A novel statistical signal processing approach for analysing high volatile expression profiles

Zara Ghodsi

A thesis submitted in partial fulfilment of the requirements of Bournemouth University for the degree of *Doctor of Philosophy*

April 2017

I would like to dedicate this thesis to my loving parents for their everlasting encouragement, love and patience. I also dedicate this thesis to my fabulous nephew Sooshyant for bringing so much happiness, joy and love to my life.

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other University. This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration, except where specifically indicated in the text. This dissertation contains less than 65,000 words including appendices, bibliography, footnotes, tables and equations and has less than 150 figures.

Zara Ghodsi
Supervisors:
Dr. George Filis, Dr. Demetra Andreou and Dr. Kevin McGhee
2017

Acknowledgements

Foremost, I would like to express my sincere gratitude to my advisor Dr. Hossein Hassani, for the continuous support of my research, for his patience, motivation, enthusiasm, and immense knowledge. Also, I really enjoyed our wide-ranging conversations in which I learnt a lot more than just science. A special thank you to my sister, Dr. Mansi Ghodsi for her guidance, expertise and endless support and love.

I would like to acknowledge and note my appreciation to my supervisors: Dr. George Filis, Dr. Kevin McGhee and Dr. Demetra Andreou for their encouragement, insightful comments, and hard questions.

List of Publications

- Silva, E.S., Ghodsi, Z. and Hassani, H., 2016. Optimising Bicoid signal Extraction. Digital Signal Processing. Submitted.
- Ghodsi, Z., Hassani, H., 2016. Evaluating the Analytical Distribution of *bicoid* Gene Expression Profile. Briefings in Functional Genomics. Submitted.
- Ghodsi, Z., Huang, X. and Hassani, H., 2016. Causality analysis detects the regulatory role of maternal effect genes in the early Drosophila embryo. Genomics Data.
- Hassani, H., Ghodsi, Z., Silva, E.S. and Heravi, S., 2016. From nature to maths: Improving forecasting performance in subspace-based methods using genetics Colonial Theory. Digital Signal Processing, 51, pp.101-109.
- Alharbi, N., Ghodsi, Z. and Hassani, H., 2016. Noise correction in gene expression data: a new approach based on subspace method. Mathematical Methods in the Applied Sciences.
- Ghodsi, Z., Hassani, H. and McGhee, K., 2015. Mathematical approaches in studying bicoid gene. Quantitative Biology, 3(4), pp.182-192.
- Ghodsi, Z., Silva, E.S. and Hassani, H., 2015. Bicoid signal extraction with a selection of parametric and nonparametric signal processing techniques. Genomics, proteomics & bioinformatics, 13(3), pp.183-191.

- Hassani, H. and Ghodsi, Z., 2015. A glance at the applications of Singular Spectrum Analysis in gene expression data. Biomolecular detection and quantification, 4, pp.17-21.
- Hassani, H. and Ghodsi, Z., 2014. Pattern recognition of gene expression with singular spectrum analysis. Medical Sciences, 2(3), pp.127-139.

Abstract

The aim of this research is to introduce new advanced statistical methods for analysing gene expression profiles to consequently enhance our understanding of the spatial gradients of the proteins produced by genes in a gene regulatory network (GRN). To that end, this research has three main contributions.

In this thesis, the segmentation Network (SN) in *Drosophila melanogaster* and the *bicoid* gene (*bcd*) as the critical input of this network are targeted to study. The first contribution of this research is to introduce a new noise filtering and signal processing algorithm based on Singular Spectrum Analysis (SSA) for extracting the signal of bicoid gene. Using the proposed SSA algorithm which is based on the minimum variance estimator, the extraction of *bcd* signal from its noisy profile is considerably improved compared to the most widely accepted model, Synthesis Diffusion Degradation (SDD). The achieved results are evaluated via both simulation studies and empirical results.

Given the reliance of this research towards introducing an improved signal extraction approach, it is mandatory to compare the proposed method with the other well-known and widely used signal processing models. Therefore, the results are compared with a range of parametric and non-parametric signal processing methods. The conducted comparison study confirmed the outperformance of the SSA technique.

Having the superior performance of SSA, in the second contribution, the SSA signal extraction performance is optimised using several novel computational methods including window length and eigenvalue identification approaches, Sequential and Hybrid SSA and

SSA based on Colonial Theory. Each introduced method successfully improves a particular aspect of the SSA signal extraction procedure.

The third and final contribution of this research aims at extracting the regulatory role of the maternal effect genes in SN using a variety of causality detection techniques. The hybrid algorithm developed here successfully portrays the interactions which have been previously accredited via laboratory experiments and therefore, suggests a new analytical view to the GRNs.

Table of contents

List of figures

List of tables

Nomenclature

Acronyms / Abbreviations

$2D - SSA$	Two Dimensional SSA
ADF	Augmented Dickey Fuller test
AIC	Akaike Information Criterion
AP	Anterior Posterior axis
AR	Autoregressive
$ARFIMA$	Autoregressive Fractionally Integrated Moving Average
$ARIMA$	Autoregressive Integrated Moving Average
bcd	bicoid gene or mRNA
Bcd	Bicoid Protein
Bcd	Bicoid protein
CCM	Convergent Cross Mapping
CT	Colonial Theory
CV	Coefficient of Variation

D.melanogaster	*Drosopila melanogaster*
Dpp	Decapentaplegic gene
DV	Dorso Ventral axis
ETS	Exponential Smoothing
eve	even-skipped gene
FPE	Final Prediction Error Information Criterion
FRAP	Fluorescence Recovery After Photobleaching
FRET	Förster Resonance Energy Transfer
ftz	fushi-tarazu gene
GC	Granger Causality
GO	Gene Ontology
GRNs	Gene Regulatory Networks
h	hairy gene
HQ	Hannan Quinn Information Criterion
IDC	Intraerythrocytic Developmental Cycle
kni	knirps gene
Kr	kruppel gene
KS	Kolmogorov-Smirnov test
L	Window length

LS	Least Squares
MAPE	Mean Absolute Percentage Error
MSE	Mean Square Error
MSSA	Multivariate SSA
NN	Neural network
odd	odd-skipped gene
ODEs	Ordinary Differential Equations
PCA	principal component analysis
prd	paired gene
r	Number of eigenvalues
RMSE	Root Mean Squared Error
ROI	Regions of Interest
RRMSE	Ratio of Root Mean Squared Error
run	runt gene
SDD	Synthesis Diffusion Degradation model
Shh	Sonic hedgehog homolog gene
SIC	Schwarz Information Criterion
slp	sloppypaired gene
SN	Segmentation Network

SNR	Signal to Noise Ratio
$SSmethods$	subspace methods
SSA	Singular spectrum analysis
SSA$_{MV}$	SSA based on the minimum variance estimator
SVD	Singular Value Decomposition
SW	Shapiro-Wilk test
TFs	Transcription Factors
tll	tailless gene
VAR	Vector Auto-regression Model
w-correlation	Weighted Correlation

Chapter 1

Introduction

Most of our knowledge about the molecular basis of biological development has come from the magnificent studies on *Drosopila melanogaster* (*D. melanogaster*), also known as fruit fly.

Recent advances in methods and techniques that generate gene expression data have provided an excellent possibility to investigate the development of *D. melanogaster*.

D. melanogaster life cycle consists of several stages: embryogenesis, three larval stages, a pupal stage and the adult stage [Arbeitman et al., 2002]. Developmental studies are mostly focused on the principals underlying the genetic regulation mechanisms in the embryogenesis stage. According to the studies conducted in the field of system biology, the embryogenesis principles which are uncovered following the investigations on *D. melanogaster* may apply to other eukaryotes as well.

During the embryogenesis stage of *D. melanogaster*, several Gene Regulatory Networks (GRNs) act in the embryo which among them Segmentation Network (SN) is a well known and mostly studied GRN. Such studies are aimed at explaining and characterising the complex relationships between primary developmental transcription factors (TFs) and their target genes. The TFs control the amount and appearance timing of the proteins produced by the genes in a GRN.

Therefore, the insight gained by studying the gene expression profiles of the SN as a fundamental network at the beginning of cell differentiation process in *D. melanogaster* has considerably enhanced our knowledge of transcriptional mechanisms and gene-gene interactions.

Undoubtedly, given the big size, noisy feature and complicated characteristics of gene expression data sets, a computationally efficient and statistically robust analysing method plays a central role to reinforce the validity and reliability of new findings in this area. Hence, this research introduces a number of novel statistical methods to improve the computational analysis of cell differentiation process during the segmentation stage. The presented methods play a significant role in improving the precision of gene expression signal processing and network analysing techniques. Furthermore, the new methods proposed in this research introduce novel approaches for modelling the segmentation gene expression profiles which are present in the SN.

It is imperative to note that the theoretical contribution presented in this thesis results not only in the improvement of the accuracy of gene expression studies, but it is also applicable to other fields of applied statistics where signal extraction, noise filtering, forecasting and causality detection goals are aimed to study.

1.1 Motivation

At the beginning of life, there were only single cells. Today, after many millions of years, most plants, animals, fungus, and algae consist of multiple cells that work together as an individual being. Accordingly, the questions of how a simple fertilised cell develops into a complex multicellular organism and how different cells cooperate to carry out specific tasks and perform particular roles have attracted the interest of many researchers during the last century.

Being very much fascinated by the questions above, in this thesis, we focused our research on a particular aspect of these subjects. Accordingly, we introduce and integrate several computational techniques and develop a new analytical view to study the gene expression profiles and to analyse the GRNs. The principal aim of our study is to improve the noise filtering algorithms applied for extracting the signal of gene expression data set and to give a statistical representation of gene expression profiles. To elucidate the important contribution of a useful gene expression signal processing approach to address the questions mentioned above, an overview of the phenomenon lied in those questions is given below.

The process of developing a simple zygote to a complex system of tissues and cell types is known as cell differentiation [Van den Berg et al., 1997]. Like any other scientific question, to investigate this phenomenon and to acquire new details, a scientific method needs to be adopted.

Understanding the genetic process of cell differentiation demands to apply several advanced methodologies and tools as well as a broad knowledge of the other areas of study such as chemistry, physics and mathematics. More importantly, there is a great amount of studies focused on this phenomenon over the past decade. Therefore, to effectively improve the knowledge in this area, a research study should be narrowed down to one particular aspect of cell differentiation. As it has been famously quoted, a good research question is not too broad nor too narrow.

According to the extensive studies carried out during the last century, now, it is widely believed that cellular differentiation leading to the development of a multicellular organism is a response to the distinct spatial order of morphogen gradients [Rivera-Pomar and Jãckle, 1996; Scott et al., 1987]. Hence, morphogens, their gradients and interactions have been targeted for further exploration by developmental biologists. A summary of these studies is presented later Chapter 2.

Having considered the morphogens as the substance under study, a possible way to facilitate the research is the use of a model organism. Besides the fact that studies on certain characteristics of morphogens initially started with studying *D. melanogaster*, the main reason behind applying a model organism in genetic studies is an evolutionary principle stating that due to the common ancestry fact, all organisms have some degree of relatedness and genetic similarity. For example, several basic biological, physiological, and neurological characteristics are conserved between mammals and *D. melanogaster*, and approximately 75% of human disease-causing genes are believed to have a functional homolog in this fly [Pandey and Nichols, 2011].

Being easy to control and breed in a laboratory setting, simple genetic accessibility and short generation time can be addressed as the other important features which credit applying *D. melanogaster* in genetic studies.

Given these factors, *D. melanogaster* was chosen to be an excellent candidate for screening the function of morphogenes process. In addition to the reasons mentioned above, the small size of this fly, transparent embryo and large embryo size can be addressed as the other key factors which further encouraged scientists to adopt *D. melanogaster* to investigate the characteristics of the morphogens [Powell, 1997].

Having considered *D. melanogaster* as a very useful model organism, Bicoid (Bcd) [1] , the first identified protein to act as a morphogen was discovered in 1988. Therefore, the *bcd* gene in *D.melanogaster* soon became a hot research spot for developmental researches.

Bcd is a homeodomain transcriptional factor whose protein concentration gradient plays a crucial role in patterning along the future head to tail axis of an adult fruit fly and is a key maternal input of the SN. As previously mentioned, SN is one of the first developmental gene networks in *D. melanogaster* which establishes only within several hours after fertilisation [Rivera-Pomar and Jãckle, 1996; Scott et al., 1987]. It is widely accepted that the

[1] In what follows, the italic lower-case *bcd* represents either the gene or the mRNA and Bcd refers to the protein. This presentation is applied for all the genes mentioned in this thesis.

SN is responsible for forming the segmented pattern of *D. melanogaster* body along the anterior–posterior (AP) axis , Fig 1.1 [Surkova et al., 2008b].

Over the years, the extensive studies on different aspects of pattern formation and segment determination processes in SN have enhanced our understanding of how genes cooperate to pattern the body after fertilisation.

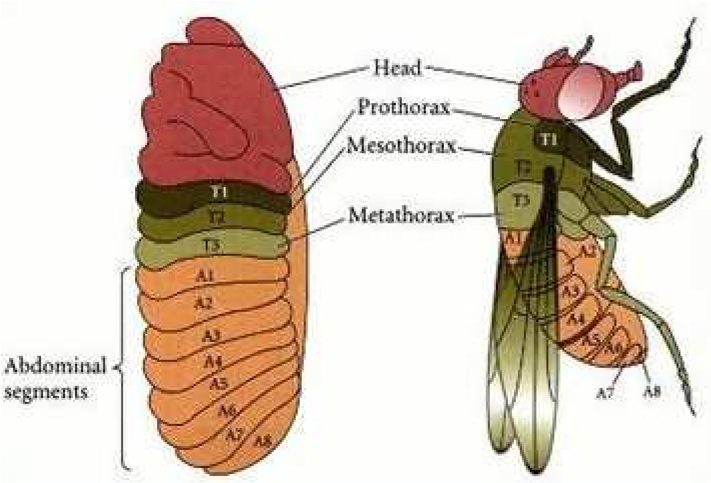


Fig. 1.1 The segmented body of *D. melanogaster*. Figure adopted from [Alberts et al., 2002]

Now we know that the genetic cascade in the SN comprised of maternal, gap, pair-rule and segment polarity genes and the sequential activation in this network starts with the expression of *bcd* gene [Zamparo and Perkins, 2009].

Accordingly, the SN and in particular the *bcd* gene as the principal input of this network were considered as the premier system for system biology studies. Hence, over many years of study, different experimental datasets have been coupled to various computational models to portray the biological processes of this network. [Hengenius et al., 2011; Jaeger et al., 2004; Papatsenko and Levine, 2011; Poustelnikova et al., 2004; Reinitz and Sharp, 1995].

It is evident that the prediction reliability and success of the proposed models greatly depend on the availability of the methods and techniques which can generate improved and accurate quantitative gene expression data at cellular resolution. To extract the quantitative data set from cell-specific gene expression, immunofluorescence technique using confocal scanning microscopy of fluorescently tagged molecules is widely applied. This technique is a popular method which among different applicable approaches has recently transformed itself into a practical toolkit for developmental biologists [Zamparo and Perkins, 2009].

Immunofluorescence method is based on the interactions created between antibodies labelled with fluorescent dyes and the appropriate antigens [Minsky, 1988] (Fig 1.2). The fluorescent signals are then detected by fluorescent microscopy using laser and confocal scanning microscopes. The achieved signals can accordingly provide great insight into the cellular world by visualising the distribution and the concentration of the proteins under study [Surkova et al., 2008b]. It should be noted that although in studying a single-cell system it is possible to perform an *in vivo* screening, in multicellular organisms, the study is usually carried out on fixed tissues [Wu et al., 2007].

There are two classes of immunofluorescence technique:

- **Primary (direct)**: Which applies a single, primary antibody. This antibody is chemically attached to a fluorochrome. The primary antibody binds to the target molecule (i.e. antigen). A fluorescent microscope then identifies the attached fluorochrome. Although the direct Immunofluorescence method is less common than the indirect one, it has several powerful advantages including the reduced number of analytical steps, quickness and less non-specific background signals [Fritschy and Härtig, 2001].

- **Secondary (indirect)**: Which applies two antibodies. The primary antibody from one side binds the antigen while from the other side binds the secondary antibody which is attached to the fluorochrome. In this case, the signal is amplified considerably by multiplying the number of fluorochrome per antigens. Nevertheless, this method is rather time-consuming, it is more flexible because for any given primary antibody, different secondary antibodies can be utilised, [Fritschy and Härtig, 2001], (Fig 1.3).

Later in this chapter, we will see that the gene expression data used in this study is achieved via an indirect immunofluorescence technique.

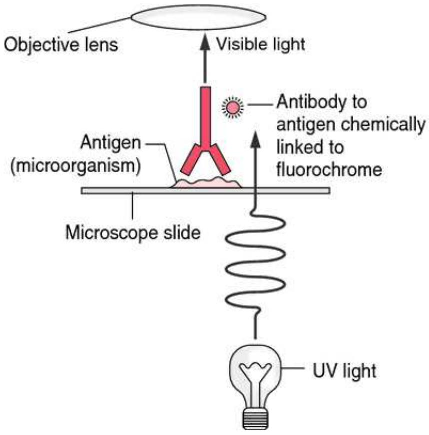


Fig. 1.2 Direct immunofluorescence. In this method the object is visualised using a fluorescence-tagged antibody. Figure adapted from [Dubreux, 1998].

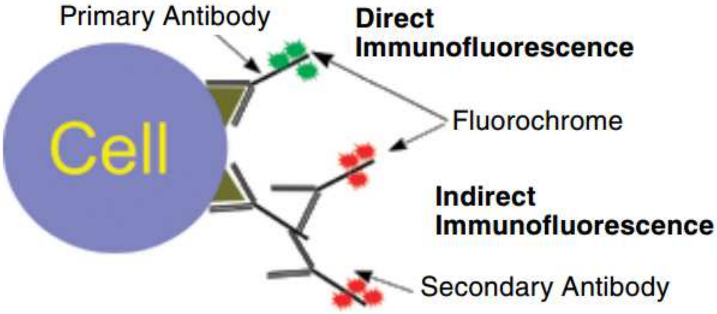


Fig. 1.3 The direct immunofluorescence techniques versus the indirect immunofluorescence.

The data achieved by means of immunofluorescence technique can be applied in a wide variety of studies (e.g. Studying GRNs, determining co-expressed genes or co-located gene products, sub cellular localisation of proteins, determining the effects of gene knockouts, over/under gene expression, promoter manipulations, inferring protein functions determining the concentration of specific cellular ions) [Zamparo and Perkins, 2009].

Immunofluorescence technique can quantify both the spatial and the temporal measurements of the fluorescent molecules in a biological specimen. In the spatial measurement, the information on features such as distance, areas and velocity are aimed to achieve as opposed to the temporal analyses where the intensity levels attributed to the fluorochromes are measured at different time points [Ji et al., 2009].

Moreover, depending on the aim of the study, different techniques can be applied in conjunction with immunofluorescence technique. (e.g. Comparing the relative intensities of two fluorescent specimens, fluorescence recovery after photobleaching (FRAP) and Förster resonance energy transfer (FRET) [Waters, 2009].)

Without any doubt, the accuracy and precision of the data achieved by immunofluorescence technique play a significant role in the success of the computational models proposed for gene expression profiles. Therefore, it is of critical importance to note that regardless of the general advantages of immunofluorescence technique, like any other computational procedure, this method has its limitations which have been previously raised by several studies [van der Loos, 2008]. According to [Wu et al., 2007], the existence of considerable amounts of noise can be addressed as one of the most important limitations of this technique.

Confocal scanning microscopes provide digital images of the fluorescence signals which are used for numerical measurements. To this aim, the detected photons at each pixel of a digital image are converted to an intensity value. It should be noted that this value is only an estimation of the correlation between the number of the photons and the intensity level and is not exactly equal to the number of detected photons [Pawley, 2006].

However, this is not the only flaw of the numerical measurements of immunofluorescence technique (the outline of these issues is addressed later in this chapter). Therefore, to increase the accuracy and precision of the measurements, this technique has been notably improved over the past decade. For example, the advanced digital cameras and genetically encoded fluorochromes have led to a considerable enhancement of the resolution and contrast of the images.

Since the last decade, the analysis of noise in achieved gene expression profiles has attracted the attention of researchers and theoreticians from different fields of study [Golyandina et al., 2012]. The main focus of such studies has been to uncover the source of the errors introduced to the data and to evaluate the possible ways to overcome this issue [Myasnikova et al., 2009]. According to these studies, noise in quantitative measurements might arise from the specimen, the microscope, or the detector [Pawley, 2006].

However, noise in the gene expression data enters not only from the data acquisition and processing procedures [Wu et al., 2007] but also the fluctuations seen in an expression pattern can be a consequence of biological noise which may also introduce error into the data [Myasnikova et al., 2009]. Therefore, the source of the natural biological variability is different from the system noise [Myasnikova et al., 2009]. In Chapter 5, we will show that by adopting a new hybrid signal processing algorithm developed in this thesis, different sources of error can be effectively discriminated and modelled in analysing gene expression profiles.

Accordingly, the contribution of error introduced to a gene expression profile which causes variance in intensity can be categorised as follows:

- **Biological noise**

  arises from the active molecular transport, compartmentalisation, and the mechanics of cell division [Spirov et al., 2012]. Biological noise is categorised in two different groups of intrinsic and extrinsic noises [Fedoroff and Fontana, 2002].

Intrinsic noise is related to the inherent randomness of biochemical processes such as transcription and translation and is highly dynamic in time [Longo and Hasty, 2006]. Intrinsic noise arises from the stochastic nature of the biochemical process of gene expression and causes identical copies of a gene to express at different levels [Golyandina et al., 2012]. According to [Wu et al., 2007], this source of error is the major contribution of fluctuations.

Extrinsic noise arises from the fluctuations in cellular components such as regulatory proteins and polymerases that indirectly cause variation in the expression pattern of a gene [Longo and Hasty, 2006].

- **Observational noise**

  This source of error is attributed to the fluorescence measurement uncertainty in the acquisition of quantitative gene expression data which might be due to several reasons:

  Errors introduced by confocal microscope [Myasnikova et al., 2009], chemical causes such as fluctuations in the number of primary and secondary antibody molecules binding to proteins which according to [Wu et al., 2007] can explain the multiplicative observed noise, possible errors in instrument functionality, sample preparation and mathematical treatment of data [Myasnikova et al., 2009], the quantification of blurred images [Myasnikova et al., 2009], errors due to over- and under-saturation, errors due to image blurring [Myasnikova et al., 2009], background staining, non-accurate embryo orientation, image and data processing errors. Evidently, all these procedures add further errors in data [Myasnikova et al., 2009].

In affecting the precision and accuracy of the obtained images and quantitative measurements, the structural damage or pessimal immunolabeling produced by poor sample preparation should not be underestimated [North, 2006]. “Garbage in = garbage out” can be addressed as the best description for the noise introduced after a poor sample preparation step.

As noted in [Verveer et al., 1999], inaccurate accumulation of fluorescent tags on organelles also creates a signal bias.

It is of note that in some projects repeated measurement is addressed to be an effective way to reduce the noise. However, in some studies like a one-time point measurement in a live-cell time-lapse experiment, the analysis is often bounded to make only one measurement.

Accordingly, one of the challenging issues in analysing gene expression profiles is the association of the data with both biological and observational noise, which limits the performance of the models and techniques, and consequently affects the accuracy of results [Gregor et al., 2007; Hilfinger and Paulsson, 2011]. Therefore, employing an effective method to deal with noisy profiles is essential.

Since in this thesis, our main focus is on analysing the *bcd* gene, from now we consider the Bcd profile as the gene expression profile under study. In Chapter 5, we will show how the outlined principles are adjustable and applicable to the other gene expression profiles of SN.

There are two main approaches for fitting a model to a noisy Bcd profile:

- Ignoring the presence of noise and directly fit a model to the noisy Bcd data (A model like the Synthesis Diffusion Degradation (SDD) model).

- Reducing the level of noise in Bcd profile by using a filtering method and then fit a model to the filtered data [Hassani et al., 2009b].

The second approach increases the precision of the study by reducing the noise to the lowest level possible and therefore, it is believed to be more efficient than the first approach. As would expect, the selection of an appropriate filtering method is crucial for the effectiveness of this approach.

In adopting a standard filtering method, assuming the stationarity of data, linearity of the model and normality of residuals are essential [Sanei and Hassani, 2015]. However, a method with these assumptions provides only an approximation of the actual situation [Hassani et al., 2013c]. Therefore, a method that does not depend on the mentioned assumptions can be very useful for modelling and filtering not only the Bcd data but the other segmentation gene expression profiles.

Singular spectrum analysis (SSA) is an excellent time series analysing and noise filtering technique. SSA is a nonparametric method and therefore does not require any assumptions [Hassani et al., 2010]. Accordingly, These features of SSA technique have motivated us to consider this method as the foremost analytical tool of interest for this research. The reason behind this selection and the main contributions and characteristics of SSA technique are further elaborated later in Chapter 2.

However, before going to the details of SSA, it is necessary to point out the main objectives targeted in this research.

1.2 Research Aim and Objectives

The main aim of this research is to introduce an improved statistical approach for signal processing of gene expression profiles and network analysing in SN. The emphasis is placed on providing an analytical tool for signal extraction and noise filtering process. To that end, this research has four main objectives.

- To introduce SSA based on minimum variance estimator as a new and enhanced signal extraction method for gene expression profiles. To compare and evaluate the reliability of different parametric and nonparametric signal processing techniques with SSA based on minimum variance estimator as the new introduced method and an excellent candidate of non parametric techniques.

- To improve the performance of the signal processing procedure by optimising specialising and automating the introduced noise filtering algorithm and expand the application to different segmentation genes of SN.
- Evaluate the applicability of a hybrid analytical tool comprised of SSA and a variety of causality detection techniques as a novel analytical approach for studying the SN and GRNs in general.

Accordingly, this research is organised in such direction to achieve the mentioned objectives. The rational, methodology and results of each objective are further expanded and elaborated in sufficient detail under separate chapters.

As it will be further discussed the realisation of these objectives will indeed result in considerable theoretical advancements for SSA and also the field of signal processing and time series analysis.

The conclusions, limitations and pathways for future research are also presented in the final chapter of this thesis.

1.3 Why Singular Spectrum Analysis?

Having the research aims and objectives specifically identified above, in this section, the motivation for the selection of SSA technique is further elaborated.

Accordingly, different characteristics which motivated the choose of SSA as the principal method of interest for this research are described below.

After the introduction of SSA as a time series analysis technique in 1986, this method transformed itself into a practical tool for analysing biomedical data and recently it has also been applied to genetic studies where it illustrated its strong potential for such studies [Du et al., 2008; Tang and Yan, 2012]. Noise filtering, smoothing, and forecasting a given time series are among the most important applications of SSA technique. Since only the signal processing capabilities of SSA including smoothing and filtering are beneficial for this research, Chapter 3, provides the details on these features of the method and therefore the forecasting aspect is not discussed in this thesis.

As already mentioned, SSA is a nonparametric technique which does not rely on the assumptions of normality of the residuals, stationarity of the data, or linearity of the model [Hassani et al., 2013a] which are highly unlikely to hold in the real world applications. As such, applying SSA enables modelling a given data set with a true approximation of the situation without losing any information [Hassani and Mahmoudvand, 2013].

It is also worthy to ask if the SSA technique is a powerful method for time series analysis how is it possible to adopt this method for a one-time point measured gene expression profile? To answer this question, it should be noted that as we will discuss later in this chapter, the gene expression can be traced either in time or space. The data points used in this study are the intensity levels attributed to the nuclei along the AP axis which can be considered as a sequence series. Therefore, SSA with all of its capabilities can be applied to a sequence exactly the same way as it is implemented on a time series.

Since the introduction of SSA, the power of this technique has been illustrated by several studies. Having said that, it should be noted that SSA is a relatively young technique and therefore there is still a huge scope for further improving and enhancing this method as a viable and efficient tool for modelling and processing different types of data sets.

1.4 Database

without any doubt, availability of a database which provides researchers with access to up-to-date and accurate quantitative data specified to *D. melanogaster* plays a significant role in developing new models to analyse the Bcd gradient and to infer the regulatory links between the segmentation genes in the SN.

There are several databases which provide comprehensive information on genomics and developmental processes in *D. melanogaster* including FlyEx http://urchin.spbcas.ru/flyex/, FlyBase http://flybase.org/, Flymove http://flymove.uni-muenster.de/, BDGP http://www.fruitfly.org/, FlyTF http://www.flytf.org/, FlyMine http://www.flymine.org/, FlyAtlas http://flyatlas.org/atlas.cgi and FlyCircuit http://www.flycircuit.tw/. Each database mentioned above, is specifically designed to provide essential information for studies focused on a particular feature of *D. melanogaster* and other related species.

FlyEx is the most popular database in providing quantitative data on SN and is the main database used in this research. Therefore, in what follows the data characteristics and the methodology applied to achieve FlyEx quantitative data are exploited. Thereafter, we also briefly touch upon the other databases which can provide powerful insight into *D. melanogaster* and in parallel to FlyEx can be successfully adopted for the future studies.

1.4.1 FlyEx

FlyEx is a large database for the expression profiles of *D. melanogaster* segmentation genes in cleavage cycles 10-–13 and all temporal classes of cycle 14A. In total, FlyEx contains 4716 images of 14 segmentation genes. The images are obtained from 1580 wild-type (OregonR) *D. melanogaster* embryos immunostained for segmentation proteins which provide 9500000 quantitative data records.

FlyEx is designed as a spatiotemporal atlas of fourteen segmentation genes listed as follows:

- The maternal coordinate genes: bicoid (*bcd*) and caudal (*cad*).
- The gap genes: Kruppel (*kr*), knirps (*kni*), giant, hunchback and tailless (*tll*).
- The pair-rule genes: even-skipped (*eve*), fushi-tarazu (*ftz*), hairy (*h*) , runt (*run*), odd-skipped (*odd*), paired (*prd*) and sloppypaired (*slp*) .

A short description of the methodology used to gather the quantitative data in FlyEx is given below. Since in SN, our principal focus is on the *bcd* gene, the data-acquisition procedure is specifically outlined for this gene and in doing so, we mainly follow [Pisarev et al., 2009; Poustelnikova et al., 2004] where a more detailed description is made available. The most steps of this procedure are similar to what is carried on to obtain the expression profile for the other segmentation genes.

Step 1: Confocal imaging

Confocal scanning microscopy provides images of gene expression patterns. As described in [Kosman et al., 1998], gene expression is measured by using fluorescence-tagged antibodies technique. The final image of a single-gene expression pattern in an embryo is achieved by averaging, cropping and rotating the obtained image.

Step 2: Image segmentation

In this step to improve the accuracy, the size differences among the various embryos are normalised in size.

At a given time point, each embryo is scanned for the expression of three segmentation genes. This process is followed by combining three embryo images. The final image is then segmented to make a binary nuclear mask. The advantage of using the binary nuclear mask is determining the average x and y related to each nucleus to estimate the average fluorescence level of each three gene expression. A detailed explanation of the nuclear mask characteristics can be found in [Janssens et al., 2005].

To improve the quality and to characterise the achieved quantitative data for each single embryo, the following two steps are applied:

Step 3: Data normalization

In this step, the background noise is minimised by re-scaling the *bcd* expression data by fitting an exponential curve [Myasnikova et al., 2005]. The exponential curve used here is derived from the SDD model which described before. The improvements made in this step following the studies provided in this thesis are reported in details in Chapters 4 and 5.

Step 4: Temporal characterization

To achieve a more reliable estimation of the quantitative data, determining the developmental age of each embryo and temporal dynamics of gene expression data is of critical importance. To this end, for the embryos prior to cycle 14A the number of nuclei is computed. Since cleavage cycle 14A is about 50 min long, in addition to the nuclei number computing, other morphological markers such as the blastoderm morphology and the membrane/cortex ratio are also examined.

Step 5: Registration

In this step, the small size variation among the embryos within a specific temporal class is removed. As noted in [Kozlov et al., 2002; Myasnikova et al., 2001] there are two registration methods: the spline and or wavelet method. The central difference between these two methods lies in the procedure applied to extract the ground control points. The data obtained in this step is called the integrated data.

1.4.2 FlyBase

FlyBase is a large informative database for *D. melanogaster* fundamental genomics information. In this database, the three attributes of wild-type gene products including the molecular function, biological processes and subcellular location are provided by adopting the Gene Ontology (GO) terms.

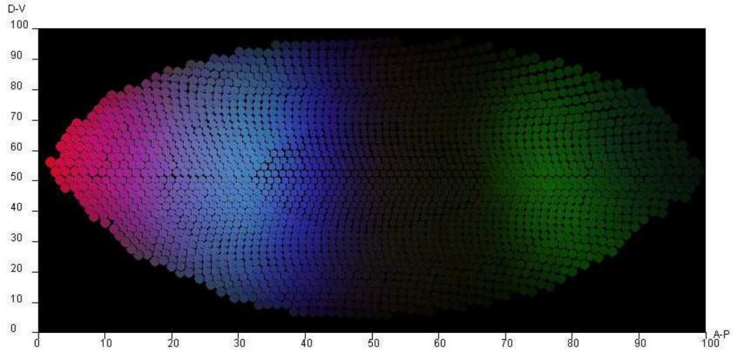


Fig. 1.4 An example of a gene expression image from Flyex. Different colors of pink, blue and green are respectively related to Bcd , Hb and Gt. Figure adapted from http://urchin.spbcas.ru/flyex/.

Moreover, the FlyBase database has the responsibility of recurrent or re-evaluated annotation of the *D. melanogaster* genome. This database has provided valuable insights into the genome of *D. melanogaster*. The facilities available at this platform are helpful for users with different levels of knowledge. There is also a wide variety of modes of access to download the files.

1.4.3 FlyMove

FlyMove is a helpful resource for studying the developmental process of *D. melanogaster*. This database provides users with a variety of multimedia presentations such as images, movies and interactive shockwaves, which effectively facilitate the understanding of the developing stages in this fly. All the media provided by FlyMove is free to download.

1.5 The Bcd Data

Real data

The activity of *bcd* gene can be examined either by studying its mRNA profile or its gene expression profile. A gene expression profile is a concentration measurement of the proteins produced by that particular gene. This profile is achieved by fluorescence imaging technique

[Kosman et al., 1998]. The applied antibody allows the visualisation of the Bcd proteins. Such quantification relies on the assumption that the actual protein concentration detected by the antibodies and the measured fluorescence intensity are linearly related.

In this thesis, the quantitative *bcd* gene expression data in wild-type *D. melanogaster* embryos is obtained from FlyEx database [Pisarev et al., 2009]. This database is widely used as a valuable source of information for studying the dynamics of segment determination in the early *D. melnogaster* development [Poustelnikova et al., 2004].

Data acquisition in this dataset is based on the confocal scanning microscopy of fixed embryos immunostained for segmentation proteins [Pisarev et al., 2009]. To that aim, a 1024 × 1024 pixel confocal image with 8 bits of fluorescence data is achieved for each embryo which then transformed into an ASCII table.

The ASCII table contains the fluorescence intensity levels attributed to each nucleus in the %10 of longitudinal strips (i.e. only the nuclei correspondents to the central 45—55% of the DV axis)along in the AP direction) and is unprocessed for any noise reduction methods. Fig 1.5 shows the %10 longitudinal strips of a confocal image [Holloway et al., 2006]. It is of note that Ap and DV axis make the first and the second dimension of the data respectively.

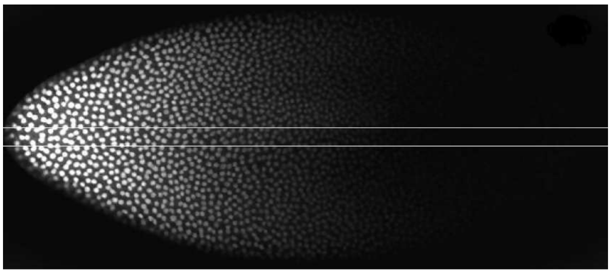


Fig. 1.5 Confocal image of an embryo related to time class 14(1). White horizontal lines depict the 10% strip utilised to collect data. Figure adapted from [Surkova et al., 2008a].

Since the segment determination starts from cleavage cycle 10 and lasts until the end of cleavage cycle 14A (when proteins synthesised from maternal transcripts begin to appear up to the onset of gastrulation) the data is categorised to five main cycles of 10 to 14A. Additionally, as the cleavage cycle 14A is considerably longer in time, to facilitate the analysis, temporal classes 1 to 8 are considered as the subgroups of this cleavage cycle. Hence, all the embryos are classified into the following time classes: 10, 11, 12, 13, 14(1), 14(2), 14(3), 14(4), 14(5), 14(6), 14(7) and 14(8). It should be noted that each class of data contains a different number of embryos.

Fig 1.6 shows a typical example of Bcd gradient along the embryo at cleavage cycle 14(3), the effect of noise (i.e., fluctuations visible in Fig 1.6) in this gradient can be seen as the high volatile pattern.

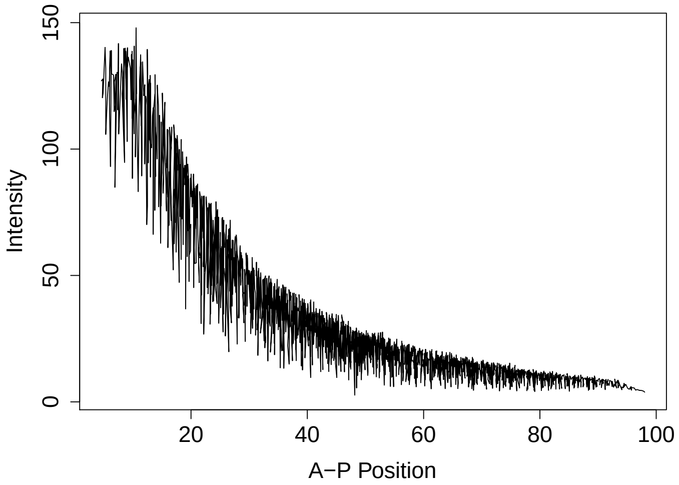
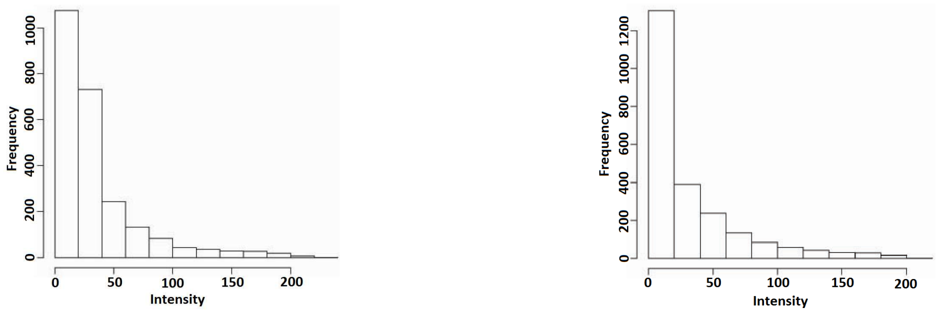


Fig. 1.6 A typical example of Bcd gene expression profile. Y-axis shows the fluorescence intensities and X-axis shows the position along the A-P axis of the embryo.

Table 1.1 presents the descriptive statistics of Bcd. As it can be seen, each time class has a dissimilar number of embryos and the length of the series obtained for each embryo is different where a large series length indicates that a greater number of nuclei was presenting the fluorescence intensity. In other words, Bcd protein molecules were produced in a higher number of nuclei along the AP axis. For each time class, the average of variance, series length, mean, skewnwss and kurtosis are presented separately. Due to the considerable variation present in the data, the median is chosen as a measure of central tendency.

In Table 1.1, the fourth column shows the variation of each profile within a time class. For example, in time class 10, the minimum variance is 928.8, however, the maximum variance for this cycle is more than 2000. Hence, here, we deal with two kinds of variation; within a cycle and between a cycle variation. The skewness was also tested, and the results confirm that there is a statistically significant skewness (at 5% level) indicating that almost all the series have values towards the lower end of the series. In this table, Min, Median and Maximum of variance, length size, mean, median, minimum, maximum, skewness and kurtosis have been presented separately for each time class.

Fig 1.7 shows the histogram of two Bcd profiles [2]. As it is apparent from this figure, Bcd profile is asymmetric and notably right-skewed suggesting that it may be poorly characterised by its mean and variance.



(a) Time Class 14(3)

(b) Time Class 14(4)

Fig. 1.7 The histogram of Bcd related to time class 14(3) and 14(4).

1.5.1 Simulated Data

In this thesis, a series of simulated data is used to evaluate the performance of different models adopted to study. The baseline of the simulation procedure applied in this thesis is presented below. Some adjustments in this process are made upon the aim of the study of

[2] Histograms of cleavage cycles 10-13 and all temporal classes of cleavage cycle 14A are presented in Appendix 1.

Time class	N		Var.	length	Mean	Med.	Min.	Max.	Skew.	Kurt.
10	7	Min	928.8	79.00	16.95	2.970	0	134.4	1.09	0.33
		Med	1294	124.0	46.33	37.55	7.420	163.9	1.550	1.950
		Max	2358	146.0	70.83	54.68	20.95	209.5	1.670	4.020
11	14	Min	693.5	238.0	34.41	21.39	3.67	152.91	1.190	0.300
		Med	1780	288.5	46.30	26.92	6.160	185.8	1.460	1.120
		Max	2999	308.0	77.07	63.64	20.96	223.2	1.860	2.980
12	31	Min	1160	394.0	35.15	17.63	1.570	147.20	0.66	-1.27
		Med	2422	524.0	51.09	28.41	7.110	206.4	1.420	0.980
		Max	5224	607.0	165.5	174.6	87.62	239.7	1.860	2.850
13	98	Min	412.6	738.0	21.97	13.05	0	131.0	0.660	-0.32
		Med	1578	1276	42.08	26.21	4.740	197.8	1.810	2.701
		Max	2795	1550	77.87	65.39	16.14	240.5	2.640	7.430
14(1)	58	Min	705.4	1085	24.07	12.00	0	143.7	1	-0.15
		Med	2041	2257	43.25	24.73	4.110	223.67	1.890	2.880
		Max	2968	2548	148.66	131.6	67.19	252.9	2.600	6.840
14(2)	30	Min	1344	2043	32.18	16.54	0	190.9	1.680	2.250
		Med	1921	2315	42.93	25.16	4.430	225.6	1.690	3.400
		Max	1887	2678	51.64	36.20	11.84	245.63	2.710	7.240
14(3)	38	Min	480.8	1642	17.62	6.460	0	147.0	0.680	-0.66
		Med	1490	2280	40.30	23.19	6.070	216.02	2.18	4.48
		Max	2654	2783	138.9	125.9	56.07	252.7	2.560	6.540
14(4)	28	Min	697.7	1741	33.49	16.39	0	170.8	1.390	1.110
		Med	1578	2275	42.17	25.88	7	212.1	1.990	3.510
		Max	2324	2520	55.33	42.64	13.91	234.6	2.250	5.110
14(5)	25	Min	439.6	1707	22.87	13.69	0	113.0	0.520	-0.90
		Med	1195	2297	38.17	23.99	4.400	192.64	2.020	3.850
		Max	2263	2453	154.0	137.7	71.08	236.60	2.270	5.450
14(6)	29	Min	84.37	1583	27.85	15.64	0	134.20	0.980	0.820
		Med	1131	2266	39.26	25.81	7.620	194.3	1.920	3.300
		Max	2057	2584	93.40	83.62	40.43	235.2	2.460	6.700
14(7)	15	Min	141.0	1535	17.48	12.70	0	81.58	0.670	0.090
		Med	443.5	2109	40.00	34.94	8.140	133.6	1.670	2.110
		Max	18060	2423	108.7	101.3	52.56	220.6	2.460	6.700
14(8)	12	Min	397.9	1245	26.05	14.79	2.170	133.6	0.630	-0.36
		Med	636.2	1521	64.68	56.00	21.55	170.0	1.470	2.120
		Max	818.0	2195	134.1	128.9	80.26	202.0	2.060	5.130

Note: N= Number of embryos studied per time class, Var.= Variance in each profile, Length= Length of data in each expression profile, Mean=The average of intensity levels, Med.= Median of intensity levels, Min.= The minimum value of intensity levels, Max.= The maximum value of intensity levels, Skew. =Skewness, and Kurt.= Kurtosis.

Table 1.1 Descriptive statistics of Bcd profile.

each chapter where in that case the modifications are presented in details under the simulation section of that chapter.

To start the simulation, an exponential curve is drawn from the simple SDD model which is a commonly used model for studying the Bcd profile [Bergmann et al., 2007; Driever and Nüsslein-Volhard, 1988; Gregor et al., 2005; Houchmandzadeh et al., 2002]. This curve is then considered as the benchmark for comparison studies. Recall that the concentration of Bcd in SDD model follows:

$$B = Ae^{-\frac{x}{\lambda}}, \tag{1.1}$$

where A is the amplitude, x is the distance from the anterior [Driever and Nüsslein-Volhard, 1988], and λ is the length parameter obtained by fitting an exponential model to the Bcd intensity profile and computing the position at which the concentration has dropped to $1/exp$ of the maximal value at the anterior (at $x = 0$) [Bergmann et al., 2007; Driever and Nüsslein-Volhard, 1988; Gregor et al., 2005; Houchmandzadeh et al., 2002]. The initial value of these parameters obtained by following the study conducted by Gregor et al. [Gregor et al., 2005].

To obtain a noisy series as close as possible to the real one, random errors ε of a normal distribution with zero mean and variance σ_ε^2 with different amplitudes which are an estimation from the real dataset, are added to various parts of this curve. This simulation is repeated $1,000$ times.

From now, in this thesis, the length parameter of SDD model λ will be shown as H. This is just to make a discrimination between the length parameter of SDD model and eigenvalue of SSA technique.

Chapter 2

Background

In order to display the power of SSA technique, in what follows, almost all recently published articles associated with the application of SSA and Singular Value Decomposition (SVD) [1] in genetic studies are categorised and summarised. It is important to note that the following section is not aimed at showing the details of the presented studies. Instead we are mainly interested in showing how SSA technique can be applied as a useful method for variety of experiments with different aims of study.

2.1 Applications of SSA in Signal Extraction and Filtering

Here, the existing signal extraction and noise filtering applications of SSA in genetic studies are identified.

The first such application was reported in 2006 where SSA was used for signal extraction of *D. melanogaster*'s gene expression profiles [Holloway et al., 2006]. in that study, the relation between the maternal gradients and *D. melanogaster* segmentation process was investigated by quantifying the spatial precision in protein distribution patterns. The patterns were extracted by applying the basic SSA method.

[1] The SVD algorithms used here are either based on Hankel or Teoplitz matrix.

The idea of using SSA for signal extraction was then followed in an approximately similar study in [Surkova et al., 2008a] where the characteristic features of the expression domains were studied. The intensity levels were used to classify the domains. However, in prior to this classification the intensity values were refined by applying SSA as a filtering method.

Having depicted the importance of the signal extraction step in such studies, in 2008, a more technical study was conducted on the methodology of signal extraction from the noisy Bcd protein profile in [Alexandrov et al., 2008]. According to that study, the signal extraction process on noisy gene expression profiles (Bcd in particular) was complicated by two facts: (i) the data contained outliers and (ii) that the data was exceedingly noisy and the noise consisted of an unknown structure.

To reconstruct a precise signal, the author examined two approaches:

- Applying a small window length.
- Improving the signal and noise separability by adding a constant to the series.

Following the presented approaches, the researchers could obtain an analytical representation of the signal as a sum of two exponential functions; the well-known exponential pattern of the Bcd, and the background approximated by an exponential or linear function [Alexandrov et al., 2008].

Although these solutions are functional, in practice they may not be applicable for every profile. Moreover, by considering a small window length signal noise perturbation can seriously degrade the performance of the method. These approaches will be discussed in more detail in Chapter 4 where the Bcd signal extraction using SSA is comprehensively studied.

In addition, the activation of *hb* gene in response to different concentrations of Bcd gradient was studied in [Lopes et al., 2008]. In that study SSA was applied for filtering two types of noise; experimental noise and the noise caused by variability in nuclear order (As previously noted, this noise is categorised under the biological noise class) [Lopes et al., 2008].

2.2 SSA Combined with AR Model

Microarray technique is a very useful method for acquiring quantitative data in genetic studies. Today researchers are conducting most of their studies using this method. The main advantage of microarray is the capability of studying thousands of genes simultaneously. However, microarray data usually contains a high level of noise, which can reduce the performance of the results considerably [Klebanov and Yakovlev, 2007].

Among many applications of the microarray technique, studying rhythmic cellular processes can be addressed as an important application. The rhythmic cellular process is regulated by different gene products, and can be measured by using multiple DNA microarray experiments. Having a group of gene experiments over a time period, a time series of gene expression related to the rhythmic behavior of that specific gene can be achieved.

The characteristics of a rhythmic gene expression are summarised as follows:

- The number of time points and cycles related to a profile is usually very few. For example, the 14 time point elutriation observations may be in constitution of just one cell-cycle [Spellman et al., 1998].

- The dataset usually contains many missing values which need to be determined in prior to the analysis [Spellman et al., 1998].

- The intervals spaced between time points are not equal and the gene expression data is considerably noisy [Liew and Yan, 2009].

To overcome these issues and to extract the dominant trend from the noisy expression profiles, in 2008, Du et al. applied SSA on the obtained microarray results [Du et al., 2008]. The authors have also proposed a new procedure for analysing the periodicity of the transcriptome of the Intraerythrocytic Developmental Cycle (IDC) by combining Autoregressive (AR) model and SSA technique. This hybridisation of SSA and AR resulted in successfully identifying almost 90% of genes (4496 periodic profiles) in *P. falciparum*, which was a noticeable improvement regarding detecting 777 additional periodic oligonucleotides in comparison to the previous results reported in [Bozdech et al., 2003]. Fig 2.1 depicts the improvement achieved following applying SSA for analysing the rhythmic cellular processes.

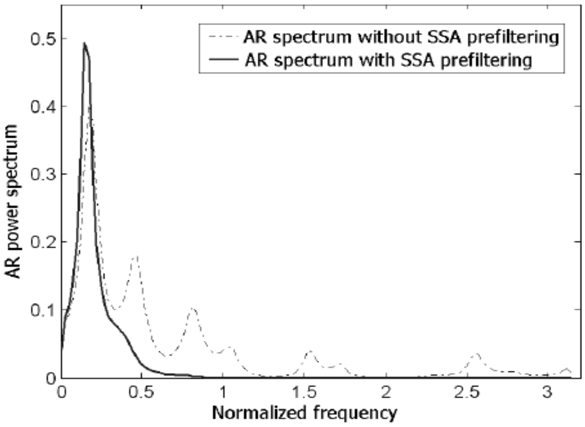


Fig. 2.1 The AR spectra of the expression profile of Dihydrofolate Reductase-Thymidylate Synthase with and without SSA filtering [Du et al., 2008].

Four subsequent studies followed this procedure and successfully improved the capability of detecting periodicity from 60% to 80% [Tang et al., 2010; Tang and Yan, 2010, 2012; Vivian et al., 2010].

According to [Liew and Yan, 2009], periodic profiles can be detected by combining SSA and AR as follows:

- At the first step SSA is applied to reconstruct each expression dataset. To this aim only those expression profiles with sum of first two eigenvalues over the sum of all eigenvalues greater than 0.6 are selected for reconstruction.

- The second step is devoted to the calculation of the AR spectrum. The frequency f_i at peak value point and the ratio of the power f_i in Regions of Interest (ROI) of the reconstructed profiles achieved in the first step are also computed. It should be noted that to get the ratio of the power, the frequency band [f_{i-1}, f_{i+1}] to the total power are used according to $S = \frac{power_i}{power_{total}}$.

- If the obtained power ratio S is larger than 0.7, the corresponding profile is classified as periodic [Liew and Yan, 2009].

2.3 Two Dimensional SSA

The Two Dimensional SSA (2D-SSA) approach was used to process two-dimensional scalar fields [Golyandina and Zhigljavsky, 2013]. The first difference between 2D-SSA and basic SSA is the window length. In 2D-SSA we need to choose two different values for the window length L (L_1, L_2), whilst in basic (also known as univariate) SSA we only require to select one window length. Note that if $L_2 = 1$, then 2D-SSA is equivalent to basic SSA. By choosing $L_2 = M$, the interaction among different series can be taken into account for further analysis (For more information, see [Golyandina and Usevich, 2010; Zhang et al., 2014]).

As mentioned before, Bcd is a transcriptional regulator for downstream segmentation genes. Therefore, the alteration in Bcd gradient expects to shift the downstream gene expression patterns [Porcher and Dostatni, 2010]. However, according to a study conducted by [Spirov and Holloway, 2003], the zygotic gene products are positioned more precisely than the gradients related to the maternal genes which indicates an embryonic error reduction or adjusting process.

In order to determine how gene regulation dynamics controls this different levels of noise, the AP segmentation in *D. melanogaster* was studied in [Holloway et al., 2011] .In that study, the activation of the *hb* gene in response to the Bcd protein gradient of the anterior part of the embryo was monitored by modelling the noise observed in Hb protein profile using a chemical master equation approach. To solve this model, the MesoRD software package was used http://mesord.sourceforge.net. The solution is found following a stochastic modeling approach. According to the results of that study, the noise in Hb greatly depends on the transcription and translation dynamics of its own expression.

Moreover, the multiple Bcd binding sites located on the *hb* promoter improve was claimed to have a critical role in the Hb pattern formation [Holloway et al., 2011].
The noise in that study was calculated using the following formula:

$$\sqrt{\frac{\Sigma[(data - trend)/trend]^2}{m-1}}, \tag{2.1}$$

where data is background-removed intensity of an energid, trend is the signal extracted using 2D SSA (AP and DV and m indicates the number of positions. This measurement was obtained for the activated region (15-45% egg length).

the 2D SSA method was also adopted by Golyandina et al. to measure the between-nucleus variability (the part of the noise which was targeted to study) seen in the gradient of Bcd in *D. melanogaster* embryos [Golyandina et al., 2012]. In that study, To measure and compare the noise in fixed immunostained embryos with the live embryos with fluorescent Bcd (Bcd-GFP), 2D SSA was adopted. The result of that study is depicted in Fig 2.2, as can be seen the nucleus-to-nucleus noise in Bcd intensity, is signal-independent for live and fixed immunstained embryos. Therefore the authors claimed that this noise is primarily sensitive to the nuclear masking technique used to extract the intensity levels.

2.4 SVD

SVD has proved to be a very useful method and has already become an appropriate tool for analysing the data achieved by microarray technique (see, for example, [Vikalo et al., 2008]). Such studies are mostly aimed at eliminating cross-hybridisation in time microarray data.

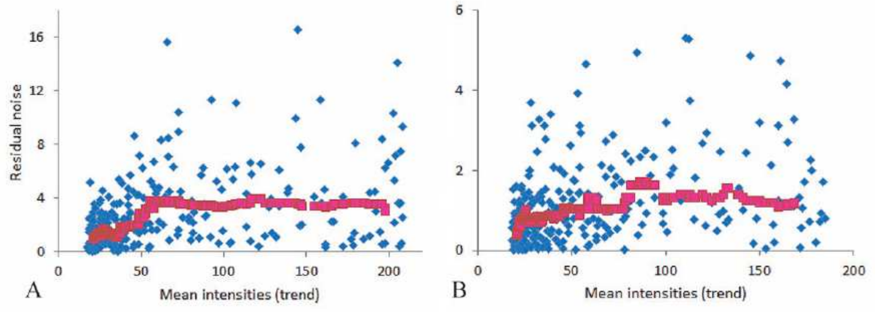


Fig. 2.2 Between-nucleus noise and the its corresponding signal dependency at different ROI, the between-nucleus noise is mostly correlated with the signal level. A) Nuclear intensity of Bcd, obtained by small ROI. B) Bcd nuclear intensity with large ROI [Golyandina et al., 2012].

To separate the components of a compound signal which results in a better estimation of the amounts of both hybridising and cross-hybridising targets, Vikalo et al. in [Vikalo et al., 2008] evaluated multiple techniques including SVD.

Nevertheless, the microarray technology has become a widely used method, the data is still very complicated to analyse. In 2010 Rau et al. presented an algorithm to infer the structure of gene regulatory networks using an empirical Bayes estimation procedure for the hyperparameters of a linear feedback state-space model [Rau et al., 2010]. In that model, to select the model, SVD of the block-Hankel matrix was applied. This approach enabled Rau et al. to significantly reduce the computation time required for running the algorithm mainly because in the new computational model, there is no need to run the algorithm over a wide range of values for the hidden state dimension. Moreover, as discussed in [Bremer and Doerge, 2009] the numerical computations in SVD even for a huge dataset such as microarray data is not time extensive compared to the other clustering techniques.

In the study conducted by [Bremer and Doerge, 2009] a new approach for gene ranking was introduced which was based on gene degree of regulation. According to that study, the gene ranking according to the regulation was genetically more meaningful than using the absolute expression or variation over time. Therefore, the introduced model was presented as a valuable aid to find the regulatory pathways and networks. The SVD in that method was applied on the block-Hankel matrix of the observation autocovariance estimated from the gene expression profiles. The number of singular values attributed to the greatest magnitude was considered as the best estimation of for the state space dimension. The singular values of the estimated Hankelautocovariance matrix were also calculated and normalised to a $0-1$ scale. The number of singular values of magnitude larger than the threshold was consequently considered as an estimation for the state space dimension.

Since the gist of this thesis concentrates on developing a new computational model for analysing the segmentation gene expression profiles and *bcd* gene in particular, the following section is devoted to the *bcd* gene, its function, characteristics and previous introduced methods and techniques to model this morphogen which offers further reasons for extensive interest in studying *bcd* gene.

2.5 bicoid Gene

As mentioned before, the fundamental question of how a simple fertilised cell develops into a complex multicellular organism has attracted the interest of many researchers since the last century. The first noteworthy attempt to answer this question was by Driesch in 1891 when he tried to separate two sea urchin blastomeres [Driesch, 1929]. Taking Driesch by surprise, each of the separated blastomeres developed to a half-sized blastula rather than a half sea urchin which was the expectation [Driesch, 1929]. This result furnished Driesch with the idea of considering a coordinate system which dictates the positional information along the embryo [Driesch, 1929]. Later on, this initial view by Driesch was perceived as a fundamental step in the discovery of what we know today as a morphogen. Today, it is widely accepted that the pattern of morphogen products play a key role in directing the embryo into a complex multicellular organism.

A prime example of morphogens is Bcd which is the first known morphogen identified by Nüsslein-Volhard [Driever and Nüsslein-Volhard, 1988]. Bcd is a homeodomain transcription factor which plays a crucial role in patterning head and thorax of the adult *D. melanogaster* [Driever and Nüsslein-Volhard, 1988]. The important role of Bcd in developing the anterior structure of *D. melanogaster* has been proven by several studies [Berleth et al., 1988; Driever and Nüsslein-Volhard, 1988; Frohnhöfer and Nüsslein-Volhard, 1986]. For example, Frohnhöfer et al. [Frohnhöfer and Nüsslein-Volhard, 1986] showed that embryos receiving different doses of *bcd* have differently sized anterior structures (Fig 2.3) and in the absence of *bcd*, anterior structures of body are replaced with posterior regions.

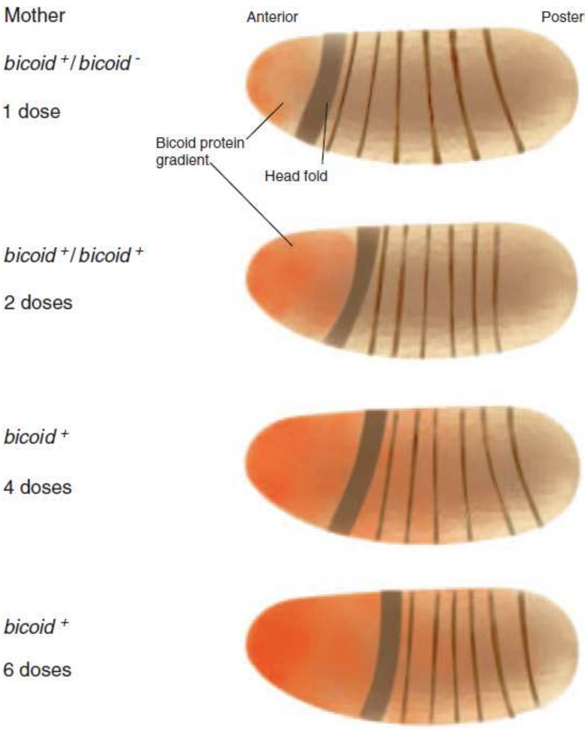


Fig. 2.3 Different doses of *bcd* affects the size of the head in *D. melanogaster*. Figure adapted from [Harwell and Leroy Goldberg, 2004]

During the oogenesis, *bcd* mRNA molecules which have been previously synthesised in the nurse cells are localised at the anterior end of the egg [Berleth et al., 1988].

As can be seen in Fig 2.4, the embryonic period starts with fertilisation. In that stage, different length of two main axes of *D. melanogaster* embryo; AP (Adult head to tail) and DV (Adult back to abdomen) causes the ellipsoidal shape of the embryo.

The translation of *bcd* mRNA starts immediately after fertilisation. Consequently, the Bcd protein molecules distribute along the AP axis of the embryo which forms a concentration gradient of Bcd [Driever and Nüsslein-Volhard, 1988]. Meanwhile, the nuclear division without division of the cytoplasm begins in the early *D. melanogaster* embryo which unlike most of the animal embryos results in a multinucleate cell called a syncytium.

A syncytial blastoderm contains about 6,000 nuclei in the cortical layer. Lack of cell membranes makes the Bcd diffusion much easier during this stage [Driever and Nüsslein-Volhard, 1988; Houchmandzadeh et al., 2002]. The very rapid nuclear divisions in the syncytium which is also known as cleavage, further rise the number of nuclei [Houchmandzadeh et al., 2002].

By the ninth division majority of nuclei migrate to the periphery and leave the yolk cells in the middle of the embryo. This stage is known as the blastoderm [Houchmandzadeh et al., 2002].

Later on, the blastoderm undergoes four more cleavage cycles (10-14A). By invaginating the cell membrane to enclose each nucleus the critical stage of cellularisation begins. Up to the cellular blastoderm stage, development depends largely—although not exclusively—on stocks of maternal mRNA and protein that accumulated in the egg before fertilisation. Mid blastula transition starts by the end of cycle 14A when zygotic transcription initiates rapidly after maternal mRNA and protein degradation. Then after, the gastrulation stage completes after forming the germ layers.

The whole process of 14 different cleavage cycles occurs during the first three hours after fertilisation. The final result of this process is the segmented body of *D. melanogaster* [Driever and Nüsslein-Volhard, 1988; Grimm et al., 2010; Grimm and Wieschaus, 2010]. This period is known as the early stage of the development in *D. melanogaster* (Fig 2.5).

the segmented pattern contains fourteen repeating units; three segments in the head (mandibular, maxillary and labial), three in the thorax and eight in the abdomen.

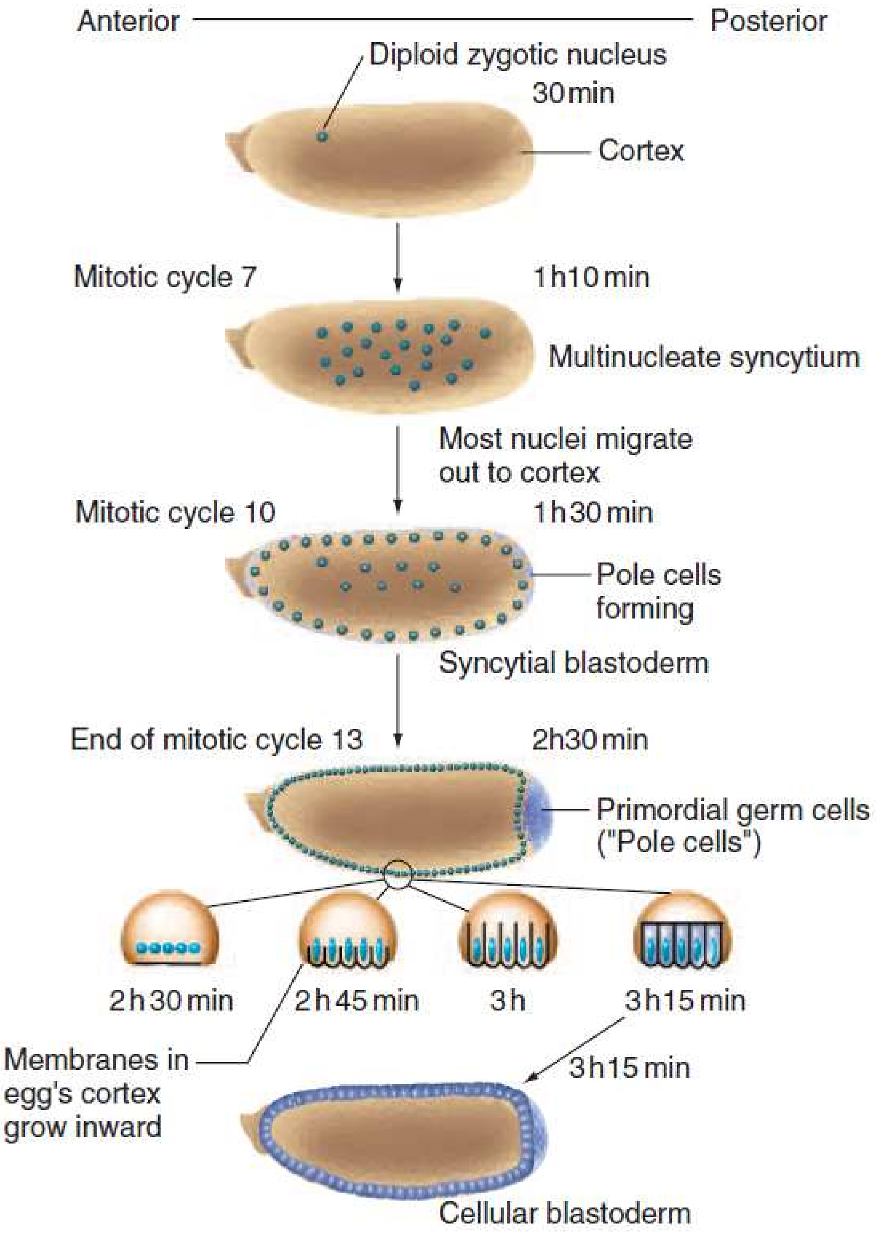


Fig. 2.4 Nuclear divisions in 14 cleavage cycles produce a syncytial blastoderm. Figure adapted from [Harwell and Leroy Goldberg, 2004].

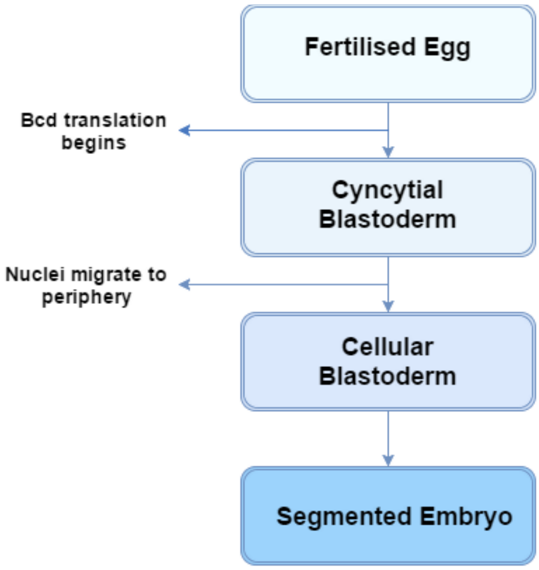


Fig. 2.5 Early development in *D. melanogaster*. From fertilised egg to the segmented embryo.

Efforts to introduce a model for Bcd gradient have brought together many researchers from different areas, including biology, physics, and statistics. Hence, the various aspects of this gradient, including molecular and functional have been explored in different studies.

Here, we classify and explore those studies mainly focused on Bcd gradient. To that aim, this review starts with a brief introduction on models which are representing a general morphogen gradient formation.

In 1902, Morgan published a book titled “Regeneration". In that book, Morgan suggested that the fate of the cells in an organism is determined by a gradient of “formative substances” [Morgan, 1901]. These substances specify the pattern formation. Later on, through an experiment, Spemann [Spemann and Mangold, 2003] discovered that in a gastrula, a secondary body axis is formed by grafting a group of dorsal cells onto the opposite ventral pole. Spemann’s experiment indicated that the gradient responsible for pattern formation was released by a group of localised cells [Spemann and Mangold, 2003]. According to [Child, 1941], the gradient of formative substances contains metabolic functions.

In 1952, Turing named these chemical substances as morphogens [Turing, 1952]. He also introduced the reaction—diffusion model. This model shows how different morphogens with slightly different diffusion properties interact with each other to generate spatial patterns for various concentrations of morphogens [Liu et al., 2011].

The establishment of morphogen gradient then was described by French Flag Model [Wolpert, 1969]. According to that model, a morphogen is extracted from a group of source cells to provide positional information. Different target genes are consequently expressed above the specific concentration thresholds of that morphogen, Fig 2.6.

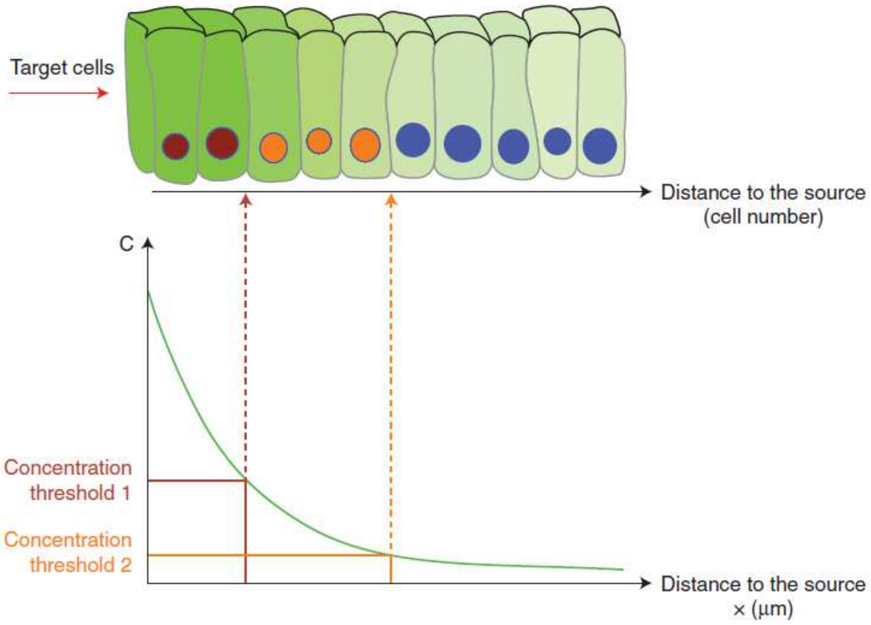

Target cells
Distance to the source (cell number)
C
Concentration threshold 1
Concentration threshold 2
Distance to the source x (μm)

Fig. 2.6 French flag model. Based on this model the concentration is a function of the positions (x). The morphogen activates different target genes above different concentration thresholds (brown and orange) [Liu et al., 2011].

Source-sink model [Crick, 1970] was proposed on the foundation of the French-flag model. Based on Source-sink model, morphogen molecules are produced in source cells and destroyed at the other end of the embryo by sink cells. However, nowadays it is believed that considering sink cells is not requisite since a morphogen gradient can be formed even if it is not degraded at all [Liu et al., 2011].

In 1988, for the first time, the gradient of Bcd in D. melanogaster was visualised by the antibody staining technique [Driever and Nüsslein-Volhard, 1988]. This visualisation along with the inputs of the previous models supported Driever and Nusslein-Volhard to introduce the SDD model which is the most widely accepted model for Bcd gradient. According to SDD model, Bcd gradient in the embryo follows an exponential curve until it reaches a steady state. In the steady state, the production and degradation of Bcd molecules are in the balance.

Since then, studying the Bcd gradient has become popular, and many quantitative models have been developed to portray the function of this gradient and its interactions with other morphogens.

Most of the models accounted for Bcd gradient establishment rely on three principal concepts: Bcd synthesis at the anterior end of the embryo, diffusion and degradation. Interestingly, what makes these models different from each other is the assumptions held to find the parameters.

2.5.1 Modeling Bcd using Differential Equation

This equation is the first mathematical model proposed for describing the Bcd gradient in 1985 which was introduced before visualising this gradient [Rice, 1985]. Based on this model by assuming the space interval as Δx, the concentration of Bcd molecules over the period of Δt follows the equation 2.2.

$$\begin{aligned} C(x,t+\Delta t) = C(x,t) - kC(x,t)\Delta t \\ + d\big[-C(x,t) + \frac{1}{2}C(x+\Delta x,t) + \frac{1}{2}C(x-\Delta x,t)\big] \\ + j(x,t)\Delta t. \end{aligned} \tag{2.2}$$

where C is the concentration of Bcd [molecules/μm^3] and is a function of space and time. It should be noted that for a very small Δt the concentration of Bcd at time $t+\Delta t$ is equal to the concentration at t. Unlike the sink model which considers a localized sink, in this model in order to generate the steady-state gradients morphogens are degraded with a constant degradation rate $k[1/s]$ and the terms inside the bracket describes the Bcd movement which is respectively related to the diffusion, leaving and coming Bcd molecules from the adjacent positions $x \mp \Delta x$. The one-half here shows the equal probability of Bcd movement to the left and right, d is used for scaling and is related to D which is diffusion coefficient and j is the source function accounts for describing the spatial distribution and the rate of Bcd production [Grimm et al., 2010].

Accordingly, the rate of alteration in Bcd concentration can also be obtained by reordering Equation 2.2:

$$\begin{aligned}\frac{C(x,t+\Delta t)-C(x,t)}{\Delta t} &= -kC(x,t) \\ &+ \frac{d(\Delta x)^2}{2\Delta t}\frac{-2C(x,t)+C(x+\Delta x,t)+C(x-\Delta x,t)}{(\Delta x)^2} \\ &+ j(x,t),\end{aligned} \tag{2.3}$$

by assuming the limit Δt, $\Delta x \rightarrow 0$ equal to a finite value when $d(\Delta x)^2/2\Delta t \equiv D$, Differential Equation is obtained.

$$\frac{\partial C}{\partial t} = D\frac{\partial^2}{\partial x^2} - kC + j. \tag{2.4}$$

From now for the simplicity we can show $C(x,t)$ as C.

It is worthy to mention that when using this model different assumptions should be taken into account and also the parameters used in this model are not tractable thorough analytical solutions since the parameters in equation 2.4 are the representor of the macroscopic properties of the embryo which sometimes are not a good representative of Bcd molecules (for more information see, [Crick, 1970; Grimm et al., 2010; Liu et al., 2011]).

2.5.2 Synthesis Diffusion Degradation model

As mentioned earlier this model was proposed by Nüsslein-Volhard and is widely known as the SDD model.

Since the $j(x,t)$ for any other space except the source is equal to zero, by having diffusion and degradation together, the diffusion equation with linear degradation is achieved. Therefore, the changes in Bcd concentration C over time t can be written as:

$$\frac{\partial C}{\partial t} = D\frac{\partial^2 C}{\partial x^2} - kC. \tag{2.5}$$

Now by considering the fact that in the steady state, $\frac{\partial C}{\partial t} = 0$, the diffusion equation with linear degradation can be written as:

$$\frac{\partial^2 C}{\partial x^2} = \frac{k}{D}C. \tag{2.6}$$

Either an exponential functions or sums of hyperbolic sines and cosines can be considered as the solution for the second derivative of C. Since the sums of hyperbolic sines and cosines can also be written as exponentials, the general solution for C follows:

$$C = C_0 e^{-\frac{x}{\lambda}}, \tag{2.7}$$

where λ is a constant value. Assuming $C = C_0$ at $x = 0$, therefore C_0 is the source concentration and at $x = \lambda$, C will be equal to $C_0 e^{-1}$, thus λ is the distance to the source at which the concentration has dropped to $1/e$ of the maximal value at the anterior (at $x = 0$) Fig 2.7(A). It should be noted that the Equation 2.7 is only true when the limit of $L \gg \lambda$ and $\lambda = \sqrt{\frac{D}{k}}$. This exponential decaying gradient is a general way of presenting the SDD model [Bergmann et al., 2007; Driever and Nüsslein-Volhard, 1988; Gregor et al., 2005; Houchmandzadeh et al., 2002]. This model can also be written as:

$$C = \frac{C_0}{k\lambda} e^{-\frac{x}{\lambda}} \tag{2.8}$$

Based on this model the activation of gap genes such as *hb* and *kr* is regulated by the Bcd gradient and their transcription begins at different concentration of Bcd [Driever and Nüsslein-Volhard, 1988].

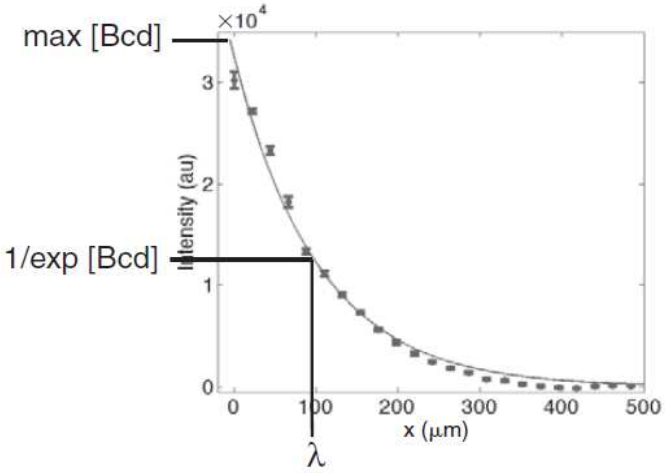


Fig. 2.7 Bcd gradient along the A-P axis. Calculating the length constant [Grimm et al., 2010], λ..

In 2005, the SDD model was tested by injecting Dextran in living embryos and which resulted in added support for the SSD model [Gregor et al., 2005]. In that study, the relation between the egg size during evolution and gene expression pattern scaling was also investigated and the authors showed that even in embryos with very different sizes, diffusion constants are the same in closely related dipteran species. However, it should be noted in different species, the pattern of gene expression scales with egg length and it also greatly depends on the Bcd lifetime [Gregor et al., 2005].

Among the quantitative approaches, Gregor et al. experiment [Gregor et al., 2007] on nuclear dynamics and stability of the Bcd gradient must be addressed as a successful study.

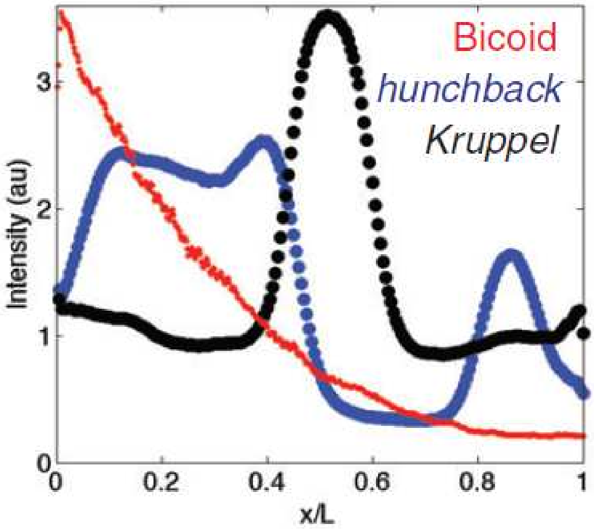


Fig. 2.8 Expression the different target genes based on Bcd gradient (find the detailed explanation in text). Figures adapted from [Gregor et al., 2007].

The issues outlined by Gregor et al. is mentioned below in detail and discussed further using the experiments achieved by other related studies.

According to [Gregor et al., 2005, 2007], the Bcd gradient establishes within the first hour after fertilisation. This pattern stays consistent during blastodermal stages (by 10% precision). However, since Bcd is a DNA-binding protein, the nuclear Bcd concentration has an up and down pattern during the mitosis caused by forming and deforming the nuclear membrane [Gregor et al., 2007]. This observation raised the question of how the stable Bcd gradient is established in the embryo in the presence of the repeated dynamical structural change during the mitoses.

Another important part of Gregor et al. experiment is measuring the diffusion coefficient of Bcd using fluorescence recovery after photobleaching method. According to this measurement, $D \sim 0.3\mu m^2/s$ and $\lambda \sim 100\mu m$. Therefore, the author concluded that using the SDD model, by considering theSDD model, the time needed to achieve the steady-state concentration profile ($> 9h$) is much greater than the syncytial embryonic development time ($\sim 2h$) and protein lifetime τ. Moreover, the length constant λ is much smaller than the length of the embryo [Gregor et al., 2007]. Considering the slow Bcd diffusion, now the question is: How the Bcd gradient is formed so quickly after fertilisation?

Therefore, taking the nuclear degradation element into account, Gregor et al. [Gregor et al., 2007] proposed a variant of the SDD model and suggested that Bcd gradient is founded following the intranuclear degradation of Bcd protein molecules:

In that model, transporting the molecules into the nucleus depends on the Bcd concentration in the cytoplasm ($k_{in}C_{out}$). However, disappearing the Bcd molecules from the nucleus is due to both their active transportation back into the cytoplasm (at k_{out}) and degradation (at k_{d-nuc}). This dynamic for $n(t)$ nucleus is presented as:

$$\frac{\partial n(t)}{t} = k_{in}C_{out} - \frac{1}{\tau_n}n(t), \tag{2.9}$$

$$\frac{1}{\tau_n} = k_{out} + k_{d-nuc}, \tag{2.10}$$

where τ_n is the life time related to the Bcd molecules in the nucleus. As cytoplasmic concentration approximated to be constant, hence we have:

$$n(t) = n_\infty - \Delta n_0 e^{(-t/\tau_n)}, \tag{2.11}$$

where Δn_0 is the molecules reduction, and $n_\infty = k_{in}\tau_n C_{out}$ is the number of molecules at nucleus steady state. Following this equation, the τ_n is approximated to be around 1 or 2 minutes. considering this equation along with the physical factors of the nucleus, equation 2.12 is achieved:

$$\frac{C_{in}}{C_{in}} = \frac{3D\tau_n}{r_n^2}. \tag{2.12}$$

where r_n is the nucleus radius.

Gregor et al. suggested two possible reasons for the paradox seen between the slow diffusion constant and the rapid formed steady state; diffusion coefficient may not be consistent at different temporal and spatial conditions, or there might be and an element of nuclear contribution function or active transportation which should be taken into account [Gregor et al., 2007].

These suggestions encouraged more researchers to investigate the pointed out possibilities:

2.5.3 Pre-steady-state decoding

According to the previous studies, Bcd induces the cell's fate only after it reaches a steady state. However, in 2007 the pre-steady state decoding hypothesis raised by Bergmann et al. noting that Bcd gradient is decoded even before reaching a steady state [Bergmann et al., 2007].

Bergmann et al. showed that changes in Bcd dosage induces small shifts in the gap and pair-rule gene expression domains suggesting that the observed shifts can be explained by the positions of gap genes and their distance to the source of Bcd since those genes at the posterior part found to be less sensitive to Bcd changes which is inconsistent with the steady state assumption. The proposed model in [Bergmann et al., 2007], explains the small shifts in the gap and pair-rule gene expression domains. Based on this model, the solution of Eq 2.4 is time independent and Bcd concentration changes with time:

$$C = \frac{C_0}{k\lambda}\left[e^{-\frac{x}{\lambda}} - \frac{e^{-\frac{x}{\lambda}}}{2} erfc(\frac{2Dt/\lambda - x}{\sqrt{4Dt}}) - \frac{e^{\frac{x}{\lambda}}}{2} erfc(\frac{2Dt/\lambda - x}{\sqrt{4Dt}})\right]. \tag{2.13}$$

Based on this model, the λ is not constant and is due to change by both degradation and diffusion time.

2.5.4 Nuclear trapping

The idea of nuclear trapping proposed by Coppey considers a substitute for Bcd degradation [Coppey et al., 2007; Liu et al., 2011].

This model explains the temporarily stable Bcd gradients in the absence of Bcd degradation. Based on this model, nuclear division not only increases the nuclear density in the syncytium, but it also raises the shuttling of Bcd molecules in and out of nuclei at each cleavage cycle. This, in turn decreases the effective Bcd diffusivity. Hence, the nuclear Bcd concentration gets stable over several cleavage cycles, despite the continuous expansion of total Bcd concentration. Therefore, the steady Bcd concentration in the nuclei is read out by the target genes.

$$C_{(tot)} = \frac{C_0}{D}\left[\frac{\lambda(t)}{sqrt\pi}e^{-\frac{x^2}{\lambda(t)^2}} - x.erfc\left(\frac{x}{\lambda(t)}\right)\right], \tag{2.14}$$

2.5.5 *bcd* mRNA based approaches

As discussed before, based on [Ephrussi and St Johnston, 2004; Johnston et al., 1989] *bcd* mRNA is transcribed in the mother and localised at the anterior end of the egg and encodes Bcd. The Bcd molecules distribute and establish a gradient along A-P axis. Surdej and Jacobs-Lorena in 1998 suggested that the stability of maternal *bcd* mRNA is under a systematically regulation [Surdej and Jacobs-Lorena, 1998]. Years later, following this idea, an entirely different approach for describing the Bcd gradient establishment proposed by Spirov et al. [Spirov et al., 2009].

Spirov et al. state that Bcd gradient is formed mainly by transporting the *bcd* mRNA along the cortex and not by Bcd diffusion. Using FISH method and confocal microscopy authors showed that the gradient related to *bcd* mRNA and Bcd are virtually identical at all time classes. Moreover, their proposed Active RNA Transport and Synthesis model (ARTS) could successfully explain the small measured diffusion coefficient by Gregor et al. Following the ARTS model, Berezhkovskii et al. state since there is no production of mRNA at the anterior part of the embryo and all the *bcd* is provided by the mother, the *bcd* diffusion follows a Gaussian rather than an exponential distribution [Berezhkovskii et al., 2009]. Taking this distribution into account the source function of Equation 2.4 is rewritten as:

$$j = \frac{C_0}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}}, \tag{2.15}$$

where σ is the length constant attributed to the mRNA Gaussian distribution (i.e $\lambda_{RNA} \equiv \sigma$), now another solution for the Equation 2.4 can be obtained:

$$C(x) = \frac{C_0}{2k\lambda} e^{\frac{\sigma^2}{2\lambda^2}} \left(e^{\frac{x}{\lambda}} erfc\left(\frac{\sigma^2 + \lambda x}{\sqrt{2}\lambda\sigma} \right) + e^{-\frac{x}{\lambda}} erfc\left(\frac{\sigma^2 - \lambda x}{\sqrt{2}\lambda\sigma} \right) \right) \tag{2.16}$$

Following this solution Bcd gradient is not just a simple exponential curve and both λ and σ should be considered in gradient formation model.

According to these findings, another computational model proposed by Dilao and Muraro [Dilão and Muraro, 2010]. In that model, the *bcd* mRNA diffuses along the A-P axis and Bcd protein production occurs in the ribosomes localised near the syncytial nuclei. A scaling relation is also considered between the *bcd* diffusion and degradation coefficient rate and the embryo length. The steady state achieved when the *bcd* translation is completed. Hence, the idea of Bcd degradation step is no longer a necessary to consider in this model.

However, in 2011 Little et al. by using a novel method of fluorescent in situ hybridization found that approximately 90% of all *bcd* mRNA localised within the 20% anterior part of the embryo. Hence this result is inconsistent with the Dilao and Muraro work because the mRNA spatial distribution is not sufficient to provide the Bcd protein gradient especially at the posterior end of the egg therefore either the active or passive Bcd movement is necessary for the gradient formation.[Little et al., 2011].

Accordingly, the result illustrated by Little et al. experiment adds more confidence and value to the reliability of adopting the Bcd expression profile in studying the Bcd gradient.

Gregor et al. also studied the precision and reproducibility of positional information of Bcd. in [Gregor et al., 2007] where a 10% concentration difference was reported to be necessary to distinguish individual nuclei from their neighbours. However, the absolute concentration of Bcd molecules at half-embryo length, where Bcd targets Hb, found to be too low to be distinguished by nuclei [Gregor et al., 2007].

This, made the author quantify three other measures of precision; limits set by the random arrival of Bcd molecules at their targets, input—output relationship between Bcd and Hb and reproducibility in the spatial profile of Bcd from embryo to embryo. Using a combination of different experiments, the precision in all the quantities mentioned above found to be 10% [Gregor et al., 2007]. This quantification suggests that mechanisms other than small noise reduction should be involved to achieve a 10% accuracy and it was concluded that the pattern formation in the embryo is precisely and reproducibly controlled [Gregor et al., 2007].

To understand the mechanisms that control this extreme precision and reproducibility, several studies were conducted by other researchers [Crauk and Dostatni, 2005; Lewis, 2008; Wang and Leu, 1996]. For example, using stochastic simulations, Wang et al. challenged the traditional models for gene regulation systems and suggest that activators regulate the gene transcription by recruiting several proteins such as general transcription factors and RNA polymerase II (Pol II) [Wang and Leu, 1996].

The stochastic approaches conducted to explain the dynamics of Bcd gradient are discussed in more details in the following section.

2.5.6 Stochastic models

The deterministic models presented so far have limitations when uncertainty analysis is required. Thus, to provide a more detailed understanding of the chemical reactions in a morphological system, several researchers studied a discrete chemical reaction system, based on the stochastic simulation [Andrews and Bray, 2004; Gibson and Bruck, 2000; Liu, 2013].

According to [Liu, 2013], the stochastic simulation tries to answer two fundamental questions: "which reaction occurs next and when does it occur?". Stochastic simulation is considered as a new successful method for simulating evolution at the genomic level. However, when the system contains many reactions, the stochastic simulation can be computationally expensive [Liu, 2013].

To understand the dynamics of Bcd gradient formation at the molecular number level and to identify the source of the nucleus-to-nucleus expression variation Wu et al. formulated a chemical master equation by considering the embryo as a finite number of sub volumes and presented the stochastic simulations for molecules undergoing transitions between those compartments [Wu et al., 2007]. The simulations then obtained by using a publicly available software called MesoRD [Hattne et al., 2005].

Another stochastic approach can be found in [Dewar et al., 2010] where a bayesian inference approach presented to solve both the parameter and the state estimation problem for a stochastic reaction-diffusion system.

To address that how Bcd gradient is generated and read out precisely to form the Hb gradient with a small embryo-to-embryo fluctuation, OkabeOho et al. developed one- and three-dimensional stochastic models. In the three-dimensional model, the consequences of nuclei dynamical change is examined by considering a cylinder of length $L \approx 490\mu m$ extending along the AP axis of the embryo [Okabe-Oho et al., 2009].

The achieved results by these two models reconfirmed that the stable profile of Bcd establishes by the stochastic processes of synthesis, diffusion and degradation as well as the rapid movements of the Bcd molecules to the nuclei [Okabe-Oho et al., 2009]. Moreover, the authors claimed that the fluctuations in *hb* gene profile are due to the Bcd random arrival at the *hb* enhancer [Okabe-Oho et al., 2009].

In another model presented in [Deng et al., 2010], two forms of Bcd molecules are reflected in the model; free-diffusing molecules, Bfree, that are in the cytoplasm, and immobile molecules, Bbound, that are bound to the low-affinity DNA sites inside the nuclei. According to this two-dimensional stochastic model, Bbound plays a significant role in the formation of the Bcd gradient [Deng et al., 2010] confirming the results reported in [Gregor et al., 2007] which showed that the Bcd and Hb random diffusions are the dominant sources to limit the precision of readout.

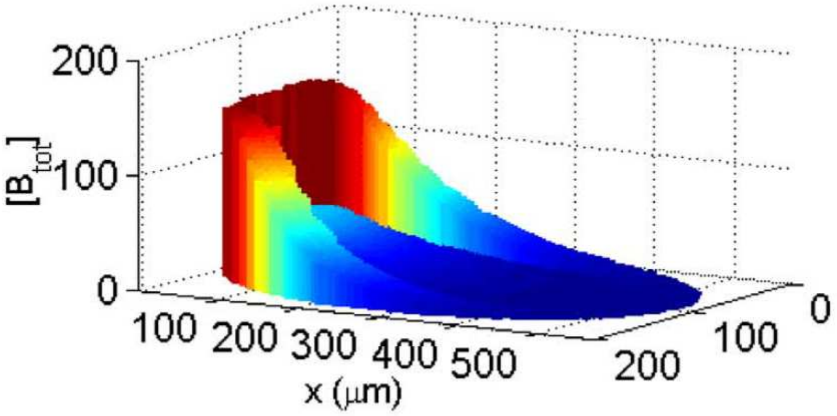


Fig. 2.9 A simulated embryo at cleavage cycle 14 showing the local total Bcd concentration [Btot] (arbitrary units). Figure adopted from [Deng et al., 2010].

The quantitative and theoretical analysis of Bcd gradient has enhanced our knowledge on several key parameters; the generation mechanism of Bcd profile, the readout system for downstream genes. However, still a simple quantitative model with more molecular and cell biological depth is encouraged to be developed. Such a less detailed mechanism, in which avoids using too many equations and parameters gives a more reliable image of *D. melanogaster* morphogen system.

Furthermore, the fundamental questions of how morphogens and their receptors are moving in and extracellular matrix and how the precision and robustness of developmental control is achieved at the molecular number level are yet to be addressed.

Chapter 3

Methodology

Presented in this chapter is the methodology relevant to this thesis. The primary focus is on the theory underlying the basic SSA and principles associated with this technique. Therefore, the different steps of SSA are clearly illustrated here. The potential points to improve the performance of the SSA technique are also outlined and consequently tackled under the following chapters of this thesis.

In this thesis, to determine whether the proposed techniques are equally/out or under performing the current models in use, extensive comparison studies to other well-known signal processing techniques are conducted. The benchmark models and the metrics applied to compare the performance of the models are concisely articled in this chapter.

3.1 Singular Spectrum Analysis

The origin of SSA goes back to 1986 with a publication by Broomhead and King [Broomhead and King, 1986]. Later on, this technique was independently developed in Russia, UK and the USA.

The notably versatile and flexible nature of this method soon made SSA as an excellent candidate and powerful technique in time series analysis. Compared to the other well established time series analysis techniques, SSA is a young method, but it has been developing at a noteworthy pace. Therefore, the application of this technique as the main analytical method is reported in a wide range of areas: from mathematics and physics to economics and medicine. As once quoted by Golyandina et al. "*Any seemingly complex series with a potential structure could provide another example of a successful application of SSA*" [Golyandina et al., 2001b].

SSA is a powerful non-parametric technique which combines some elements of classical time series analysis, multivariate statistics, multivariate geometry, dynamical systems and signal processing [Hassani, 2007]. Unveiling the dynamics underlying a given time series and removing the random, unexplained component of noise are among the most widely adopted applications of the SSA technique.

As it was mentioned before, an important feature of SSA lies in the fact that this method can be applied to a study without making any assumptions about stationarity and normality of the data under study [Sanei and Hassani, 2015]. Therefore, in the process of the analysis, the data transformation is no longer needed. This privilege can be considered as a significant advantage of this technique because in adopting parametric methods, data transformation is an inevitable step of the analysis in most cases. Although transforming the data facilitates the analysis in several ways, it may also cause loss of information [Hassani et al., 2013c].

SSA technique is further classified as univariate (also known as basic SSA) and multivariate SSA (MSSA). The univariate SSA considers a single time series and is applied in this research which will be explained in details later in this chapter. Some recent examples of the application of the univariate type can be found in [Ghodsi et al., 2015; Hassani et al., 2010, 2013b, 2009a; Hassani and Thomakos, 2010; Rodrıguez-Aragón and Zhigljavsky, 2010; Sanei et al., 2011; Silva and Hassani, 2015]. The multivariate version of this method considers modelling multiple information about the same series or a particular time series with a time lag into the past [Hassani and Mahmoudvand, 2013].

Although in comparison to its univariate counterpart, the MSSA is a relatively younger approach, its diverse capabilities have recently transferred it into a useful method for a wide range of areas especially forecasting applications, see for example [Groth and Ghil, 2011; Hassani et al., 2013a; Hassani and Mahmoudvand, 2013; Kapl and Müller, 2010; Oropeza and Sacchi, 2011; Patterson et al., 2011].

The selection between the two versions of SSA greatly depends on the aim of the study. Given the data characteristics previously portrayed in chapter 1 (i.e. The intensity levels attributed to the nuclei along the AP axis which are considered as a sequenced series) and the principal objective of this research (i.e. Enhancing the gene expression signal processing) our focus is mainly on the univariate SSA.

SSA comprised of two complementary stages known as decomposition and reconstruction. Each stage consists of two additional steps [Hassani, 2007]. In summary, the main aim of SSA technique is to decompose a time series to identify the trend, harmonic and noise components. A less noisy series then is reconstructed by estimating the trend and harmonic components [Golyandina et al., 2001a].

Before going to the details of each step, let us illustrate the whole process with a simple example:

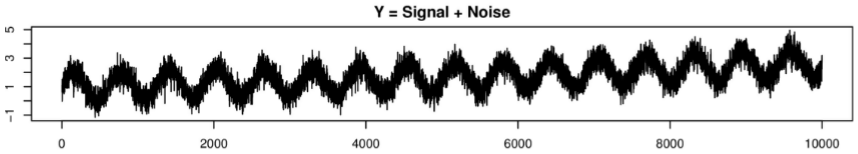


Fig. 3.1 An example of a noisy time series called Y.

In Figure 3.1, a simple series called Y is depicted. According to a general concept in time series analysis, any given time series can be represented as a composition of signal and noise.

It should be noted noise in statistics is defined as the colloquialism for recognised amounts of unexplained variation in a sample [Hida et al., 2013]. According to its statistical characteristics, noise can be categorised in two classes of White Noise and Coloured Noise [Arnold et al., 1978; Hida et al., 2013].

White noise is a discrete signal whose samples are considered as a sequence of serially uncorrelated random variables with zero mean and finite variance [Chichilnisky, 2001; Hida et al., 2013]. In other words, white noise is defined as an uncorrelated noise process with equal power at all frequencies [Kuo, 1996]. If the variables in the series are drawn from a Gaussian distribution, the series is called Gaussian white noise [Kuo, 1996].

However, the coloured noise has correlations which are not satisfied by white noise. Therefore, coloured noise can be described by having a different integrated power at different frequency bands [Chichilnisky, 2001; Hida et al., 2013]. The power spectrum is used to classify the coloured noise into several groups known by different colours [Chichilnisky, 2001; Hida et al., 2013].

Therefore, considering the general concept in time series analysis, Y consists of signal and noise. Nevertheless, this may not be distinctly apparent in that figure.

To extract the signal from Y, the noise components need to be identified and removed first. Fig 3.2 shows the extracted signal and what is left from the Y series known as the residuals or noise.

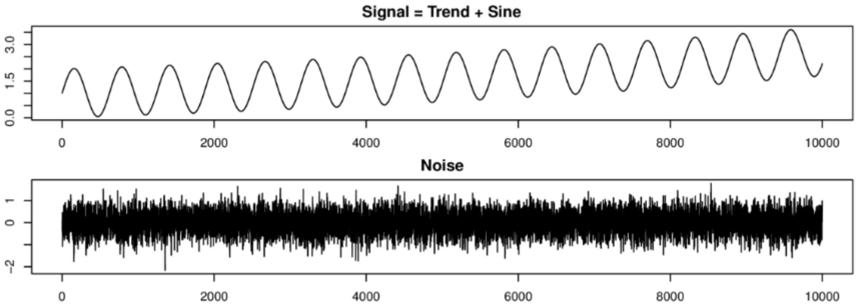


Fig. 3.2 Signal and noise components in Y.

In this particular example, the signal can be further decomposed into a trend component and sine component as shown in Fig 3.3.

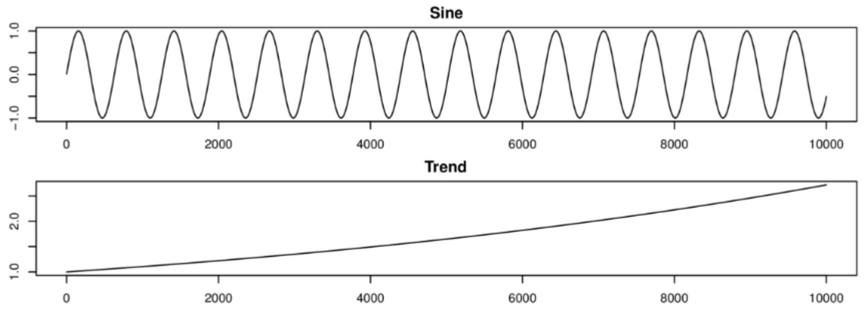


Fig. 3.3 Sine and trend components in Y.

The analytical process of this example follows several steps which are depicted in Fig 3.4 below. According to this flow chart, the SSA algorithm is made upon making two important inputs:

- Window length L.
- Number of eigenvalues r.

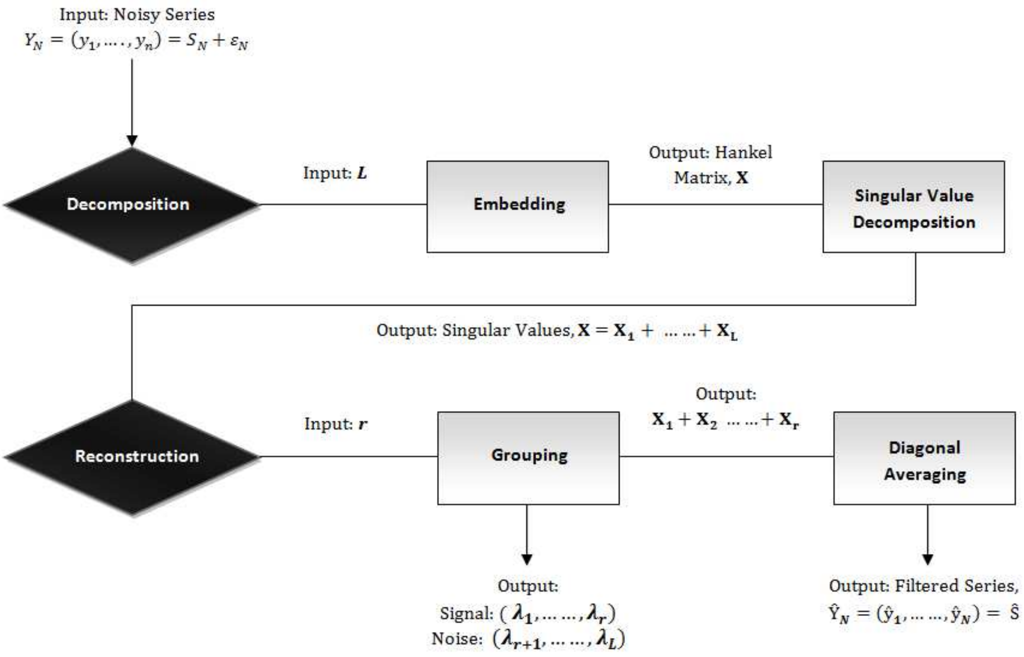

Input: Noisy Series
$Y_N = (y_1, \dots, y_n) = S_N + \varepsilon_N$
Decomposition
Input: L
Embedding
Output: Hankel Matrix, $\mathbf{X}$
Singular Value Decomposition
Output: Singular Values, $\mathbf{X} = \mathbf{X}_1 + \dots\dots + \mathbf{X}_L$
Reconstruction
Input: r
Grouping
Output: $\mathbf{X}_1 + \mathbf{X}_2 \dots\dots + \mathbf{X}_r$
Diagonal Averaging
Output:
Signal: $(\lambda_1, \dots\dots, \lambda_r)$
Noise: $(\lambda_{r+1}, \dots\dots, \lambda_L)$
Output: Filtered Series,
$\hat{Y}_N = (\hat{y}_1, \dots\dots, \hat{y}_N) = \hat{S}$

Fig. 3.4 A summary of the basic SSA process [Sanei and Hassani, 2015].

SSA analysis initiates with a noisy time series Y_N. The first SSA choice (L), is the input for decomposition stage. Following the embedding step, a Hankel matrix $\mathbf{X}$ is achieved which is then forwarded as the input of the SVD step. The SVD step provides the singular values of the series. Further examining the singular values leads to identifying and differentiating between signal and noise components.

At the reconstruction stage, the singular values are grouped according to the second and final choice of SSA, r. This step results in the grouping matrices $\mathbf{X}_1, \ldots, \mathbf{X}_L$. The r is the cutting point between signal and noise of the grouping matrices. Eventually, diagonal averaging is applied to transform the defined signal matrices into a Hankel matrix so that it can be consequently converted into a less noisy time series.

As it will be discussed in more details later in this chapter, the effectiveness and performance of SSA technique greatly depend upon making the two choices of L and r. The selection of L considerably relates to the aim of study and the data under investigation, while the choice of r is imperative to provide the noise free series following the reconstruction stage [Sanei and Hassani, 2015]. The choice of r has a critical role in determining the efficiency of the signal extraction process. An inaccurate estimation of r results in signal-noise perturbation or loss of information in some parts of the signal(s). In this thesis, several approaches are introduced to optimise the SSA performance. The improvements made upon the L and r selections are discussed in details in Chapter 5.

3.2 Basic SSA

The theory underlying basic SSA is exploited and discussed below. In doing so, we mainly follow [Hassani, 2007] and [Sanei and Hassani, 2015].

Before discussing the methodology of the SSA technique, the general idea of applying SSA in gene expression analysis is briefly outlined here.

As noted before, gene expression can be traced either in time or space. The data points used in this study represent the intensity levels attributed to the nuclei along the AP axis and are considered as a sequenced series. Therefore, the analysis starts with a one-dimensional gene expression profile and the aim is to extract the signal of this profile.

Let us consider a noisy gene expression Y_N with any arbitrary series length N, such that:

$$Y_N = (y_1, \ldots, y_N). \tag{3.1}$$

This series can be the gene expression profile of any of the segmentation genes. However, from now for simplicity, we consider that Y_N represents the Bcd profile.

As it was already discussed Bcd expression profile is exceedingly noisy. Therefore, we assume that Y_N comprises of signal and noise. Hence, Y_N can be represented as:

$$Y_N = S_N + E_N = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix} = \begin{pmatrix} s_1 \\ s_2 \\ \vdots \\ s_N \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_N \end{pmatrix}, \tag{3.2}$$

where S_N and E_N stand for the signal and noise respectively.

Next, SSA is applied to estimate the signal and leave aside the approximated E_N. It should be noted that the term ‘approximated’ is used here as it is not possible to affirm if the extracted signal is the true signal present in the real situations.

Below we provide a brief discussion on the methodology of the SSA technique.

3.2.1 Stage 1: Decomposition

At the decomposition stage, Window Length L which is the first choice of SSA is defined. The L choice transfers a one-dimensional expression profile Y_N into a multidimensional series.

It is of note that L is an integer such that $2 \leq L \leq N-1$. The first step of the SSA algorithm produces a Hankel trajectory matrix. The step providing the hankel trajectory matrix is of critical importance since the other steps of the SSA technique greatly depend on the structure of this matrix which consequently provides the eigenvalues [Hassani and Mahmoudvand, 2013].

Step 1: Embedding

Embedding can be defined as the mapping operation that transfers a one-dimensional profile Y_N into a multidimensional series $X_1,\ldots,X_K$ with $i = 1,2,\ldots,K$ vectors:

$$X_i = [y_i, y_{i+1}, y_{i+2}, \cdots, y_{i+L-1}]^T, \tag{3.3}$$

T represents transposition. As discussed later in this chapter, there is not a single universal rule for choosing L [Sanei and Hassani, 2015]. According to [Hassani, 2007] and [Golyandina et al., 2001a] L should be large enough to provide a sufficient separation between components but no greater than $N/2$. [Hassani, 2007] also showed that it is practical to choose L proportional to the periodicity and cycle duration of the data,

The outcome of the embedding step is the trajectory matrix **X** which is a Hankel matrix. In **X**, all the elements along the diagonal $i+j = const$ are constant [Hassani, 2007]:

$$\mathbf{X} = [X_1,\ldots,X_K] = (x_{ij})_{i,j=1}^{L,K} = \begin{pmatrix} y_1 & y_2 & y_3 & \cdots & y_K \\ y_2 & y_3 & y_4 & \cdots & y_{K+1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ y_L & y_{L+1} & y_{L+2} & \cdots & y_T \end{pmatrix}. \tag{3.4}$$

Step 2: Singular Value Decomposition

At the second step of the decomposition stage, the singular values of the trajectory matrix **X** are calculated. These singular values which are also known as eigenvalues capture all the information of a profile. To perform the SVD, the matrix $\mathbf{XX}^T$ needs to be found first. This matrix provides us with positive eigenvalues $\lambda_1,\ldots,\lambda_L$ in decreasing order of size. The SVD of $\mathbf{X}$ can be shown as:

$$\mathbf{X} = \mathbf{X}_1 + \ldots + \mathbf{X}_L, \tag{3.5}$$

where $\mathbf{X}_i$ are rank-one bi-orthogonal elementary matrices, $\mathbf{X}_i = \sqrt{\lambda_i} U_i V_i^T$, and $V_i = \mathbf{X}^T U_i / \sqrt{\lambda_i}$. Here, U_i and V_i are also known as principal components which represents the left and right eigenvectors of the trajectory matrix $\mathbf{X}$.

The $\sqrt{\lambda_i}$ is also known as the singular value of $\mathbf{X}$, whilst $\{\sqrt{\lambda}_1, \sqrt{\lambda}_2, \ldots, \sqrt{\lambda}_L\}$ is called the spectrum of singular values. "Singular Spectrum Analysis" is derived from this feature of the SSA process which describes the idea of achieving and analysing the spectrum of singular values for any given time series to distinguish and discriminate between the signal and noise.

3.2.2 Stage 2: Reconstruction

The reconstruction stage corresponds to the analysis of the spectrum of singular values to identify the signal and noise components. The eventual outcome of this stage is a filtered gene expression profile. As it was previously mentioned the second and final choice of SSA which is the number of eigenvalues, r is required to be made at the first step of this stage known as grouping step.

Step 1: Grouping

At the first step of the reconstruction stage the elementary matrices $\mathbf{X}_i$ are split into several groups and summing the matrices within each group. As discussed in [Silva and Hassani, 2015], if we denote $I = \{i_1, \ldots, i_p\}$ as a group of indices $i_1, \ldots, i_p$, then the matrix $\mathbf{X}_I$ corresponding to the group I is defined as $\mathbf{X}_I = \mathbf{X}_{i_1} + \cdots + \mathbf{X}_{i_p}$. The spilt of the set of indices $\{1, \ldots, L\}$ into disjoint subsets $I_1, \ldots, I_m$ corresponds to the representation $\mathbf{X} =$

$\mathbf{X}_{I_1} + \cdots + \mathbf{X}_{I_m}$. The process of choosing the sets $I_1, \ldots, I_m$ is called the grouping. For any given group of I, the contribution of the component $\mathbf{X}_I$ can be estimated by the share of the corresponding eigenvalues: $\sum_{i \in I} \lambda_i / \sum_{i=1}^{d} \lambda_i$. In separating the signal and noise of a profile, one then considers two groups of indices, $I_1 = \{1, \ldots, r\}$ and $I_2 = \{r+1, \ldots, L\}$ and associate the group $I = I_1$ with the signal component and the group I_2 with noise.

The grouping step should be regarded as the most important step of SSA which determines the accuracy of the analysis. Therefore, several options are introduced to assist the analysis for differentiating between signal and noise in a given profile including the possibility of examining the periodogram, scatterplot of right eigenvectors or the eigenvalue functions graph (see, [Hassani, 2007] or [Sanei and Hassani, 2015]). Once the selection of eigenvalues corresponding to signal and noise is made, the effectiveness of the separation needs to be evaluated. This can be performed by using a statistical technique known as the weighted correlation (*w*-correlation). This technique is further discussed under section 3.5.1 of this chapter.

Step 2: Diagonal averaging

The last step of SSA technique transforms a matrix into a Hankel matrix which can afterwards be converted to a new series. Following [Sanei and Hassani, 2015] this process can be briefly explained as follows:

If we consider the z_{ij} for an element of a matrix $\mathbf{Z}$. Then, the k-th term of the resulting series is calculated by averaging z_{ij} over all i, j such that $i + j = k + 1$. By performin the diagonal averaging over all matrix components of $\mathbf{X}_{I_j}$ in the expansion of $\mathbf{X}$ above, another expansion of: $\mathbf{X} = \widetilde{\mathbf{X}}_{I_1} + \ldots + \widetilde{\mathbf{X}}_{I_m}$, where $\widetilde{\mathbf{X}}_{I_j}$ is the diagonalized version of the matrix $\mathbf{X}_{I_j}$ is found.

Note that the SVD of the trajectory matrix $\mathbf{X}$ can be represented as:

$$\mathbf{X} = \sum_{i=1}^{d} \sqrt{\lambda_i} U_i V_i^T = \mathbf{X}_1 + \ldots, \mathbf{X}_d = \sum_{i \in I} \mathbf{X}_i + \sum_{i \notin I} \mathbf{X}_i,$$

where $d = \max\{i; i = 1, \ldots, L | \lambda_i > 0\}$ (rank $\mathbf{X} = d$), $V_i = \mathbf{X}^T U_i / \sqrt{\lambda_i}$ $(i = 1, \ldots, d)$, $\mathbf{X}_i = \sqrt{\lambda_i} U_i V_i^T$ and $I \subset \{1, \ldots, d\}$. The noise reduced series is reconstructed by $\mathbf{X}_I = \sum_{i \in I} \mathbf{X}_i$ after selecting a set of indices I. However, $\mathbf{X}_I$ does not have a Hankel structure and is not the trajectory matrix of some time series. By performing diagonal averaging over the diagonals $i + j = const$ which corresponds to averaging the matrix elements over the 'antidiagonals' $i + j = k + 1$, the aforementioned issue is overcome: the choice $k = 1$ gives $y_1 = y_{1,1}$, for $k = 2$, $y_2 = (y_{1,2} + y_{2,1})/2$, $y_3 = (y_{1,3} + y_{2,2} + y_{3,1})/3$ and so on. Applying diagonal averaging to the matrix $\mathbf{X}_I$ yeilds a reconstructed signal s_t, and provides the SSA decomposition of the original series y_t as follows $y_t = s_t + \varepsilon_t$ $(t = 1, 2, \ldots, N)$, where ε_t is the residual series following signal extraction.

3.3 On the Selection of L and r in SSA

As will be discussed in more details in chapter 5, improving the performance of the SSA (SSA optimisation) can occur at different levels and in various steps which greatly depends on the aim of the study. For example, if increasing the accuracy of signal extraction is the main objective of the study, the parameter selection step should be conducted with regards to some precision criteria. However, in some studies, the priority is given to reducing the amount of time that running a program takes. Therefore, in such cases, one might deliberately choose a less accurate, but faster algorithm.

Given the critical role of L and r in SSA technique, several noteworthy studies have attempted to introduce a particular theoretical justification for choosing these parameters (see, for example, [Golyandina et al., 2001b]). Since the selection of the window length L

significantly depends on the structure of the data as well as the purpose of the analysis, such studies could only bring several suggestions which are only practical for a particular type of data like a seasonal time series (see, for example, [Golyandina et al., 2001b] and [Hassani et al., 2009a]). Therefore, there is not a universally best method for L and r selection. The situation is even vaguer when it comes to the application of SSA in gene expression analysis. Because in analysing gene expression profiles, the pattern of the profiles is significantly diverse between different genes. Moreover, in gene expression analysis, we deal with a noticeable amount of noise in the profiles which makes the problem even more complex.

Therefore, defining a value for L for a particular study is both critical and challenging. Although there is some recommendations on selecting $L = N/4$, [Elsner and Tsonis, 1996], the most common practice is to select the window length fairly large but not greater than $N/2$ [Ghodsi et al., 2009; Golyandina et al., 2001a; Hassani, 2007].

When L is too large, the covariance matrix of the L variables is calculated with only a few observations. This, in turn, extends the imprecision of the result [Hassani et al., 2011a]. Moreover, a too large L results in mixing some parts of the noise with the signal.

However, a rather small L leads to convey some parts of the signal to the noise [Golyandina et al., 2001a]. According to [Hassani, 2007], in a seasonal or periodic time series with an integer period, to achieve a better separation of the periodic components, L should be selected proportional to the period present in the data.

According to the outcome of a simulation study, Golyandina et al. claimed that setting L close to the half of the time series length gives the optimal signal-noise separation [Golyandina et al., 2001a]. However, [Atikur Rahman Khan and Poskitt, 2013] showed that this claim is not constantly valid.

It should be noted that in the selection of L, the criterion of comparison is of critical importance. [Hassani et al., 2011a] and [Hassani et al., 2012] suggested this criterion can be the separability between signal and noise. Hence, setting $L = [\frac{N+1}{2}]$, the minimum value for the w-correlation statistic which is a natural measure of the similarity between two series can be attained [Hassani et al., 2011a]. However, according to [Atikur Rahman Khan and Poskitt, 2013] setting L much shorter than the upper bound $N/2$ results in an improved SSA forecasts. In other words, $L \leq\leq\leq\leq N/2$ and $L = (\log N)^c$, where $c > \log(2)/\log\log(N)$ [Atikur Rahman Khan and Poskitt, 2013].

The selection of the correct number of eigenvalues r which has a direct effect on the reconstruction stage is another open question in working with SSA.

Hassani et al. noted setting r greater than what it accurately should be, enters some levels of noise into the reconstructed series. However, a smaller estimation for r in comparison with its exact value results in losing some parts of the signal [Hassani and Mahmoudvand, 2013].

Also, analysing the scree plot and pairwise scatter plots are suggested to be a helpful guide to select r [Hassani, 2007]. Having said that, according to [Atikur Rahman Khan and Poskitt, 2013], there are no established statistical decision rules for using these approaches.

So far, we have highlighted the importance of the accuracy of gene expression signal processing studies and the critical role of L and r in improving the quality of those studies.

Accordingly, in this thesis, the process of enhancing and adjusting the SSA technique to analyse the segmentation gene profiles is organised as follows:

- SSA based on the minimum variance estimator is presented in Chapter 4. The MV estimator is the optimal linear estimator, which gives the minimum total residual power [De Moor, 1993].

- In Chapter 5, several approaches for optimising the SSA algorithm including SSA hybrid algorithm with parametric and non-parametric signal processing models, identifying L and r according to the foundations of genetics Colonial Theory are proposed. Moreover, a new approach for determining the eigenvalues related to the signal of different gene expression profiles is presented. The latter method is based on the distribution of the eigenvalues of a scaled Hankel matrix.

3.4 Benchmark Models

In this thesis, the performance of the SSA technique is compared with those of several popular, benchmark models. As such, these models are concisely described below and in doing so [Ghodsi et al., 2015] is mainly followed.

3.4.1 Autoregressive Integrated Moving Average

An optimized version of the ARIMA model is provided through the forecast package in R. Referred to as *auto.arima*, a detailed description of the algorithm can be found in [Hyndman and Khandakar, 2007]. In brief, the number of differences is defined as d, and this may be determined using either a KPSS test, Augmented Dickey Fuller (ADF) test or the Phillips-Perron test. The algorithm then minimises the Akaike Information Criterion (AIC) to determine the values for the order of autoregressive terms p, and the order of the moving average process q. The optimal model is chosen to be the model which represents the smallest AIC. The decision on the inclusion or exclusion of the constant c is made depending on the value of d.

To expand on the above summary, we provide the following modelling equations for ARIMA based on [Hyndman and Athanasopoulos, 2014]. A non-seasonal ARIMA model may be written as:

$$(1-\phi_1 B-\ldots\phi_p B^p)(1-B)^d y_t = c+(1+\phi_1 B+\ldots+\phi_q B^q)e_t, \tag{3.6}$$

or

$$(1-\phi_1 B-\ldots\phi_p B^p)(1-B)^d (y_t-\mu t^d/d!) = (1+\phi_1 B+\ldots+\phi_q B^q)e_t, \tag{3.7}$$

where μ is the mean of $(1-B)^d(y_t, c=\mu(1-\phi_1-\ldots-\phi_p)$ and B is the backshift operator. In the R software, the inclusion of a constant in a non-stationary ARIMA model is equivalent to inducing a polynomial trend of order d in the forecast function. It should be noted that when d=0, μ is the mean of y_t.

3.4.2 Autoregressive Fractionally Integrated Moving Average

The general form of an ARFIMA(p,d,q) model shares the same form as an ARIMA process shown in equations (2). However, in contrast to the ARIMA models, here the d is allowed to take the form of non-integer values. Moreover, there is evidence of ARFIMA models been applied in the medical field, see for example [Leite et al., 2007, 2006]. The ARFIMA model used here is estimated automatically using the Hyndman and Khandakar [Hyndman and Khandakar, 2007] algorithm explained above, and the Haslett and Raftery [Haslett and Raftery, 1989] algorithm for estimating the parameters including d. Moreover, this ARFIMA algorithm combines the functions of $fracdiff$ and $auto.arima$ to automatically select and estimate an ARFIMA model. Initially, the fractional differencing parameter is assumed to be an ARFIMA(2,d,0) model. Thereafter the data are fractionally differenced using this estimated d and an ARMA model is selected for the resulting time series using $auto.arima$. Finally, the full ARFIMA(p,d,q) model is re-estimated using the $fracdiff$ function.

3.4.3 Exponential Smoothing

The ETS technique overcomes a limitation found in earlier exponential smoothing models which did not provide a method for easy calculation of prediction intervals [Makridakis et al., 2008]. The ETS model from the forecast package considers the error, trend and seasonal components along with over 30 possible options for choosing the best exponential smoothing model via optimization of initial values and parameters using the MLE and selecting the best model based on the AIC. A detailed description of ETS can be found in [Hyndman and

Athanasopoulos, 2014]. Those interested in the several ETS formula's that are evaluated through the forecast package when selecting the best model to fit the data are referred to Table 7.8 in [Hyndman and Athanasopoulos, 2014].

3.4.4 Neural Networks

The NN model has been successfully used for gene expression profiling, clustering and also gene identification (see for example [Herrero et al., 2001; Xu et al., 1996]) and this is the first time that this model been applied for gene expression signal extraction. This model chosen is referred to as *nnetar* and provided through the forecast package for *R*. A detailed description of the model can be found in [Hyndman and Athanasopoulos, 2014] along with an explanation on the underlying dynamics. In brief, the *nnetar* function trains 25 neural networks by adopting random starting values and then obtains the mean of the resulting predictions to compute the forecasts. The neural network takes the form

$$\hat{y}_t = \hat{\beta}_0 + \sum_{j=1}^{k} \hat{\beta}_j \psi(x_t^T . \hat{\gamma}_j), \tag{3.8}$$

where x_t consist of p lags of y_t and ^{T} denotes transpose. Then, the function ψ has the logistic form

$$\psi(x_t^{'} . \hat{\gamma}_j) = [1 + exp(-\hat{\gamma}_{j0} + \sum_{i=1}^{p} \hat{\gamma}_{ji} . y_{t-1})]^{-1} j = 1, \ldots, k \tag{3.9}$$

This form of neural networks is referred to as a one hidden layer feed forward neural network model and is the default version in the package. However, we consider the use of multiple hidden layers in order to select the best NN model for this type of data. The nonlinearity arises through the lagged y_t entering in a flexible way through the logistic functions of Eq 3.9. The number of logistic functions (k) included, is known as the number of hidden nodes.

The neural network models in this study are estimated using the automatic forecasting model, *nnetar* which is provided through the forecast package in R. For a detailed explanation on how the *nnetar* model operated, see (http://cran.r-project.org/web/packages/forecast/forecast.pdf). The parameters in the neural network model are selected based on a loss function embedded into learning algorithm. This loss function could be for example Root Mean Square Error (RMSE). It may be noted that in all cases the selected neural network model has only *k*=1 hidden node, *p*=2 lags and we adopt annual difference specifications.

3.5 Metrics

Presented in this section are the various metrics which are adopted to evaluate and compare the results obtained in different approaches applied in this thesis.

3.5.1 w-correlation

According to [Golyandina et al., 2001a], the *w*-correlation statistic shows the dependence between two series and can be calculated as:

$$\rho_{12}^{(w)} = \frac{\left(Y_N^{(1)}, Y_N^{(2)}\right)_w}{\parallel Y_N^{(1)} \parallel_w \parallel Y_N^{(2)} \parallel_w,} \tag{3.10}$$

where $Y_N^{(1)}$ and $Y_N^{(2)}$ are two time series, $\parallel Y_N^{(i)} \parallel_w = \sqrt{\left(Y_N^{(i)}, Y_N^{(i)}\right)_w}$, $\left(Y_N^{(i)}, Y_N^{(j)}\right)_w = \sum_{k=1}^{N} w_k y_k^{(i)} y_k^{(j)} \quad (i, j = 1, 2)$, w_k=min$\{k, L, N-k\}$ (here, assume $L \leq N/2$).

Therefore, if the obtained *w*-correlation is close to 0, this confirms that the corresponding series are *w*-orthogonal and that the two components are adequately separable [Hassani et al., 2009a]. However, if the *w*-correlation between the two reconstructed series is large, this confirms that the series are far from being *w*-orthogonal, and are therefore not very well separable and the components should be considered as one group.

Fig 3.5 depicts the *w*-correlations for an example in which 24 reconstructed components are computed. The 20-grade grey scale from white to black attributes to the absolute values of correlations from 0 to 1.

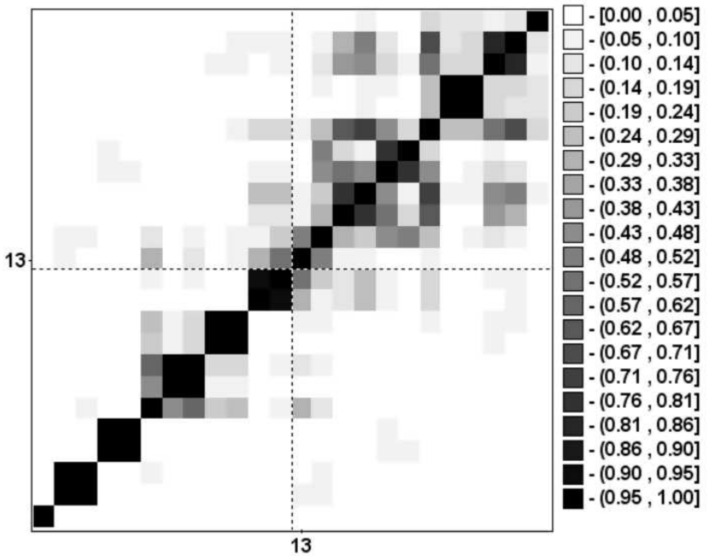


Fig. 3.5 An example of *w*-correlations Matrix for 24 reconstructed components.

3.5.2 Mean Square Error

Mean Square Error (MSE) is a widely used measure to assess the quality of an estimator. The MSE always yields a non-negative value. The closer measure of MSE to zero defines the better performance of the estimator.

It is more common to use the Root mean square error (RMSE) which simply is the square root of MSE. RMSE is also a popular measure of accuracy to find the average of the squares of the errors or deviations. RMSE has the same unit of measurement as the quantity under study.

In comparing two estimators (e.g. two noise reduction methods in this study), it is advised to adopt the Ratio of Root Mean Square Error (RRMSE) [Hassani et al., 2009a]. For example:

$$\text{RRMSE} = \frac{RMSE(SSA)}{RMSE(AlternativeModel)} = \frac{\left(\sum_{i=1}^{N}(s_i - \widehat{s_i})^2\right)^{1/2}}{\left(\sum_{i=1}^{N}(s_i - \widetilde{s_i})^2\right)^{1/2}}$$

where, $\widehat{s_i}$ are the estimated values of s_i, obtained with *SSA* and $\widetilde{s_i}$ are the estimated values of s_i obtained by the alternative model and N is the series length. If RRMSE < 1, then the *SSA* procedure outperforms the alternative method. Conversely, RRMSE > 1 would indicate that the performance of the corresponding *SSA* procedure is worse than the competing method.

3.5.3 Mean Absolute Percentage Error

Mean Absolute Percentage Error (MAPE) also known as Mean Absolute Percentage Deviation (MAPD) is a well-known measure of accuracy in statistics. Similar to RMSE, MAPE is also widely used in trend estimation, signal processing and forecasting studies. In comparison studies, a lower MAPE value indicates a better signal extraction performance. MAPE represents the accuracy as a percentage:

$$\text{MAPE} = \frac{1}{N}\sum_{t=1}^{N}|100 \times \frac{y_T - \widehat{y}_T}{y_T}|,$$

where y_T represents the actual noisy data, and $\widehat{y}_T$ is the reconstructed point obtained via a particular noise filtering method. Nevertheless, this measure of accuracy is a popular comparison criterion, its application presents a major drawback as it cannot be applied in outcomes with zero values because in that case there would be a division by zero.

3.5.4 Pearson coefficient of skewness

The Pearson coefficient of skewness (Also known as the Pearson mode skewness) estimates the skewness of a given distribution [Joanes and Gill, 1998]. This quantity is based on the notion of the moment of the distribution and is defined as:

$$skewness = \frac{\bar{x} - mode}{S} \tag{3.11}$$

where $\bar{x}$ is the arithmetic mean and S is the standard deviation. Apparently, the skewness of a normal distribution is equal to zero.

3.5.5 Coefficient of Variation

The Coefficient of Variation (CV), also known as relative standard deviation (RSD) is a measure of dispersion of a frequency distribution and is defined as the ratio of the standard deviation over mean:

$$CV = \frac{SD}{\bar{x}} \tag{3.12}$$

CV is usually presented as a percentage and is only computed for a data with non-negative ratio scale of the values.

3.5.6 Pearson correlation coefficient

The pearson correlation coefficient, also known to as the Pearson's r or pearson product-moment correlation coefficient, is a measure of the linear dependence between two variables X and Y [Benesty et al., 2009].

$$\rho_{(X,Y)} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} \tag{3.13}$$

where cov is the covariance, σ_X is the standard deviation of X and σ_Y is the standard deviation of Y.

3.5.7 Kendall rank correlation coefficient

Kendall rank correlation coefficient, is a statistic applied to measure the ordinal association between two measured quantities. Kendall is a non-parametric hypothesis test for statistical dependence based on the tau coefficient [Kendall, 1938].

Let $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ be a set of observations of the joint random variables X and Y respectively, and all the values of (x_i) and (y_i) are unique. Any pair of observations (x_i, y_i) and (x_j, y_j), where $i \neq j$, are said to be concordant if the ranks for both elements agree: that is, if both $x_i > x_j$ and $y_i > y_j$; or if both $x_i < x_j$ and $y_i < y_j$. They are said to be discordant, if $x_i > x_j$ and $y_i < y_j$; or if $x_i < x_j$ and $y_i > y_j$. If $x_i = x_j$ or$y_i = y_j$, the pair is neither concordant nor discordant [Kendall, 1938]. The Kendall τ coefficient is defined as:

$$\tau = \frac{(number of concordant pairs) - (number of discordant pairs)}{n(n-1)/2} \tag{3.14}$$

3.5.8 Spearman correlation coefficient

The Spearman correlation coefficient is defined as the Pearson correlation coefficient between the ranked variables [Spearman, 1904]. For a sample of size n, the n raw scores X_i, Y_i are converted to ranks rgx_i, rgY_i and r_s is calculated following:

$$r_s = \rho_{rgX,rgY} = \frac{cov(rgX, rgY)}{\sigma_{rgX} \sigma rgY} \tag{3.15}$$

where, ρ is the pearson correlation coefficient which is applied to the rank variables. $cov(rgX, rgY)$ is the covariance of the ranked variables and σ_{rgX} and σrgY are the standard deviations of the ranked variables.

Chapter 4

Bicoid signal extraction using SSA based on minimum variance

4.1 Introduction

Having discussed the models and techniques which have been previously proposed for studying the Bcd profile in Chapter 2, here, we introduce a new signal processing algorithm for filtering the Bcd profile and extract the signal consequently.

The new method is an enhanced version of SSA technique and is known as SSA based on the minimum variance estimator (SSA$_{MV}$). SSA$_{MV}$ was firstly introduced in [Hassani, 2010] where it was shown to be more efficient than the basic version of SSA [Hassani, 2010].

Accordingly, in this study, to improve the precision of signal extraction and noise filtering of gene expression profiles, we adopt SSA$_{MV}$ algorithm and evaluate its performance for *bcd* gene expression profile.

At the outset, it is important to note that we use the SDD model as the overall benchmark as it is the most commonly used approach for signal extraction in Bcd.

The detail of the SSA$_{MV}$ technique is provided in Section 4.2 which follows with the rest of the chapter organised in two phases:

- Comparison study I

- Comparison study II

At the first phase, we start the analysis with evaluating the application of SSA$_{MV}$ on Bcd profile. Recall Fig 1.6 illustrating the Bcd profile achieved using the immunofluorescence technique. As it was apparent in that figure, Bcd profile contains different levels of noise which need to be removed first.

To validate the theoretical results, extensive simulation study is carried out. Analysing the real data is also performed on all cleavage cycles in which Bcd is present in the embryo (i.e. cleavage cycles 10-14(A)). The empirical results are presented and compared with SDD model as the commonly used model for analysing Bcd profile.

At the second phase, we evaluate the use of a range of powerful and popular signal processing techniques from both parametric and nonparametric backgrounds to provide a sound extraction of Bcd signal. Our aim here is to ascertain whether the other signal processing models can provide a more accurate signal for Bcd in comparison to the SDD and SSA$_{MV}$ techniques.

Similar to the first phase, here, the evaluation consists of the analysis on both simulated and real data sets.

The selections of the signal processing models in the second phase representing both parametric and nonparametric methods are important for several reasons. Firstly, as will be discussed, the residuals following signal extraction in Bcd is nonstationary. Secondly, whilst it is well known that parametric models rely on the underlying assumptions of normality and stationarity, interestingly, the most widely used method for Bcd signal extraction, the SDD model is also parametric. Thirdly, as noted in [Hassani et al., 2013c], for the parametric methods, assuming stationarity for the data, the linearity of the model and normality of the residuals can provide only an approximation of the true situation. Therefore, a method that does not depend on these assumptions could be very useful for modelling and extract the signal in Bcd data.

Moreover, the selection of the models in this chapter has also considered former applications in solving signal extraction problems.

As we are interested in ascertaining the applicability, and introducing a suitable signal processing method for signal extraction in gene expression profiles, it is common to consider an Autoregressive Integrated Moving Average (ARIMA) model as a benchmark for comparing with the other signal processing models. In addition, ARIMA-based models have been applied for signal extraction in various fields both historically and recently (see, for example, [Alexandrov et al., 2012; Burman, 1980; Gonzalez-Romera et al., 2006; Halliday, 2004; Harvey and Koopman, 2000].)

It should be noted that Autoregressive Fractionally Integrated Moving Average is mainly recognised as a method suitable for long memory processes where the decay is slower than in an exponential process. However, the results justify the consideration given to an ARFIMA model as explained later. The use of state space models such as Exponential Smoothing (ETS) does not require added justification as the benchmark model SDD in itself follows and exponential curve [Liu et al., 2011]. Furthermore, the ETS method used here fits the best model for signal extraction by evaluating between both Holt-Winters and Exponential Smoothing. Neural network (NN) models are increasingly popular in the modern world and have been applied for solving signal extraction problems in the past (see for example, [Halliday, 2004; Lee et al., 2003]).

Accordingly, the signal processing models used in the second phase of this chapter include an optimised version of ARIMA, ARFIMA, ETS and NN model with one hidden layer. These models are fully described in Chapter 3.

In what follows, we show how the selected signal processing techniques compare and compete against each other, and the widely accepted SDD model. Any efforts at finding a

Model	Nature
SDD	Parametric
ARIMA	Parametric
ARFIMA	Parametric
ETS	Non-parametric
NN	Non-parametric
SSA	Non-parametric

Note:SDD – Synthesis Diffusion Degradation, ARIMA – Autoregressive Integrated Moving Average, ARFIMA – Autoregressive Fractionally Integrated Moving Average, ETS – Exponential Smoothing, NN – Neural Networks, SSA – Singular Spectrum Analysis.

Table 4.1 Parametric or non-parametric nature of algorithms.

universally optimal model for this purpose would require more extensive research which considers a wide range of filtering techniques and that objective is beyond the mandate of this research.

4.2 SSA based on minimum variance

The theory underlying SSA$_{MV}$ technique is given below. In doing so we mainly follow [Hassani, 2010] where a more detailed description is made available.

Consider a noisy signal vector $Y_N = (y_1, \ldots, y_N)^T$ of length N.

$$Y_N = S_N + E_N \tag{4.1}$$

where S_N represents the signal component and E_N the noise component. Let $K = N - L + 1$, where L is some integer called the window length (we can assume $L \leq N/2$). Define the so-called "trajectory matrix":

$$\mathbf{X} = (x_{ij})_{i,j=1}^{L,K} = \begin{pmatrix} y_1 & y_2 & y_3 & \cdots & y_K \\ y_2 & y_3 & y_4 & \cdots & y_{K+1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ y_L & y_{L+1} & y_{L+2} & \cdots & y_N \end{pmatrix} \tag{4.2}$$

The columns X_j of $\mathbf{X}$, considered as vectors, lie in an L-dimensional space $\mathbb{R}^L$. It is obvious that:

$$\mathbf{X} = \mathbf{S} + \mathbf{E} \tag{4.3}$$

where $\mathbf{S}$ and $\mathbf{E}$ represent Hankel matrices of the signal S_N and noise E_N, respectively. The SVD of the trajectory matrix $\mathbf{X}$ can be written as:

$$\mathbf{X} = \mathbf{U}\Sigma\mathbf{V}^T \tag{4.4}$$

where $\mathbf{U} \in \mathbb{R}^{L\times K}$ is the matrix consists of the normalized eigenvector U_i corresponding to the eigenvalue λ_i $(i = 1,\ldots,L)$, $\mathbf{V} \in \mathbb{R}^{K\times K}$, is the matrix contains the principal components defined as $V_i = \mathbf{X}^T U_i/\sqrt{\lambda}_i$, and $\Sigma = \mathrm{diag}(\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_L)$. The diagonal elements of Σ are called singular value of $\mathbf{X}$, and their set is called the singular value spectrum.

The signal subspace methods are based on the assumption that the vector space of the noisy time series (signal) can be split in mutually orthogonal noise and "signal+noise" subspaces. Thus, by adapting the weights of the different singular components, an estimate of the Hankel matrix $\mathbf{X}$, which corresponds to noise reduced series, can be achieved:

$$\mathbf{X} = \mathbf{U}(\mathbf{W}\Sigma)\mathbf{V}^T \tag{4.5}$$

where **W** is the diagonal matrix containing the weights. Now, the problem is choosing the weight matrix **W**. Next we consider the problem of choosing this matrix using different criteria. The SVD of the matrix **X** can be written as:

$$\mathbf{X} = [\mathbf{U}_1 \;\; \mathbf{U}_2] \begin{bmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{bmatrix} \begin{bmatrix} \mathbf{V}_1^T \\ \mathbf{V}_2^T \end{bmatrix} \tag{4.6}$$

where $\mathbf{U}_1 \in \mathbb{R}^{L\times r}$, $\Sigma_1 \in \mathbb{R}^{r\times r}$ and $\mathbf{V}_1 \in \mathbb{R}^{K\times r}$. We can also represent the SVD of the Hankel matrix of the signal $\mathbf{s}_T$ as:

$$\mathbf{S} = [\mathbf{U}_{1s} \;\; \mathbf{U}_{2s}] \begin{bmatrix} \Sigma_{1s} & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \mathbf{V}_{1s}^T \\ \mathbf{V}_{2s}^T \end{bmatrix} \tag{4.7}$$

It is clear that the Hankel matrix **S** cannot be reconstructed exactly if it is perturbed by noise.

4.2.1 Least Square Estimate of S

If the matrix $\mathbf{X}_{L\times K}$ is rank deficient, *i.e.*, rank $\mathbf{X} = r$ and $r < L < K$. The simplest estimate of **S** is obtained when we approximate **S** by a matrix of rank r in the LS sense:

$$\min \parallel \mathbf{X} - \hat{\mathbf{S}}_{LS} \parallel_F^2 \tag{4.8}$$

where $\parallel . \parallel_F$ is the *Frobenius* norm. That is, the LS estimate is obtained by setting the smallest singular value to zero ($\lambda_{r+1} = 0, \ldots, \lambda_L = 0$) in Equation (4.6):

$$\hat{\mathbf{S}}_{LS} = [\mathbf{U}_1 \;\; \mathbf{U}_2] \begin{bmatrix} \Sigma_1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \mathbf{V}_1^T \\ \mathbf{V}_2^T \end{bmatrix} = \mathbf{U}_1 \Sigma_1 \mathbf{V}_1^T \tag{4.9}$$

The $\mathbf{S}_{LS}$ estimate removes the noise subspace but keeps the noisy signal uncorrelated in the “signal+noise” subspace. The disadvantage of LS is that the performance of the LS estimator is crucially dependent on the estimation of the signal rank r. That is, selecting singular values in LS is a binary approach. The main advantage of the LS estimate is that one does not need to consider any assumptions either about the signal or the noise.

4.2.2 MV Estimate of S

The aims of the noise reduction can be considered as follows: (1) separate the “signal+noise” subspaces from the noise-only subspace; (2) remove the noise-only subspace; (3) ideally, remove the noise components in the “signal+noise” subspaces. The first two steps can be achieved by the least squares estimate, while the MV estimate allows us to achieve the third aim as well. However, one should consider some assumptions to obtain the MV estimate [Hassani, 2010] and if the assumptions are met, one can obtain the MV estimate as follows [De Moor, 1993; Van Huffel, 1993]. Given the matrix $\mathbf{X}$, with rank $\mathbf{X}$ = rank $\mathbf{N} = L$ and also rank $\mathbf{S} = r$. Find the matrix $\mathbf{P} \in \mathbb{R}^{K\times K}$ that minimises min $\| \mathbf{XP} - \mathbf{S} \|_F^2$ and the solution is obtained by $\mathbf{P} = (\mathbf{X^T X})^{-1}\mathbf{X^T S}$. Therefore, the MV estimate of $\mathbf{S}$ is $\mathbf{XP} = \mathbf{X}(\mathbf{X^T X})^{-1}\mathbf{X^T S}$. Using the SVD of the $\mathbf{X}$, we can obtain $\mathbf{XP} = \mathbf{UU}^T\mathbf{S}$.

That is, the MV estimate of $\mathbf{S}$ can be interpreted as an orthogonal projection of $\mathbf{S}$ onto the column space of $\mathbf{X}$ because $\mathbf{UU}^T$ is the associated projection matrix. Note also that rank ($\mathbf{XP}$) = rank ($\mathbf{S}$) $= r$. Let us now consider an alternative form of the SVD of the matrix $\mathbf{X}$ using the SVD of $\mathbf{S}$ Equation (4.7) as follows:

$$\begin{aligned} \mathbf{X} &= \mathbf{S} + \mathbf{N} = \mathbf{U}_{1s}\Sigma_{1s}\mathbf{V}_{1s}^T + \mathbf{N}\mathbf{V}_{1s}\mathbf{V}_{1s}^T + \mathbf{N}\mathbf{V}_{2s}\mathbf{V}_{2s}^T \\ &= \begin{bmatrix} (\mathbf{U}_{1s}\Sigma_{1s} + \mathbf{N}\mathbf{V}_{1s})(\Sigma_{1s}^2 + \sigma_{noise}^2\mathbf{I})^{-1/2} & \sigma_{noise}^{-1}\mathbf{N}\mathbf{V}_{2s} \end{bmatrix} \\ &\times \begin{bmatrix} (\Sigma_{1s}^2 + \sigma_{noise}^2\mathbf{I})^{1/2} & \mathbf{0} \\ \mathbf{0} & \sigma_{noise}\mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{V}_{1s}^T \\ \mathbf{V}_{2s}^T \end{bmatrix} \end{aligned} \tag{4.10}$$

As it appears from Equation (4.10), the middle matrix is diagonal, and the left and right matrices have orthonormal columns. Therefore, Equation (4.10) can be considered as an alternative form of the SVD of **X**, and the singular values of **X** are:

$$\begin{aligned}\Sigma_1 &= (\Sigma_{1s}^2 + \sigma_{noise}^2 \mathbf{I})^{1/2} \\ \Sigma_2 &= \sigma_{noise}\mathbf{I}\end{aligned} \tag{4.11}$$

Hence, the singular values in Σ_2 can be interpreted as a noise threshold, which permits estimating σ_{noise} from Σ_2 in Equation (4.11). We can also consider the following submatrices:

$$\begin{aligned}&\mathbf{U}_1 = (\mathbf{U}_{1s}\Sigma_{1s} + N\mathbf{V}_{1s})(\Sigma_{1s}^2 + \sigma_{noise}^2\mathbf{I})^{-1/2} \\ &= (\mathbf{U}_{1s}\Sigma_{1s} + \mathbf{N}\mathbf{V}_{1s})\Sigma_1^{-1} \\ &\mathbf{U}_2 = \sigma_{noise}^{-1}\mathbf{N}\mathbf{V}_{2s} \\ &\mathbf{V}_1 = \mathbf{V}_{1s} \\ &\mathbf{V}_2 = \mathbf{V}_{2s}\end{aligned} \tag{4.12}$$

Now, the MV estimate of **S**:

$$\mathbf{U}_1\Sigma_1^{-1}(\Sigma_1^2 - \sigma_{noise}^2\mathbf{I})\mathbf{V}_1^T \tag{4.13}$$

4.2.3 Weight Matrix W

Let us consider again the weight matrix **W** based on the LS and MV estimates. As it is appears from Equations (4.9) and (4.13), the left and right singular vectors, $\mathbf{U}_1$ and $\mathbf{V}_1$, of LS and MV estimates are the same, but the singular values are different. The LS and MV estimates can be defined based on the weight matrix $\mathbf{W}_{r\times r}$ as follows:

$$\begin{aligned}\hat{\mathbf{S}}_{LS} &= \mathbf{U}_1(\mathbf{W}_{LS}\Sigma_1)\mathbf{V}_1^T \\ \hat{\mathbf{S}}_{MV} &= \mathbf{U}_1(\mathbf{W}_{MV}\Sigma_1)\mathbf{V}_1^T\end{aligned} \tag{4.14}$$

where

$$\begin{aligned} \mathbf{W}_{LS} &= \mathbf{I}_{r\times r} \\ \mathbf{W}_{MV} &= \mathrm{diag}\left((1-\frac{\sigma^2_{noise}}{\lambda^2_1}),\ldots,(1-\frac{\sigma^2_{noise}}{\lambda^2_r})\right) \end{aligned} \tag{4.15}$$

However, it should be noted than in achieving the obtained weight matrix it has been assumed that the variance of the signal is always greater than the variance of noise.

Next, we compare the performance of the SSA$_{MV}$ and SSA$_{LS}$ on the simulated series drawn from an exponential curve.

The detail of the simulation procedure was provided in chapter 1. The only difference to note here is the number of data points generated per simulated series which is in total 200 data points.

Different values of window length L are considered to examine the sensitivity of the SSA technique for different L. Based on the theoretical results L should be large enough but not greater than $N/2$ [Golyandina and Zhigljavsky, 2013]. Here, for reconstruction stage, the first two eigenvalues are selected. RRMSE is used to compare two noise reduction methods [Hassani et al., 2009a]:

Fig 4.1 shows the RRMSE values obtained for SSA$_{MV}$/SSA$_{LS}$. As it appears, the SSA$_{MV}$ has a better performance in small window lengths where this value tends to 1 as the window length increases showing that both methods have a similar performance for large window lengths. Also, as it can be seen, there is a gradual increase in RRMSE with regards to the window length. For example, for a window length equal to 3, the performance of the SSA$_{MV}$ in series reconstruction is up to 14% better than SSA$_{LS}$. However, there is no significant discrepancy between the performance of SSA$_{MV}$ and SSA$_{LS}$ in window lengths greater than 36.

Having illustrated the detail of SSA$_{MV}$ algorithm and its superior performance over the basic SSA in this study, in what follows, in section 4.3 this method is applied to extract the

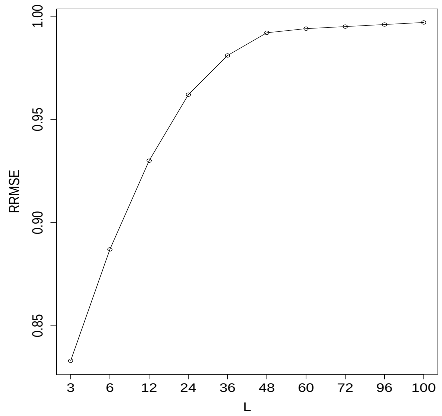


Fig. 4.1 The RRMSE of SSA$_{MV}$/SSA$_{LS}$ in the reconstruction of noisy exponential series.

Bcd signal and is compared against SDD model, while in section 4.4 the comparison study is further extended to a range of parametric and non-parametric signal processing methods.

It is of note that the evaluation is conducted on both simulated and real data separately in each section.

4.3 Comparison study I

4.3.1 Simulation study

Here, a series of simulated data is used to evaluate the performance of two different possible approaches for extracting the Bcd signal from its noisy profile.

In the first approach, the performance of the SDD model is evaluated before and after filtering the series using the SSA technique. Recall that in SDD model $B = Ae^{-x/H}$, is assumed to be the Bcd protein diffusion during segmentation which follows an exponential curve. By estimating the two parameters of this model before and after filtering, one can depict the effect of the noise filtering method on the performance of the SDD model.

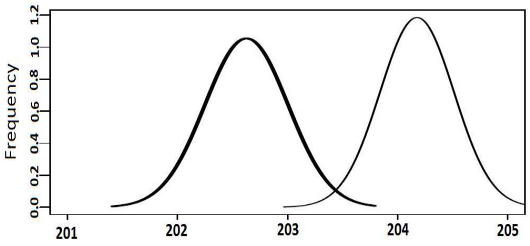
In the second approach, SSA$_{MV}$ is directly applied to the Bcd profiles and its signal extraction capability is compared with SDD model using RMSE criterion.

Let us start the first approach by comparing the performance of the SDD model before and after filtering. Table 4.2 presents the values of the estimated parameters and the related standard deviations. As it appears, by adopting SSA, the estimated values are more robust and accurate than the values obtained before filtering.

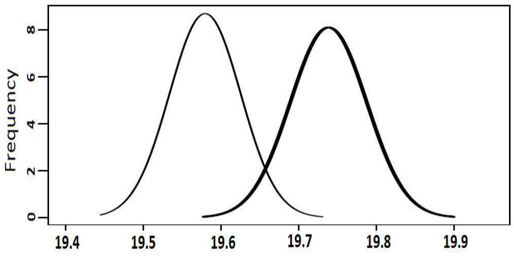
Parameter	**Original value**	**Average**		**S.D**		**Ratio**
		before	**after**	**before**	**after**	
A	200	204.2	202.6	4.19	2.64	$\frac{2.64}{4.19} = 0.63$
H	20	19.57	19.76	0.41	0.26	$\frac{0.26}{0.41} = 0.62$

Table 4.2 The estimated parameters and their standard deviation before and after applying SSA.

Fig 4.2 shows the normal distribution of the estimated parameters A and H before and after filtering. As it appears, parameters show different values in noisy and noise-free series, which confirms that applying a noise reduction method does help the SDD model to give an improved result.



(a) Parameter A



(b) Parameter H

Fig. 4.2 Distribution of the estimated parameters of A and H for noisy Bcd and noise-reduced Bcd (thick lines).

It should be noted that noise has been added to the signal. Therefore, the distribution of the parameters are not a function of noise distribution.

Next, we compare the performance of the SSA$_{MV}$ against the SDD model. In this approach SSA$_{MV}$ and SDD are separately fitted to Bcd profiles and their signal extraction performance is compared using the RRMSE value.

For the SSA analysis, since the length of each simulated series is assumed to be 100 (one observation per egg length percentage), the window length of 50 is chosen to extract the signal and reconstruct the series.

Choosing $L = 50$ and performing the SVD step of the trajectory matrix **X**, 50 eigentriples are obtained, ordered by their contribution (share) in the decomposition stage. We shall say that the series Y_N is not complex if Y_N is well approximated by a series with small rank d. For example series $y_i = e^{\alpha i}$ (i = 1, . . . , N) has rank 1. For all $2 \leq L \leq N-1$, $y_i = by_{i-1}$ where $b = e^{\alpha}$ where α is a model parameter. It should be noted that the number of eigentriples selected as corresponding to the series S_N has to be at least d. For example, if $y_i = e^{\alpha} + \varepsilon_i$ then the window length L should be at least 2 and the first eigentriple is enough for reconstructing the original series if $||S_N|| \gg ||E_N||$. Accordingly, the first eigentriple is selected for filtering and trend extraction.

Furthermore, RRMSE is used to measure the difference between the estimated values (fitted by SDD model and reconstructed by SSA) and the actual values.

Table 4.3 shows the results. As it appears from this table, a significant reduction in the RMSE value is achieved by SSA, confirming that these results are more accurate than the points estimated by SDD model.

For all simulation runs (5 of which are shown here) SSA outperforms the SDD model ranging from 42% to 46%. For instance, regarding the first case of this table, the extracted signal by SSA method is 45% more accurate than the SDD model.

Simulation run	**RMSE**		**RRMSE** ($\frac{SSA}{SDD}$)
	SSA	**SDD**	
1	2.72	4.87	0.55
2	2.71	4.83	0.56
3	2.69	3.98	0.54
4	2.72	4.68	0.58
5	2.73	4.96	0.54

Table 4.3 The RMSE results achieved by SSA and SDD using noisy simulated data.

4.3.2 Real data

Here, the performance of the newly introduced model is evaluated on the real data and is compared against the SDD model. This data is fully described in Chapter 1, where a more detailed information on the characteristics of the data and the data acquisition procedure is made available.

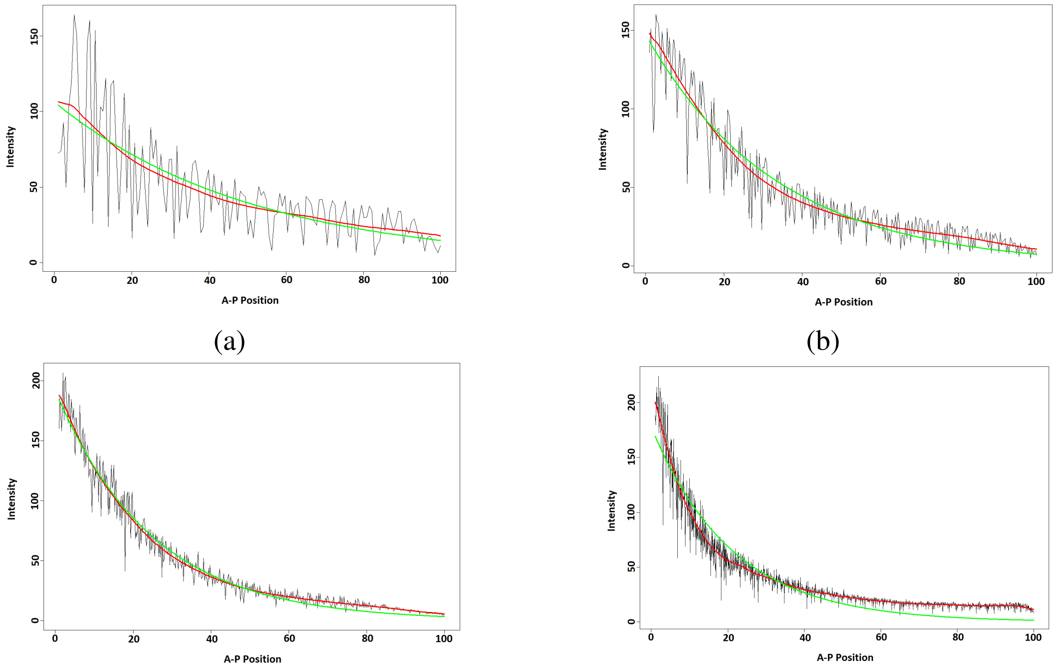
As it was previously noted in Chapter 3, the matrix of the absolute values of w-correlations is a great criterion to decide about the efficiency of a noise filtering technique. Therefore, the calculated w-correlations for different cleavage cycles and time classes are presented in Table 4.4.

If the absolute value of the w-correlations is small, then Bcd signal and its corresponding noise are almost w-orthogonal, but if this value is large, then they are far from being w-orthogonal, and are therefore not very well separable. As it can be seen from the results, the Bcd signal can be effectively separated from the noise component confirming the technique proposed in this study performs well in noise reduction and pattern recognition of Bcd profile.

Class	w-correlation	**Class**	w-correlation
10	0.008	14(3)	0.001
11	0.005	14(4)	0.002
12	0.007	14(5)	0.001
13	0.003	14(6)	0.001
14(1)	0.001	14(7)	0.001
14(2)	0.001	14(8)	0.001

Table 4.4 w-correlations between signal and residuals for *bcd* gene expression profile.

Figs 4.3 and 4.4 show the extracted signal obtained by SSA$_{MV}$ and SDD. The black, red and green colours indicate the original data, the signal extracted by SSA and SDD respectively. Here we can see that the signal component is precisely extracted for cleavage cycles 10-13 and all temporal classes of 14(A), indicating the robustness of the SSA$_{MV}$ to different levels of noise and outliers.



(a)

(b)

(c)

(d)

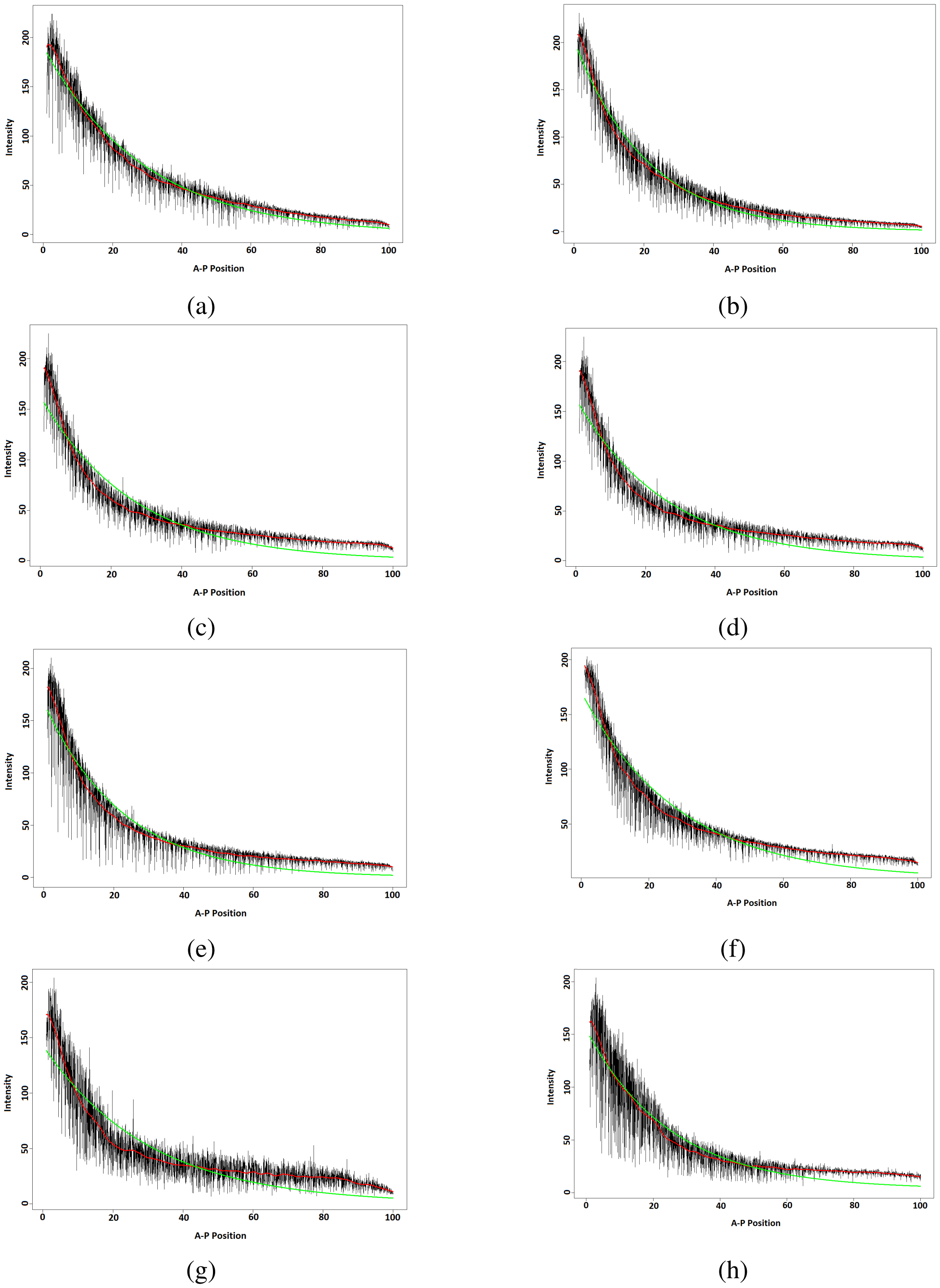
Fig. 4.3 Bcd expression profile in cleavage cycles 10-13. Black, red and green depict the original data, signal extracted using SSA$_{MV}$ and signal extracted using SDD model. a) cleavage cycle 10, b) cleavage cycle 11, c) cleavage cycle 12 and d) cleavage cycle 13

4.4 Comparison study II

4.4.1 Simulation study

In this section, different signal processing models which are fully described in Chapter 1 are fitted to the noisy simulated Bcd profiles.

The following metrics are calculated in order to measure the accuracy of signal extraction:



(a) (b)

(c) (d)

(e) (f)

(g) (h)

Fig. 4.4 Bcd expression profile in cleavage cycle 14(A). Black, red and green depict the original data, signal extracted using SSA$_{MV}$ and signal extracted using SDD model. a) cleavage cycle 14(1), b) cleavage cycle 14(2), c) cleavage cycle 14(3), d) cleavage cycle 14(4), e) cleavage cycle 14(5), f) cleavage cycle 14(6), g) cleavage cycle 14(7) and h) cleavage cycle 14(8)

- Mean Absolute Error (MAE)
- Mean Absolute Percentage Error (MAPE)
- Ratio of the Root Mean Squared Error (RRMSE)

It is of note that the RRMSE has been calculated to measure the accuracy of signal extraction as follows:

$$\text{RRMSE} = \frac{RMSE(AlternateModel)}{RMSE(SDD)},$$

If RRMSE < 1, then the alternate model outperforms the SDD method. In contrast, RRMSE > 1 would indicate that the performance of the corresponding alternate procedure is worse than the SDD model.

Table 4.5 reports the average RMSE values attained by each model following 1000 iterations and some other descriptives relating to the performance of each model.

Model	RMSE	MAPE	MAE	RRMSE	Min	Max
ARIMA	4.69	6.00%	3.21	0.42	3.00	7.70
ARFIMA	5.34	5.80%	3.42	0.48	3.54	10.19
ETS	3.74	4.40%	2.50	0.34	1.89	6.51
NN	6.03	6.27%	4.11	0.55	4.02	11.29
SSA$_{MV}$	2.25	2.00%	1.58	0.20	1.07	3.87
SDD	10.96	23.00%	9.30	N/A	10.92	11.27

Note: RMSE – Root Mean Squared Error, MAPE – Mean Absolute Percentage Error, MAE – Mean Absolute Error, RRMSE – Ratio of the Root Mean Squared Error, Min – Minimum, Max – Maximum.

Table 4.5 Average loss functions and RRMSE for signal extraction using noisy simulated data.

In brief, as it appears from Table 4.5, a significant reduction in the RMSE value is achieved by SSA$_{MV}$, confirming that these results are more accurate than the points estimated by SDD and other models.

Based on this table we can also reach the following conclusions:

Firstly, the parametric SDD model reports the worst performance in comparison to the other parametric and nonparametric models considered in this simulation. The SDD model is outperformed by 58% and 52% by the ARIMA and ARFIMA models respectively whilst the ETS model is 66 % better than the SDD model. The performance of the ARFIMA model is noteworthy mainly because it is famously recognized as a model suitable for long memory processes. These results justify the consideration given to ARFIMA in this study.

Interestingly, the nonparametric feed-forward NN model is the second worst performer and outperforms the SDD model by 45%. It should be noted that we have examined the use of multiple hidden layers and selected a NN model with two hidden layers as the most appropriate for this data based on the lowest RMSE and MAE.

The simulation results indicate that SSA$_{MV}$ provides the best signal extraction and is successful at outperforming the SDD model by 80%. The MAE and MAPE criterions too confirm that SSA$_{MV}$ is the best model in comparison to SDD, ARIMA, ARFIMA, ETS and NN and that SDD is infact the worst performer in this case. The minimum and maximul columns clearly indicate that there is less variation in the results reported by SSA$_{MV}$ and accordingly we can confirm that SSA$_{MV}$ is the most stable model in this case.

In order to confirm the statistical significance of the simulation results, we opted for the nonparametric two sample Wilcoxon test which in this case is able to indicate whether the values of the RMSE's attained via simulation from two given methods actually differ in terms of the size. Based on this test it was confirmed that there exists a statistically significant difference between the RMSE's obtained via SDD and all other models at a p-value of 0.01, under the null hypothesis of equality of RMSE values, further confirming the validity of the results. The p-value is defined as the probability of making the type one error which is rejecting the null hypothesis when it should not have been rejected [Fisher et al., 1949].

Given this superior performance portrayed by the SSA$_{MV}$ technique, we believe it is important to briefly comment on the factor underlying this result. The SSA$_{MV}$ model is a specialised filtering technique with the ability of decomposing a given time series and analysing the eigenvalues for accurately identifying, and separating the noise from the signal.

The appropriateness of the separation between signal and noise obtained via SSA_{MV} was confirmed by the very small values of w-correlation which confirms that the signal and its corresponding noise are almost w-orthogonal.

It is also interesting to note that the minimum and maximum errors reported by SSA_{MV} over the 1,000 simulations are significantly lower than the minimums and maximums reported by the other models. Accordingly, it is clear that the SSA_{MV} technique is more reliable and suitable for signal extraction in Bcd as the average RMSE, MAPE and MAE values are significantly lower than those reported by the other models over the 1,000 iterations.

4.4.2 Real data

The analysis on *bcd* gene expression profile is usually carried on either the whole data set from FlyEx or a sample of embryos introduced by Alexandrov et al. [Alexandrov et al., 2008]. Therefore, having used the FlyEx in the first phase of this chapter, in this section the evaluation is performed on the data which consists of 17 embryos (http://urchin.spbcas.ru/flyex/) and were presented by Alexandrov et al. [Alexandrov et al., 2008]. A complete explanation of the method and biological characteristics of this data can be found in [Alexandrov et al., 2008].

For each embryo the expression level of Bcd protein was measured by using fluorescently tagged antibodies. The nuclear intensities were obtained from a rectangle of 50% of the DV height of the embryo, centred on the AP axis. This data presents the gene expression of the AP coordinate between 20 and 80% egg length. Similar to the first comparison study, this work seeks to extract the signal from one-dimensional gene expression data, hence, the second spatial coordinate (DV axis) has not been considered.

First we seek to extract the signal in the actual data using the various signal processing techniques. The examples of the output from these efforts for embryo *hz29* can be found in Fig 4.5. It is evident that in comparison to the other models, the SSA_{MV} method provides a comparatively smooth signal line. Moreover, it is clear that the SDD signal appears to be least accurate in comparison to the other models even though SDD provides a smooth line as opposed to ARIMA, ARFIMA, and NN models. Overall, the results from the application to real data appear to be consistent with the simulation findings based on what is visible through Fig 4.5.

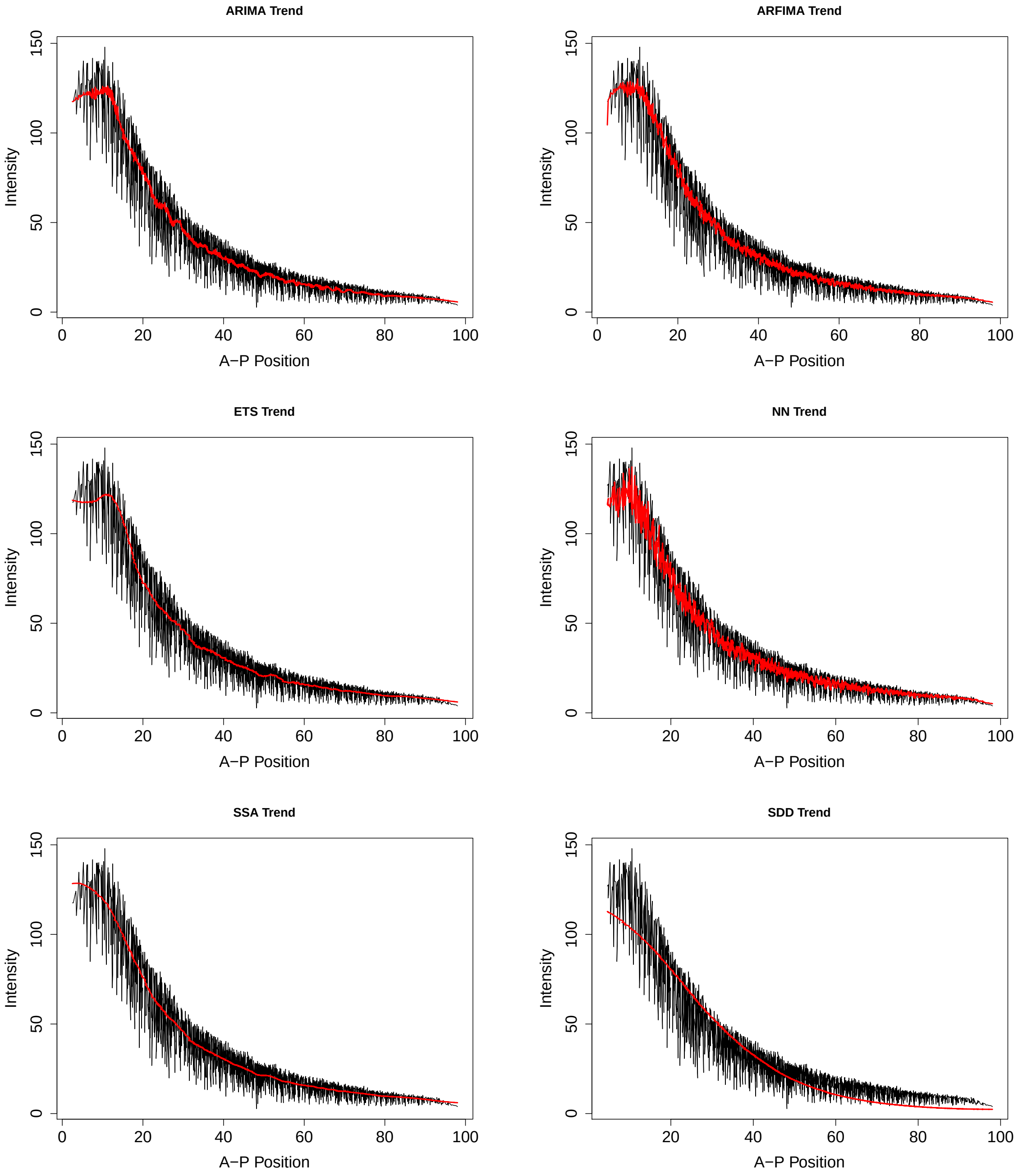


Fig. 4.5 Signal extraction using various time series models and SDD.

Let us now consider the extracted signal lines from each model. A close look at Fig 4.5 suggests that the SDD signal line is the smoothest one out of the evaluated options. However, it is very clear that the SDD signal extraction is also the worst fit in this case as it fails to accurately model the signal amidst the fluctuations, regardless of the fact that it appears to have filtered these fluctuations out. The feed-forward NN model with two hidden layers is showing signs of difficulties with filtering the fluctuations to accurately capturing the signal in Bcd. We can notice similar issues to a certain extent in both the ARIMA and ARFIMA models (with ARFIMA being comparatively worse than ARIMA). In contrast, the ETS and SSA$_{MV}$ signal lines are the most effective in this case. Yet, a close observation of the ETS signal extraction graph makes it evident that in the middle stages the ETS line loses its smoothness to a certain extent. In contrast, the SSA$_{MV}$ model is able to provide a smooth signal line right throughout and thus based on the smoothness of the extracted signal we conclude that SSA$_{MV}$ does indeed capture the signal in Bcd comparatively better than the other methods evaluated here.

Fig 4.6 shows the residuals from each model following signal extraction. Firstly, it is clear from Fig 4.6 that the residuals for the Bcd data following signal extraction is nonstationary. In order to confirm this nonstationary aspect, we tested each of the residuals using the Augmented Dickey-Fuller (ADF) test for unit roots [Cheung and Lai, 1995].

Based on the ADF test is was confirmed that the residuals are in fact nonstationary at a p-value of 0.01. As we have used parametric methods which assume stationarity, we believe it is important to briefly comment on the results obtained through the simulation.

Interestingly, the parametric models of ARIMA and ARFIMA are able to provide a comparatively sound signal extraction for Bcd even with the data being nonstationary (the ARIMA algorithm used in this paper automatically considers taking the number of differences until the series becomes stationary.

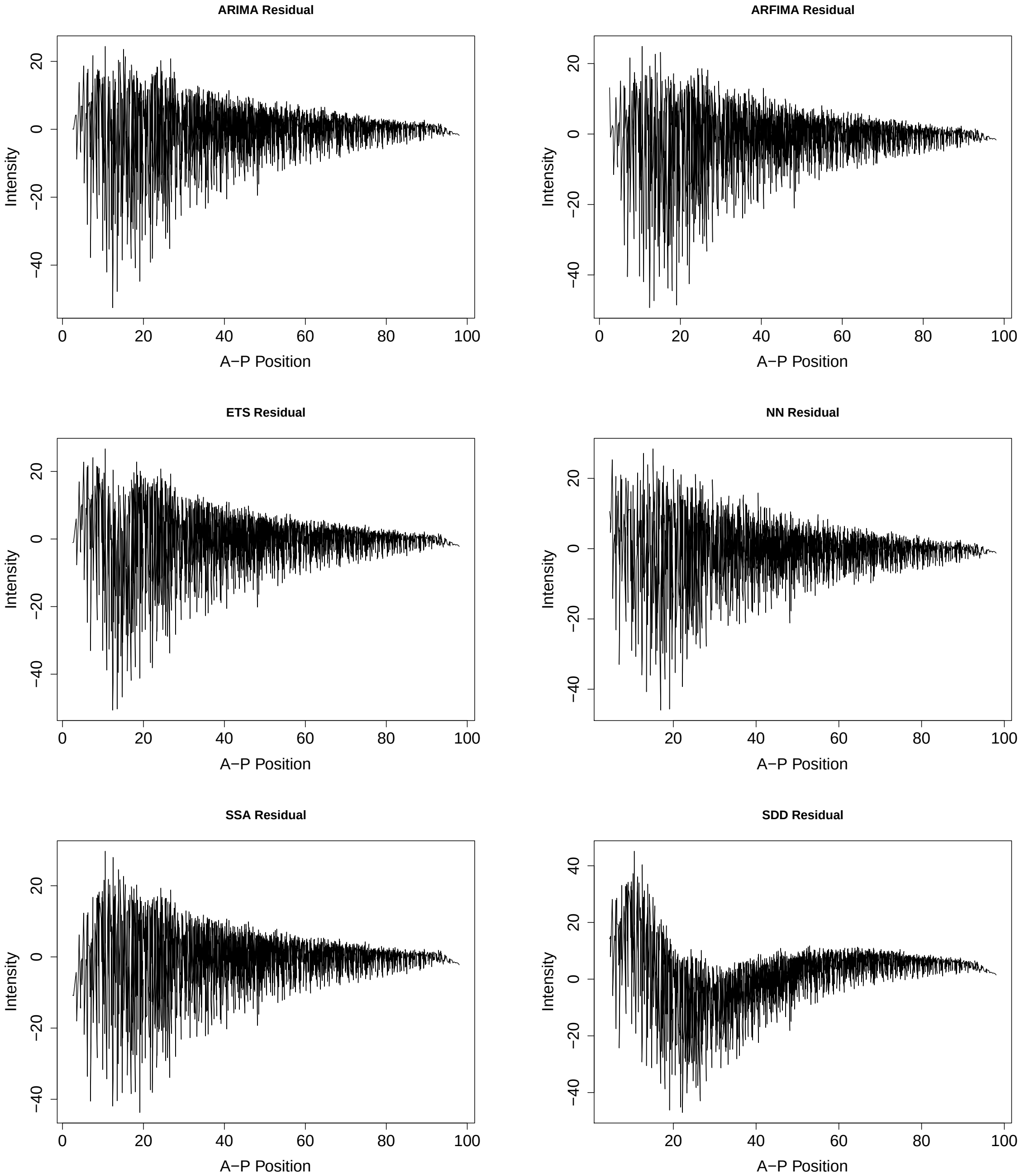


Fig. 4.6 Residuals following Bcd signal extraction.

For the ARFIMA model we evaluated a log transformation which worsened the signal extraction). However, residuals from the widely accepted SDD model does not appear to be white noise and we can see that it has a clear signal. Accordingly, we tested the residuals from the parametric models for white noise using the Ljung-Box test which proved the SDD residuals are not white noise at a p-value of 0.01. This further explains the comparatively mediocre performance shown by SDD when applied to actual data.

Finally, we consider the correlation between the signal and noise extracted from each model in order to analyse the noise separation capabilities. For this purpose we consider correlation between signal and noise based on three different methods; pearsons, spearman and kendall. These results which consider all 17 datasets are reported in Table 4.6. From this table it is evident that all models here have attained a satisfactory level of separation between noise and signal with SSA$_{MV}$ providing correlations below 0.10 in 15 out of the 17 cases. Moreover in general the correlations reported by SSA$_{MV}$ are comparatively smaller with the exception of a few cases.

These results provide further support for the comparatively sound performance that SSA$_{MV}$ has portrayed in extracting the signal in Bcd. It is clear that SSA$_{MV}$'s impressive filtering capabilities are indeed advantageous for this purpose.

4.5 Conclusion

Even though the extraction of Bcd signal appears to be simple, in practice it is an arduous and complicated task as a result of Bcd profile characteristics including lack of normality, being highly volatile and posseisng a heavy tail.

The derived result from this chapter is based on both simulated and actual data application and highlights the following conclusions:

First, SSA$_{MV}$ outperforms the basic SSA and therefore can be adopted as a new analytical algorithm for gene expression signal extraction.

Embryo		ARIMA	ARFIMA	ETS	NN	SSA$_{MV}$	SDD
ac2	pearson	0.1026	0.0716	0.145	0.047	0.035	0.057
	kendall	0.125	0.087	0.065	0.001	0.011	0.176
	spearman	0.146	0.104	0.077	0.005	0.03	0.187
ad36	pearson	0.225	0.147	0.025	0.041	0.051	0.124
	kendall	0.161	0.144	0.054	0.019	0.078	0.452
	spearman	0.207	0.193	0.08	0.022	0.106	0.51
as15	pearson	0.129	0.125	0.125	0.03	0.057	0.161
	kendall	0.088	0.086	0.018	0.012	0.006	0.387
	spearman	0.114	0.11	0.032	0.024	0.007	0.438
as18	pearson	0.129	0.125	0.041	0.03	0.057	0.161
	kendall	0.088	0.086	0.018	0.012	0.006	0.387
	spearman	0.114	0.11	0.032	0.024	0.007	0.438
as19	pearson	0.157	0.13	0.015	0.037	0.034	0.065
	kendall	0.12	0.099	0.018	0	0.054	0.16
	spearman	0.181	0.151	0.024	0.003	0.079	0.203
as22	pearson	0.148	0.124	0.069	0.338	0.03	0.085
	kendall	0.029	0.008	0.126	0.252	0.093	0.235
	spearman	0.03	0.018	0.168	0.388	0.14	0.252
as27	pearson	0.175	0.183	0.053	0.003	0.015	0.195
	kendall	0.105	0.073	0.03	0.014	0.006	0.439
	spearman	0.139	0.101	0.042	0.025	0.015	0.473
cb22	pearson	0.204	0.198	0.025	0.003	0	0.076
	kendall	0.069	0.083	0.014	0.011	0.011	0.159
	spearman	0.086	0.111	0.031	0.023	0.035	0.163
cb23	pearson	0.156	0.092	0.041	0.026	0.028	0.06
	kendall	0.082	0.037	0.051	0.011	0.039	0.157
	spearman	0.12	0.057	0.065	0.01	0.053	0.199
hx8	pearson	0.156	0.092	0.041	0.026	0.028	0.06
	kendall	0.082	0.037	0.051	0.011	0.039	0.157
	spearman	0.12	0.057	0.065	0.01	0.053	0.199
hz19	pearson	0.186	0.206	0.008	0.001	0.038	0.156
	kendall	0.092	0.111	0.014	0.009	0.017	0.335
	spearman	0.141	0.168	0.024	0.007	0.019	0.399
hz20	pearson	0.152	0.086	0.063	0.006	0.005	0.071
	kendall	0.013	0.064	0.177	0.119	0.14	0.123
	spearman	0.02	0.091	0.244	0.162	0.195	0.133
hz29	pearson	0.229	0.188	0.191	0.001	0.074	0.232
	kendall	0.062	0.033	0.116	0.04	0.02	0.463
	spearman	0.09	0.054	0.147	0.042	0.025	0.543
iz4	pearson	0.202	0.23	0.022	0.028	0.003	0.105
	kendall	0.071	0.006	0.108	0.034	0.073	0.21
	spearman	0.086	0.01	0.16	0.052	0.119	0.212
iz13	pearson	0.244	0.039	0.026	0.035	0.027	0.149
	kendall	0.137	0.008	0.029	0.008	0.046	0.334
	spearman	0.178	0	0.037	0.024	0.051	0.387
iz15	pearson	0.228	0.172	0.075	0.029	0.043	0.149
	kendall	0.119	0.115	0.049	0.02	0.006	0.426
	spearman	0.17	0.162	0.056	0.026	0.013	0.523
ms19	pearson	0.207	0.201	0.102	0.028	0.017	0.154
	kendall	0.141	0.106	0.069	0.023	0.049	0.348
	spearman	0.185	0.142	0.096	0.04	0.048	0.385

Table 4.6 Correlation absolute values between Bcd signal and noise in 17 different embryos.

Second, SDD model provides a more reliable signal with more accurate parameters if it applies to a filtered profile. Therefore, if a user is inclined to apply a parametric method such as SDD, it is strongly suggested to reduce the level of noise with any noise reduction method (here for example the basic SSA$_{MV}$ or) before fitting the model to the data.

Third, the obtained results illustrate that the SSA$_{MV}$ filtering method proposed in this study outperforms the SDD model and various other methods considered for filtering noisy Bcd.

The ETS technique was found to be the next best alternative followed by ARIMA. Moreover, given that both parametric and non-parametric algorithms have been evaluated in this investigation, it is pertinent to note that non-parametric algorithms are not explicitly better than parametric algorithms. The performance depends and varies on the nature of the data in question. However, in the case of Bcd signal extraction, we find the SSA$_{MV}$ model (which is nonparametric) outperforming the rest.

Interestingly the simple parametric model of ARIMA was seen outperforming the non-parametric NN algorithm considered in our study. The poor performance of the NN model can be attributed to its proneness to overfitting.

In conclusion, the results confirm that filtering is critical for Bcd curve fitting and the SSA$_{MV}$ technique yields a promising result for Bcd analysis. The consistent superior performance of the SSA$_{MV}$ method over different cleavage cycles and time classes also suggests that SSA$_{MV}$ is a more flexible technique and therefore can be a valuable aid in analysing spatially inhomogeneous noisy data. The feasibility of capturing the signal of the *bcd* gene in *D. melanogaster* embryos suggests that SSA$_{MV}$ may be of general use in evaluating other expressional systems.

Also, comparing with other signal processing methods, using SSA for signal extraction provides the ability to use both dimensions (AP and DV) which expects to give a more reliable result.

Regarding future research it would be insightful to consider various other filtering approaches, such as nonparametric linear filtering, wavelets and the NN models in [Hinton et al., 2006] and [Hinton and Salakhutdinov, 2006] for signal extraction in Bcd.

Chapter 5

Optimising Bcd signal extraction

Signal extraction is critical and at times challenging task in the fields of time series analysis and gene expression signal processing which enables analysts to smooth out a time series by removing the seasonal and cyclical variations and determining the long-run behaviour of the underlying data.

A signal can be formally defined as a smooth additive component which contains information relating to the global change in a time series [Alexandrov, 2008], and the term ‘smooth’ is a vital characteristic of any given signal. In the field of genetics and gene expression in particular, signal extraction and noise reduction are crucial as genetic data is often characterised by the existence of significant noise [Hassani and Ghodsi, 2014].

As outlined in the previous chapters, in achieving the profiles using fluorescence imaging technique, the quantification relies on the assumption that the actual protein concentrations detected by the fluorescence imaging technique are linearly related to the embryo’s natural protein concentration [Surkova et al., 2008a]. However, this is just an estimation and the obtained profiles contain different levels of noise which need to be removed first.

Besides the effect of the observational noise, the Bcd signal extraction process is complex as the data is also associated with biological noise. Moreover, the extracted residual is not normally distributed as required by parametric techniques. The distribution of Bcd follows an exponential trend, and the high volatility depicted in Fig 1.6 of chapter 1, ensures that the extraction of this signal remains an arduous task.

Recall that Bcd is a morphogen localised at the anterior end of the egg. After fertilisation, the distribution of Bcd along the embryo –the signal under study in this chapter– determines the cell's destiny in a concentration-dependent mode.

The findings in chapter 3 which showed SSA as a nonparametric approach produces the most efficient extraction of the Bcd signal (in relation to SDD, ETS, ARIMA, ARFIMA and NN) motivated us to step further in this chapter and enhance this signal processing procedure in different ways.

Accordingly, the aim of this chapter is to introduce and define several new criteria for optimising Bcd signal extraction. At present, there exist no definitive criterion to aid researchers and scientists interested in extracting the Bcd signal for analysis. Characterising the Bcd signal as a critical segmentation gene expects to enhance our knowledge of the developmental processes such as embryogenesis, regional specification and canalisation since the Bcd signal defines what positional information actually is available for morphogen readout. Hence, a small level of noise remained in the extracted signal may considerably affect our understanding of the developmental fate of an embryo. The importance of defining such criteria is further evidenced by the fact that SSA has been applied for extracting the Bcd and other segmentation gene's signal since 2006 (see for example [Holloway et al., 2006; Spirov et al., 2012]). Therefore, it is clear that researcher's and scientists alike can benefit from some formal criteria for the selection of SSA choices when using same for Bcd signal extraction.

As mentioned before, the SSA technique is a nonparametric filtering technique that is highly dependent upon its choice of window length L and the number of eigenvalues r. SSA was successfully introduced for Bcd signal extraction in [Holloway et al., 2006] and exploited in more details in Chapter 4 of this thesis. According to Chapter 4, the residual following signal extraction in Bcd is not normally distributed or stationary, and also the residual itself has a complex pattern which adds further to the difficulty in smoothing and signal extraction.

In this chapter, we present several new approaches for optimising Bcd signal extraction with SSA and provide justification for the process. Accordingly, this chapter presents five different approaches for enhacing Bcd signal extraction:

- Section 5.1 introduces a new criterion to determine *L*.
- Section 5.2 introduces the concept of sequential SSA and its application in gene expression signal processing.
- Section 5.3 evaluates the performance of SSA method in conjunction with parametric and non parametric signal processing techniques.
- Section 5.4 introduces a new approach for eigenvalues identification in extracting the signal from gene expression profiles.
- Section 5.5, inspired by Colonial Theory, introduces a new approach for the grouping step of SSA.

The approaches presented in section 5.1, 5.2, 5.4 and 5.5 are suitable for those who wish to rely on a single model for Bcd signal extraction. We have tailored the criteria presented in this chapter to enable a swift and accurate Bcd signal extraction using the nonparametric approach identified as best in Chapter 4.

Should the extracted signal appear to have captured some unnecessary fluctuations, then the sequential process described in section 5.2 can be applied to the original signal to generate a refined and smoother signal line. Even though the findings in Chapter 4 suggests that the Bcd residual is skewed, we appreciate that statistician who subscribes to classical methods would find it difficult to agree with such outcomes. Therefore, as the third approach, we propose a hybrid parametric signal extraction process which can ensure that the residual is in fact white noise. For those who wish to exploit hybrid modelling from a purely nonparametric perspective with the possibility of capturing the maximum variation via a smooth signal line, we present the hybrid nonparametric approach and show that it can produce far better results when combined with the optimising signal extraction criteria presented herewith.

It should be noted that the set criteria in sections 5.1 and 5.3 are tailored for the sole purpose of extracting an accurate Bcd signal based on the knowledge disseminated through the work in Chapter 4 with regard to the distribution of the residual following Bcd signal extraction. However, the process introduced in sections 5.2, 5.4 and 5.5 can be adopted to extract the signal from a wide ranges of gene expression profiles.

5.0.1 Data

The evaluation in this chapter is performed on 17 *D. melanogaster* embryos introduced by Alexandrov et al. [Alexandrov et al., 2008] which is avaliable via (http://urchin.spbcas.ru/flyex/). The quantitative Bcd data in these wild-type embryos was obtained using the confocal scanning microscopy of fixed embryos immunostained for segmentation proteins [Pisarev et al., 2009]. To that aim, A 1024 × 1024 pixel confocal image with 8 bits of fluorescence data was achieved for each embryo which then transformed into an ASCII table. The ASCII table contains the fluorescence intensity levels attributed to each nucleus in the 10% of longitudinal strips (i.e. only the nuclei correspondent to the central 45-55% of DV axis along the 20-80% of the AP direction). This data is unprocessed for any noise reduction methods. A complete explanation of the method and biological characteristics of this data can be found in [Alexandrov et al., 2008; Surkova et al., 2008a]. The analysis has considered both large and small sample sizes of profiles ranging from $N = 79$ to $N = 2570$.

5.1 On the selection of window length

SSA is a unique technique as it can extract several signals for any given series depending on the chosen value of L. In fact, the choice could be any L such that $2 \leq L \leq N/2$ where N is the length of the series. In this section, to exploit the best L, the number of the eigenvalues correspond to the signal is considered to be one (in some cases, $r = 1,2$). This is due to the general application of SSA in extracting the signal in which the first eigenvalue is considered as the signal component and the remainder as noise. Thereafter, the diagonal averaging is performed to transform the matrix containing the first eigenvalue into a series which will now provide the extracted signal from Bcd.

In this section, to optimise the Bcd signal extraction process using SSA, a new criterion upon making the choice of L is introduced. The proposed criterion is developed as follows.

1) The extracted Bcd signal must be smooth. This is in accordance with the widely accepted definition of a signal which states that it must be a 'smooth' additive component [Alexandrov, 2008].

2) Setting L sufficiently large enables the first eigenvalue, i.e. $r = 1$ (in some cases, $r = 1,2$) to extract a smooth signal for a given series, however the value of L must not be too small or too large. By theory, L must lie between $2 \leq L \leq N/2$ [Sanei and Hassani, 2015]. Yet, when it comes to Bcd signal extraction, setting L at $N/2$ can have negative implications, as with setting L too small.

For example, let us first consider the scenario in Fig 5.1 whereby in a series with length 301 we consider SSA choices of $L = 2$ and $r = 1$ for Bcd signal extraction. Notice how the extracted signal fails to meet the 'smooth' criteria as per the definition of a signal in [Alexandrov, 2008]. Accordingly, it is evident that setting L too small fails to achieve an optimal signal extraction with SSA for Bcd.

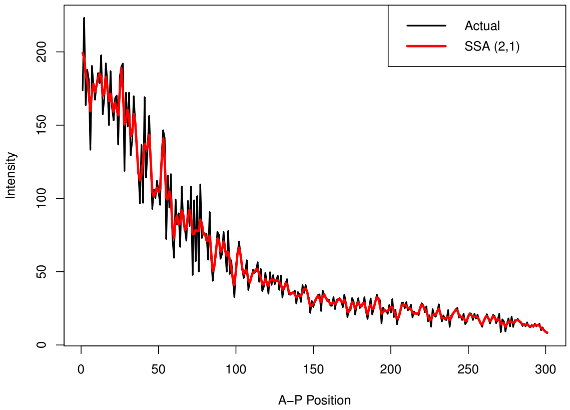


Fig. 5.1 signal extraction from noisy Bcd with SSA choices of $L = 2$ and $r = 1$.

Secondly, let us consider what happens when we set L too large for the same data set. Here, the maximum possible value of L is 150. As such, we set $L = 150$ and seek to extract the signal in our data. Fig 5.2 shows the resulting outcome. In this case, notice how the signal line is smooth (confirming that setting L large can provide a smoother line) but the extracted signal fails to fit well to the actual data especially towards the tail of the series.

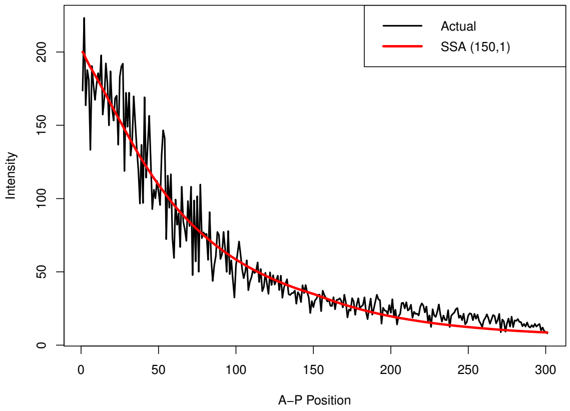


Fig. 5.2 signal extraction from noisy Bcd with SSA choices of $L = 150$ and $r = 1$.

3) Based on points 1) and 2), we suggest the following threshold for the selection of L for Bcd signal extraction purposes. The window length L should be some value between $10 \leq L \leq N/4$. Whilst this assumption helps restrict the selection of L, on its own it fails to provide the researcher with an exact value for L. Therefore, we call upon the nonparametric nature of SSA to provide the final closing argument for the criteria.

4) As a nonparametric technique, the SSA residual can be skewed. Based on the findings in Chapter 4 which is an extensive study into signal extraction in Bcd, the residual from the process is in fact found to be skewed. As such, we propose using the skewness statistic as an indicator, and finding L which corresponds to the minimum skewness for a given Bcd series within the threshold $10 \leq L \leq N/4$.

The presented justifications are further explored and evaluated on the real dataset later in this chapter.

5.1.1 Residual Analysis

In order to save space, via Fig 5.3 we only show the residuals corresponding to the signal extractions shown in Fig 5.4. All remaining residuals are shown in Appendix 2 for those interested. A first look at the structure and distribution of the residual over time helps us to understand the difficulty in extracting the signal from Bcd profiles. This is largely to do with the the highly volatile nature of the data which results in fluctuating amplitudes over time in a particular pattern. In fact, the general patterns appears such that all residuals portray amplitudes which are initially high and then gradually decrease. This in turn means that the techniques adopted for Bcd signal extraction should be able to cope well with such variation and fluctuations in data if it is to accurately perform its task. Moreover, it appears to the naked eye that there is indeed some signal contained within these residuals. Whilst it is expected that a residual following signal extraction would result in capturing the other signals, in some instances there also appears to be a small signal pattern hidden within this data.

However, as visual inspections fall short of providing sound evidence, we also consider some statistics for analysing the residuals further. These are reported via Table 5.1 for all the Bcd data considered in this study. The residuals are initially tested for normality via the Kolmogorov-Smirnov (KS) test for normality. The choice of KS test as opposed to using the popular Shapiro-Wilk (SW) test for normality was because when faced with large samples the KS test is likely to be comparatively more accurate than the SW test [Silva et al., 2016]. As expected, all residuals failed to pass the normality test reporting probability values of less than 0.001, and thereby leading to a rejection of the null hypothesis of normality. This lets us conclude with 99% confidence that the Bcd residuals following signal extraction are in fact skewed and these results are consistent with the findings in Chapter 4.

Finally, we go a step further and fit optimal ARIMA models [Hyndman and Athanasopoulos, 2014] to the residuals. This was done in order to ascertain the randomness of the residuals following Bcd signal extraction with optimised SSA. Statisticians who rely on classical signal extraction techniques would be overly concerned with the parametric assumptions of normality and stationarity of the residuals. Whilst we have assessed the normality of residuals via the KS test and justified based on Chapter 4 that the residuals from this signal extraction exercise should be skewed, fitting of optimal ARIMA models enables us to easily show whether the residuals meet the stationary criteria. We fit automated and optimised ARIMA models (as provided via the forecast package in R) on the residuals and report the outcomes in Table 5.1. If the data is nonstationary, then within the ARIMA(p,d,q) process the value of $d \geq 1$. If the data is stationary, then no differencing is required, and so $d = 0$. In this case, we notice that $d = 0$ in all instances, and thereby proves that the residuals are indeed stationary.

However, the fitting of ARIMA models on the residuals also highlight another interesting point. Notice how for 27 Bcd residuals there have been a variety of 14 different ARIMA models which have been fitted. This in turn indicates the complexity and difficulty associated with the selection of a single technique for extracting Bcd signal, and most certainly highlights the difficulties which any technique would when seeking to extract a signal from data with such complex fluctuations. In addition, except for where the model reads $ARIMA(0,0,0)$, in all other instances we notice that the residuals are not white noise. We discuss this, and provide a possible solution later in this chapter.

Embryo	N	KS	ARIMA
ab2	138	<0.001	ARIMA(0,0,1) with zero mean
hz15	85	<0.001	ARIMA(0,0,0) with zero mean
hz28	79	<0.001	ARIMA(2,0,2) with zero mean
ad14	301	<0.001	ARIMA(2,0,5) with zero mean
ad22	294	<0.001	ARIMA(4,0,3) with zero mean
ad23	308	<0.001	ARIMA(1,0,3) with non-zero mean
ab17	485	<0.001	ARIMA(1,0,3) with non-zero mean
ad4	556	<0.001	ARIMA(4,0,4) with zero mean
ad6	566	<0.001	ARIMA(2,0,2) with non-zero mean
ab12	2284	<0.001	ARIMA(4,0,2) with zero mean
ab10	2263	<0.001	ARIMA(1,0,2) with zero mean
ac5	2404	<0.001	ARIMA(4,0,4) with non-zero mean
ab1	2570	<0.001	ARIMA(4,0,4) with zero mean
ac7	2268	<0.001	ARIMA(1,0,2) with zero mean
ad13	2235	<0.001	ARIMA(4,0,2) with non-zero mean
ad29	2193	<0.001	ARIMA(1,0,2) with zero mean
ad32	2183	<0.001	ARIMA(2,0,1) with zero mean
ab7	2346	<0.001	ARIMA(1,0,2) with zero mean
ac3	2356	<0.001	ARIMA(0,0,1) with zero mean
ac9	2215	<0.001	ARIMA(4,0,1) with zero mean
ms14	2305	<0.001	ARIMA(4,0,2) with zero mean
ab11	2355	<0.001	ARIMA(4,0,2) with zero mean
ac4	2383	<0.001	ARIMA(3,0,1) with zero mean
ab14	2218	<0.001	ARIMA(1,0,2) with zero mean
ab9	2369	<0.001	ARIMA(2,0,1) with zero mean
dq2	2423	<0.001	ARIMA(2,0,4) with zero mean
ms36	2239	<0.001	ARIMA(5,0,1) with zero mean

Note: Embryo: name of the embryo under study. N: length of the Bcd profile. KS: p-values obtained using the KS test of normality.

Table 5.1 Residual analysis for Bcd signal extractions.

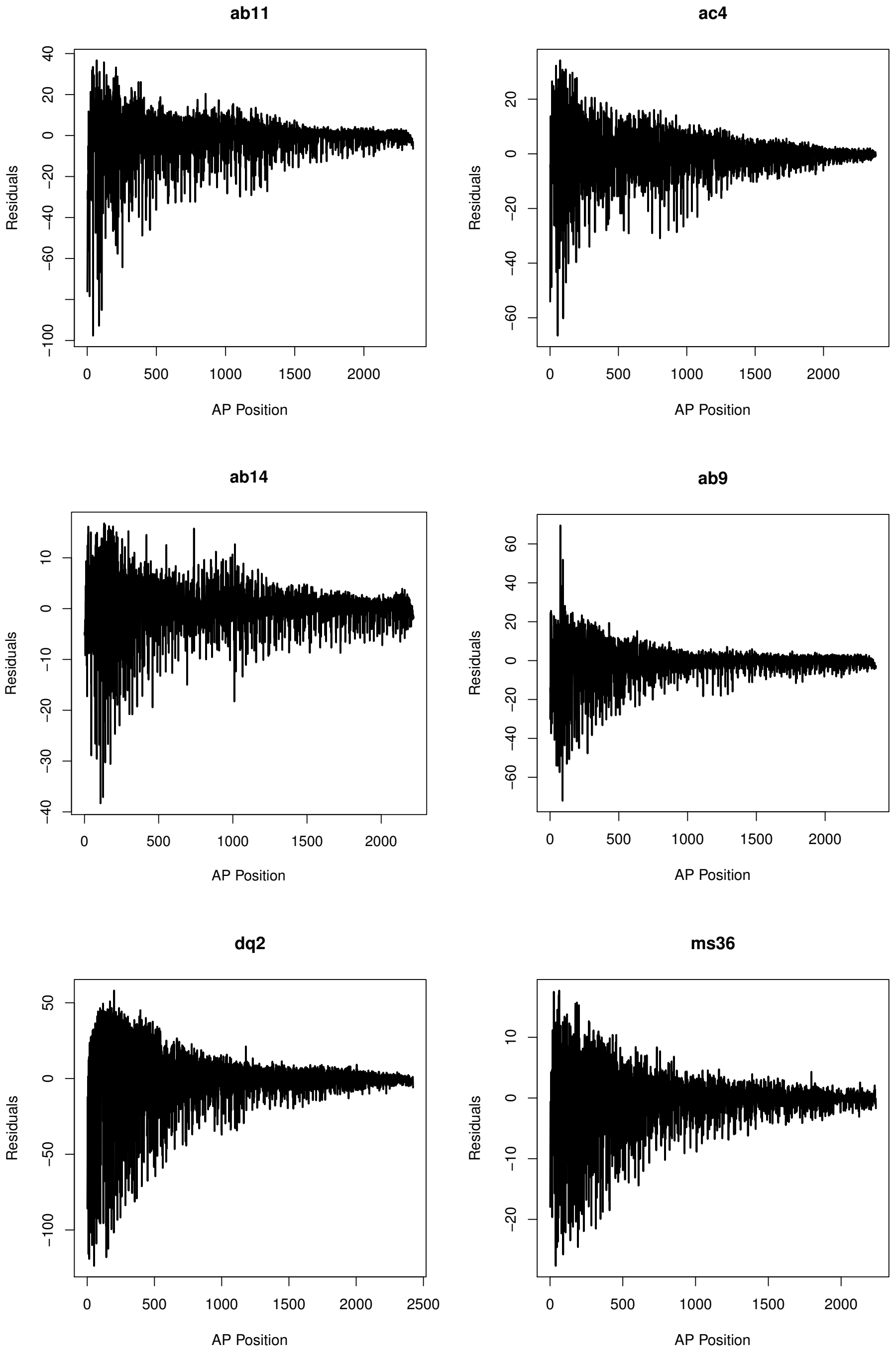


Fig. 5.3 Residuals following optimised signal extraction with SSA for a selection of Bcd data.

Here, we consider real Bcd data and seek to extract the signal with SSA using the newly proposed criterion as outlined in Section 5.1. Fig 5.4 below portrays a selection of the actual data and extracted signal with the optimised SSA algorithm. Table 5.2 outlines the SSA choices which have been used in each case.

Embryo	N	L,r	**Embryo**	N	L,r
ab2	138	25,1	ad13	2235	557,1-2
hz15	85	21,1	ad29	2193	547,1-2
hz28	79	19,1	ad32	2183	544,1-2
ad14	301	11,1	ab7	2346	585,1-2
ad22	294	73,1	ac3	2356	11,1
ad23	308	10,1	ac9	2215	552,1-2
ab17	485	10,1	ms14	2305	575,1-2
ad4	556	139,1	ab11	2355	588,1-2
ad6	566	130,1	ac4	2383	39,1
ab12	2284	570,1-2	ab14	2218	52,1
ab10	2263	564,1-2	ab9	2369	591,1-2
ac5	2404	600,1-2	dq2	2423	604,1-2
ab1	2570	642,1-2	ms36	2239	558,1-2
ac7	2268	566,1-2			

Note: Embryo: name of the embryo under study. N: length of the Bcd profile. L,r: Optimal L and r.

Table 5.2 Optimal L and r found for different embryos under study.

All results for the remaining signal extractions can be found in the Appendix 2. For the examples in Fig 5.4, note how the extracted signal is not only smooth, but also well centred around the data, thereby providing the reader with a very accurate outlook for the long term prospects of the Bcd gradient. However, it is evident that on its own, SSA appears to have difficulties in accurately capturing the signal curve initially when it is faced with very high levels of fluctuations as clearly visible within the first few observations of the Bcd profile. We consider this aspect further in Section 5.3.

Even though signal extraction is the primary focus of this study, it is no secret that the residual can often enlighten us to crucial information pertaining to any given data set. As such, we follow up the signal extractions with a sound residual analysis.

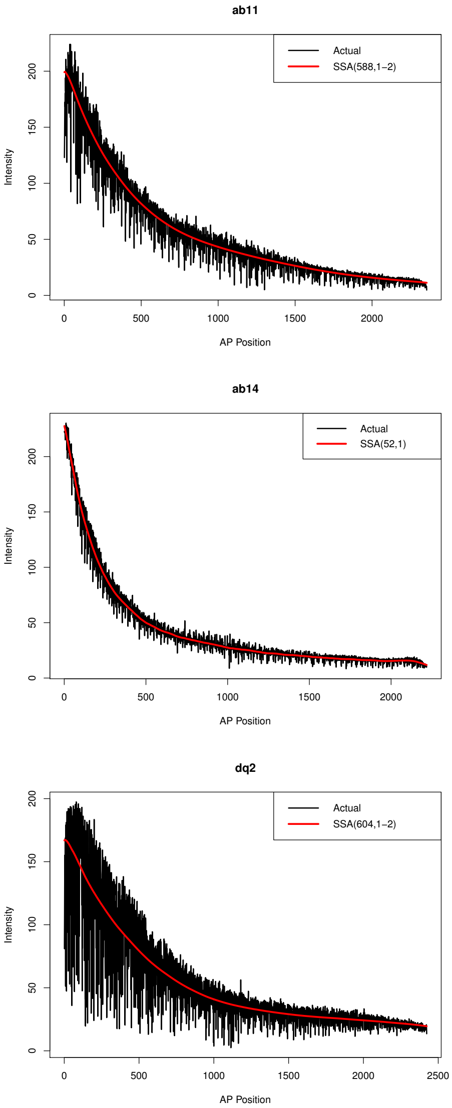


Fig. 5.4 Optimised signal extraction with SSA for a selection of Bcd data.

5.2 Sequential SSA

Signal extraction in Bcd data can be an arduous task owing to the complex structure portrayed by the data. Sequential SSA is a relatively new concept which is of great benefit when faced with weak separability between signal and noise as a result of such complexities. For example, when faced with problems in separating a signal of complex form and seasonality, sequential SSA can be exploited to obtain a more accurate decomposition from the residual after signal extraction [Golyandina and Shlemov, 2013]. Whilst historically, sequential SSA was performed on a residual, in this chapter we suggest the use of sequential SSA for further refining the Bcd signal.

The basic idea underlying sequential SSA is to perform a second round of SSA based decomposition and reconstruction on data that has already undergone an initial round of SSA, with the aim to refine the signal of interest further. Suppose that we exploit the optimised Bcd signal extraction algorithm explained above and extract some signal line. However, if the Bcd data in question has a highly complex structure, it is possible to end up with a signal line that is not as smooth as one would like. In such instances, we suggest exploiting sequential SSA, not on the residual, but on the extracted signal to smooth it further and obtain a new and refined signal curve. This approach is greatly beneficial to those who wish to rely on a single model for Bcd signal extraction and enjoy the benefits of a nonparametric technique.

Note how the signal extraction in ac3 (see, Appendix 2) appears to have captured some other fluctuations apart from the signal alone. As such, this extraction, in particular, fails to meet our criteria for a smooth signal. When faced with such situations, we are able to find a solution via sequential SSA. Sequential SSA enables users to take the extracted signal (the Bcd signal in our example) and filter same with SSA once more to obtain a more refined output. In what follows we have applied sequential SSA on the initially extracted Bcd signal.

As visible via Fig 5.5, following sequential SSA we have been able to extract a smoother signal. In this instance, we used the signal extracted via the optimised SSA signal extraction algorithm for Bcd and refined this signal further via sequential SSA. Here we have used $L = N/2$ and $r = 1$ for signal extraction with Sequential SSA. In line with good practice, the residual was once again tested for normality via the KS test which indicated that the residual is skewed at a 1% significance level, and fitting of an ARIMA model showed that the residual is stationary as well.

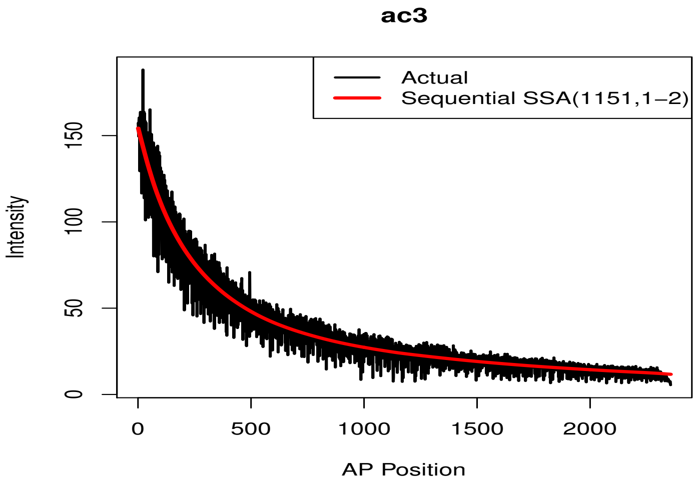


Fig. 5.5 Refined signal extraction with sequential SSA on ac3 signal.

5.3 Hybrid signal Extraction Techniques

It is possible that some statisticians may not be convinced or used to subspace-based methods such as SSA. Therefore, we find it pertinent to present the possibility of obtaining a hybrid signal extraction process which will combine the optimised SSA signal extraction algorithm for Bcd with other automated signal processing techniques from both parametric and nonparametric backgrounds.

The basic idea underlying the hybrid signal extraction process is as follows:

1. Extract the Bcd signal via the optimised SSA signal extraction algorithm.

2. Fit a different signal processing model to the residuals following SSA signal extraction and obtain the fitted values.

3. Add the fitted values to the original SSA signal to create the Hybrid SSA signal.

5.3.1 Hybrid SSA signal: Parametric Approach

The idea underlying hybriding SSA signal extraction method with a parametric approach is to combine the nonparametric SSA signal with the fitted values on residuals from a parametric signal processing model. As most classical statisticians welcome and subscribe to the ARIMA model, here we choose an automated ARIMA model as provided via the forecast package in R [Hyndman and Khandakar, 2007]. This approach is useful as it ensures that the residual following hybrid signal extraction will indeed be white noise.

The residual analysis in Table 5.1 indicates that ARIMA models could be fitted to all but one of the residuals following signal extraction with the optimised SSA signal algorithm. This means that only one of the residuals are pure white noise as it stands. Whilst some might argue that this is acceptable given that the objective is to extract the signal component alone, there may be others who subscribe to an alternate view along the lines of obtaining a random residual following signal extraction. The first hybrid SSA signal approach we present is one which enables users who wish to obtain white noise achieve this following Bcd signal extraction with SSA. We begin by fitting the ARIMA models as identified via Table 5.1 to the data and extract the fitted values which are then combined with our original SSA Bcd signal to create a hybrid SSA-ARIMA signal for Bcd. We consider the examples discussed in text so far and generate the following results. Fig 5.6 shows the hybrid SSA-ARIMA signals for Bcd data. In comparison to the optimised SSA signals in Fig 5.4, the hybrid SSA signal with ARIMA fit fails to meet the smooth criteria. As such, it is evident that on its own, the hybrid SSA-ARIMA approach is only beneficial for those who wish to capture all the signal in the data whilst ensuring that the residual following Bcd signal extraction is white noise. It clearly comes at a cost of lost smoothness in signal curves. However, it is of note that as previously mentioned, noise in gene expression data enters not only from the data acquisition and processing procedures [Wu et al., 2007] but also the fluctuations seen in an expression pattern can be a consequence of biological noise which may also introduce error into the data [Myasnikova et al., 2009]. Therefore, the source of the natural biological variability is different from the experimental noise [Myasnikova et al., 2009]. Biological noise arises from the active molecular transport, compartmentalization, and the mechanics of cell division [Spirov et al., 2012]. Therefore, the hybrid SSA with the ARIMA model can be applied in studies such as segmentation network analysis where the combination of Bcd signal with its biological noise needs to be considered as an input to the system.

5.3.2 Hybrid SSA signal: Nonparametric Approach

Whilst the underlying idea remains the same, in this instance, as opposed to relying on a parametric signal processing model, we can combine the nonparametric and optimised SSA signal with fitted values on residuals from a nonparametric signal processing model in order to obtain the hybrid SSA signal. The benefits of this approach would be that it enables to overcome the parametric restrictions of normality and stationarity of residuals of which the former condition was found to be irrelevant in the case of Bcd data where the residual following signal extraction is skewed according to Chapter 4. In this case we rely on the automated Exponential Smoothing (ETS) model found in the forecast package in R. Those interested in the several ETS formula's that are evaluated through the forecast package when selecting the best model to fit the residuals are referred to Chapter 7, Table 7.8 in [Hyndman and Athanasopoulos, 2014].

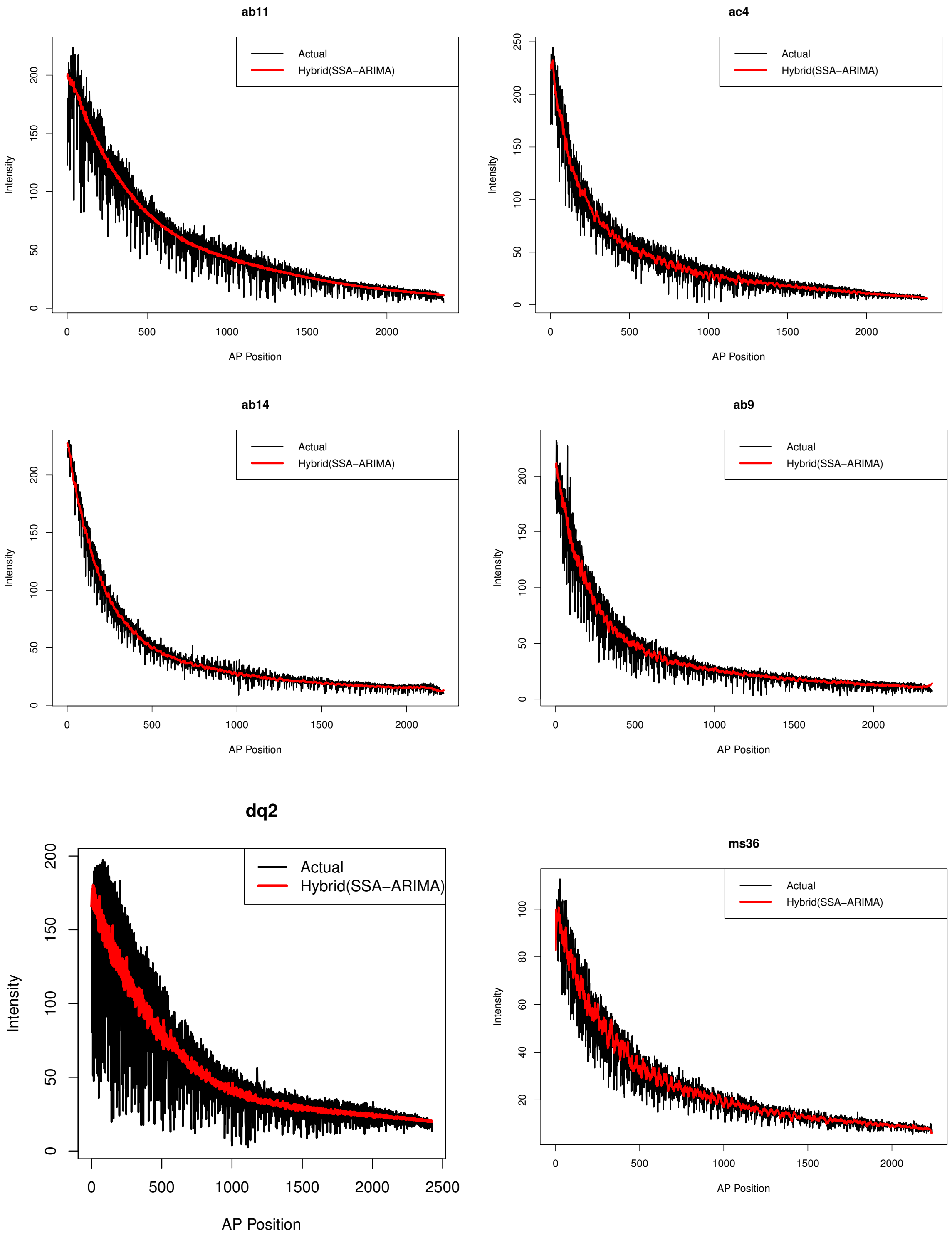


Fig. 5.6 Hybrid SSA signal with ARIMA fit for Bcd data.

Here, we apply the same process as above, but instead of ARIMA, we rely on the nonparametric time series analysis model of ETS. This enables the entire hybrid SSA signal approach to remain nonparametric in nature. The resulting hybrid SSA signals with ETS fit are shown via Fig 5.7.

There are few interesting points to note here. Firstly, in comparison to the parametric hybrid signal extraction approach, it is clear that the nonparametric hybrid approach has resulted in much smoother signal curves as one would expect and like to see following a signal extraction exercise. As such, out of the two hybrid approaches, for the purposes of Bcd signal extraction, it is likely that users will prefer the nonparametric approach over the parametric approach.

Secondly, and perhaps most importantly, the hybrid SSA signal with ETS fit for Bcd data is able to overcome an issue we experienced in the optimised SSA signal extractions. Recall that the optimised SSA signals in Fig 5.4 failed to capture the initial curve in Bcd appropriately. This primary curve attributes to the concentration of Bcd in nuclei at the anterior position along the AP axis which shows that Bcd concentration initially reaches a maximum value before decaying along the embryo. Considering the fact the data used in this study presents only the interval of the AP coordinate between 20% and 80% egg lengths (%EL), detecting the initial curve in this interval suggests that the mechanism of gradient readout might be more complicated than reading a particular concentration of the morphogen.

If we compare the results for the very first observations of Bcd data in Fig 5.7 and Fig 5.4, we will notice how the hybrid nonparametric signals are able to capture the initial curve in Bcd data more accurately than before, thereby providing a refined and improved signal extraction for Bcd. In order to make this clearer for the reader, we present in Fig 5.8 few examples of signal extraction whereby we have zoomed into the initial signal extraction chart areas for both optimised SSA signal and hybrid SSA-ETS signals.

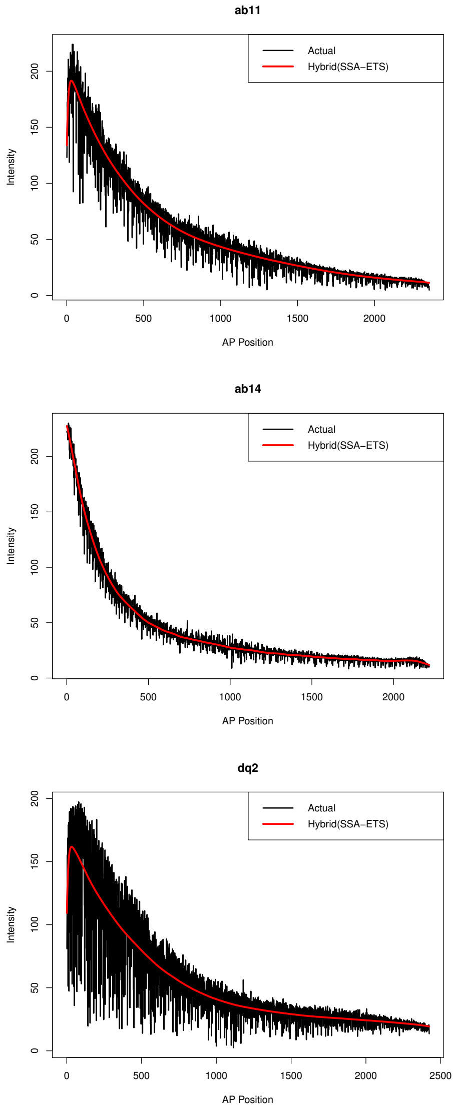


Fig. 5.7 Hybrid SSA signal with ETS fit for bicoid data.

This result is of utmost importance as the problem of accurately fitting the SSA signal to the initial variation in Bcd data was also visible in Chapter 4, and we have succeeded in providing a solution to this issue via this study. Moreover, results not reported here (but available upon request) showed that one will not be able to capture the initial curve using the ETS technique alone. This further portrays a positive aspect of our proposed optimised SSA signal extraction process as it can produce superior results when combined with the smoothing capabilities of ETS.

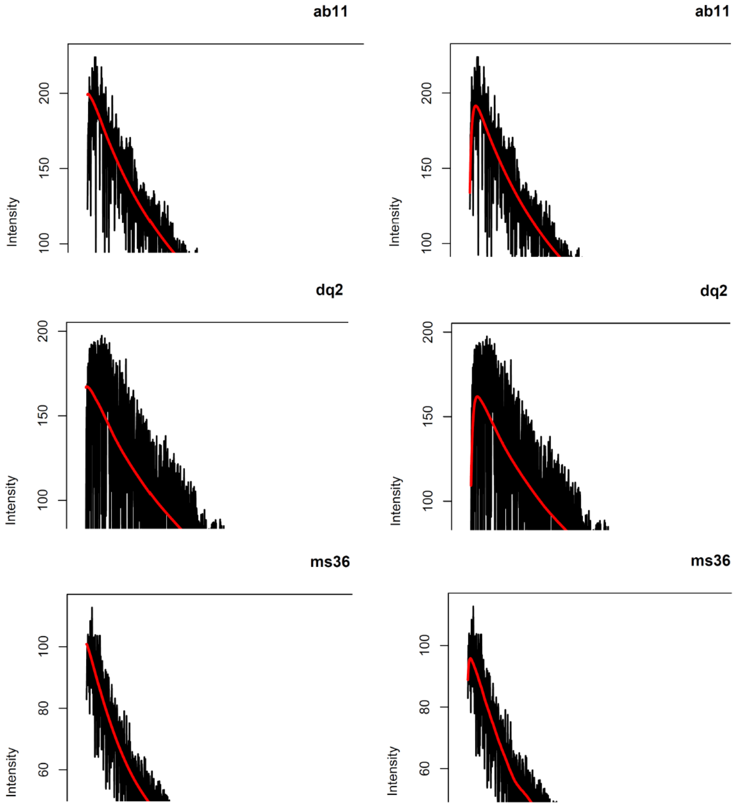


Fig. 5.8 Zoomed in view of optimized SSA signals [left] and hybrid SSA-ETS signals [right].

5.4 Eigenvalues identification approach

As discussed before, due to the presence of noise, finding the expression pattern of the segmentation factors in an embryo is not a simple task [Alexandrov et al., 2008; Houchmandzadeh et al., 2002; Myasnikova et al., 2005]. It is of critical importance that having portrayed the process underlying finding the optimal *L*, in applying the SSA algorithm for signal extraction purposes, there is still an open question related to the identification of the number of eigenvalues required for series reconstruction. Two principal reasons are underlying the essential need for an enhanced algorithm which enables us to decide the appropriate number of eigenvalues related to the signal:

- Improving the signal extraction performance by minimising the signal-noise perturbation.
- finding the optimal number of eigenvalues not only for Bcd but for different gene expression profiles of the SN.

Accordingly, to address this issue, this section provides a new approach to find the number of eigenvalues needed for removing the nonspecific noise from *D. melanogaster* segmentation genes. This approach mainly relies on the distribution of the scaled Hankel matrix eigenvalues.

As it is shown below, gene expression pattern is greatly diverse among the various segmentation genes. Therefore, we expect to find a different number of eigenvalues required for signal reconstruction in each profile. Consequently, given the considerable variation between different gene expression profiles, eigenvalues identification can be considered as the biggest challenge for the signal extraction process using SSA technique.

The performance of the newly introduced signal extraction method is evaluated on four different genes of SN; *bcd*, *cad*, *gt* and *eve* which are among the most important maternal and zygotic segmentation genes. The *bcd* gene is maternal, *cad* has both maternal and zygotic origin and *gt* and *eve* are respectively related to gap and pair rule category of zygotic genes [Holloway et al., 2006; Surkova et al., 2008a].

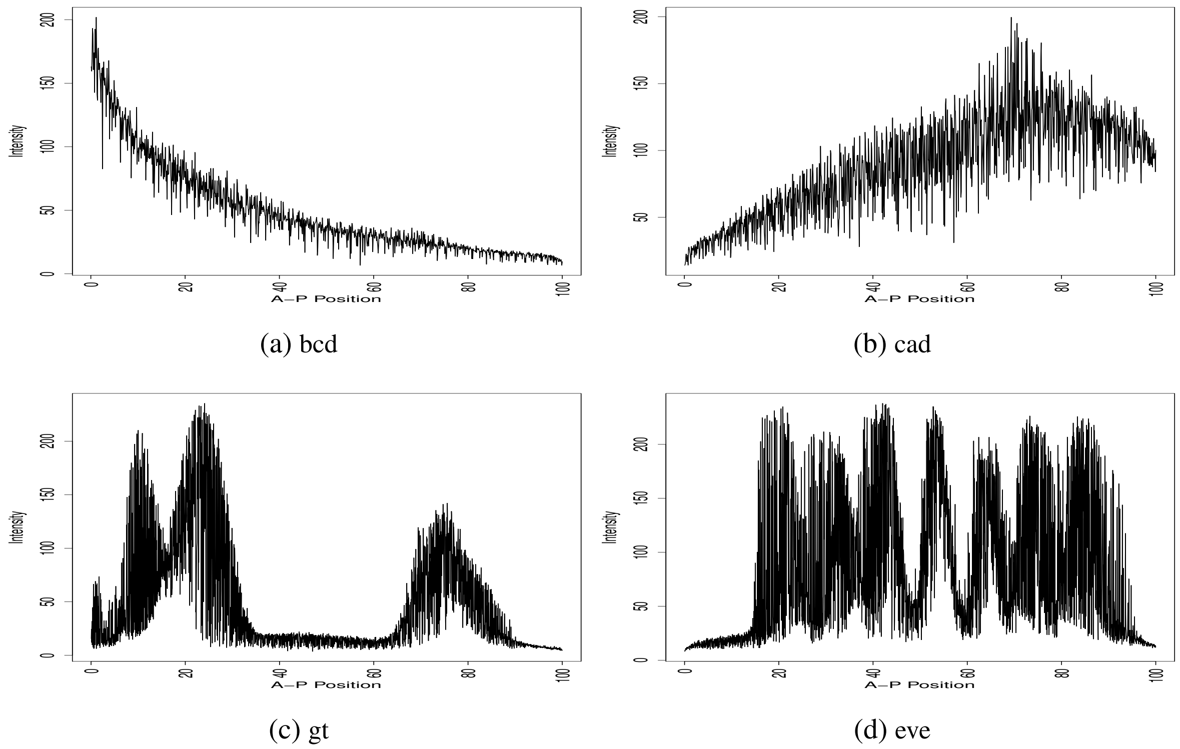
Same as the data set used in the previous section, the gene expression data of *cad*, *gt* and *eve* was achieved by immunofluorescence technique and is available via http://urchin.spbcas.

ru/flyex/. This data is also extracted from the nuclear intensities of %10 longitudinal strips and is not processed by any other noise removal technique.

Recall that *bcd* mRNA is completely maternal and the Bcd protein gradient is formed at cleavage cycle 10 [Holloway et al., 2006; Surkova et al., 2008a].

The *cad* mRNA has both maternal and the zygotic origin and the maternal transcripts begin to translate immediately after fertilisation. However, proteins encoded by *gt* and *eve* appear at cycle 12 and 10 respectively.

Fig 5.9 depicts the gene expression profile of these four different segmentation genes. As can be seen, the profiles are highly volatile and diverse in the pattern.



(a) bcd

(b) cad

(c) gt

(d) eve

Fig. 5.9 Experimental data from *D. melanogaster* embryo; (a): *bcd*, (b): *cad*, (c): *gt*, (d): *eve* .

Here, the method introduced in this chapter is explained in details. The central aim of the study is to enhance analysing the original series by decomposing it into a sum of series so that each component can be identified as either a signal component or noise. The proposed approach is a novel step that lies between the first and the second stages of SSA (i.e. decomposition and reconstruction) to select the proper value for the number eigenvalues r.

In doing so, let us consider a one-dimensional series $Y_N = (y_1, \ldots, y_N)$ of length N. Mapping this series into a multi-dimensional series $X_1, \ldots, X_K$ where $X_i = (y_i, \ldots, y_{i+L-1})^T \in \mathbf{R}^L$ provides $\mathbf{X} = (x_{i,j})_{i,j=1}^{L,K}$, where L $(2 \leqslant L \leqslant N/2)$ and $K = N - L + 1$. The matrix $\mathbf{X}$ is a Hankel matrix, which means all the elements along the diagonal $i + j = const$ are equal. Set $\mathbf{A} = \mathbf{X}\mathbf{X}^T$ and denote by λ_i $(i = 1, \ldots, L)$ the eigenvalues of $\mathbf{A}$ taken in the decreasing order of magnitude ($\lambda_1 \geq \ldots \geq \lambda_L \geq 0$) and by $U_1, \ldots, U_L$ the orthonormal system of the eigenvectors of the matrix $\mathbf{A}$ corresponding to these eigenvalues. Set

$$d = rank\,\mathbf{X} = \max(i, \text{ such that} \lambda_i > 0).$$

The SVD of the trajectory matrix can be written as:

$$\mathbf{X} = \mathbf{X}_1 + \cdots + \mathbf{X}_d, \tag{5.1}$$

where $\mathbf{X}_i = \sqrt{\lambda}_i U_i {V_i}^T$. The elementary matrices $\mathbf{X}_i$ have rank 1, U_i and V_i are the left and right eigenvectors of the trajectory matrix. Note that the collection $(\sqrt{\lambda_i}, U_i, V_i)$ is called the ith eigentriple of the SVD. Note also that $||\mathbf{X}||_F^2 = tr(\mathbf{X}\mathbf{X}^T) = \sum_{i=1}^{L} \lambda_i$ and $||\mathbf{X}||_F^2 = \lambda_i$, where $||\ ||_F$ denotes the Frobenius norm.

Let us now consider the step that comes between the two stages in SSA, that is to divide the matrix $\mathbf{A}$ by its trace, $\mathbf{A}/tr(\mathbf{A})$. Let $\zeta_1, \ldots, \zeta_L$ denote the eigenvalues of the matrix $\mathbf{A}/tr(\mathbf{A})$ in decreasing order of magnitude $(1 \geq \zeta_1 \geq \ldots \geq \zeta_L \geq 0)$. In this step, we perform the simulation technique to gain the distribution of ζ_i, so we can understand the behaviour of each eigenvalue, which can be useful for obtaining the proper value of r. In this section, our aim is to ascertain the distribution of ζ_i and its related forms that can be used directly for choosing the optimal value of r for the genes signal extraction.

Once r is obtained, the grouping step splits the matrices $\mathbf{X}_i$ into two groups. Therefore, (5.1) can be written as

$$\mathbf{X} = \mathbf{S} + \mathbf{E}, \tag{5.2}$$

where $\mathbf{S} = \sum_{i=1}^{r} \mathbf{X}_i$ is the signal matrix and $\mathbf{E} = \sum_{i=r+1}^{d} \mathbf{X}_i$ is the noise matrix. At the final step, we use the diagonal averaging to transform the matrix $\mathbf{S}$ into a new series of length N (for more information see [Ghodsi et al., 2015; Golyandina et al., 2001b]).

5.4.1 Algorithm

The Algorithm is composed of two stages. At the first stage, we use skewness coefficient and coefficient of variation of ζ_i as the main indicators to find the optimal value of r for the separability between signal and noise, and then at the second stage, we reconstruct the time series.

Stage 1:

1. Transfer a one-dimensional time series $Y_N = (y_1, \ldots, y_N)$ into the multi-dimensional series $X_1, \ldots, X_K$ with vectors $X_i = (y_i, \ldots, y_{i+L-1})^T \in \mathbf{R}^L$, where $K = N - L + 1$, and the window length L is an integer such that $2 \leq L \leq N/2$. This steps provide the trajectory matrix $\mathbf{X} = [X_1, \ldots, X_K] = \left(x_{ij}\right)_{i,j=1}^{L,K}$.

2. Computing the matrix $\mathbf{A} = \mathbf{X}\mathbf{X}^T / tr(\mathbf{X}\mathbf{X}^T)$.

3. Compute the eigenvalues and eigenvectors of the matrix $\mathbf{A}$ and represent it in the form $\mathbf{A} = \mathbf{P}\Gamma\mathbf{P}^T$. Here, $\Gamma = diag(\zeta_1, \ldots, \zeta_L)$ is the diagonal matrix of the eigenvalues of $\mathbf{A}$ that has the order $(1 \geq \zeta_1 \geq \zeta_2, \ldots, \zeta_L \geq 0)$ and $\mathbf{P} = (P_1, P_2, \ldots, P_L)$ is the corresponding orthogonal matrix of the eigenvectors of $\mathbf{A}$.

4. Simulate the original series m times and calculate the eigenvalues for each series. We simulate y_i from a uniform distribution with boundaries $y_i - a$ and $y_i + b$, where

$a =\mid y_{i-1} - y_i \mid$ and $b =\mid y_i - y_{i+1} \mid$. In order to obtain a noisy series similar to the real one, random error ε with a normal distribution with zero mean and variance σ^2_ε with different amplitudes were added to different parts of the series.

5. Calculate the coefficient of skewness for each eigenvalue, $skew(\zeta_i)$. If $skew(\zeta_c)$ is the maximum, then select $r = c - 1$.

6. Calculate the coefficient of variation, $CV(\zeta_i)$. This can split the eigenvalues in two groups, from ζ_1 to ζ_{c-1} which are corresponding to the signal and the rest which has almost a U shape which are corresponding to the noise component.

7. Calculate the absolute values of the correlation between ζ_i and ζ_{i+1}, and plot them in one figure (ρ). If $\rho(\zeta_{c-1}, \zeta_c)$ is the minimum, and the pattern for $\rho(\zeta_c, \zeta_{c+1})$ to $\rho(\zeta_{L-1}, \zeta_L)$ has the same pattern for the white noise, then choose $r = c - 1$.

Stage 2

1. Use the number of the eigenvalues r obtained in the first stage to calculate the approximated signal matrix $\widetilde{\mathbf{S}}$, that is $\widetilde{\mathbf{S}} = \sum_{i=1}^{r} \mathbf{X}_i$, where $\mathbf{X}_i = \sqrt{\lambda}_i U_i V_i^T$, U_i and V_i stands for the left and right eigenvectors of the trajectory matrix.

2. Transition to the one dimensional series can now be achieved by averaging over the diagonals of the matrix $\widetilde{\mathbf{S}}$.

Although the gene expression profiles are slightly different from embryo to embryo, as the obtained results in terms of number of eigenvalues are similar, we only consider ten different embryos for studying each gene. In this regard, each copy of gene expression data was simulated 10^4 times. Studying the distribution of each eigenvalue provides the capacity to obtain an accurate and deep intuitive understanding of selecting the proper value of r. The first data for each gene is analysed and discussed in more detail whilst the results of the other data are summarised based on the outcomes of the skewness, variation and correlation coefficients. The window length used for analysing the *bcd*, *cad*, *gt* and *eve* genes series is 200 (for more information for the selection of the window length refer to [Golyandina and Zhigljavsky, 2013]).

We mainly focus on the third moment, that is the skewness of the distribution for each eigenvalue:

$$skew(\zeta_i) = \frac{\frac{1}{m}\sum\limits_{n=1}^{m}\left(\zeta_{i,n} - \overline{\zeta}_i\right)^3}{\left[\frac{1}{m-1}\sum\limits_{n=1}^{m}\left(\zeta_{i,n} - \overline{\zeta}_i\right)^2\right]^{3/2}}, \tag{5.3}$$

and the coefficient of variation, $CV(\zeta_i)$, which is defined as the ratio of the standard deviation $\sigma(\zeta_i)$ and $\overline{\zeta}_i$:

$$CV_i = \frac{\sigma(\zeta_i)}{\overline{\zeta}_i}. \tag{5.4}$$

In addition, the Spearman correlation ρ between ζ_i and ζ_{i+1} is also evaluated to enhance the results obtained by *skew* and CV measures. The absolute value of the correlation between ζ_i and ζ_{i+1} is considered, 1 indicates that ζ_i and ζ_{i+1} have perfect positive correlation whilst 0 shows there is no correlation between them.

Fig 5.10 illustrates the results of $skew(\zeta_i)$ (left) and $CV(\zeta_i)$ (right) for the first data series for each gene type. It can be seen from the left column that the maximum value of *skew* is obtained for ζ_2 in both *bcd* and *cad* data. Whereas, $skew(\zeta_4)$ is the maximum for both *eve* and *gt* series. In the right column, the results of CV splits the eigenvalues into two groups for each data; the second group looks like a U shape which is related to the noise component. The results indicate that $r = 2,2,3,3$ for extracting the *bcd*, *cad*,*eve* and *gt* signal, respectively.

Furthermore, the result of ρ can be used o decide the *skew* and CV measures give different results. However, in these typical examples, the results of those two measures are the same which also supported by the results of the correlation coefficient. It is obvious that the minimum value of ρ are emerged between (ζ_2,ζ_3), (ζ_2,ζ_3), (ζ_3,ζ_4) and (ζ_3,ζ_4) for *bcd*, *cad*, *eve* and *gt*, respectively. Therefore, the results enhance that $r = 2,2,3,3$ for the first data for each gene (see Fig. 5.11).

Tables 5.3, 5.4, 5.5, and 5.6 depict the results of r based on those three measures for all 40 series. For the *bcd* signal extraction, all the outputs show $r = 2$ for all *bcd* data (see Table. 5.3). Similar results was emerged for extracting the *cad* signal, most of the outcomes indicate $r = 2$.

For the *eve* data, $r = 3$ for five series as all the three measures give the same result. However, for example; for series 2, the results of *skew* and CV are different, $r = 3$ and $r = 4$, respectively. To overcome this, we look at the result of ρ, which confirms $r = 4$. In this regards, the decision that $r = 3$ is for six series of ten *eve* data. Table. 5.6 demonstrates that $r = 3$ for all *gt* series except the last series, because all measures have the same results. As a result, for $L = 200$, the required eigenvalues to extract the *bcd*, *cad*, *eve*, and *gt* signals are 2, 2, 3, 3, respectively. Table. 5.7 shows the final results for all four genes along with the most frequent reported *skew*, CV and ρ.

Series	r (*skew*)	r (CV)	r (ρ)	Series	r (*skew*)	r (CV)	r (ρ)
1	2	2	2	6	2	2	2
2	2	2	2	7	2	2	2
3	2	2	2	8	2	2	2
4	2	2	2	9	2	2	2
5	2	2	2	10	4	2	2

Table 5.3 The values of r based on *Skew* and CV for the ten *bcd* series.

Series	r (*skew*)	r (CV)	r (ρ)	Series	r (*skew*)	r (CV)	r (ρ)
1	2	2	2	6	1	2	1
2	2	2	2	7	2	2	2
3	2	2	2	8	2	2	2
4	1	2	1	9	2	1	2
5	2	2	2	10	3	3	3

Table 5.4 The values of r based on *skew* and CV for the ten *cad* series.

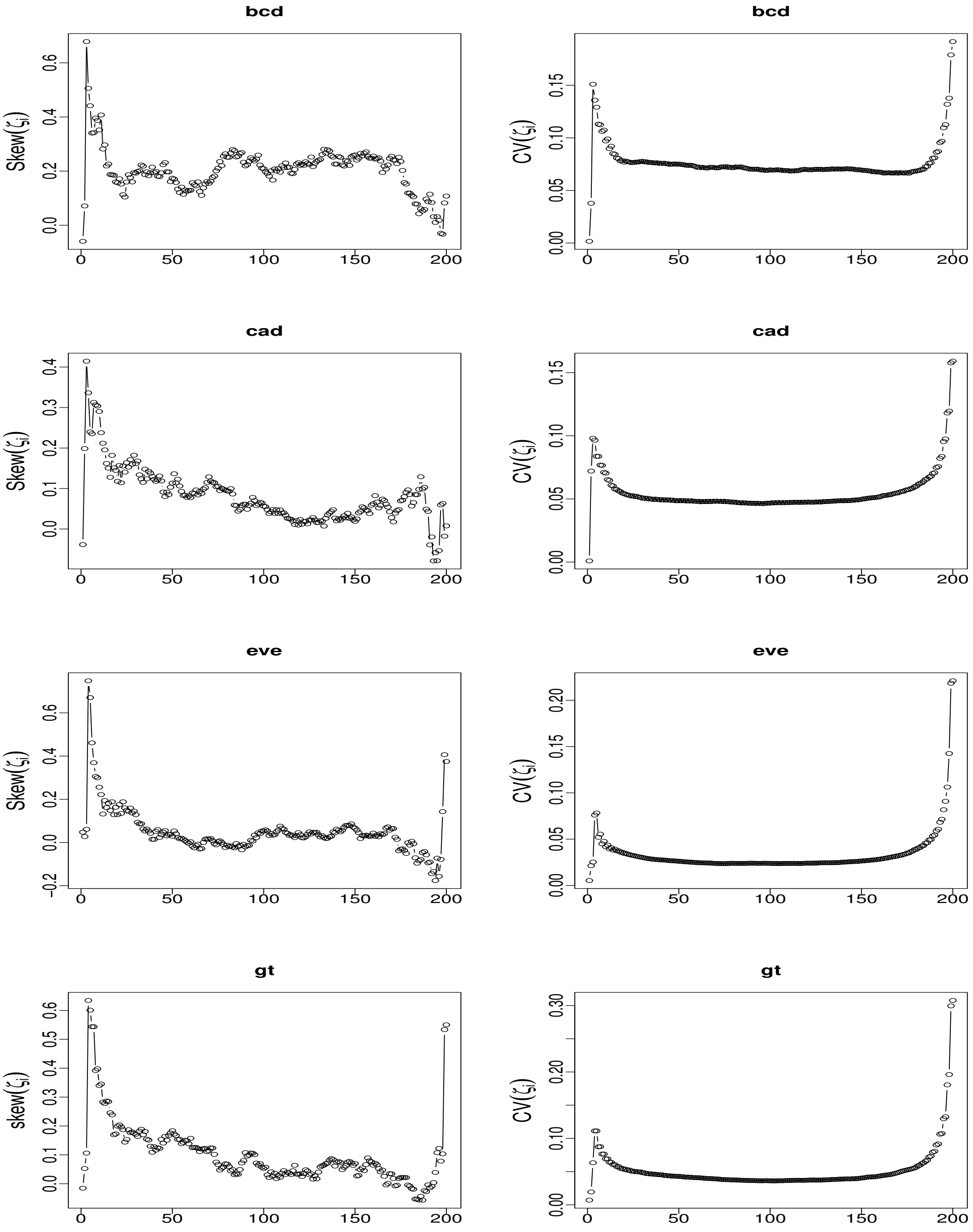


Fig. 5.10 The skewness coefficient (left) and the variation coefficient of ζ_i (right) for the first series of *bcd*, *cad*, *gt* and *eve* data.

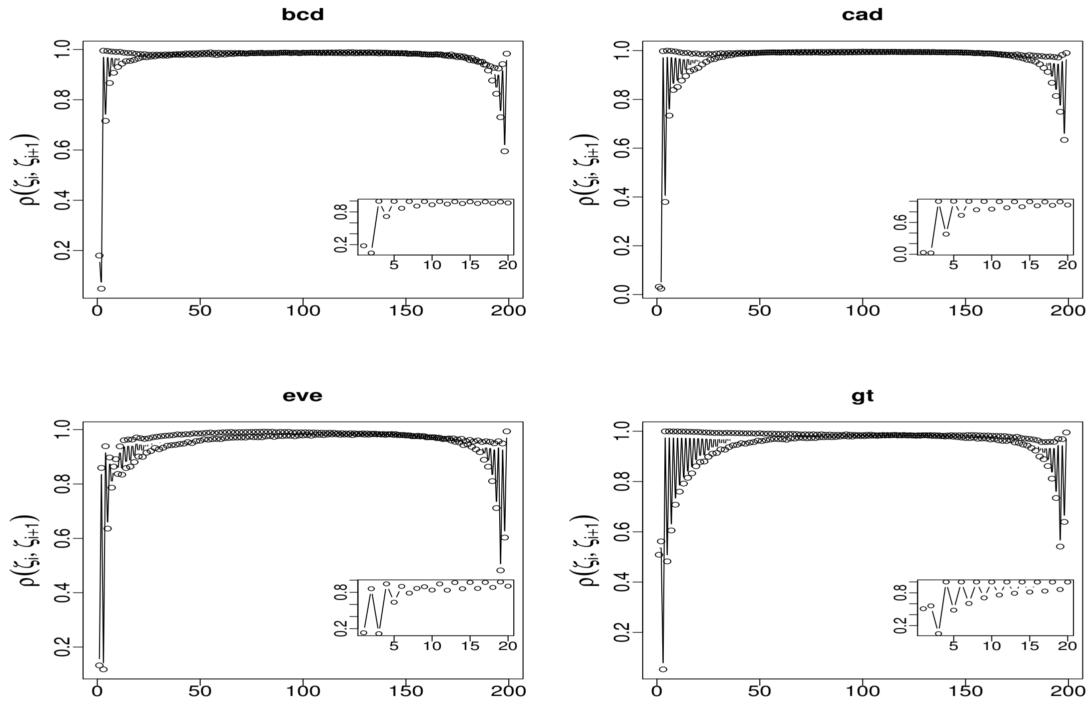


Fig. 5.11 The correlation between ζ_i and ζ_{i+1} for the first series from each data.

Series	$r\,(skew)$	$r\,(CV)$	$r\,(\rho)$	Series	$r\,(skew)$	$r\,(CV)$	$r\,(\rho)$
1	3	3	3	6	3	4	4
2	4	3	4	7	3	3	3
3	6	6	6	8	4	4	4
4	6	4	4	9	3	3	3
5	3	3	3	10	3	3	3

Table 5.5 The values of r based on *skew* and *CV* for the ten *eve* series.

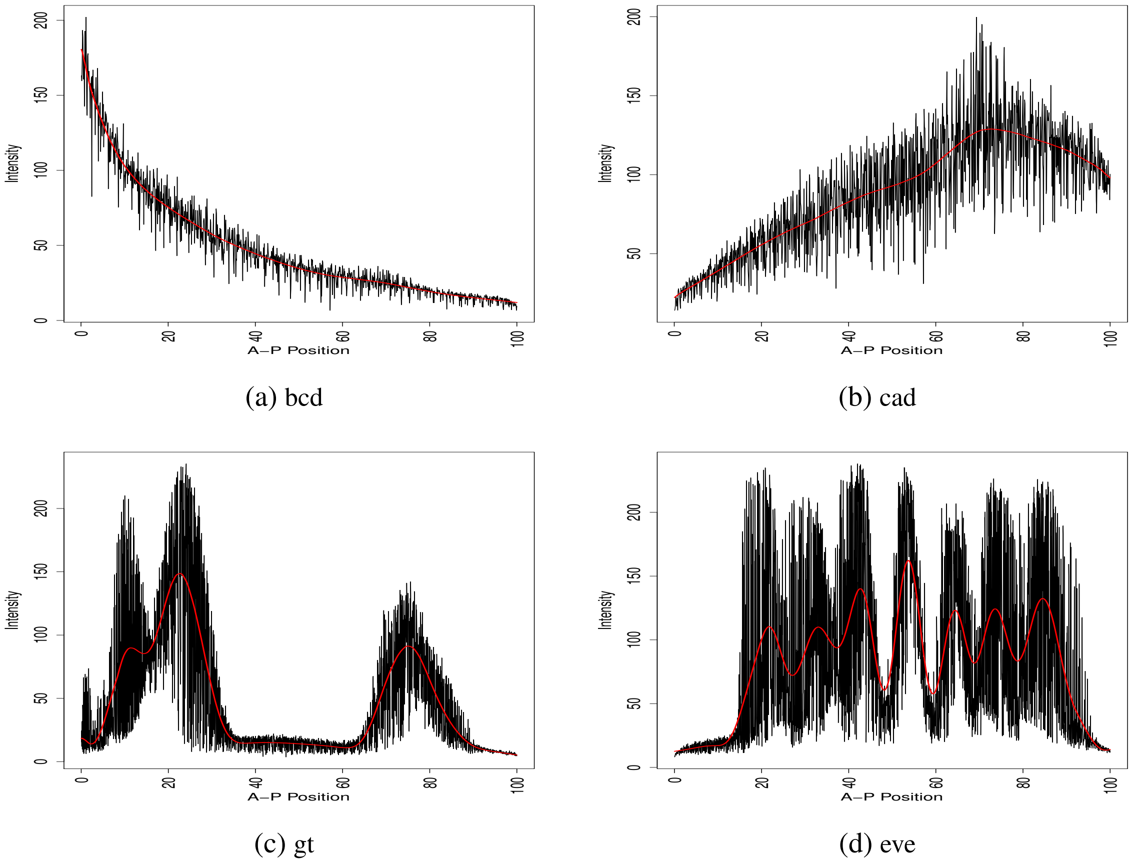
Series	$r\,(skew)$	$r\,(CV)$	$r\,(\rho)$	Series	$r\,(skew)$	$r\,(CV)$	$r\,(\rho)$
1	3	3	3	6	3	3	3
2	3	3	3	7	3	3	3
3	3	3	3	8	3	3	3
4	3	3	3	9	3	3	3
5	3	3	3	10	5	3	5

Table 5.6 The values of r based on *skew* and *CV* for the ten *gt* series.

Gene type	r (*skew*)	r (CV)	r (ρ)
bcd	2	2	2
cad	2	2	2
eve	3	3	3
gt	3	3	3

Table 5.7 The final result obtained in noise-signal separation study.

After the step of identifying the value of r, we can use those leader eigenvalues in the second stage of the SSA approach (Grouping and Diagonal averaging) to reconstruct the first typical data for each gene. Fig. 5.12 shows the result of the gene signal extraction or reconstruction series without noise. The red and the black lines correspond to the reconstructed series and the original series respectively. As a results, the considered r for the reconstruction of the original series is obtained properly.



(a) bcd

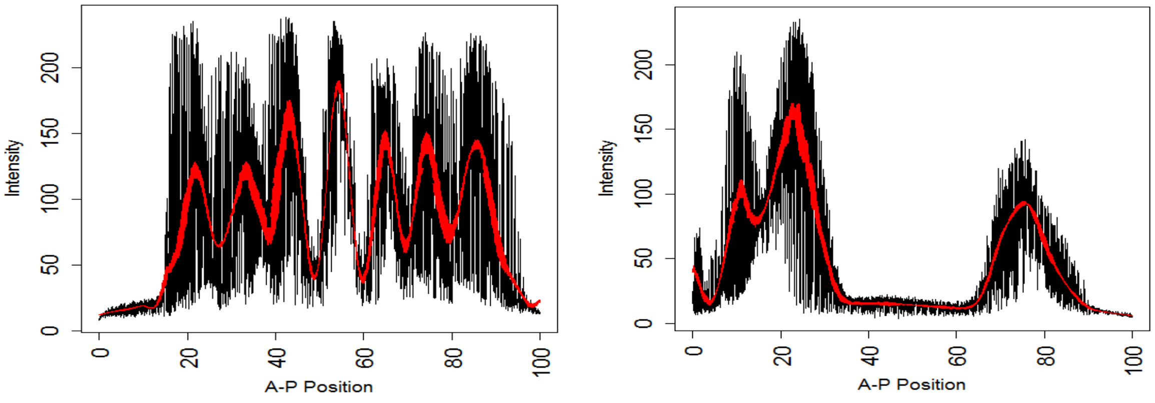
(b) cad

(c) gt

(d) eve

Fig. 5.12 Original (black) and reconstructed (red) series.

Taking a closer look at Fig 5.12, it is imperative to note that the extracted signal profiles of *eve* and *gt* do not follow the expression data satisfactorily when the data series changes sharply. Therefore, in order to solve this issue and capture the peaks of the profiles, we used sequential SSA. As opposed to the sequential SSA introduced in section 5.2, here, the main idea underlying this approach is to apply SSA recursively on the residuals with different window length L [Lahiri et al., 1995]. By doing so we extract some components of the initial series using basic SSA and then extract the remaining components related to the signal by applying SSA on residuals. Such a recursive SSA application produces a gradual extraction of the signal present in the noise. Fig 5.13 shows the result after applying sequential SSA. As can be seen signal extraction and peak capturing has been improved accordingly.



(a) bcd (b) cad

Fig. 5.13 Improving signal extraction using sequential SSA. Original (black) and extracted signal (red);(a): *eve*, (b): *gt*

5.5 SSA based on Colonial Theory

As previously mentioned in Chapter 3, the mathematical optimisation of an algorithm can be defined as the selection of the best inputs from some set of possible choices. This selection is made in respect to a criterion which is set upon the aim of the study (e.g. the precision of the extracted signal in signal processing studies, the accuracy of forecasted points in forecasting analysis or the cost of time for running the programme). Therefore, it is possible to introduce various approaches to optimise a signal processing algorithm according to different objectives of interest. Presented in this section is an alternative method for choosing the eigenvalues related to the signal of any given time series and gene expression profiles. The idea behind the newly proposed technique has driven from a developmental genetics theory called "Colonial theory” which aims at exploring the grouping step of the SSA technique in details.

As a great source of inspiration, nature holds the key to many questions we face on a daily basis. Therefore, it is not entirely surprising that much of the novel problem-solving techniques were initially inspired by nature (see, for example [Bonabeau et al., 1999; Dorigo, 1992; Eiben et al., 1994; McCulloch and Pitts, 1943; Nielsen and Chuang, 2010]). Even though credit is seldom given, In developing such intelligent solutions, nature provides us with effective background knowledge which follows from the profound observation and questioning of a natural phenomenon.

Quantum computing [Nielsen and Chuang, 2010], genetic algorithms [Eiben et al., 1994], neural networks[McCulloch and Pitts, 1943], swarm algorithms [Bonabeau et al., 1999] and ant colony optimization algorithms [Dorigo, 1992] are among the most established nature inspired models which seek to imitate specific phenomenon from nature in order to provide simple solutions to complex problems. Although attempting to model natural phenomena has a long history, the recent application of nature inspired algorithms like firefly algorithm [Yang, 2010], neuro fuzzy technique [Jang et al., 1997] and genetic programming [Koza, 1999] in the area of soft computing and also recent improvements in forecasting approaches [Benedetto et al., 2015; Martens and Zein, 2004] must be addressed as successful studies with great impacts.

However, improving signal extraction using bio–inspired algorithms is a relatively new area of research. It is noteworthy that in most of the nature-inspired algorithms, the natural phenomenon of interest is the strategy taken by biological organisms after facing an environmental change. Because the environmental setting makes the organism to exploit an ingenious solution to meet the new specific conditions.

Even though such biological solutions have been taken many times by organisms over the evolution process, the most prominent one can be referred to as the multicellularity phenomenon which describes how multicellular organisms arose from a single cell and generated multi-celled organisms. Despite there being various theories that may be able to explain this mechanism, Colonial Theory (CT) has received most credit by developmental scientists [Wolpert and Szathmáry, 2002].

Inspired by CT and by identifying certain characteristics of this theory, in this section, we draw a line between nature and mathematics. The mathematical procedures which we specifically seek to link with nature consist of not only SSA technique, but the SVD based methods and signal subspace (SS) methods which form the basis of a general class of subspace-based noise reduction algorithms. The superior performance of this class of algorithms in noise reduction has been proved by several studies (see, for example, [Hassani et al., 2011b; Soofi and Cao, 2012]).

SSA technique, is a SVD and SS based method which has been considered as a powerful nonparametric tool in several studies [Hassani, 2007]. Recall that SSA technique begins by decomposing the original series into the sum of a small number of independent components. Thereafter, the selected components are used to reconstruct the less noisy series which can be used for forecasting the future data points. However, due to the nature of least squares (LS) estimation method used in the current SSA procedure, the signal and noise separation is not optimum and the reconstructed series continues to hold some part of initial noise whilst the residual is not completely signal free. This section considers an alternative approach which is based on CT in order to provide a more efficient outcome for the signal and noise separation issue in SSA.

The exploitation of CT towards improving the SSA process is made possible via our identification of a general similarity between CT and SSA. However, this similarity was not visible in the grouping step of the basic SSA process, and therein lies our focus as we intend on defining a new approach to grouping in SSA by imitating one of the steps followed in CT. It is expected that this novel CT based approach to grouping enables a more efficient separation of signal from noise which in turn enhances the signal extraction and forecasting results. Since the forecasting aspect of the SSA technique is not of the main application in this thesis, we briefly touch upon the forecasting approach as the steps of the improved algorithm are explained.

5.5.1 Similarities between SSA and CT

This section focusses on providing a clear view on the similarities between SSA and CT as portrayed in Fig 5.14. In what follows, the information contained in Fig 5.14 is expanded upon as we present a detailed explanation of the linkages between SSA and CT.

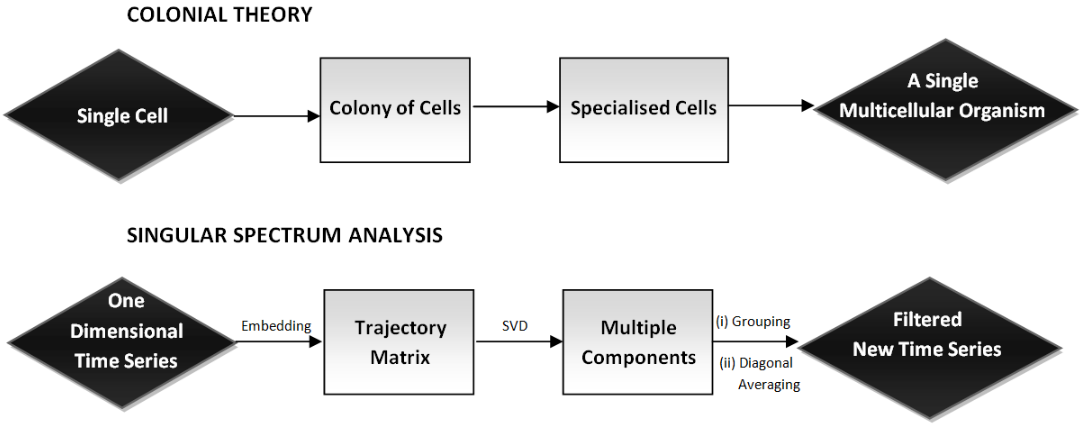


Fig. 5.14 The linkage between CT and SSA.

As presented in details in Chapter 3, the SSA method is made up of two complementary stages: Decomposition and Reconstruction; each stage consists of two compatible steps. At the first stage a group of small number of independent and interpretable components is achieved by decomposing the main series [Hassani, 2007], which is followed by the reconstruction of a less noisy series at the second stage [Hassani, 2007].

Stage 1: Decomposition

We begin with a one dimensional time series, $Y_N = (y_1, \ldots, y_N)$ where N is the length of the series. The SSA technique consists of two choices, the window length L and the number of eigenvalues r [Hassani and Mahmoudvand, 2013]. In SSA the number of components are related to the selection of the proper window length L which should be defined such that it minimises the signal distortion and maximises the residual noise level. However, as discussed before, we cannot impose a general rule in selecting L for different time series with different structure. For example, in instances where there is a periodic component with an integer period like a seasonal component, to obtain a higher separability the tradition is to select L proportional to that period [Hassani, 2007].

Likewise, the starting point of CT is a single cell which evolves over time and generates a multi-celled organism. Similar to SSA, there is also a limit on the number of cell types and different kinds of organisms necessitate different numbers of cell types. It is assumed that this number is determined by the balance between selective pressure and functional requirements, whilst variety is favoured by selection, functional needs limit the number of cell types [Grosberg and Strathmann, 2007].

1st step: Embedding

Here we take the one dimensional time series Y_N, and map it in order to create a multi-dimensional variable of $X_1, \ldots, X_K$ where $X_i = (y_i, \ldots, y_{i+L-1})' \in \mathbf{R}^L$. It is clear that this can be viewed as the creation of the colony from the initial single cell as we take Y_N and create multiple dimensions from the same series. This step provides us with a trajectory matrix, **X** which is a Hankel matrix that captures all information contained in Y_N.

$$\mathbf{X} = (x_{ij})_{i,j=1}^{L,K} = \begin{pmatrix} y_1 & y_2 & y_3 & \cdots & y_K \\ y_2 & y_3 & y_4 & \cdots & y_{K+1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ y_L & y_{L+1} & y_{L+2} & \cdots & y_N \end{pmatrix}. \tag{5.5}$$

It should be noted that unlike the Symbiotic theory [Hickman et al., 2001], which assumes that the symbiosis of various species caused a multicellular organism, in CT it is the symbiosis of many cells of the same species that forms a multicellular organism. This point is interesting as it can be referred to as the first and main difference between SSA and principal component analysis (PCA) . In the latter, the obtained matrix is achieved by considering different time series (multiple cells) whilst in SSA we consider one time series (single cell).

Moreover, transferring a one dimensional time series into a trajectory matrix will enable us to significantly reduce the computation time required for running the algorithm, as it eliminates the need for running the algorithm over a wide range of values for the hidden state dimension. Furthermore, by analysing the eigenvalues with the aim of filtering the signal and noise, the Signal to Noise Ratio (SNR) will be optimised in the newly reconstructed time series. Likewise, increasing in size is initially favoured by individual cells since multicellular organisms do not have the size limit which is mainly imposed by diffusion. As the surface-to-volume ratio decreases in a given single cell, with increased size they will experience difficulty in obtaining the required nutrients and transporting the cellular waste products out the cell [Grosberg and Strathmann, 2007; Kirk, 2005].

2nd step: Singular Value Decomposition

SVD is a procedure which is performed on **X** and provides us with several eigenvalues or components. The components obtained via this step are identified as trend, periodic, quasi-periodic component, or noise. In CT, **X** can represent the entire colony of cells generated from the original single cell. By increasing the interdependency level in a colony some of the cells specialise to do different tasks and by obtaining ever more complexity level, cells form tissues and then organs [Grosberg and Strathmann, 2007].

5.5.2 Stage 2: Reconstruction

In SSA, this is the stage where we seek to analyse the eigenvalues extracted via the SVD step to differentiate between noise and signal in a time series. In CT this would correspond to identifying which of the specialised cells are able to successfully carry out the reproductive task and which cells are responsible for viability.

1st step: Grouping

Grouping is a very important step in SSA as the quality of the filtering achieved via this technique depends on the successful analysis of eigenvalues and selection of appropriate groups of them to rebuild the less noisy time series. In brief, this step involves grouping together the eigenvalues with similar characteristics i.e. signal and harmonic components whilst leaving out those corresponding to noise. Likewise, the grouping step plays a significant role in CT as it determines which of the specialised cells are successful in carrying out the reproduction task and which of these fail along that way. Here it is imperative to note that in spite of the general similarity between SSA and CT in the grouping step, there is an important fundamental difference between the current version of SSA and CT which is discussed in detail in Section 5.5.3.

2nd step: Diagonal Averaging

The diagonal averaging step in SSA transforms the matrix of grouped eigenvalues back into a Hankel matrix which can later be converted into a time series. The resulting time series will be the less noisy, filtered time series corresponding to the original one-dimensional time series that was applied to the SSA process at the beginning. This step is important and similar to the final stage of CT where a single multiple organism is formed after defining the productive cells in order to compensate for the increased cost of reproduction imposed by increasing the size of the colony size.

According to the role of a small variant in CT, adding a single cell will only have a slight impact on the performance of a large organism [Bell and Mooers, 1997] which means after achieving the major functional specializations, there would presumably be a decline in adding capabilities by increasing the number of cell lines [Grosberg and Strathmann, 2007]. Similarly, in a time series, after extracting different components related to the trend, oscillation and noise, increasing the number of observations gives more components but all of these are categorised in the previously defined groups of components.

5.5.3 A New Approach for Grouping

It is widely accepted that the first grouping in nature happened when multicellular organisms arose from a single cell and generated a multi-celled organism [Michod, 2007]. At the very beginning of life there were only single cells. Today, after millions of years, most animals, plants, fungi, and algae are made up of multiple cells that work together as a single being [Adl et al., 2005].

Presented here is a brief explanation on developing functional specialisation and grouping the specialised cells of CT. Following this approach which is called changes in the level of complexity [Herron and Michod, 2008], we describe a novel approach for grouping in SSA.

In order to present a clearer view, we consider Volvocalean green algae as a model system. Volvocalean green algae are well suited for studying the transition as they provide different ranges from unicells to multicellular organisms with the explicit specification between germ and soma cells. In Volvocalean green algae, multicellular organisms are formed clonally from a single cell [Michod and Nedelcu, 2003]. Here, we mainly focus on one member of this lineage named *V. carteri* which exhibits a range of development in different cell types [Herron and Michod, 2008].

In [Kirk, 2005] a twelve-step program for the grouping step in *V. carteri* is considered. In this process, motility and mitosis activity compete for the same cellular machinery and cell destiny is determined finally by the location of microtubule organizing centre. The activity of the microtubule organizing centre mainly depends on its location in the cell and can serve either as a basal body which is related to the flagellar synthesis and consequently cell motion or a mitotic spindle which aids the segregation of the chromosomes during mitosis [King, 2004]. This germ–soma dichotomy is generated during the early embryogenesis and results in specifying two cell types; the somatic cells which are non-reproductive and only vegetative and the germ cell which performs the exclusively reproductive task [Kirk, 2005].

Here, as the germ cells execute the reproductive function and cell divisions which are required to produce a new daughter colony we consider them as those signal components which are later selected for series reconstruction. Differentiating the somatic and germ cells in *V. carteri* is largely dependent on the genetic differentiation which happens in somatic cells. During this process *regA* which is a regulatory gene and encodes a transcriptional repressor begins to express [Kirk et al., 1999]. As a result, several nuclear genes which are responsible for coding the chloroplast proteins are suppressed [Meissner et al., 1999]. Consequently, these somatic cells will not go under the cell growth or division. Lacking the division ability, they do not participate in the reproductive functions and offspring but accomplish the survival task by flagellar action [Kirk, 2005].

When comparing the grouping step between CT and SSA, two important points must to be highlighted:

1. Following differentiation germ cells do not not stay attached together. In a colony, those cells underpinning reproductive functions may divide on (a) the colony surface, (b) introgress or (c) in the interior of the colony as shown in Fig 5.15 [King, 2004].

2. In each round of reproduction, germ cells produce both future germ cells and soma cells.

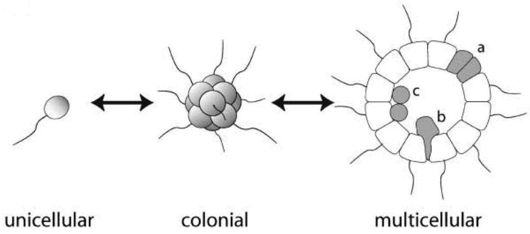


Fig. 5.15 Genetic variants produced differentiated cells in colonial flagellates [King, 2004].

However, the current grouping stage of the SSA method is based on the LS estimator i.e., choosing the leading eigentriple which describes the general tendency of the series [Hassani, 2007]. Accordingly:

- Selecting the signal components follows a binary approach. In other words, by estimating the signal rank r, the first r eigenvalues are always selected as signal components and the rest is considered to be noise.

- The leading components of $I_1,\ldots,I_m$ are related only to signal, hence, it is assumed that the reconstructed signal is not perturbed by noise.

Taking into consideration these points of comparison and our assumption which states germs cell from CT are equal to signal components, we can conclude that following the LS estimator the first r selected eigenvalues will not produce a clear signal because germ cells will produce both future germ cells and soma cells. This statement makes sense when we take the noise perturbation into account. Even in the leading components of $I_1, \ldots, I_m$, we have to exclude some eigenvalues with less information about the series, seek to find the other related signal components which capture more information and use all of them for reconstructing the series.

5.5.4 SSA–CT Algorithm

Presented below is a concise explanation of the SSA–CT algorithm which has also been depicted in Fig 5.16. Here, we use the RMSE criterion to determine the optimal choices of L and r (it is also possible to use any other criteria to determine the error as explained below in the algorithm). Accordingly, relying on the simulation procedure, in each L we are looking for a combination of eigenvalues r which provides the lowest RMSE, and this in turn represents the optimal decomposition and reconstruction choices for the SSA model. The automated SSA-CT code is able to perform this task by evaluating all possible L and r choices for a given time profile. [1]

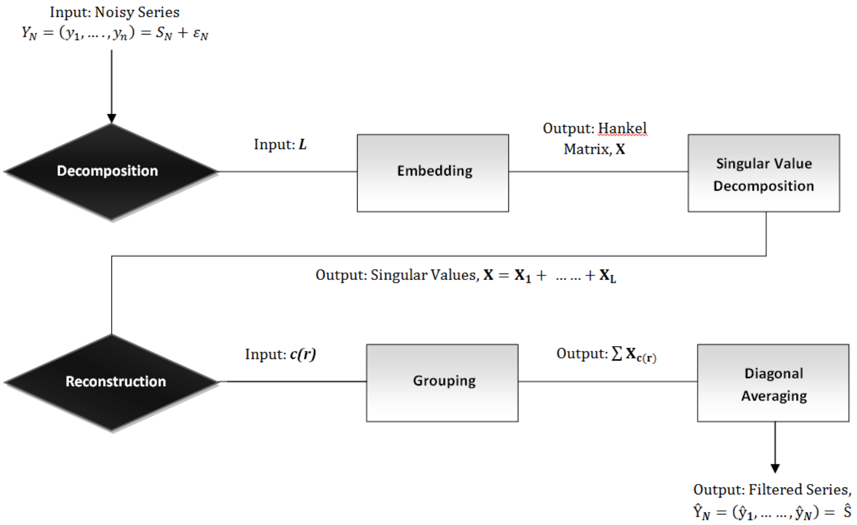

Input: Noisy Series
$Y_N = (y_1, \dots, y_n) = S_N + \varepsilon_N$
Decomposition
Input: L
Embedding
Output: Hankel Matrix, $\mathbf{X}$
Singular Value Decomposition
Output: Singular Values, $\mathbf{X} = \mathbf{X}_1 + \dots\dots + \mathbf{X}_\mathrm{L}$
Reconstruction
Input: $c(r)$
Grouping
Output: $\sum \mathbf{X}_{\mathrm{c(r)}}$
Diagonal Averaging
Output: Filtered Series,
$\hat{Y}_N = (\hat{y}_1, \dots\dots, \hat{y}_N) = \hat{S}$

Fig. 5.16 SSA-CT flowchart which depicts different computational steps of this algorithm.

[1] The SSA-CT code used in this thesis is available upon request.

1. Consider a real-valued nonzero time series $Y_N = (y_1, \ldots, y_N)$ of length N.

2. Use the training data to construct the trajectory matrix $\mathbf{X} = (x_{ij})_{i,j=1}^{L,K} = [X_1, ..., X_K]$, where $X_j = (y_j, ..., y_{L+j-1})^T$ and $K = N - L + 1$. Initially, we begin with $L = 2$ $(2 \leq L \leq \frac{N}{2})$ and in the process, evaluate all possible values of L for Y_N.

3. Obtain the SVD of $\mathbf{X}$ by calculating $\mathbf{X}\mathbf{X}^T$ for which $\lambda_1, \ldots, \lambda_L$ denotes the eigenvalues in decreasing order $(\lambda_1 \geq \ldots \lambda_L \geq 0)$ and by $U_1, \ldots, U_L$ the corresponding eigenvectors. The output of this stage is $\mathbf{X} = \mathbf{X}_1 + \ldots + \mathbf{X}_L$ where $\mathbf{X}_i = \sqrt{\lambda}_i U_i V_i^T$ and $V_i = \mathbf{X}^T U_i / \sqrt{\lambda}_i$.

4. Adapt a diagonal matrix containing the weights $\mathbf{W}_{K \times K}$ as follows:

$$\hat{\mathbf{S}}_{CT} = \mathbf{U}_1(\mathbf{W}_{CT}\Sigma_1)\mathbf{V}_1^{'} \quad .$$

5. Choose the weight matrix $\mathbf{W}_{K \times K}$.

$$\mathbf{W}_{CT} = \text{diag}\left((1 \vee 0), \ldots, (1 \vee 0)\right) \quad .$$

6. Evaluate all possible combinations of $\mathbf{W}_{CT}$ (step by step) for the selected L and split the elementary matrices $\mathbf{X}_i$ $(i = 1, \ldots, L)$ into several groups and sum the matrices within each group.

7. Perform diagonal averaging to transform the matrix into a Hankel matrix which can then be converted into a time series.

8. Find the RMSE value for each reconstructed series and report the L and the selected combination of r attributed to the minimum RMSE as optimal choices.

9. The output is a filtered series that can be used for forecasting.

The optimal choices of L and r can be used for forecasting via vector or recurrent SSA (SSA-V,SSA-R) techniques. More detailed information of these techniques can be found in [Sanei and Hassani, 2015].

Accordingly, to assess the performance of the newly proposed technique we conduct a simulation study. The Bcd simulation process follows the procedure explained in chapter 1. To start the simulation, an exponential curve is drawn from the simple SDD model. Thereafter, random errors ε of a normal distribution with zero mean and variance σ_{ε}^{2} with different amplitudes are added to various parts of this curve. This simulation is repeated $1,000$ times.

SSA-CT algorithm gives the ability to evaluate all possible values of L with different combinations of r. Although this method assures to provide a signal with the least RMSE, the computation process for a simulation study covering different noise levels and lengths of series can be considerably time taking. However, for forecasting objective, a set of training data is utilised and therefore the computation time required for running the algorithm is considerably reduced.

Accordingly, we have applied the SSA-CT algorithm for Bcd signal extraction. Table 5.8 shows the obtained result where N depicts the length of the Bcd profile, (L,r) the optimal L and r and f shows the frequency out of 1000 iteration. As can be seen, for the case of Bcd profile the best-extracted signal with the least RMSE and the highest frequency always achieved following the binary approach. This result indicates that to reconstruct the Bcd signal as correctly assumed before only the first eigenvalue (in some cases, $r = 1,2$) is required which may be due to the overall exponential pattern of the Bcd profile. However, when the SSA-CT algorithm is applied for both signal extraction and forecasting purposes on a wide range of time series, the obtained result (Not reported here but available upon request) showed that for the more complicated series, one will not be able to capture the best result using the binary approach.

N	L,r	f	**N**	L,r	f
50	5,1	947	700	205,1	689
100	11,1	952	1000	256,1-2	863
150	25,1	856	1200	153,1	745
200	25,1	984	1400	542,1-2	923
300	74,1	866	1800	512,1-2	685
500	124,1-2	954	2000	532,1-2	981
600	212,1	978	2500	585,1-2	714

Table 5.8 Optimal L and r found for simulated Bcd profiles with different lengths.

5.6 Conclusion

This chapter begins with the core aim of introducing new criteria for optimising Bcd signal extraction. Motivated by the findings in Chapter 4, we opt to tailor the new Bcd signal extraction criteria for use with the SSA technique which found to be the best option for Bcd signal extraction in relation to SDD, ARIMA, ETS, ARFIMA and NN models. In line with our aim, we initially produce an algorithm for optimising the Bcd signal extraction process with SSA. In brief, the algorithm is optimised based on minimising the skewness statistic for the SSA residual. We suggest that setting L equal to the minimum skewness within the threshold $10 \geq L \geq N/4$ and combine this SSA choice with $r = 1$ or $r = 1,2$ as appropriate will enable users to obtain the optimal Bcd signal extraction with SSA.

Through this chapter, we have succeeded in presenting several contributions to the field of Bcd signal extraction. The first and most important of which deals with the application of the newly proposed algorithm to 27 real Bcd data to show that it can enable researchers to select the appropriate SSA choices to extract a smooth and accurate Bcd signal quickly and easily without the need to spend an increased amount of time for the selection of L for decomposing the data. However, we notice that given the highly complex nature of the Bcd data, on one occasion the SSA algorithm fails to extract an absolutely smoothed signal. As a solution to this problem, we introduce for the first time, the concept of Sequential SSA on signals in section 5.2. Via this approach, we are able to refine and smoothen further the initial signal which had captured some of the observational and biological noise in Bcd data.

In line with good practice, in addition to evaluating the signal extractions alone, this study also pays attention to the residuals. The analysis of the residuals motivated us to introduce hybrid SSA based signal extraction processes for Bcd in section 5.3. In brief, when extracting the signal from any given data set, one would reasonably expect other signals to end up within the noise component. However, this would mean that the residual is no longer random and some statisticians could find it difficult to accept such techniques. Accordingly, the first hybrid SSA signal process is focussed on providing a Bcd signal extraction procedure which will ensure the residual is white noise. This was achieved by combining the optimised SSA signal with optimised ARIMA models being fitted to the residuals. Whilst the results did provide the necessary outcomes in terms of residuals with white noise, it comes at a cost (i.e., a loss in the smoothness of the extracted signal).

The SSA-ARIMA hybrid approach is a combination of parametric and nonparametric techniques. For those who wish to rely on nonparametric techniques alone so that one is not restricted by the parametric assumptions, we present the SSA-ETS hybrid Bcd signal extraction approach. This process also produces an important contribution of this research as we find a solution to the problem of modelling accurately the initial curve in Bcd data which was not only experienced in Chapter 4. Accordingly, we are able to present the hybrid SSA-ETS process which is a combination of the optimised SSA signal extraction algorithm with an optimised ETS algorithm as the most efficient approach for Bcd signal extraction.

Also, in this chapter, a new approach for removing noise and signal extraction of Bcd is introduced which also applied and evauted for four different *D. melanogaster* segmentation genes. The approach was based on the distribution of the eigenvalues of a scaled Hankel matrix. The skewness and variation coefficients of the eigenvalue distribution are used as new criteria and indicator for the cut-off point in the eigenvalue spectra between signal and noise components. The results confirm that the proposed approach gives a promising output for the gene expression signal extraction.

The results obtained by applying the introduced algorithm in section 5.4, emphasises that when extracting signal from different expression gene profiles, for optimised signal and noise separation, a different number of eigenvalues need to be chosen for each gene.

In Section 5.5, a novel approach for enhancing the accuracy of SSA signal extraction is introduced based on the foundations of CT. Initially, we draw upon the general similarity between CT and SSA, and then exploit these similarities, particularly certain characteristics of CT in the grouping step which is the most important step in the SSA procedure.

In brief, we suggest that relying on a binary approach of differentiating between signal and noise at the grouping step is not necessarily the best approach as this assumes there is no useful information contained in the selected noise components. Instead, based on CT we propose a different grouping approach which considers analysing all eigenvalues and selecting those which have useful information for grouping in SSA. The result shows that the new idea of grouping has the potential to enable us to obtain a more efficient signal in comparison to the existing approach for grouping in SSA which is based on LS.

Chapter 6

A novel hybrid method to study the regulatory interactions

6.1 Introduction

As previously noted in chapter 1, segmentation in *D. melanogaster* is a particularly well-studied process which highlights the role of GRNs in the earliest stage of development [Lewis, 1978].

In SN, there are three fundamental types of genes which play crucial roles in the development of *D. melanogaster* : maternal effect genes, gap genes and pair-rule genes [Bieler et al., 2011]. Among them, the maternal effect genes including *bcd* and *cad* must be addressed as the most important factors since they respectively determine most aspects of anterior and posterior axis of an adult fruit fly and more importantly, they commence the sequential activation of SN. [Berleth et al., 1988; Bieler et al., 2011; Copf et al., 2004]

The SN is perhaps the best-studied transcriptional network in *D. melanogaster* development. Therefore, there are considerable attempts to portrait a picture of the interactions presented between regulators in this GRN. Quantitatively, it is common to model GRNs using Ordinary Differential Equations (ODEs) or stochastic ODEs [De Jong, 2002; Karlebach and

Shamir, 2008]. Even though, the substantial progress which has been made in modeling transcriptional regulations using these models in recent years is not deniable, the enormous number of regulatory functions obtained by these models and the estimation of parameters which are difficult to assess experimentally are still considered as two major drawbacks of these methods [Schlitt and Brazma, 2007; Wilczynski and Furlong, 2010].

Recently, the availability of more data on molecular mechanisms of regulatory interactions has made it possible to study these interactions in more quantitative depth. however, to the best of our knowledge, there is not a particular study which evaluates the dynamic interactions of this system from a statistical causality point of view [Chaves et al., 2005; Frigerio et al., 1986; Levine and Davidson, 2005].

Hence, this chapter considers an alternative approach to evaluate the possibility of ratifying the validity and reliability of genetic inferences derived from experimental evidence by using proper analytical tools. It is of note that the detected regulatory link can be either inductive (i.e. increasing the protein concentration of one gene raises the protein concentration of the other gene), or inhibitory (i.e. increasing the protein concentration of one gene decreases the protein concentration of the other gene) [Davidson and Levin, 2005]. Any efforts for identifying the nature of the detected interaction would require more extensive research and that objective is beyond the mandate of this thesis.

The analytical methods developed in this chapter is a novel hybrid method comprised of a variety of causality detection methods and SSA technique.

The causality detection methods adopted here consist of time and frequency domain Granger Causality detection (GC) [Geweke, 1982] approaches and an advanced non-parametric method - Convergent Cross Mapping (CCM) [Sugihara et al., 2012]. Time domain causality test [Granger, 1969] and its developed versions are the most common and generally accepted methods in causal inference analysis. Frequency domain causality test is the extension of time domain causality test on identifying causality for each individual frequency component instead of computing a single measure for the entire causal association. CCM is an advanced non-parametric method that is designed for a dynamical system involving complex interactions. The fundamental concept of CCM is that the information of the driver variable can be recovered from the predator variable, but not vice versa.

It is imperative to note that since providing robust genetic evidence is an important step in reporting genetic regulatory, among all the interactions between regulators in SN, we have narrowed down this study to the interactions between *bcd* and *cad*, *bcd* and *kr* and *cad* and *kr* which their interactions have been previously accredited via laboratory experimental evidence. Accordingly, extracting this link using mentioned causality detection techniques gives us credit to step further and apply this method to find the unknown regulatory links between other genes.

The regulatory role of *bcd* has been unveiled by several studies [Berleth et al., 1988; Lopes et al., 2012]. According to [Baird-Titus et al., 2006] Bcd is one of few proteins which binds both RNA and DNA targets and can be involved in both transcriptional and post-transcriptional regulation. Bcd enhances the transcription of anterior gap genes such as *hb* and represses the translation of *cad* in the anterior region of the embryo. In 2002, through an experimental approach, Niessing et al. showed translational repression of *cad* mRNA by Bcd depends on a functional eIF4E-binding motif [Niessing et al., 2002]. The *cad* and *kr* genes are also required for a normal segmentation of the embryo. As noted in [Liu and Jack, 1992], the interaction of Cad and Kr gene is an important input of the segmentation genetic network.

In applying the causality detection techniques, it should also be noted that several studies have previously shown that these methods are sensitive to noise [Ancona et al., 2004; Hiemstra and Jones, 1994; Zou and Feng, 2009] and gene expression profiles are exceedingly noisy [Golyandina et al., 2012]. As it has been shown in Fig 6.1, the profile achieved by fluorescence antibodies technique is highly volatile and in such cases, establishing a cause-

and-effect relationship is more challenging and demands to apply a noise filtering step before regulatory detection studies. Therefore, to overcome this issue we adopt the SSA technique to filter the noise of the profiles. The main reason behind choosing SSA does not require added justification here as the superior performance of this method is illustrated through the previous chapters of this thesis.

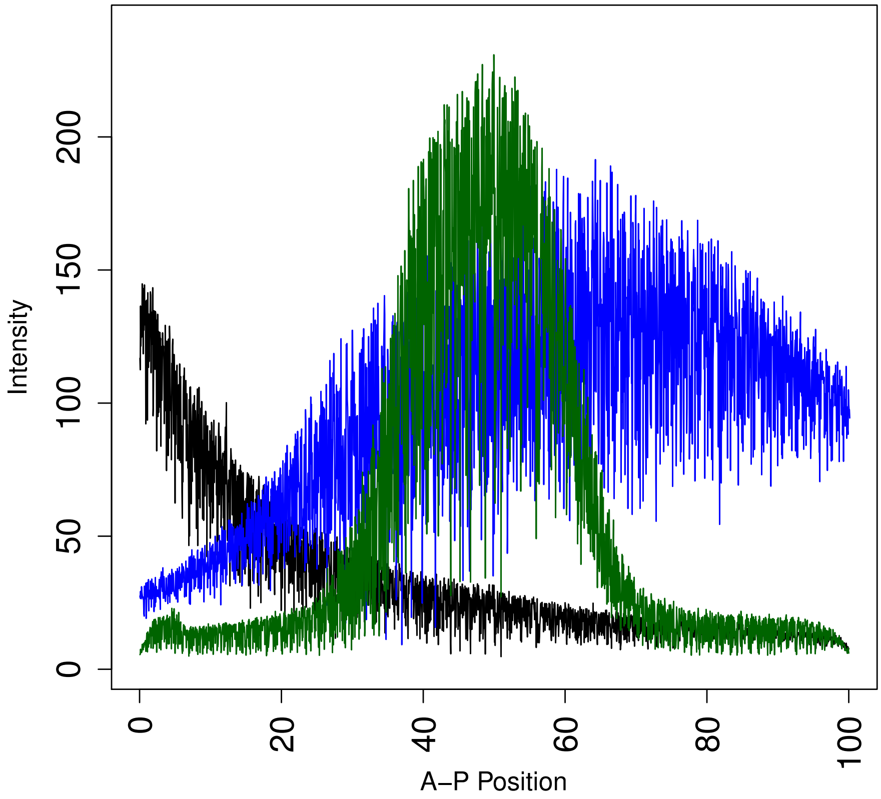


Fig. 6.1 A typical example of noisy Bcd, Cad and Kr for embryo *ms26* at time class 14(1). Black, blue and green colours depict Bcd, Cad and Kr profiles respectively. The x-axis shows the position of the nuclei along the A-P axis of the embryo and Y-axis shows the fluorescence intensity levels.

6.2 Causality detection techniques

6.2.1 Time domain Granger causality

Granger causality test [Granger, 1969] is the most generally accepted and significant method for causality analyses in various disciplines. Various applications and developments of this technique, also more specifically in the biomedical area, can be found in [Chen et al., 2006; Deshpande et al., 2010; Gow et al., 2008; Hsiao, 1981; Sims, 1972; Sims et al., 1990]. The regression formulation of Granger causality states that vector X_i is the cause of vector Y_i if the past values of X_i are helpful in predicting the future value of Y_i, two regressions are considered as follows:

$$X_i = \sum_{t=1}^{T} \alpha_t Y_{i-t} + \varepsilon_{1i}, \tag{6.1}$$

$$Y_i = \sum_{t=1}^{T} \alpha_t Y_{i-t} + \sum_{t=1}^{T} \beta_t X_{i-t} + \varepsilon_{2i}, \tag{6.2}$$

where $i = 1, 2, \cdots, N$ (N is the number of observations), T is the maximal time lag, α and β are vectors of coefficients, ε is the term. The first regression is the model that predicts X_i by using the history of Y_i only, while the second regression represents the model of Y_i is predicted by the past information of both X_i and Y_i. Therefore, the conclusion of existing causality is conducted if the second model is a significantly better model than the first one.

6.2.2 Frequency domain causality

The frequency domain causality test is the extension of time domain GC test that identifys the causality between different variables for each frequency. In order to briefly introduce the methodology, we mainly follow [Ciner, 2011; Croux and Reusens, 2013; Geweke, 1982]. More details can be found in [Breitung and Candelon, 2006; Faes and Nollo, 2013; Lemmens et al., 2008].

It is assumed that X_i and Y_i are two centred stationary time series (where $i = 1, 2, \cdots, N$ and N is the number of observations) with a finite-order Vector Auto-regression Model (VAR) representative of order p,

$$\Theta(R)\begin{pmatrix} Y_i \\ X_i \end{pmatrix} = \begin{pmatrix} \Theta_{11}(R) & \Theta_{12}(R) \\ \Theta_{21}(R) & \Theta_{22}(R) \end{pmatrix}\begin{pmatrix} Y_i \\ X_i \end{pmatrix} + \mathscr{E}_i, \tag{6.3}$$

where $\Theta(R) = I - \Theta_1 R - ... - \Theta_p R_p$ is a 2×2 lag polynomial and $\Theta_1, ..., \Theta_p$ are 2×2 autoregressive parameter matrices, with $R^k X_i = X_{i-k}$ and $R^k Y_i = Y_{i-k}$. The error vector vector $\mathscr{E}$ is assumed to be multivariate white noise with zero mean, and $E(\mathscr{E}_i \mathscr{E}_i') = \mathbf{Z}$, where $\mathbf{Z}$ is positive definite matrix. The moving average (MA) representative of the system is

$$\begin{pmatrix} Y_i \\ X_i \end{pmatrix} = \Psi(R)\eta_i = \begin{pmatrix} \Psi_{11}(R) & \Psi_{12}(R) \\ \Psi_{21}(R) & \Psi_{22}(R) \end{pmatrix}\begin{pmatrix} \eta_{1i} \\ \eta_{2i} \end{pmatrix}, \tag{6.4}$$

with $\Psi(R) = \Theta(R)^{-1}\mathbf{G}^{-1}$ and $\mathbf{G}$ is the lower triangular matrix of the Cholesky decomposition $\mathbf{G}'\mathbf{G} = \mathbf{Z}^{-1}$, such that $E(\eta_t \eta_t') = I$ and $\eta_i = \mathbf{G}\mathscr{E}_i$. The causality test developed in [Geweke, 1982] can be written as:

$$C_{X \Rightarrow Y}(\gamma) = log\left[1 + \frac{|\Psi_{12}(e^{-i\gamma})|^2}{|\Psi_{11}(e^{-i\gamma})|^2}\right]. \tag{6.5}$$

However, according to this framework, no Granger causality from X_i to Y_i at frequency γ corresponds to the condition $|\Psi_{12}(e^{-i\gamma})| = 0$, this condition leads to

$$|\Theta_{12}(e^{-i\gamma})| = |\Sigma_{k=1}^{p}\Theta_{k,12}\cos(k\gamma) - i\Sigma_{k=1}^{p}\Theta_{k,12}\sin(k\gamma)| = 0, \tag{6.6}$$

where $\Theta_{k,1,2}$ is the $(1,2)th$ element of Θ_k, such that a sufficient set of conditions for no causality is given by [Breitung and Candelon, 2006]

$$\begin{array}{c}\Sigma_{k=1}^{p}\Theta_{k,1,2}\cos(k\gamma)=0\\ \Sigma_{k=1}^{p}\Theta_{k,1,2}\sin(k\gamma)=0\end{array}, \tag{6.7}$$

Hence, the null hypothesis of no Granger causality at frequency γ can be tested by using a standard F-test for the linear restrictions (6.7), which according to Breitung and Candelon, follows an $F(2,B-2p)$ distribution, for every γ between 0 and π, with B begin the number of observations in the series [Breitung and Candelon, 2006].

6.2.3 Convergent Cross Mapping

Convergent Cross Mapping (CCM) is firstly introduced in [Sugihara et al., 2012] that aimed at detecting the causation among time series and provide a better understanding of the dynamical systems that have not been covered by other well established methods like Granger causality. CCM has proven to be an advance non-parametric technique for distinguishing causality in a dynamic system that contains complex interactions in biological studies and ecosystems, more details can be found in [Clark et al., 2015; Deyle et al., 2013; Sugihara et al., 2012; Ye et al., 2015]. CCM is briefly introduced below by mainly following [Sugihara et al., 2012].

Assume there are two variables X_i and Y_i, for which X_i has a causal effect on Y_i. CCM test will test the causality by evaluating whether the historical record of Y_i can be used to get reliable estimates of X_i. Given a library set of n points (not necessarily to be the total number of observations N of two variables) and here set $i=1,2,\cdots,n$, the lagged coordinates are adopted to generate an E-dimensional embedding state space [Sugihara and Mayf, 1990; Takens, 1981], in which the points are the library vector X_i and prediction vector Y_i

$$X_i \; : \; \{x_i, x_{i-1}, x_{i-2}, \cdots, x_{i-(E-1)}\}, \tag{6.8}$$

$$Y_i \; : \; \{y_i, y_{i-1}, y_{i-2}, \cdots, y_{i-(E-1)}\}, \tag{6.9}$$

The $E+1$ neighbors of Y_i from the library set X_i will be selected, which actually form the smallest simplex that contains Y_i as an interior point. Accordingly, the forecast is then conducted by this process, which is the nearest-neighbour forecasting algorithm of simplex projection [Sugihara and Mayf, 1990]. The optimal E will be evaluated and selected based on the forward performances of these nearby points in an embedding state space.

Therefore, by adopting the essential concept of Empirical Dynamic Modeling (EDM) and generalized Takens' Theorem [Takens, 1981], two manifolds are conducted based on the lagged coordinates of the two variables under evaluation, which are the attractor manifold M_Y constructed by Y_i and respectively, the manifold M_X by X_i. The causality will then be identified accordingly if the nearby points on M_Y can be employed for reconstructing observed X_i. Note that the correlation coefficient ρ is used for the estimates of cross map skill due to its widely acceptance and understanding, additionally, leave-one-out cross-validation is considered a more conservative method and adopted for all evaluations in CCM.

6.3 Data

In selecting the appropriate type of data, it is imperative to note that unlike *bcd* mRNA, *cad* mRNA has a uniform distribution along the embryo, which indicates that Bcd does not regulate the transcription of *cad* but inhibits the translation of the *cad* mRNA [Latchman et al., 2007]. Therefore, Cad can only be produced where the concentration of Bcd is low suggesting that the protein profiles of these two genes should be targeted to study.

Accordingly, we use the dataset which is fully described in chapter 1. Similar to Bcd, Cad Profile in wild-type *D. melanogaster* embryos achieved by fluorescently tagged antibodies technique and is available via FlyEx database.

Here, we evaluate the method for cleavage cycles 10-14A (when proteins synthesised from maternal transcripts begin to appear up to the onset of gastrulation) and all the time classes of cleavage cycle 14A. As expected, the profiles from different cleavage cycles vary in expression level and expression pattern.

Table 6.1 presents the number of embryos studied per each time class. It is of note the expression profile of each embryo has a different length of data where the third column in this table reports the average.

Time class	N	Length	SD
10	5	127	18.83
11	12	276	25.83
12	15	489	97.18
13	47	1224	78.56
14(1)	28	2318	143.87
14(2)	15	2315	86.83
14(3)	20	2367	141.05
14(4)	17	2309	119.16
14(5)	14	2301	126.96
14(6)	18	2347	103.74
14(7)	13	2007	229.61
14(8)	12	1600	311.21

Table 6.1 Different time classes and the embryos studied per each time class. Note: N= Number of embryos studied per each time class, Length= The average length of data of expression profiles, SD= Standard deviation of length of data.

In order to extract the signals from the profiles, we adopt the SSA technique. To that aim, number of the eigenvalues needed for signal reconstruction is obtained by following the method presented in section 5.4 of Chapter 5.

Fig 6.2 illustrates the output from this effort. It is evident that the SSA method provides a relatively smooth signal line with correlation below 0.10 which credits the satisfactory level of separation between noise and signal using this technique (See, Chapter 4).

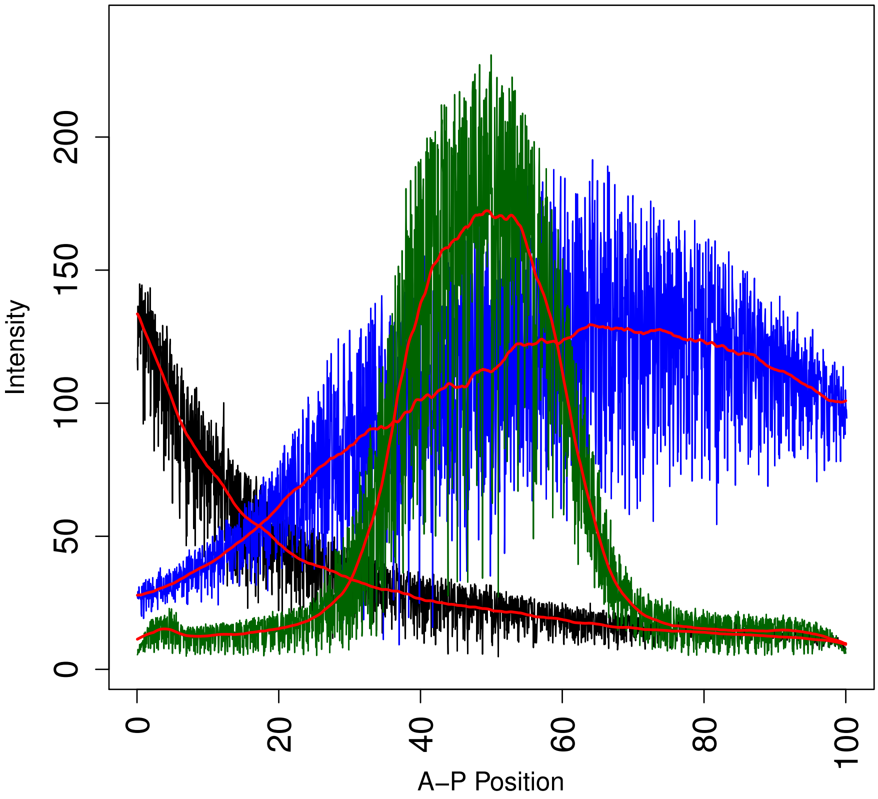


Fig. 6.2 A typical example of noisy Bcd, Cad and Kr along with the extracted signals in red for embryo *ms26* at time class 14(1). Black, blue and green colours depict Bcd, Cad and Kr profiles respectively. The x-axis shows the position of the nuclei along the A-P axis of the embryo and Y-axis shows the fluorescence intensity level.

6.4 Empirical Results

This section provides the result following the application of three causality detection approaches before and after filtering the expression profiles using SSA. For all evaluations, we ensure that all the test requirements are satisfied by choosing the optimal indices. Table 6.2 illustrates the findings of the regulatory relationships where “Yes" stands for the detected regulatory relationship by the adopted test. It should be noted that for each time class, all the embryos available at FlyEx representing expression profiles for Bcd, Cad and Kr are studied. The p-values reported for time domain GC test are the average p-values attained for each time class. Moreover, for time domain GC test, the co-integration test is conducted only for those variables having one unit root. Since none of the tested groups showed a significant result in indicating the co-integration, the co-integration test result is not reported here. The optimal lag for each Vector Auto-regression Model (VAR) model is selected by comparing the information criteria matrix, which includes results based on the AIC [Akaike, 1973], HQ [Hannan and Quinn, 1979], SIC [Schwarz et al., 1978] and FPE [Akaike, 1969] criteria.

According to Table 6.2, it is evident that there is a significant difference in results before and after reducing the noise from the profiles. The regulatory link between Bcd and Cad can be detected by neither time domain nor frequency domain tests in presence of noise. Accordingly, it is clear that the filtering capability displayed by SSA is indeed advantageous for regulatory detection. Nevertheless, as can be seen, the feasibility of capturing the regulatory link for CCM method is not affected by noise and the results achieved by this test confirm the regulatory relationship between Bcd and Cad in expression profiles with and without noise. However, regardless of the developing time, the index representing the ability of cross mapping is relatively smaller on average for noisy series than filtered series.

It is of note that the length of the data under study vary between different cleavage cycles and time classes. Cleavage cycle 10 to 13 and time classes 14(7-8) have shorter lengths comparing to the time class 14(1-6), which may be the reason of getting slightly smaller p-values for cleavage cycles 11 to 13 and 14(8) comparing to the rest of the sub classes of time class 14. Yet, the frequency domain test shows less sensitivity to the data length possibly because this method identifies the possible regulatory link for each individual frequency component rather than the entire series.

	Time Domain GC				**Frequency Domain GC**		**CCM**	
Time Class	Noisy Series		Filtered Series		Noisy Series	Filtered Series	Noisy Series	Filtered Series
	YES/NO	p-value	YES/NO	p-value	YES/NO	YES/NO	YES/NO	YES/NO
10	NO	0.68	NO	0.45	NO	YES	YES	YES
11	NO	0.71	NO	0.33	NO	YES	YES	YES
12	NO	0.89	NO	0.32	NO	YES	YES	YES
13	NO	0.89	NO	0.24	NO	YES	YES	YES
14(1)	NO	0.95	YES	0.05	NO	YES	YES	YES
14(2)	NO	0.98	YES	0.04	NO	YES	YES	YES
14(3)	NO	0.98	YES	0.01	NO	YES	YES	YES
14(4)	NO	0.94	YES	0.01	NO	YES	YES	YES
14(5)	NO	0.95	YES	0.00	NO	YES	YES	YES
14(6)	NO	0.96	YES	0.00	NO	YES	YES	YES
14(7)	NO	0.81	YES	0.00	NO	YES	YES	YES
14(8)	NO	0.79	YES	0.04	NO	YES	YES	YES

Table 6.2 Summary of causality test results of Bcd on Cad.Note: Differentiations are taken accordingly for stationarity prior to the tests; Optimal lag lengthes are chosen based on the AIC, HQ, SIC and FPE criteria. "Yes" stands for the detected regulatory link and "No" means the regulatory link could not be detected by the adopted test.

Furthermore, the p-values obtained for both noisy and filtered data of all the embryos in different time classes are summarised in Figs 6.3 and 6.4 as box and whisker diagram respectively. These diagrams follow the standard format of box plot in displaying the distribution of the p-values based on maximum, upper quartile, median, lower quartile, and minimum. A close look at Figs 6.3 and 6.4 suggests that the time domain GC test cannot detect any regulatory links in the presence of noise, while the results for filtered series are significant and more consistent especially for those time classes after 14(1). Moreover, it is evident that the length of the profiles and level of intensities affect the noisy profiles more than the filtered ones as the p-values in Fig 6.3 are getting more insignificant for the final subclasses of time class 14, where the length and intensity levels of the profiles tend to decrease.

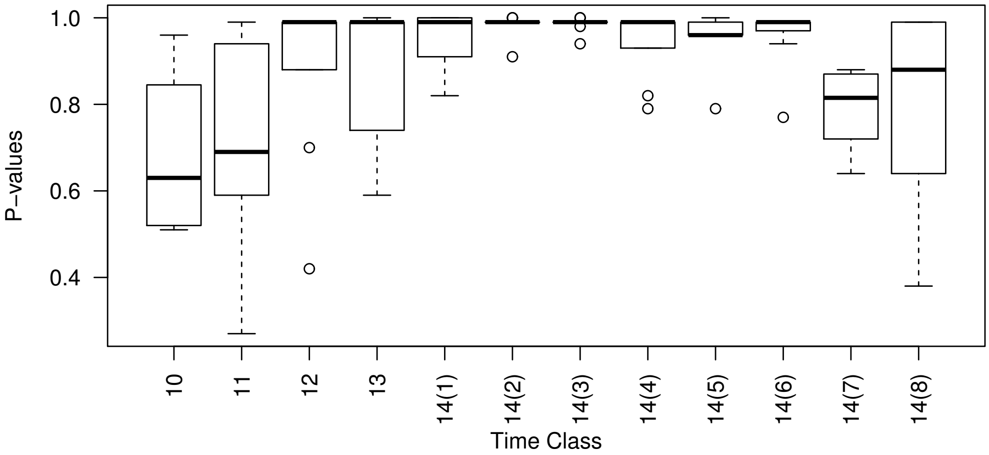


Fig. 6.3 Box Plots of Time Domain GC Test P-values for Noisy Series. (Circle refers to the corresponding outlier that is more/less than 1.5 times of upper/lower quartile; the central rectangle spans the upper quartile to the lower quartile; the segment inside the rectangle indicates the median; whiskers above and below the box refer to the maximum and minimum.)

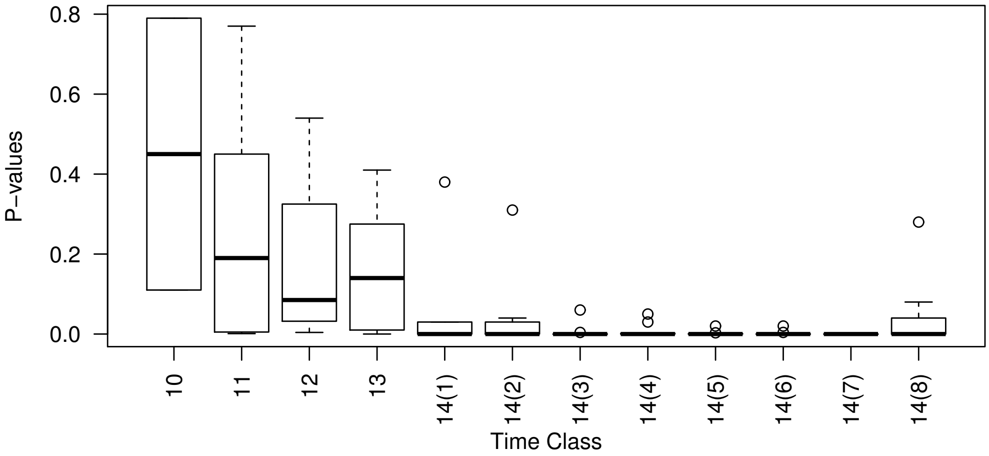


Fig. 6.4 Box Plots of Time Domain GC Test P-values for Filtered Series. (Circle refers to the corresponding outlier that is more/less than 1.5 times of upper/lower quartile; the central rectangle spans the upper quartile to the lower quartile; the segment inside the rectangle indicates the median; whiskers above and below the box refer to the maximum and minimum.)

Tables 6.3 and 6.4 present the results of the conducted analysis to detect the regulatory link between Bcd–Kr profiles and Cad–kr profiles respectively. As can be seen, reducing the noise level is an essential step in detecting the regulatory link using the time domain and frequency domain tests. Similar to the results reported in Table 6.2, CCM method can again efficiently identify the regulatory relationship even in the presence of noise.

	Time Domain GC				**Frequency Domain GC**		**CCM**	
Time Class	Noisy Series		Filtered Series		Noisy Series	Filtered Series	Noisy Series	Filtered Series
	YES/NO	p-value	YES/NO	p-value	YES/NO	YES/NO	YES/NO	YES/NO
12	NO	0.71	NO	0.15	NO	YES	YES	YES
13	NO	0.66	YES	0.04	NO	YES	YES	YES
14(1)	NO	0.89	YES	0.03	NO	YES	YES	YES
14(2)	NO	0.93	YES	0.01	NO	YES	YES	YES
14(3)	NO	0.97	YES	0.01	NO	YES	YES	YES
14(4)	NO	0.94	YES	0.00	NO	YES	YES	YES
14(5)	NO	0.95	YES	0.00	NO	YES	YES	YES
14(6)	NO	0.92	YES	0.00	NO	YES	YES	YES
14(7)	NO	0.81	YES	0.00	NO	YES	YES	YES

Table 6.3 A summary of the causality tests results for Bcd on Kr profiles. Note: Differentiations are taken accordingly for stationarity prior to the tests; Optimal lag lengthes are chosen based on the AIC, HQ, SIC and FPE criteria. "Yes" stands for the detected regulatory link and "No" means the regulatory link could not be detected by the adopted test.

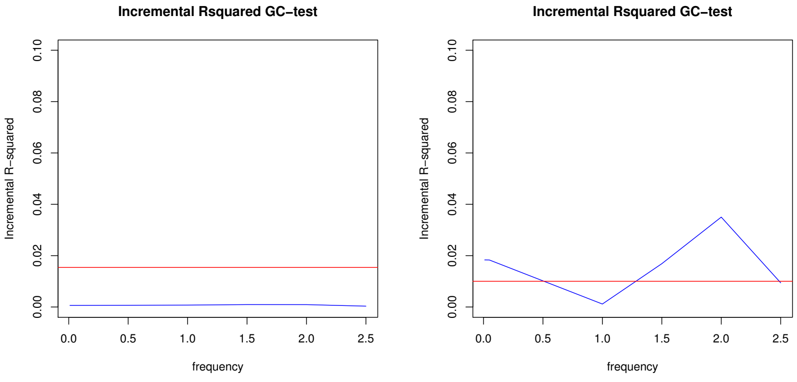
	Time Domain GC				**Frequency Domain GC**		**CCM**	
Time Class	Noisy Series		Filtered Series		Noisy Series	Filtered Series	Noisy Series	Filtered Series
	YES/NO	p-value	YES/NO	p-value	YES/NO	YES/NO	YES/NO	YES/NO
12	NO	0.39	NO	0.25	NO	YES	YES	YES
13	NO	0.78	NO	0.11	NO	YES	YES	YES
14(1)	NO	0.84	YES	0.05	NO	YES	YES	YES
14(2)	NO	0.89	YES	0.03	NO	YES	YES	YES
14(3)	NO	0.94	YES	0.01	NO	YES	YES	YES
14(4)	NO	0.91	YES	0.01	NO	YES	YES	YES
14(5)	NO	0.87	YES	0.00	NO	YES	YES	YES
14(6)	NO	0.82	YES	0.00	NO	YES	YES	YES
14(7)	NO	0.75	YES	0.00	NO	YES	YES	YES

Table 6.4 A summary of the causality tests results for Cad on Kr profiles.Note: Differentiations are taken accordingly for stationarity prior to the tests; Optimal lag lengthes are chosen based on the AIC, HQ, SIC and FPE criteria. "Yes" stands for the detected regulatory link and "No" means the regulatory link could not be detected by the adopted test.

Figures 6.5, 6.6 and 6.7 depict examples of the results obtained by frequency domain GC test for Bcd–Cad, Bcd–Kr and Cad–Kr profile pairs respectively [1]. In these Figs, the

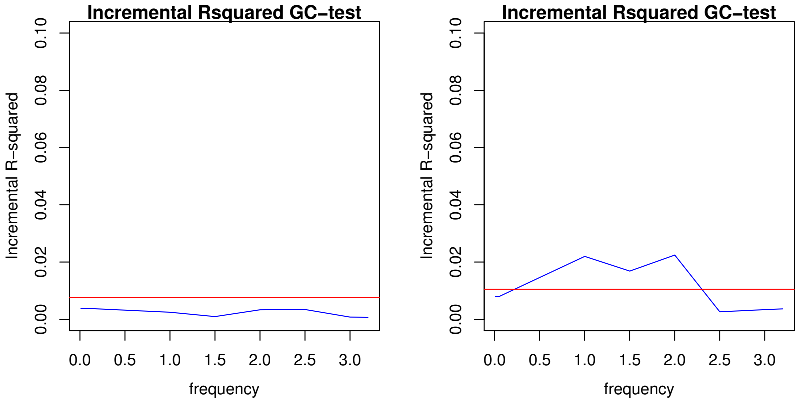
[1] The frequency domain GC test results for all considered pairs of genes related to all different time classes can be found in Appendix 3.

blue line represents the statistic test of each specific frequency, and the red line represents the 5% critical value for all the frequencies. The horizontal axis gives the parameter w to calculate the corresponding frequency f by $f = 2\pi/w$. Therefore, when the test statistics is above or very close to the 5% critical value, the regulatory is detected for that corresponding frequency.



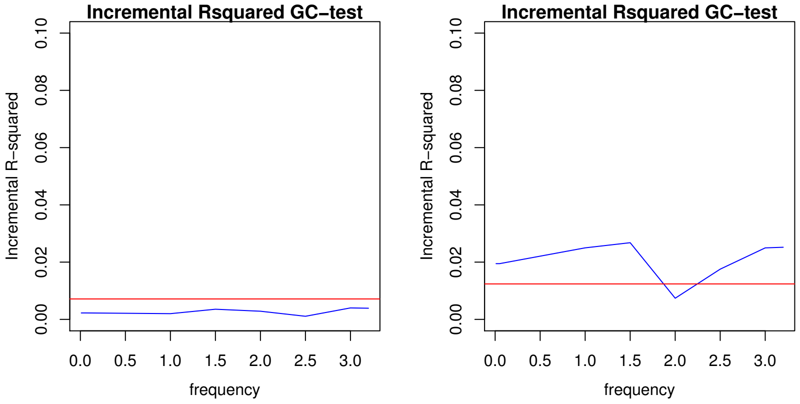
(a) Noisy-t11-bcd on cad (b) Filtered-t11-bcd on cad

Fig. 6.5 Frequency domain causality test results for Bcd and Kr before and after filtering (time class 12). The blue line represents the statistic test of each specific frequency, and the red line represents the 5% critical value for all the frequencies.



(a) Noisy-t12-bcd on kr (b) Filtered-t12-bcd on kr

Fig. 6.6 Frequency domain causality test results for Bcd and Kr before and after filtering (time class 12). The blue line represents the statistic test of each specific frequency, and the red line represents the 5% critical value for all the frequencies.



(a) Noisy-t12-cad on kr

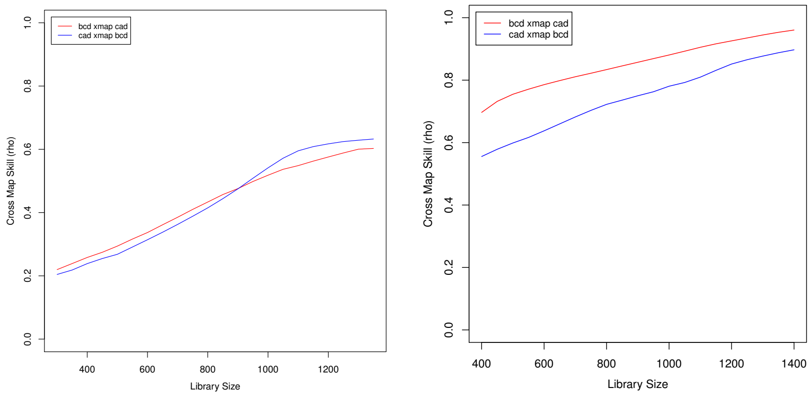
(b) Filtered-t12-cad on kr

Fig. 6.7 Frequency domain causality test results for Cad and Kr before and after filtering (time class 12). The blue line represents the statistic test of each specific frequency, and the red line represents the 5% critical value for all the frequencies.

For CCM test, the optimal embedding dimension E is selected for each pair of gene expression profiles based on the nearest neighbor forecasting performance by simplex projection. Figs 6.8, 6.9 and 6.10 represent the examples of the CCM test result for Bcd–Cad, Bcd–Kr and Cad–Kr before and after filtering the profiles [2], where for example regrding the Fig 6.8, the red line indicates the reconstruction ability of Bcd cross mapping Cad, while the blue line represents the performance of using historical information of Cad on cross mapping Bcd. In general, the higher ability of factor X on reconstructing the attractor reflects more significant regulatory effects of the attractor on X. The results of CCM reflect close relationships between Bcd and Cad with and without filtering, whilst Bcd shows more significant relationship with Kr comparing to Cad for both original and filtered data. The crossmap abilities of Bcd and Cad on Kr are fairly similar, however, Kr clearly indicates higher reconstruction ability on Bcd comparing to Cad. In more details regarding the relationship between Bcd and Cad, considering the average reconstruction ability represented by ρ, it is suggested that CCM is not affected by the smaller length of the series related to the initial time. However, the increasing pattern of the average level of cross-mapping ability

[2] The CCM test results for all considered pairs of genes related to all different time classes can be found in Appendix 3.

up to time class 14(3), which follows by a decreasing trend for the rest of the subclasses, indicates less accuracy of the results for higher time classes. The approximate average value of ρ over 0.5 for noisy series indicates significant cross-mapping (or reconstruction) ability to identify the regulatory links. Correspondingly, an average is found to be approximately over 0.8, which reflects stronger regulatory links detected between Bcd and Cad after filtering. Regarding the relationships between Bcd and Kr, both original and filtered series indicate stronger cross-mapping ability from Kr to Bcd, which means that Bcd shows a more powerful regulatory effect on Kr than the other way around. However, this link is slightly more significant in the filtered profiles. In the case of Cad and Kr, the regulatory relationship identified is less significant comparing to the other pairs of genes considered in this study and the average of 0.4 for filtered profiles compared to the average of 0.2 for original series highlights the role of the SSA in improving the achieved results.

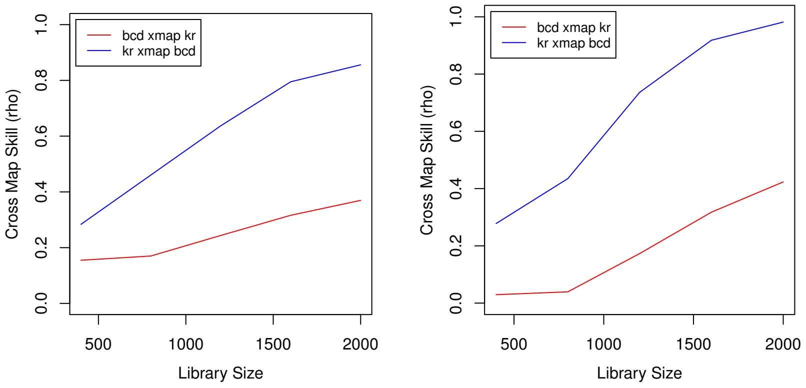


(a) Noisy-t14(8)-ccm

(b) Filtered-t14(8)-ccm

Fig. 6.8 CCM test results for Bcd and Cad before and after filtering (time class 14(8)). The red line indicates the reconstruction ability of Bcd crossmap Cad, while the blue line represents the performance of Cad on crossmapping Bcd.

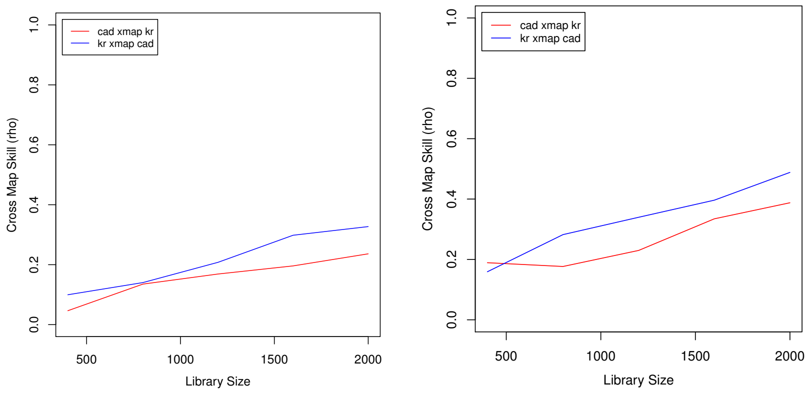
It is of note that the overall findings of this research are consistent with the previous efforts in mathematical modelling the segmentation network [Gursky et al., 2011; Surkova et al., 2009]. For example, [Surkova et al., 2009] presents a succesful canalization study of



(a) Noisy-t14(7)-ccm

(b) Filtered-t14(7)-ccm

Fig. 6.9 CCM test results for Bcd and Kr before and after filtering (time class 14(7)). The red line indicates the reconstruction ability of Bcd crossmap Kr, while the blue line represents the performance of Kr on crossmapping Bcd.



(a) Noisy-t14(5)-ccm

(b) Filtered-t14(5)-ccm

Fig. 6.10 CCM test results for Cad and Kr before and after filtering (time class 14(5)). The red line indicates the reconstruction ability of Cad crossmap Kr, while the blue line represents the performance of Kr on crossmapping Cad.

four gap genes *hb*, *gt kni* kr using the gene circuit method which uses the concentration of *bcd*, *cad*, *tll* and genes as outside inputs.

6.5 Conclusion

In developmental studies, inferring the regulatory interactions in SN plays an important role in unveiling the mechanism responsible for pattern formation. As such, there exists an opportune demand for theoretical developments and new mathematical models to achieve a more accurate illustration of this genetic network. Accordingly, this chapter introduces a new method to extract the meaningful regulatory role of maternal effect genes using a variety of causality detection techniques and also explores whether this method can suggest a new analytical view to study the gene regulatory networks.

To that aim, three causality detection approaches before and after filtering the expression profiles using SSA are applied. According to the obtained results, data accuracy is of critical importance to the success of the causality detection studies. Using time domain and frequency domain GC tests, the regulatory link can be detected only after removing the noise from the expression profiles which indicates even having a small amount of noise in mean intensities may lead us to obtain a false negative result.

It is also imperative to note that for all pairs of genes considered in this study, the time domain GC fails to detect the regulatory link in time classes 10-13. The poor performance of this model here can be attributed to either the length of the data or low expression level of the gene under study for those time classes. The protein molecules synthesised from maternal transcripts just begin to appear from time class 10 and the number of these morphogens, in the areas where they were concentrated, is at a lower amount for time classes 10-13 comparing to the higher time classes.

In summary, Our findings show that the regulatory role of maternal effect genes is detectable in different cleavage cycles and time classes and thereby the introduced method can be applied to infer the possible regulatory interactions present among the other genes of this network.

Chapter 7

Conclusions and Future Research Directions

This thesis begins with an overall aim of improving signal processing step in analysing gene expression profiles using SSA technique. To that aim, *D. melanogaster* is considered as the model organism and the protein profile of *bcd* gene which is the first and most critical input of the SN is chosen to be the primer gene expression profile under study [Berleth et al., 1988; Driever and Nüsslein-Volhard, 1988; Frohnhöfer and Nüsslein-Volhard, 1986]. Chapter one begins with an introduction to the subject highlighting the noteworthy contribution of this research not only in the field of gene expression studies but also in the area of applied statistics where signal extraction, noise filtering, forecasting and network analysis goals are aimed to study. An introduction to the methodology in addition to concise portrays of benchmark signal processing models and metrics adopted in this study are presented in Chapter 3. The remaining Chapters (up until Chapter 7) successfully present the four outlined objectives of this research.

here talk about bcd The significantly stochastic, highly volatile and non-normal nature of Bcd

This Chapter is divided into two sections. Section 7.1 brings the thesis to a conclusion by presenting the contribution of the current work and section 7.2 discusses the future work.

7.1 Summary of Findings

The first contribution of this thesis is to introduce a new method for reducing the noise and accordingly to extracting the Bcd signal from its noisy profile. Hence, in Chapter 4 an enhanced version of SSA known as SSA based on minimum variance estimator SSA_{MV} is introduced. Both the simulation studies and empirical results presented in that Chapter shows SSA_{MV} to be more efficient than the basic version of SSA.

The performance of the SSA_{MV} is compared against several parametric and non-parametric well established signal processing techniques (ARIMA, ARFIMA, ETS and NN models) as well as SDD model as the most commonly used model for analysing the Bcd profile [Bergmann et al., 2007; Driever and Nüsslein-Volhard, 1988; Gregor et al., 2005; Houchmandzadeh et al., 2002]. In that Chapter, the efficiency of adopted noise filtering techniques is evaluated using several criteria including RMSE, MAPE and MAE for simulation study and w, pearsons, spearman and kendall correlations for empirical results. The attained RMSE, MAE and MAPE values in the simulation results indicate that SSA_{MV} provides the best signal extraction and is successful at outperforming the SDD, ARIMA, ARFIMA, ETS and NN models. Moreover, in analysing the real data sets a satisfactory level of separation between noise and signal was achieved by SSA_{MV} providing an average signal–noise correlation value below 0.10 in 90% of the studied profiles.

Because of examining the applicability of several parametric methods in addition to non-parametric ones, the ADF test is used to examine for stationarity where the obtained results show that the residuals are in fact nonstationary. More importantly, approved by Ljung-Box test, the SDD residuals are not white noise suggesting that there are some components of signal remained in the residuals which motivated us to thoroughly explore the residuals in in the next Chapter.

Having depicted the superior performance of the SSA technique in Chapter 4, the second contribution of this thesis provides five different approaches to optimise this algorithm.

As discussed before the performance of the SSA technique highly depends upon its choice of window length L and the number of eigenvalues r. Accordingly, we initially produce an algorithm for optimising the Bcd signal extraction following setting a new criterion for choosing the best value for L. In brief, the algorithm is optimised based on minimising the skewness statistic for the SSA residual. We suggest that setting L equal to the minimum skewness within the threshold $10 \geq L \geq N/4$ and combine this SSA choice with $r = 1$ or $r = 1,2$ as appropriate will enable users to obtain the optimal Bcd signal extraction with SSA. The main advantage of this algorithm is producing a smooth and accurate Bcd signal quickly and easily without the need to spend an increased amount of time for the selection of L.

It is of note that in some applications such as gene expression signal processing, reducing the running time of an algorithm is of critical importance because some of the profiles achieved at specific developmental times possess a high degree of series length. As was reported in Chapter 1, the length of the data in some Bcd profiles exceeds 3000 observations giving a range between 2 to 1500 as the possible values for L. Therefore, setting a new criterion which successfully allows exploring a precise signal in a less amount of time would be highly beneficial. The application of the introduced algorithm on the real data which produces smooth and well-centred signals has illustrated the feasibility of the proposed theory.

In some cases, the extracted signal using SSA contained small levels of fluctuations. In such instances, we suggest exploiting sequential SSA, not on the residual, but on the extracted signal to smooth it further. Sequential SSA provides users with a single non-parametric algorithm for Bcd signal extraction. The results obtained from different Bcd profiles confirm the capability of the sequential SSA in capturing the maximum variation via a smooth signal line and also highlights the usefulness and applicability of the proposed test.

Furthermore, exploring the residuals in more details has led us to introduce the hybrid Bcd signal extraction methods. The first hybrid SSA signal process is a combination of optimised SSA with optimised ARIMA (popular in signal processing literature) being fitted to the residuals and ensures providing the white noise residuals. The extracted signal achieved following this procedure is no longer smooth. Since the small fluctuations modelled using ARIMA may be attributed to the biological noise, the hybrid SSA–ARIMA model can be applied in studies where a combination of Bcd signal with its biological noise is considered as an input to the network simulation system.

The second hybrid model introduced in this thesis is a combination of non-parametric models. The SSA–ETS hybrid presents a great contribution to Bcd signal extraction by finding a solution to the problem of modelling the initial curve in Bcd data. The initial curve in Bcd profiles attributes to the concentration of Bcd in nuclei at the anterior end of the AP axis which shows that Bcd concentration initially reaches a maximum value before decaying along the embryo and suggests a more complicated mechanism of morphogen gradient readout.

The third approach of optimising the Bcd signal extraction introduced in Chapter 5 focuses on finding the number of eigenvalues r as the second and final choice of the SSA technique. This approach is based on the distribution of the eigenvalues of a scaled Hankel matrix. To that aim, a new step is added to the algorithm which comes between the two stages of SSA and divides the matrix $\mathbf{A}$ by its trace, $\mathbf{A}/tr(\mathbf{A})$. Let $\zeta_1,\ldots,\zeta_L$ denote the eigenvalues of the matrix $\mathbf{A}/tr(\mathbf{A})$ in decreasing order of magnitude ($1 \geq \zeta_1 \geq \ldots \geq \zeta_L \geq 0$). In this step, we perform the simulation technique to gain the distribution of ζ_i, so we can understand the behaviour of each eigenvalue, which can be useful for obtaining the proper value of r. In this study, our aim is to ascertain the distribution of ζ_i and its related forms that can be used directly for choosing the optimal value of r for the genes signal extraction. Therefore, the skewness and variation coefficients of the eigenvalue distribution are used as new criteria and indicator for the cut-off point in the eigenvalue spectra between signal and noise components. Thereafter, this procedure is evaluated for three other members of the SN, *cad*, *gt* and *eve*. The obtained results confirm that in the extracting signal from different expression gene profiles, for optimised signal and noise separation, a different number of eigenvalues need to be chosen for each gene. The new introduced algorithm enables both automation and optimisation of SSA.

Chapter 5 also introduces a novel approach for enhancing the accuracy of SSA signal extraction following studying the foundations of Colonial Theory. In brief, the proposed algorithm suggests that the binary approach for differentiating between signal and noise at the grouping step does not necessarily give the best outcome as the binary approach assumes that there is no useful information contained in the selected noise components. Instead, based on CT we propose a different grouping approach which considers analysing all L and r combinations and according to the precision criteria of choose, selecting those eigenvalues which have useful information for reconstructing the signal.

The final contribution of this research (as discussed in Chapter 6) is the development of a new approach for extracting the meaningful regulatory links between the genes present in the SN using a hybrid model comprised of optimised SSA and a variety of causality detection techniques. The principal aim here is to explore the possibility of determining the genes in SNwhich are regulated in response to Bcd signal. Since the data accuracy is of critical importance to the success of the causality detection studies, the role of SSA in this model is refining the data and providing the next step of the analysis with the required input (i.e signals of gene expression profiles). According to the obtained results, using time domain and frequency domain GC tests, the regulatory link can be detected only after removing the noise from the expression profiles which indicates even having a small amount of noise in mean intensity may lead us to obtain a false negative result. However, the CCM method presents more resistant to the noise and fluctuations of data.

7.2 Future Directions

While this thesis demonstrates innovative computational techniques for analysing segmentation gene expression profiles, in particular, the gradient of Bcd during *D. melanogaster* embryonic development, many opportunities for extending the scope of this thesis remain. This section presents some of these directions.

As already mentioned in Chapter 1, FlyEx provides the most commonly used database for quantitative segmentation gene expression profiles in *D. melanogaster*. However, in addition to having a considerable level of noise, FlyEx presents a number of challenges which addressing them would be highly advantageous when developing a new database for future studies:

Firstly, since the fixing process causes the death of the embryos prior to imaging, in FlyEx, the gene expression profiles are obtained from nonliving embryos [Van Soom et al., 2003]. The achieved information from a single time point (i.e. the time that the embryo was fixed) makes the data to be far from ideal [Van Soom et al., 2003]. Moreover, the confocal scans are achieved from flattened embryos, squeezed under a cover glass causing an arbitrary orientation of embryos under the cover glass.

Secondly, following the procedure developed in FlyEx, a single embryo can be imaged only for up to three segmentation genes. This issue limits the analysis of the SN, where the interaction between different segmentation genes is explored. As would expect a database containing the quantitative gene expression data of several genes obtained from the same embryo will greatly improve the accuracy of the analysis presented in Chapter 6.

This thesis thoroughly explores the expression profile of *bcd* gene. However, there are several other morphogen gradients, which control the gene expression similar to the way that *bcd* does, including Decapentaplegic (*Dpp*) which makes a contribution to the *Drosophila* DV axis and the wing imaginal disc [Ferguson and Anderson, 1992] and Sonic hedgehog homolog (Shh), a critical gene for the growth of digits on limbs and the brain in humans [Roelink et al., 1995]. Therefore, a more thorough analysis of these genes can also be performed in the future by the modelling tools introduced in our work.

Regarding the future studies, more extensive research into hybrid signal extraction processes is likely to result in positive, vital and interesting outcomes as clearly shown via Chapter 5. Researchers should evaluate a variety of different signal extraction techniques within the hybrid framework proposed in this thesis to ascertain whether outcomes could be further improved.

To obtain a more accurate illustration of the SN, there exists an opportune demand for theoretical developments and new mathematical models representing this network. Therefore, we believe that the findings presented in Chapter 5, open up several avenues for future research. Introducing the statistical distribution and model parameters of the other members of this network lays the necessary groundwork for the future goal of modelling a dynamic SN and an automated mutant recognition system.

Chapter 6 of this thesis opens up an entirely new area of research for analysing the regulatory links between the genes cooperating in a GRN. Hence, it would be insightful to assess the cross-regulatory links existing between the entire members of this network. To extend this line of research, new optimised algorithms should be developed which can successfully address the limitation of the implementation time.

References

Adl, S. M., Simpson, A. G., Farmer, M. A., Andersen, R. A., Anderson, O. R., Barta, J. R., Bowser, S. S., Brugerolle, G., Fensome, R. A., Fredericq, S., et al. (2005). The new higher level classification of eukaryotes with emphasis on the taxonomy of protists. *Journal of Eukaryotic Microbiology*, 52(5):399–451.

Akaike, H. (1969). Fitting autoregressive models for prediction. *Annals of the institute of Statistical Mathematics*, 21(1):243–247.

Akaike, H. (1973). Maximum likelihood identification of gaussian autoregressive moving average models. *Biometrika*, 60(2):255–265.

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., and Walter, P. (2002). Drosophila and the molecular genetics of pattern formation: Genesis of the body plan.

Alexandrov, T. (2008). A method of trend extraction using singular spectrum analysis. *arXiv preprint arXiv:0804.3367*.

Alexandrov, T., Bianconcini, S., Dagum, E. B., Maass, P., and McElroy, T. S. (2012). A review of some modern approaches to the problem of trend extraction. *Econometric Reviews*, 31(6):593–624.

Alexandrov, T., Golyandina, N., and Spirov, A. (2008). Singular spectrum analysis of gene expression profiles of early drosophila embryo: exponential-in-distance patterns. *Research letters in signal processing*, 2008:12.

Ancona, N., Marinazzo, D., and Stramaglia, S. (2004). Radial basis function approach to nonlinear granger causality of time series. *Physical Review E*, 70(5):056221.

Andrews, S. S. and Bray, D. (2004). Stochastic simulation of chemical reactions with spatial resolution and single molecule detail. *Physical biology*, 1(3):137.

Arbeitman, M. N., Furlong, E. E., Imam, F., Johnson, E., Null, B. H., Baker, B. S., Krasnow, M. A., Scott, M. P., Davis, R. W., and White, K. P. (2002). Gene expression during the life cycle of drosophila melanogaster. *Science*, 297(5590):2270–2275.

Arnold, L., Horsthemke, W., and Lefever, R. (1978). White and coloured external noise and transition phenomena in nonlinear systems. *Zeitschrift für Physik B Condensed Matter*, 29(4):367–373.

Atikur Rahman Khan, M. and Poskitt, D. (2013). A note on window length selection in singular spectrum analysis. *Australian & New Zealand Journal of Statistics*, 55(2):87–108.

Baird-Titus, J. M., Clark-Baldwin, K., Dave, V., Caperelli, C. A., Ma, J., and Rance, M. (2006). The solution structure of the native k50 bicoid homeodomain bound to the consensus taatcc dna-binding site. *Journal of molecular biology*, 356(5):1137–1151.

Bell, G. and Mooers, A. O. (1997). Size and complexity among multicellular organisms. *Biological Journal of the Linnean Society*, 60(3):345–363.

Benedetto, F., Giunta, G., and Mastroeni, L. (2015). A maximum entropy method to assess the predictability of financial and commodity prices. *Digital Signal Processing*, 46:19–31.

Benesty, J., Chen, J., Huang, Y., and Cohen, I. (2009). Pearson correlation coefficient. In *Noise reduction in speech processing*, pages 1–4. Springer.

Berezhkovskii, A. M., Coppey, M., and Shvartsman, S. Y. (2009). Signaling gradients in cascades of two-state reaction-diffusion systems. *Proceedings of the National Academy of Sciences*, 106(4):1087–1092.

Bergmann, S., Sandler, O., Sberro, H., Shnider, S., Schejter, E., Shilo, B.-Z., and Barkai, N. (2007). Pre-steady-state decoding of the bicoid morphogen gradient. *PLoS Biol*, 5(2):e46.

Berleth, T., Burri, M., Thoma, G., Bopp, D., Richstein, S., Frigerio, G., Noll, M., and Nüsslein-Volhard, C. (1988). The role of localization of bicoid rna in organizing the anterior pattern of the drosophila embryo. *The EMBO journal*, 7(6):1749.

Bieler, J., Pozzorini, C., and Naef, F. (2011). Whole-embryo modeling of early segmentation in drosophila identifies robust and fragile expression domains. *Biophysical journal*, 101(2):287–296.

Bonabeau, E., Dorigo, M., and Theraulaz, G. (1999). *Swarm intelligence: from natural to artificial systems*. Number 1. Oxford university press.

Bozdech, Z., Llinás, M., Pulliam, B. L., Wong, E. D., Zhu, J., and DeRisi, J. L. (2003). The transcriptome of the intraerythrocytic developmental cycle of plasmodium falciparum. *PLoS Biol*, 1(1):e5.

Breitung, J. and Candelon, B. (2006). Testing for short-and long-run causality: A frequency-domain approach. *Journal of Econometrics*, 132(2):363–378.

Bremer, M. and Doerge, R. (2009). The km-algorithm identifies regulated genes in time series expression data. *Advances in bioinformatics*, 2009.

Broomhead, D. S. and King, G. P. (1986). Extracting qualitative dynamics from experimental data. *Physica D: Nonlinear Phenomena*, 20(2-3):217–236.

Burman, J. P. (1980). Seasonal adjustment by signal extraction. *Journal of the Royal Statistical Society. Series A (General)*, pages 321–337.

Chaves, M., Albert, R., and Sontag, E. D. (2005). Robustness and fragility of boolean models for genetic regulatory networks. *Journal of theoretical biology*, 235(3):431–449.

Chen, Y., Bressler, S. L., and Ding, M. (2006). Frequency decomposition of conditional granger causality and application to multivariate neural field potential data. *Journal of neuroscience methods*, 150(2):228–237.

Cheung, Y.-W. and Lai, K. S. (1995). Lag order and critical values of the augmented dickey–fuller test. *Journal of Business & Economic Statistics*, 13(3):277–280.

Chichilnisky, E. (2001). A simple white noise analysis of neuronal light responses. *Network: Computation in Neural Systems*, 12(2):199–213.

Child, C. M. (1941). Patterns and problems of development.

Ciner, C. (2011). Eurocurrency interest rate linkages: A frequency domain analysis. *International Review of Economics & Finance*, 20(4):498–505.

Clark, A. T., Ye, H., Isbell, F., Deyle, E. R., Cowles, J., Tilman, G. D., and Sugihara, G. (2015). Spatial convergent cross mapping to detect causal relationships from short time series. *Ecology*, 96(5):1174–1181.

Copf, T., Schröder, R., and Averof, M. (2004). Ancestral role of caudal genes in axis elongation and segmentation. *Proceedings of the National Academy of Sciences of the United States of America*, 101(51):17711–17715.

Coppey, M., Berezhkovskii, A. M., Kim, Y., Boettiger, A. N., and Shvartsman, S. Y. (2007). Modeling the bicoid gradient: diffusion and reversible nuclear trapping of a stable protein. *Developmental biology*, 312(2):623–630.

Crauk, O. and Dostatni, N. (2005). Bicoid determines sharp and precise target gene expression in the drosophila embryo. *Current Biology*, 15(21):1888–1898.

Crick, F. (1970). Diffusion in embryogenesis.

Croux, C. and Reusens, P. (2013). Do stock prices contain predictive power for the future economic activity? a granger causality analysis in the frequency domain. *Journal of Macroeconomics*, 35:93–103.

Davidson, E. and Levin, M. (2005). Gene regulatory networks. *Proceedings of the national academy of sciences of the United States of America*, 102(14):4935–4935.

De Jong, H. (2002). Modeling and simulation of genetic regulatory systems: a literature review. *Journal of computational biology*, 9(1):67–103.

De Moor, B. (1993). The singular value decomposition and long and short spaces of noisy matrices. *IEEE transactions on signal processing*, 41(9):2826–2838.

Deng, J., Wang, W., Lu, L. J., and Ma, J. (2010). A two-dimensional simulation model of the bicoid gradient in drosophila. *PLoS One*, 5(4):e10275.

Deshpande, G., Sathian, K., and Hu, X. (2010). Effect of hemodynamic variability on granger causality analysis of fmri. *Neuroimage*, 52(3):884–896.

Dewar, M. A., Kadirkamanathan, V., Opper, M., and Sanguinetti, G. (2010). Parameter estimation and inference for stochastic reaction-diffusion systems: application to morphogenesis in d. melanogaster. *BMC Systems Biology*, 4(1):1.

Deyle, E. R., Fogarty, M., Hsieh, C.-h., Kaufman, L., MacCall, A. D., Munch, S. B., Perretti, C. T., Ye, H., and Sugihara, G. (2013). Predicting climate effects on pacific sardine. *Proceedings of the National Academy of Sciences*, 110(16):6430–6435.

Dilão, R. and Muraro, D. (2010). mrna diffusion explains protein gradients in drosophila early development. *Journal of theoretical biology*, 264(3):847–853.

Dorigo, M. (1992). Optimization, learning and natural algorithms. *Ph. D. Thesis, Politecnico di Milano, Italy*.

Driesch, H. (1929). The science & philosophy of the organism.

Driever, W. and Nüsslein-Volhard, C. (1988). The bicoid protein determines position in the drosophila embryo in a concentration-dependent manner. *Cell*, 54(1):95–104.

Du, L., Wu, S., Liew, A. W.-C., Smith, D. K., and Yan, H. (2008). Spectral analysis of microarray gene expression time series data of plasmodium falciparum. *International journal of bioinformatics research and applications*, 4(3):337–349.

Dubreux, S. (1998). Atlas de poche de microbiologie. *Lyon Pharmaceutique*, 2(49):113.

Eiben, A. E., Raue, P.-E., and Ruttkay, Z. (1994). Genetic algorithms with multi-parent recombination. In *International Conference on Parallel Problem Solving from Nature*, pages 78–87. Springer.

Elsner, J. B. and Tsonis, A. A. (1996). *Singular Spectrum Analysis: A New Tool in Time Series Analysis*. Plenum.

Ephrussi, A. and St Johnston, D. (2004). Seeing is believing: the bicoid morphogen gradient matures. *Cell*, 116(2):143–152.

Faes, L. and Nollo, G. (2013). Measuring frequency domain granger causality for multiple blocks of interacting time series. *Biological cybernetics*, 107(2):217–232.

Fedoroff, N. and Fontana, W. (2002). Small numbers of big molecules. *Science*, 297(5584):1129–1131.

Ferguson, E. L. and Anderson, K. V. (1992). Decapentaplegic acts as a morphogen to organize dorsal-ventral pattern in the drosophila embryo. *Cell*, 71(3):451–461.

Fisher, R. A. et al. (1949). The design of experiments. *The design of experiments.*, (Ed. 5).

Frigerio, G., Burri, M., Bopp, D., Baumgartner, S., and Noll, M. (1986). Structure of the segmentation gene paired and the drosophila prd gene set as part of a gene network. *Cell*, 47(5):735–746.

Fritschy, J.-M. and Härtig, W. (2001). *Immunofluorescence*. John Wiley & Sons, Ltd.

Frohnhöfer, H. G. and Nüsslein-Volhard, C. (1986). Organization of anterior pattern in the drosophila embryo by the maternal gene bicoid. *Nature*, 324:120–125.

Geweke, J. (1982). Measurement of linear dependence and feedback between multiple time series. *Journal of the American statistical association*, 77(378):304–313.

Ghodsi, M., Hassani, H., Sanei, S., and Hicks, Y. (2009). The use of noise information for detection of temporomandibular disorder. *Biomedical Signal Processing and Control*, 4(2):79–85.

Ghodsi, Z., Silva, E. S., and Hassani, H. (2015). Bicoid signal extraction with a selection of parametric and nonparametric signal processing techniques. *Genomics, proteomics & bioinformatics*, 13(3):183–191.

Gibson, M. A. and Bruck, J. (2000). Efficient exact stochastic simulation of chemical systems with many species and many channels. *The journal of physical chemistry A*, 104(9):1876–1889.

Golyandina, N., Nekrutkin, V., and Zhigljavsky, A. (2001a). *Analysis of time series structure: SSA and related techniques*. CRC Press.

Golyandina, N., Nekrutkin, V., and Zhigljavsky, A. A. (2001b). Analysis of time series structure: Ssa and related techniques (chapman & hall crc monographs on statistics & applied probability).

Golyandina, N. and Shlemov, A. (2013). Variations of singular spectrum analysis for separability improvement: non-orthogonal decompositions of time series. *arXiv preprint arXiv:1308.4022*.

Golyandina, N. and Usevich, K. (2010). 2d-extension of singular spectrum analysis: algorithm and elements of theory. *Matrix Methods: Theory, Algorithms and Applications*, pages 449–473.

Golyandina, N. and Zhigljavsky, A. (2013). *Singular Spectrum Analysis for time series*. Springer Science & Business Media.

Golyandina, N. E., Holloway, D. M., Lopes, F. J., Spirov, A. V., Spirova, E. N., and Usevich, K. D. (2012). Measuring gene expression noise in early drosophila embryos: nucleus-to-nucleus variability. *Procedia computer science*, 9:373–382.

Gonzalez-Romera, E., Jaramillo-Moran, M. A., and Carmona-Fernandez, D. (2006). Monthly electric energy demand forecasting based on trend extraction. *IEEE Transactions on power systems*, 21(4):1946–1953.

Gow, D. W., Segawa, J. A., Ahlfors, S. P., and Lin, F.-H. (2008). Lexical influences on speech perception: a granger causality analysis of meg and eeg source estimates. *Neuroimage*, 43(3):614–623.

Granger, C. W. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society*, pages 424–438.

Gregor, T., Bialek, W., van Steveninck, R. R. d. R., Tank, D. W., and Wieschaus, E. F. (2005). Diffusion and scaling during early embryonic pattern formation. *Proceedings of the National Academy of Sciences of the United States of America*, 102(51):18403–18407.

Gregor, T., Wieschaus, E. F., McGregor, A. P., Bialek, W., and Tank, D. W. (2007). Stability and nuclear dynamics of the bicoid morphogen gradient. *Cell*, 130(1):141–152.

Grimm, O., Coppey, M., and Wieschaus, E. (2010). Modelling the bicoid gradient. *Development*, 137(14):2253–2264.

Grimm, O. and Wieschaus, E. (2010). The bicoid gradient is shaped independently of nuclei. *Development*, 137(17):2857–2862.

Grosberg, R. K. and Strathmann, R. R. (2007). The evolution of multicellularity: a minor major transition? *Annual Review of Ecology, Evolution, and Systematics*, pages 621–654.

Groth, A. and Ghil, M. (2011). Multivariate singular spectrum analysis and the road to phase synchronization. *Physical Review E*, 84(3):036206.

Gursky, V. V., Panok, L., Myasnikova, E. M., Manu, M., Samsonova, M. G., Reinitz, J., and Samsonov, A. M. (2011). Mechanisms of gap gene expression canalization in the drosophila blastoderm. *BMC systems biology*, 5(1):1.

Halliday, R. (2004). Equity trend prediction with neural networks.

Hannan, E. J. and Quinn, B. G. (1979). The determination of the order of an autoregression. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 190–195.

Harvey, A. and Koopman, S. J. (2000). Signal extraction and the formulation of unobserved components models. *The Econometrics Journal*, 3(1):84–107.

Harwell, L. H. H. and Leroy Goldberg, M. L. (2004). *Genetics from genes to genomes*. Number 576.5 G328g. McGraw-Hill,.

Haslett, J. and Raftery, A. E. (1989). Space-time modelling with long-memory dependence: Assessing ireland's wind power resource. *Applied Statistics*, pages 1–50.

Hassani, H. (2007). Singular spectrum analysis: methodology and comparison. *Journal of Data Science*, 5(2):239–257.

Hassani, H. (2010). Singular spectrum analysis based on the minimum variance estimator. *Nonlinear Analysis: Real World Applications*, 11(3):2065–2077.

Hassani, H., Dionisio, A., and Ghodsi, M. (2010). The effect of noise reduction in measuring the linear and nonlinear dependency of financial markets. *Nonlinear Analysis: Real World Applications*, 11(1):492–502.

Hassani, H. and Ghodsi, Z. (2014). Pattern recognition of gene expression with singular spectrum analysis. *Medical Sciences*, 2(3):127–139.

Hassani, H., Heravi, S., Brown, G., and Ayoubkhani, D. (2013a). Forecasting before, during, and after recession with singular spectrum analysis. *Journal of Applied Statistics*, 40(10):2290–2302.

Hassani, H., Heravi, S., Brown, G., and Ayoubkhani, D. (2013b). Forecasting before, during, and after recession with singular spectrum analysis. *Journal of Applied Statistics*, 40(10):2290–2302.

Hassani, H., Heravi, S., and Zhigljavsky, A. (2009a). Forecasting european industrial production with singular spectrum analysis. *International journal of forecasting*, 25(1):103–118.

Hassani, H. and Mahmoudvand, R. (2013). Multivariate singular spectrum analysis: A general view and new vector forecasting approach. *International Journal of Energy and Statistics*, 1(01):55–83.

Hassani, H., Mahmoudvand, R., and Zokaei, M. (2011a). Separability and window length in singular spectrum analysis. *Comptes rendus mathematique*, 349(17):987–990.

Hassani, H., Mahmoudvand, R., Zokaei, M., and Ghodsi, M. (2012). On the separability between signal and noise in singular spectrum analysis. *Fluctuation and Noise Letters*, 11(02):1250014.

Hassani, H., Soofi, A. S., and Zhigljavsky, A. (2013c). Predicting inflation dynamics with singular spectrum analysis. *Journal of the Royal Statistical Society: Series a (Statistics in Society)*, 176(3):743–760.

Hassani, H. and Thomakos, D. (2010). A review on singular spectrum analysis for economic and financial time series. *Statistics and its Interface*, 3(3):377–397.

Hassani, H., Xu, Z., and Zhigljavsky, A. (2011b). Singular spectrum analysis based on the perturbation theory. *Nonlinear Analysis: Real World Applications*, 12(5):2752–2766.

Hassani, H., Zokaei, M., von Rosen, D., Amiri, S., and Ghodsi, M. (2009b). Does noise reduction matter for curve fitting in growth curve models? *Computer methods and programs in biomedicine*, 96(3):173–181.

Hattne, J., Fange, D., and Elf, J. (2005). Stochastic reaction-diffusion simulation with mesord. *Bioinformatics*, 21(12):2923–2924.

Hengenius, J. B., Gribskov, M., Rundell, A. E., Fowlkes, C. C., and Umulis, D. M. (2011). Analysis of gap gene regulation in a 3d organism-scale model of the drosophila melanogaster embryo. *PLoS One*, 6(11):e26797.

Herrero, J., Valencia, A., and Dopazo, J. (2001). A hierarchical unsupervised growing neural network for clustering gene expression patterns. *Bioinformatics*, 17(2):126–136.

Herron, M. D. and Michod, R. E. (2008). Evolution of complexity in the volvocine algae: transitions in individuality through darwin's eye. *Evolution*, 62(2):436–451.

Hickman, C. P., Roberts, L. S., Larson, A., Ober, W. C., and Garrison, C. (2001). *Integrated principles of zoology*, volume 15. JSTOR.

Hida, T., Kuo, H.-H., Potthoff, J., and Streit, W. (2013). *White noise: an infinite dimensional calculus*, volume 253. Springer Science & Business Media.

Hiemstra, C. and Jones, J. D. (1994). Testing for linear and nonlinear granger causality in the stock price-volume relation. *The Journal of Finance*, 49(5):1639–1664.

Hilfinger, A. and Paulsson, J. (2011). Separating intrinsic from extrinsic fluctuations in dynamic biological systems. *Proceedings of the National Academy of Sciences*, 108(29):12167–12172.

Hinton, G. E., Osindero, S., and Teh, Y.-W. (2006). A fast learning algorithm for deep belief nets. *Neural computation*, 18(7):1527–1554.

Hinton, G. E. and Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507.

Holloway, D. M., Harrison, L. G., Kosman, D., Vanario-Alonso, C. E., and Spirov, A. V. (2006). Analysis of pattern precision shows that drosophila segmentation develops substantial independence from gradients of maternal gene products. *Developmental Dynamics*, 235(11):2949–2960.

Holloway, D. M., Lopes, F. J., da Fontoura Costa, L., Travençolo, B. A., Golyandina, N., Usevich, K., and Spirov, A. V. (2011). Gene expression noise in spatial patterning: hunchback promoter structure affects noise amplitude and distribution in drosophila segmentation. *PLoS Comput Biol*, 7(2):e1001069.

Houchmandzadeh, B., Wieschaus, E., and Leibler, S. (2002). Establishment of developmental precision and proportions in the early drosophila embryo. *Nature*, 415(6873):798–802.

Hsiao, C. (1981). Autoregressive modelling and money-income causality detection. *Journal of Monetary economics*, 7(1):85–106.

Hyndman, R. and Khandakar, Y. (2007). Automatic time series forecasting: The forecast package for r 7. 2008. *URL: https://www. jstatsoft. org/article/view/v027i03 [accessed 2016-02-24][WebCite Cache].*

Hyndman, R. J. and Athanasopoulos, G. (2014). *Forecasting: principles and practice*. OTexts.

Jaeger, J., Blagov, M., Kosman, D., Kozlov, K. N., Myasnikova, E., Surkova, S., Vanario-Alonso, C. E., Samsonova, M., Sharp, D. H., Reinitz, J., et al. (2004). Dynamical analysis of regulatory interactions in the gap gene system of drosophila melanogaster. *Genetics*, 167(4):1721–1737.

Jang, J.-S. R., Sun, C.-T., and Mizutani, E. (1997). Neuro-fuzzy and soft computing; a computational approach to learning and machine intelligence.

Janssens, H., Kosman, D., Vanario-Alonso, C. E., Jaeger, J., Samsonova, M., and Reinitz, J. (2005). A high-throughput method for quantifying gene expression data from early drosophila embryos. *Development genes and evolution*, 215(7):374–381.

Ji, C., Merl, D., Kepler, T. B., and West, M. (2009). Spatial mixture modelling for unobserved point processes: Examples in immunofluorescence histology. *Bayesian analysis (Online)*, 4(2):297.

Joanes, D. and Gill, C. (1998). Comparing measures of sample skewness and kurtosis. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 47(1):183–189.

Johnston, D. S., Driever, W., Berleth, T., Richstein, S., and Nüsslein-Volhard, C. (1989). Multiple steps in the localization of bicoid rna to the anterior pole of the drosophila oocyte. *Development*, 107(Supplement):13–19.

Kapl, M. and Müller, W. G. (2010). Prediction of steel prices: A comparison between a conventional regression model and mssa. *Statistics and Its Interface*, 3:369–275.

Karlebach, G. and Shamir, R. (2008). Modelling and analysis of gene regulatory networks. *Nature Reviews Molecular Cell Biology*, 9(10):770–780.

Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika*, 30(1/2):81–93.

King, N. (2004). The unicellular ancestry of animal development. *Developmental cell*, 7(3):313–325.

Kirk, D. L. (2005). A twelve-step program for evolving multicellularity and a division of labor. *BioEssays*, 27(3):299–310.

Kirk, M. M., Stark, K., Miller, S. M., Muller, W., Taillon, B. E., Gruber, H., Schmitt, R., and Kirk, D. L. (1999). rega, a volvox gene that plays a central role in germ-soma differentiation, encodes a novel regulatory protein. *Development*, 126(4):639–647.

Klebanov, L. and Yakovlev, A. (2007). How high is the level of technical noise in microarray data? *Biology Direct*, 2(1):1.

Kosman, D., Reinitz, J., and Sharp, D. H. (1998). Automated assay of gene expression at cellular resolution. In *Proceedings of the 1998 Pacific Symposium on Biocomputing*, pages 6–17.

Koza, J. R. (1999). *Genetic programming III: Darwinian invention and problem solving*, volume 3. Morgan Kaufmann.

Kozlov, K., Myasnikova, E., Pisareva, A., Samsonova, M., and Reinitz, J. (2002). A method for two-dimensional registration and construction of the two-dimensional atlas of gene expression patterns in situ. *In silico biology*, 2(2):125–141.

Kuo, H.-H. (1996). *White noise distribution theory*, volume 5. CRC press.

Lahiri, K., Vaughan, D. R., and Wixon, B. (1995). Modeling ssa's sequential disability determination process using matched sipp data. *Soc. Sec. Bull.*, 58:3.

Latchman, D. et al. (2007). *Gene regulation*. Taylor & Francis.

Lee, J.-M., Cho, S., and Baek, J. (2003). Trend detection using auto-associative neural networks: Intraday kospi 200 futures. In *CIFEr*, pages 417–420.

Leite, A., Rocha, A., Silva, M., Gouveia, S., Carvalho, J., and Costa, O. (2007). Long-range dependence in heart rate variability data: Arfima modelling vs detrended fluctuation analysis. In *2007 Computers in Cardiology*, pages 21–24. IEEE.

Leite, A. S., Rocha, A. P., Silva, M. E., and Costa, O. (2006). Modelling long-term heart rate variability: an arfima approach. *Biomedizinische Technik*, 51(4):215–219.

Lemmens, A., Croux, C., and Dekimpe, M. G. (2008). Measuring and testing granger causality over the spectrum: An application to european production expectation surveys. *International Journal of Forecasting*, 24(3):414–431.

Levine, M. and Davidson, E. H. (2005). Gene regulatory networks for development. *Proceedings of the National Academy of Sciences of the United States of America*, 102(14):4936–4942.

Lewis, E. B. (1978). A gene complex controlling segmentation in drosophila. In *Genes, Development and Cancer*, pages 205–217. Springer.

Lewis, J. (2008). From signals to patterns: space, time, and mathematics in developmental biology. *Science*, 322(5900):399–403.

Liew, A. W.-C. and Yan, H. (2009). Reliable detection of short periodic gene expression time series profiles in dna microarray data. In *Systems, Man and Cybernetics, 2009. SMC 2009. IEEE International Conference on*, pages 4274–4279. IEEE.

Little, S. C., Tkačik, G., Kneeland, T. B., Wieschaus, E. F., and Gregor, T. (2011). The formation of the bicoid morphogen gradient requires protein movement from anteriorly localized mrna. *PLoS Biol*, 9(3):e1000596.

Liu, J., He, F., and Ma, J. (2011). Morphogen gradient formation and action. *Fly*.

Liu, S. and Jack, J. (1992). Regulatory interactions and role in cell type specification of the malpighian tubules by the cut, krüppel, and caudal genes of drosophila. *Developmental biology*, 150(1):133–143.

Liu, W. (2013). *Machine learning approaches to modelling bicoid morphogen in Drosophila melanogaster*. PhD thesis, University of Southampton.

Longo, D. and Hasty, J. (2006). Dynamics of single-cell gene expression. *Molecular systems biology*, 2(1):64.

Lopes, F. J., Spirov, A. V., and Bisch, P. M. (2012). The role of bicoid cooperative binding in the patterning of sharp borders in drosophila melanogaster. *Developmental biology*, 370(2):165–172.

Lopes, F. J., Vieira, F. M., Holloway, D. M., Bisch, P. M., and Spirov, A. V. (2008). Spatial bistability generates hunchback expression sharpness in the drosophila embryo. *PLoS Comput Biol*, 4(9):e1000184.

Makridakis, S., Wheelwright, S. C., and Hyndman, R. J. (2008). *Forecasting methods and applications*. John Wiley & Sons.

Martens, M. and Zein, J. (2004). Predicting financial volatility: High-frequency time-series forecasts vis-à-vis implied volatility. *Journal of Futures Markets*, 24(11):1005–1028.

McCulloch, W. S. and Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4):115–133.

Meissner, M., Stark, K., Cresnar, B., Kirk, D. L., and Schmitt, R. (1999). Volvox germline-specific genes that are putative targets of rega repression encode chloroplast proteins. *Current genetics*, 36(6):363–370.

Michod, R. E. (2007). Evolution of individuality during the transition from unicellular to multicellular life. *Proceedings of the National Academy of Sciences*, 104(suppl 1):8613–8618.

Michod, R. E. and Nedelcu, A. M. (2003). On the reorganization of fitness during evolutionary transitions in individuality. *Integrative and Comparative Biology*, 43(1):64–73.

Minsky, M. (1988). Memoir on inventing the confocal scanning microscope. *Scanning*, 10(4):128–138.

Morgan, T. (1901). Regeneration macmillan. *New York [PubMed]*.

Myasnikova, E., Samsonova, A., Kozlov, K., Samsonova, M., and Reinitz, J. (2001). Registration of the expression patterns of drosophila segmentation genes by two independent methods. *Bioinformatics*, 17(1):3–12.

Myasnikova, E., Samsonova, M., Kosman, D., and Reinitz, J. (2005). Removal of background signal from in situ data on the expression of segmentation genes in drosophila. *Development genes and evolution*, 215(6):320–326.

Myasnikova, E., Surkova, S., Panok, L., Samsonova, M., and Reinitz, J. (2009). Estimation of errors introduced by confocal imaging into the data on segmentation gene expression in drosophila. *Bioinformatics*, 25(3):346–352.

Nielsen, M. A. and Chuang, I. L. (2010). *Quantum computation and quantum information*. Cambridge university press.

Niessing, D., Blanke, S., and Jäckle, H. (2002). Bicoid associates with the 5-cap-bound complex of caudal mrna and represses translation. *Genes & development*, 16(19):2576–2582.

North, A. J. (2006). Seeing is believing? a beginners' guide to practical pitfalls in image acquisition. *The Journal of cell biology*, 172(1):9–18.

Okabe-Oho, Y., Murakami, H., Oho, S., and Sasai, M. (2009). Stable, precise, and reproducible patterning of bicoid and hunchback molecules in the early drosophila embryo. *PLoS Comput Biol*, 5(8):e1000486.

Oropeza, V. and Sacchi, M. (2011). Simultaneous seismic data denoising and reconstruction via multichannel singular spectrum analysis. *Geophysics*, 76(3):V25–V32.

Pandey, U. B. and Nichols, C. D. (2011). Human disease models in drosophila melanogaster and the role of the fly in therapeutic drug discovery. *Pharmacological reviews*, 63(2):411–436.

Papatsenko, D. and Levine, M. (2011). The drosophila gap gene network is composed of two parallel toggle switches. *PLoS One*, 6(7):e21145.

Patterson, K., Hassani, H., Heravi, S., and Zhigljavsky, A. (2011). Multivariate singular spectrum analysis for forecasting revisions to real-time data. *Journal of Applied Statistics*, 38(10):2183–2211.

Pawley, J. B. (2006). Fundamental limits in confocal microscopy. In *Handbook of biological confocal microscopy*, pages 20–42. Springer.

Pisarev, A., Poustelnikova, E., Samsonova, M., and Reinitz, J. (2009). Flyex, the quantitative atlas on segmentation gene expression at cellular resolution. *Nucleic acids research*, 37(suppl 1):D560–D566.

Porcher, A. and Dostatni, N. (2010). The bicoid morphogen system. *Current biology*, 20(5):R249–R254.

Poustelnikova, E., Pisarev, A., Blagov, M., Samsonova, M., and Reinitz, J. (2004). A database for management of gene expression data in situ. *Bioinformatics*, 20(14):2212–2221.

Powell, J. R. (1997). *Progress and prospects in evolutionary biology: the Drosophila model*. Oxford University Press.

Rau, A., Jaffrézic, F., Foulley, J.-L., and Doerge, R. W. (2010). An empirical bayesian method for estimating biological networks from temporal microarray data. *Statistical Applications in Genetics and Molecular Biology*, 9(1).

Reinitz, J. and Sharp, D. H. (1995). Mechanism of eve stripe formation. *Mechanisms of development*, 49(1):133–158.

Rice, S. A. (1985). *Diffusion-limited reactions*, volume 25. Elsevier.

Rivera-Pomar, R. and Jãckle, H. (1996). From gradients to stripes in drosophila embryogenesis: filling in the gaps. *Trends in Genetics*, 12(11):478–483.

Rodrıguez-Aragón, L. J. and Zhigljavsky, A. (2010). Singular spectrum analysis for image processing. *Statistics and Its Interface*, 3(3):419–426.

Roelink, H., Porter, J., Chiang, C., Tanabe, Y., Chang, D., Beachy, P., and Jessell, T. (1995). Floor plate and motor neuron induction by different concentrations of the amino-terminal cleavage product of sonic hedgehog autoproteolysis. *Cell*, 81(3):445–455.

Sanei, S., Ghodsi, M., and Hassani, H. (2011). An adaptive singular spectrum analysis approach to murmur detection from heart sounds. *Medical Engineering & Physics*, 33(3):362–367.

Sanei, S. and Hassani, H. (2015). *Singular spectrum analysis of biomedical signals*. CRC Press.

Schlitt, T. and Brazma, A. (2007). Current approaches to gene regulatory network modelling. *BMC bioinformatics*, 8(6):1.

Schwarz, G. et al. (1978). Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464.

Scott, M. P., Carroll, S. B., et al. (1987). The segmentation and homeotic gene network in early drosophila development. *Cell*, 51(5):689–698.

Silva, E. S., Ghodsi, M., Hassani, H., and Abbasirad, K. (2016). A quantitative exploration of the statistical and mathematical knowledge of university entrants into a uk management school. *The International Journal of Management Education*, 14(3):440–453.

Silva, E. S. and Hassani, H. (2015). On the use of singular spectrum analysis for forecasting u.s. trade before, during and after the 2008 recession. *International Economics*, 141:34–49.

Sims, C. A. (1972). Money, income, and causality. *The American economic review*, 62(4):540–552.

Sims, C. A., Stock, J. H., and Watson, M. W. (1990). Inference in linear time series models with some unit roots. *Econometrica: Journal of the Econometric Society*, pages 113–144.

Soofi, A. S. and Cao, L. (2012). *Modelling and forecasting financial data: techniques of nonlinear dynamics*, volume 2. Springer Science & Business Media.

Spearman, C. (1904). The proof and measurement of association between two things. *The American journal of psychology*, 15(1):72–101.

Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D., and Futcher, B. (1998). Comprehensive identification of cell cycle–regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. *Molecular biology of the cell*, 9(12):3273–3297.

Spemann, H. and Mangold, H. (2003). Induction of embryonic primordia by implantation of organizers from a different species. 1923. *International Journal of Developmental Biology*, 45(1):13–38.

Spirov, A., Fahmy, K., Schneider, M., Frei, E., Noll, M., and Baumgartner, S. (2009). Formation of the bicoid morphogen gradient: an mrna gradient dictates the protein gradient. *Development*, 136(4):605–614.

Spirov, A. V., Golyandina, N. E., Holloway, D. M., Alexandrov, T., Spirova, E. N., and Lopes, F. J. (2012). Measuring gene expression noise in early drosophila embryos: the highly dynamic compartmentalized micro-environment of the blastoderm is one of the main sources of noise. In *European Conference on Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics*, pages 177–188. Springer.

Spirov, A. V. and Holloway, D. M. (2003). Making the body plan: precision in the genetic hierarchy of drosophila embryo segmentation. *In silico biology*, 3(1, 2):89–100.

Sugihara, G., May, R., Ye, H., Hsieh, C.-h., Deyle, E., Fogarty, M., and Munch, S. (2012). Detecting causality in complex ecosystems. *science*, 338(6106):496–500.

Sugihara, G. and Mayf, R. (1990). Nonlinear forecasting as a way of distinguishing chaos from measurement error in time series.

Surdej, P. and Jacobs-Lorena, M. (1998). Developmental regulation of bicoid mrna stability is mediated by the first 43 nucleotides of the 3 untranslated region. *Molecular and cellular biology*, 18(5):2892–2900.

Surkova, S., Kosman, D., Kozlov, K., Myasnikova, E., Samsonova, A. A., Spirov, A., Vanario-Alonso, C. E., Samsonova, M., Reinitz, J., et al. (2008a). Characterization of the drosophila segment determination morphome. *Developmental biology*, 313(2):844–862.

Surkova, S., Spirov, A. V., Gursky, V. V., Janssens, H., Kim, A.-R., Radulescu, O., Vanario-Alonso, C. E., Sharp, D. H., Samsonova, M., Reinitz, J., et al. (2009). Canalization of gene expression in the drosophila blastoderm by gap gene cross regulation. *PLoS Biol*, 7(3):e1000049.

Surkova, S. Y., Myasnikova, E., Kozlov, K., Samsonova, A., Reinitz, J., and Samsonova, M. (2008b). Methods for acquisition of quantitative data from confocal images of gene expression in situ. *Cell and tissue biology*, 2(2):200–215.

Takens, F. (1981). Detecting strange attractors in turbulence. In *Dynamical systems and turbulence, Warwick 1980*, pages 366–381. Springer.

Tang, T. Y., Liew, A. W.-C., and Yan, H. (2010). Analysis of mouse periodic gene expression data based on singular value decomposition and autoregressive modeling. In *Proceedings of the International MultiConference of Engineers and Computer Scientists*, volume 1. Citeseer.

Tang, T.-Y. and Yan, H. (2010). Identifying periodicity of microarray gene expression profiles by autoregressive modeling and spectral estimation. In *2010 International Conference on Machine Learning and Cybernetics*, volume 6, pages 3062–3066. IEEE.

Tang, V. T. and Yan, H. (2012). Noise reduction in microarray gene expression data based on spectral analysis. *International Journal of Machine Learning and Cybernetics*, 3(1):51–57.

Turing, A. M. (1952). The chemical basis of morphogenesis. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 237(641):37–72.

Van den Berg, C., Willemsen, V., Hendriks, G., Weisbeek, P., and Scheres, B. (1997). Short-range control of cell differentiation in the arabidopsis root meristem. *Nature*, 390(6657):287–289.

van der Loos, C. M. (2008). Multiple immunoenzyme staining: methods and visualizations for the observation with spectral imaging. *Journal of Histochemistry & Cytochemistry*, 56(4):313–328.

Van Huffel, S. (1993). Enhanced resolution based on minimum variance estimation and exponential data modeling. *Signal processing*, 33(3):333–355.

Van Soom, A., Mateusen, B., Leroy, J., and de Kruif, A. (2003). Assessment of mammalian embryo quality: what can we learn from embryo morphology? *Reproductive biomedicine online*, 7(6):664–670.

Verveer, P. J., Gemkow, M. J., and Jovin, T. M. (1999). A comparison of image restoration approaches applied to three-dimensional confocal and wide-field fluorescence microscopy. *Journal of Microscopy*, 193(1):50–61.

Vikalo, H., Hassibi, B., and Hassibi, A. (2008). Modeling and estimation for real-time microarrays. *IEEE Journal of Selected Topics in Signal Processing*, 2(3):286–296.

Vivian, T.-Y. T., Liew, A. W.-C., and Yan, H. (2010). Periodicity analysis of dna microarray gene expression time series profiles in mouse segmentation clock data. *Statistics and Its Interface*, 3(3):413–418.

Wang, J.-H. and Leu, J.-Y. (1996). Stock market trend prediction using arima-based neural networks. In *Neural Networks, 1996., IEEE International Conference on*, volume 4, pages 2160–2165. IEEE.

Waters, J. C. (2009). Accuracy and precision in quantitative fluorescence microscopy. *The Journal of cell biology*, 185(7):1135–1148.

Wilczynski, B. and Furlong, E. E. (2010). Challenges for modeling global gene regulatory networks during development: insights from drosophila. *Developmental biology*, 340(2):161–169.

Wolpert, L. (1969). Positional information and the spatial pattern of cellular differentiation. *Journal of theoretical biology*, 25(1):1–47.

Wolpert, L. and Szathmáry, E. (2002). Multicellularity: evolution and the egg. *Nature*, 420(6917):745–745.

Wu, Y. F., Myasnikova, E., and Reinitz, J. (2007). Master equation simulation analysis of immunostained bicoid morphogen gradient. *BMC systems biology*, 1(1):52.

Xu, Y., Mural, R. J., Einstein, J. R., Shah, M. B., and Uberbacher, E. C. (1996). Grail: a multi-agent neural network system for gene identification. *Proceedings of the IEEE*, 84(10):1544–1552.

Yang, X.-S. (2010). *Nature-inspired metaheuristic algorithms*. Luniver press.

Ye, H., Deyle, E. R., Gilarranz, L. J., and Sugihara, G. (2015). Distinguishing time-delayed causal interactions using convergent cross mapping. *Scientific reports*, 5.

Zamparo, L. and Perkins, T. J. (2009). Statistical lower bounds on protein copy number from fluorescence expression images. *Bioinformatics*, 25(20):2670–2676.

Zhang, J., Hassani, H., Xie, H., and Zhang, X. (2014). Estimating multi-country prosperity index: A two-dimensional singular spectrum analysis approach. *Journal of Systems Science and Complexity*, 27(1):56–74.

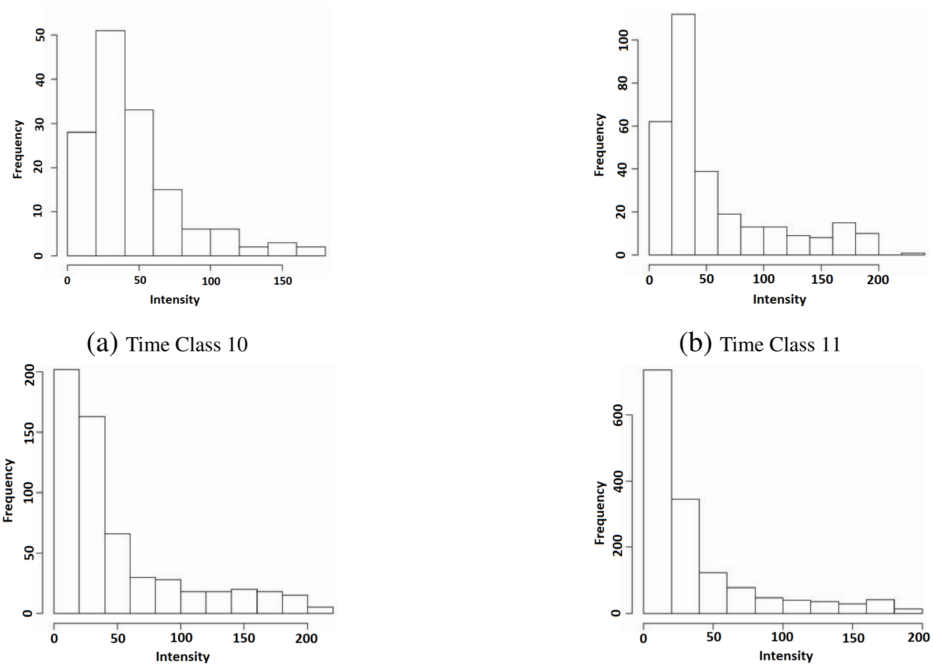
Zou, C. and Feng, J. (2009). Granger causality vs. dynamic bayesian network inference: a comparative study. *BMC bioinformatics*, 10(1):1.

Appendix A

Bicoid data illustration

Histograms

Presented below are the histograms of cleavage cycles 10-13 and all time classes of cleavage cycle 14A. It should be noted that each histogram is related to one particular embryo which is selected as a sample representing that cycle or temporal class.

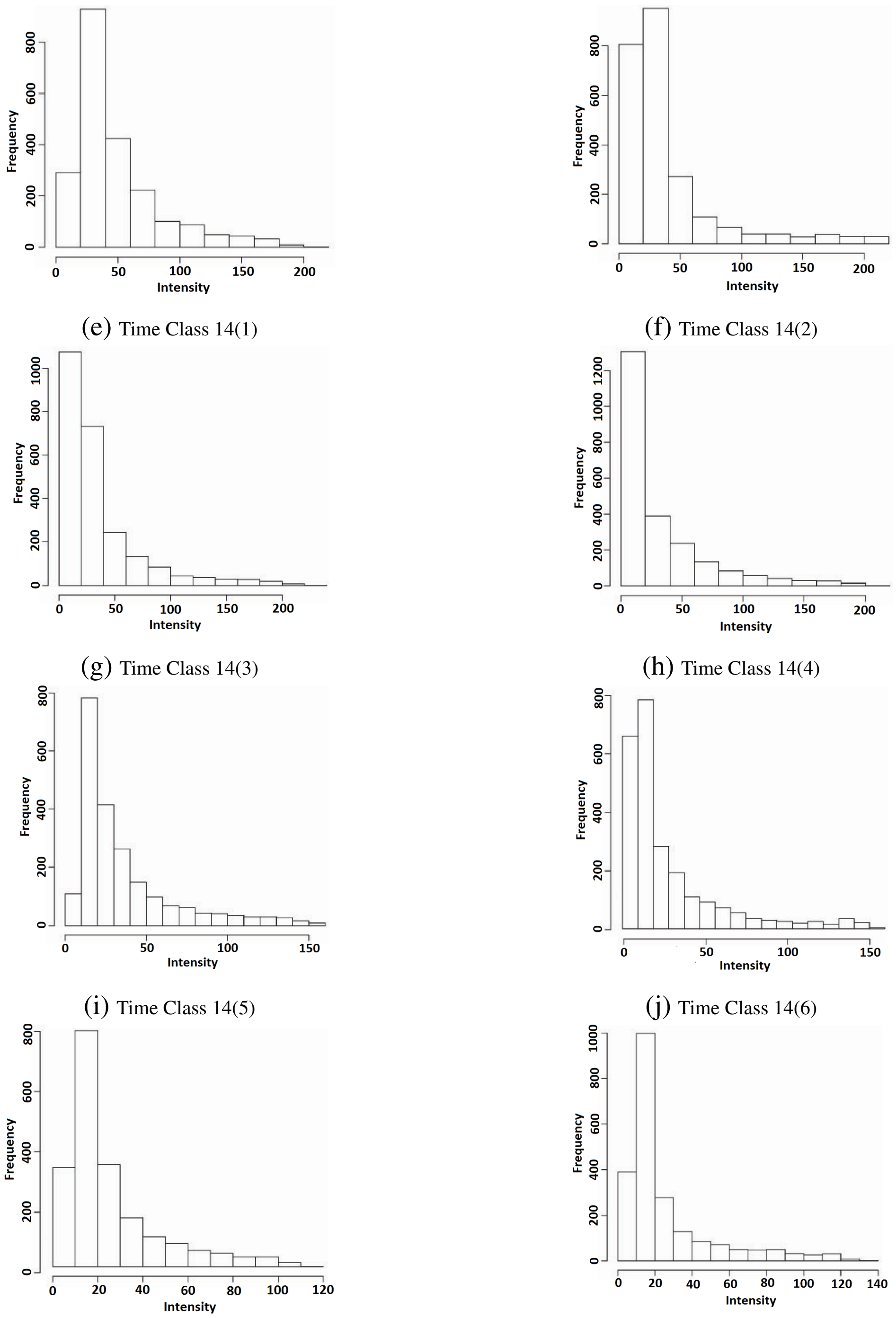


(a) Time Class 10

(b) Time Class 11

(c) Time Class 12

(d) Time Class 13



(e) Time Class 14(1)

(f) Time Class 14(2)

(g) Time Class 14(3)

(h) Time Class 14(4)

(i) Time Class 14(5)

(j) Time Class 14(6)

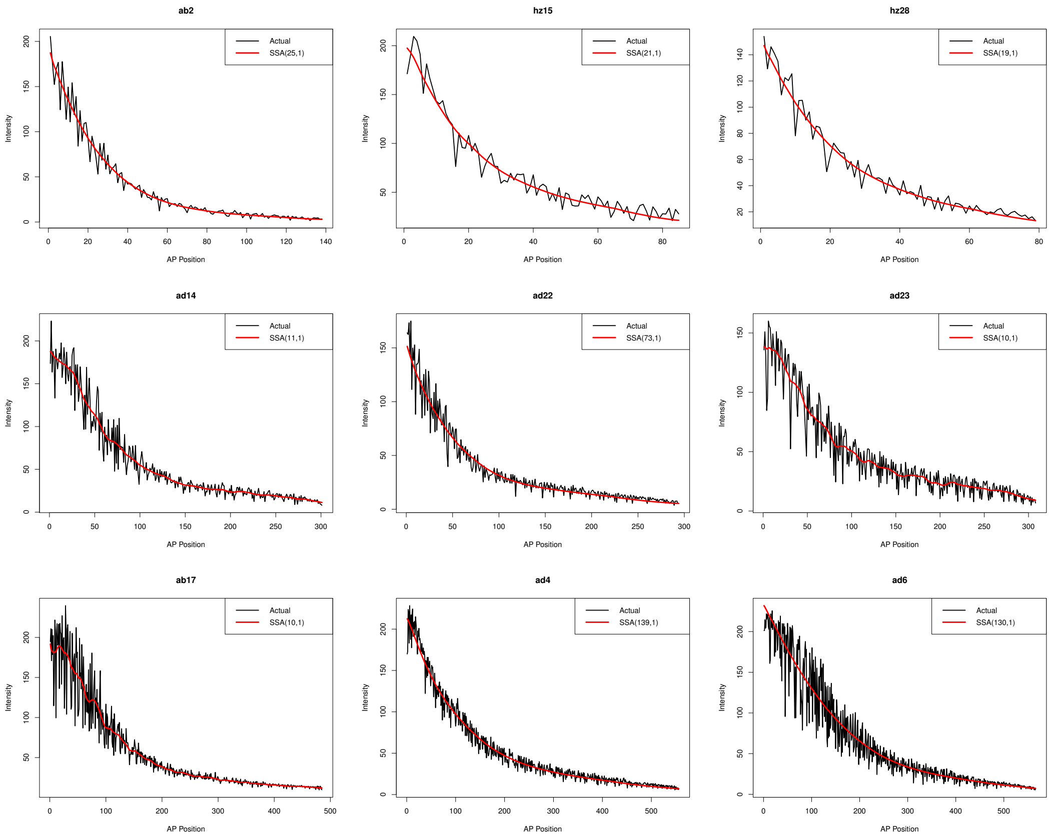
(k) Time Class 14(7)

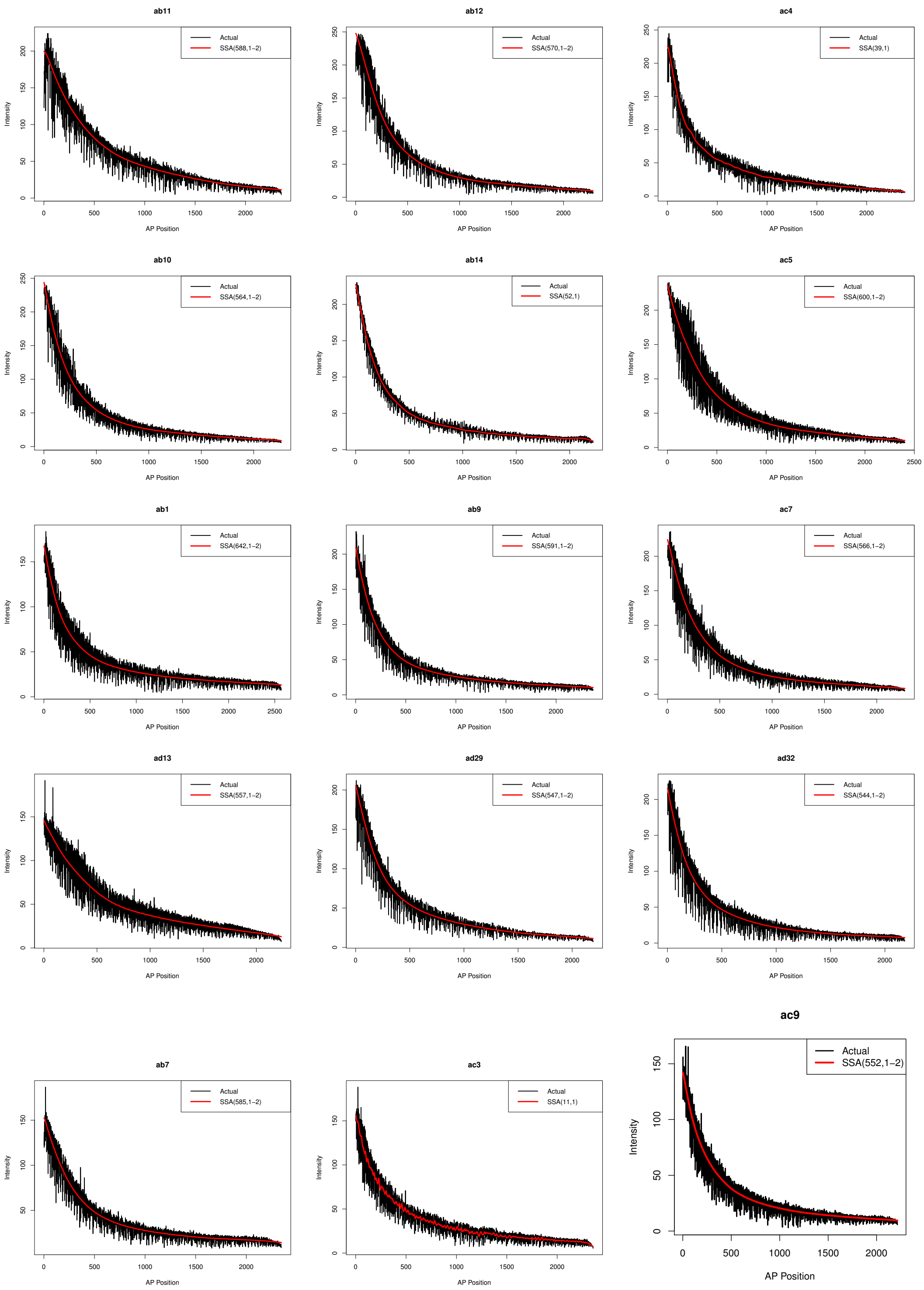
(l) Time Class 14(8)

Fig. A.1 The histogram of Bcd related to cleavage cycles 10-13 and all the time class of cleavage cycle 14A.

Appendix B

Bcd signal extraction




ab11
Actual
SSA(588,1–2)
ab12
Actual
SSA(570,1–2)
ac4
Actual
SSA(39,1)
ab10
Actual
SSA(564,1–2)
ab14
Actual
SSA(52,1)
ac5
Actual
SSA(600,1–2)
ab1
Actual
SSA(642,1–2)
ab9
Actual
SSA(591,1–2)
ac7
Actual
SSA(566,1–2)
ad13
Actual
SSA(557,1–2)
ad29
Actual
SSA(547,1–2)
ad32
Actual
SSA(544,1–2)
ac9
Actual
SSA(552,1–2)
ab7
Actual
SSA(585,1–2)
ac3
Actual
SSA(11,1)
Intensity
AP Position

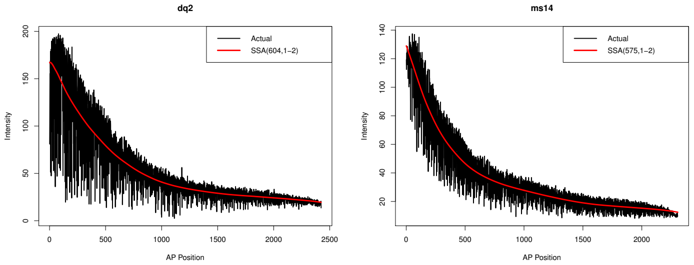
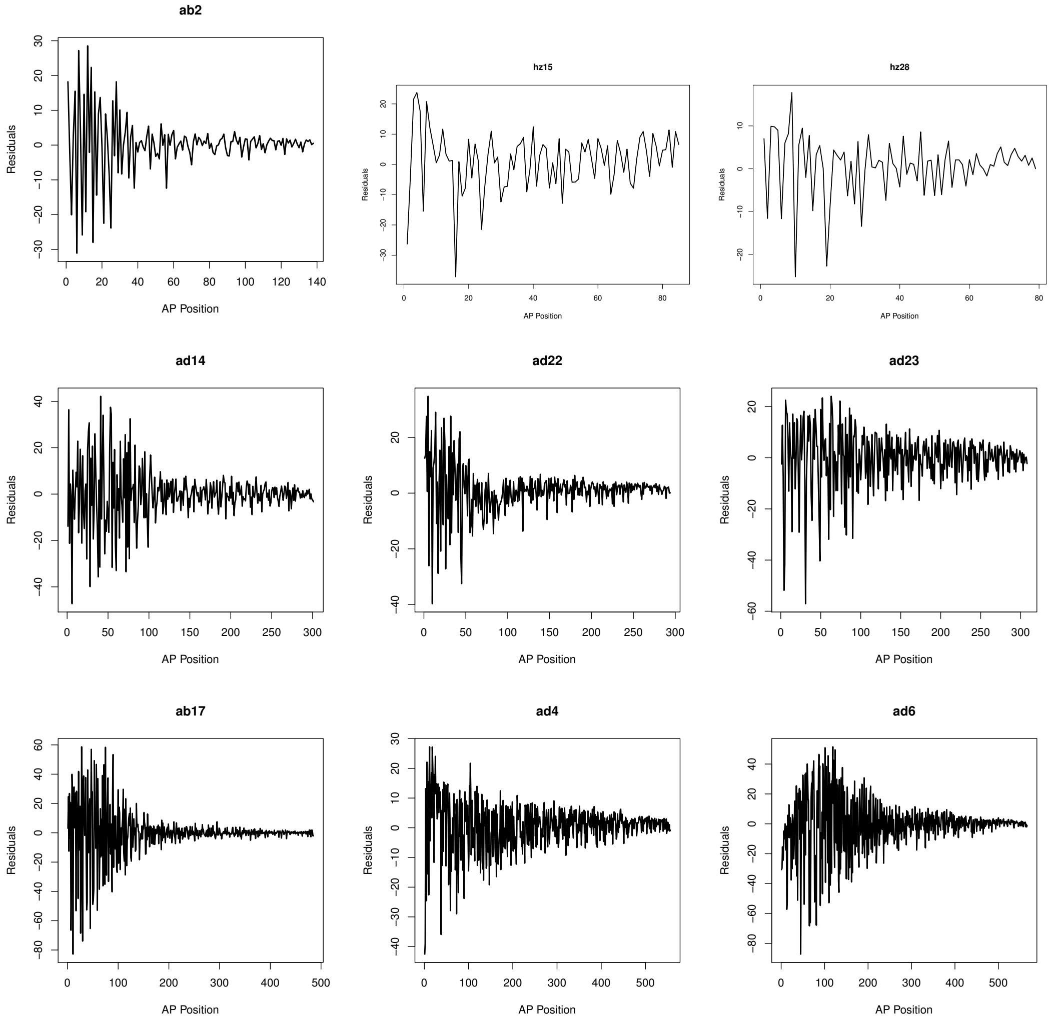
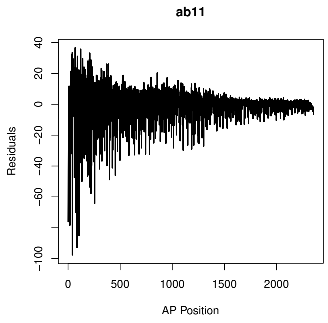
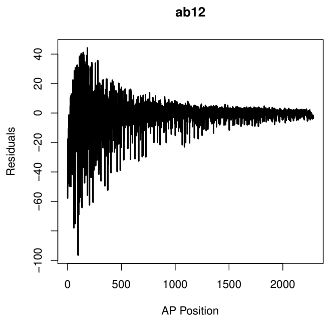
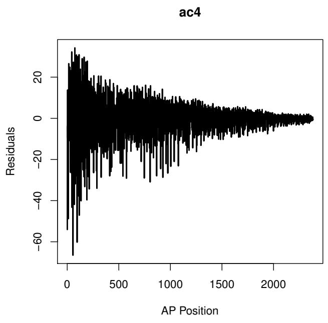


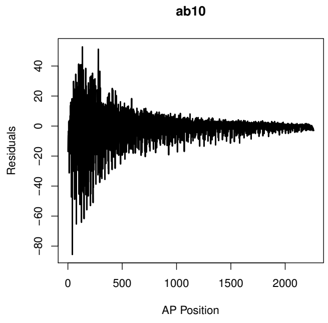
Fig. B.1 SSA based optimal trend extraction for Bcd.

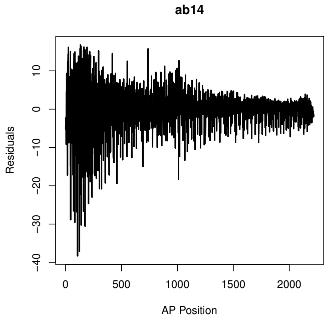


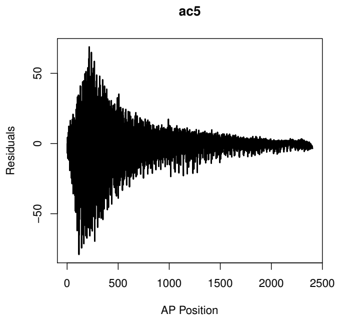

ab11
Residuals
40 20 0 −20 −40 −60 −100
AP Position
0 500 1000 1500 2000

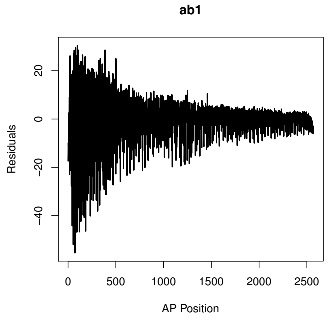

ab12
Residuals
40 20 0 −20 −40 −60 −100
AP Position
0 500 1000 1500 2000

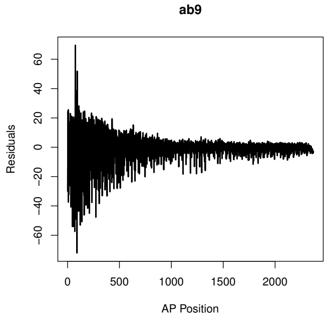

ac4
Residuals
20 0 −20 −40 −60
AP Position
0 500 1000 1500 2000

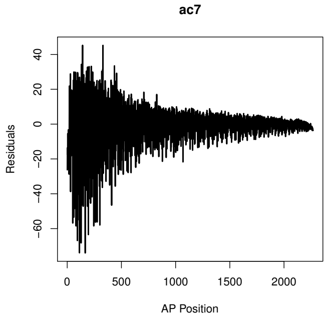

ab10
Residuals
40 20 0 −20 −40 −60 −80
AP Position
0 500 1000 1500 2000

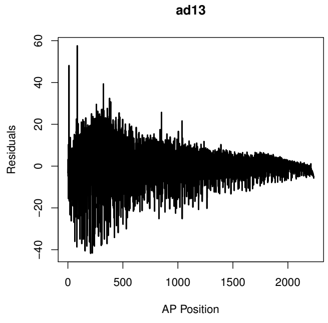

ab14
Residuals
10 0 −10 −20 −30 −40
AP Position
0 500 1000 1500 2000

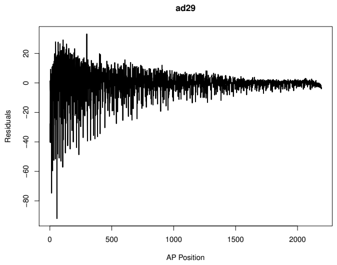

ac5
Residuals
50 0 −50
AP Position
0 500 1000 1500 2000 2500

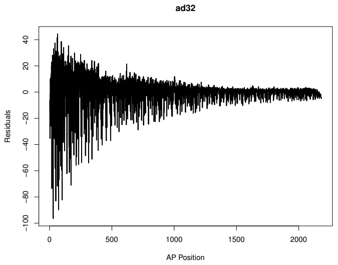

ab1
Residuals
20 0 −20 −40
AP Position
0 500 1000 1500 2000 2500

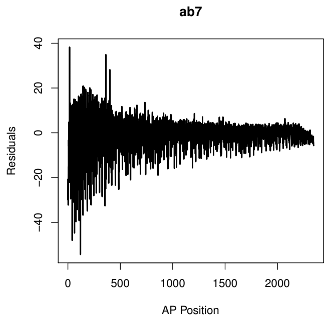

ab9
Residuals
60 40 20 0 −20 −40 −60
AP Position
0 500 1000 1500 2000

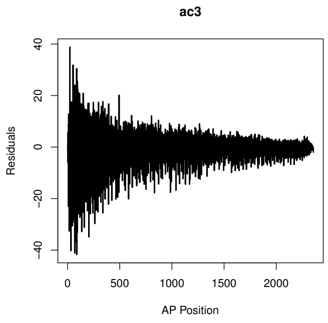

ac7
Residuals
40 20 0 −20 −40 −60
AP Position
0 500 1000 1500 2000

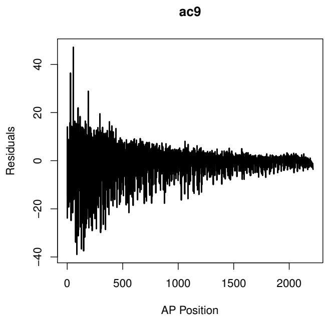

ad13
Residuals
60 40 20 0 −20 −40
AP Position
0 500 1000 1500 2000


ad29
Residuals
20 0 −20 −40 −60 −80
AP Position
0 500 1000 1500 2000


ad32
Residuals
40 20 0 −20 −40 −60 −80 −100
AP Position
0 500 1000 1500 2000


ab7
Residuals
40 20 0 −20 −40
AP Position
0 500 1000 1500 2000


ac3
Residuals
40 20 0 −20 −40
AP Position
0 500 1000 1500 2000


ac9
Residuals
40 20 0 −20 −40
AP Position
0 500 1000 1500 2000

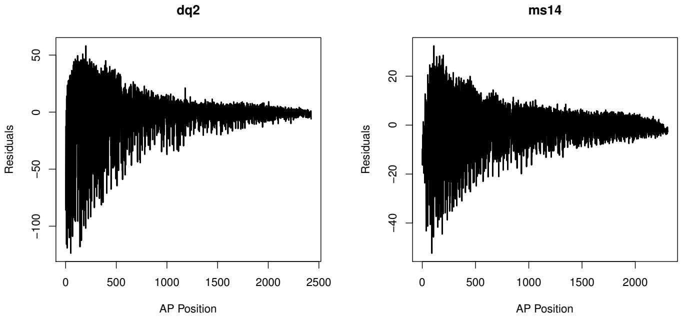
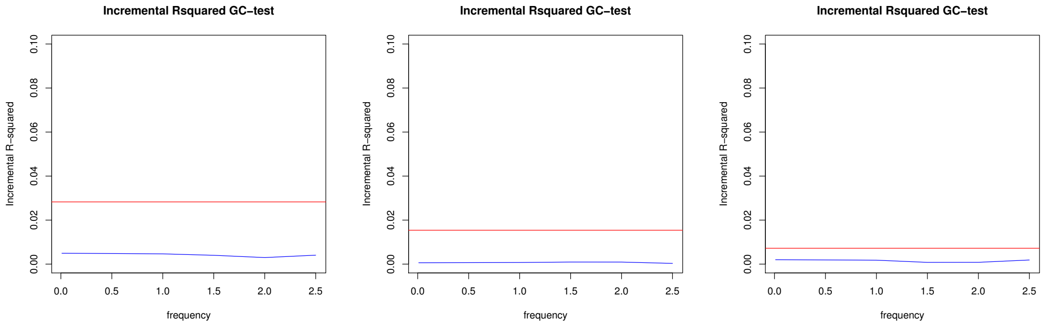


Fig. B.2 Residuals following SSA based optimal trend extraction for Bcd.

Appendix C

Frequency domain GC and CCM Tests Results

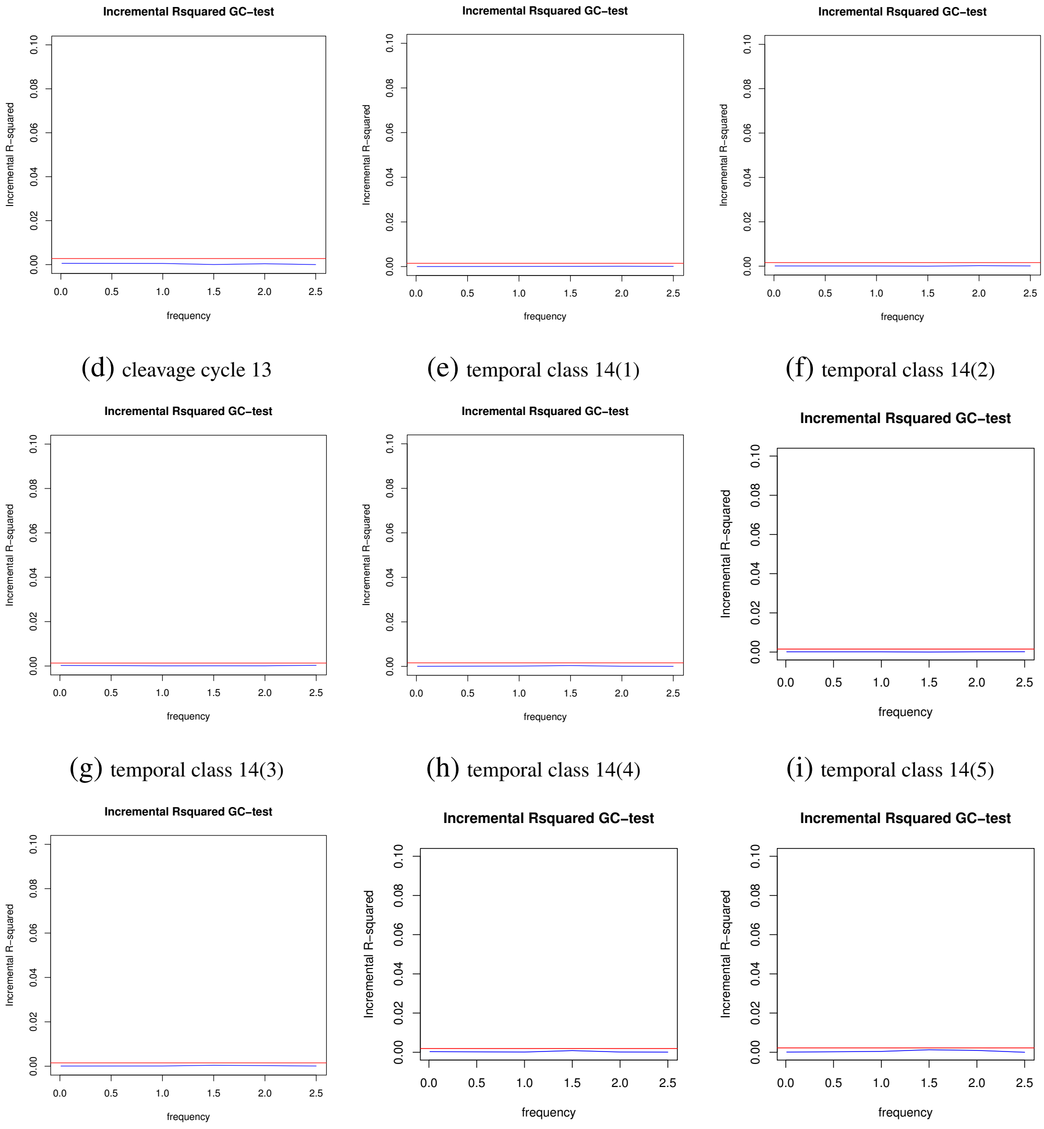
In this appendix, the obtained results following adopting the Frequency domain and CCM Tests are presented. Each plot is related to one particular embryo which is selected as a sample representing that cycle or temporal class. Note that some results of filtered series show a minor difference between the test statistics and the 5% critical value, which is hard to depict in the outcome test plots when considering the same legend for comparison.



(a) cleavage cycle 10

(b) cleavage cycle 11

(c) cleavage cycle 12



(d) cleavage cycle 13

(e) temporal class 14(1)

(f) temporal class 14(2)

(g) temporal class 14(3)

(h) temporal class 14(4)

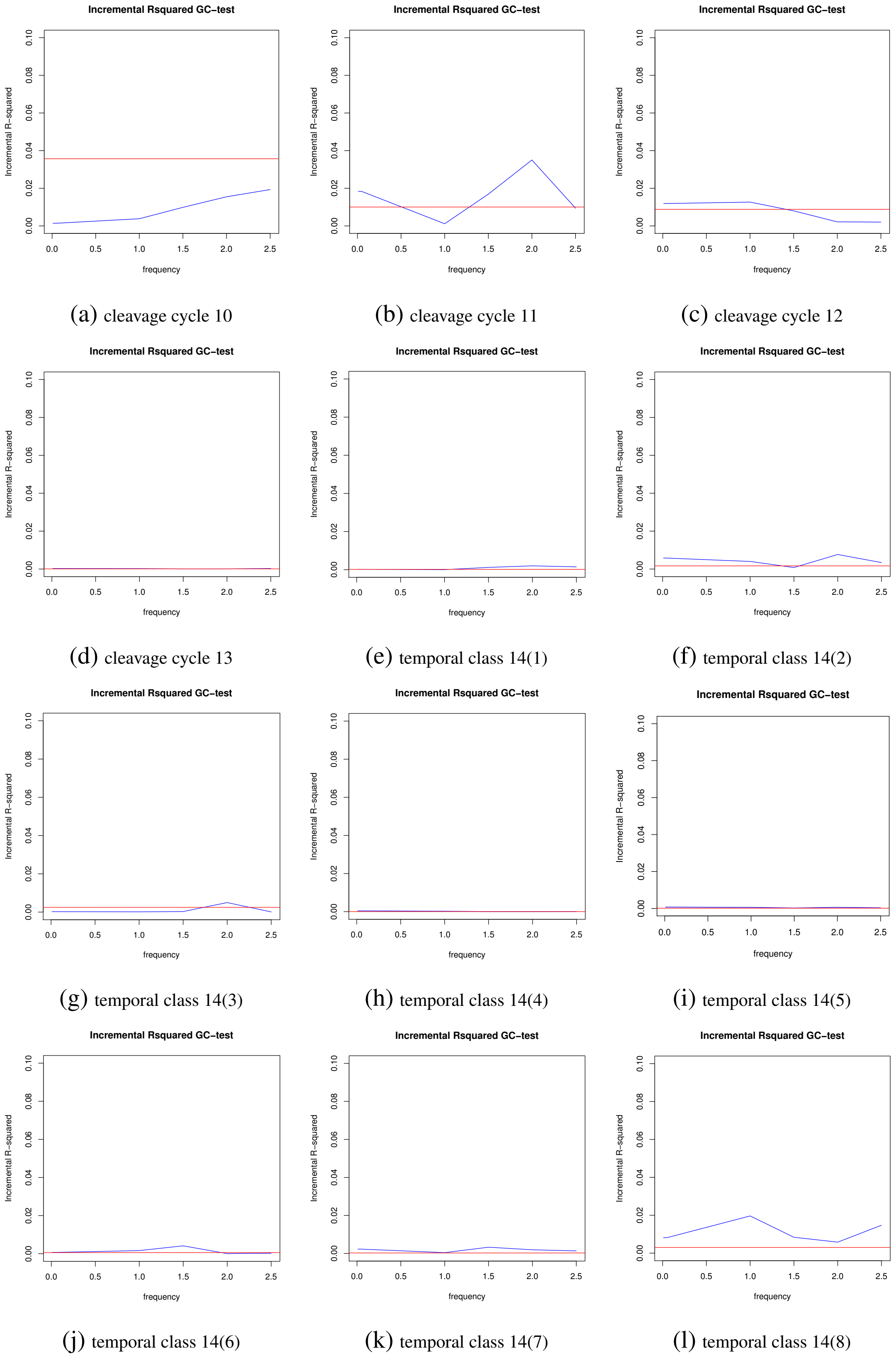
(i) temporal class 14(5)

(j) temporal class 14(6)

(k) temporal class 14(7)

(l) temporal class 14(8)

Fig. C.1 Frequency domain causality test results for Bcd and Cad at cleavage cycles 10-14A (Noisy Series).



(a) cleavage cycle 10

(b) cleavage cycle 11

(c) cleavage cycle 12

(d) cleavage cycle 13

(e) temporal class 14(1)

(f) temporal class 14(2)

(g) temporal class 14(3)

(h) temporal class 14(4)

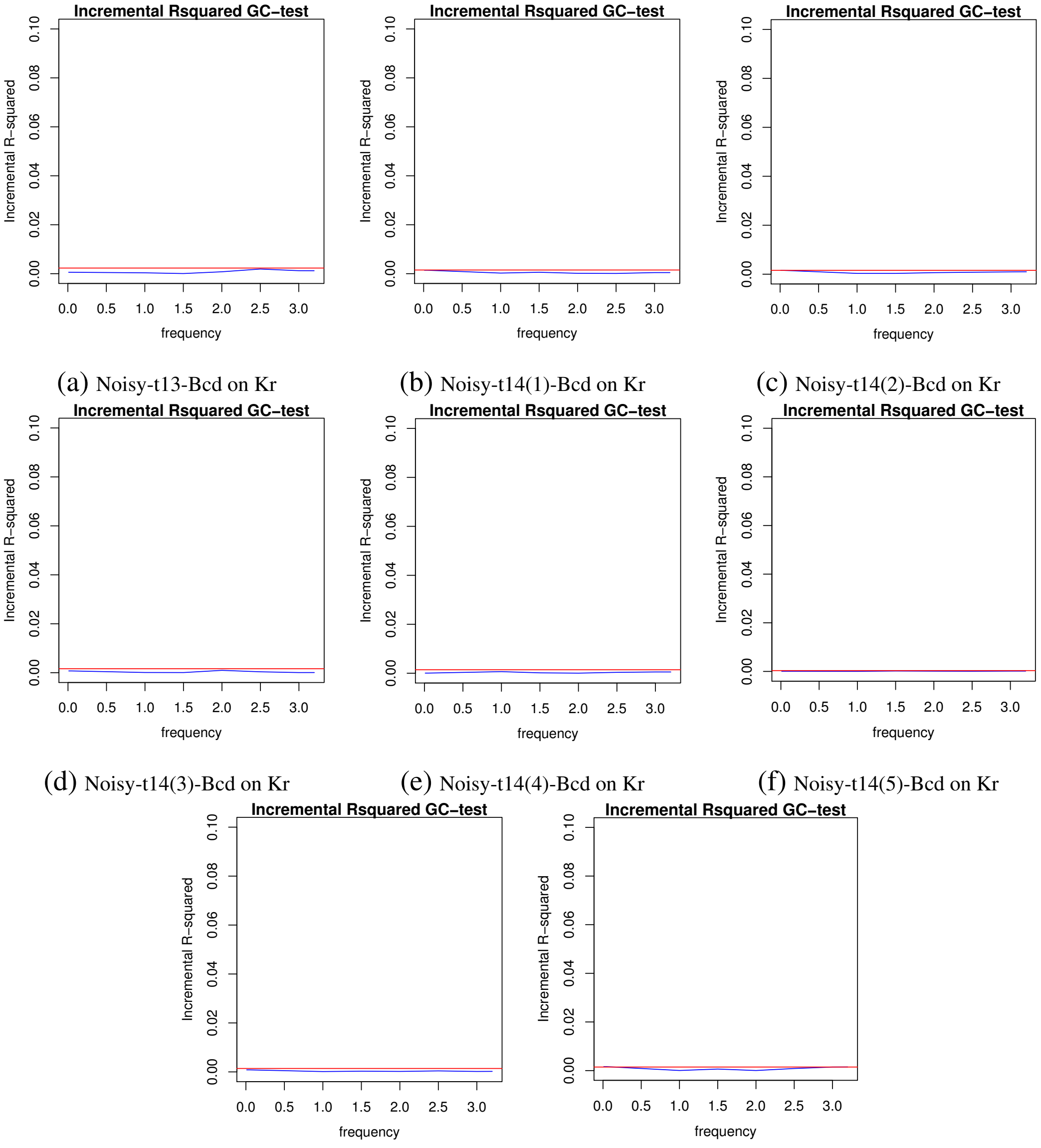
(i) temporal class 14(5)

(j) temporal class 14(6)

(k) temporal class 14(7)

(l) temporal class 14(8)

Fig. C.2 Frequency domain causality test results for Bcd and Cad at at cleavage cycles 10-14A (Filtered Series).



(a) Noisy-t13-Bcd on Kr

(b) Noisy-t14(1)-Bcd on Kr

(c) Noisy-t14(2)-Bcd on Kr

(d) Noisy-t14(3)-Bcd on Kr

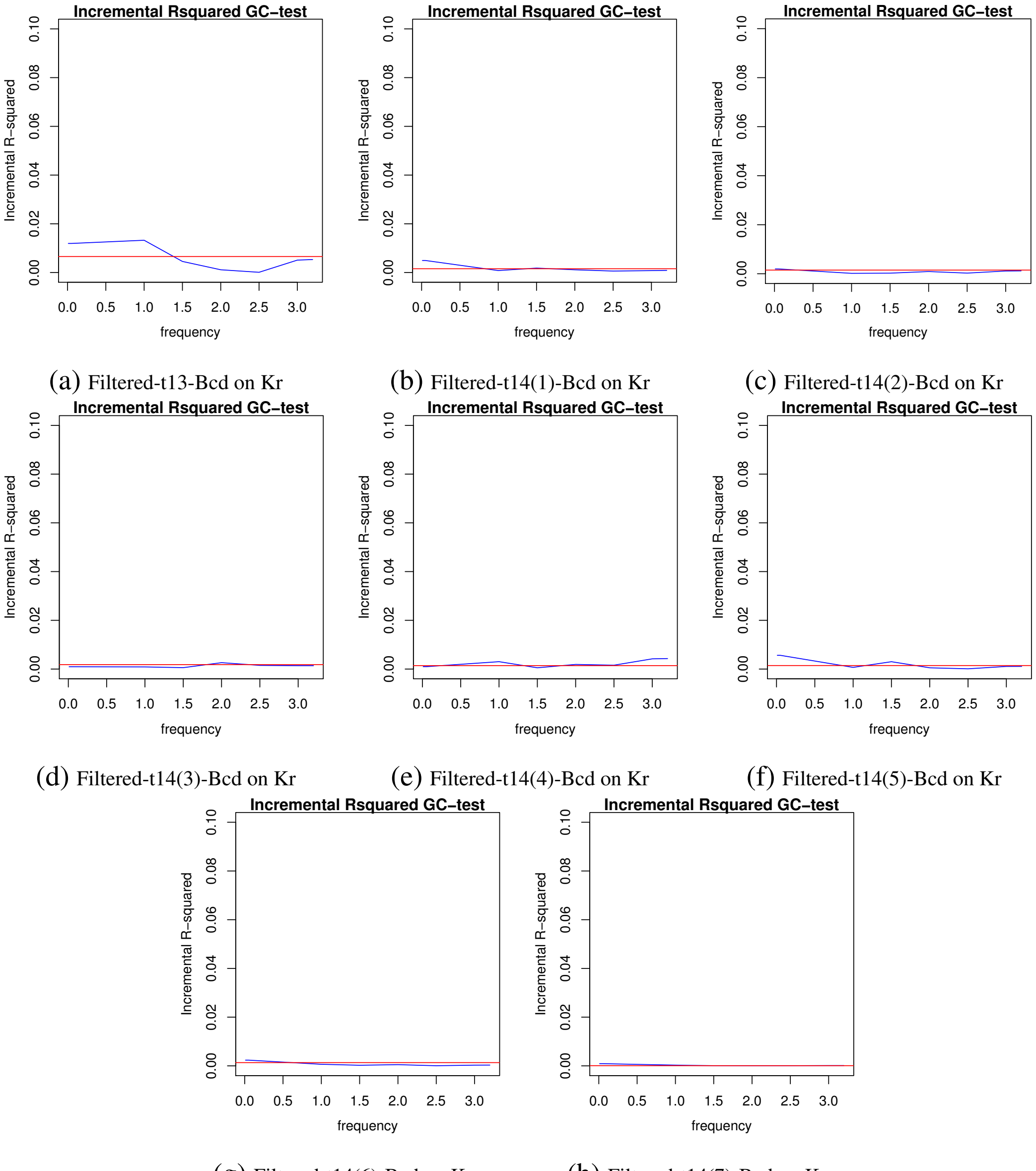
(e) Noisy-t14(4)-Bcd on Kr

(f) Noisy-t14(5)-Bcd on Kr

(g) Noisy-t14(6)-Bcd on Kr

(h) Noisy-t14(7)-Bcd on Kr

Fig. C.3 Frequency domain causality test results for Bcd and Kr at cleavage cycles 10-14A (Noisy Series).



(a) Filtered-t13-Bcd on Kr

(b) Filtered-t14(1)-Bcd on Kr

(c) Filtered-t14(2)-Bcd on Kr

(d) Filtered-t14(3)-Bcd on Kr

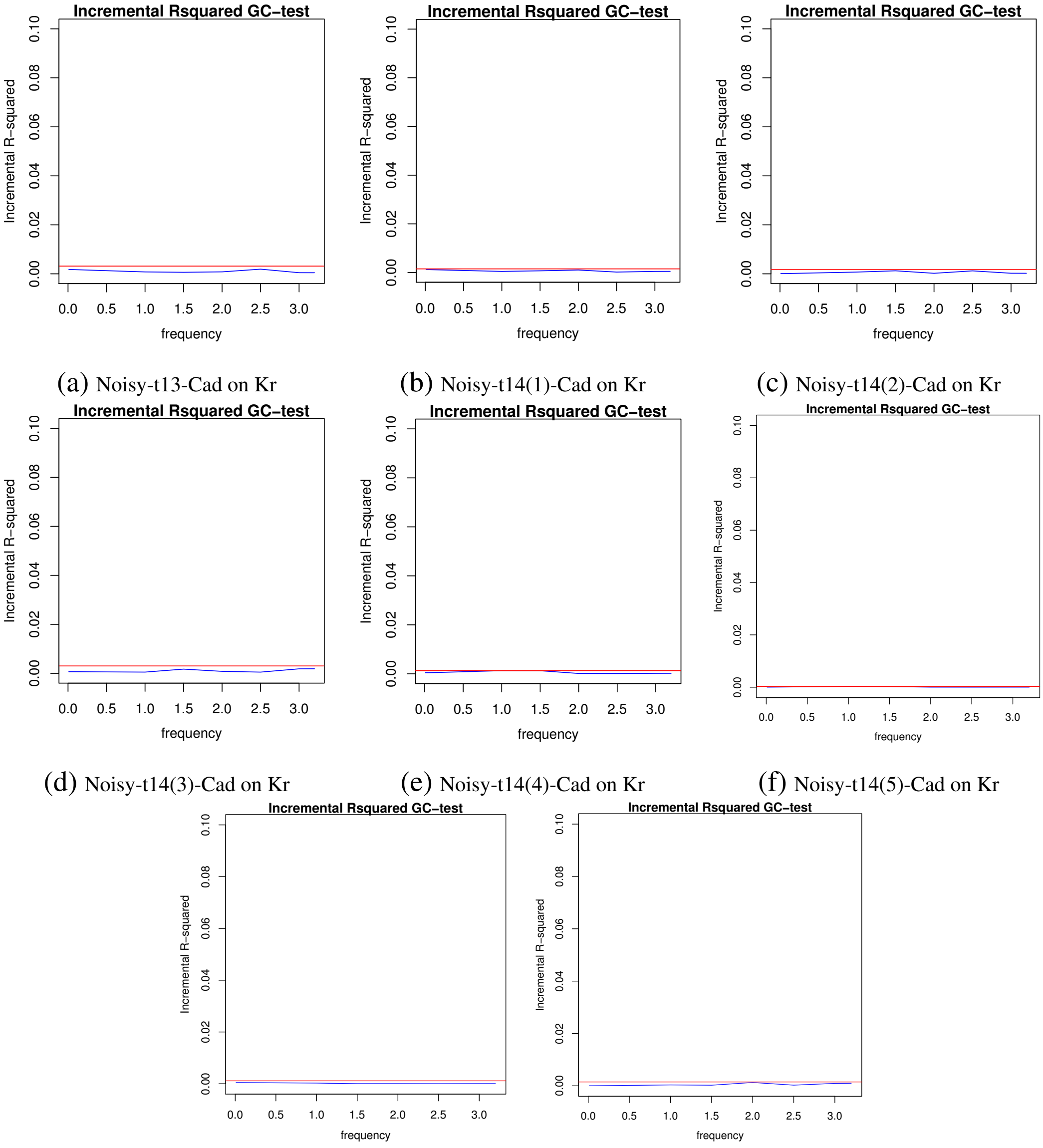
(e) Filtered-t14(4)-Bcd on Kr

(f) Filtered-t14(5)-Bcd on Kr

(g) Filtered-t14(6)-Bcd on Kr

(h) Filtered-t14(7)-Bcd on Kr

Fig. C.4 Frequency domain causality test results for Bcd and Kr at cleavage cycles 10-14A (Filtered Series).



(a) Noisy-t13-Cad on Kr

(b) Noisy-t14(1)-Cad on Kr

(c) Noisy-t14(2)-Cad on Kr

(d) Noisy-t14(3)-Cad on Kr

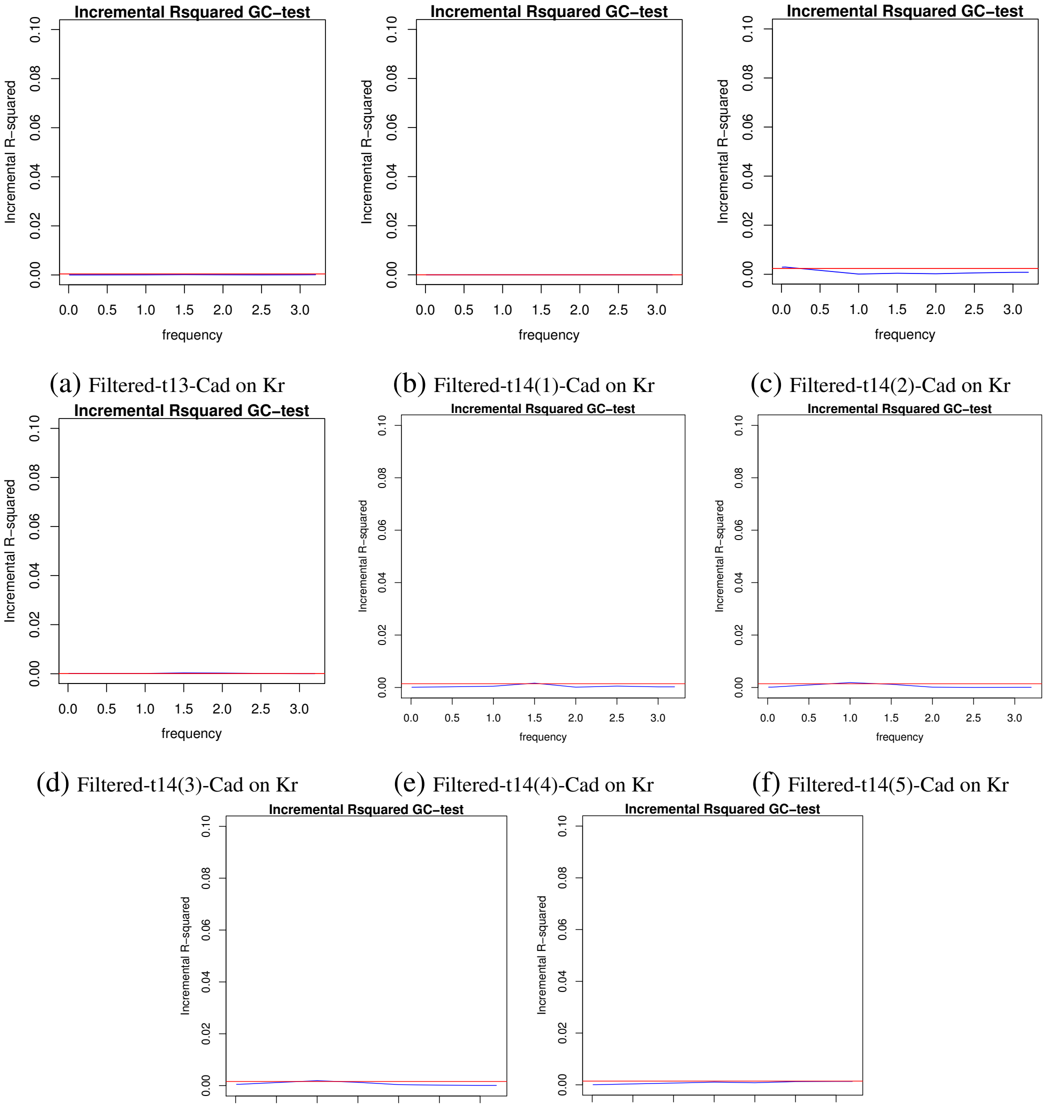
(e) Noisy-t14(4)-Cad on Kr

(f) Noisy-t14(5)-Cad on Kr

(g) Noisy-t14(6)-Cad on Kr

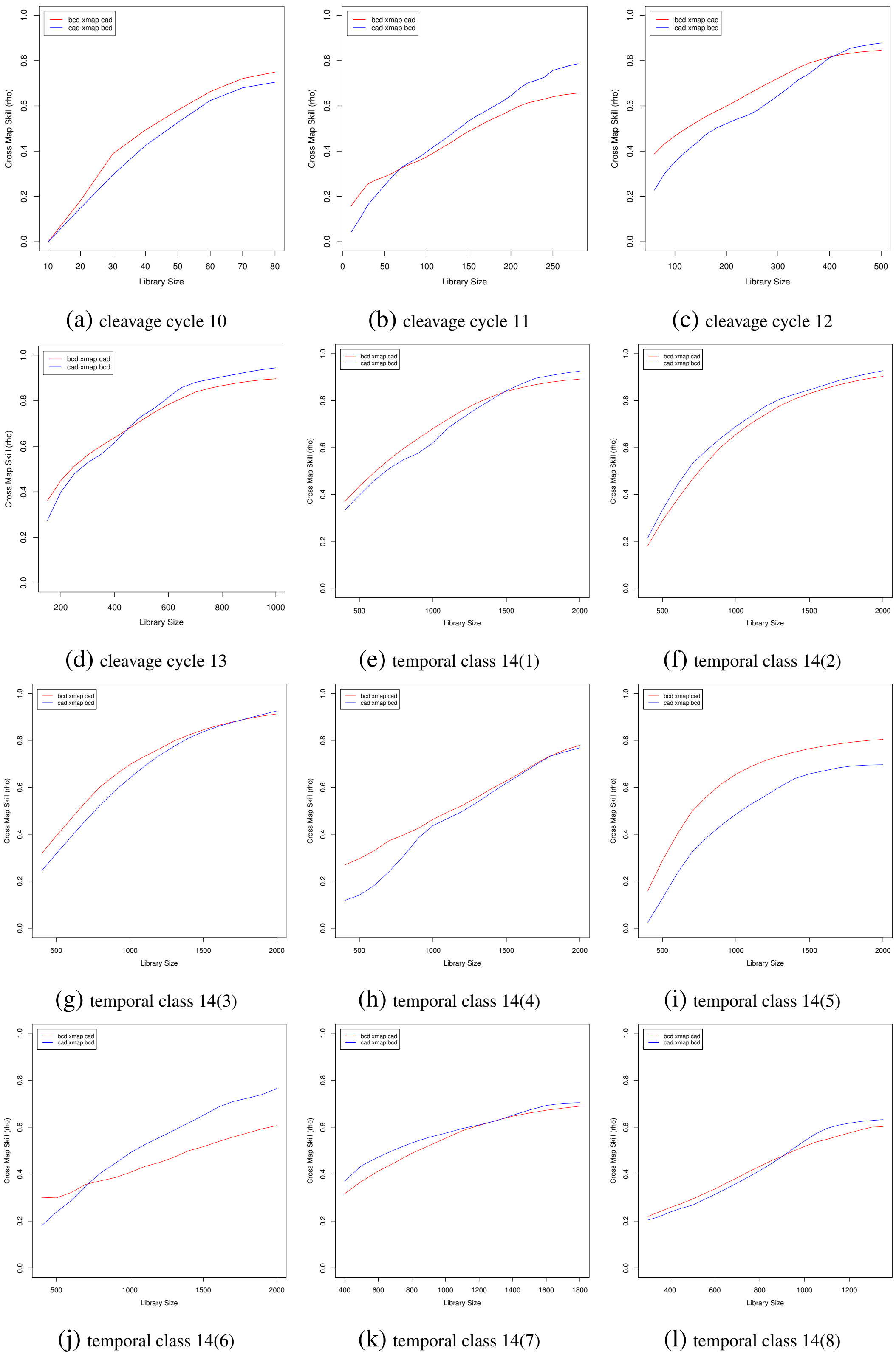
(h) Noisy-t14(7)-Cad on Kr

Fig. C.5 Frequency domain causality test results for Cad and Kr at cleavage cycles 10-14A (Noisy Series).



(a) Filtered-t13-Cad on Kr
(b) Filtered-t14(1)-Cad on Kr
(c) Filtered-t14(2)-Cad on Kr
(d) Filtered-t14(3)-Cad on Kr
(e) Filtered-t14(4)-Cad on Kr
(f) Filtered-t14(5)-Cad on Kr
(g) Filtered-t14(6)-Cad on Kr
(h) Filtered-t14(7)-Cad on Kr

Fig. C.6 Frequency domain causality test results for Cad and Kr at cleavage cycles 10-14A (Filtered Series).



(a) cleavage cycle 10

(b) cleavage cycle 11

(c) cleavage cycle 12

(d) cleavage cycle 13

(e) temporal class 14(1)

(f) temporal class 14(2)

(g) temporal class 14(3)

(h) temporal class 14(4)

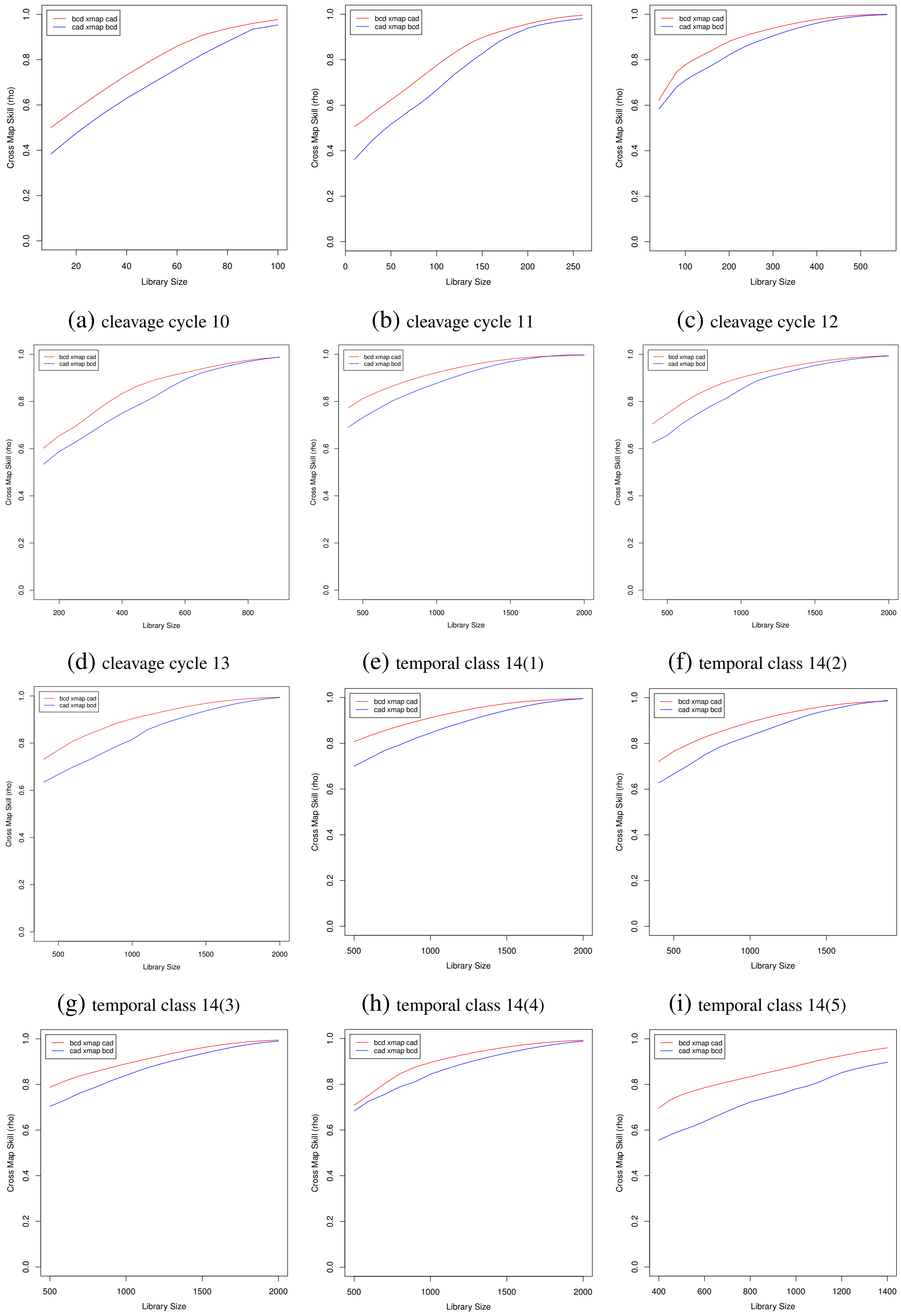
(i) temporal class 14(5)

(j) temporal class 14(6)

(k) temporal class 14(7)

(l) temporal class 14(8)

Fig. C.7 CCM test results for Bcd and Cad at cleavage cycle at cleavage cycles 10-14A (Noisy Series).



(a) cleavage cycle 10

(b) cleavage cycle 11

(c) cleavage cycle 12

(d) cleavage cycle 13

(e) temporal class 14(1)

(f) temporal class 14(2)

(g) temporal class 14(3)

(h) temporal class 14(4)

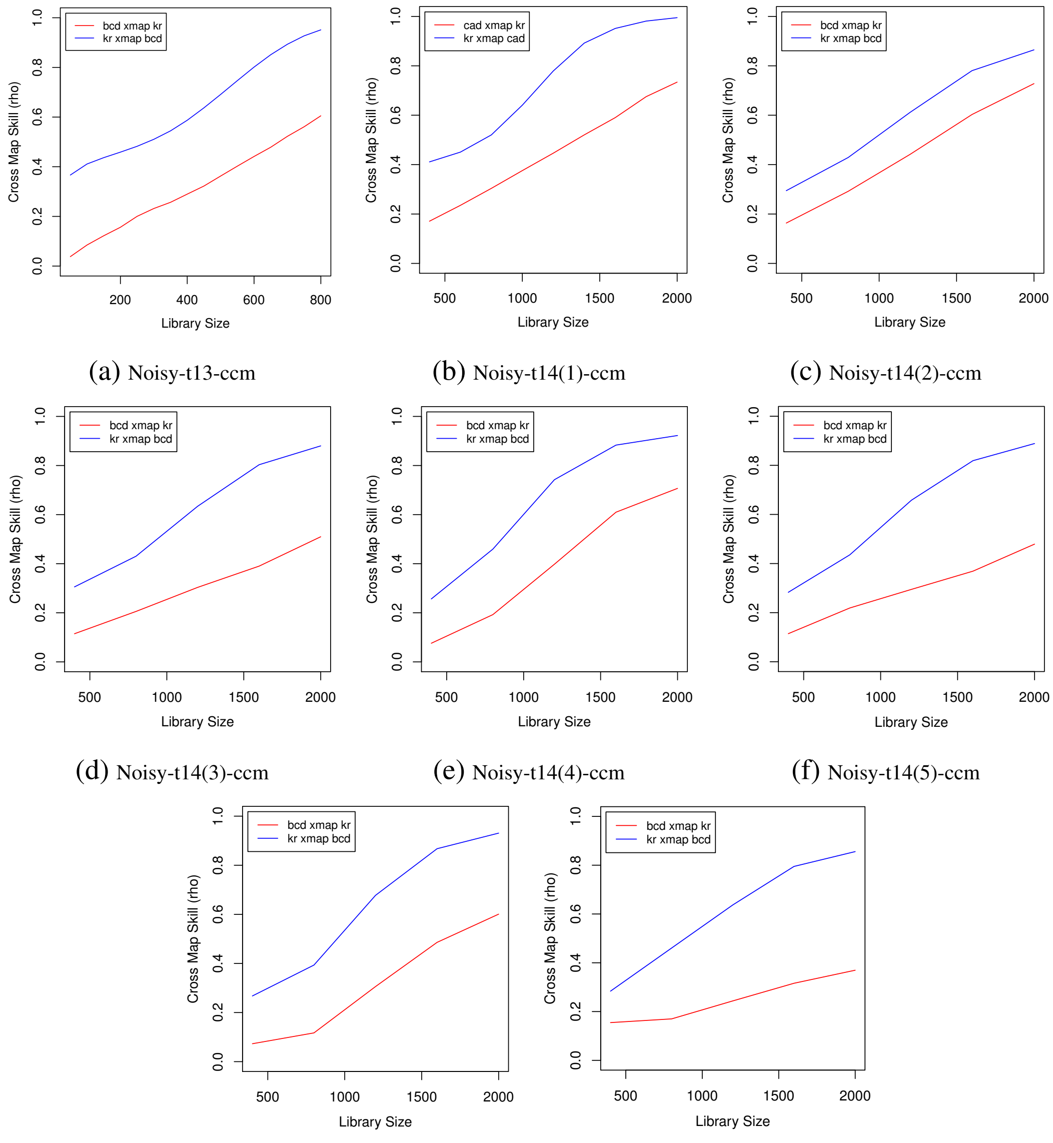
(i) temporal class 14(5)

(j) temporal class 14(6)

(k) temporal class 14(7)

(l) temporal class 14(8)

Fig. C.8 CCM test results for Bcd and Cad at cleavage cycles 10-14A (Filtered Series).



(a) Noisy-t13-ccm

(b) Noisy-t14(1)-ccm

(c) Noisy-t14(2)-ccm

(d) Noisy-t14(3)-ccm

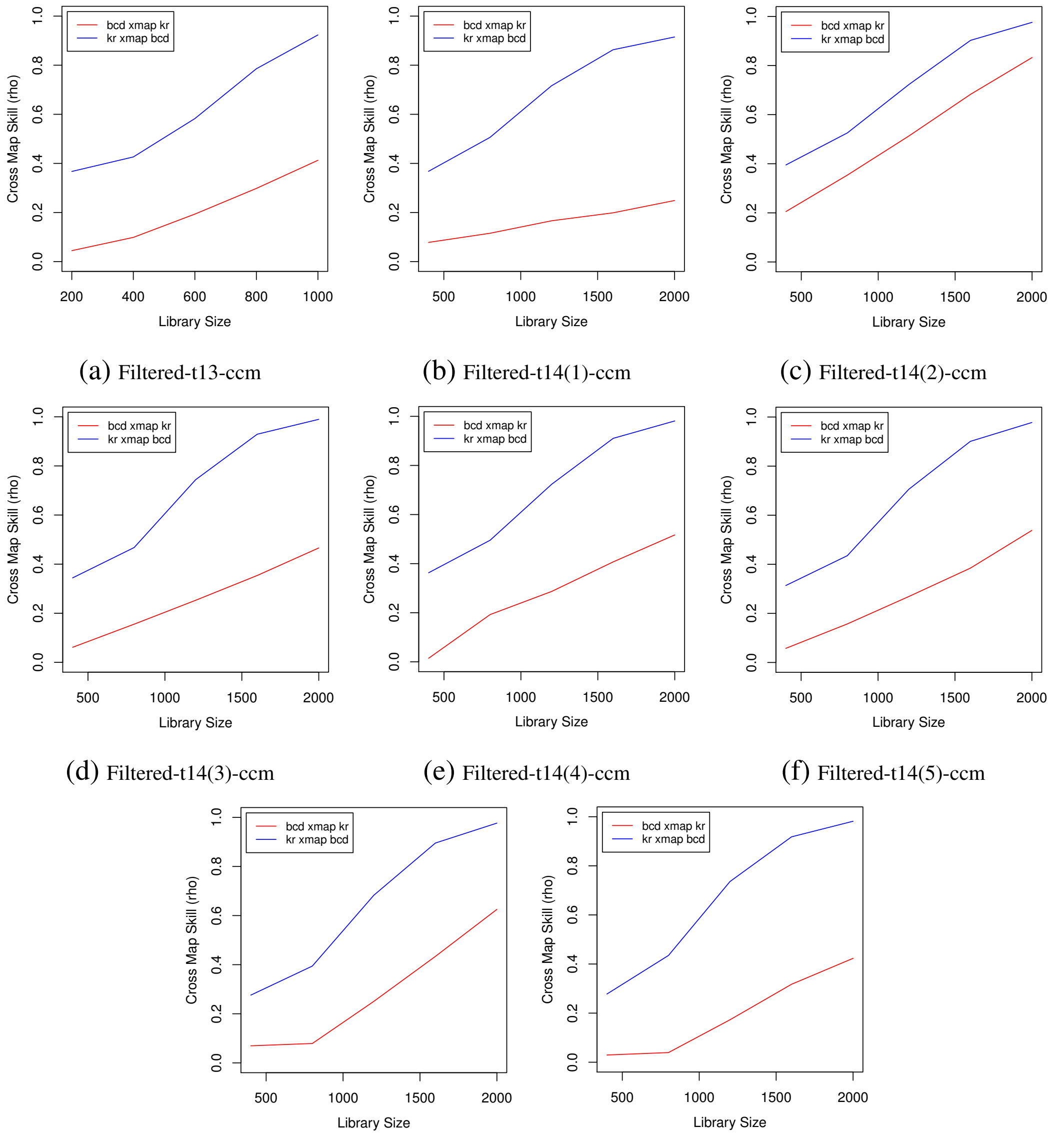
(e) Noisy-t14(4)-ccm

(f) Noisy-t14(5)-ccm

(g) Noisy-t14(6)-ccm

(h) Noisy-t14(7)-ccm

Fig. C.9 CCM test results for Bcd and Kr at cleavage cycles 10-14A (Noisy Series).



(a) Filtered-t13-ccm

(b) Filtered-t14(1)-ccm

(c) Filtered-t14(2)-ccm

(d) Filtered-t14(3)-ccm

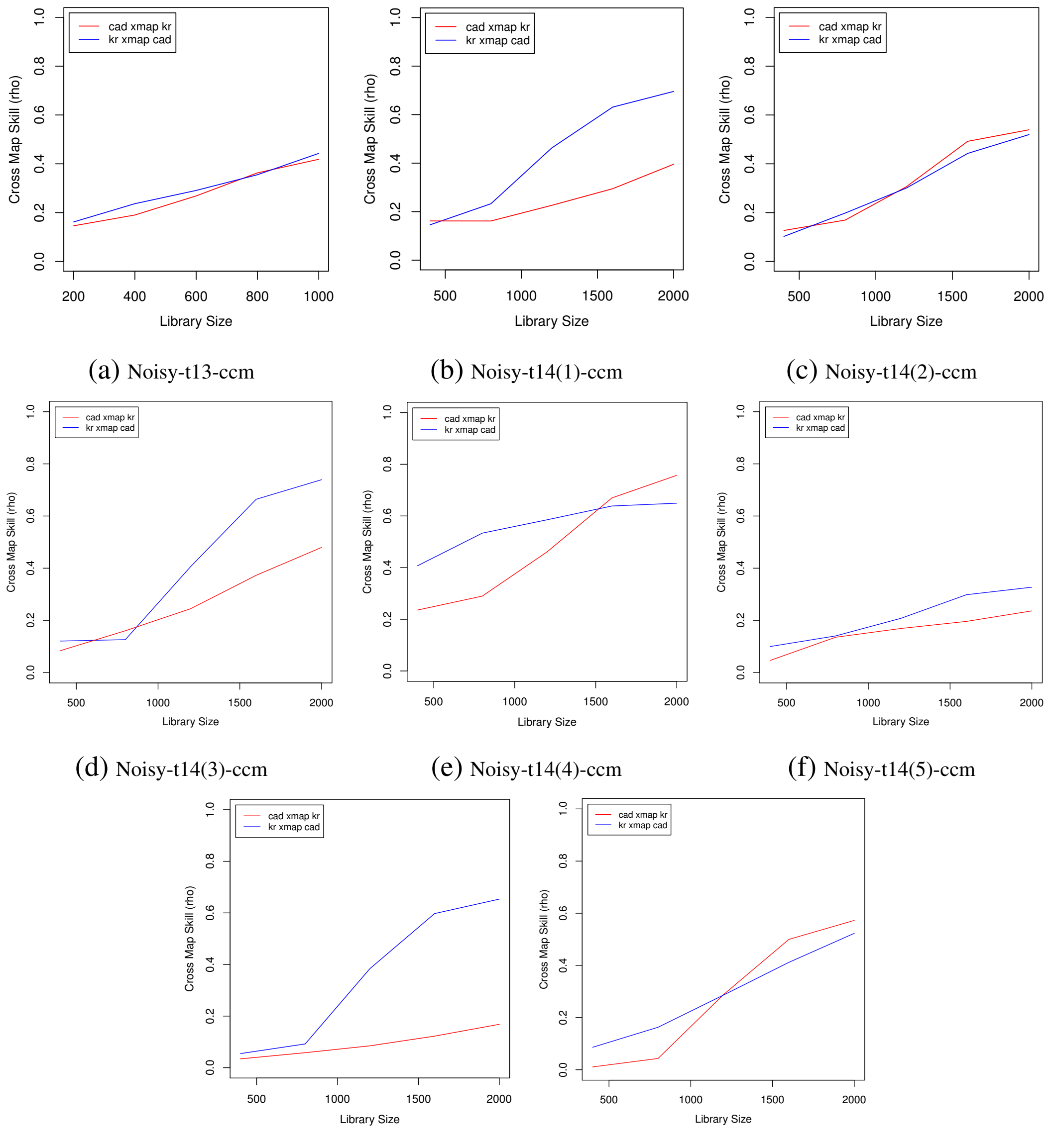
(e) Filtered-t14(4)-ccm

(f) Filtered-t14(5)-ccm

(g) Filtered-t14(6)-ccm

(h) Filtered-t14(7)-ccm

Fig. C.10 CCM test results for Bcd and Kr at cleavage cycles 10-14A (Filtered Series).



(a) Noisy-t13-ccm

(b) Noisy-t14(1)-ccm

(c) Noisy-t14(2)-ccm

(d) Noisy-t14(3)-ccm

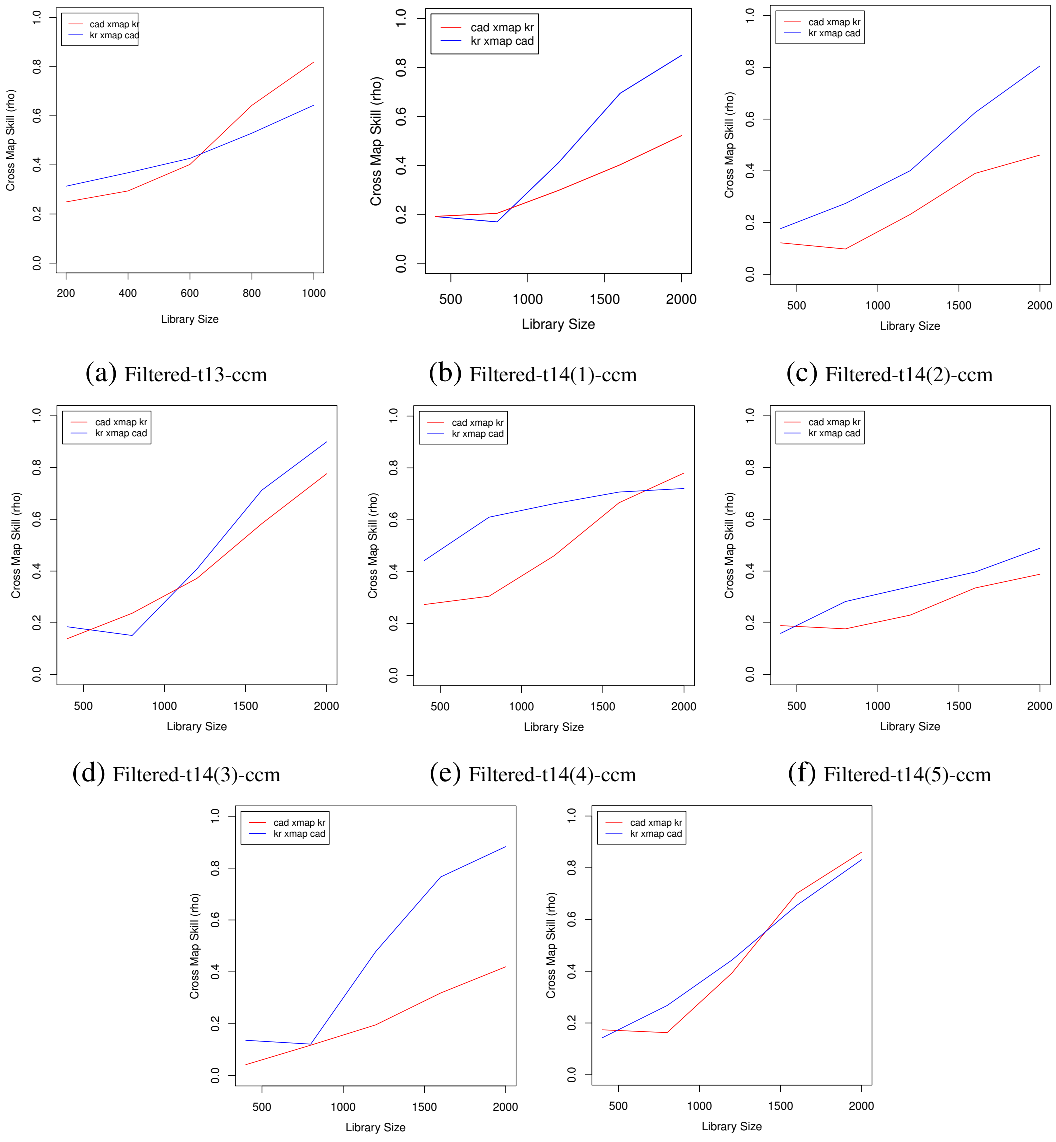
(e) Noisy-t14(4)-ccm

(f) Noisy-t14(5)-ccm

(g) Noisy-t14(6)-ccm

(h) Noisy-t14(7)-ccm

Fig. C.11 CCM test results for Cad and Kr at cleavage cycles 10-14A (Noisy Series).



(a) Filtered-t13-ccm

(b) Filtered-t14(1)-ccm

(c) Filtered-t14(2)-ccm

(d) Filtered-t14(3)-ccm

(e) Filtered-t14(4)-ccm

(f) Filtered-t14(5)-ccm

(g) Filtered-t14(6)-ccm

(h) Filtered-t14(7)-ccm

Fig. C.12 CCM test results for Cad and Kr at cleavage cycles 10-14A (Filtered Series).