \rangle_{i,l} \\ \mathsf{x}\mathsf{x}' &\Longleftarrow \langle xx' \rangle_{i,l}.\end{aligned} \tag{3.47}$$

The following analysis between (3.48) and (3.52) is inspired by [14].

### 3.7.1 E-step

Note that $\langle x \rangle_{i,l}$ is the mean of the Gaussian distribution $\mathsf{p}^{\mathsf{x}|\mathsf{y},\theta}(\mathsf{x} \mid y_l, \theta_i)$, which is evaluated in Appendix B.2 to be

$$\begin{aligned} \langle x \rangle_{i,l} &= \Omega_i W_i' (\Gamma_i^{\mathsf{z}})^{-1} y_l, \\ \Omega_i &= (I_q + W_i' (\Gamma_i^{\mathsf{z}})^{-1} W_i)^{-1}, \end{aligned} \tag{3.48}$$

where $\Omega_i \in \mathbb{R}^{q \times q}$. From the classical relation of mean and covariance of any distribution, it is known that $\langle xx' \rangle_{i,l}$ is the sum of $\langle x \rangle_{i,l} \langle x' \rangle_{i,l}$ and covariance of $\mathsf{x} \mid y_l, \theta_i$, i.e.,

$$\langle xx' \rangle_{i,l} = \langle x \rangle_{i,l} \langle x' \rangle_{i,l} + \Omega_i. \tag{3.49}$$

This completes the E-step of the EM Algorithm.

### 3.7.2 M-step

Towards the M-step of the EM algorithm, the substitutions in (3.47) give

$$\begin{aligned} \mathsf{E}^{\mathsf{x}|\mathsf{y},\theta}[\mathsf{f}(\theta_i, \mathsf{x})] = -n \log_e(\det(\Gamma_i^{\mathsf{z}})) - \sum_{l=1}^{n} \mathsf{tr}((\Gamma_i^{\mathsf{z}})^{-1} W_i \langle xx' \rangle_{i,l} W_i') \\ - 2 y_l' (\Gamma_i^{\mathsf{z}})^{-1} W_i \langle x \rangle_{i,l} + y_l' (\Gamma_i^{\mathsf{z}})^{-1} y_l. \end{aligned} \tag{3.50}$$

Now $\mathsf{E}^{\mathsf{x}|\mathsf{y},\theta}[\mathsf{f}(\theta_i, \mathsf{x})]$ may be maximized to update the parameters $W_i$ and $\Gamma_i^{\mathsf{z}}$:

**Update $W_i$ :** The problem that has to be solved is

$$\begin{aligned} W_{i+1} &= \underset{W_i}{\mathsf{argmax}}\ \mathsf{E}^{\mathsf{x}|\mathsf{y},\theta}[\mathsf{f}(\theta_i, \mathsf{x})] \\ &= \underset{W_i}{\mathsf{arg}} \left( \frac{\partial}{\partial W_i} \mathsf{E}^{\mathsf{x}|\mathsf{y},\theta}[\mathsf{f}(\theta_i, \mathsf{x})] = 0 \right). \end{aligned} \tag{3.51}$$

It is easy to see using matrix differentiation rules, e.g., refer [98], that

$$\frac{\partial}{\partial W_i} \mathsf{E}^{\mathsf{x}|\mathsf{y},\theta}[\mathsf{f}(\theta_i, \mathsf{x})] = -\sum_{l=1}^{n} 2(\Gamma_i^{\mathsf{z}})^{-1} W_i \langle xx' \rangle_{i,l} - 2(\Gamma_i^{\mathsf{z}})^{-1} y_l \langle x' \rangle_{i,l},$$

which when equated to zero gives

$$W_{i+1} = \left( \sum_{l=1}^{n} y_l \langle x' \rangle_{i,l} \right) \left( \sum_{l=1}^{n} \langle xx' \rangle_{i,l} \right)^{-1}. \tag{3.52}$$

**Update $\Gamma_i^{\mathsf{z}}$ :** The access to the updated $W_{i+1}$ is available and if

$$v_{i,l} = W_{i+1} \langle x \rangle_{i,l},$$

then $\mathsf{E}^{\mathsf{x}|\mathsf{y},\theta}[\mathsf{f}(\theta_i,\mathsf{x})] = -n\log_e(\det(\Gamma_i^{\mathsf{z}})) - \sum_{l=1}^{n}\mathsf{M}(y_l, v_{i,l}, \Gamma_i^{\mathsf{z}})$. Now, consider the update

$$\Gamma_{i+1}^{\mathsf{z}} = \underset{W_i}{\arg}\left(\frac{\partial}{\partial W_i}\mathsf{E}^{\mathsf{x}|\mathsf{y},\theta}[\mathsf{f}(\theta_i,\mathsf{x})] = 0\right). \tag{3.53}$$

For $\Gamma_i^{\mathsf{z}} = \mathsf{diag}(\gamma_i^{\mathsf{z}_1},\ldots,\gamma_i^{\mathsf{z}_r})$ it can be seen that

$$\mathsf{E}^{\mathsf{x}|\mathsf{y},\theta}[\mathsf{f}(\theta_i,\mathsf{x})] = -n\sum_{k=1}^{r}[\log_e(\gamma_i^{\mathsf{z}_k}) + \frac{1}{\gamma_i^{\mathsf{z}_k}}a_i^{\mathsf{z}_k}]$$

where

$$a_i^{\mathsf{z}_k} = \frac{1}{n}\sum_{l=1}^{n}(y_l - v_{i,l})^2.$$

Then, $\partial\mathsf{E}^{\mathsf{x}|\mathsf{y},\theta}[\mathsf{f}(\theta_i,\mathsf{x})]/\partial\gamma_i^{\mathsf{z}_k} = 0$ at

$$\gamma_{i+1}^{\mathsf{z}_k} = a_i^{\mathsf{z}_k}. \tag{3.54}$$

### Factor model estimation via EM algorithm

Given the dataset, in Algorithm 2, the results of the analysis of the iterative parametric estimation of the factor model using the EM algorithm are summarized.

**Algorithm 2**: EM algorithm for the factor model

**Input**: $\mathcal{D} = \{y_l\}, l = 1,\ldots n$
**Output**: $\widehat{W}, \widehat{\Gamma}^{\mathsf{z}} = \mathsf{diag}(\hat{\gamma}^{\mathsf{z}_1},\ldots,\hat{\gamma}^{\mathsf{z}_r})$
initialize $i = 0$;
initialize randomly $W_i, \Gamma_i^{\mathsf{z}}$ ;
**do**
  **E-step**:
  **for** $l = 1$ *to* $n$ **do**
    compute
      $\langle x\rangle_{i,l}$ using (3.48);
      $\langle xx'\rangle_{i,l}$ using (3.49);
  **end**
  **M-step**: update
    $W_{i+1}$ using (3.52);
    $\gamma_{i+1}^{\mathsf{z}_k}\ \forall k = 1,\ldots,r$ using (3.54);
  $i \longleftarrow i+1$;
  $\epsilon \longleftarrow \mathsf{E}^{\mathsf{x}|\mathsf{y},\theta}[\mathsf{f}(\theta_i,\mathsf{x})] - \mathsf{E}^{\mathsf{x}|\mathsf{y},\theta}[\mathsf{f}(\theta_{i-1},\mathsf{x})]$ using (3.50);
**while** $\epsilon > 10^{-8}$ **and** $i < 20$ ;
$\widehat{W} \longleftarrow W_i, \hat{\gamma}^{\mathsf{z}_k} \longleftarrow \gamma_i^{\mathsf{z}_k}\ \forall k = 1,\ldots,r$;

A major drawback of the EM algorithm is the possibility that the estimation might get trapped in a local maximum of the log-likelihood and hence might require random restarts or other heuristic measures to be more certain regarding the estimates.

## 3.8 Summary

Two possibilities of modeling an $r$-dimensional measured vector random variable y were considered, viz., $(i)$ the linear model where a measured variable $x \in \mathbb{R}^q, q < r$ is transformed to y and $(ii)$ the factor model where a latent $q$-variate random variable x is transformed to y. Essentially, a factor model transforms a latent vector random variable of known probability distribution to a measured vector random variable of higher dimensionality that is perturbed by independent and uncorrelated noise. For the linear model, an efficient estimator of the transformation matrix was presented; whereas for the factor model there is no unique transformation. However, by restricting the variances unique to each of the measured variable, it is possible to estimate meaningful transformations. Thus, from a parametric modeling perspective, the transformation matrix and the unique variances are the parameters of the factor model.
In order to estimate the factor model parameters, two approaches based on the principle of maximum likelihood were discussed: The analytical estimation approach involves approximating the covariance structure of the measured variables using that of the transformed variables. For the iterative approach based on the EM algorithm, the log-likelihood function being lower bound by the *a posteriori* expectation of the logarithm of the joint probability density of the measured variables and the latent variables was exploited. Starting from guesses of the parameters, the EM algorithm maximizes the complete log-likelihood function of the latent variables and the measured variables by iteratively converging to the log-likelihood with every update of the parameters.

# Chapter 4

# Dynamic and spectral factor models

Recall the **linear model** $\mathsf{y} = Wx + \mathsf{z}$ of (3.7). The intention there was to linearly relate the set of $r$-variate independent samples $\{y_l\}$, $l = 1, \ldots, n$, which are thought to be **realizations** of a **vector random variable** y to the corresponding set of $q$-variate samples $\{x_l\}$, $l = 1, \ldots, n$. However, both $\{x_l\}$ and $\{y_l\}$ were measured and available in a given dataset.

Now, contrast the linear model with the **factor model** $\mathsf{y} = W\mathsf{x} + \mathsf{z}$ of (3.11), where x is a $q$-variate hidden or **latent vector random variable** while y is an $r$-variate **measured vector random variable**.

The noticeable similarities between the linear model and the factor model are assumptions that $r > q$, transformation matrix $W$ is non-random, and z is an $r$-variate vector random variable with uncorrelated components.

In this chapter, the assumption of Chapter 2 that a sequence of vector random variables $\{\mathsf{y}_t\}$ are temporally correlated is underpinned. Thence, based on the motivations presented in Chapter 1, existence of a $q$-variate time series $\{\mathsf{x}_t\}$ which gets transformed by a **non-stochastic** matrix $\{W_t\}$ to obtain an $r$-variate time series $\{\mathsf{y}_t\}$ $\forall t \in \mathbb{Z}$ is assumed. The objective of this chapter is to define such a model and enable it for learning problems.

In Section 4.1, the time-domain definition of the **dynamic factor model** and the **commonalities** it represents are defined. In doing so, the assumptions made with respect to the model are emphasized and the relations between the parameters of the model, viz., the ***acvf***s of the **measured**, **latent**, and **idiosyncratic time series** are analyzed. Then, the dynamic factor model is defined.

In Section 4.2, the analysis is switched to **Fourier-domain**: Frequency-domain counterparts of the measured, latent, and idiosyncratic time series are defined and the frequency-domain equivalent of the dynamic factor model called the **spectral factor model** is defined.

The following situates the developments in this chapter with respect to the state-of-the-art:

- ▷ Definition 4.2 and Definition 4.3 define the dynamic and spectral factor models, respectively.
  These definitions include all model assumptions and the model objectives. In [100, 104, 43, 36], both time and frequency domain analyses are called "dynamic

factor model"; for convenience, "spectral factor model" is a term introduced here to emphasize the frequency-domain analysis. The properties and assumptions of interacting linear processes of the model used here are standard practice in literature.

- ▷ Definition 4.1 introduces commonalities; relations (4.9) and (4.18) state the criterion for inheriting them from the measured variables.
  Unlike in the existing literature, cross-correlations are emphasized in the definitions here of dynamic and spectral factor models through the concept of commonalities. Also, the existing literature does not specifically relate the dynamic transformation to the commonalities nor its maximal inheritance as defined here.

## 4.1 Dynamic factor model

A **multivariate time series model** is to be designed where a $q$-variate **latent time series** $\{\mathsf{x}_t\}$ is transformed by a sequence of $r \times q$ non-stochastic **transformation matrices** $\{W_t\}$ to a measured $r$-variate time series $\{\mathsf{y}_t\}$, where $r > q$. Only those actions of $W_t$ in transforming $\mathsf{x}_t$ to $\mathsf{y}_t$ that is intuitively appealing, theoretically valid, practically feasible, and analytically sound will be allowed. In the simplest of forms, such a time series model may be written as $\mathsf{y}_t = f(W_t, \mathsf{x}_t) + \mathsf{z}_t$, where $f$ is some linear function of $\{W_t\}$ and $\{\mathsf{x}_t\}$ $\forall t \in \mathbb{Z}$, where $\{\mathsf{z}_t\}$ is a vector random variable independent of $\{\mathsf{x}_t\}$ that offers itself as the error in the transformation.

It is also to be ensured that the transformation will take advantage of the **frequency-based** techniques discussed in Chapter 2. In that case, existence of the **spectral density function** of $f(W_t, \mathsf{x}_t)$ is a necessity. As discussed in Section 2.4, for a weakly stationary vector random variable sequence $\{\mathsf{x}_t\}$, the Fourier transform of any linear relation $f(W_t, \mathsf{x}_t)$ between $W_t$ and $\mathsf{x}_t$ does not exist for no guarantee of $\sum_{t\in\mathbb{Z}} |f(W_t, \mathsf{x}_t)| < \infty$ could be made. But as long as the *acvf* of $f(W_t, \mathsf{x}_t)$ exists and that *acvf* is absolutely summable as in (2.20), techniques of Fourier transform could be pushed.

Take a look at one of the simplest linear operations for $f(W_t, \mathsf{x}_t) \triangleq \mathsf{v}_t$, which is an $r$-variate vector random obtained when $W_t$ is convolved with $\mathsf{x}_t$, i.e., $\forall t \in \mathbb{Z}$,

$$\mathsf{v}_t = \sum_{j\in\mathbb{Z}} W_j \mathsf{x}_{t-j}. \tag{4.1}$$

If both $\{W_t\}$ and the *acvf* $\{\Gamma_h^{\mathsf{x}}\} \forall h \in \mathbb{Z}$ of $\mathsf{x}_t$ are absolutely summable, then $\mathsf{v}_t$ according to (4.1) exists; refer Theorem 2.7.1 of [18] for this result. Let $r$-variate linear processes $\{\mathsf{y}_t\}$, $\{\mathsf{v}_t\}$, and $\{\mathsf{z}_t\}$ be related according to

$$\mathsf{y}_t = \mathsf{v}_t + \mathsf{z}_t. \tag{4.2}$$

Further, if $\mathsf{v}_t$ and $\mathsf{z}_t$ are independent, then they have their *acvf*s related as

$$\Gamma_h^{\mathsf{y}} = \Gamma_h^{\mathsf{v}} + \Gamma_h^{\mathsf{z}}, \tag{4.3}$$

where $\forall h \in \mathbb{Z}$ is the lag parameter of the *acvf*s. It is further assumed that

$$\mathsf{rank}(\Gamma_h^{\mathsf{z}}) = r. \tag{4.4}$$

Thus, the measured $r$-variate vector random variable $\mathsf{y}_t$ is assumed to be obtained by adding two independent $r$-variate vector random variables $\mathsf{v}_t$ and $\mathsf{z}_t$. And, $\mathsf{v}_t$ is a **dynamic transformation** of a latent $q$-variate vector random variable $\mathsf{x}_t$ as per (4.1).

Recall that Chapter 1 hoped to dynamically transform a latent vector random variable of known or presumed characteristics and that the dynamic transformation is the one that is unknown. The similarity between the form of dynamic transformation in (4.1) and the form of **linear process** defined in Definition 2.11 is evident. This similarity entices to assume that $\{\mathsf{x}_t\}$ is a $q$-variate zero mean white noise and that

$$\sum_t |W_t| < \infty \quad \forall t \in \mathbb{Z}. \tag{4.5}$$

It will deliver a $\{\mathsf{v}_t\}$ that is a linear process resulting from a linear transformation of $\{\mathsf{x}_t\}$ by the sequence of parameters $\{W_t\}$. Further, to simplify the analysis, it is assumed that $\{\mathsf{x}_t\}$ is a unit variance white noise process, i.e.,

$$\Gamma^{\mathsf{x}}_h = I_q \quad \forall h \in \mathbb{Z}. \tag{4.6}$$

Such an assumption is admissible because it is not intended to estimate $\Gamma^{\mathsf{x}}_h$ anyway. Then, referring back again to Definition 2.11, it is easy to see that

$$\Gamma^{\mathsf{v}}_h = \sum_{j \in \mathbb{Z}} W_{j+h} W'_j \quad \forall h \in \mathbb{Z}, \tag{4.7}$$

and

$$\mathsf{rank}(\Gamma^{\mathsf{v}}_h) = q. \tag{4.8}$$

The objective is to enable $\{\mathsf{v}_t\}$ to maximally inherit the commonalities in the measured time series $\{\mathsf{y}_t\}$. And, in Chapter 1, commonalities of the measured time series $\{\mathsf{y}_t\}$ were regarded to be the temporal covariations of the $r$ measured components of $\mathsf{y}_t = [\mathsf{y}_{1_t}\, \mathsf{y}_{2_t}\, \ldots\, \mathsf{y}_{r_t}]'$. A good measure of the commonalities should be the expected value of a suitable function combining the $r$ random variables, e.g., their mean product. Now, the following definition is arrived at:

**Definition 4.1.** *For a weakly stationary time series $\{\mathsf{y}_t\}$, the **commonalities** are the off-diagonal elements of its acvf $\Gamma^{\mathsf{y}}_h$.*

**Appropriateness of Definition 4.1**: Cross-covariances describe all the mutual characteristics of the components of a zero-mean multivariate time series linear process. The pairwise commonality between any two components $\mathsf{y}_{i_t}$ and $\mathsf{y}_{j_t}$ are the off-diagonal elements of the *acvf* $\Gamma^{\mathsf{y}}_h$ of the measured time series $\{\mathsf{y}_t\}$.

In Chapter 1, it was envisaged to estimate parameters of a model that will maximize the measured commonalities. Earlier in this section, the role of measured cross-covariances as a suitable measure of the commonalities was confirmed. As a result, the commonalities are retained in the cross-covariance terms of $\Gamma^{\mathsf{v}}_h$ upon a dynamic transformation of $\{\mathsf{y}_t\}$ to $\{\mathsf{v}_t\}$ as discussed earlier in this section using $\{W_t\}$. the

proposed measure for the inheritance of the commonalities of $\Gamma_h^{\mathsf{y}}$ using $\Gamma_h^{\mathsf{v}}$ is the sum of square differences of the covariances across all measured dimensions and lags, i.e.,

$$g = \sum_{h \in \mathbb{Z}} \|\Gamma_h^{\mathsf{v}} - \Gamma_h^{\mathsf{y}}\|_F^2. \tag{4.9}$$

**Appropriateness of** $g$: These reason the choice of the quality of approximation of the commonalities:

1. Since $\Gamma_h^{\mathsf{y}} - \Gamma_h^{\mathsf{v}}$ is positive definite and $\Gamma_h^{\mathsf{v}}$ is of lower rank than positive definite $\Gamma_h^{\mathsf{y}}$, trace of $\Gamma_h^{\mathsf{v}}$ will also be lower than $\Gamma_h^{\mathsf{y}}$, i.e., the unique variance terms of $\Gamma_h^{\mathsf{y}}$ will be also affected and has to approximated in a low-rank sense.

2. It has a direct equivalence in the frequency domain; this is through relation (4.18).

3. It provides the properties of the residual process $\mathsf{y}_t - \mathsf{v}_t$ easily; this is due to (5.7) enabling (5.12) and (5.13).

4. Its analytical conveniences and properties are well-known; refer [84].

The optimal parameters using the measure in (4.9), with reference to (4.7) and (4.3), are given by

$$\widetilde{W}_t, \widetilde{\Gamma}_h^{\mathsf{z}} := \underset{W_t, \Gamma_h^{\mathsf{z}}}{\operatorname{argmin}}\ g. \tag{4.10}$$

Since orthogonal rotations of $W_j\, \forall j \in \mathbb{Z}$ lead to same $\Gamma_h^{\mathsf{v}}\, \forall h \in \mathbb{Z}$ in (4.7), there is no unique solution to the minimization problem (4.10).

**Note 4.1.** *In this thesis, the choice of the latent dimensionality $q$ is made arbitrarily. No theoretical effort is spent towards the important problem of determining an optimal $q$. In the experiments, however, performance of the dynamic factor model across $q$ will be evaluated. Asymptotic properties of the dynamic factors in the latent space with respect to larger measured number of samples and dimensionality is available in [37].*

The dynamic model defined by (4.1) - (4.8) implies that the measured vector random variable $\mathsf{y}_t$ is an addition of an independent linear process to another linear process formed by the dynamic transformation of a lower dimensional unit variance white noise. Apart from the time series aspect of the measured variables, the dynamic model bears much resemblance to the factor model: In (3.11), a latent vector random variable is transformed to an unobserved higher dimensional vector random variable which is perturbed by independent noise resulting in the measured vector random variable. This similarity invites the following definition:

**Definition 4.2.** *Let a $q$-variate latent zero mean unit variance white noise $\{\mathsf{x}_t\}$ be dynamically transformed by non-stochastic $\{W_t\}$ to an $r$-variate linear process $\{\mathsf{v}_t\}$. Suppose an independent $r$-variate linear process $\{\mathsf{z}_t\}$ is added to $\{\mathsf{v}_t\}$ to obtain an $r$-variate weakly stationary measured time series $\{\mathsf{y}_t\}$. Such a vector time series model which satisfies the conditions (4.1) - (4.8) and solves (4.10) is called a* ***dynamic factor model****.*

**Model assumptions**: Recall the original list of model assumptions in Section 1.2. With the dynamic factor model as per Definition 4.2, they may be concretely restated as follows:

1. the measured time series is a linear process,
2. the measured time series is a dynamic transformation of a zero-mean unit variance white noise of a dimensionality lower than that of the measured time series,
3. the *acvf* of the dynamically transformed process is a low rank approximation in a Frobenius norm sense of the measured *acvf*,
4. the residual time series is a linear process independent of the latent time series and has finite unique variances.

## 4.2 Spectral factor model

The objective of the dynamic factor model is to estimate the optimal parameters that maximize the commonalities of the measured time series $\{\mathsf{y}_t\}$ inherited by the unobserved time series $\{\mathsf{v}_t\}$. However, what stands out is the concern regarding how to perform such a maximization that adheres to the transformation of $\Gamma_h^{\mathsf{v}}$ to $W_t$ as per (4.7).

**Motivation for a Fourier-domain approach**: It is clear from (4.7) that $\Gamma_h^{\mathsf{v}}$ is the correlation of the sequence $\{W_t\}$ in the time domain. According to the **autocorrelation theorem** of Fourier transform, which is also known as the Wiener-Khinchin-Einstein theorem, the autocorrelation of a function and power spectrum of that function are Fourier transform pairs; refer §10.1.1 in [77]. Then, $\forall h \in \mathbb{Z}, -\frac{1}{2} \leq \omega < \frac{1}{2}$

$$\Gamma_h^{\mathsf{v}} \overset{\mathcal{F}}{\longleftrightarrow} \mathsf{S}^{\mathsf{v}}(\omega) = \mathbf{W}(\omega)\mathbf{W}^*(\omega), \tag{4.11}$$

where $W_t \overset{\mathcal{F}}{\longleftrightarrow} \mathbf{W}(\omega)$ refers to the **discrete time Fourier transform** as per Definition 2.15 and $\mathsf{S}^{\mathsf{v}}(\omega)$ is the spectral density function of $\mathsf{v}$ whose $(i,j)$-th matrix element is $\mathsf{s}^{\mathsf{v}_i,\mathsf{v}_j}(\omega)$ $\forall h \in \mathbb{Z}, -\frac{1}{2} \leq \omega < \frac{1}{2}$. Note that applying the definition of the Fourier transform to the relation (4.3) gives

$$\mathsf{S}^{\mathsf{y}}(\omega) = \mathsf{S}^{\mathsf{v}}(\omega) + \mathsf{S}^{\mathsf{z}}(\omega). \tag{4.12}$$

It is further assumed that

$$\mathsf{rank}(\mathsf{S}^{\mathsf{z}}(\omega)) = r. \tag{4.13}$$

Also, it emerges from Property 2.7 that $|\mathsf{S}^{\mathsf{v}}(\omega)| = |\mathbf{W}(\omega)\mathbf{W}^*(\omega)| \neq 0 \quad \forall \omega \in [-\frac{1}{2}, \frac{1}{2}]$, i.e.,

$$|\mathbf{W}(\omega)| \neq 0 \quad \forall \omega \in [-\frac{1}{2}, \frac{1}{2}]. \tag{4.14}$$

For a finite $\tau$-length realization of the process, combining (4.1) and (4.2) gives $y_t = \sum_{j=1}^{\tau} W_j x_{t-j} + z_t$. As per Definition 2.17, the **discrete Fourier transform** of these

realizations are $\mathbf{y}(\omega_k) = \mathbf{W}(\omega_k)\mathbf{x}(\omega_k) + \mathbf{z}(\omega_k)$, at frequencies $-\frac{1}{2} \leq \omega_k < \frac{1}{2}$, $k = 1, \ldots, \tau$.

To proceed, recall Theorem 2.5 where the spectral density function $\mathsf{S}^{\mathsf{y}}(\omega_j)$ becomes the covariance matrix of the **complex Gaussian distribution** of the discrete Fourier transform components sufficiently close to $\omega_j$ so that

$$\mathsf{y}(\omega_j) = \mathbf{W}(\omega_j)\mathsf{x}(\omega_j) + \mathsf{z}(\omega_j), \tag{4.15}$$

are complex Gaussian vector random variables $\{\mathsf{x}(\omega_j), \mathsf{y}(\omega_j), \mathsf{z}(\omega_j)\}$. Then, $\mathsf{x}(\omega_j)$ becomes a vector random variable whose complex Gaussian distribution has a covariance matrix

$$\mathsf{S}^{\mathsf{x}}(\omega_j) = I_q. \tag{4.16}$$

Note that the equivalent for (4.8) is

$$\mathsf{rank}(\mathsf{S}^{\mathsf{v}}(\omega)) = q. \tag{4.17}$$

Note that the form in (4.15) is very much reminiscent of the factor model, where $\mathsf{x}(\omega_j)$ is a latent factor of known Fourier characteristics transformed by a non-stochastic $\mathbf{W}(\omega_j)$ perturbed by independent vector random variable $\mathsf{z}(\omega_j)$. Hence, the motivation for pursuing a Fourier domain approach for the solution of the dynamic factor model is the possibility that classical factor model methods as reviewed in Chapter 3 might be availed to solve for $W_t$ in (4.1).

**Dynamic factor model equivalent in the Fourier-domain**: Armed with a Gaussian probability distribution for the measured discrete Fourier transform $\mathbf{y}(\omega_j)$, the maximum likelihood estimation for the factor modeling should follow naturally. In that pursuit, the hope is to attain a relation connecting the maximum likelihood spectral density function $\mathsf{S}^{\mathsf{v}}(\omega_j)$, $\mathsf{S}^{\mathsf{z}}(\omega_j)$, and $\mathsf{S}^{\mathsf{y}}(\omega_j)$. Certainly, their inverse Fourier transform should yield their unique *acvf*s of $\Gamma_h^{\mathsf{v}}$, $\Gamma_h^{\mathsf{z}}$, and $\Gamma_h^{\mathsf{y}}$, respectively, which are also of interest.

However, some estimate of the parameters of the interest is not satisfactory because the objective is to find those that will maximize the commonalities. Next, applying Theorem 2.1, the sum in (4.9) may be written as

$$g = \int_{-\frac{1}{2}}^{\frac{1}{2}} \|\mathsf{S}^{\mathsf{y}}(\omega) - \mathsf{S}^{\mathsf{v}}(\omega)\|_F^2 \, \mathsf{d}\omega. \tag{4.18}$$

Thereafter, in line with the arguments for (4.10), it could be deduced that the optimal parameters in the Fourier domain are

$$\widetilde{\mathbf{W}}(\omega), \widetilde{\mathsf{S}}^{\mathsf{z}}(\omega) := \underset{\mathbf{W}(\omega), \mathsf{S}^{\mathsf{z}}(\omega)}{\mathsf{argmin}} \; g. \tag{4.19}$$

Since orthogonal rotations of $\mathbf{W}(\omega)$ lead to same $\mathsf{S}^{\mathsf{v}}(\omega)$ in (4.11), there is no unique solution to the minimization problem in (4.19).

Due to the Fourier domain similarities of the dynamic factor model with the classical factor model, the following definition arrives:

**Definition 4.3.** *Let a $q$-variate latent zero mean unit variance discrete Fourier transform vector random variable* $\mathsf{x}(\omega_j)$ *be transformed by non-stochastic* $\mathbf{W}(\omega_j)$ *to an $r$-variate zero mean discrete Fourier transform vector random variable* $\mathsf{v}(\omega_j)$*. Suppose an $r$-variate discrete Fourier transform vector random variable* $\mathsf{z}(\omega_j)$ *that is independent of* $\mathsf{x}(\omega_j)$ *is added to* $\mathsf{v}(\omega_j)$ *resulting in an $r$-variate measured vector random variable* $\mathsf{y}(\omega_j)$*. Such a vector discrete Fourier transform model which satisfies the conditions (4.11) - (4.17) and solves (4.19) is called a* ***spectral factor model****.*

**Model assumptions**: Recall the list of model assumptions in Section 1.2 and subsequent to Definition 4.2. With the spectral factor model as per Definition 4.3, they may be restated as follows:

1. the measured discrete Fourier transform components ('spectra') are asymptotically Gaussian within small subbands,

2. the measured spectra are transformations of a zero-mean unit variance Gaussian spectra of lower dimensionality,

3. the spectral density function of the transformed spectra is a low rank approximation in a Frobenius norm sense of the measured spectra,

4. the residual spectra is a Gaussian independent of the latent spectra and has finite unique variances.

**Basic goal of the spectral factor model**: The dynamic and spectral factor models and the accompanying problem of maximization of the commonalities of a measured multivariate time series were defined in this chapter. The maximum commonalities transformation matrix is the best approximation, in a Frobenius norm sense, using a lower number of variables of the cross-covariances of the measured time series. Since there exist problems, as the examples in Section 1.2 show, where commonalities will aid learning, the goal is to adapt the transformation matrix for classification and prediction problems. This will be done by deriving the required parameters of the spectral factor model in Chapter 5 using the principle of maximum likelihood fostered by the constraint of maximum commonalities. Classification and prediction algorithms will be developed in Chapter 6.

## 4.3 Summary

In this chapter, the dynamic factor model and the spectral factor model were introduced. Conceptually, the dynamic and spectral factor models transform a latent vector random process by maximally inheriting the measured commonalities. It was discussed why the cross-covariances could be called as commonalities. A criterion based on approximating the *acvf*s in a Frobenius norm sense such that it will correspond to maximizing the commonalities was formulated. It was claimed that the inheritance of the commonalities of a vector random process by another increases if the Frobenius norm of the difference between their autocovariance functions across all lags decreases; an equivalent criterion for the spectral density function was also formulated. It was

assessed how the spectral factor model for measured discrete Fourier transform components in a 'small' bandwidth resembles the classical factor model.
The impediments of complex-valued parametric estimation should be overcame to extend the classical factor model estimation techniques reviewed in Chapter 3 to maximize the commonalities of the spectral factor model.

# Chapter 5

# Maximum likelihood commonalities

The objective of this chapter is to solve the maximization problem defined as part of the spectral factor model in Section 4.2. That problem refers to maximizing the commonalities retained by the latent spectral factor transformation. It will be shown that its solution requires estimating the **maximum likelihood spectral density function**. Two methods are developed to arrive at the maximum likelihood spectral density function estimates: The first method is analytical and is the topic of Section 5.1; it gains traction from the estimation procedures summarized in Section 3.4. The second method discussed in Section 5.2 is iterative; it is along the lines of the EM algorithm presented in Section 3.6.

In Section 5.1, as part of the analytical method, optimal parameters of the spectral factor model are made available in (5.10) and (5.13). In order to arrive at those results, the expression for the log-likelihood function of the spectral factor model is written. Due to difficulties in maximizing such a real-valued function of complex-valued parameters, Wirtinger relaxation rules of complex differentiation are sought. Such an approach gives the relation (5.5) connecting the spectral density functions of the latent and the idiosyncratic processes to the sample measured spectral density function. Sadly, it evades delivering a unique solution. Therefore and subsequently, a much restricted class of maximum likelihood solutions is pursued where the commonalities will be maximized as well. Towards the end of that pursuit, the low-rank approximation technique of Section 3.4.2 is used to arrive at a suitable solution.

In Section 5.2, the objective is to iteratively solve the commonality maximization problem defined as part of the spectral factor model in Section 4.2. The optimal parameters of the spectral factor model are made available in (5.33) and (5.34). Just as with the analytical method in Section 5.1, first the maximum likelihood parameters of the spectral factor model are obtained; here it is done iteratively using the EM algorithm. In doing so, the line of the estimation approach in Section 3.6 for Section 5.2 is towed by which the definition of the 'E' and 'M' steps are laid out. For this purpose, the formulae for the *a posteriori* expectation and the maximum likelihood parameters are carried out just as they were derived in Sections 3.7.1 and 3.7.2. However, the analysis is tedious because of the non-analytic nature of the real-valued log-likelihood function of complex-valued parameters. As in Section 5.1, this difficulty is overcame by employing Wirtinger relaxations. The equations (5.25) and (5.29) give the maxi-

mum likelihood parameters of the spectral factor model at each iteration of the EM algorithm. Once the EM algorithm has converged, the parameters that maximize the commonalities are found in Section 5.2.3 using the idea of an efficient unbiased estimator reviewed in Section 3.2.

**Note 5.1.** *For the analysis in this chapter, the focus is on any one and only one target frequency in the set of target frequencies obtained on application of Theorem 2.5. Hence, for brevity of notations in this chapter, the index specifying different subbands will be dropped. Therefore, the sans-serif script without any subscripts as in* $\mathsf{y}$ *will be used to refer to the discrete Fourier transform vector random variable at the target frequency of interest. As a result, the spectral density function at the target frequency is simply* $\mathsf{S}^{\mathsf{y}}$ *and the transformation matrix is* $\mathbf{W}$.

## Maximum likelihood estimation of linear processes in time-domain

The attempt in this thesis is to use the principle of maximum likelihood to estimate parameters of the dynamic factor model in the frequency domain. Despite the challenges posed by complex-valued parameters of the model, such an approach was motivated by the established route of maximum likelihood in factor analysis.
An alternative route that should easily be motivated by the maximum likelihood principle is the estimation of the dynamic factor model in the time domain [78]; however, it is not pursued in this thesis. It involves expressing the large sample approximation of the likelihood function in terms of finite-order vector autoregressive moving average process parameters; the maximum likelihood parameters are known to be consistent and asymptotically Gaussian. The derivative of the likelihood function with respect to the parameters are typically non-linear. Hence, iterative algorithms such as Newton - Raphson scoring algorithm [8] or state-space Expectation - Maximization algorithm [83] are used to maximize the log-likelihood. These iterative procedures in the time-domain for vector autoregressive moving average processes are complicated owing to a large set of parameters requiring reliable initial values as well as convergence issues requiring robust estimates of model orders; refer Chapter 12 of [78]. The efficacy of adopting these methods to dynamic factor model estimation in the time-domain is yet to be seen.
On the other hand, the frequency-domain method as presented in this thesis exploits proven methodologies to solve the estimation problem. The analytical approach of Section 5.1 offers an intuitive computationally stable closed-form solution; it uses low-rank approximation theorem and Weyl's theorem to arrive at maximum commonalities parameters. The iterative approach of Section 5.2 uses the EM algorithm for complex-Gaussian estimation and Gauss-Markov theorem. Beyond the known-issue of local minima, it does not suffer from over-parameterization and, as presented in this thesis, is computational stable for Gaussian factor model estimation [14].

The following situates the developments in this chapter with respect to the state-of-the-art:

- ▷ The analytical solution for spectral factor model is derived in (5.10) using low-rank approximation theorem.
  The solution, which involves the principal components of the sample spectral

density function, coincides with that of the the projection theorem solution of [36]; they have a motivation and approach to dynamic factor model not dependent on commonalities.

- ▷ An iterative solution for spectral factor model is derived in Section 5.2 using the Expectation - Maximization algorithm. The converged maximum likelihood parameters in Section 5.2.3 that maximally inherit the commonalities are extracted by applying the Gauss - Markov theorem.
  Iterative solutions recommended by [104] and [100] were based on Fletcher-Powell-Davidon numerical methods.

- ▷ Mild cross-correlation property of the difference between the maximally inherited commonalities and the measured variables in Property 5.1 is obtained via Weyl's theorem.
  In [37], a similar result is obtained via "monotone convergence theorem".

- ▷ Wirtinger relaxations are used for maximizing log-likelihood.
  Relations (5.4) - (5.6) in Section 5.1 states a well-known fact that the sample spectral density maximizes the log-likelihood; e.g., [104] calls it the "unobservable index model". They are retold here using Wirtinger relaxations to emphasize nonexistence, in the Cauchy-Riemann sense, of a non-trivial derivative of the real-valued log-likelihood function of complex-valued variables. The relaxations are introduced in the very familiar setting of Section 5.1 in anticipation of its use in Section 5.2. An alternative of using the isomorphic relations of a complex-Gaussian with that of a real-Gaussian as discussed in Section 2.5 could prove tedious for the purposes in Section 5.2.

## 5.1 Analytical estimation of maximum likelihood commonalities

**Note 5.2.** *Since the selected target frequency represents a subband of frequencies near it, the realization of* $\mathsf{y}$ *corresponding to the* $l$*-th frequency sample within the subband near the target frequency is referred to by* $\mathbf{y}(\omega_l)$.

In Theorem 2.5, an asymptotic property of the discrete Fourier transform was reviewed. It involved treating the discrete Fourier transform at a target frequency $\omega$ as a complex vector random variable y whose realizations are asymptotically the discrete Fourier transform samples $\mathbf{y}(\omega_l) \in \mathbb{C}^r$ at appropriately spaced frequencies $\omega_l$ near $\omega$. It was observed there that these samples may be thought to have been generated from a complex Gaussian probability density

$$\mathsf{p}^{\mathsf{y}}(\mathbf{y}(\omega_l)) = \pi^{-r}(\det(\mathsf{S}^{\mathsf{y}}))^{-1}\exp(-\mathbf{y}'(\omega_l)(\mathsf{S}^{\mathsf{y}})^{-1}\mathbf{y}(\omega_l)), \tag{5.1}$$

where $\mathsf{S}^{\mathsf{y}} \in \mathbb{C}^{r\times r}$ is the spectral density function at frequency $\omega$. For $n$ such discrete Fourier transform samples $\mathbf{y}(\omega_l)$, $l = 1,\ldots,n$, their log-likelihood function may be written as $-rn\log(\pi) - n\log(\det(\mathsf{S}^{\mathsf{y}})) - \sum_{l=1}^{n}\mathbf{y}'(\omega_l)(\mathsf{S}^{\mathsf{y}})^{-1}\mathbf{y}(\omega_l)$. The terms which are independent of $\mathsf{S}^{\mathsf{y}}$ may be discarded and the effective log-likelihood is written as

$$\mathsf{L}(\mathsf{S}^{\mathsf{y}}) = -\log(\det(\mathsf{S}^{\mathsf{y}})) - \mathrm{tr}((\mathsf{S}^{\mathsf{y}})^{-1}\check{\mathsf{S}}^{\mathsf{y}}), \tag{5.2}$$

where $\check{\mathsf{S}}^{\mathsf{y}} \in \mathbb{C}^{r\times r}$ is the sample spectral density function as per (2.30). Note that the inner product of two vectors is converted to the trace of their outer product.

The log-likelihood function $\mathsf{L}(\mathsf{S}^{\mathsf{y}})$ is a real-valued function of complex valued variables in $\mathsf{S}^{\mathsf{y}}$. Hence, it is a non-analytical function and its stationary points have to be found from its vanishing differential $\mathsf{dL}(\mathsf{S}^{\mathsf{y}})$. Presenting the details of deriving the differentials of common real-valued functions of complex-valued matrices is skipped. A comprehensive treatment starting from the basic idea mentioned in Appendix A.1 to a full-fledged multivariate complex calculus is beyond the scope of this thesis. Instead, among many good references, the reader is referred to [57]. Referring to Tables II and V of [57], it is easy to verify that $\mathsf{d}\log(\det(\mathsf{S}^{\mathsf{y}})) = \mathsf{tr}((\mathsf{S}^{\mathsf{y}})^{-1}\mathsf{dS}^{\mathsf{y}})$ and $\mathsf{d}\,\mathsf{tr}((\mathsf{S}^{\mathsf{y}})^{-1}\check{\mathsf{S}}^{\mathsf{y}}) = -\mathsf{tr}((\mathsf{S}^{\mathsf{y}})^{-1}\check{\mathsf{S}}^{\mathsf{y}}(\mathsf{S}^{\mathsf{y}})^{-1}\mathsf{dS}^{\mathsf{y}})$, and their sum may be written as

$$\mathsf{dL}(\mathsf{S}^{\mathsf{y}}) = -\mathsf{tr}(((\mathsf{S}^{\mathsf{y}})^{-1} - (\mathsf{S}^{\mathsf{y}})^{-1}\check{\mathsf{S}}^{\mathsf{y}}(\mathsf{S}^{\mathsf{y}})^{-1})\mathsf{dS}^{\mathsf{y}}).$$

Based on this differential and from the trace form of the differentials in Table III of [57],

$$\frac{\partial}{\partial \mathsf{S}^{\mathsf{y}}}\mathsf{L}(\mathsf{S}^{\mathsf{y}}) = -(\mathsf{S}^{\mathsf{y}})^{-1} + (\mathsf{S}^{\mathsf{y}})^{-1}\check{\mathsf{S}}^{\mathsf{y}}(\mathsf{S}^{\mathsf{y}})^{-1}. \tag{5.3}$$

As mentioned in Appendix A.1, the stationary points of $\mathsf{L}(\mathsf{S}^{\mathsf{y}})$ occur wherever $\mathsf{dL}(\mathsf{S}^{\mathsf{y}})$ vanishes. Since $(\mathsf{S}^{\mathsf{y}})^{-1} = 0$ is prohibited for the existence of $\mathsf{S}^{\mathsf{y}}$, the maximum likelihood solution is

$$\widehat{\mathsf{S}}^{\mathsf{y}} = \check{\mathsf{S}}^{\mathsf{y}} \tag{5.4}$$

wherever $\frac{\partial}{\partial \mathsf{S}^{\mathsf{y}}}\mathsf{L}(\mathsf{S}^{\mathsf{y}}) = 0$. Now substitute (4.12) in the maximum likelihood solution for $\mathsf{S}^{\mathsf{y}}$ in (5.4); it follows that

$$\mathsf{S}^{\mathsf{v}} + \mathsf{S}^{\mathsf{z}} = \check{\mathsf{S}}^{\mathsf{y}}, \tag{5.5}$$

where the check denotes the sample estimate of the spectral density function. Based on (4.11) the maximum likelihood estimates may be further rewritten as

$$\mathbf{W}\mathbf{W}^{*} + \mathsf{S}^{\mathsf{z}} = \check{\mathsf{S}}^{\mathsf{y}}. \tag{5.6}$$

Since maximum likelihood solution for the parameters $\mathbf{W}$ and $\mathsf{S}^{\mathsf{z}}$ have to be gleaned from just one relation in (5.6), there will not be any unique solution.

In order to find the parameters that maximize the commonalities of y amongst the maximum likelihood parameters $\mathbf{W}$ and $\mathsf{S}^{\mathsf{z}}$ of (5.6), further restrictions on the quality of the solutions will have to be imposed. Recall that Definition 4.1 of the commonalities led the formulation of (4.19) which meant v will inherit the covariation in y maximally according to relation (4.18).

However, note that the trivial solution that the diagonal matrix $\mathsf{S}^{\mathsf{z}}(\omega) = 0\ \forall\omega \in [-\pi,\pi]$ and $\check{\mathsf{S}}^{\mathsf{y}}(\omega) = \mathsf{S}^{\mathsf{v}}(\omega)$ is forbidden because $\mathsf{rank}(\check{\mathsf{S}}^{\mathsf{y}}) = r \neq \mathsf{rank}(\mathsf{S}^{\mathsf{v}}) = q$.

### Parameters due to commonalities

Note that the function to be minimized in (4.19) is nonnegative for every $\omega$ in the integral in (4.18). Hence, $\|\check{\mathsf{S}}^{\mathsf{y}}(\omega) - \mathsf{S}^{\mathsf{v}}(\omega)\|_F^2$ may be minimized for each $\omega$ individually

and specifying the variable $\omega$ may be dropped for brevity. Therefore, the maximum commonalities maximum likelihood solution should solve

$$\begin{aligned} \widetilde{\mathbf{W}} &= \underset{\mathbf{W}}{\operatorname{argmin}} \ \|\check{\mathsf{S}}^{\mathsf{y}} - \mathsf{S}^{\mathsf{v}}\|_F^2, \\ \operatorname{rank}(\mathsf{S}^{\mathsf{v}}) &= q < \operatorname{rank}(\check{\mathsf{S}}^{\mathsf{y}}) = r. \end{aligned} \tag{5.7}$$

Recall that according to Theorem 3.1, for the eigenvalue decomposition

$$\check{\mathsf{S}}^{\mathsf{y}} = U \operatorname{diag}(\lambda_1, \cdots, \lambda_r) U^*, \tag{5.8}$$

where $U = [u_1 \cdots u_r]$ is unitary and $\lambda_1 \geq \lambda_r > 0$ are the eigenvalues of $\check{\mathsf{S}}^{\mathsf{y}}$, the best $q$ rank approximation in the Frobenius norm sense is

$$\widetilde{\mathsf{S}}^{\mathsf{v}} = [u_1 \cdots u_q] \operatorname{diag}(\lambda_1, \cdots, \lambda_q) [u_1 \cdots u_q]^*. \tag{5.9}$$

Then it is straightforward to observe that for $\mathsf{S}^{\mathsf{v}} = \mathbf{W}\mathbf{W}^*$ in (4.11), a possible decomposition for the optimal $\mathbf{W}$ is

$$\widetilde{\mathbf{W}} = [u_1 \cdots u_q] \operatorname{diag}(\sqrt{\lambda_1}, \cdots, \sqrt{\lambda_q}). \tag{5.10}$$

Comparing the result (5.10) with that of classical principal components analysis [88], it can be seen that columns of $\widetilde{\mathbf{W}}$ are seeking directions in which the sample measured variances are maximally retained.

## Properties of non-commonalities

Note at this juncture from (5.5) that

$$\operatorname{tr}(\widetilde{\mathsf{S}}^{\mathsf{v}}) = \sum_{i=1}^{q} \lambda_i, \tag{5.11}$$

which enables

$$\operatorname{tr}(\widetilde{\mathsf{S}}^{\mathsf{z}}) = \operatorname{tr}(\check{\mathsf{S}}^{\mathsf{y}}) - \operatorname{tr}(\widetilde{\mathsf{S}}^{\mathsf{v}}) = \sum_{i=q+1}^{r} \lambda_i. \tag{5.12}$$

For the following lemma, refer Chapter 1 §4.4 of [116]:

**Theorem 5.1.** *(Weyl's theorem) For $A, B, C \in \mathbb{C}^{r\times r}$ whose eigenvalues are $a_1 \geq \cdots \geq a_r$, $b_1 \geq \cdots \geq b_r$, and $c_1 \geq \cdots \geq c_r$, respectively, if $A = B + C$, then $b_i + c_1 \geq a_i \geq b_i + c_r$.*

Let $\check{\mathsf{S}}^{\mathsf{y}}$, $\widetilde{\mathsf{S}}^{\mathsf{v}}$, $\widetilde{\mathsf{S}}^{\mathsf{z}}$ correspond to $A, B, C$, respectively, in Theorem 5.1. Recall that the least $r - q$ eigenvalues of $\widetilde{\mathsf{S}}^{\mathsf{v}}$ are equal to zero. Then, it follows that any $\widetilde{\mathsf{S}}^{\mathsf{z}}$ satisfying (5.12) may be chosen such that

$$\operatorname{tr}(\widetilde{\mathsf{S}}^{\mathsf{v}}) > c_1 \geq (r-q)^{-1} \operatorname{tr}(\widetilde{\mathsf{S}}^{\mathsf{z}}) \geq c_r > 0. \tag{5.13}$$

This inequality establishes the following property for z while preserving the Frobenius norm criterion for the inheritance of commonalities defined through 5.7.

**Property 5.1.** *The variables forming the non-commonalities* $\{\mathsf{z}_t\}$ *of the dynamic factor model in (4.2) may be 'mildly cross-correlated' as per (5.12) and (5.13).*

Property 5.1 suggests that z need not be strictly idiosyncratic.

Suppose the discrete Fourier transform components $\mathcal{D} = \{\mathbf{y}(\omega_l)\}$, $l = 1, \ldots n$ within a subband as per Algorithm 1 are obtained. The solution proposed in Algorithm 3 provides the analytical solution of the spectral factor model within a subband.

**Algorithm 3**: Analytical solution for the spectral factor model in a subband

**Input**: $\mathcal{D} = \{\mathbf{y}(\omega_l)\}, l = 1, \ldots n$
**Output**: $\widetilde{\mathbf{W}}$
estimate $\widehat{\mathsf{S}}^{\mathsf{y}}$ using (5.4) and (2.30);
compute pairs $(\lambda_k, u_k)$ $\forall k = 1, \ldots, r$ using (5.8);
estimate $\widetilde{\mathbf{W}}$ as in (5.10);

## 5.2 Iterative estimation of maximum likelihood commonalities

In Section 3.6, an iterative solution for the parameters of the classical factor model of (3.11) was developed. Based on the EM algorithm presented therein, in this section, an iterative procedure for the estimation of the maximum likelihood parameters which is also enforced to maximally inherit the measured commonalities will be developed. Such a motivation to do so is due to the similarity of the relations of the classical factor model in (3.11) and the spectral factor model (4.15). This similarity is obvious if it is supposed that the parameters of the factor model are $\theta \triangleq \{\mathbf{W}, \mathsf{S}^{\mathsf{z}}\}$ and that the random vectors y and z at a target frequency realize $\mathbf{y}$ and $\mathbf{z}$ at nearby frequencies according to Theorem 2.5.

**Note 5.3.** *As in the previous section, the realization of* y *corresponding to the* $l$*-th frequency sample within the subband near the target frequency is denoted by* $\mathbf{y}(\omega_l)$*. In addition, in this section,* $\mathsf{S}_i^{\mathsf{y}}$ *and* $\mathbf{W}_i$ *are used to refer to the* $i$*-th iterative estimate of the spectral density function* $\mathsf{S}^{\mathsf{y}}$ *and the transformation matrix* $\mathbf{W}$ *at the target frequency, respectively.*

As in Section 3.6, first notice that the spectral factor model equivalent of (3.39) is

$$\mathsf{p}^{\mathsf{y}|\mathsf{x},\theta}(\mathbf{y} \mid \mathsf{x}, \theta) = \mathcal{N}_{\mathbb{C}}(\mathbf{y} \mid \mathbf{W}\mathsf{x}, \mathsf{S}^{\mathsf{z}}). \tag{5.14}$$

Let a dataset $\mathcal{D}$ render the discrete Fourier transform samples $\mathbf{y}(\omega_l)$, $l = 1, \ldots, n$ at frequencies within the subband represented by the random vector y. At the target

frequency under consideration, the likelihood of $\mathcal{D}$ to correspond to the spectral factor model is

$$\mathsf{p}^{\mathsf{y}|\mathsf{x},\theta}(\mathcal{D} \mid \mathsf{x}, \theta) = \prod_{l=1}^{n} \mathsf{p}^{\mathsf{y}|\mathsf{x},\theta}(\mathbf{y}(\omega_l) \mid \mathsf{x}, \theta). \tag{5.15}$$

Now, in line with (3.41), (3.43), and (3.42), the direct extension of the estimation of the spectral factor model parameters in the $i$-th iteration of the EM algorithm gives:

$$\hat{\theta}_{i+1} = \underset{\theta_i}{\mathsf{argmax}}\ \mathsf{E}^{\mathsf{x}|\mathsf{y},\theta}\big[\mathsf{log}_e \mathsf{p}^{\mathsf{y}|\mathsf{x},\theta}(\mathcal{D} \mid \mathsf{x}, \theta_i)\big] \tag{5.16}$$

and for

$$\begin{aligned} \theta_i &\triangleq \{\mathbf{W}_i, \mathsf{S}_i^{\mathsf{z}}\}, \\ \hat{\theta}_{i+1} &= \underset{\theta_i}{\mathsf{argmax}}\ \mathsf{E}^{\mathsf{x}|\mathsf{y},\theta}[\mathsf{f}(\theta_i, \mathsf{x})], \\ \mathsf{f}(\theta_i, \mathsf{x}) &\triangleq \sum_{l=1}^{n} \mathsf{log}_e \mathcal{N}_{\mathbb{C}}(\mathbf{y}(\omega_l) \mid \mathbf{W}_i \mathsf{x}, \mathsf{S}_i^{\mathsf{z}}). \end{aligned}$$

Expanding $\mathsf{f}(\theta_i, \mathsf{x})$ will lead to terms in $\mathsf{x}$ and $\mathsf{x}\mathsf{x}^*$. So, as with (3.46), first define

$$\begin{aligned} \langle \mathbf{x} \rangle_{i,l} &\triangleq \mathsf{E}^{\mathsf{x}|\mathsf{y},\theta}[\mathsf{x} \mid \mathbf{y}(\omega_l), \theta_i], \\ \langle \mathbf{x}\mathbf{x}^* \rangle_{i,l} &\triangleq \mathsf{E}^{\mathsf{x}|\mathsf{y},\theta}[\mathsf{x}\mathsf{x}^* \mid \mathbf{y}(\omega_l), \theta_i]. \end{aligned} \tag{5.17}$$

Note that $\langle \mathbf{x} \rangle_{i,l} \in \mathbb{C}^q$ and $\langle \mathbf{x}\mathbf{x}^* \rangle_{i,l} \in \mathbb{C}^{q \times q}$; whereas their estimation will define the E-step of the EM-algorithm. Then, as in (3.51), it may be written that

$$\mathbf{W}_{i+1} = \underset{\mathbf{W}_i}{\arg} \left( \frac{\partial}{\partial \mathbf{W}_i} \mathsf{E}^{\mathsf{x}|\mathsf{y},\theta}[\mathsf{f}(\theta_i, \mathsf{x})] = 0 \right). \tag{5.18}$$

Similarly, as in (3.53),

$$\mathsf{S}_{i+1}^{\mathsf{z}} = \underset{\mathsf{S}_i^{\mathsf{z}}}{\arg} \left( \frac{\partial}{\partial \mathsf{S}_i^{\mathsf{z}}} \mathsf{E}^{\mathsf{x}|\mathsf{y},\theta}[\mathsf{f}(\theta_i, \mathsf{x})] = 0 \right). \tag{5.19}$$

These optimizations complete the M-step of the EM-algorithm.

Hence, starting from initial guesses, the $i$-th iteration alternates between:

1. **Expectation-step** Evaluate $\langle \mathbf{x} \rangle_{i,l}$ and $\langle \mathbf{x}\mathbf{x}^* \rangle_{i,l}$ using (5.17), and
2. **Maximization-step** Update $\mathbf{W}_{i+1}$ using (5.18) and $\mathsf{S}_{i+1}^{\mathsf{z}}$ using (5.19).

It is clear that the EM algorithm leads to non-unique maximum likelihood solutions depending on the starting conditions.

### 5.2.1 EM steps and form of the maximum likelihood parameters

As in the previous sections, note that $\omega_l$ corresponds to the $l$-th frequency sample within the subband near the target frequency. Also, $\mathsf{S}_i^{\mathsf{y}}$ and $\mathbf{W}_i$ refer to the $i$-th iterative estimate of the spectral density function $\mathsf{S}^{\mathsf{y}}$ and the transformation matrix $\mathbf{W}$ at the target frequency, respectively.

In this section, the solutions encountered in the two steps of the algorithm will be analyzed and the usability of an iterative solution in lieu of or complementing an analytical solution assessed. In doing so, the derivations due to Section 3.6 will be of sufficient aid and will be the main reference.

First, note from Appendix B.3 and relations (3.48) and (3.49) that the E-step of the EM algorithm is simply

$$\langle \mathbf{x} \rangle_{i,l} = \mathbf{\Omega}_i \mathbf{W}_i^* (\mathsf{S}_i^{\mathsf{z}})^{-1} \mathbf{y}(\omega_l) \tag{5.20}$$

$$\mathbf{\Omega}_i = (I_q + \mathbf{W}_i^* (\mathsf{S}_i^{\mathsf{z}})^{-1} \mathbf{W}_i)^{-1}, \tag{5.21}$$

$$\langle \mathbf{x}\mathbf{x}^* \rangle_{i,l} = \langle \mathbf{x} \rangle_{i,l} \langle \mathbf{x} \rangle_{i,l}^* + \mathbf{\Omega}_i, \tag{5.22}$$

where the inverse of $\mathbf{\Omega}_i \in \mathbb{C}^{q \times q}$, in general, exists. For evaluating $\mathbf{W}_{i+1}$ according to (5.18), first write

$$\begin{aligned} \mathsf{E}^{\mathsf{x}|\mathsf{y},\theta}[\mathsf{f}(\theta_i, \mathsf{x})] &= \sum_{l=1}^{n} \mathsf{E}^{\mathsf{x}|\mathsf{y},\theta}[\log_e(\mathcal{N}_{\mathbb{C}}(\mathbf{y}(\omega_l) \mid \mathbf{W}_i \mathsf{x}, \mathsf{S}_i^{\mathsf{z}}))] \\ &= -n \log_e(|\mathsf{S}_i^{\mathsf{z}}|) \\ &\quad - \sum_{l=1}^{n} \mathsf{tr}((\mathsf{S}_i^{\mathsf{z}})^{-1} \mathbf{W}_i \langle \mathbf{x}\mathbf{x}^* \rangle_{i,l} \mathbf{W}_i^*) + \mathbf{y}^*(\omega_l)(\mathsf{S}_i^{\mathsf{z}})^{-1} \mathbf{y}(\omega_l) \\ &\quad\quad - 2\Re(\mathbf{y}^*(\omega_l)(\mathsf{S}_i^{\mathsf{z}})^{-1} \mathbf{W}_i \langle \mathbf{x} \rangle_{i,l}), \end{aligned} \tag{5.23}$$

where eliminated are terms independent of $\mathbf{W}_i$ or $\mathsf{S}_i^{\mathsf{z}}$. The reader is referred to [57] to verify using Wirtinger relaxations that

$$\frac{\partial}{\partial \mathbf{W}_i} \mathsf{E}^{\mathsf{x}|\mathsf{y},\theta}[\mathsf{f}(\theta_i, \mathsf{x})] = (\mathsf{S}_i^{\mathsf{z}})^{-1} \sum_{l=1}^{n} (\overline{\mathbf{W}}_i \langle \mathbf{x}\mathbf{x}^* \rangle'_{i,l} - \overline{\mathbf{y}}(\omega_l) \langle \mathbf{x} \rangle'_{i,l}). \tag{5.24}$$

Then, due to (5.18),

$$\mathbf{W}_{i+1} = \left( \sum_{l=1}^{n} \mathbf{y}(\omega_l) (\langle \mathbf{x} \rangle_{i,l})^* \right) \left( \sum_{l=1}^{n} \langle \mathbf{x}\mathbf{x}^* \rangle_{i,l} \right)^{-1}. \tag{5.25}$$

Just as in Section 3.7.2, let

$$\mathbf{v}_i(\omega_l) \triangleq \mathbf{W}_{i+1} \langle \mathbf{x} \rangle_{i,l}. \tag{5.26}$$

For $\mathsf{S}_i^{\mathsf{z}} = \mathsf{diag}(\mathsf{s}_i^{\mathsf{z}_1}, \ldots, \mathsf{s}_i^{\mathsf{z}_r})$ it can easily be seen that

$$\mathsf{E}^{\mathsf{x}|\mathsf{y},\theta}[\mathsf{f}(\theta_i, \mathsf{x})] = -n \sum_{k=1}^{r} [\log_e(\mathsf{s}_i^{\mathsf{z}_k}) + \frac{1}{\mathsf{s}_i^{\mathsf{z}_k}} b_i^{\mathsf{z}_k}] \tag{5.27}$$

where

$$b_i^{\mathsf{z}_k} = \frac{1}{n}\sum_{l=1}^{n} |\mathbf{y}_k(\omega_l) - \mathbf{v}_{ki}(\omega_l)|^2. \tag{5.28}$$

Note that $\mathsf{S}_i^{\mathsf{z}}$ is a real-valued diagonal matrix and the derivative with respect to it is straightforward. Then $\partial \mathsf{E}^{\mathsf{x}|\mathsf{y},\theta}[\mathsf{f}(\theta_i, \mathsf{x})]/\partial \mathsf{s}_i^{\mathsf{z}_k} = 0$ at

$$\begin{aligned} \mathsf{s}_{i+1}^{\mathsf{z}_k} &= b_i^{\mathsf{z}_k}, \\ \mathsf{S}_{i+1}^{\mathsf{z}} &= \mathsf{diag}(\mathsf{s}_{i+1}^{\mathsf{z}_1}, \cdots, \mathsf{s}_{i+1}^{\mathsf{z}_r}). \end{aligned} \tag{5.29}$$

The relations (5.25) and (5.29) stand for the M-step of the EM algorithm for the maximum likelihood parameters of the spectral factor model.

### 5.2.2 EM algorithm for spectral factor model

The following pseudocode of the EM algorithm for the maximum likelihood spectral factor model may now be provided; this is in line with Algorithm 2 in Section 3.7.2. In Algorithm 2, the input was the dataset $\mathcal{D}$ of iid data samples; whereas here it is assumed that $\mathcal{D}$ is a set of discrete Fourier transform components near a target frequency as recommended by the asymptotic requirements of Theorem 2.5.

**Algorithm 4**: EM algorithm for the spectral factor model in a subband

**Input**: $\mathcal{D} = \{\mathbf{y}(\omega_l)\}, l = 1, \ldots n$
**Output**: $\widehat{\mathbf{W}}, \widehat{\mathsf{S}}^{\mathsf{z}} = \mathsf{diag}(\hat{\mathsf{s}}^{\mathsf{z}_1}, \ldots, \hat{\mathsf{s}}^{\mathsf{z}_r})$
initialize $i = 0$;
randomize $\mathbf{W}_i, \mathsf{S}_i^{\mathsf{z}}$;
**do**
  **E-step**:
  **for** $l = 1 \,\mathrm{to}\, n$ **do**
    compute
      $\langle \mathbf{x} \rangle_{i,l}$ using (5.20);
      $\langle \mathbf{x}\mathbf{x}^* \rangle_{i,l}$ using (5.22);
  **end**
  **M-step**: update
    $\mathbf{W}_{i+1}$ using (5.25);
    $\mathsf{s}_{i+1}^{\mathsf{z}_k}$ $\forall k = 1, \ldots, r$ using (5.29);
  $i \longleftarrow i + 1$;
  $\epsilon \longleftarrow \mathsf{E}^{\mathsf{x}|\mathsf{y},\theta}[\mathsf{f}(\theta_i, \mathsf{x})] - \mathsf{E}^{\mathsf{x}|\mathsf{y},\theta}[\mathsf{f}(\theta_{i-1}, \mathsf{x})]$ using (5.23);
**while** $\epsilon > 10^{-8}$ **and** $i < 20$;
$\widehat{\mathbf{W}} \longleftarrow \mathbf{W}_i$, $\hat{\mathsf{s}}^{\mathsf{z}_k} \longleftarrow \mathsf{s}_i^{\mathsf{z}_k}$ $\forall k = 1, \ldots, r$;

Suppose the discrete Fourier transform components $\mathcal{D} = \{\mathbf{y}(\omega_l)\}$, $l = 1, \ldots n$ within a subband are obtained as per Algorithm 1. Algorithm 4 demonstrates how the E and M steps may be alternated, starting with a random initialization of the parameters corresponding to a target frequency, to output the converged parameters $\widehat{\mathbf{W}}$ and $\widehat{\mathsf{S}}^{\mathsf{z}}$ of the spectral factor model.

**Note 5.4.** *For EM algorithm in Algorithm 4 converging towards a local maximum of the log-likelihood is possible, converged parameters $\theta$ corresponding to the largest $\mathsf{E}^{\mathsf{x}|\mathsf{y},\theta}[\mathsf{f}(\theta,\mathsf{x})]$ from a number of random restarts will be chosen.*

### 5.2.3 Maximizing commonalities in spectral factor model

**Note 5.5.** *In this section, it is assumed that the EM steps have converged. Therefore, for notational brevity, any indexing of the iteration is dropped and the updated parameters will be denoted by $\theta \triangleq \{\widehat{\mathbf{W}}, \widehat{\mathsf{S}}^{\mathsf{z}}\}$. As in the previous sections, $\omega_l$ corresponds to the $l$-th frequency sample within the subband near the target frequency.*

As seen, at the end of the iterations of a converged EM algorithm, access is available to the estimate of the transformed factor to get

$$\mathbf{v}(\omega_l) = \widehat{\mathbf{W}}\mathbf{x}(\omega_l)$$

corresponding to the $l$-th realization $\mathbf{y}(\omega_l)$ $\forall l \in 1, \ldots, n$, where

$$\mathbf{x}(\omega_l) = \mathsf{E}^{\mathsf{x}|\mathsf{y},\theta}[\mathsf{x} \mid \mathbf{y}(\omega_l), \theta] \tag{5.30}$$

as in (5.17) and computed in (5.20). Thus, in the context of (4.15),

$$\mathbf{y}(\omega_l) = \widehat{\mathbf{W}}\mathbf{x}(\omega_l) + \mathbf{z}(\omega_l) \quad \forall l = 1, \ldots, n \tag{5.31}$$

where $\mathbf{z}(\omega_l)$ is the error in regressing $\mathbf{x}(\omega_l)$ towards $\mathbf{y}(\omega_l)$. The regression errors are due to zero mean isotropic Gaussian vector random variable $\mathsf{z}$; this is not the assumption but the result of Theorem 2.5.

Now it shall be seen why the same situation as in the linear model of Section 3.2 persists. From the form of (4.18) for inheritance by $\mathsf{S}^{\mathsf{v}}$ of the commonalities of $\mathsf{S}^{\mathsf{y}}$, it is clear that $\|\mathsf{S}^{\mathsf{y}}(\omega) - \mathsf{S}^{\mathsf{v}}(\omega)\|_F^2$ may be minimized for each $\omega$ individually. Then, (4.19) implies that the optimal $\mathsf{S}^{\mathsf{z}}$ is given by $\widetilde{\mathsf{S}}^{\mathsf{z}} \triangleq \underset{\mathsf{S}^{\mathsf{z}}}{\operatorname{argmin}}\, g = \underset{\mathsf{S}^{\mathsf{z}}}{\operatorname{argmin}}\, \mathsf{tr}(\mathsf{S}^{\mathsf{z}})$, or for each of the diagonal elements $\mathsf{s}^{\mathsf{z}_k}$ of $\mathsf{S}^{\mathsf{z}}$

$$\tilde{\mathsf{s}}^{\mathsf{z}_k} \triangleq \mathsf{min}(\mathsf{s}^{\mathsf{z}_k}) \quad \forall k = 1, \ldots, r. \tag{5.32}$$

But $\mathsf{s}^{\mathsf{z}_k}$ is the variance of the zero mean Gaussian error in approximating $\mathbf{y}_k(\omega_l)$ using $\mathbf{v}_k(\omega_l)$. Hence, a minimum variance unbiased regression of $\mathbf{x}(\omega_l)$ towards $\mathbf{y}(\omega_l)$ is sought using $\mathbf{v}(\omega_l) = \mathbf{W}\mathbf{x}(\omega_l)$.

As seen, maximizing the commonalities upon convergence of the EM algorithm requires an efficient estimator of $\mathbf{W}$. Therefore, if the Gauss-Markov solution of (3.8) is used, the efficient estimator got is

$$\widetilde{\mathbf{W}} = [\mathbf{y}(\omega_1) \cdots \mathbf{y}(\omega_n)]\mathbf{X}^* \left(\mathbf{X}\mathbf{X}^*\right)^{-1}, \tag{5.33}$$

where, using $\mathbf{x}(\omega_l)$ referred to in (5.30 ) and computed via (5.20), the $q \times n$ matrix $\mathbf{X} = [\mathbf{x}(\omega_1) \cdots \mathbf{x}(\omega_n)]$ having $\mathsf{rank}(\mathbf{X}) = q$ is a maximum likelihood 'latent data matrix.' And, as per (3.9), an unbiased estimate of $\mathsf{S}^{\mathsf{z}}$ is

$$
\begin{aligned}
\widetilde{\mathsf{S}}^{\mathsf{z}} &= \mathsf{diag}(\tilde{\mathsf{s}}^{\mathsf{z}_1}, \ldots, \tilde{\mathsf{s}}^{\mathsf{z}_r}), \\
\tilde{\mathsf{s}}^{\mathsf{z}_k} &= \frac{1}{n-q} \sum_{l=1}^{n} |\mathbf{y}_k(\omega_l) - \tilde{\mathbf{v}}_k(\omega_l)|^2, \; k = 1, \ldots, r, \\
\tilde{\mathbf{v}}(\omega_l) &= \widetilde{\mathbf{W}} \mathbf{x}(\omega_l),
\end{aligned}
\tag{5.34}
$$

where $\mathbf{x}(\omega_l)$ referred to in (5.30 ) is computed via (5.20). It is important to understand that although the EM algorithm gives the maximum likelihood solution, the maximization of the commonalities was achieved through (5.33).

Suppose the discrete Fourier transform components $\mathcal{D} = \{\mathbf{y}(\omega_l)\}, l = 1, \ldots n$ within a subband as per Algorithm 1 are obtained. Then, as per Algorithm 5, the procedure for estimating the maximum commonalities spectral factor model parameters utilizing the EM algorithm developed in Algorithm 4 could be compiled.

**Algorithm 5**: Maximum commonalities spectral factor model via EM algorithm

**Input**: $\mathcal{D} = \{\mathbf{y}(\omega_l)\}, l = 1, \ldots n$
**Output**: $\widetilde{\mathbf{W}}, \widetilde{\mathsf{S}}^{\mathsf{z}}$
estimate $\{\widehat{\mathbf{W}}, \widehat{\mathsf{S}}^{z}\}$ with input $\mathcal{D}$ to Algorithm 4;
compute $\mathbf{x}(\omega_l)$ as in (5.30);
estimate $\widetilde{\mathbf{W}}$ using (5.33);
estimate $\widetilde{\mathsf{S}}^{\mathsf{z}}$ using (5.34);

## 5.3 Summary

The form of the spectral factor model in (4.15) is similar to the classical factor model in (3.11). In this chapter, as reviewed for the classical factor model in Chapter 3, two approaches for maximum likelihood estimation of the spectral factor model were developed and within each of them the commonality maximization parameters were found:

In the analytical approach put forth, the sample spectral density function computed from the discrete Fourier transform samples of a measured vector random process near a target frequency is the maximum likelihood spectral density function of the process. The maximum likelihood maximum commonalities solution provided by (5.10) is similar in interpretation to the low-rank approximation of the classical factor model solution and (5.13) provides the leverage to choose idiosyncrasies the way the user wants without destroying the rank stringencies of the transformation matrix. The commonality maximizing maximum likelihood transformation was found to direct the latent spectra along the principal components of the measured spectra. This analytical solution was presented in Algorithm 3.

Again, as with the classical factor model, Algorithm 4 was designed to estimate the maximum likelihood spectral factor model in an iterative fashion. The parameters of the model thus estimated were improved in (5.33) and (5.34) by treating the maximum

likelihood transformation of the *a posteriori* mean of the latent variables of the spectral model as a regression towards the measured spectral components. This enabled the transformed latent spectra to maximally inherit the commonalities of the measured spectra through the Gauss-Markov theorem.

# Chapter 6

# Learning via spectral factor model

In Chapter 1, the objective of learning a time series process was discussed with examples. Two challenges to prove the learning worth of the spectral factor model were proposed there. Firstly, a given measured time series has to be classified as belonging to one of the several possible processes that could have generated it. In this chapter, classification is done based on the proximity of the optimal spectral factor model parameters of the unclassified time series with that of the time series of various classes of possible processes. Secondly, prediction of the future evolution of a current measured time series is done. In this chapter, prediction is performed by enriching classical vector autoregression parameters of the measured time series in the prediction equation with commonalities.

In Section 6.1, before moving to either of those learning applications, it is necessary to consider the computational requirements of the spectral factor model estimation. In particular, strategies to choose the best of the two possible estimation procedures developed in Chapter 5 are considered from a practical perspective of using them in a learning problem.

In Section 6.2, the classification problem is defined concretely. The strategy involves comparing projection of the subspace spanned by the transformation matrix of the test time series episodes onto those of a number of training time series episodes. An approach based on the nearest neighbors in terms of the projection is used to decide whether a test episode belongs to one class or another; this is made available in Algorithm 7.

In Section 6.3, the prediction problem is taken up. The strategy there is simple: The measured *acvf* is an addition of two *acvf*s, one of them inheriting the commonalities and the other not. All occurrences of the measured *acvf* in the classical vector autoregression prediction equations are replaced with the part of the measured *acvf* that inherits the commonalities. This is demonstrated in Algorithm 8.

The following situates the developments in this chapter with respect to the state-of-the-art:

- ▷ Spectral factor model based classification.
  The classification metric of (6.4) compares the maximum commonality transformations of any two multivariate time series. The metric quantifies the overlap of maximum commonality subspaces despite (i) multiplicity of maximum likeli-

hood solution due to orthogonal rotations, (ii) the transformation matrices being complex-valued, and (iii) the transformations for all the subbands are to be compared. The closest work in the literature to this is that of [66] who struggle to achieve a proper metric that will compare two classes of spectral densities despite working with their full-rank sample estimates.

- ▷ Commonalities driven multivariate time series prediction.
  For predicting measured multivariate time series believed to consist of substantial commonalities, an estimate of the *acvf* is obtained by inverting its commonalities enriched spectral density function. Classical vector autoregression on current and past samples with orthogonal errors, as prevalent in literature [102, 51], is used to obtain the predictions. Except, here, the measured *acvf* is replaced by that of the commonalities estimate.
  On the other hand, the focus of eminent works in dynamic factor model literature such as [36] is in the prediction of the commonalities, which is typically unmeasured. E.g., [104] wants to model business cycles whereas [118] predicts diffusion index based on other measurable indicators.

## 6.1 Practicalities of spectral factor model estimation

It is clear that learning problems would require estimation of the spectral factor model parameters that inherit commonalities maximally. Hence, as a prelude to using commonalities for learning problems, Algorithm 6 is followed to estimate these parameters given a finite $\tau$ length time series $\{y_t\}$, $t = 1, 2, \ldots, \tau$. The output of the algorithm is the set of spectral model parameters $\{\mathbf{W}(\omega_j), \mathsf{S}^{\mathsf{z}}(\omega_j)\}$ at $\hat{\jmath}$ target frequencies $\omega_j \in [0, 1)$, $j = 1, \ldots, \hat{\jmath}$.

**Algorithm 6**: Estimate optimal spectral factor model per subband

**Input**: $\{y_t\}, t = 1, \ldots, \tau$; $y_t \in \mathbb{R}^r$;
**Output**: $\{\mathbf{W}(\omega_j)\}$; $j = 1, \ldots, \hat{\jmath}$
compute $\{\mathbf{y}(\omega_{j,l})\}$; $j = 1, \ldots, \hat{\jmath}$; $l = 1, \ldots, n$ using Algorithm 1;
**foreach** $j = 1, \ldots, \hat{\jmath}$ **do**
  gather $\mathcal{D} = \{\mathbf{y}(\omega_{j,l})\}$, $l = 1, \ldots, n$;
  estimate $\{\mathbf{W}(\omega_j), \mathsf{S}^{\mathsf{z}}(\omega_j)\}$ with input $\mathcal{D}$ to Algorithms 3 or 5;
**end**

The following observations regarding Algorithm 6 may be noted:

1. The procedures of Chapter 5 estimated the maximum commonalities spectral factor model within a spectral subband as per the asymptotic theory discussed in Section 2.5. Hence, the discrete Fourier transform components are split into $\hat{\jmath}$ subbands using Algorithm 1.

2. Each subband should have a sufficiently large $n$ number of samples for a reliable estimation of the spectral factor model parameters; this may typically be set to $n \approx r^2$ to ensure consistency of sample estimates without inviting the curse of dimensionality issues [106, 12].

3. Although informative, the parameter $\mathsf{S}^{\mathsf{z}}$ is needed for neither the classification nor the prediction exercises because it contains no commonalities which are all available through $\mathbf{W}$.

4. Depending on the computational demands and application, either Algorithms 3 or 5 may be chosen for computing the optimal parameters of the spectral factor model.

The last of the above requires further discussion. Theoretically, the analytical solution of Algorithm 3 is elegant and unique till orthogonal rotations of the transformation matrix. However, in favor of the iterative Algorithm 5 are the following practical aspects:

- ▷ For an $r$-variate measured time series, computing its spectral density function as well as computing its eigenvalue decomposition are typically $O(r^3)$ operations [32]. This makes Algorithm 3 very prohibitive as the number $r$ of measured variables grows. For $q$-variate latent time series, the intensive operations of the EM algorithm-based estimation in Algorithm 5 are $q \times q$ matrix inverses; they are typically $O(q^3)$ operations and $q \ll r$ is the practical choice. Note that $(\mathsf{S}^{\mathsf{z}})^{-1}$ in the EM Algorithm would involve only scalar reciprocals of its diagonal. Hence, practically, for online or real-time implementations where complexity is always a constraint, spectral factor model updates could be done better using Algorithm 5.
  On the down side, as mentioned in Note 5.4, the issue of local minima in the EM algorithm poses some risk. Hence, it is desirable to confirm iterative estimates with an occasional update via Algorithm 3. Or, the randomization of the parameters in the beginning of the EM algorithm might be replaced by analytical estimates.

- ▷ In many time series, especially in econometrics, seasonality leads to distinct spikes in the spectral components. Their adjusting or correction leads to undesired consequences including elimination of true and introduction of misleading non-seasonal characteristics as well as distortion of commonalities [91]. Suppose the discrete Fourier transform components of the unadjusted seasonal time series corresponding to the suspected seasonalities are assumed missing. EM algorithm could be extended to impute the missing values using approaches such as Monte Carlo EM [123] and Stochastic Approximation EM [31]. This allows the possibility to model and learn the commonalities without inviting unnecessary pre-processing.

## 6.2 Multivariate time series classification

Let an $r$-variate measured time series be denoted by $\{\mathsf{y}_t\}$. The objective of the classification problem is to assign $\{\mathsf{y}_t\}$ to one and only one of the $c$ exhaustive classes of time series $\mathcal{C}_i$, $i = 1, \ldots, c$. It is necessary to clarify what a class of time series means. A **class** of time series is a stochastic process, which is distinct from other processes according to an expert who has measured the time series. Such a distinction might be due to some dynamic characteristics of the time series the class is associated with that

is objectively or subjectively obvious to the expert. Or, the expert might believe that the physical process that generated a class of time series is dissimilar to others.

To ease the discussion on classification of time series, revisit the first of the two examples in Section 1.2. There, the computer gamer has to make joystick movements which require her to position the cursor from the center of the screen to any one of the four corners. During the game, the magnetoencephalography sequences corresponding to ten spatial spots in the brain were recorded via a magnetoencephalography scanner. Existence of a set of two latent signals, viz., her cognition and reaction sequences, of known general characteristics which generate the measured time series is presumed. When a joystick is moved, these latent signals must undergo a dynamic transformation corresponding to that particular class of joystick movements. In this example, an expert might have witnessed several episodes of the gamer making these four movements and understood the dynamic characteristics of the measured time series. Each episode is a finite length multivariate time series realization. Suppose access is available to a historical database of many such episodes which have been classified by the expert; they may be called the **training episodes**. It is wished to classify more episodes without the aid of the expert one by one; each of them will be called a **test episode**.

The challenge to reliably classify a test episode of a multivariate time series process based on the dynamic characteristics of a given dataset of classified training episodes is the time series **classification problem**. Hence, the classification process will have two phases: In the training phase, the summary of the dynamic characteristics of many training episodes are extracted. Obviously, the summary here implies the parameters of the spectral factor model. In the testing phase, the test episode is fed as input to the classification system. Its dynamic characteristics are compared with the dynamic characteristics of all the classes, and the most appropriate class label is given as the output of the system. Effectively, the spectral factor model parameters of the test episode is compared with those of the training episodes.

Such a classification system is indeed a learning system because of two reasons: First, the essential dynamic characteristics from all training episodes have to be appropriately summarized, which in this thesis's context will be in model parameters. Second, the classification system demonstrates the ability to use past experiences of training episodes to respond to a new test episode which it has not witnessed earlier.

### Proposal for a classification system

The motivation so far has been that, firstly, each of the components of a multivariate measured time series contribute towards the commonalities shared amongst them, and, secondly, the dynamic transformation should maximally inherit the commonalities. Then, the following steps are devised in the time series classification strategy:

1. Estimate the optimal dynamic transformation for the test episode and all training episodes whereby the latent dimensionality maximally inherits the commonalities. From this thesis's perspective, this is equivalent to estimating the maximum commonalities spectral factor model as done by Algorithm 6.

2. Create a neighborhood for the test episode by computing its proximity with each of the training episodes with respect to their optimal dynamic transformation parameters.
   For this step, the proximity of the test episode has to be compared to the training

episodes with respect to their maximal inheritance of commonalities, which are believed to distinguish one class from another. The spectral factor model parameters $\mathbf{W}$ are known to correspond to maximal inheritance of commonalities. Hence, the proximity of the optimal $\mathbf{W}$ of the test episode ought to be compared with the optimal $\mathbf{W}$ of the training episodes.

3. Classify test episode to the class in which majority of the training episodes in the former's immediate neighborhood belong to.
   For this step, one should be able to design a distance metric between the transformation matrices of any two time series. Then, with respect to the distance metric, the concepts of 'neighborhood' of the transformation matrices as well as the 'closeness' between them may be used. Specifically, one may decide in favor of the class which has $\kappa$ training episodes closer to the test episode than any other class; this strategy is generally known as the $\kappa$-nearest neighbor classification [28].

The last two steps above beg elaboration. Suppose access is available to time series from $c$ classes $\mathcal{C}_i, i = 1, \ldots, c$ each with $|\mathcal{C}_i|$ examples and an unclassified time series episode. To proceed further, let the following notations be compiled:

| | |
|---|---|
| $\{y_t\}_{l@i},\ l = 1, \ldots, \lvert\mathcal{C}_i\rvert$ | $l$-th example time series in the class $\mathcal{C}_i$ |
| $\{\mathbf{W}(\omega_j)\}_{l@i}$ | spectral transformation of $\{y_t\}_{l@i}$ at $\omega_j$ |
| $\{y_t\}_?$ | unclassified test episode |
| $\{\mathbf{W}(\omega_j)\}_?$ | spectral transformation of $\{y_t\}_?$ at $\omega_j$ |
| $\delta\left(\{\mathbf{W}(\omega_j)\}_{l@i}, \{\mathbf{W}(\omega_j)\}_?\right)$ | similarity between $\{\mathbf{W}(\omega_j)\}_{l@i}$ and $\{\mathbf{W}(\omega_j)\}_?$ |
| $\rho(l@i, ?)$ | proximity between $\{y_t\}_?$ and $\{y_t\}_{l@i}$ |
| $\rho(\downarrow l@i, ?)$ | decreasing sort of $\rho(l@i, ?)$ over $l$ |

Since the spectral factor models at $\hat{\jmath}$ target frequencies are independent of one another, it is proposed that

$$\rho(l@i, ?) = \sum_{j=1}^{\hat{\jmath}} \delta\left(\{\mathbf{W}(\omega_j)\}_{l@i}, \{\mathbf{W}(\omega_j)\}_?\right). \tag{6.1}$$

Then, $\{y_t\}_?$ may be associated with $\mathcal{C}_{\hat{\imath}}$, if

$$\hat{\imath} = \underset{i}{\text{argmax}} \sum_{l=1}^{\kappa} \rho(\downarrow l@i, ?), \tag{6.2}$$

where a tie is broken at random and $\kappa$ is a suitable integer, e.g., $\kappa = 5$.

## Classification metric

Recall from solution (5.10) that columns of $\mathbf{W} \in \mathbb{C}^{r \times q}$ form a set of scaled unitary vectors which define a $q$-dimensional space embedded in $\mathbb{C}^r$. These vectors carve a hyperparallelopiped in $\mathbb{C}^r$ whose sides are norms of these columns [112]. Then, a possible measure of disparity or similarity between any two transformation matrices is to compare the overlap of the volumes.

What the overlap of volumes really implies is specified now. The overlap for $a, b \in \mathbb{C}^r$ is defined as $\delta(a,b) = |a^*\, b|$, which is the absolute 2-norm of the unitary projection of $a$ onto the span of $b$. Consider a set of linearly independent columns vectors of some matrix $A$ spanning a subspace $\mathcal{M}_A \subset \mathbb{C}^r; \mathsf{rank}(\mathcal{M}_A) = q$ which is unitarily projected onto a subspace $\mathcal{M}_B \subset \mathbb{C}^r; \mathsf{rank}(\mathcal{M}_B) = q$ of another matrix $B$. This projection may be thought of as carving a volume measured as the absolute determinant $|\det(A^*B)|$ of the unitary projection of the vectors spanning $\mathcal{M}_A$ onto $\mathcal{M}_B$ [75, 84]. In [85], it is available that

$$|\det(A^*B)| = \mathsf{vol}(A)\mathsf{vol}(B)\cos\{\mathsf{R}(A), \mathsf{R}(B)\} \tag{6.3}$$

where $\mathsf{vol}(A) \triangleq \det(\mathsf{R}(A))$, where $\mathsf{R}(A)$ is the range space of $A$ and $\cos\{\mathsf{R}(A), \mathsf{R}(B)\}$ refers to the product of the principal angles between compatible matrices $A$ and $B$. In [40], it is shown that $\cos\{\mathsf{R}(A), \mathsf{R}(B)\} = \prod_{k=1}^{q} |a_k^*\, b_k|$, where $a_k$ and $b_k$ correspond to the $k$-th principal singular vector pair of $A$ and $B$, respectively. For the purposes here, it is appropriate to use (6.3) to find

$$\delta\left(\{\mathbf{W}(\omega_j)\}_{l@i}, \{\mathbf{W}(\omega_j)\}_{?}\right) \triangleq |\det(\{\mathbf{W}(\omega_j)\}_{l@i}^*\, \{\mathbf{W}(\omega_j)\}_{?})|. \tag{6.4}$$

**Salient features of the classification metric**: The metric due to (6.1) and (6.4) is superior to those proposed by [66] for multivariate time series classification because (i) it evaluates the latent structure (ii) is invariant to orthogonal rotations of the transformation matrix, (iii) applicable in rank-deficient spectral density functions, and (iv) scalable with the number of subbands.

## Classification algorithm

It is now ready to classify a test series $\{y_t\}_?$ based on class affiliations and distances to the $\kappa$-nearest neighbor training series from classes $\mathcal{C}_i, i = 1, \ldots, c$ each with $|\mathcal{C}_i|$ training series whose $l$-th example is $\{y_t\}_{l@i}$, $l = 1, \ldots, |\mathcal{C}_i|$. The classification procedure is simple and is given in Algorithm 7:

**Algorithm 7**: Spectral factor model classification

**Input**: $\{y_t\}_?$, $\{y_t\}_{l@i}, i = 1, \ldots, c; l = 1, \ldots, |\mathcal{C}_i|$;
**Output**: $\hat{\imath} : \{y_t\}_? \in \mathcal{C}_{\hat{\imath}}$
choose Algorithm 3 in Algorithm 6 and for
$j = 1, \ldots, \hat{\jmath}$
    estimate output $\{\mathbf{W}(\omega_j)\}_?$ with input $\{y_t\}_?$;
    estimate output $\{\mathbf{W}(\omega_j)\}_{l@i}$ with input $\{y_t\}_{l@i}$;
compute $\rho(l@i, ?)$ using (6.1) and (6.4);
compute $\hat{\imath}$ using (6.2);

**Note 6.1.** *Algorithm 3 was insisted in Algorithm 7 because the solution based on EM algorithm of Algorithm 5 does not guarantee orthogonal columns for the spectral transformation matrix* $\mathbf{W}$ *for the metric (6.4) to be directly applicable.*

**Note 6.2.** *The optimal parameters of the training episodes* $\{\mathbf{W}(\omega_j)\}_{l@i}$, $j = 1, \ldots, \hat{\jmath}$, $l = 1, \ldots, |\mathcal{C}_i|$ *may be computed offline and only once.*

## 6.3 Multivariate time series prediction

The **prediction problem**, as introduced in Chapter 1, meant reliable estimation of the future evolution of a given time series realization. Subsequently, through the spectral factor model, a parametric time series model was developed; it assumes existence of latent time series that could be dynamically transformed to imitate a higher dimensional multivariate measured time series by inheriting the commonalities of the measured variables. As a solution, it is hoped to drive the future evolution of a given realization by using the commonalities and avoiding the idiosyncarsies.

### Prediction methodology

Insofar as to validate the robustness of the spectral factor model and its underlying assumptions, the intention is to predict the evolution of the time series using the commonalities the spectral factor model could extract from the data. In order to validate that the predictions are benchmarked appropriately, it is necessary to compare the prediction accuracy of the spectral factor model with those of the state-of-the-art multivariate time series models. Then, it seems reasonable to modify the parameters of the state-of-the-art models to be dependent on the commonalities only and assess the accuracy upon that modification.

Fortunately, the aforementioned modification of the state-of-the-art model in the context of this thesis is easy. This is because the spectral factor model was built on the spectral density function or equivalently on the *acvf* of stationary processes; whereas the *acvf*s decompose into parts which are commonalities-dependent and commonalities-independent as per (4.3).

Predicting a multivariate measured time series using commonalities dependent state-of-the-art prediction models accurately should strongly hint that the evolution of the time series is driven by the commonalities. Then, the assumption regarding a latent time series will stand vindicated. On the contrary, if the component time series are all uncorrelated there will not be much to gain in prediction through this approach.

### Classical vector autoregressive prediction

One of the most widely used family of equivalent time series models based on classical **vector autoregressive modeling** of linear processes will be used [102, 51]. This is because the prediction framework in that model is simple to comprehend, popularly tested, and easy to implement. Later, the classical model will be adapted such that its parameters are maximal carriers of commonalities.

The basic principle of vector autoregression is to estimate a future sample of a given realization as a weighted sum of the current and past samples. One may refer [51] among many references to pick from a wide ranging approaches ranging from maximum likehihood estimation, Kalman filter, Bayesian analysis, etc. to time series prediction; in moving forward, just one of those approaches based on linear projections is used. For now, the classical vector autoregressive model may be summarized as follows: For an $r$-variate linear process $\{\mathrm{y}_t\}$ up to the current sample $\mathrm{y}_t$, a simple and valuable version of the prediction problem involves estimating the $s$-th next sample

$y_{t+s|t}$ as a linear function of a finite number $p$ of the present and past samples as

$$\mathsf{y}_{t+s|t} = \epsilon_{t+s} + \sum_{j=0}^{p-1} \phi^{\mathsf{y}}_{j+1,s} \mathsf{y}_{t-j}, \tag{6.5}$$

where $\epsilon_{t+i} = \mathsf{y}_{t+i} - \mathsf{y}_{t+i|t}\, \forall i = 1, 2, \ldots$ is the estimation error and $\phi^{\mathsf{y}}_{l,s} \in \mathbb{R}^{r\times r}\, \forall l = 1, 2, \ldots, p$ are the autoregression coefficient matrices. The condition that ensures minimum mean square error are when the errors $\epsilon_{t+i}$ above are uncorrelated, i.e., $\mathsf{E}[\epsilon_{t+i}\mathsf{y}'_{t-j}] = 0\ \forall j = 0, \ldots, p-1$; refer, e.g., Theorem 4.5 of [48], for this well-known result. It gives rise to the relation between the *acvf*s and the coefficient matrices:

$$\Phi^{\mathsf{y}}_{p,s} = [\phi^{\mathsf{y}}_{1,s}\, \phi^{\mathsf{y}}_{2,s}\, \cdots\, \phi^{\mathsf{y}}_{p,s}]' = (\Xi^{\mathsf{y}}_p)^{-1} \rho^{\mathsf{y}}_{p,s}, \tag{6.6}$$

where

$$\rho^{\mathsf{y}}_{p,s} = \left[\Gamma^{\mathsf{y}}_s\, \Gamma^{\mathsf{y}}_{s+1} \cdots \Gamma^{\mathsf{y}}_{s+p-1}\right]', \tag{6.7}$$

and

$$\Xi^{\mathsf{y}}_p = \begin{bmatrix} \Gamma^{\mathsf{y}}_0 & \Gamma^{\mathsf{y}}_1 & \cdots & \Gamma^{\mathsf{y}}_{p-1} \\ \Gamma^{\mathsf{y}}_{-1} & \Gamma^{\mathsf{y}}_0 & \cdots & \Gamma^{\mathsf{y}}_{p-2} \\ \vdots & \vdots & \ddots & \vdots \\ \Gamma^{\mathsf{y}}_{-p+1} & \Gamma^{\mathsf{y}}_{-p+2} & \cdots & \Gamma^{\mathsf{y}}_0 \end{bmatrix}. \tag{6.8}$$

In practical problems of interest $(\Xi^{\mathsf{y}}_p)^{-1}$ will exist. Therefore, for any given $p$-length subsequence of $\{\mathsf{y}_t\}$ written as

$$\vec{y}_{t,p} = \mathsf{vec}(y_t, y_{t-1}, \ldots, y_{t-p+1}) \in \mathbb{R}^{pr}, \tag{6.9}$$

for the classical vector autoregression on past samples as per (6.5), referring to §4.3 of [51], the minimum mean square error prediction is

$$\hat{y}_{t+s|t} = \Phi^{\mathsf{y}}_{p,s}\, \vec{y}_{t,p}. \tag{6.10}$$

## Spectral factor model prediction

Based on the prediction methodology envisaged in Section 6.3, $\Gamma^{\mathsf{y}}_h$ may be replaced in (6.7) and (6.8) by the part of the *acvf* which inherits the commonalities. The spectral factor model was developed based on the decomposition of the measured multivariate time series $\mathsf{y}_t$ as per (4.2) to $\mathsf{v}_t$ and $\mathsf{z}_t$, which inherit the commonalities and the idiosyncrasies, respectively. The decomposition (4.3) of the *acvf* of $\Gamma^{\mathsf{y}}_h$ into $\Gamma^{\mathsf{v}}_h$ and $\Gamma^{\mathsf{z}}_h$ was also seen. It was further found out that the best approximation of $\Gamma^{\mathsf{y}}_h$ in the sense of inheriting the commonalities is $\Gamma^{\mathsf{v}}_h$ obtained via the spectral factor model. The optimal spectral factor model parameter $\mathsf{S}^{\mathsf{v}}$ is related to $\Gamma^{\mathsf{v}}_h$ through (4.11).

Suppose a maximum commonalities spectral factor model is computed based on a training set of measured time series either via the analytical approach of Section 5.1 or the iterative approach of Section 5.2 according to Algorithm 6. As a result, the optimal transformation matrices $\{\mathbf{W}(\omega_j)\}, j = 1, \ldots, \hat{\jmath}$ at $\hat{\jmath}$ target frequencies may be assumed available. Then, given any subsequence $\vec{y}_{t,p}$ of the measured time series, by replacing $\Gamma^{\mathsf{y}}_h$ with $\Gamma^{\mathsf{v}}_h$ in the prediction equations, predictions may be performed as per Algorithm 8:

**Algorithm 8**: Spectral factor model prediction

**Input**: $\vec{y}_{t,p}$; $\{\mathbf{W}(\omega_j)\}, j = 1, \ldots, \hat{j}$
**Output**: $\hat{y}_{t+s|t}$
compute $\Gamma_h^{\mathsf{v}}$ using (4.11);
compute $\rho_{p,s}^{\mathsf{v}}$ replacing $\Gamma_h^{\mathsf{y}} \to \Gamma_h^{\mathsf{v}}$ in (6.7);
compute $\Xi_p^{\mathsf{v}}$ replacing $\Gamma_h^{\mathsf{y}} \to \Gamma_h^{\mathsf{v}}$ in (6.8);
compute $\Phi_{p,s}^{\mathsf{v}} = (\Xi_p^{\mathsf{v}})^{-1} \rho_{p,s}^{\mathsf{v}}$;
estimate $\hat{y}_{t+s|t} = \Phi_{p,s}^{\mathsf{v}} \, \vec{y}_{t,p}$;

## 6.4 Summary

In practical learning problems one is bound to use spectral factor model with limited computational resources. In Section 6.1, choosing the estimation procedure was discussed; it was based on either (i) the cheaper EM algorithm but with necessary caution to evade local optimum traps or (ii) the accurate but expensive analytical formulas of the low-rank approximation.

For classification of multivariate time series based on the similarities of their commonalities, a metric in (6.1) and a $\kappa$-nearest neighbor classification rule in (6.2) was designed. A test multivariate time series may be classified as belonging to the class of training multivariate time series for which the subspaces spanned by their optimal spectral factor transformation matrices overlap maximally.

For prediction of multivariate time series based on the spectral factor model, the classical vector autoregression prediction models was modified by replacing the measured *acvf* with the *acvf* corresponding to the optimal commonalities.

# Chapter 7

# Experiments

The notion of multivariate time series learning was introduced in Chapter 1 using two practical example problems. One was classification of human MEG signals and the other was prediction of share prices in a portfolio. It was posited that measured multivariate signals in both these problems were generated by dynamic transformation of a low-dimensional latent time series whose *acvf* characteristics are assumed known or given. For convenience, it was assumed that the latent time series is a zero-mean unit-variance white noise.
In designing a modeling framework for multivariate time series in Chapter 4, many merits and challenges in estimating the dynamic transformation in Fourier spectral domain were seen and the modeling framework was called the spectral factor model. In deriving an optimal model in Chapter 5, the following points were considered :

(a) From all possible spectral factor models, a model that is the most likely to have generated the available measured time series according to the principle of maximum likelihood was found. For the maximum likelihood spectral transformation matrix $\mathbf{W}$, through (5.5) it was found that a unique analytical solution is infeasible; whereas an iterative solution in (5.25) was obtained.

(b) From all possible maximum likelihood spectral factor models, the one which maximizes the commonalities inherited by the dynamic transformation was sought. To attain that model, a solution each for the analytical and the iterative procedures via Algorithms 3 and 4, respectively, were formulated.

Through the design of a learning framework in Chapter 6, the following were provided:

(i) A classifier, in Algorithm 7, based on $\kappa$-nearest neighbor proximity of the projection cast by the subspace defined by the optimal spectral factor model transformation of a test time series with the training examples from various classes.

(ii) A vector autoregression prediction scheme, in Algorithm 8, that replaces the *acvf* of the measured time series in the classical prediction equations with the *acvf* corresponding to the commonalities.

Each of these learning objective, viz., classification and prediction, will be experimented with in Sections 7.1 and 7.2, respectively. In both experiments, their data acquisition scheme and the general characteristics of the measured variables will be briefly explained. Importantly, limitations and advantages of these experiments with respect to

the data will be discussed.
One important aspect of a spectral factor model that was taken for granted in the theoretical development was the choice of the latent dimensionality. Hence, in the experiments, its influence on classification and prediction accuracies will be tested. Another aspect of the modeling framework that will be tested is the optimal number of subbands as required by Theorem 2.5 of the asymptotic theory of spectral estimates.

**Implementation**

In addition to the practicalities discussed in Section 6.1, certain implementation aspects of the experiments need to be highlighted. In conducting these experiments, the learning capabilities of the spectral factor model are to be demonstrated. To allow appropriate benchmarking, publicly available data from live fields of study will be used without much expert insights on the processes for which the data was collected. All experiments were conducted using a standard laptop with Intel Dual Core T7200 CPU (2.00GHz). Implementation of the entire estimation and the learning experiments were written using the R language [101]; the codes are intended to be made publicly available through the Comprehensive R Archive Network [4].

## 7.1 Classification of magnetoencephalography signals

In the first of the introductory examples in Section 1.2, the problem of dynamic factor model using the exercise of classification of wrist movements based on magnetoencephalogram (MEG) measurements was described. The task was originally part of a prestigious international competition which has concluded; its solutions have already been published and the winners were announced [1]. The typical approach of participants in the competition involved processing time series to extract certain static time and frequency domain signatures which are then fed to state-of-the-art classifiers. Nevertheless, the competition is attempted here to demonstrate the capability of the spectral factor model in utilizing much of the commonalities captured by the the latent time series presumed for the measured MEG variables for the purpose of determining the particular class of wrist movements responsible for modulating the MEG.
Briefly recap the discussion in Chapter 1 regarding what classification of time series implies: A class of time series may be regarded as an ensemble of finite length time series episodes if they are realizations of the same dynamic transformation of the same latent time series. The dynamic transformation thus represents a class of measured time series process. But remember that the dynamic transformation is such that it allows inheritance of the commonalities maximally from the measured time series. Hence, by comparing the dynamic or spectral transformation matrix of any two measured time series processes, it should be possible to decide which among them a new unclassified measured time series is closest to.

Detailed description and information of the task are available in the competition website of [1]; the data was contributed by [2]. In summary, there are $c = 4$ classes of wrist movements for which $10$ MEG time series are recorded. All movements are appropriately resampled to have $\tau = 400$ samples and have similar stimulus cues and movement procedures. Independent data sets $\mathcal{D}_1$ and $\mathcal{D}_2$ are available for two human subjects; each subject produces $40$ example movements per class and with $73$ and $74$ unlabeled test movements, respectively. The number of test movements per class per

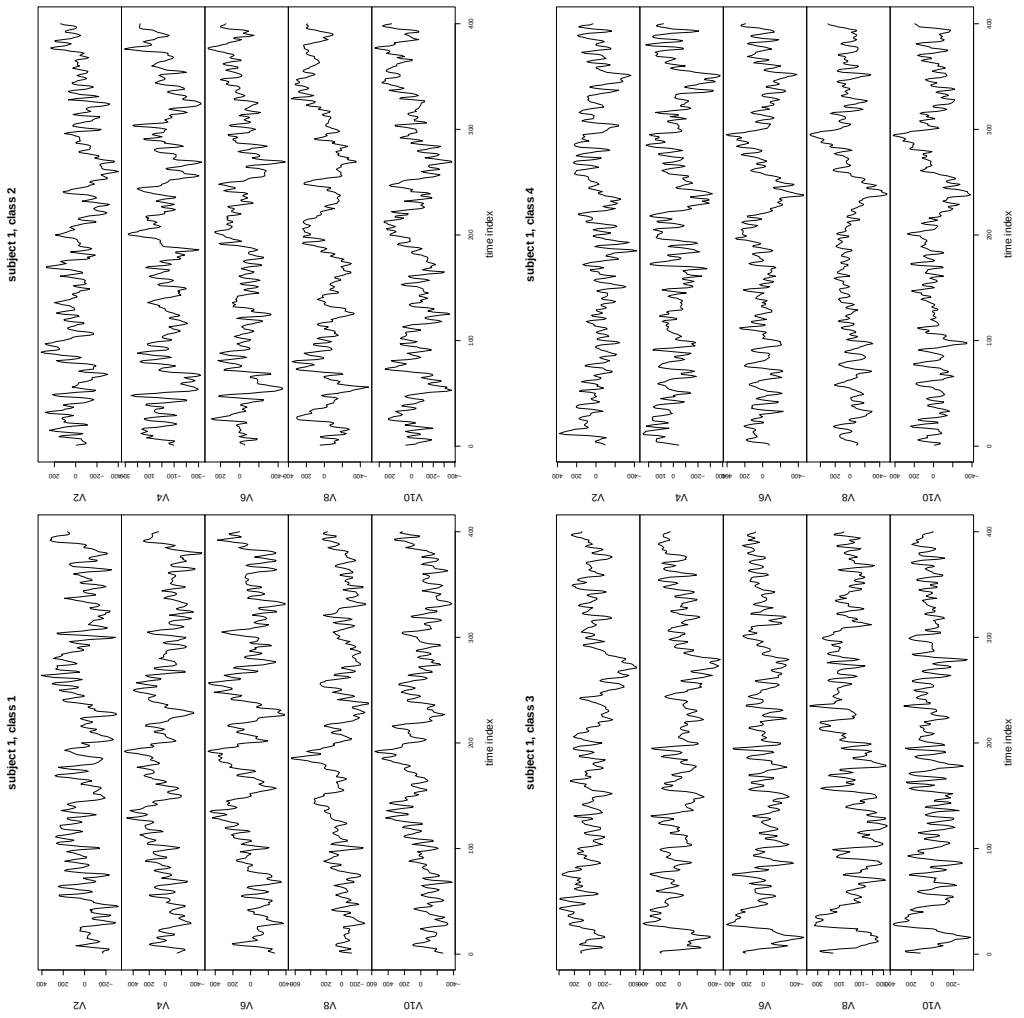


**Figure 7.1:** MEG signals of a human subject $\mathcal{D}_1$ corresponding to five brain spots $(V2, V4, V6, V8, V10)$ during four classes of wrist movements.

subject is also unknown. For neither learning nor testing, there is a need to mix the data from $\mathcal{D}_1$ and $\mathcal{D}_2$ whereas tests are assessed on their average count of classification accuracies, $a_1$ and $a_2$, respectively.

**Testing latent dimensionality**: Since it is a prerogative to estimate an appropriate latent dimensionality $q$ for a given measured dimensionality $r$, the classification accuracy on all possibilities, viz., $q = 1, \ldots, r-1$ will be tested. However, as discussed in Section 6.1, there ought to be sufficient number of samples $n$ within a subband of the discrete Fourier transform of the measured factor model for enhancing reliability of the estimated parameters $\mathbf{W} \in \mathbb{C}^{r \times q}$.

**Testing number of target frequencies**: Yet another constraint that was summarized in Section 6.1 was the number $\hat{j}$ of target frequencies; the sampling rate should be high enough so that sufficiently large $\hat{j}$ number of target frequencies may be assigned to meet the conditions of the asymptotic theory of spectral estimates. Unfortunately,

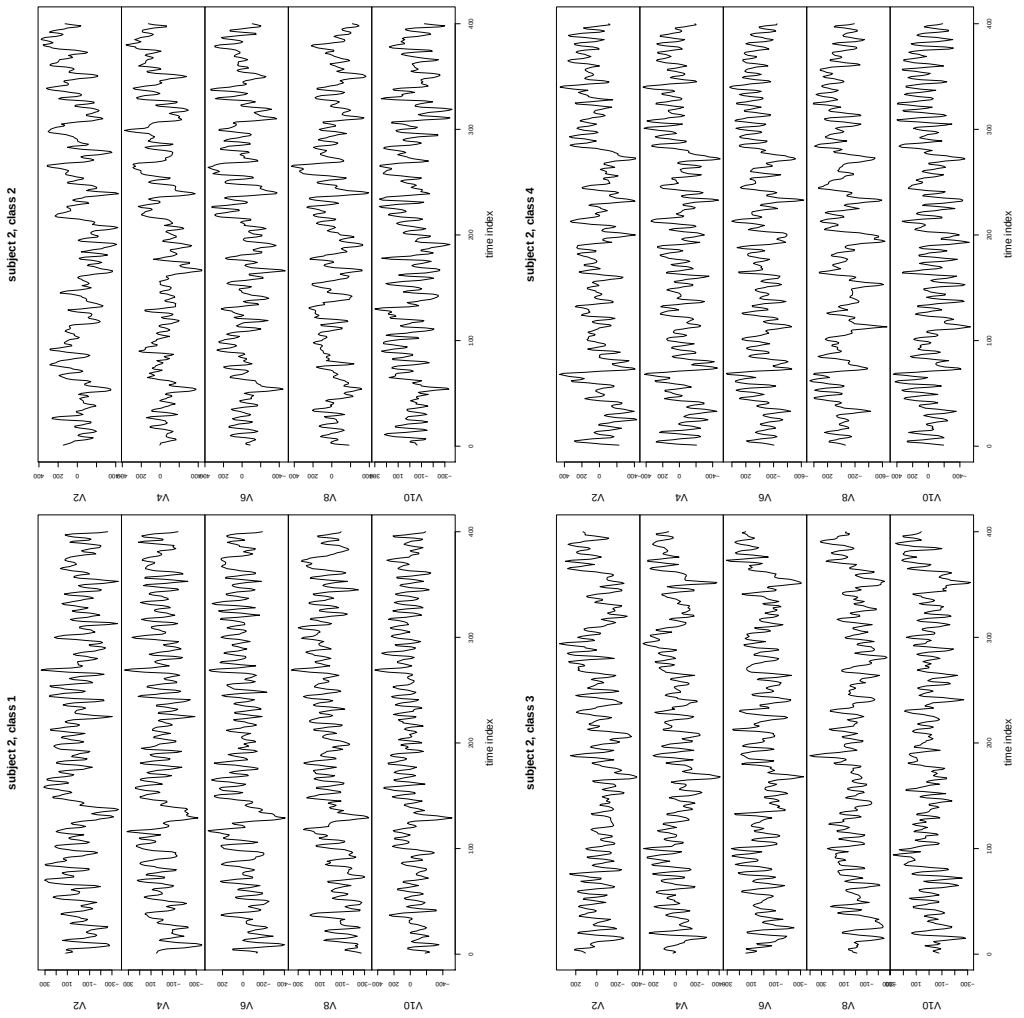


**Figure 7.2:** MEG signals of a human subject $\mathcal{D}_2$ corresponding to five brain spots $(V2, V4, V6, V8, V10)$ during four classes of wrist movements.

the number of samples for the data was just $400$.

**Balancing asymptotic Gaussanity and curse of dimensionality**: A balance has to be struck between the demands for a large number of samples $n$ within a subband for estimating the parameters for a latent dimensionality up to $q = r - 1$ while ensuring that increasing $n$ would not hamper the large number $\hat{j}$ of target frequencies required. It is not the intention to pre-process the data to increase the sample rate or perform other modifications that might lead to explainable bias in comparison of spectral factor model performance with others. As a result, it was decided to use $r = 5$ measured signals only from among the 10 measured signals. In Figures 7.1 and 7.2, these are marked as $(V2, V4, V6, V8, V10)$ instead of $(V1, \ldots, V10)$ of Figure 1.3. The signals $V1, \ldots, V10$ correspond to spatially adjacent parts of the brain; other than that no set of signals seem qualitatively more similar to another set of signals and no particular criteria was used to select the set $(V2, V4, V6, V8, V10)$ of five

measured signals. Obviously, using only part of the measured variables for such a tedious classification exercise invites the risk of losing information rich data that might reflect in poor classification accuracy. As a validation, however, the exercise with the other set $(V1, V3, V5, V7, V9)$ of measured signals will also be carried out.

| | $\hat{\jmath} = 20$ | | $\hat{\jmath} = 25$ | | $\hat{\jmath} = 30$ | |
|---|---|---|---|---|---|---|
| $q$ | $a_1$ | $a_2$ | $a_1$ | $a_2$ | $a_1$ | $a_2$ |
| 1 | 40.54 | 32.88 | 40.54 | 32.88 | 40.54 | 26.03 |
| 2 | 40.54 | 32.88 | 40.54 | 32.88 | 40.54 | 26.03 |
| 3 | 40.54 | 32.88 | 40.54 | 32.88 | 39.19 | 31.51 |
| 4 | 40.54 | 32.88 | 40.54 | 32.88 | 39.19 | 30.14 |

**Table 7.1:** Percentages of average accuracies $a_1$ and $a_2$ in classifying $c = 4$ classes of wrist movements on two subjects $\mathcal{D}_1$ and $\mathcal{D}_2$, respectively, based on their 5-variate MEG $(V2, V4, V6, V8, V10)$. The classifier was based on Algorithm 7 using $\kappa = 3$ for various values for the latent dimensionality $q$ and number of target frequencies $\hat{\jmath}$.

| | $\hat{\jmath} = 20$ | | $\hat{\jmath} = 25$ | | $\hat{\jmath} = 30$ | |
|---|---|---|---|---|---|---|
| $q$ | $a_1$ | $a_2$ | $a_1$ | $a_2$ | $a_1$ | $a_2$ |
| 1 | 40.54 | 32.88 | 35.14 | 32.88 | 40.54 | 32.88 |
| 2 | 39.19 | 32.88 | 36.49 | 32.88 | 41.89 | 26.03 |
| 3 | 39.19 | 32.88 | 36.49 | 32.88 | 41.89 | 31.51 |
| 4 | 35.14 | 32.88 | 35.14 | 32.88 | 40.54 | 30.14 |

**Table 7.2:** Results of the experiments for the 5-variate MEG $(V1, V3, V5, V7, V9)$ with the same setup as in Table 7.1.

The accuracy of the classification are available in Tables 7.1 and 7.2. Note that the data obtained for both those tables are from the same set of processes with a different set of measured variables. However, within a table there are some accuracies which do not seem to change with dimensionality $q$ or number of subbands $\hat{\jmath}$. Explaining such results is attempted below:

**Class imbalance:** The number of test episodes per class was unequal. Note that, had the classes were balanced, the classification is considered to be worse than random classification if accuracies $a_1$ and $a_2$ were below $\frac{1}{c} = 25\%$; whereas perfect classification will imply $100\%$ in any case.

**Nearest neighbours:** Tests on $\kappa = 5$ proved to be not significantly different from those presented in Tables 7.1 and 7.2 for $\kappa = 3$. Whereas for larger $\kappa$, the accuracies were poorer especially for larger $q$ possibly due to the sparsity of consistent training samples against a larger set of features.

**Asymptotic Gaussanity** The subbands tend to lose their distinct Gaussanity with increasing bandwidth, e.g., $\hat{\jmath} = 25$ and $\hat{\jmath} = 20$. In such situations, the classification accuracy becomes invariant as more Gaussian subbands are merged. Subbands could not be increased disproportionately because of the following reason.

| Rank | $a_1$ | $a_2$ | Competing methods |
|---|---|---|---|
| 1 | 59.5 | 34.3 | Reported access to 'bipolar' time series unavailable to others.<br>Fourier and wavelet features selected via genetic algorithm.<br>Support vector and linear discriminant classifiers. |
| 2 | 31.1 | 19.2 | 0-0.5 s segment with 0.5-8 Hz + 20 Hz subsampling<br>Principal Fisher discriminant time and Fourier features.<br>Fisher discriminant classifiers. |
| 3 | 16.2 | 31.5 | Fourier, wavelet features selected via genetic algorithm.<br>Support vector classifiers. |
| 4 | 23.0 | 17.8 | 0-0.5 s segment with 0.5-8 Hz.<br>Principal Fisher discriminant time and Fourier features.<br>Fisher discriminant classifiers. |

**Table 7.3:** Percentage of average accuracies of the winners published by [1].

**Curse of dimensionality:** With $\tau = 400$ and $20$ being the number of transformation matrix parameters for $q = 4$, the subbands with $n = 20; \hat{\jmath} = 20$; $n = 16; \hat{\jmath} = 20$; and $n \approx 13; \hat{\jmath} = 20$ will all challenge the asymptotic theory and suffer from the curse of dimensionality.

**Competition:** It is noteworthy that had the spectral factor model competed in [1] with any $q$ and $\hat{\jmath}$ setting, as shown in Table 7.3, the spectral factor model would have bettered all reported accuracies except against the topper. The topper of the competition seemingly had an advantage of prior knowledge or extra information regarding the time series. Also, no pre-processing of the time series was done unlike the competitors; this is because expertise on the scientific procedure of the data acquisition was lacking nor was it desired to skew the benchmarking of the spectral factor model through unexplainable effects of data pre-processing. However, a basic Bartlett-Hann windowing [50] is performed. This is a standard procedure for discrete Fourier transform techniques to reduce the Gibbs phenomenon as the theoretically periodic finite length realization of a time series is truncated [52].

Moreover, based on available results at [1], the spectral factor model results are a clear front runner despite not requiring any of the advanced process knowledge and preprocessing of the competitors. Also, it is very likely that there was a handicap in the accuracy of the classification due to the inability to use all the measured MEG variables due to the low data sampling rate as explained earlier. Nevertheless, the results obtained demonstrate sufficient classification capabilities of the spectral factor model.

## 7.2 Prediction of yield rates of shares

In Section 1.2, the discussion on the setup of the prediction experiments was initiated through the example of a portfolio of shares obtained from [6]. The same motivation, data and setup are continued here. There is access to a multivariate time series consisting of synchronously sampled daily share prices of 6 German companies over a

period of 2747 trading days during 01/01/1983 - 30/12/1993. As shown in Figure 1.5, the component time series demonstrate similar dynamic covariations when they increase or decrease with observable patterns which are not necessarily readily quantifiable.

It may be verified from Figure 1.5 that there exist increasing and decreasing general trend patterns over a substantial number of samples. Hence, regression detrending on the training series [22] will be performed. The current test series subject to prediction is detrended using the parameters of the regression; the result is displayed in Figure 7.3. Despite this detrending, there still exists obvious non-stationarity in the data.

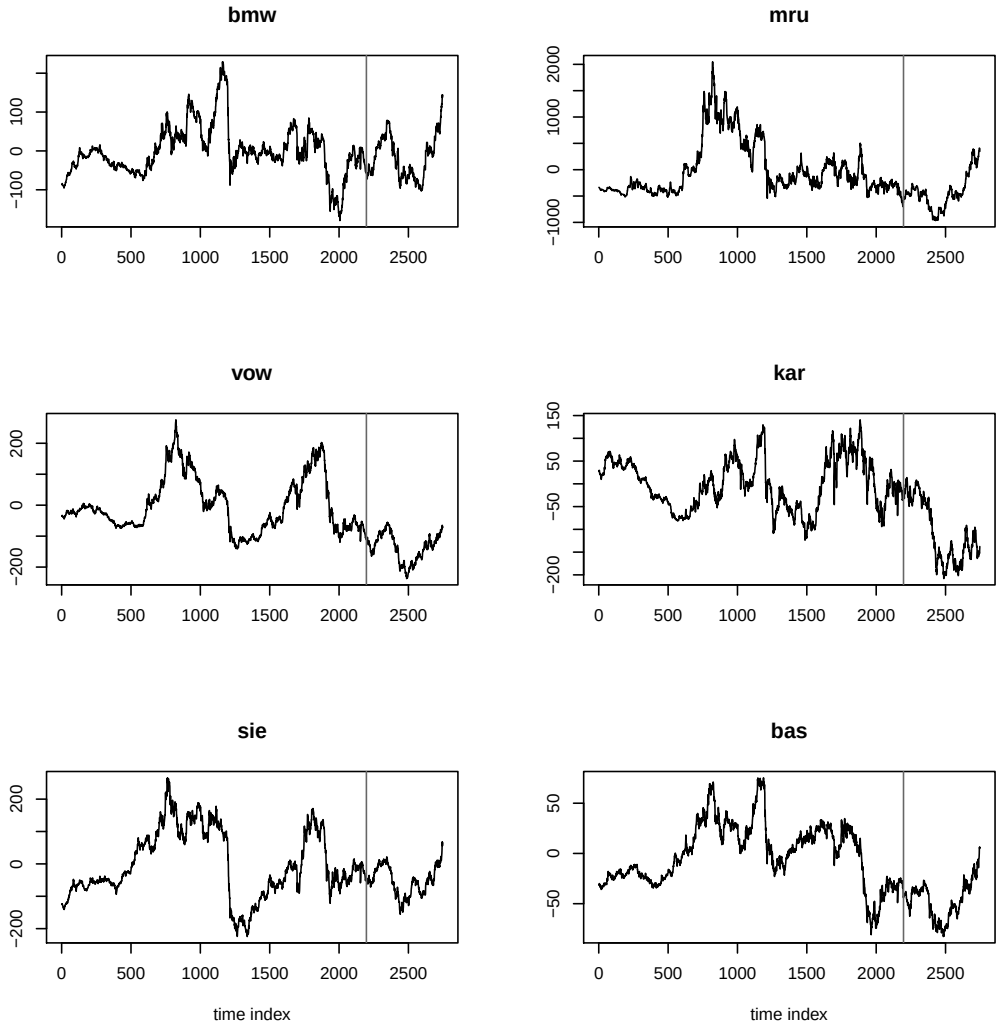


**Figure 7.3:** Original share prices shown in Figure 1.5 are regression detrended and split (by the gray vertical line) into the training series followed by the test series. The test series is corrected using the regression parameters of the training series.

Hence, this prediction exercise will implicitly also test the robustness of the spectral factor model in deviations from the assumption of weak stationarity.

Another pre-processing is effected in the frequency-domain for the robustness of spectral estimates. Prior to estimation of the spectral factor model, windowing of the

time series is performed to reduce the Gibbs phenomenon arising due to disparities between the ends of the finite length realization of the time series [52]. As with the previous experiment, a basic Bartlett-Hann windowing [50] will be performed to the measured time series.

How hard is this chosen problem of prediction of multivariate time series? To answer this question, the predictability of each component series of the multivariate time series has to be checked. In Algorithm 8, the estimate for the time series $\{\mathrm{y}_t\}$ for a horizon $s$ given the current sample $y_t$ and the past $p-1$ samples of the series was developed. On the other hand, a very naïve prediction is to assume that the future evolution is held on the current value. Obviously, the naïvety will incur errors given the stochastic nature of the time series. To measure the accuracy of the prediction, the ratio of the mean of the square errors normalized to the variance of the true time series, called the normalized mean square error (NMSE), is used. The sample counterpart of the population NMSE will be used to assess predictive performance.

naïve prediction

| $s$ | bmw | mru | vow | kar | sie | bas |
|---|---|---|---|---|---|---|
| 1 | 1.24 | 0.83 | 1.33 | 0.96 | 1.53 | 1.8 |
| 2 | 2.84 | 1.77 | 2.66 | 1.97 | 3.11 | 3.57 |
| 3 | 4.65 | 2.78 | 3.94 | 2.97 | 4.8 | 5.25 |
| 4 | 6.54 | 3.83 | 5.24 | 3.83 | 6.36 | 6.66 |
| 5 | 8.58 | 4.84 | 6.55 | 4.72 | 8.14 | 8.14 |
| 6 | 10.57 | 5.97 | 7.99 | 5.69 | 10.04 | 9.71 |
| 7 | 12.47 | 7.07 | 9.35 | 6.73 | 11.81 | 11.12 |
| 8 | 14.37 | 8.25 | 10.77 | 7.75 | 13.55 | 12.61 |
| 9 | 16.36 | 9.48 | 12.34 | 8.86 | 15.40 | 13.91 |
| 10 | 18.30 | 10.78 | 13.90 | 9.95 | 17.20 | 15.09 |
| 20 | 37.09 | 22.56 | 30.79 | 21.08 | 37.44 | 31.08 |

**Table 7.4:** NMSE% of the naïve prediction $\hat{y}_{t+s|t} = y_t$ of each component share price of the portfolio for various horizons $s$.

Table 7.4 gives the NMSE for the naïve prediction of each component measured time series for various horizons $s$. Note that for $s = 1$, i.e., for the next trading day, the naïve prediction is reasonable as the NMSE registers just about 1% prediction error of the variance of their true evolution. For $s = 5$, which generally corresponds to a week-ahead prediction records individual prediction error NMSE averaging between 4 - 9 %, which is neither trivial nor grossly incorrect. For $s = 10$ and $s = 20$ in Table 7.4, it may be seen that the naïve prediction deteriorates substantially for larger horizons.

The spectral factor model prediction methodology is due to Algorithm 8. Following the notations in earlier chapters, the measured dimensionality is $r = 6$ and a latent dimensionality $q < r$ for the spectral factor model is presumed. Within the sufficiency of the number of samples required for a reliable estimation of transformation matrix $\mathbf{W} \in \mathbb{C}^{r\times q}$, an optimal setting of the spectral factor model was tested in trials using a part of the time series dataset for training and another for testing.

As input to Algorithm 8, the $\hat{\jmath}$ spectral factor transformation matrices $\{\mathbf{W}(\omega_j)\}, j = 1, \ldots, \hat{\jmath}$ could be provided via either Algorithms 3 or 5. As mentioned earlier, the EM algorithm requires multiple restarts and the parameters that correspond to the maxi-

mum of the converged likelihood could be chosen for maximizing the commonalities. A numerical log-likelihood convergence difference of $10^{-8}$ and a maximum of 20 iterations were considered appropriate [7]. For the share price portfolio dataset, the EM algorithm typically converged in less than 10 iterations and a maximum of 20 restarts were typically found appropriate in discovering a transformation matrix that is close to to $1\%$ of the log-likelihood of the analytical solution. For the ease of reporting and with the focus on the prediction methodology, the experimental results presented here were carried out with Algorithm 3.

The following observations were made on the prediction accuracy of Algorithm 8 using the spectral factor model as measured by the NMSE on those trials:

(i) an autoregression of order $p = 2$ performed consistently much better than other orders. Hence, $p = 2$ was chosen for the experiments and presenting the results of the tests with orders $p \neq 2$ is skipped.

(ii) increasing the number $\hat{\jmath}$ of subbands of frequencies as stipulated by the asymptotic theory enhanced the prediction accuracy significantly only with $q = 1$. Hence, $\hat{\jmath} = 60$ was picked for the experiments; it corresponds to $n = 36$ discrete frequency transform components per subband which is reasonable for the estimation of the spectral factor model parameters for $r = 6$ measured time series.

It is wished to do predictions of the share prices in terms of a number of trading days, i.e., for the next day ($s = 1$), one week ahead ($s = 5$), a fortnight ahead ($s = 10$), and a month ahead ($s = 22$). Table 7.5 gives the results of the prediction exercise using the spectral factor model as per Algorithm 8 for horizons $s = 1$, $s = 5$, and $s = 10$.

It shows that increasing the latent dimensionality $q$ increases the prediction accuracy with $q = 1$ substantially worse than others and $q = 5$ being the best. This is a logical progression of accuracy that as you increase the latent dimensionality, the commonalities of the measured time series that the spectral transformation could inherit is larger. Hence, higher the latent dimensionality, higher the accuracy or lower the NMSE.

It is the aim to pick a suitable latent dimensionality $q$ by trading accuracy of the prediction NMSE for the number of parameters $rq$. It is numerically obvious from Table 7.5 that there is a significant advantage in terms of the NMSE in picking $q \notin \{1, 2\}$ but $q \in \{3, 4, 5\}$. Moreover, picking $q > 3$ seems not to improve the accuracy much. On comparing the NMSE from Table 7.4 for various horizon with Table 7.5, it is evident that the spectral factor model for $q \in \{3, 4, 5\}$ is a much more accurate long-term predictor than sample *acvf*-based classical autoregressive predictor.

Algorithm 8 recommended replacing the *acvf* $\Gamma_h^{\mathsf{y}}$ of the measured time series $\{\mathsf{y}_t\}$ with the *acvf* $\Gamma_h^{\mathsf{v}}$ of the dynamically transformed latent variables obtained through the spectral factor model estimation. As a result, spectral factor model predictions are assessed with the accuracy of the original predictions with the sample *acvf* using the classical vector autoregression of (6.10). Table 7.6 gives the NMSE% of $\hat{y}_{t+1|t}$ according to (6.10) for various orders $p$ of autoregression. The sharp decline in the prediction of most of the component time series with increasing orders shows that the sample *acvf* estimates are very unreliable; the predictions $\hat{y}_{t+5|t}$ in Table 7.6 also corroborate such a conclusion. Moreover, on comparing Table 7.6 with Table 7.5, it is seen that for $s = 1$ the performance of the spectral factor model with $q \in \{3, 4, 5\}$ is similar in performance to the classical vector autoregression with $p = 1$. On the

spectral factor model-based vector autoregression

| $q$ | bmw | mru | vow | kar | sie | bas |
|---|---|---|---|---|---|---|
| $s = 1$ | | | | | | |
| 1 | 97.79 | 70.25 | 73.77 | 58.32 | 60.39 | 12.65 |
| 2 | 5.66 | 9.78 | 6.05 | 6.36 | 3.12 | 5.68 |
| 3 | 3.47 | 3.99 | 1.60 | 3.69 | 1.52 | 2.36 |
| 4 | 2.45 | 2.14 | 1.20 | 3.41 | 1.19 | 1.69 |
| 5 | 2.36 | 2.05 | 1.17 | 3.32 | 1.17 | 1.64 |
| $s = 5$ | | | | | | |
| 1 | 39.75 | 94.31 | 205.65 | 15.35 | 174.06 | 15.96 |
| 2 | 29.11 | 28.55 | 22.78 | 26.06 | 11.10 | 19.52 |
| 3 | 9.53 | 9.57 | 4.26 | 9.97 | 4.06 | 5.40 |
| 4 | 7.36 | 5.56 | 3.24 | 9.42 | 3.36 | 4.19 |
| 5 | 7.28 | 5.44 | 3.23 | 9.58 | 3.36 | 4.19 |
| $s = 10$ | | | | | | |
| 1 | 92.17 | 236.98 | 143.46 | 168.84 | 195.85 | 29.74 |
| 2 | 52.52 | 30.86 | 9.70 | 63.44 | 11.35 | 10.91 |
| 3 | 13.96 | 10.27 | 7.05 | 17.02 | 7.83 | 7.95 |
| 4 | 13.96 | 9.61 | 6.62 | 16.81 | 6.98 | 7.90 |
| 5 | 13.51 | 9.43 | 6.36 | 16.73 | 6.74 | 7.72 |
| $s = 22$ | | | | | | |
| 1 | 191.34 | 390.07 | 213.13 | 310.42 | 458.14 | 96.42 |
| 2 | 89.85 | 82.62 | 23.09 | 112.64 | 32.94 | 25.56 |
| 3 | 37.23 | 24.44 | 17.46 | 33.91 | 23.74 | 18.37 |
| 4 | 33.37 | 20.90 | 15.11 | 32.29 | 17.68 | 18.15 |
| 5 | 32.51 | 20.47 | 14.62 | 33.44 | 17.02 | 17.54 |

**Table 7.5:** NMSE% of the predictions for the next day ($s = 1$), one week ahead ($s = 5$), a fortnight ahead ($s = 10$), and a month ahead ($s = 22$) for each component share price of the portfolio for various latent dimensions $q$; $\hat{j} = 60$ and $p = 2$ were chosen.

other hand, for $s = 5$, spectral factor model with $q \in \{3, 4, 5\}$ is clearly outperforming the classical vector autoregression. For even higher horizons of $s = 10$ and $s = 20$, the classical vector autoregression is immensely worse in performance than the spectral factor model and the results are, hence, not presented.

classical vector autoregression

| $p$ | bmw | mru | vow | kar | sie | bas |
|---|---|---|---|---|---|---|
| $s = 1$ | | | | | | |
| 1 | 1.26 | 0.88 | 2.05 | 1.79 | 1.83 | 2.82 |
| 2 | 5.26 | 0.99 | 18.32 | 10.58 | 23.43 | 87.87 |
| 3 | 8.83 | 1.03 | 35.63 | 18.58 | 41.21 | 183.35 |
| 4 | 12.18 | 1.10 | 59.40 | 30.03 | 60.94 | 287.96 |
| 5 | 17.71 | 1.24 | 90.14 | 44.09 | 85.69 | 437.76 |
| $s = 5$ | | | | | | |
| 1 | 9.60 | 5.75 | 28.25 | 30.88 | 17.35 | 46.99 |
| 2 | 17.38 | 5.77 | 60.54 | 46.12 | 50.35 | 191.55 |
| 3 | 23.60 | 5.78 | 91.18 | 58.54 | 75.25 | 323.58 |
| 5 | 38.84 | 5.89 | 175.62 | 89.47 | 136.50 | 668.26 |
| 10 | 111.52 | 7.15 | 498.85 | 161.82 | 330.19 | 2438.70 |

**Table 7.6:** NMSE% of one day ahead ($s = 1$) and one week ahead ($s = 5$) predictions of each component share price of the portfolio for various orders $p$ of autoregression.

# Chapter 8

# Extensions

We had set out to build a model for multivariate time series that will enable developing strategies to predict or classify unseen data. In Chapter 4, the spectral factor model as the frequency-domain counterpart of the time-domain dynamic factor model was proposed. It was argued how dynamic transformation of a latent time series to a multivariate measured time series may be modeled to imbibe the characteristics that are common to any two measured variables. Those characteristics were called the commonalities. Maximally inheriting the maximum-likelihood cross-covariations was the estimation strategy adopted in Chapter 5 to model the commonalities. There, analytical formulae as well as an iterative algorithm for estimating the spectral factor model parameters corresponding to the optimal commonalities were presented. In Chapter 6, a classification rule was derived and a prediction methodology for multivariate time series based on the spectral factor model was designed. The experiments presented in Chapter 7 validate that the commonalities as defined, designed, and determined for the spectral factor model possess substantial classification and prediction capabilities for many real-world multivariate time series problems.
This thesis is concluded by highlighting a large number of possibilities that await in extending the spectral factor model. As pointed out, there are some improvements to the presented work that could be attained by overcoming the limitations and relaxing the assumptions.

The spectral factor model was developed aiming for applications involving multivariate time series learning. As seen in earlier chapters, the spectral factor model transformation matrix was the parameter that took the most focus in classification and prediction experiments that were carried out. There could be many more other applications possible through the spectral factor model along the same concepts as were presented. However, there are certain strong assumptions the spectral factor model is grounded on; they might pose some challenges for its widespread use. Therefore, in what follows, possibilities of developing the spectral factor model beyond the current design are investigated; some of the essential further investigations that are wished to be performed more formally are listed.

## 8.1 Challenges

Certain aspects that have come across as limitations and annoyances for the spectral factor model are as follow:

**Linearity**: The spectral factor model is rooted on the assumption that the measured multivariate time series is a linear weakly stationary process; this was essential for utilizing Theorem 2.5. In practice, given the samples of a realization of a time series, it is not easy to validate its linearity [13, 67]. This important, but broader scoped, issue was not addressed much in this thesis. The appropriateness of using Fourier domain methods is also at the mercy of this assumption. Hence, in using spectral factor model for the purposes dealt with in this thesis, it is recommended to tie the prediction or classification results with some appropriate test of nonlinearity of the measured multivariate time series data. Alternatively, lazy learning formalism of the spectral factor model could also be pursued whereby the model parameters will adapt to reflect the locality of the operating regime [16].
**Commonalities**: In developing the spectral factor model, existence of physically valid cross-covariation between the measured variables was naïvely assumed. It should, therefore, be borne in mind that applying it to independent or uncorrelated variables might show up numerically non-trivial off-diagonal *acvf* but whose interpretations might most certainly be illogical. In that respect, more robust estimation procedures of the sample spectral density of the type in [106] would be a path forward.
**Pre-processing**: As seen in the experiments, certain pre-processing of the time series was required to rectify obvious deviations of measured data from the assumption of weak stationarity. A rigorous and detailed study is envisaged for assessing the impact of the two preprocessing steps that used, viz., detrending and windowing [22, 65]; this is beyond the realms of this thesis.
**Sample size**: It is required a sufficient number of samples within a subband be maintained in order to obtain a reliable estimate of the spectral factor model parameters without inviting the curse of dimensionality [12]. Meanwhile, as dealt with in the experiments and in Algorithm 1, a sufficiently large number of subbands as required by Theorem 2.5 is to be maintained too. This balancing act was performed by testing on an array of choices regarding the number of parameters to be estimated and the sample size. This approach is perhaps not suitable when it is not clear whether the spectral factor model has learned. E.g., had the future series in the prediction exercise or true class labels in the classification experiments were lacking, evaluation of the prediction or classification accuracy, and therefore, the quality of the estimation would not have been possible. In essence, more theoretical efforts have to proceed beyond experimental validation and benchmarking towards determining an appropriate latent dimensionality $q$ and the $\hat{\jmath}$ number of target frequencies for learning problems.

## 8.2 Further work

Certain realistic extensions to the spectral factor model beyond the objectives originally meant for it are listed below:
**Process understanding**: With an agreeable performance in a learning problem, it might be inferred that the presumed latent dimensionality $q$ is credible. This hint regarding the dimensionality of the presumed latent process $\{\mathsf{x}_t\}$ is a preliminary step towards better understanding of a complicated high-dimensional time series generating system. In addition, there is a possibility to assert any process knowledge gained from human experience through the *acvf* $\Gamma_h^{\mathsf{x}}$ of the latent linear process. Such an *acvf* could replace the default assumption in (4.6) of $\{\mathsf{x}_t\}$ being a zero mean unit variance white noise.

Two other quantities that certainly will aid better understanding of the process under investigation are $\mathbf{x}(\omega_l)$ as per (5.30) and computed in (5.20) as well as $\tilde{\mathbf{v}}(\omega_l)$ estimated in (5.34). An inverse discrete Fourier transform of these quantities should enable now-casting [10], which is to provide a better assessment of the present and the past of the latent characteristics of the process.

**Clustering**: Suppose there are no class labels for an ensemble of various episodes from various time series processes and it is the intention to cluster them based on the characteristics of their commonalities. Then, a scheme similar to $\kappa$-means clustering [64] or any variations thereof might be adopted. Towards such a purpose, $\rho(i,k) = \sum_{j=1}^{\hat{j}} \delta\left(\{\mathbf{W}(\omega_j)\}_i, \{\mathbf{W}(\omega_j)\}_k\right)$ may be used as the distance between any two time series episodes $\{y_t\}_i$ and $\{y_t\}_k$ computed across all the $\hat{j}$ spectral factor model subbands where $\delta\left(\{\mathbf{W}(\omega_j)\}_i, \{\mathbf{W}(\omega_j)\}_k\right) = |\det(\{\mathbf{W}(\omega_j)\}_i^* \{\mathbf{W}(\omega_j)\}_k)|$ is the overlap for the $j$-th subband according to (6.1).

**Real-time implementation**: The computational aspects of a practical implementation of the spectral factor model was discussed in Section 6.1. There exist many multivariate monitoring applications, e.g., algorithmic trading [21], industrial plant monitoring [76], automated anesthesia [56], where frequent assessment and update of the model are necessary but prohibitive. In such problems, it is envisaged to use either the inexpensive EM-algorithm for incremental updates or approaches such as with online principal component analysis [73] for a real-time update to the analytical spectral factor model solution.

## 8.3 Summary

In everyday life, in business, health, search engines, etc., we are witnessing an immensely increasing demand for robust and efficient models for machine learning. Such models are necessary to meet objectives ranging from real-time computational decision support to scalable pattern recognition based on the multivariate time series they generate. It was demonstrated through reviews, contributions, and experiments that the dynamic and spectral factor models are live and active fields of research due to the simplified understanding of a complicated process they offer. The improvements and extensions to the modeling and learning frameworks of this thesis are near and feasible for them to deal with more diverse and real-world time series challenges. The spectral factor model promises to be a viable way forward for mastering the process generating the increasing volumes of multivariate streaming data.

# Bibliography


[1] www.bbci.de/competition/iv, 2008.

[2] www.bmi.uni-freiburg.de, 2008.

[3] www.cnsorg.org, 2011.

[4] www.cran.r-project.org, 2011.

[5] www.elekta.com, 2011.

[6] www.stat.uni-muenchen.de, 2011.

[7] R. Abbi, E. El-Darzi, C. Vasilakis, and P. Millard. Analysis of stopping criteria for the EM algorithm in the context of patient grouping to length of stay. In *Proc. of the 4th Intl. IEEE Conf. on Intelligent Systems*, 2008.

[8] F. Ahrabi. Maximum likelihood estimation of the autoregressive coefficients and moving average covariances of vector autoregressive moving average models. Technical Report 39, Department of Statistics, Stanford University, 1979.

[9] E. J. Aubert, I. A. Lund, and A. Thomasell Jr. Some objective six-hour predictions prepared by statistical methods. *J. Meteor.*, 16:436–446., 1959.

[10] M. Bánbura, D. Giannone, and L. Reichlin. Nowcasting. In *Working Paper Series*. European Central Bank, European Central Bank, December 2010.

[11] D. J. Bartholomew, M. Knott, and I. Moustaki. *Latent Variable Models and Factor Analysis: A Unified Approach*. Wiley Series in Probability and Statistics. Wiley, 3rd edition, 2011.

[12] R. E. Bellman. *Adaptive Control Processes: A Guided Tour*. Princeton University Press, 1961.

[13] P. J. Bickel and P. Bühlmann. What is a linear process? *Proc. Natl. Acad. Sci.*, 93:12128–12131, 1996.

[14] C. M. Bishop and M. E. Tipping. Probabilistic principal component analysis. *Journal of the Royal Statistical Society, Series B*, 21(3):611–622, 1999.

[15] V. I. Bogachev. *Gaussian Measures*. Mathematical Surveys and Monographs. American Mathematical Society, 1998.

[16] G. Bontempi. *Local Learning Techniques for Modeling, Prediction and Control*. PhD thesis, Université Libre de Bruxelles, 1999.

[17] W. L. Briggs and V. E. Henson. *The DFT: An Owner's Manual for the Discrete Fourier Transform*. SIAM, Philadelphia, 1995.

[18] D. R. Brillinger. *Time Series: Data Analysis and Theory*. Holden-Day, San Francisco, 1981.

[19] P. J. Brockwell and R. A. Davis. *Introduction to Time Series and Forecasting*. Springer, 2nd edition, 2002.

[20] P. J. Brockwell and R. A. Davis. *Time Series: Theory and Methods*. Springer,

2nd edition, 2009.

[21] J. A. Brogaard. High frequency trading and its impact on market quality. Technical report, Kellogg School of Management, Northwestern University, July 2010.

[22] R. Chandler and M. Scott. *Statistical Methods for Trend Detection and Analysis in the Environmental Sciences*. John Wiley & Sons, Inc., 2011.

[23] U. Cherubini, G. Della Lunga, P. Rossi, and S. Mulinacci. *Fourier Transform Methods in Finance*. John Wiley & Sons, Inc., 2010.

[24] S. Choi, A. Cichocki, H.-P. Park, and S.-Y Lee. Blind Source Separation and Independent Component Analysis: A Review. *Neural Information Processing - Letters and Reviews*, 6(1):1–57, 2005.

[25] T. Chonavel. *Statistical Signal Processing: Modelling and Estimation*. Springer, 2002.

[26] A. Cichocki and R. Thawonmas. On-line algorithm for blind signal extraction of arbitrarily distributed, but temporally correlated sources using second order statistics. *Neural Processing Letters*, 12(1):91–98, 2000.

[27] P. Comon. Independent component analysis, a new concept? *Signal Processing*, 36:287–314, 1994.

[28] T. Cover. Estimation by the nearest neighbor rule. *IEEE Tran. Info. Theory*, 14(1):50–55, 1968.

[29] D. R. Cox and H. D. Miller. *The Theory of Stochastic Processes*. Chapman and Hall, 1977.

[30] K. R. Davidson and A. P. Donsig. *Real Analysis with Real Applications*. Prentice Hall, 2001.

[31] B. Delyon, M. Lavielle, and E. Moulines. Convergence of a stochastic approximation version of the EM algorithm. *The Annals of Statistics*, 27(94):128, 1999.

[32] J. W. Demmel. *Applied Numerical Linear Algebra*. SIAM, 1997.

[33] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.

[34] W. J. den Haan and A. T. Levin. Vector autoregressive covariance matrix estimation. *University of California at San Diego*, February 1998.

[35] I. Elishakoff. *Probabilistic Theory of Structures*. Wiley, 2nd edition, 1999.

[36] M. Forni, M. Hallin, M. Lippi, and L. Reichlin. The Generalized Dynamic Factor Model: One-Sided Estimation and Forecasting. *Journal of the American Statistical Association*, 100(471):830–840, September 2005.

[37] M. Forni and M. Lippi. The Generalized Dynamic Factor Model: Representation Theory. *Econometric Theory*, 17:1113–1141, 2001.

[38] Y. Fujikoshi, V. V. Ulyanov, and R. Shimizu. *Multivariate Statistics : High-Dimensional and Large-Sample Approximations*. John Wiley & Sons, Inc., 2010.

[39] W. Fuller. *Introduction to Statistical Time Series*. Wiley, New York, 1996.

[40] A. Galántai and Cs. J. Hegedűs. Jordan's principal angles in complex vector spaces. *Numerical Linear Algebra with Applications*, 13:589–598, 2006.

[41] C. Gasquet and P. Witomski. *Fourier Analysis and Applications: Filtering, Numerical Computation, Wavelets*. Springer-Verlag, New York, 1999.

[42] J. S. George, C. J. Aine, J. C. Mosher, et al. Mapping function in the human

brain with magnetoencephalography, anatomical magnetic resonance imaging, and functional magnetic resonance imaging. *J Clin Neurophysiolo.*, 12(5):406–31, September 1995.

[43] J. F. Geweke and K. J. Singleton. Maximum likelihood "confirmatory" factor analysis of economic time series. *International Economic Review*, 22(1):37–54, 1981.

[44] S. Ghosh. *Signals and Systems*. Pearson Education, 2006.

[45] N. R. Goodman. Statistical analysis based on a certain multivariate complex Gaussian distribution (An introduction). *The Annals of Mathematical Statistics*, 34(1):152–177, 1963.

[46] R. L. Gorsuch. *Factor Analysis*. Psychology Press, 2nd edition, 1983.

[47] R. M. Gray. *Probability, Random Processes, and Ergodic Properties*. Springer, 1987.

[48] R. M. Gray and L. D. Davisson. *An Introduction to Statistical Signal Processing*. Cambridge University Press, 2005.

[49] G. R. Grimmett and D. R. Stirzaker. *Probability and Random Processes*. Oxford University Press, 3rd edition, 2001.

[50] Y. H. Ha and J. A. Pearce. A new window and comparison to standard windows. *IEEE Trans. on Acoustics, Speech, and Signal Processing*, 37(2):298–301, 1999.

[51] J. D. Hamilton. *Time Series Analysis*. Princeton University Press, 1994.

[52] R. W. Hamming. *Digital Filters*. Dover Publications, 3rd edition, 1998.

[53] E. J. Hannan and P. J. Thomson. Spectral inference over narrow bands. *Journal of Applied Probability*, 8(1):157–169, 1971.

[54] W. K. Härdle and L. Simar. *Applied Multivariate Statistical Analysis*. Springer, 2007.

[55] H. H. Harman. *Modern Factor Analysis*. The University of Chiacago Press, 1976.

[56] T. Hemmerling. *Automatic control system and method for control of anesthesia (CA2009/001491)*. Canadian Intellectual Property Office, 2010.

[57] A. Hjørungnes and D. Gesbert. Complex-valued matrix differentiation: Techniques and key results. *IEEE Trans. Signal Processing*, pages 55(6):2740–46, June 2007.

[58] P. J. Huber and E. Ronchetti. *Robust statistics*. Wiley, 2nd edition, 2009.

[59] R. Hunger. An Introduction to Complex Differentials and Complex Differentiability. Technical Report TUM-LNS-TR-07-06, Technische Universität München, 2007.

[60] A. Hyvärinen. *Independent component analysis*. John Wiley & Sons, Inc., 2001.

[61] A. Hyvärinen and Y. Kano. Independent component analysis for non-normal factor analysis. Technical report, Neural Networks Research Centre, Helsinki University of Technology, 2003.

[62] A. J. Izenman. *Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning*. Springer, 2008.

[63] L. B. Jackson. *Digital Filters and Signal Processing: With MATLAB Exercises*. Kluwer Academic, 3rd edition, 1995.

[64] A. K. Jain. Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, 31(8):651–666, 2010.

[65] A. H. Kaiser. *Digital Signal Processing using the Fast Fourier Transform (FFT)*.

GRIN Verlag, 2007.

[66] Y. Kakizawa, R. H. Shumway, and M. Taniguchi. Discrimination and Clustering for Multivariate Time Series. *Journal of the American Statistical Association*, 93(441):328–340, 1998.

[67] H. Kantz and T. Schreiber. *Nonlinear Time Series Analysis*. Cambridge University Press, 2nd edition, 2004.

[68] S. Kay. *Intuitive Probability and Random Processes using MATLAB*. Springer, 2005.

[69] B. F. King. Market and industry factors in stock price behavior. *The Journal of Business*, 39(1):pp. 139–190, 1966.

[70] K. Knight. *Mathematical Statistics*. Chapman and Hall / CRC, 1999.

[71] S. G. Krantz. *Function Theory of Several Complex Variables*. American Mathematical Society, 1992.

[72] S. Kumaresan. *Short Courses in Mathematics*. University Press, Hyderabad, 2003.

[73] D. Kuzmin and M. K. Warmuth. *Online kernel PCA with entropic matrix updates*. Intl. Conf. on Machine Learning, 2007.

[74] T. Lancaster. *An Introduction to Modern Bayesian Econometrics*. Blackwell, Malden, 2004.

[75] S. Lang. *Introduction to Linear Algebra*. Springer, 2nd edition, 1985.

[76] T. Larsson and S. Skogestad. Plantwide control - a review and a new design procedure. *Modeling, Identification and Control*, 21(4):209–240, 2000.

[77] A. Leon-Garcia. *Probability, Statistics, and Random Processes For Electrical Engineering*. Prentice Hall, 3rd edition, 2008.

[78] H. Lütkepohl. *New Introduction to Multiple Time Series Analysis*. Springer, 2005.

[79] H. Madsen. *Time Series Analysis*. Chapman and Hall / CRC, 2007.

[80] K. V. Mardia, J. T. Kent, and J. Bibby. *Multivariate Analysis*. Academic Press, 1979.

[81] A. L. McCutcheon. *Latent Class Analysis*. SAGE Publications Inc, 1987.

[82] J. McLachlan and D. Peel. *Finite Mixture Models*. Wiley Series in Probability and Statistics. Wiley-Interscience, 2000.

[83] K. Metaxoglou and A. Smith. Maximum likelihood estimation of VARMA models using a state-space EM algorithm. *Journal of Time Series Analysis*, 28(5):666–685, 2007.

[84] C. D. Meyer. *Matrix Analysis and Applied Linear Algebra*. SIAM, 2001.

[85] J. Miao and A. Ben-Israel. On Principal Angles between Subspaces in $\mathbb{R}^n$. *Lin. Algeb. Appl.*, 171:81–98, 1992.

[86] S. Miller and D. Childers. *Probability and Random Processes: With Applications to Signal Processing and Communications*. Academic Press, 2nd edition, 2004.

[87] T. P. Minka. Expectation-Maximization as lower bound maximization. Technical report, Microsoft Research, 1998.

[88] A. A. Miranda, Y-.A. Le Borgne, and G. Bontempi. New Routes from Minimal Approximation Error to Principal Components. *Neural Processing Letters*, pages 197–207, 2008.

[89] S. H. Mneyney. *An Introduction to Digital Signal Processing*. River Publishers

ApS, Aalborg, 2008.

[90] P. Molenaar. A Dynamic Factor Model for the Analysis of Multivariate Time Series. *Psychometrika*, 50(2):181–202, June 1985.

[91] M. Nerlove. Spectral Analysis of Seasonal Adjustment Procedures. *Econometrica*, 32(3):241–286, 1964.

[92] S. Nsiri and R. Roy. On the invertibility of multivariate linear processes. *Journal of Time Series Analysis*, 14:305–316, 1993.

[93] A. V. Oppenheim and R. W. Schafer. *Discrete-Time Signal Processing*. Prentice Hall Press, 2009.

[94] M. J. Panik. *Advanced Statistics from an Elementary Point of View*. Academic Press, 2005.

[95] A. Papoulis. *Probability, Random Variables, and Stochastic Processes*. McGraw-Hill, 1965.

[96] Y. Pawitan. *In All Likelihood: Statistical Modelling and Inference Using Likelihood*. Oxford University Press, 2001.

[97] T. M. Peters and J. C. Williams. *The Fourier Transform in Biomedical Engineering*. Applied and Numerical Harmonic Analysis. Birkhäuser Boston, 1998.

[98] K. B. Petersen and M. S. Pedersen. *The Matrix Cookbook*. Technical University of Denmark, 2008.

[99] M. B. Priestley. *Spectral Analysis and Time Series*, volume I and II. Academic Press, 1983.

[100] M. B. Priestley, T. Subba Rao, and H. Tong. Applications of Principal Component Analysis and Factor Analysis in the Identification of Multivariable Systems. *IEEE Transactions on Automatic Control*, 19(6):730–734, December 1974.

[101] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2011.

[102] G. C. Reinsel. *Elements of Multivariate Time Series Analysis*. Springer, 2nd edition, 2003.

[103] R. A. Reyment and K. G. Jöreskog. *Applied Factor Analysis in the Natural Sciences*. Cambridge University Press, 2nd edition, 1996.

[104] T. J. Sargent and C. A. Sims. Business cycle modelling without pretending to have too much a priori economic theory. In C. A. Sims, editor, *New Methods in Business Research*, Minneapolis, 1977. Federal Reserve Bank of Minneaopolis.

[105] T. Särkämö, M. Tervaniemi, S. Soinila, et al. Auditory and cognitive deficits associated with acquired amusia after stroke: A magnetoencephalography and neuropsychological follow-up study. *PLoS ONE*, 5(12):e15157, 12 2010.

[106] J. Schäfer and K. Strimmer. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology*, 4(1), 2005.

[107] A. M. Scher, A. C. Young, and W. M. Meredith. Factor Analysis of the Electrocardiogram. *Circulation Research*, 8:519–526, 1960.

[108] M. J. Schervish. *Theory of Statistics*. Springer-Verlag, 1995.

[109] A. Sen and M. Srivastava. *Regression Analysis: Theory, Methods, and Applications*. Springer, 1990.

[110] A. Shiryaev. *Probability Theory and Stochastic Processses*, volume 95 of *Graduate Texts in Mathematics*. Springer, 2nd edition, 1996.

[111] R. H. Shumway and D. S. Stoffer. *Time Series Analysis and Its Applications*. Springer, 2nd edition, 2006.

[112] G. E. Silverman and R. A. Shilov. *Linear Algebra*. Dover Publications, 1977.

[113] D. S. Sivia and J. Skilling. *Data Analysis – A Bayesian Tutorial*. Oxford Science Publications, 2nd edition, 2006.

[114] C. Spearman. General Intelligence Objectively Determined and Measured. *American Journal of Psychology*, 15:201–293, 1904.

[115] H. Stefan, G. Scheler, C. Hummel, et al. Magnetoencephalography (meg) predicts focal epileptogenicity in cavernomas. *Journal of Neurology, Neurosurgery & Psychiatry*, 75(9):1309–1313, 2004.

[116] G. W. Stewart. *Matrix Algortithms: Basic Decompositions*, volume 1. SIAM, 1998.

[117] J. H. Stock and M. W. Watson. New Indexes of Coincident and Leading Economic Indicators. *NBER Macroeconomics Annual*, 4:351–394, 1989.

[118] J. H. Stock and M. W. Watson. Macroeconomic Forecasting using Diffusion Indexes. *Journal of Business & Economic Statistics*, 20(2):147–162, 2002.

[119] J. H. Stock and M. W. Watson. *Introduction to Econometrics*. Addison Wesley, 2003.

[120] L. N. Trefethen and D. Bau III. *Numerical Linear Algebra*. SIAM, 1997.

[121] W. J. van der Linden and R. K. Hambleton, editors. *Handbook of Modern Item Response Theory*. Springer, 1997.

[122] A. W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2000.

[123] G.C.G Wei and M.A. Tanner. A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association*, 85:699–704, 1990.

# Appendix A

## A.1 Differentiation of real-valued functions of complex variables

Some properties of functions which map complex-valued variables to real-valued images is reviewed here. For details and applications of such an analysis, [57] is referred to. Suppose $\mathcal{A} \subset \mathbb{C}$ is an open set and a complex function $f(u) : \mathcal{A} \to \mathbb{C}$ is defined. The function $f(u)$ is said to be differentiable at $\hat{u} \in \mathcal{A}$ if its derivative at $\hat{u}$ defined as

$$\frac{\mathsf{d}}{\mathsf{d}u} f(u)\bigg|_{\hat{u}} = \lim_{u \to \hat{u}} \frac{f(u) - f(\hat{u})}{u - \hat{u}}, \tag{A.1}$$

exists. The function $f(u)$ is said to be analytical if the derivative exists for all $\hat{u} \in \mathcal{A}$. For analytical functions, the stationary points are located wherever

$$\frac{\mathsf{d}}{\mathsf{d}u} f(u) = 0. \tag{A.2}$$

The differential of an analytical $f(u)$ is given by

$$\mathsf{d}f(u) = \frac{\partial}{\partial u} f(u) \mathsf{d}u + \frac{\partial}{\partial \bar{u}} f(u) \mathsf{d}\bar{u}, \tag{A.3}$$

where $\bar{u} = u_1 - \mathsf{i}u_2$ is the complex conjugate of $u = u_1 + \mathsf{i}u_2$, where $u_1, u_2 \in \mathbb{R}$ and

$$\begin{aligned} \frac{\partial}{\partial u} &= \frac{1}{2}\left(\frac{\partial}{\partial u_1} - \mathsf{i}\frac{\partial}{\partial u_2}\right), \\ \frac{\partial}{\partial \bar{u}} &= \frac{1}{2}\left(\frac{\partial}{\partial u_1} + \mathsf{i}\frac{\partial}{\partial u_2}\right) \end{aligned} \tag{A.4}$$

are called Wirtinger derivatives. Also, note a direct consequence of (A.4) that

$$\frac{\partial}{\partial \bar{u}} u = \frac{\partial}{\partial u} \bar{u} = 0, \tag{A.5}$$

or $\bar{u}$ may be regarded as a constant when differentiating with respect to $u$, and vice-versa.
For any $f(u)$ that is not necessarily analytical, based on the condition (A.2), the stationary points may now be found by searching where

$$\mathsf{d}f(u) = 0. \tag{A.6}$$

Let $f(u) = f_1(u_1, u_2) + \mathrm{i} f_2(u_1, u_2)$, where $f_1, f_2 : \mathbb{R}^2 \to \mathbb{R}$. For $f(u)$ to be analytical, it is necessary that it satisfies the Cauchy-Riemann conditions

$$\frac{\partial}{\partial u_1} f_1 = \frac{\partial}{\partial u_2} f_2, \qquad \frac{\partial}{\partial u_2} f_1 = -\frac{\partial}{\partial u_1} f_2. \tag{A.7}$$

Now, focus the situation in which $f(u) : \mathcal{A} \to \mathbb{R}$. Firstly, the conditions (A.7) show that $f(u)$ is analytical if and only if $f(u)$ is constant. Secondly, $\mathrm{d}f = 2\Re(\frac{\partial}{\partial u} f(u) \mathrm{d}u) = 2\Re(\frac{\partial}{\partial \bar{u}} f(u) \mathrm{d}\bar{u})$, which vanishes if and only if

$$\frac{\partial}{\partial u} f(u) = 0. \tag{A.8}$$

Hence, for finding the stationary points of a non-analytical function, the trick involves writing the differential in the form of (A.3) and set the term corresponding to $\frac{\partial}{\partial u} f(u)$ to zero.
In the multivariate case [71, 59], for the complex-valued function $f(u) : \mathcal{A} \subset \mathbb{C}$ with $\mathcal{A} \subset \mathbb{C}^r$,

$$\mathrm{d}f = \frac{\partial}{\partial u'} f(u)\, \mathrm{d}u + \frac{\partial}{\partial u^*} f(u)\, \mathrm{d}(\bar{u}), \tag{A.9}$$

where $u^* \equiv \bar{u}'$ is the conjugate transpose of $u$. It then easily follows that the differential $\mathrm{d}f$ of a real-valued function $f(u) : \mathcal{A} \to \mathbb{R}\, \forall u \in \mathcal{A} \subset \mathbb{C}^n$ vanishes if and only if the Wirtinger derivative is zero, i.e.,

$$\mathrm{d}f(u) = 0 \Leftrightarrow \frac{\partial}{\partial u} f(u) = 0. \tag{A.10}$$

# Appendix B

## B.1 Certain details of the EM Algorithm

To enable a smooth reading of the EM Algorithm developed in Section 3.5, certain details are let to reside separately. They are elucidated here:

### B.1.1 Log-likelihood as summation of logarithms

The following lemma is well-known; refer §16.5.4 of [30]:

**Lemma B.1.** *Suppose that $u_1, \ldots, u_m$ are points in the interval $\mathcal{U}$ and $c_1, \ldots, c_m \geq 0$ are such that $\sum_{l=1}^{m} c_l = 1$ and $f$ is a concave function in $\mathcal{U}$. According to* ***Jensen's inequality*** *$f(c_1u_1 + \cdots + c_mu_m) \geq c_1 f(u_1) + \cdots + c_m f(u_m)$.*

With $f \leftarrow \log_e$, $c_l \leftarrow g(x_l)$, and $u_l \leftarrow \mathsf{p}^{\mathsf{y},\mathsf{x}_l|\theta}(\mathcal{D}, x_l \mid \theta)/g(x_l)$, (3.32) is got.

### B.1.2 Decomposition of the complete log-likelihood

Using Theorem B.1, $\mathsf{p}^{\mathsf{y},\mathsf{x}|\theta}(\mathcal{D}, x \mid \theta) = \mathsf{p}^{\mathsf{y}|\theta}(\mathcal{D} \mid \theta)\, \mathsf{p}^{\mathsf{x}|\mathsf{y},\theta}(x \mid \mathcal{D}, \theta)$ is obtained. Hence, the right side of (3.32) may be factorized so that

$$\mathsf{L}(\theta, g) \geq \sum_{\mathsf{x}} g(x) \log_e \mathsf{p}^{\mathsf{y}|\theta}(\mathcal{D} \mid \theta) + \sum_{\mathsf{x}} g(x) \log_e \frac{\mathsf{p}^{\mathsf{x}|\mathsf{y},\theta}(x \mid \mathcal{D}, \theta)}{g(x)},$$

where the first term reduces to $\mathsf{L}(\theta)$ due to (3.5) and (3.30).

### B.1.3 Maximization of an expectation

If $\hat{g}_i$ of (3.35) is substituted in (3.32)

$$\mathsf{L}(\theta, \hat{g}_i) = \sum_{x} \mathsf{p}^{\mathsf{x}|\mathsf{y},\theta}(x \mid \mathcal{D}, \theta_i) \log_e \frac{\mathsf{p}^{\mathsf{y},\mathsf{x}|\theta}(\mathcal{D}, x \mid \theta)}{\mathsf{p}^{\mathsf{x}|\mathsf{y},\theta}(x \mid \mathcal{D}, \theta_i)},$$

where the denominator in the logarithm being independent of $\theta$ may be eliminated. As a result, (3.37) boils down to

$$\begin{aligned}\hat{\theta}_{i+1} &= \underset{\theta_i}{\text{argmax}}\ \mathsf{L}(\theta_i, \hat{g}_i) \\ &= \underset{\theta_i}{\text{argmax}} \sum_x \mathsf{p}^{\mathsf{x}|\mathsf{y},\theta}(x \mid \mathcal{D}, \theta_i) \log_e \mathsf{p}^{\mathsf{y},\mathsf{x}|\theta}(\mathcal{D}, x \mid \theta_i), \\ &= \underset{\theta_i}{\text{argmax}}\ \mathsf{E}^{\mathsf{x}|\mathsf{y},\theta}\big[\log_e \mathsf{p}^{\mathsf{y},\mathsf{x}|\theta}(\mathcal{D}, x \mid \theta_i)\big].\end{aligned}$$

## B.2 Posterior density with a Gaussian prior

Refer [113, 74] and §6.2 of [95] for the following theorem:

**Theorem B.1.** *According to the* ***Bayes theorem*** *for continuous probability density functions, the conditional distribution of a random variable* y *with any realization* $y$ *given a set of random variables* x *with any realization* $x$ *is related to the conditional distribution of* x *given* y *according to*

$$\mathsf{p}^{\mathsf{x}}(x)\mathsf{p}^{\mathsf{y}|\mathsf{x}}(y \mid x) = \mathsf{p}^{\mathsf{y}}(y)\,\mathsf{p}^{\mathsf{x}|\mathsf{y}}(x \mid y) \equiv \mathsf{p}^{\mathsf{y},\mathsf{x}}(y, x).$$

Due to Theorem B.1, $\mathsf{p}^{\mathsf{x}|\mathsf{y}}(x \mid y) = \frac{\mathsf{p}^{\mathsf{x}}(x)\mathsf{p}^{\mathsf{y}|\mathsf{x}}(y|x)}{\mathsf{p}^{\mathsf{y}}(y)}$; so, given the parameters $\theta$, it follows that $\mathsf{p}^{\mathsf{x}|\mathsf{y},\theta}(x \mid y, \theta) = \frac{\mathsf{p}^{\mathsf{x}|\theta}(x|\theta)\mathsf{p}^{\mathsf{y}|\mathsf{x}}(y|x,\theta)}{\mathsf{p}^{\mathsf{y}}(y|\theta)}$. While a Gaussian has been accepted for the denominator $\mathsf{p}^{\mathsf{y}}(y \mid \theta)$ according to (3.18), $\mathsf{p}^{\mathsf{y}|\mathsf{x}}(y \mid x, \theta)$ in the numerator is also a Gaussian as per (3.39). Assuming yet another Gaussian for

$$\mathsf{p}^{\mathsf{x}|\theta}(x \mid \theta) = \mathsf{p}^{\mathsf{x}}(x) = \mathcal{N}(x \mid 0, I_q).$$

Therefore,

$$\mathsf{p}^{\mathsf{x}|\mathsf{y},\theta}(x \mid y, \theta) = \frac{\mathcal{N}(x \mid 0, I_q)\mathcal{N}(y \mid Wx, \Gamma^{\mathsf{z}})}{\mathcal{N}(y \mid \mu^{\mathsf{y}}, \Gamma^{\mathsf{y}})}. \tag{B.1}$$

Suppose $c_1, \ldots, c_4$ are factors independent of $x$ such that

$$\begin{gathered}\mathcal{N}(x \mid 0, I_q) = c_1 \exp(-0.5\, x'x), \\ \mathcal{N}(y \mid Wx, \Gamma^{\mathsf{z}}) = c_2 \exp(-0.5\, x'W'(\Gamma^{\mathsf{z}})^{-1}Wx + x'W'(\Gamma^{\mathsf{z}})^{-1}y), \\ \mathcal{N}(y \mid \mu^{\mathsf{y}}, \Gamma^{\mathsf{y}}) = c_3, \\ c_4 = \frac{c_1 c_2}{c_3}.\end{gathered}$$

Then, (B.1) may be written as

$$\mathsf{p}^{\mathsf{x}|\mathsf{y},\theta}(x \mid y, \theta) = c_4 \exp(-0.5\, x'\Omega^{-1}x + x'\Omega^{-1}\Omega W'(\Gamma^{\mathsf{z}})^{-1}y),$$

where

$$\Omega^{-1} = I_q + W'(\Gamma^{\mathsf{z}})^{-1}W.$$

The probability density function of a Gaussian $\xi$ with mean $a$ and covariance matrix $B$ may be written as $\mathcal{N}(\xi \mid a, B) = c\exp(-0.5\,\xi' B^{-1}\xi + \xi' B^{-1}a)$, where $c$ is a factor independent of $\xi$. Thus, $\mathsf{p}^{\mathsf{x}|\mathsf{y},\theta}(x \mid y, \theta)$ is a Gaussian with mean $\Omega W'(\Gamma^{\mathsf{z}})^{-1}y$ and covariance matrix $\Omega$. It can be seen that

$$\mathsf{p}^{\mathsf{x}|\mathsf{y},\theta}(x \mid y, \theta) = \mathcal{N}(x \mid \Omega W'(\Gamma^{\mathsf{z}})^{-1}y, \Omega).$$

## B.3 Posterior density with a complex Gaussian prior

The extension of Section B.2 to complex Gaussian densities is straightforward. In that order of equations and interpretations therein, the following relations hold:

$$\mathsf{p}^{\mathsf{x}|\theta}(\mathbf{x} \mid \theta) = \mathcal{N}_{\mathbb{C}}(\mathbf{x} \mid 0, I_q)$$

$$\mathsf{p}^{\mathsf{x}|\mathsf{y},\theta}(\mathbf{x} \mid \mathbf{y}, \theta) = \frac{\mathcal{N}_{\mathbb{C}}(\mathbf{x} \mid 0, I_q)\mathcal{N}_{\mathbb{C}}(\mathbf{y} \mid \mathbf{W}\mathbf{x}, \mathcal{S}^{\mathsf{z}})}{\mathcal{N}_{\mathbb{C}}(\mathbf{y} \mid 0, \mathcal{S}^{\mathsf{y}})}. \tag{B.2}$$

Suppose $c_1, \ldots, c_4$ are factors independent of $\mathbf{x}$ such that

$$\mathcal{N}_{\mathbb{C}}(\mathbf{x} \mid 0, I_q) = c_1\exp(-\mathbf{x}^*\mathbf{x}),$$

$$\mathcal{N}_{\mathbb{C}}(\mathbf{y} \mid \mathbf{W}\mathbf{x}, \mathcal{S}^{\mathsf{z}}) = c_2\exp(-\mathbf{x}^*\mathbf{W}^*(\mathcal{S}^{\mathsf{z}})^{-1}\mathbf{W}\mathbf{x} + 2\Re(\mathbf{x}^*\mathbf{W}^*(\mathcal{S}^{\mathsf{z}})^{-1}\mathbf{y})),$$

$$\mathcal{N}_{\mathbb{C}}(\mathbf{y} \mid 0, \mathcal{S}^{\mathsf{y}}) = c_3,$$

$$c_4 = \frac{c_1 c_2}{c_3}.$$

Then, (B.2) may be written using

$$\mathbf{\Omega}^{-1} = I_q + \mathbf{W}^*(\mathcal{S}^{\mathsf{z}})^{-1}\mathbf{W}.$$

as

$$\mathsf{p}^{\mathsf{x}|\mathsf{y},\theta}(\mathbf{x} \mid \mathbf{y}, \theta) = c_4\exp(-\mathbf{x}^*\mathbf{\Omega}^{-1}\mathbf{x} + 2\,\Re(\mathbf{x}^*\mathbf{\Omega}^{-1}\mathbf{\Omega}\mathbf{W}^*(\mathcal{S}^{\mathsf{z}})^{-1}\mathbf{y}))$$

The probability density function of a complex Gaussian $\xi$ with mean $a$ and covariance matrix $B$ may be written as $\mathcal{N}_{\mathbb{C}}(\xi \mid a, B) = c\exp(-\xi^* B^{-1}\xi + 2\Re(\xi^* B^{-1}a))$, where $c$ consists of the normalization factor of the distribution independent of $\xi$. This shows that $\mathsf{p}^{\mathsf{x}|\mathsf{y},\theta}(\mathbf{x} \mid \mathbf{y}, \theta)$ above is a complex Gaussian with mean $\mathbf{\Omega}\mathbf{W}^*(\mathcal{S}^{\mathsf{z}})^{-1}\mathbf{y}$ and covariance matrix $\mathbf{\Omega}$, i.e.,

$$\mathsf{p}^{\mathsf{x}|\mathsf{y},\theta}(\mathbf{x} \mid \mathbf{y}, \theta) = \mathcal{N}_{\mathbb{C}}(\mathbf{x} \mid \mathbf{\Omega}\mathbf{W}^*(\mathcal{S}^{\mathsf{z}})^{-1}\mathbf{y}, \mathbf{\Omega}).$$